OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

FACULTY OF
COMPUTER SCIENCE

Domain-Specific Forensic Process Models for Media Forensics - A Discussion based on the Example Application Domains of Face Morph Attack Detection, DeepFake Detection and Forensic Steganalysis

Kumulative Habilitationsschrift
zur Erlangung der Venia legendi für
das Fach Informatik

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von: Dr.-Ing. Christian Krätzer
geboren am 12. Dezember 1978 in Halberstadt, Deutschland

Gutachterinnen/Gutachter:

Prof. Dr. Jana Dittmann, Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

Prof. Dr. Stefan Katzenbeisser, Universität Passau, Passau, Germany

Prof. Dr. Edward J. Delp, Purdue University, West Lafayette, IN, United States of America

Magdeburg, Germany
06. November 2024

Abstract

This **cumulative habilitation treatise** is a moderated collection of previously published scientific work in the domain of media forensics. The newly written chapters 1 to 4 provide a framework to project the content of the individual original papers onto the common perspective of this habilitation project.

Media forensics is a young sub-discipline of digital forensics and focuses on the examination of media and multimedia objects in different contexts, ranging from general digital evidence to analyses of specific media types (e.g., video or image) to specific semantic analyses, such as facial identification or speaker recognition. Since this field is so wide, and media forensics in general is still considered by many as rather immature, there is a need to expand **domain-specific forensic process models**. The intention behind this modelling work is to create and maintain the trustworthy and validated forensic procedures required in up-to-date investigations that face various technical challenges in terms of significantly growing amounts of devices and data to be analysed, new file types and data formats, and an ever increasing number of potential data sources. In addition to these technical challenges, the forensic practitioners also face organisational issues that influence the evolution of forensic process models. Such issues include, among other things, new requirements for the reproducibility, auditability and contestability of forensic results that have been obtained using any form of investigation method based on machine learning.

In this treatise, twelve previously published papers on media forensics methods are aggregated into a wider perspective on the creation and expansion of domain-specific forensic process models. With face morphing attack detection, DeepFake detection and forensic steganalysis, these papers cover **three application domains in media forensics selected as illustrative examples**. For these application domains, the previously published work used in this treatise is re-iterated and put into perspective using five requirements that focus on the following considerations:

- Describing necessary conditions for using a media forensics method
- Standards for the evaluation of new methods
- Standardisation of investigation processes
- Causes and standards for the re-evaluation of methods
- Publication of methods and processes

The **results presented** and discussed on the basis of these requirements provide a common perspective on the conceptual and operational modelling work of the author and his co-authors in the three application domains mentioned above. This modelling work is based on established best practices (in the case of this treatise, the code of practice for IT forensics provided by the German Federal Office for Information Security (BSI) as well as selected European Network of Forensic Science Institutes (ENFSI) Best Practice Manuals) and expands these at various points by adding important aspects, such as a domain-adapted data model for media forensics as well as a proposal for the fine-grained operational modelling of media forensics (sub-)processes.
The descriptive summary of the modelling aspects is then followed by a structured set of conclusions and considerations on potential future work. The latter address aspects of future research and development as well as recommendations for improved operations.

In order to manage expectations, it must be clearly stated here that this work does not pretend to present a complete overview of the current state of the art in IT forensics, its sub-discipline of media forensics or the selected application domains of face morphing attack detection, DeepFake detection and forensic steganalysis. This treatise does not provide a guideline for the development of media forensics methods from academic research into industry-strength forensic tools. Furthermore, an academic publication such as this habilitation treatise cannot propose guidelines or even standard operational procedures for a forensics sub-domain. This is not the author's intention and would require standards published by the corresponding authority, e.g. the German BSI. What a publication like this treatise might achieve is to provide stakeholders like ENFSI (representing forensic practitioners) or policy makers in executive systems with arguments and recommendations for updating established best practices or policy documents.

Deutschsprachige Version des Abstracts

Diese **kumulative Habilitationsschrift** ist eine moderierte Sammlung von bereits veröffentlichten wissenschaftlichen Arbeiten im Kontext der Medienforensik. Die neu verfassten Kapitel 1 bis 4 bieten einen Rahmen, um die Inhalte der einzelnen Originalarbeiten auf die gemeinsame Perspektive dieses Habilitationsprojekts zu projizieren.

Die **Medienforensik** ist eine relativ junge Teildisziplin der digitalen Forensik und befasst sich mit der Untersuchung von Medien- und Multimediaobjekten in verschiedenen Zusammenhängen, die von allgemeinen digitalen Beweisen über Analysen spezifischer Medientypen (z. B. Video oder Bild) bis hin zu spezifischen semantischen Analysen wie Gesichtserkennung oder Sprechererkennung reichen. Da dieser Bereich so breit gefächert ist und die Medienforensik als Wissenschaftsdisziplin im Allgemeinen von vielen noch als recht unausgereift angesehen wird, besteht die Notwendigkeit, **domänenspezifische forensische Prozessmodelle** zu erarbeiten. Die Absicht hinter diesen Modellierungsarbeiten ist es, die vertrauenswürdigen und validierten forensischen Verfahren zu schaffen und zu erhalten, die in forensischen Untersuchungen benötigt werden. Diese müssen sich dabei verschiedenen technischen Herausforderungen in Bezug auf die wachsenden Mengen an zu analysierenden Geräten und Daten, neue Dateitypen und Datenformate und eine immer größer werdende Anzahl von potenziellen Datenquellen stellen. Neben diesen technischen Herausforderungen sehen sich Forensiker auch mit organisatorischen Fragen konfrontiert, welche die Entwicklung forensischer Prozessmodelle beeinflussen. Zu diesen Fragen gehören unter anderem neue Anforderungen an die Reproduzierbarkeit, Überprüfbarkeit und Anfechtbarkeit forensischer Ergebnisse, die mit einer auf maschinellem Lernen basierenden Untersuchungsmethode erzielt wurden.

In dieser kumulativen Habilitationsschrift werden zwölf bereits veröffentlichte wissenschaftliche Arbeiten zusammengeführt, um darauf aufbauend eine umfassendere Sichtweise auf die Erstellung bzw. Erweiterung von domänenspezifischen forensischen Prozessmodellen zu ermöglichen. Die wissenschaftlichen Arbeiten behandeln mit der Erkennung von Face-Morphing-Angriffen, der DeepFake-Erkennung und der forensischen Steganalyse **drei ausgewählte Anwendungsbereiche der Medienforensik**, die hier der Veranschaulichung dienen sollen. Für diese Anwendungsbereiche werden ausgewählte Inhalte der verwendeten Arbeiten auszugsweise zusammengefasst und anhand der folgenden fünf Anforderungen neu in Zusammenhang gesetzt:

- Beschreibung notwendiger Bedingungen für die Anwendung einer medienforensischen Methode
- Standards für die Bewertung neuer Methoden
- Standardisierung von Untersuchungsprozessen
- Auslöser und Standards für die Neubewertung von Methoden
- Veröffentlichung von Methoden und Verfahren

Die auf Grundlage dieser Anforderungen zusammengefassten und diskutierten **Ergebnisse** bieten eine gemeinsame Perspektive auf die konzeptionellen und operativen Modellierungsarbeiten des Autors und seiner Mitautoren in den drei oben genannten Anwendungsbereichen. Diese Modellierungsarbeiten basieren auf etablierten Best Practices (im Falle dieser kumulativen Habilitationsschrift der 'Leitfaden IT-Forensik' des Bundesamtes für Sicherheit in der Informationstechnik (BSI) sowie ausgewählte Best Practice Manuals des European Network of Forensic Science Institutes (ENFSI)) und erweitern diese an verschiedenen Stellen um wichtige Aspekte wie ein domänenangepasstes Datenmodell für die Medienforensik sowie einen Vorschlag zur feingranularen operativen Modellierung von medienforensischen (Teil-)Prozessen.
Auf die Zusammenfassung der Modellierungsaspekte folgt in der vorliegenden Arbeit eine strukturierte Reihe von Schlussfolgerungen sowie Überlegungen zu möglichen künftigen Arbeiten. Letztere umfassen potenzielle Aspekte für die künftige Forschung und Entwicklung sowie Empfehlungen für eine verbesserte Umsetzung im Rahmen der Entwicklung und der Ausführung von forensischen Prozessen.

Um die Erwartungen an die Inhalte dieser Arbeit zu steuern, muss an dieser Stelle erwähnt werden, dass diese nicht den Anspruch erhebt, einen vollständigen Überblick über den aktuellen Stand der Technik in der IT-Forensik, in der Teildisziplin der Medienforensik oder in den ausgewählten Anwendungsbereichen

der Erkennung von Face-Morphing-Angriffen, der DeepFake-Erkennung und der forensischen Steganalyse zu geben. Diese Habilitationsschrift bietet keinen Leitfaden für die komplette (Software-)Entwicklung der Ergebnisse akademischer Forschung zu medienforensischen Methoden hin zu industrietauglichen forensischen Werkzeugen. Auch kann eine akademische Publikation wie diese Habilitationsschrift keine Leitlinien oder gar verbindliche Standards für einen forensischen Teilbereich vorschlagen. Dies ist nicht die Absicht des Autors und würde normative Verfahren erfordern, die von einer entsprechenden Behörde, wie z. B. dem deutschen BSI, durchgeführt werden müssten. Eine Publikation wie diese Habilitationsschrift könnte höchstens Gremien wie dem ENFSI oder Entscheidungsträgern Argumente und Empfehlungen für die Aktualisierung von Best Practices oder Vorgaben liefern.

Erklärungen

Hiermit versichere ich, Christian Krätzer (geboren am 12. Dezember 1978 in Halberstadt, Deutschland), an Eides statt, dass die hier vorliegende Habilitationsschrift von mir selbstständig verfasst wurde.
Für die Verwendung bereits veröffentlichter Arbeiten im Rahmen dieser kumulativen Habilitationsschrift liegt das Einverständnis der jeweiligen Koautoren vor. Die Anteile („shares“ / „author contributions“) der jeweiligen Koautoren sowie genutzte Hilfsmittel sind angegeben.

Hiermit erkläre ich, dass mir die Habilitationsordnung der Fakultät für Informatik der Otto-von-Guericke-Universität Magdeburg vom 07. April 1999 bekannt ist und dass keine früheren Habilitationsversuche meinerseits vorliegen.

Declarations

I, Christian Krätzer (born 12. December 1978 in Halberstadt, Germany), hereby declare under penalty of perjury that I have composed the presented habilitation treatise independently on my own.
I have obtained the consent of the respective co-authors for the use of previously published work in the context of this cumulative habilitation. The shares (author contributions) of the respective co-authors and the resources used are indicated.

I hereby declare that I am familiar with the habilitation regulations of the Faculty of Computer Science at Otto-von-Guericke University Magdeburg dated 7. April 1999 and that I have not previously attempted to habilitate.

Magdeburg, 19. February 2024

Dr. Christian Krätzer

Contents

Acronyms

ABC	automated border control
AI	Artificial Intelligence
AIA	Artificial Intelligence Act
ANANAS	Project 'Anomalieerkennung zur Verhinderung von Angriffen auf gesichtsbildbasierte Authentifikationssysteme'
ASTM	American Society for Testing and Materials
AUC	area under curve
BKA	German Federal Criminal Police Office
BMBF	German Federal Ministry of Education and Research
BPM	Best Practice Manual
BSI	German Federal Office for Information Security
CERT	Computer Emergency Response Team
CISO	Chief Information Security Officer
CoC	chain of custody
CSI	crime scene investigation
CSM	cover-source mismatch
DA	Data Analysis
DCEA	Data Centric Examination Approach
DCNN	Deep Convolutional Neural Networks
DCT	discrete cosine transform
DD	Forensic Data Type
DF	digital forensics
DFU	Digital Forensic Unit
DG	Data Gathering
DI	Data Investigation
DIA	Digital Image Authentication
DFaaS	Digital-Forensics-as-a-Service
DO	Documentation phase
DPE	Data processing and evaluation
EACTDA	European Anti-Cybercrime Technology Development Association
EMID	Explicit means of intrusion detection
eMRTD	Electronic Machine Readable Travel Documents
ENF	electric network frequency

ENFSI	European Network of Forensic Science Institutes
EPE	Europol Platform for Experts
ETR	Europol Tool Repository
EU	European Union
EXIF	Exchangeable Image File Format
EWG	Expert Working Group
FAKE-ID	Project ‘Videoanalyse mit Hilfe künstlicher Intelligenz zur Detektion von falschen und manipulierten Identitäten’
FAR	false alarm rate
FAVIAU	Forensic Audio, Video, and Image Analysis Unit of the FBI
FBI	Federal Bureau of Investigation
FIT	Forensic Examination of Digital Technology
FKZ	Förderkennzeichen
FMA	face morphing attack
FMR	false missing rate
FRE	Federal Rules of Evidence
FRE702	Federal Rules of Evidence - rule 702
FSA	Digital Audio Authenticity Analysis
ESR	early-stage researcher
FS	file system
FSA	Digital Audio Authenticity Analysis
FSAAWG	Forensic Speech and Audio Analysis Working Group
FSR	Forensic Science Regulator
GDPR	General Data Protection Regulation
ICAO	International Civil Aviation Organization
ICS	industrial control systems
ICT	information and communications technology
IEC	International Electrotechnical Commission
ILAC	International Organisation for Accreditation Bodies
INTERPOL	International Criminal Police Organization
IOCE	International Organisation on Computer Evidence
ID	identity
ISMS	information security management systems
ISO	International Standards Organization

IT	information technology
ITA	IT application
JPEG	Joint Photographic Experts Group
LEA	law enforcement agencies
LR	likelihood ratio
LSB	least significant bit
MFDT	Media Forensic Data Type
ML	machine learning
NFI	Netherlands Forensic Institute
NIJ	National Institute of Justice
NIST	National Institute of Standards and Technology
NPCC	National Police Chiefs' Council
OP	Operational Preparation
OSAC	Organization of Scientific Area Committees for Forensic Science
OS	operating system
OSI	Open System Interconnection
PhD	Doctor of Philosophy
PRNU	Photo Response Non-Uniformity
R-UE/EU-R	RESTREINT UE/EU RESTRICTED
RFC	Request for Comments
SAB	security advisory board
SDK	Software Development Kit
SMG	Scaling of methods for evidence gathering
SOP	standard operational procedures
SP	Strategic Preparation
SWG	Scientific Working Group
SWGDE	Scientific Working Group on Digital Evidence
SWGFAST	Scientific Working Group on Friction Ridge Analysis, Study, and Technology
TRL	Technology Readiness Level
UK	United Kingdom of Great Britain and Northern Ireland
UKAS	United Kingdom Accreditation Service
UNCOVER	Project 'Development of an efficient steganalysis framework for uncovering hidden data in digital media'
UNICRI	United Nations Interregional Crime and Justice Research Institute
US	United States
ZITiS	Central Office for Information Technology in the Security Sector

List of Figures

1 Introduction

This cumulative habilitation treatise[1] aggregates research conducted on process modelling in media forensics between 2013 and 2024. In Chapters 5 and following, the original papers ('feeder paper(s)'[2]) by the author used in this moderated collection are presented as they have originally been published with co-authors. The only alterations made are the addition of continuous page numbering as well as the insertion of copyright notices (where required).
Chapters 1 to 4 provide a framework to project the content of the individual original papers onto the common perspective of this habilitation project.

1.1 Motivation

Digital forensics (DF) is the sub-discipline of the forensic[3] sciences that focuses on the investigation of digital assets. In contrast to many other sub-disciplines of forensics, like forensic pathology, it is rather young, having gained relevance only with the growing use of information technology (IT) at the end of the 20th century. A very illustrative statement on the increasing relevance of digital forensics (DF) science and techniques is given in [Vaughan20]:

> "*digital forensics (DF) science - examining digital evidence to support investigations and prosecutions - was once niche but is now very much mainstream. Over 90% of all crime is recognised as having a digital element, and society's accelerating use of technology means the critical role DF science plays will only grow.*"

In its 'Digital Forensic Science Strategy' [Vaughan20] of 2020, the National Police Chiefs' Council (NPCC) of the United Kingdom of Great Britain and Northern Ireland (UK) also emphasises why it so important to invest into research and development activities in this field:

> "*Digital Forensic Science sits at the heart of delivering justice in the 21st century, spanning the entire criminal justice system, from crime scene to courtroom. It shapes policy, offers a range of capabilities that better enable us to counter new and emerging threats, and is central to achieving our shared outcomes around reducing crime and increasing public safety.*"

Within digital forensics, the field of **media forensics**, which is the focus of this habilitation project, is an even younger research domain and has to be clearly distinguished from other DF sub-categories like computer forensics (a.k.a. IT forensics). In [Böhme09], the authors argue that the difference between media and IT forensics lies in particular in "*the reliability of the extracted probative facts: it is harder to forge media data undetectably than to manipulate other digital evidence*".

[1] A cumulative habilitation treatise does not present research findings as a monograph, but in the form of a moderated collection of previously published scientific work (usually journal, conference and workshop papers).

[2] Papers from the moderated collection of already published scientific work that constitute this cumulative habilitation treatise are here referred to as feeder paper.

[3] The Merriam-Webster dictionary defines the adjective 'forensic' as "*belonging to, used in, or suitable to courts of judicature or to public discussion and debate*", and the noun 'forensics' as: "*the application of scientific knowledge to legal problems.*" [Pollitt19] expands the latter definition by adding: "*This includes investigative activities performed in support of legal problems, as well as development of testimony for use in courts of law.*"

At the core of media forensics lies the examination of media and multimedia objects in different contexts, ranging from general digital evidence to analyses of specific media types (e.g., video or image) to specific semantic analyses, such as facial identification or speaker recognition. The document 'A Framework for Harmonizing Forensic Science Practices and Digital/Multimedia Evidence' [Pollitt19], published by the Organization of Scientific Area Committees for Forensic Science (OSAC) in 2019, summarises the focus of media forensics as follows:

> "*In practice, digital/multimedia evidence serves investigative, procedural, and scientific functions, and the outcomes of these multiple modalities are synthesized into expert opinions and conclusions.*"

The interesting part here is that the OSAC experts point out that such forensic science is "*not limited to legal problems in civil and criminal justice systems (courtroom contexts)*" [Pollitt19]. This reflects very well the fact that media forensics results are not only relevant in court cases but also in many other contexts of public debate. An example for this would be a reliable media forensics 'fake news' detector which would not only help police investigators and prosecutors in their work but also every media outlet or, in an even wider context, everyone relying on news feeds.

As a result of these discussions, three different application contexts are usually discussed for media forensic methods: intended courtroom use, intelligence for police investigations and public use.

The first of these three is the typical forensics application context and also the most prominent one, usually demanding the highest degree of maturity from methods and procedures. For this, one ENFSI BPM [ENFSI15] summarises the requirements for the practitioner and the application of forensic techniques as follows:

> "*Good practice requires an understanding of both the processes*[4] *selected to perform a forensic examination of digital technology, and an understanding of the expected knowledge of the intended recipient of any report generated.*"

The second application context, addressing especially the investigative activities of law enforcement agencies (LEA), will also require a high degree of maturity, because the result might be used to scale up an investigation, e.g., to obtain an additional warrant from a judge. The third application context requires the lowest degree of maturity, but methods used in this category still need to be trusted by the user, who might use them to confirm a suspicion (e.g., that an image was manipulated) and then use this outcome to file a complaint with the police.

To achieve the necessary maturity for the first (and to some extent also the second) application context, there is a strong need to standardise and formalise the forensic processes involved in all DF sub-disciplines to enable them to withstand the scrutiny of the other stakeholders in legal proceedings, first and foremost the judges (who have to act as 'gatekeepers' for their trials and decide upon the admissibility or inadmissibility of evidence), but also the prosecutors and the defence counsel. Section 2.3.1 provides a short overview of quality criteria and corresponding assurance methods. In summary, these state that creating and maintaining trustworthy and validated forensic procedures requires (among other things) detailed **forensic process modelling** and documenting. The need for process modelling increases even further if more than one forensic laboratory is involved in an investigation and the results have to be exchanged in a way that also enables reproducibility, auditability and contestability.

It is widely assumed that the necessary process models exist in most countries, at least on a national level. However, the following very open statement, made in 2020 by the NPCC of the UK, presents a different picture:

> "*Digital forensic services across UK policing are fragmented and disjointed. At present there are 40* [Digital Forensic Units[5]] *serving the 43 territorial forces. Collaboration between*

[4] [ENFSI15]: "*A forensic process requires that analysts understand and report the known limitations of their processes and specific tools selected using proven scientific methods and practice. In other words, they should not use; or incorrectly assert, assumptions if they do not understand the operation and/or limitations of the system used.*"

[5] A Digital Forensic Unit (DFU) is defined in [Vaughan20] as "*a department within a police force dedicated to digital forensic analysis of devices and/or data, staffed by practitioners or specialists. In some forces, may be known as a digital forensic laboratory.*"

> *forces, either formal or informal, is limited, and a 'typical force' DFU does not do work for others, nor does it have the capacity to do so. Individual units develop their own methods, procure and deploy their own hardware and software and manage their individual quality accreditations in line with the* [Forensic Science Regulators[6]] *requirements. All of this involves substantial duplication of effort and inevitable waste of resources. Despite this, the core functions of force DFUs have striking commonalities, and there is clear scope to standardise methods and share support services between units. To improve operational efficiency, we need to 'industrialise' and streamline this fragmented landscape, redesigning DF from the ground up.*" [Vaughan20]

This statement, coming from the highest levels of authority in a large national police force in Europe, does not represent an outlier but is (in the author's opinion) to some extend symptomatic for many countries world-wide. In Chapter 2, the corresponding situations in the United States of America and in Germany are briefly reflected, confirming similar problems as in the UK and thus showing a significant need for work on the modelling of (media) forensic processes.

In this context, this cumulative habilitation treatise addresses selected issues of such modelling work by aggregating published work by the author on deriving domain-specific forensic process models for media forensics in three example application domains.

1.2 Research Project Contexts Represented in this Habilitation Treatise

In the papers contributing to this cumulative habilitation treatise, the three application domains of face morphing attack (FMA) detection (for digital images), DeepFake detection (for digital videos) and forensic steganalysis (for digital images) are used as examples.

The **face morphing attacks (FMAs)** under consideration are summarised in feeder paper **[Neubert19]** as follows:

> "*In 2014 Ferrara et al. present an identity theft scheme for those scenarios in* [Ferrara14]. *They describe an approach allowing two or more persons to pass a face image based authentification scenario with only one artificially weakened Photo-ID template. For the presented attack, a so called face morphing is created, which melts two or more face images of different persons. This morphed face image is used for the document creation performed by a corresponding authority. This Photo-ID document is able to successfully pass all subjective and biometric checks in a border control scenario.*"

Figure 1 in feeder paper **[Neubert19]** (see page 99 in Chapter 7) illustrates the FMA.

The application scenario of FMA detection represents the author's involvement in the German nationally-funded project ANANAS ('Anomaly detection to prevent attacks on facial image-based authentication systems', German title: '*Anomalieerkennung zur Verhinderung von Angriffen auf gesichtsbildbasierte Authentifikationssysteme*'; funded in part by the German Federal Ministry of Education and Research (BMBF) under the contract no. FKZ: 16KIS0509K; `https://www.forschung-it-sicherheit-kommunikationssysteme.de/projekte/ananas`).

The focus of the author's involvement in this project, which ran from 2016 until 2019, lay in conceptual modelling, supporting the empirical work of the research group leader, another post-doc colleague and a PhD student of the same research group as well as academic and industry partners at four different partner institutions and companies within the project consortium. The corresponding results on conceptual modelling are aggregated in this cumulative habilitation treatise.

[6] A Forensic Science Regulator (FSR) is defined in [Vaughan20] as: "*A government appointee responsible for ensuring that the provision of forensic science services across the criminal justice system is subject to an appropriate regime of scientific quality standards.*"

In the context of the application scenario of **DeepFake detection**, the term 'DeepFake' is defined in **[Siegel21]** as follows:

> "*DeepFakes (a neologism combining the terms 'deep learning' and 'fake') are synthetic videos (or images) in which a person's face (and optionally also voice) is replaced with someone else's likeness using deep learning technologies.*"

An more recent, comprehensive description is presented in [Tahraoui23]:

> "*The term deepfake commonly refers to visual content that is artificially generated, manipulated, or distorted by using artificial intelligence tools to alter or replace a person or selected attributes of that person in the content. That content can be not only visual (i.e., pictures and videos) but also aural (i.e., sounds and noises).*"

This novel threat emerged with the widespread availability of neural network-driven technologies from around 2017. It first received news coverage due to its use in pornographic contexts using face-swaps, where primarily women became victims of targeted defamation. Later, it was associated with manipulated video footage of politicians (e.g., former US president Barack H. Obama, and Nancy Pelosi during her time as Speaker of the United States House of Representatives) used to spread misinformation and potentially influence political opinion. In this context, DeepFakes are discussed in [Tahraoui23] as a potential threat to digital sovereignty that has been recognised as such, with the paper providing a summary of the European Union's regulatory reactions to this threat.
It is important to point out that DeepFakes can have such malicious as well as non-malicious application scenarios.[7] These are discussed in the feeder paper **[Kraetzer22]**, included as Chapter 12 of this cumulative habilitation treatise (see page 191).

The application scenario of DeepFake detection represents the author's involvement in the German nationally-funded project FAKE-ID ('Video analysis using artificial intelligence to detect false and manipulated identities', German title: '*Videoanalyse mit Hilfe künstlicher Intelligenz zur Detektion von falschen und manipulierten Identitäten*'; funded in part by the German Federal Ministry of Education and Research (BMBF) under the contract no. FKZ: 13N15736; `https://www.sifo.de/sifo/shareddocs/Downloads/files/projektumriss_fake_id.html?nn=248456`).
The focus of the author's involvement in FAKE-ID (duration: 2020-2024) lies in operational modelling, supporting the empirical work of the research group leader, two PhD students from the same research group as well as academic and industry partners at five different partner institutions and companies within the project consortium. In the context of this ongoing research effort, various abstractions of forensic processes were presented in published work, refining an approach for operational modelling based on the German BSI 'Leitfaden IT-Forensik' [BSI11] and other sources extending these German national guidelines for IT-forensic investigations, such as the Data Centric Examination Approach (DCEA) presented in [Kiltz20].

[7] Both, black- and white-hat application scenarios, and one recent trend how to counter the issue of manipulations are well summarised in [Rathgeb22] by the following statement:

> "*Face manipulation brings an array of complex legal issues. There is no comprehensive legislation on the use of manipulated images, yet several aspects are already regulated in various countries. It should hence not surprise that the development of new manipulation technology and the detection thereof also leads to new issues and questions from a legal perspective which deserve further research. If it is used to mislead, manipulated images can cause significant harm* [...] *In some countries, altered (body) images used for commercial purposes (such as the fashion industry) need to be labelled. More generally, legislative proposals in several countries try to tackle the transparency issue by imposing an obligation to inform users that they interact with AI-generated content (such as DeepFakes).*"

But this holds true only for white-hat application of methods like DeepFakes. No (criminal or other) threat actor will adhere to such an obligation when spreading fake news or other media-related manipulations. As a consequence, entities such as news agencies strongly relying on media objects submitted from external sources would also require mature manipulation detection mechanisms that would have to be integrated into their already established source (material) verification routines. The exact extend and scope of such analysis methods and 'filters', their transparency and fairness, as well as their potential impact to public and politic debates are currently a hot debate topic, especially in Europe (see for example [Øe21] for the discussion of free speech implications of Article 17 (regulating upload filters) of the EU 'Directive on copyright and related rights in the Digital Single Market' as adopted in 2020).

In the context of this cumulative habilitation treatise, selected other application scenarios were considered in addition to the two main application scenarios of face morphing attack detection and DeepFake detection. From these additional scenarios, results for the media forensics sub-domain of **forensic steganalysis** for media objects are included here. The results in this sub-domain are published in the context of the EU-funded project UNCOVER ('Development of an efficient steganalysis framework for uncovering hidden data in digital media'; funded by the European Union's Horizon 2020 research and innovation program under grant agreement no. 101021687; https://cordis.europa.eu/project/id/101021687).
The focus of the author's involvement in UNCOVER (duration: 2021-2024) lies in conceptual modelling, supporting the work of the research group leader and other colleagues from the same research group as well as 22 external partner institutions (including several European forensic institutions and LEA) on the topic of forensic steganalysis.
Since parts of the project have been classified by the EU as RESTREINT UE/EU RESTRICTED (R-UE/EU-R), publication activity on the results of this project is significantly reduced in comparison to other research projects with a similar number of academic partners.

1.3 Problem Outline for this Cumulative Habilitation Treatise

The problem outline for this habilitation treatise is presented in the feeder paper **[Kraetzer15a]**. This paper is included as Chapter 5 on page 75 of this document. Briefly summarised, it derives quality aspects for media forensics approaches from the Federal Rules of Evidence (FRE) and the so-called Daubert criteria. These rules and criteria provide a set of guidelines defined for federal-level judicial matters in the US. They were codified in the 1990s by the United States of America Supreme Court based on the decisions in the 'Daubert v. Merrell Dow Pharmaceuticals' (or short 'Daubert') court cases from 1993 on. The Daubert standard is widely regarded as a good (or even the best established) set of recommendations for judges on how to evaluate the usefulness of scientific (as well as non-scientific) expert testimony. Some details of these recommendations are presented in Section 2.3.1.
From the synopsis presented in **[Kraetzer15a]** (see Chapter 5, pages 80 and 81), the following relevant requirements are derived here for the problem outline of this treatise (the wording is adapted from **[Kraetzer15a]** to the terminology used in this treatise, the items are reordered, and two of the original requirements are combined in REQ2):

1. **REQ1 'Describing necessary conditions for using a method':** For every media forensics approach, it is important to clearly specify to which type of media and which type of content it can be applied in which context. These specifications have to be communicated clearly by the researchers and developers of the method to the corresponding forensic practitioners intending to use the method.

2. **REQ2 'Evaluation of new methods':** Very high standards have to be set for the evaluation of new (media) forensic principles and methods as well as for the documentation of the evaluation methodology and the evaluation results of the corresponding proficiency tests. Evaluation setups with a statistically significant number of samples and relevant domain coverage should be used, wherever possible, to establish the exact error rates of a forensic technique as precisely as they can be measured or estimated.

3. **REQ3 'Standardisation of investigation processes':** The successful application of forensic techniques in most contexts requires an effort to standardise the investigation process into which the method is to be integrated.

4. **REQ4 'Re-evaluation of methods':** The suitability of an established media forensics method has to re-evaluated regularly, as well as on specific occasions. Media forensics approaches will face changes in investigation contexts and investigated content (e.g., by new file formats or encoding methods gaining relevance) as well as the emergence of targeted countermeasures (i.e., anti-forensics). In case of changes in the evaluation outcomes, the affected investigation processes have to undergo a renewed standardisation.

5. **REQ5 'Publication of methods and processes':** Scientific results for media forensics techniques must be published and openly discussed with the corresponding community. They should also be communicated to a wider audience – preferably the general public – to counter the 'CSI effect' in courts.

These requirements are in the following used to structure and project the relevant contributions from the feeder papers onto the common perspective of this habilitation treatise.

1.4 Feeder Papers for this Cumulative Habilitation Treatise

The following twelve papers are included in this cumulative habilitation treatise as feeder papers. For each of these publications, a table describing the projection onto the requirements REQ1 to REQ5 introduced in Section 1.3 is included, and the main contribution towards domain-specific forensic process models for media forensics is briefly summarised. As is common practice for a cumulative habilitation, the feeder papers are included exactly as they were originally published (i.e., with the original, unedited layout and content).
A complete list of all publications by the author published during the habilitation project (i.e., since the defence of the PhD thesis) is given in Appendix A (page 253 ff.).

[Kraetzer15a] Christian Kraetzer, Jana Dittmann: *Considerations on the benchmarking of media forensics*. Proceedings of the 23rd European Signal Processing Conference (EUSIPCO 2015), Nice, France, August 31 - September 4, 2015, IEEE, pp. 61-65, 2015.
https://doi.org/10.1109/EUSIPCO.2015.7362345

REQ1: 'Necessary conditions for using a method'	The assessment of the benchmarking of media forensics methods performed in **[Kraetzer15a]** is the basis for REQ1 to REQ5 as used in this habilitation project.
REQ2: 'Evaluation of new methods'	
REQ3: 'Standardisation of investigation processes'	
REQ4: 'Re-evaluation of methods'	
REQ5: 'Publication of methods and processes'	

Main contribution of [Kraetzer15a] to domain-specific forensic process models for media forensics: Deduction of the problem outline for this habilitation treatise from FRE and Daubert criteria as well as a use case report on a media forensics procedure in court proceedings provided as an example.

[Kraetzer17] Christian Kraetzer, Andrey Makrushin, Tom Neubert, Mario Hildebrandt, Jana Dittmann: *Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing*. Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec 2017, Philadelphia, PA, USA, June 20-22, 2017, pp. 21-32, 2017.
https://dl.acm.org/doi/10.1145/3082031.3083244

REQ1: 'Necessary conditions for using a method'	**[Kraetzer17]** models the investigation contexts for FMA detection in photo-ID documents. In addition, a descriptive image editing history model is proposed as a method of the attack modelling.
REQ2: 'Evaluation of new methods'	The systematic application of the descriptive image editing history model to an FMA detection pipeline is illustrated. Testing with malicious and non-malicious image modifications is performed to provide more meaningful error rate estimates.
REQ3: 'Standardisation of investigation processes'	Context and attacker model components are defined as building blocks for standardisation work.
REQ4: 'Re-evaluation of methods'	Not addressed in this paper
REQ5: 'Publication of methods and processes'	Systematic descriptions of image-editing histories and a description of the feature space used for detection are provided.

Main contribution of [Kraetzer17] to domain-specific forensic process models for media forensics: First steps for conceptual modelling of media generation processes and corresponding source and attack models for the application example of FMA detection.

[Neubert19] Tom Neubert, **Christian Kraetzer**, Jana Dittmann: *A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on eMRTD*. Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'19), July 2019, pp. 95–100. 2019.
https://doi.org/10.1145/3335203.3335721

REQ1: 'Necessary conditions for using a method'	**[Neubert19]** discusses the pre-processing methods required to increase the robustness of an FMA detection approach adapted to an application scenario.
REQ2: 'Evaluation of new methods'	In addition to the empirical estimation of detection performances, the paper evaluates several potential influencing factors for the obtained error rates (here: three different morph generation pipelines, neutral and smiling face expressions, and impact of the file format used).
REQ3: 'Standardisation of investigation processes'	The paper uses (among other data sets) the reference data from the IHMMSEC'19 special session 'fake or real' to provide empirical results comparable to other work.
REQ4: 'Re-evaluation of methods'	Methods from **[Kraetzer17]** are re-evaluated for a different application context (here: passport-scaled images for eMRTD).
REQ5: 'Publication of methods and processes'	A description of the feature space extensions performed and the corresponding feature sub-space performance evaluations are provided.

Main contribution of [Neubert19] to domain-specific forensic process models for media forensics: As part of the conceptual modelling for FMA detection, an attack model extension is discussed and combined with work to increase the robustness of an FMA detection approach adapted to the application scenario of Electronic Machine Readable Travel Documents (eMRTD) verification. The differentiated attacks and the corresponding modelling of attack detection are used, among other research goals, to determine the impact of different morph generation pipelines on the detection performance using two different feature (sub-)spaces.

[Neubert18a] Tom Neubert, **Christian Kraetzer** and Jana Dittmann: *Reducing the False Alarm Rate for Face Morph Detection by a Morph Pipeline Footprint Detector*. Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 1002-1006, 2018.
https://doi.org/10.23919/EUSIPCO.2018.8553067

REQ1: 'Necessary conditions for using a method'	**[Neubert18a]** provides a detailed image processing and feature extraction pipeline description.
REQ2: 'Evaluation of new methods'	A three-stage detection and verification sequence is introduced to reduce the false alarm rate (FAR) of the complete FMA detection process. The impact of this conceptual extension to FMA detection pipelines on the false missing rate (FMR) is discussed.
REQ3: 'Standardisation of investigation processes'	Not addressed in this paper
REQ4: 'Re-evaluation of methods'	A set of three binary FMA detectors from previous publications is complemented by a morph pipeline footprint detector and a validation step to provide a concept for context adaptation in this approach. This makes it possible to transfer the addressed multi-class problem to a sequence of media forensics analysis methods ending in context-adapted 2-class decisions with a lowered FAR.
REQ5: 'Publication of methods and processes'	A systematic description of the three-stage detection and verification sequence is provided.

Main contribution of [Neubert18a] to domain-specific forensic process models for media forensics: The conceptual pipeline extension makes it possible to gradually generate and use knowledge about the media object under investigation and can improve the overall performance of a media forensics investigation (in this example by lowering the overall FAR).

[Kraetzer21] Christian Kraetzer, Andrey Makrushin, Jana Dittmann, Mario Hildebrandt: *Potential advantages and limitations of using information fusion in media forensics—a discussion on the example of detecting face morphing attacks*. EURASIP Journal on Information Security 2021, 9 (2021).
https://doi.org/10.1186/s13635-021-00123-4

REQ1: 'Necessary conditions for using a method'	**[Kraetzer21]** presents an in-depth discussion of the necessary conditions for as well as potential advantages and disadvantages of using the concept of information fusion in the context of media forensics. The discussion is illustrated using the application scenario of digital image authenticity and integrity analysis for FMA detection in two evaluated application contexts.
REQ2: 'Evaluation of new methods'	The paper presents empirical evaluations with a set of FMA detectors applying different fusion methods and fusion ensemble composition strategies. These evaluations illustrate why forensic practitioners are usually reluctant to rely on information fusion approaches. The results of the experiments performed show a decrease in the overall detection performance and at the same time an increased problem of explainability.
REQ3: 'Standardisation of investigation processes'	The need for benchmarking and proficiency testing for media forensics methods, especially in fusion setups (incl. the interrelation of methods and their results), are discussed.
REQ4: 'Re-evaluation of methods'	The additional issues of the diversity required of the detectors used for fusion as well as the need for deriving suitable fusion weights (in case a weighted fusion approach is intended to be used) enforce additional measures in the evaluation and re-evaluation of methods for a media forensics setup based on information fusion.
REQ5: 'Publication of methods and processes'	A systematic description of the used fusion methods and fusion ensemble composition strategies and a discussion of the different detection and fusion trends for the two evaluated application contexts is provided.

Main contribution of [Kraetzer21] to domain-specific forensic process models for media forensics: As a general contribution, **[Kraetzer21]** illustrates why the naive assumption that including fusion in the conceptual detection model will automatically make the detection more reliable can fail in practice, i.e., why fusion sometimes behaves differently in field application than in the lab. In addition, the conceptual constraints and limitations of the application of fusion are discussed, and its impact on (media) forensics is reflected upon.

[Siegel21] Dennis Siegel, **Christian Kraetzer**, Stefan Seidlitz, Jana Dittmann: *Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features*. Journal of Imaging, vol. 7(7), Special Issue Image and Video Forensics, ISSN 2313-433X, 2021.
https://doi.org/10.3390/jimaging7070108

REQ1: 'Necessary conditions for using a method'	**[Siegel21]** models the investigation contexts for (video) DeepFake detection.
REQ2: 'Evaluation of new methods'	The work focuses on three sets of hand-crafted features and three different fusion strategies to implement DeepFake detection. The results obtained with third-party reference databases show performances similar (peak area under curve (AUC) > 0.95) to those of methods using features learned by neural networks.
REQ3: 'Standardisation of investigation processes'	First steps are taken towards a projection onto a pre-existing, data-centric examination approach for conceptual and operational forensics process modelling. In addition, the evaluation relies on third-party reference databases to obtain comparable detection results.
REQ4: 'Re-evaluation of methods'	Not addressed in this paper

REQ5: ‘Publication of methods and processes’	Detailed descriptions of the modelling background (esp. the data-centric examination approach) and the implementation of the individual detectors and the fusion operators used are provided.

Main contribution of [Siegel21] to domain-specific forensic process models for media forensics: The Data Centric Examination Approach (DCEA) from [Kiltz20] (derived from the forensic process model for IT forensics published in 2011 by the German Federal Office for Information Security (BSI)) is used as a starting point for conceptual and operational modelling. It is shown that while it does not fit the needs of media forensics analyses perfectly, this starting point provides a good basis for a domain-specific adaptation for media forensics.

[Siegel22] Dennis Siegel, **Christian Krätzer**, Stefan Seidlitz, Jana Dittmann: *Forensic data model for artificial intelligence based media forensics - Illustrated on the example of DeepFake detection*. Proc. Electronic Imaging. Springfield, VA: Society for Imaging Sciences and Technology, Vol. 34 (2022), 2022.
https://doi.org/10.2352/EI.2022.34.4.MWSF-324

REQ1: ‘Necessary conditions for using a method’	**[Siegel22]** identifies typical data streams within a media forensics process, followed by a differentiation of the data streams into data types. This process follows the established best practices of the DCEA for domain adaptation of forensic process models.
REQ2: ‘Evaluation of new methods’	The new operational modelling components introduced are applied to the design and description of the empirical experiments in the paper to illustrate their suitability.
REQ3: ‘Standardisation of investigation processes’	In addition to the domain-adapted data model, three additional components of operational modelling are introduced: the operator (an atomar processing or analysis operation of a forensic process with well-defined input and output connectors), an operational concept for modelling a forensic (sub-)process as connected operators, and operational modelling of interconnection aspects of media forensics (sub-)processes. The considerations are separated into (sub-)process preparation (templating) in the DCEA phase of Strategic Preparation (SP) and usage (instantiation) in the phase of Operational Preparation (OP).
REQ4: ‘Re-evaluation of methods’	The conceptual model presented in **[Siegel21]** for a fusion-based DeepFake detection approach is re-structured and expanded using the newly introduced operational modelling components.
REQ5: ‘Publication of methods and processes’	An in-depth discussion on the new operational modelling components, esp. the derivation of the new media forensics data model from a pre-existing DCEA data model, is provided.

Main contribution of [Siegel22] to domain-specific forensic process models for media forensics: The conceptional model for a fusion-based DeepFake detection pipeline and the initial steps for operational modelling from **[Siegel21]** are used to derive a domain-adapted data model for media forensics. In addition, three essential operational modelling components are introduced with the operator, structured operator-based descriptions of media forensics (sub-)processes, and the connection of templating and instantiating of forensic processes.

[Kraetzer22] **Christian Kraetzer**, Dennis Siegel, Stefan Seidlitz, Jana Dittmann: *Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection*. MDPI Sensors 2022 (Special Issue Detecting and Preventing Deepfake Attacks), 22(9), 3137; 2022.
https://doi.org/10.3390/s22093137

REQ1: 'Necessary conditions for using a method'	**[Kraetzer22]** expands the conceptual modelling from **[Siegel22]** with two new feature spaces and semantically analyses the blinking behaviour in a video. In addition, the integration of these new features into a fusion-based DeepFake detection framework is discussed, evaluating weighted and unweighted fusion strategies.
REQ2: 'Evaluation of new methods'	The new components of operational modelling discussed in this paper are used for describing the empirical experiments in the paper. Based on this, the two new feature spaces are integrated into the framework and then evaluated. Benchmarking metrics (here, Cohens Kappa) are discussed for this evaluation.
REQ3: 'Standardisation of investigation processes'	The components of operational modelling from **[Siegel22]** are reconsidered and expanded. One of these expansions is the inclusion of an optional feedback loop from the Documentation phase (DO) to the SP phase of the phase-driven model derived from the DCEA. The other relevant expansion is a benchmarking-driven approach for fusion weight determination.
REQ4: 'Re-evaluation of methods'	The set of detectors and the components of operational modelling from **[Siegel22]** are revisited, expanded and re-evaluated.
REQ5: 'Publication of methods and processes'	Detailed description of the newly included feature spaces and the benchmarking-driven approach for fusion weight determination as well as an expanded description of the components of operational modelling are provided.

Main contribution of [Kraetzer22] to domain-specific forensic process models for media forensics: This special-issue journal paper significantly expands the descriptions of the components of operational modelling for DeepFake detection originally introduced in **[Siegel22]**. In addition, it provides new aspects of operational modelling (esp. the feedback loop from the DO to the SP phase) as well as conceptual modelling (here, esp. the benchmarking-driven approach for fusion weight determination for the DeepFake detection framework).

[Siegel23b] Dennis Siegel, **Christian Kraetzer**, Stefan Seidlitz, Jana Dittmann: *Forensic data model for artificial intelligence based media forensics - Illustrated on the example of DeepFake detection*. Electronic Imaging, Vol. 34(4), 2022. doi:10.2352/EI.2022.34.4.MWSF-324.
https://library.imaging.org/ei/articles/34/4/MWSF-324

REQ1: 'Necessary conditions for using a method'	Motivated by the EU GDPR, **[Siegel23b]** empirically evaluates the trade-off between detection performance and data minimisation for DeepFake detection.
REQ2: 'Evaluation of new methods'	The conceptual approach of reduction or minimisation of biometric data (as motivated by the EU GDPR) has an impact on DeepFake detection accuracy.
REQ3: 'Standardisation of investigation processes'	The relevance here lies in the identification of potentially relevant GDPR aspects as part of the conceptual model.
REQ4: 'Re-evaluation of methods'	In this paper, a detector from **[Siegel21]** is re-evaluated after data minimisation. It is shown that the accuracy achieved is not significantly impaired.
REQ5: 'Publication of methods and processes'	A description of the method used to estimate the impact of the video duration on the DeepFake detection accuracy achieved is provided.

Main contribution of [Siegel23b] to domain-specific forensic process models for media forensics: With the EU General Data Protection Regulation (EU GDPR) and the (upcoming) AIA, data minimisation and decision transparency are of concern also for media forensics methods. Using the example of one DeepFake detection approach, the paper shows that data minimisation can be successfully applied in this context, without significant loss of detection accuracy.

[Kraetzer23] Christian Kraetzer, Dennis Siegel, Stefan Seidlitz, Jana Dittmann. *Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models*. Electronic Imaging, vol. 35(4):pp. 379–1–379–1, 2023. doi:10.2352/EI.2023.35.4.MWSF-379.
https://library.imaging.org/ei/articles/35/4/MWSF-379

REQ1: 'Necessary conditions for using a method'	**[Kraetzer23]** expands DeepFake detection from a 2-class problem to an n-class decision problem, presenting results for the potential attribution/identification of the DeepFake generation method used.
REQ2: 'Evaluation of new methods'	An empirical estimation of the generalization power (or lack thereof) of existing DeepFake detectors in intra- and inter-data set benchmarking, using different data selection strategies and classifiers, is presented. In addition, the classification accuracies in 2-class and n-class DeepFake detection modes are compared.
REQ3: 'Standardisation of investigation processes'	The work on operational modelling from **[Kraetzer22]** is expanded to include human-in-the-loop and human-in-control aspects made necessary by changing requirements/legislation world-wide, esp. the upcoming EU AIA.
REQ4: 'Re-evaluation of methods'	The results presented in the context of the n-class DeepFake classification experiments imply significant problems of overfitting DeepFake detection models to specific DeepFake generation methods.
REQ5: 'Publication of methods and processes'	Details of the model generation and the benchmarking strategies for robustness estimations are provided.

Main contribution of [Kraetzer23] to domain-specific forensic process models for media forensics: The main novel aspect presented here is the discussion of human-in-the-loop and human-in-control aspects for the operators used in operational modelling.

[Siegel23a] Dennis Siegel, **Christian Kraetzer**, Jana Dittmann: *Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data*. Proc. The Seventeenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2023), Porto, Portugal, September, 2023, IARIA, ISBN: 978-1-68558-092-6, pp. 43-51, 2023.
https://www.thinkmind.org/index.php?view=article&articleid=securware_2023_1_80_30054

REQ1: 'Necessary conditions for using a method'	Not addressed in this paper
REQ2: 'Evaluation of new methods'	Not addressed in this paper
REQ3: 'Standardisation of investigation processes'	The modelling work presented in this paper is compared to an established best practice promoted by forensic practitioners (the ENFSI BPM for Digital Image Authentication [ENFSI21]) to validate aspects of the conceptual and operational modelling of the approach.
REQ4: 'Re-evaluation of methods'	The media forensics data types (model) from **[Siegel22]** in particular is validated against modelling from [ENFSI21].
REQ5: 'Publication of methods and processes'	An in-depth discussion of the common aspects of and differences between the approach to media forensics process modelling presented in the paper and the one from [ENFSI21] is provided.

Main contribution of [Siegel23a] to domain-specific forensic process models for media forensics: The aspects of conceptual modelling from **[Kraetzer22]** and **[Kraetzer23]** are projected onto the conceptual model of the ENFSI BPM for Digital Image Authentication [ENFSI21] to discuss the similarities and differences between both models.

[**Kraetzer24**] **Christian Kraetzer**, Mario Hildebrandt: *Explainability and Interpretability for Media Forensic Methods: Illustrated on the Example of the Steganalysis Tool Stegdetect*. In P. Radeva, A. Furnari, K. Bouatouch, and A. A. de Sousa (eds.), Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2024, Volume 4: VISAPP, Rome, Italy, February 27-29, 2024, pp. 585–592. SCITEPRESS, 2024.

REQ1: 'Necessary conditions for using a method'	The need to re-train media forensics methods based on machine learning to counter model ageing effects is discussed. In addition, experiments are described that aim at understanding which patterns are learned in the feature space used and why these are not training a steganalysis model but rather a model for distinguishing different JPEG encoders.
REQ2: 'Evaluation of new methods'	Not addressed in this paper
REQ3: 'Standardisation of investigation processes'	The conceptual model for forensic steganalysis based on [Provos02] and [Fridrich09] is summarised.
REQ4: 'Re-evaluation of methods'	The existing steganalysis detector Stegdetect is re-trained with more recent image databases and other classifiers to determine ageing effects of the model and allow for explainability and interpretability considerations.
REQ5: 'Publication of methods and processes'	Details of the re-evaluation of Stegdetect and the corresponding steps taken to shift from black-box to gray-box evaluations are provided.

Main contribution of [Kraetzer24] to domain-specific forensic process models for media forensics: Aspects of conceptual modelling for forensic steganalysis from [Provos02] and [Fridrich09] are summarised, and aspects of the re-training of media forensics methods based on machine learning to counter ageing effects, the necessity of a shift away from black-box testing towards grey-box or white-box testing as well as explainability and interpretability issues regarding the models trained by Stegdetect are addressed.

1.5 Outline of this Cumulative Habilitation Treatise

Chapters 1 to 4 provide a framework with a common perspective for the content of the individual original feeder papers. They are structured as follows: Chapter 2 presents selected background material that is considered relevant for this treatise. In Chapter 3, the relevant contributions of the feeder papers in the context of deriving domain-specific forensic process models for media forensics are discussed in detail. Chapter 4 concludes the treatise and provides a comprehensive summary, conclusions and considerations for possible future work.

In Chapters 5 and following, the original feeder papers by the author used in this cumulative habilitation treatise are presented as they were originally published as joint work together with the corresponding co-authors.

2 Background

This chapter contains a summary of selected background material that is considered relevant by the author for placing the content of Chapters 3 and 4 into the necessary context. It includes a very brief overview of **conceptual and operational modelling** in Section 2.1. In Section 2.2, relevant **types of stakeholders** in the field under consideration are identified and briefly characterised. The main part of this Chapter is dedicated to **summaries of regulations that are relevant for the task of deriving domain-specific forensic process models for media forensics**. These summaries include: the situation in the United States of America (in Section 2.3.1), the situation in Germany (Sections 2.3.2 and 2.3.3), and selected aspects of the European context (Sections 2.3.4 to 2.3.6).

For a researcher working at a German university, the modelling context for the domain-specific process models for media forensics is of course defined by the regulatory situation in Germany. In contrast to the situation in other European countries, like e.g., the United Kingdom of Great Britain and Northern Ireland (UK), where a central authority (in the UK: the United Kingdom Accreditation Service (UKAS)) is tasked with providing centralised technical advisory service as well as an accreditation service to the forensic sector[8], the situation in Germany is more complex due to the federalistic administrative structure of the country and the rights of the sixteen German states to govern entire areas by themselves, including the education sector as well as (internal) security. As a result, the national institutions existing to provide technical advisory service (in particular the German Federal Office for Information Security (BSI), but also the Central Office for Information Technology in the Security Sector (ZITiS)) are mostly limited to an advisory role in matters of (IT) forensics and do not have strong regulatory power in most domains.

For guidelines and best practice documents for the field of information technology (IT) forensics (and all its sub-domains, including media forensics), the most relevant national stakeholder in Germany is the Federal Office for Information Security (BSI; German: *Bundesamt für Sicherheit in der Informationstechnik*; https://www.bsi.bund.de; the German national cybersecurity authority). Regarding advice on IT forensics, the BSI currently offers two documents: the '*IT-Grundschutz-Baustein* DER.2.2 *Vorsorge für die IT-Forensik*'[9], and the best practice document '*Leitfaden IT-Forensik*' [BSI11]. These two documents are briefly reflected upon in Sections 2.3.2 and 2.3.3.

Regarding forensic practice, the most relevant actors on the European level were identified[10] as the European Network of Forensic Science Institutes (ENFSI https://enfsi.eu/; see Section 2.3.4) with its Expert Working Groups, and the European Union itself (e.g., with the Artificial Intelligence Act (AIA), see Section 2.3.5). The German Federal Criminal Police Office (German: *Bundeskriminalamt*; (BKA)) as well as multiple of the 16 German state-level forensic institutions are active members of European Network of Forensic Science Institutes (ENFSI)[11], so the perspectives reflected in Section 2.3.4 also represent the positions of the German practitioners.

This chapter does not intend to present a complete picture of the forensics landscape nor does it intend to provide an overview of the research field of media forensics or selected sub-domains thereof.

[8] See https://www.ukas.com/about-us/technical-advisory-committees/forensic-science/

[9] Unfortunately most BSI documents, including this one, are only available in German; at the time of writing of this habilitation document the most recent version is Edition 2023 [BSI23b].

[10] In expert interviews conducted by the author with forensic practitioners working at various laboratories in different European countries.

[11] In fact, the German Bundeskriminalamt (BKA) hosts the ENFSI secretariat.

2.1 Conceptual and Operational Modelling in the Context of this Document

Modelling work usually serves multiple purposes. Among other purposes, scientific modelling aims at making a particular (real-world or digital-domain) phenomenon or procedure easier to understand and to reproduce. It usually consists of steps to define, quantify, represent and visualise this phenomenon or procedure with the goal of simulating or describing it on the basis of existing knowledge or data. These steps include identifying and selecting relevant aspects and then using different types of models for different aims, such as conceptual models to improve understanding, operational models to operationalise, mathematical, statistical, computational or decision models to use in recognition tasks or simulations, and graphical models to visualise the phenomenon. The focus of this habilitation project lies primarily on **conceptual and operational models**.

In a wider sense, a conceptual model refers to any model formed after a conceptualisation or generalisation process. In practice, such models are usually abstractions of entities or events in the real world, whether physical or social [Apostel60].
In IT-security research, conceptual modelling performed by practitioners or scientists in the field is the basis of any security modelling. Like any other form of scientific model, it provides a simplified abstract view of a complex reality. The challenge for any relevant conceptual modelling is to find a suitable balance between level of detail and degree of abstraction of the modelled entity or event. A too complex model might be as unsuitable as a model that over-simplifies too much.
On the basis of conceptual models, operational models are then usually used to translate strategic planing into operating requirements and decisions.

Textbooks on media forensics such as [Rathgeb22] (esp. the chapter [Cozzolino22]) or [Ho15] agree on the fact that at the core of modern media forensics pipelines looking into questions of authenticity or integrity, one or more pattern recognition or anomaly detection mechanisms are to be found. After data collection and pre-processing operations, either sequences or parallel networks of such operators (in the latter case followed by fusion operators) are used to implement a set of analysis tasks. The output of the analyses will then have to be interpreted by a human expert, e.g., in the form of an expert testimony in court.
While the community agrees on the fundamental outline of analysis pipelines, the existing state of the art lacks domain-specific conceptual and operational models.

As briefly summarised in Section 1.2, the three different media forensics application scenarios of **face morphing attack (FMA) detection** (for digital images), **DeepFake detection** (for digital videos) and forensic steganalysis are addressed in the feeder papers contributing to this cumulative habilitation treatise. These three application scenarios **define the modelling contexts** for this treatise.

2.1.1 Conceptual Modelling

According to Kung and Sölvberg [Kung86], a **conceptual model**, when implemented properly, will satisfy the following four fundamental objectives:

- Enhance an individual's understanding of the represented system
- Facilitate efficient conveyance of system details between stakeholders
- Provide a point of reference for system designers to extract system specifications
- Document the system for future reference and provide a means for collaboration

A good and widely used example of a conceptual model is the ISO OSI reference model [ISO94], used to describe and compare complex communication protocols. More relevant in the context of this treatise are the context models presented for audio authenticity analysis in the Best Practice Manual (BPM) for Digital Audio Authenticity Analysis (FSA) [ENFSI22a], with its source modelling and a corresponding specification of potential traces of (malicious) post-processing operations, and the categorisation of digital image authentication methods provided in the ENFSI BPM for Digital Image Authentication (DIA) [ENFSI21].

[Mylopoulos92] defines conceptual modelling[12] as "*the activity of formally describing some aspects of the physical and social world* [...] *for purposes of understanding and communication*." The form of expression of such a description is a 'conceptual schema' and requires "*the adoption of a formal notation*", a 'conceptual model' in the terminology of [Mylopoulos92]. Conceptual schemata are supposed to capture relevant aspects of some domain and serve as points of agreement among members of a group working in that specific domain, who need to have a common understanding. In [Mylopoulos92], Mylopoulos summarises the characteristics of conceptual schemata and modelling in the following points:

- "*Conceptual schemata can* [...] *be used to communicate that common* [domain] *view to newcomers.*"
- "*Conceptual modelling has an advantage over natural language or diagrammatic notations in that it is based on a formal notation.*"
- "*It also has an advantage over mathematical or other formal notations developed in computer science because unlike them, conceptual modelling supports structuring and inferencial facilities that are psychologically grounded. After all, the descriptions that arise from conceptual modelling activities are intended to be used by humans, not machines.*"

Synopsis: In the context of this treatise, conceptual modelling is used to improve the descriptions of media forensics investigations in regard to the requirements REQ1 to REQ5, as specified in 1.3. The focus of this work lies on REQ1, 'Describing necessary conditions for using a method', and REQ2, 'Evaluation of new methods'.

2.1.2 Operational Modelling

Operational models usually translate strategic planning into operating requirements and decisions, i.e., they specify how to execute a particular strategy. An operational model that has to be mentioned in the context of this treatise is the work on modelling examination protocols provided in the ENFSI BPM on the Forensic Examination of Digital Technology (FIT) [ENFSI15]. This work provides modelling primitives to irreducibly describe investigation procedures as a basis for their validation in function verification and proficiency testing.

As specified above, operational models are considered in this treatise as the means for translating strategic planning into operating requirements and decisions. This specification of how to execute a defined strategy is of high significance in IT security modelling in many fields, especially those that have to include risk management[13]. The process of identifying and managing operational risks is known as operational risk management and is a very important factor to be addressed in any forensic process, see e.g., the sections on 'risk analysis' in all ENFSI BPMs, for example in [ENFSI15].

Synopsis: In the context of this treatise, operational modelling is used to improve the descriptions of media forensics investigations in regard to the requirements REQ1 to REQ5 as specified in 1.3. The focus here lies on REQ3, 'Standardisation of investigation processes', and REQ4, 'Re-evaluation of methods'.

[12] Mylopoulos points out in [Mylopoulos92] that conceptual modelling, knowledge representation and semantic data modelling are three similar but distinct research areas:

> "*All three activities involve capturing knowledge about a given subject matter. Knowledge representation, however, has traditionally focused on interesting reasoning patterns and how they can be accounted for semantically and computationally.* [...] *semantic data modelling introduces assumptions about the way conceptual schemata will be realized on a physical machine (the "data modelling" dimension). Thus semantic data modelling can be seen as a more constrained activity than conceptual modelling, leading to simpler notations, but also ones that are closer to implementation.*"

[13] Here, operational risks are understood as the risk of losses or errors caused by flawed or failed processes and systems or unforeseen events that disrupt the planned operations. Such risks include natural disasters, technical defects, human errors, malicious intent and other threats.

2.2 Relevant Stakeholders in the Field of (Media) Forensics

The most relevant stakeholders in the field of media forensics are without doubt the forensic practitioners and their ‘customers’ (in most European countries, usually police officers, prosecutors and judges). However, there is no universally accepted definition of what exactly a forensic practitioner is and how exactly this person is integrated into investigative procedures. Different countries have different solutions for this, usually with the forensics departments as part of the (national, regional or local) law enforcement organisations or as independent laboratories.[14] Furthermore, different (functional) hierarchies of forensic practitioners can be found depending on the size of the laboratory under consideration as well as corresponding national regulations (cf. e.g., the discussions of the different roles of the case leader and examiners in [ENFSI21]).
These two kinds of stakeholders, the forensic practitioners and the ‘customers’, do not act independently but in a complex set of relationships with other stakeholders. There are additional entities in the executive and judicatory systems which are less involved with the daily business of investigations and court cases. These entities, in the following called ‘policy makers in executive systems’ focus on adapting the overall structure of the law enforcement system to keep it in a shape to cope with the development of crimes and criminal structures. Additional relevant stakeholders are found in legislative bodies, adapting the interpretations of what is appropriate, proportionate and ethical in law enforcement and investigations to the ever-evolving societal consensus. Another group of stakeholders relevant in this context is the large number of (academic) researchers in the forensic sciences and their own microcosm of research programmes, funding bodies, networks of excellence and scientific communities, together with the associated policies and politics.

The following list presents an overview of the stakeholders considered by the author to be relevant in the context of this treatise:

- **Forensic practitioners** (usually case leader and examiners): Work cases and present them to the 'customers', e.g., in the form of an expert testimony in court. Main focus: [ENFSI15] summarises the purpose of the work of this group as “*deliver reliable results, maximize the quality of the information obtained and produce robust evidence*.” Because of the high significance of this group of stakeholders, additional discussions of their perspectives are presented in Sections 2.3.1, 2.3.3 and 2.3.4.
- **The 'customers'**: Usually police officers, prosecutors or judges initiating the investigation and using its outcomes. Main focus: Efficient and timely legal proceedings (including a need to keep ‘junk science’ and unproven methods out of the courts).
- **Policy makers in executive systems**: Evolving the law enforcement system to keep it operational and capable of coping with the ever-evolving challenges encountered. More details on this group of stakeholders are provided in Section 2.2.1.
- **Legislative bodies**: Adapting the interpretations of what is proportionate and ethical in law enforcement and investigations to the ever-evolving societal consensus. One recent and prominent example, the European Union (EU) AIA, is discussed in detail in Section 2.3.5.
- **The academic communities in multiple research domains**: Focus on success in academic research. This group of stakeholders is discussed in more detail in Section 2.2.2.
- **Industry providing forensic solutions or tools**: Main focus: Commercial success. This group is mostly outside the scope of this treatise. Nevertheless some considerations about this group are summarised in Section 2.2.1.

Synopsis: Forensic practitioners are not self-sufficient but only one group of actors in a multi-stakeholder setup. While some of the relationships in this setup (e.g., between the forensic practitioners and their 'customers') are well defined and have had a long time to evolve, the material presented in this treatise shows that others, especially the one between the academic community and forensic practitioners, still

[14] Exact details about the different forms of integrating forensic practitioners into law enforcement or legal structures in individual countries are outside the scope of this treatise.

show significant potential for improvement. This issue is related to all requirements from REQ1 to REQ5 because the publication of and communication about forensic methods, their constraints, respectively necessary conditions, their evaluation and re-evaluation as well as their standardisation should be done in an interaction between these groups.

2.2.1 Policy Makers in Executive Systems

A very recent and detailed example of a policy document regarding a national digital forensic science strategy is the UK document [Vaughan20]. In this 60-page policy document, a group of policy makers (the National Police Chiefs' Council (NPCC)) within the UK police forces outlines a national development strategy for digital forensics (DF) for the time-frame from 2020 to 2025. This group reflects the current (2020) situation as follows: "*While crime and criminals have become ever more digitally sophisticated, our response, at every level of law enforcement, has been slow, fractured and piecemeal.*" Even though the NPCC focuses its work strictly on the UK, this statement can be considered as true for most, if not all countries world-wide.
The authors of [Vaughan20] attribute technical aspects to the problem:

> "*The data from victims, witnesses and suspects – the data for digital forensics – is from non-police sources and is about 20 times the volume of all other police data combined, and demands additional consideration around how it is captured, used and stored. But it is vital that our plans for digital forensics are no longer siloed away from the wider policing digital/*[Information and communications technology] *landscape as has too often been the case in the past.*"

In their opinion, non-technical aspects also have to be addressed, e.g., the changing societal consensus on what are proportionate and ethical means in the context of law enforcement and forensics: "[We] *need to rebuild and maintain public trust through better compliance with data protection legislation.*" [Vaughan20] Recruiting and retaining a skilled workforce, which is becoming increasingly difficult, is also addressed as a goal to be achieved, e.g., by using automated (i.e., based on machine learning) analysis methods:

> "*Key to the strategy is the industrialised, consistent and standardised approach to the use of technology. For example, in harnessing automation to make sure that we not only deliver better results quicker, but do so in a way that reduces the emotional distress for our teams caused by investigating disturbing images.*"

In [Vaughan20], many different challenges and issues are highlighted. These include (among others) the increasing volume and complexity of investigations, their legitimacy and issues of public trust and confidence, an apparent lack of support services as well as issues of recruitment and retention. In addition to these issues, which are of limited relevance for this habilitation project, the fragile commercial marketplace for forensic solutions is addressed as follows:

> "[F]*orensic science funding reductions have led to a substantial decrease in the size of the supplier market* [...] *DF services in policing rely on external suppliers for additional capacity and expertise. Action to put this market on a sustainable footing is essential because policing has insufficient internal capacity to cover the gap and this risks growing backlogs and failure to harvest critical intelligence, data and evidence.*"

In addition to identifying these challenges, [Vaughan20] also discusses selected potential solution strategies. The solution suggested by the NPCC for the issue of the fragile commercial marketplace for forensic solutions lies in 'Improved commercial practices' in combination with 'Centralising science and Research and Development'. For the first item, [Vaughan20] states:

> "*Bringing how policing engages with the commercial sector onto a sure and strategic footing is key to transforming DF science service. Coordinating this engagement nationally and agreeing joint requirements, will enable policing to leverage its collective buying power and*

> *act as an 'intelligent customer'. This will allow us greater influence to ensure the market develops the capabilities we will need and ensure the supplier market is sustainable and resilient.*"

For the aspect of 'Centralising science and Research and Development', central national police bodies should

> "*coordinate and influence an R&D programme drawing in R&D effort nationally, combining casework-driven capabilities that practitioners develop with industry research by tool vendors, and multi-disciplinary academic research.*"

In a different part of the document, the NPCC requests that this research and development (R&D) programme be "*sufficiently well-funded*" and calls for an "[a]*cademic advisory group providing independent advice and guidance, to ensure policing can access full potential of ongoing R&D.*"

Synopsis: Several items relevant for this treatise are discussed in the 60-page policy document published in 2020 by the NPCC. The first item is the question of what proportionate and ethical means in changing societal contexts (in Western societies) and why public trust is important for forensic investigations. This motivates the inclusion of recent and ongoing public debates (like the current discussion on the EU AIA and the Interpol initiative 'Responsible Artificial Intelligence (AI) Innovation in Law Enforcement') in this treatise (see Section 2.3.5).
The second item is the emphasis on a consistent and standardised approach to the use of technology, quality standards and a modern forensic data model[15] for DF. This is reflected in the modelling work summarised in Chapter 3.2.
The third item to be highlighted here is the call for a forensics marketplace that is able to respond to the fast-changing requirements and needs. Here, the NPCC shows a strong orientation towards commercial solutions [Vaughan20]:

> "*Private sector providers are essential to delivering DF service and key to a nationally networked approach. So too are other organisations – including higher education institutions, specialist research organisations, start-ups and existing DF tool vendors - which can support innovation in the future.*"

With this, the UK takes a different route than many other countries in the world. The alternative trend seen in many countries, like the Netherlands or Canada, is to heavily invest in technical solutions developed by security forces or forensic institutions themselves. Prominent examples for this trend

[15] In their request for a (national) forensic data model, as enabler for the interoperability between tools and infrastructure throughout the DF workflow, the best summary on data and its evidential value in forensic investigations sencountered by the author is found in [Vaughan20]:

> "*To understand how we can best optimise how we use this data, we need to distinguish digital forensic source data – the data directly extracted from devices – from data resulting from a digital forensic analysis. Both are digital forensic data and require a quality assured process to be followed during extraction and review, but it is important to make the distinction and be clear about the extent to which investigators can rely upon data if it received no critical technical scrutiny. In some cases, the evidential value of the data comes from its content: for example, an investigation of threatening text messages sent to an ex-partner may only require the extracted data to prove the case. In this case, a level 1 examination may be appropriate (and this will be subject to quality standards, with a trained operator following a validated process). However, if a defendant alleges that the messages in question are fake, it is the provenance which is relevant – how the messages came to be there - rather than the content. This requires further digital forensic analysis, with a DF practitioner using their skills to reconstruct the chain of events.*
> *The practitioner might check corroborating details, such as metadata on when messages were received and viewed, against their own knowledge of tool capabilities and device characteristics. They might consider alternative scenarios, such as an online service being used to send the message with a faked sender number, based on their knowledge of additional possibilities to explain what they observe. They might analyse the raw source data to verify how the tool has interpreted dates and contact identity. They could conduct tests to verify the feasibility of these scenarios and establish which alternative explanation is most likely. The output of this, expressed in a digital forensic report, will be of greater evidential value than the digital forensic source data to address that specific point, because it has received scrutiny from a DF practitioner with relevant knowledge and expertise. Understanding the requirements in each case determines the level of service which is needed, and understanding the evidential value of the data in context is vital to managing DF data in future.*"

include Hansken for the Netherlands (a case management and data warehouse system developed by the Netherlands Forensic Institute (NFI) as an open digital forensic platform / a Digital-Forensics-as-a-Service (DFaaS) solution with a public Software Development Kit (SDK) for developing new plug-ins and components) and Assemblyline for Canada (a scalable Open Source file triage and malware analysis solution developed by the Canadian Centre for Cyber Security, an agency of the Communications Security Establishment, accountable to the Minister of National Defence).

The fourth item is the lack of strategic relationships between police forces (and their forensics units), academia and the industry.

These items mostly relate to the solution standardisation and publication issues addressed in REQ3 and REQ5.

2.2.2 The Academic Community

Any generalisation about the academic community and the motivations of individuals and organisations doing research is ultimately bound to fail due to many factors, including the (cultural) differences in the established science disciplines as well as the immense number of people involved world-wide.

If such a generalisation were to be attempted for the author's research domain, which is computer science (as part of the mathematics, computer science, natural sciences and technology branch of research), then the capability of acquiring research funding might be considered one of the main performance indicators in this field, and well-established expertise (in terms of publications and a history of funded research) would in many cases be a requirement for successful project acquisition. This leads, especially for early-stage researchers (ESRs), to a situation called 'publish or perish' [Grancay17], forcing them to publish at very high output rates to be able to establish enough expertise to acquire funding. As soon as a research grant has been obtained (in computer science, such grants for ESRs usually have a duration of 1 to 4 years), it has to be turned into publications to assure a good starting position for the race to the next research grant in this highly competitive field. As a result, academic research sees as its output a large number of scientific papers, based on low-TRL[16] research demonstrators, usually evaluated under 'lab conditions'. There is hardly any time (or obtainable funding) for further development work on these low-TRL (usually TRL4 to 6) research demonstrators, i.e., they are doomed to slide into the so-called 'Valley of Death' of demonstrators which are never developed further into prototypes or tools.

In the paper 'Moving Steganography and Steganalysis from the Laboratory into the Real World' [Ker13], a group of internationally acclaimed academic experts in one forensics-specific sub-domains (steganalysis[17]) summarise this situation for their own field of expertise with the following words:

> "*There has been an explosion of academic literature on steganography and steganalysis in the past two decades. With a few exceptions, such papers address abstractions of the hiding and detection problems, which arguably have become disconnected from the real world. Most published results, including by the authors of this paper, apply 'in laboratory conditions' and some are heavily hedged by assumptions and caveats; significant challenges remain unsolved in order to implement good steganography and steganalysis in practice.*"

Ker et al. [Ker13] reflect on the huge discrepancy between academic work on steganography and steganalysis and the specimens encountered 'in the wild' as follows:

> "*However, where details of real-world use of steganography are known, it is apparent that they bear little resemblance to techniques described in modern literature. Indeed, they often suffer from flaws known to researchers for more than a decade. How has practice become so disconnected from research?*"

In the opinion of the author, this statement made in 2013 is still true today for the field of steganalysis and could (to some extent) also be generalised for many other sub-domains, especially in media forensics,

[16] Technology Readiness Level (TRL); an estimate of the degree of maturity of the implementation of a method, see Section 2.3.4.8

[17] Here, steganalysis can be understood as the counter-science to steganography, which is considered to be the art and science of hidden communication. As a result, steganalysis focuses on the efficient detection of the use of steganography.

which is without doubt one of the youngest sub-disciplines of IT forensics.
In recent years, there has been a shift in perception by researchers. In a recent textbook on media forensics targeting digital face manipulations ([Rathgeb22]), the authors reflect the current academic perspective on media forensics as follows:

> "*In case manipulation detection methods are used by public authorities competent for preventing, investigating, detecting, or prosecuting criminal offences this shall be done in a lawful and fair manner. While these are broad concepts, case law further explains how to apply these concepts.*"

The mentioned characteristics are further specified in [Rathgeb22] as:

- Lawfulness: "*the need* [...] *to adopt adequate, accessible, and foreseeable laws with sufficient precision and sufficient safeguards whenever the use of the detection technology* [...] *could interfere with fundamental rights and freedoms*"
- Fairness: "*the need for being transparent about the use of the technology. Furthermore, it is obvious that the use of the detection methods should be restricted to well-defined legitimate purposes* [...]."

Regarding fairness, Rathgeb et al. point out in [Rathgeb22] that when intended for court use, explainability of the forensic algorithms used is a strong requirement. In addition, they state that

> "[f]*rom an organizational point, one should also know that decisions purely and solely based on automated processing, producing adverse legal effects or significantly effecting subjects, are prohibited, unless authorized by law, and subject to appropriate safeguards, including at least human oversight and intervention.*"

In accordance with other well-established work originating in academic media forensics research (like [Ho15]), the synopsis presented in [Rathgeb22] is that

> "[t]*he absence of a unified approach, common regulatory framework, and commonly accepted practices has resulted in a situation where different initiatives emerge across countries which share some common elements but also numerous differences that can lead to challenges related to interoperability.*"

An important step towards more mature forensics would be more mature forensic process models. They guide investigations and are intended to make them comparable, reproducible as well as certifiable. Usually, the adherence to strict guidelines (i.e., process models) is regulated within any legal system, e.g., in the US by the fourth of the Daubert criteria ("*the existence and maintenance of standards and controls*" [Champod11]).
Due to the fact that IT forensics is a rather young discipline in this field (with media forensics being an even younger sub-discipline), it is hardly surprising that the forensic process models (if they exist at all) have not yet achieved the same degree of maturity here as in other fields. Nevertheless, they would be required to achieve universal court acceptability of methods.

Another important step here is overcoming the 'Valley of Death' that swallows most academic demonstrators at low TRLs (and also many higher-TRL prototypes or tools from technology start-ups). This would require a significant demand or technology pull from potential end-user stakeholders (e.g., forensic practitioners) and, more importantly, the support from policy makers in executive and judicatory systems, including the well-funded, specialised R&D programmes called for by the NPCC (see Section 2.2.1).

One success story that began in academia in 2004 and already achieved considerable success in 2011 with a method accepted in the Daubert hearings (see Section 2.3.1) in a US court case shall be presented here: The method, which is now widely advertised, e.g., in the ENFSI BPM for Digital Image Authentication (DIA) [ENFSI21], is the usage of the so-called Photo Response Non-Uniformity (PRNU) for image source verification. It was originally introduced in [Lukas06] and is now accepted as one of the most reliable methods for digital-camera authentication based on intrinsic characteristics of the image acquisition sensor. Its development (which is discussed in depth in feeder paper **[Kraetzer15a]**, included as Chapter 5 of this cumulative habilitation treatise) can be briefly summarised as follows: In 2005, [Lukas06] described the approach at a very low TRL (3-4). The method and implementations were refined to a mid-range TRL (5-6) and underwent large-scale in-house testing until 2009 (see [Goljan09]). At that point, an interested end-user (the Forensic Audio, Video, and Image Analysis Unit of the FBI (FAVIAU)) declared its interest, and a sponsor (the US Air Force Research Laboratory, Air Force Material Command) was willing to fund further development for an end-user-ready application (a tool called 'FindCamera' used by the FBI but never publicly released). In 2011, the tool reached TRL 9 and was accepted for court room use by the presiding judge in *United States of America v. Nathan Allen Railey* (United States District Court for the Southern District of Alabama, August 2nd, 2011). In the corresponding Daubert hearing, the FBI FAVIAU expert investigating in the case was able to convince the presiding judge that the method fulfils all the requirements (see the discussion of Daubert criteria in Section 2.3.1) and gives results that are consistent with another media forensics approach (Exchangeable Image File Format (EXIF) metadata analysis and matching) applied to the same images by an independent expert. For an in-depth discussion, see feeder paper **[Kraetzer15a]**, included as Chapter 5 of this cumulative habilitation treatise.

The successful court room use was at that time considered a huge success for the whole media forensics research community, with one of the new digital investigation methods reaching not only TRL 9 but even court room acceptance in such a short time. Unfortunately, only very few other media forensics methods have been able to achieve the same success since then.

Synopsis: The academic world excels at research, but it is trapped in its own mechanics, including the race for funding summarised above, with all its positive and negative side-effects. To make the outcome of academic research (especially the methods of low-TRL research demonstrators) available to forensic practitioners (or other stakeholders in this field), it usually requires a significant demand or technology pull from potential end-users. It should not be assumed that commercial providers (e.g., DF tool vendors) would be willing to step up to provide such help because (as discussed in Section 2.2.1) the market for forensic solutions is extremely limited.

The second strong point that should be attributed to academic researchers is their knowledge of the field and the fact that they can provide insights about the gap between what can be seen as phenomena 'in the wild' (e.g., which steganographic methods are used in current steganography tools or stego-malware[18]) and the current state of the art in the field in academic research. With this knowledge, they would be able to efficiently fill the role of the independent advisory groups requested by the NPCC (see Section 2.2.1).

Therefore, the trend currently seen in academia to shift away from results that apply only 'under laboratory conditions' towards publicly evaluated methods for which the error rate estimates provided are good enough to being taken up by the industry will have to be strengthened further.

The necessity for academia to interact with all other stakeholders in forensics relates to all requirements (REQ1 to REQ5) specified in Section 1.3.

2.3 (Some) Relevant Regulations in this Field

In this section, some international and national regulations as well as related best practices are discussed to illustrate the similarities and differences between selected regional contexts. Since the author is a German citizen working in Germany, one of the selected contexts is the German perspective, which cannot be discussed in depth without reflecting also on a wider set of European influences and regulations. These German and European contexts are then compared with the situation in the United States of America, which has an extremely active judicial system and as a consequence very active forensic

[18] Stego-malware is malicious software (malware) that uses steganographic channels for unsuspicious infiltration, data exfiltration or hidden command and control communication. Examples for stego-malware families can be found at: https://attack.mitre.org/techniques/T1001/002/.

communities (in terms of practitioners, policy makers, industry as well as academia).
Three opening remarks have to be made for this section of the treatise:

1. Some parts of the presented context summarise legal regulations. Since the author has absolutely no legal training, all legal considerations made within this document are therefore a layman's interpretation of freely available material, which are made to the best of the author's knowledge. If the content of this habilitation treatise is intended to be used in any legal proceedings, the reader must consult appropriate legal counsel for the corresponding jurisdiction.
2. The excerpts presented on the governing principles for forensics in the regional contexts discussed do not present a complete, in-depth picture of the current state of regulations and practices in the US, Germany or Europe. They are high-level summaries compiled on the basis of the author's own experience in academia and are intended to highlight some of the similarities and differences in the situation in the US in contrast to the situation in Germany/Europe. This serves to pave the road to reasoning why solutions that might work in the US might not be suitable in the other considered contexts and vice versa.
3. The presentation might give the impression that entirely separate contexts are compared. This is obviously not true. Nearly all legal systems world-wide are connected in some form, and international actors like the (now dissolved) International Organisation on Computer Evidence (IOCE), INTERPOL or the United Nations Interregional Crime and Justice Research Institute (UNICRI) work hard on a consensus about internationally accepted principles as well as on the harmonisation of methods and practices among nations to guarantee the usability of digital evidence collected by one nation in the courts of another nation. The following (incomplete) list gives an impression of how many international bodies are active in this field: the International Organisation for Accreditation Bodies (ILAC; with its standard *ILAC-G19:06/2022 – Modules in a Forensic Science Process*), the INTERPOL action on digital forensics, the ASTM International (formerly known as American Society for Testing and Materials (ASTM), with its standard *E1492-11(2017) Standard Practice for Receiving, Documenting, Storing, and Retrieving Evidence in a Forensic Science Laboratory*), and the International Standards Organization (ISO) / International Electrotechnical Commission (IEC) with a large number of different standards[19].

2.3.1 The US Situation

In 2013, the author summarised the relevance of the situation in the US regarding the handling of forensic procedures in the wider context of national and international legal procedures as follows [Krätzer13]:

> "*The U.S. legal system is one of the most active in the world with large numbers of trials involving all kinds of forensic investigations being held every day. As a result, within this legal system strict rules for the integration of the results of forensic investigations have been established.*"

On the basis of this statement (which still holds true today), a layman's interpretation of the framework for court admission of forensic evidence is presented in [Krätzer13]. This framework is based on the assumption that, in general, forensic results have to be interpreted by experts for the court. The reason for this is that any judge (or jury) will presumably lack the expert knowledge necessary to interpret the findings of a forensic investigation completely on their own and that therefore expert testimony is strictly required in court proceedings. Thus, if the expert testimony helps the fact finder in understanding the significance of factual data, then the expert witness is essential for the case and their opinion evidence is admissible.
In some of the feeder papers (esp. **[Kraetzer15a]**, which is included as Chapter 5 of this cumulative habilitation treatise, and **[Kraetzer22]**, Chapter 12), the rules governing the admission of forensic

[19] These include, among others, ISO/IEC 27037 (which concerns the initial capturing of digital evidence), ISO/IEC 27041 (which offers guidance on the quality assurance aspects of digital forensics, e.g., ensuring that the appropriate methods and tools are used properly), ISO/IEC 27042:2015 (with guidelines for the analysis and interpretation of digital evidence), ISO/IEC 27043 (which covers the broader incident investigation activities, within which forensics usually occur) and ISO/IEC 27050 (electronic discovery).

methods in US federal-level courts have been summarised and discussed from a layman's perspective. With the Federal Rules of Evidence (FRE) (esp. Federal Rules of Evidence - rule 702 (FRE702)) and the Daubert criteria, judges in US federal-level courts have a well-established set of instruments to have prosecution and defence carry the burden of arguing in favour of or against the inclusion of (forensic) methods and investigations in a specific trial during the Daubert hearings (before the actual trial). This allows them to effectively act as the 'gatekeeper' without having to acquire the specific knowledge required to establish the degree of maturity of a novel method by themselves. This is considered internationally to be an efficient and fair approach. [Champod11]

While the requirements are clear, generalizable and standardised forensic process models and standard operational procedures (SOPs) are currently sought for to bridge the gap between these strict legal requirements and the current degree of (or rather lack of) maturity of many media forensics approaches originating form academic research (see Section 2.2.2).
One important attempt to homogenise forensic procedures in the fragmented US legal system is the document 'Forensic Examination of Digital Evidence: A Guide for Law Enforcement' [Ashcroft04] by the U.S. Department of Justice, National Institute of Justice (NIJ), published in 2004 (prepared under an inter-agency agreement with the National Institute of Standards and Technology (NIST)). It describes a forensic process model and corresponding best practices. Unfortunately, it has not received any update since the initial release in 2004, and the responsibilities in this field have shifted, with the NIJ taking a more passive role, focusing on roadmapping and research funding, while standardisation tasks went to the NIST. Currently, the NIST is one of the most important drivers of the ongoing work in forensics in the US. A good example of current NIST initiatives in this field is the current draft of the document 'Digital Investigation Techniques: A NIST Scientific Foundation Review' [Lyle22], which, on the one hand, gives an overview of forensic procedures (including an updated version of the widely accepted forensic process model from [Kent06]), established techniques (heavily relying on the best practices documents of the corresponding Scientific Working Group (SWG) of experts), vendor-independent training as well as proficiency testing initiatives (spearheaded by NIST), and, on the other hand, provides a rough overview of the current landscape of forensic practitioners in the US. The NIST work on forensic process modelling, established techniques, training and proficiency testing is very similar to the work discussed below in Section 2.3.4 for the role of the ENFSI in Europe. Relevant here, because differing significantly from the European situation, are the insights presented on the landscape of forensic practitioners and crime labs. After a lengthy discussion on an appropriate way to estimate their number, [Lyle22] presents the following summarizing statement:

> "*This value of 11,000 US digital forensics organizations contrasts with the 409 publicly funded crime labs reported by the Bureau of Justice Statistics* [...]. *The decentralization of the digital forensics community in the United States is apparent in where digital forensics labs are found; they are not only in federal, state, and local crime labs, but also in prosecutor's offices, private consulting firms, and corporate cybersecurity operations.*"

As a result of this extreme fragmentation of the landscape, much or the relevant work on best practices and standardisation (including certification work) is done by the so-called Scientific Working Groups (SWGs). These are usually non-profit organisations focusing on the development of guidelines and standards. For the organisation of the SWGs, (financial and administrative) support is granted by the NIST. The SWG most relevant for this treatise is the Scientific Working Group on Digital Evidence (SWGDE). It provides best practice documents on the investigation of IT forensics (including audio, imaging, photography and video) that are very similar in nature and content to the ENFSI best practice manuals discussed in Section 2.3.4, as the US SWGs and the EU ENFSI Expert Working Groups (EWGs) work in close cooperation. When analysing the lists of members / member organisations contributing to the working groups and the corresponding best practice documents, it is apparent that only about half of the contributors are public bodies, while the other half are companies, law firms or other entities of the private sector.

Synopsis: Regarding legislation, the procedures for admitting forensic methods into US court cases are considered to be state-of-the-art in publications like [Champod11]. Regarding the situation of experts, a relatively small number of publicly funded crime labs (approximately 409) in the US are facing more than 11,000 competitors in the private sector. The National Institute of Standards and Technology (NIST) plays an important role in organising the nationwide efforts on developing the forensic sciences, e.g., by their own work in process modelling or proficiency testing, but most importantly by hosting the SWGs as bodies of experts. These SWGs publish best practice manuals and standards intended to guide forensic investigations throughout the United States of America. These standards are also recognised and used outside of the US and the SWGs work in close cooperation with other entities world-wide, e.g., their counterparts in Europe (i.e., the corresponding ENFSI EWGs). In contrast to Europe, the private sector has a significant influence on the entire forensic market in the US, on the casework as well as on the SWG documents and standards.

It remains to be seen how such a highly fragmented field with a strong private sector (i.e. commercial) influence will be able to cope with the challenges currently faced in forensics: the increasing volume of investigations as well as the amount of data per case, but also the questions arising from recent debates on ethics, trust, data protection issues and the regulation of AI.

These standardisation issues all relate to REQ3, as specified in Section 1.3.

2.3.2 The German BSI Guidelines for (IT) Forensics Established in the ‘IT-Grundschutz-Baustein DER.2.2’

The German Federal Office for Information Security (BSI) guidelines document ‘*IT-Grundschutz-Baustein DER.2.2*’ is part of the recommendations family ‘*IT-Grundschutz*’ (roughly translated by the author as ‘baseline protection (modules) for IT systems’) and intends to provide the necessary guidelines to enable forensic readiness in companies or administrative bodies.

The self-proclaimed aim of the BSI is to provide, with the ‘*IT-Grundschutz*’ module catalogue, a “*systematic basis for information security*” [BSI23a]. It aims to provide a “*sound and sustainable methodology for information security management systems (ISMS)*” and cover “*technical, organisational, infrastructural and personnel aspects in* [...] *a systematic approach to information security that is compatible to ISO/IEC 27001.*” The audience addressed is identified as “[the] *information security officer of a public authority, the Chief Information Security Officer (CISO) of a large company or the managing director of a small or medium-sized enterprise.*” As a result of the target audience chosen, the module focusing on IT forensics and the preparation of IT systems with respect to a potential forensic investigation, ‘*IT-Grundschutz-Baustein DER.2.2*’, is only six pages long and only provides a management-level overview of the topic. The document contains:

- A phase-driven conceptual model for a forensic investigation, splitting it into two parts: a preparation phase called Strategic Preparation (SP) (Geman: ‘*Strategische Vorbereitung*’) and the actual investigation
- A very brief consideration of legal aspects
- A list of the most relevant actors involved (information security officer, data protection officer, etc.)
- A list of requirements that must be fulfilled for the successful preparation of IT systems with respect to a potential forensic investigation
- A list of optional requirements that should be fulfilled for the successful preparation of IT systems with respect to a potential forensic investigation
- A list of additional resources (including a reference to the BSI best practice document ‘*Leitfaden IT-Forensik*’ discussed below in Section 2.3.3)

The most important points with regard to this habilitation treatise are:

1. The emphasis on the need for the dedicated SP phase (in which processes are planned and established to ensure that an institution can forensically analyse IT security incidents, which is considered by the BSI necessary even if the institution itself does not possess forensic expertise); relates to REQ1, REQ2 and REQ3, as specified in Section 1.3

2. The requirements focusing on the operational planning of forensic processes (including tools selection and verification); relates to REQ1, REQ2, REQ3 and REQ4
3. The requirements for training of personnel to implement the aforementioned forensic processes; relates to REQ1, REQ3 and REQ4
4. The requirements for the documentation of forensic processes (including justification of the methods chosen); relates to REQ1 and REQ3
5. The recommendation to perform proficiency testing with the above-mentioned personnel; relates to REQ3 and REQ4

The references provided in the '*IT-Grundschutz-Baustein DER.2.2*' primarily contain a pointer towards the BSI best practice document '*Leitfaden IT-Forensik*' as well as four selected additional sources (standards ISO/IEC 27042:2015 'Information technology — Security techniques — Guidelines for the analysis and interpretation of digital evidence' and ISO/IEC 27043:2015 'Information technology — Security techniques — Incident investigation principles and processes', the 'Standard of Good Practice for Information Security' published by the Information Security Forum and the Request for Comments (RFC) 3227 'Guidelines for Evidence Collection and Archiving').

Synopsis: Even though this document is important for German companies and administrative bodies aiming for a BSI certificate for their services, the technical depth and the relevance are limited to the items summarised above. For more detailed considerations, which are necessary in many contexts, including in this habilitation treatise for REQ3, it points towards the much more detailed BSI '*Leitfaden IT-Forensik*'.

2.3.3 The German BSI Guidelines for (IT) Forensics Established in the 'Leitfaden IT-Forensik' and the Follow-up Work of the DCEA

With its 353 pages of content, the '*Leitfaden IT-Forensik*' [BSI11] of the BSI provides the technical details as code of practice for IT forensics that the aforementioned '*IT-Grundschutz-Baustein DER.2.2*' lacks. Its target audience also includes security management (information security officers, CISOs, etc.), but its actual focus lies on system operators with a background in incident response (e.g., members of a Computer Emergency Response Team (CERT)) and forensic practitioners.
Since the publication of the '*Leitfaden IT-Forensik*' in 2011, it has been used as basis for a number of publications extending the original concepts. A selection of these (scientific) publications is briefly discussed in Section 2.3.3.2 below.

2.3.3.1 BSI 'Leitfaden IT-Forensik'

One of the many purposes of the BSI 'Leitfaden' is to try to somehow homogenise forensic procedures in the highly fragmented system of more than 35 different German police agencies on federal and state level. In this regard, it is very similar in its intention to the document 'Forensic Examination of Digital Evidence: A Guide for Law Enforcement' [Ashcroft04] of the U.S. Department of Justice, National Institute of Justice (NIJ) published in 2004. It is slightly outdated, with the last updated version of the '*Leitfaden*' (German for "guideline") published in 2011. Nevertheless, it is still an important starting point for conceptual as well as operational modelling and has been used as such during the entire course of research for this cumulative habilitation treatise.

The forensic process model, as the core element of this guideline document, consists of three main components[20]: a **phase-driven investigation process model** (Strategic Preparation (SP), Operational Preparation (OP), Data Gathering (DG), Data Investigation (DI), Data Analysis (DA) and Documentation phase (DO)), a modelling of **forensically relevant data types** and the definition of a set of forensic **method classes** (sets of methods for the forensic process in digital forensics, provided as: Methods of the operating system (OS), Methods of the file system (FS), IT application (ITA),

[20] The '*Leitfaden*' is written in German. The English translations of its modelling components provided in [Kiltz15] and [Kiltz20] are used here.

Explicit means of intrusion detection (EMID), Scaling of methods for evidence gathering (SMG), Data processing and evaluation (DPE)). These components are presented in detail in the feeder paper **[Siegel21]**, included as Chapter 10 of this document. The discussion of the corresponding content is found on pp. 148 ff. of this cumulative hablitation treatise.

The feeder paper **[Kraetzer22]**, included as Chapter 12 of this cumulative habilitation treatise, provides the following comparison with other models:

> "*It has to be acknowledged here that these BSI guidelines on outlining a forensic process, while acknowledging established best practices in this field, significantly differ from other national guidelines, even in other EU states. This can be illustrated by comparing it for example with the model described in [Flaglien17], which very well reflects the Norwegian approach. It also builds upon a phase-driven model but with a different established phases layout: (1) Identification Phase, (2) Collection Phase, (3) Examination Phase, (4) Analysis Phase and (5) Presentation Phase. This is much closer to long-time established best practices in traditional (analogue world) forensic sciences and requires then explicit activities to achieve and maintain* "Digital Forensic Readiness" *[Flaglien17] (an equivalent to the Strategic Preparation phase in the BSI guidelines) to successfully cope with modern day digital- and digitised forensics tasks.*"

Synopsis: Due to the national relevance for Germany, the methodology presented in the BSI '*Leitfaden IT-Forensik*' has been used (and expanded) in the author's own research work since the document was published in 2011. In the context of this treatise, this document relates to REQ1, REQ2, REQ3 and REQ4, as specified in Section 1.3
Some sources, for example [FHNW21], identify the BSI '*Leitfaden IT-Forensik*' as one of the three most relevant forensic process models.[21] The author finds it difficult to support such generalizations, since the original BSI document is only available in German, which limits its use significantly. Nevertheless, derivative work is also available in English. Examples for such derivative work are discussed in Section 2.3.3.2.

2.3.3.2 Work Extending the Methodology Presented in the BSI 'Leitfaden IT-Forensik'

Similar to the author's own work (including his PhD thesis [Krätzer13] (2013)), other researchers in Germany have also based their research on the foundations provided by the '*Leitfaden IT-Forensik*'. Worth mentioning here are the PhD theses of Tobias Hoppe (2014), Stefan Kiltz (2020), Robert Altschaffel (2020) and Mario Hildbrandt (2020) (all received their degrees from the Department of Computer Science of Otto-von-Guericke University in Magdeburg, Germany).
In his thesis, Tobias Hoppe [Hoppe14] uses the model of the '*Leitfaden*' for forensic investigations of automotive malware. The thesis of Stefan Kiltz [Kiltz20] and papers related to that PhD project, like [Kiltz15], provide a comprehensive and expanded English version of the forensic process model of the BSI 'Leitfaden IT-Forensik' called Data Centric Examination Approach (DCEA) and focus (among other things) on a formal modelling of error, loss and uncertainty in forensic investigations. In the thesis of Robert Altschaffel [Altschaffel20], a domain adaptation to the forensic investigation of ICS is performed, while Mario Hildebrandt in his thesis [Hildebrandt20] expands the coverage to include applications in digitised forensics (of fingerprint traces).
As mentioned above, publications by the author also have been based on this established model. The corresponding proposals for extensions are discussed in Chapter 3.2 of this document.

[21] The German text [FHNW21] states:

> "*In der IT Forensik haben sich drei Vorgehensmodelle primär etabliert: das Modell des Bundesamts für Sicherheit in der Informationstechnik (BSI), das Modell des National Institute of Standards and Technology (NIST) und das Secure, Analyse, Present Modell, kurz S-A-P.*"

This can be translated into English as follows:

> "*Three process models have become most relevant in IT forensics: the model of the* [German] *Federal Office for Information Security (BSI), the model of the National Institute of Standards and Technology (NIST) and the Secure, Analyse, Present model, or S-A-P for short.*"

Synopsis: Even though the methodology presented in the BSI '*Leitfaden IT-Forensik*' is slightly outdated (having been released in 2011) and only available in German, it has been established (e.g., by multiple successful PhD thesis projects) that it was (and still is) a suitable foundation for forensic modelling work. As such, it not only relates to REQ1, REQ2, REQ3 and REQ4, as specified in Section 1.3, but also to REQ5.

2.3.4 Guidelines from (Selected) ENFSI Best Practice Manuals

The European Network of Forensic Science Institutes (ENFSI) is a non-profit initiative of (national) forensic laboratories in the EU with financial support from the European Commission, organised into Expert Working Groups, each curating and publishing so-called Best Practice Manuals (BPMs[22]) or Forensic Guidelines documents. The scope of the BPMs can be very wide, covering an entire application field (e.g., the BPMs on Forensic Examination of Digital Technology (FIT) [ENFSI15] and on Digital Image Authentication (DI) [ENFSI21]), or very narrow, covering exactly one specific analysis method (e.g., the BPM for ENF Analysis in Forensic Authentication of Digital Evidence [ENFSI09]).
Because of the relevance of the members of this organisation[23], ENFSI naturally plays an important role in the world-wide discussions on forensic sciences, similar to that of the US NIST and the US SWGs (see Section 2.3.3.1). There are strong ties between the European and US institutions: In the ENFSI BPM for the Forensic Examination of Digital Technology [ENFSI15], for example, the authors point out that the terminology used has been homogenised with the corresponding North American Special Working Group (SWGDE in that case) to aim for consistency between these organisations.

Of the more than 20 existing BPMs, the two that are of the highest significance for this treatise are the ENFSI Best Practice Manuals for the Forensic Examination of Digital Technology (FIT) [ENFSI15] and on Digital Image Authentication (DIA) [ENFSI21]. In [ENFSI15], the purpose of this BPM is summarised as follows:

> "*This Best Practice Manual (BPM) aims to provide a framework for procedures, quality principles, training processes and approaches to the forensic examination. This BPM can be used by Member laboratories of ENFSI, and other forensic science laboratories to establish and maintain working practices in the field of forensic IT examination that will deliver reliable results, maximize the quality of the information obtained and produce robust evidence. The use of consistent methodology and the production of more comparable results will facilitate interchange of data between laboratories.*"

Like any other BPMs, it does not intend to be a standard or other kind of regulatory document, but intends instead to be a knowledge base document written by a group of expert practitioners in the field, providing "*technical guidance to aid the design of local standard operating procedures (SOPs*[24]*) in compliance with local regulatory requirements, and international standards.*"
Regarding the general setup, the BPMs stress the strong dependence of forensic practitioners on others: The forensic process usually has a 'customer' (described in [ENFSI21] as "*usually a judge, prosecutor, or police officer, and also private persons, where the jurisdiction allows it*", i.e., the beneficiary of a forensic report) and a forensic lab tasked with performing a forensic investigation. As argued in [ENFSI21], the BPMs are in this setup supposed to describe the following:

- "*The formulation of useful propositions*[25] *based on the claims and the questions supplied by the Customer* [...]"

[22] [ENFSI21] notes for the terminology used: "*The term BPM is used to reflect the scientifically accepted practices at the time of writing* [of the corresponding BPM document]."

[23] Members originate from 39 European countries (with all of the large countries present) and include usually the national forensic laboratory as well as the national forensic science institute.

[24] In fact, all ENFSI BPMs (but also the published ENFSI Forensic Guidelines documents) stress that they are "*not a standard operational procedures (SOP) and address the requirements of the judicial systems in general terms only*" [ENFSI21].

[25] The term 'proposition' is defined in [ENFSI15] as "[s]*tatements that are either true or false, and that can be affirmed or denied. Propositions should be formulated in pairs (e.g., views put forward by the parties to the case) and against a background of information and assumptions.*"

- *"The wide selection of methods which one may use to evaluate each proposition, the principles of how to choose between them, and the sequence in which they should be applied."*
- *"The conflation of the results of each of these methods to evaluate the level of either support or rejection of the formulated propositions."*

Even though the current version ([ENFSI15]) of the BPM actually is essentially a 65-page knowledge base document, it also acknowledges its political purpose to "*provide a helpful bridge between the requirements of international and local regulatory standards, and the actual implementation within each member's laboratory environment*". In synopsis, it states that the efficient handling of international crime requires compliant forensic investigations that can assist efficient prosecution of these crimes.
Since the landscape of national forensic laboratories in Europe is very inhomogeneous due to the wide variance in judicial systems (including common-law countries like the UK and the Republic of Ireland as well as civil-law countries, i.e., most of the rest of Europe), the BPM [ENFSI15] limits its ambitions in discussing SOPs as follows:

> "*In order to ensure the maximum compatibility with the requirements of all member laboratories, the document does not describe in a step-by-step fashion how specific forensic processes should be completed, instead it details the abstract processes, the associated possible risks, and the potential size of errors that may exist.*"

In a summary of the primary goals of [ENFSI15], the document provides the following list:

- *"Promote the use of consistent methodologies;*
- *Encourage the development of new and novel methods;*
- *Facilitate information interchange;*
- *Acknowledge the existence of errors in all forensic methods; and*
- *Promote methods for use in risk analysis and risk mitigation."*

Of these five items, especially the first, second and fifth are of high significance for this habilitation project.

All recent ENFSI BPMs share a common structure, governed by a common template (most recently [ENFSI22b]): the scope of the BPM, definitions and terms, resources (including personnel), methods, validation and estimation of uncertainty of measurement, quality assurance, handling of items, initial assessment, prioritisation and sequence of examinations, evaluation and interpretation as well as presentation of results.
In the following sections, selected content from some of these categories directly relevant for this habilitation treatise is summarised from the ENFSI BPMs FIT [ENFSI15] and DIA [ENFSI21] as well as (to a much lesser extend) the BPM on ENF analysis [ENFSI09] and the BPM for Digital Audio Authenticity Analysis (FSA) [ENFSI22a].
These categories are then also used to structure the conclusions drawn in this cumulative habilitation treatise in Section 4.2.

2.3.4.1 Personnel

Following the common structure of the BPMs, the first item of relevance for this part of the habilitation treatise is the aspect of personnel (covered as part of the resources). Here, [ENFSI21] summarises the requirements for personnel as follows:

> "*All personnel participating in* [the] *examination should be proven to be qualified to perform the examination. At each organisation, the local quality management system should clearly describe how such proof can or should be provided and documented. The periodicity with which this proof and documentation should be re-evaluated should also be described.*"

To summarise the typical structure of a team performing a forensic examination, three potential roles are identified in [ENFSI21]: the 'case leader', who "*should select and prioritize the tasks, assign each task to one or more appropriate Examiners, and finally collect and interpret results before presenting them in a report and/or in court.*", the 'examiners', and 'third parties'[26]. [ENFSI22a] states that all personnel should "*have received specific forensic training* [...]. *Examples of appropriate training include: Laboratory in-house training; Training from a university or equivalent; Training from an external certified organization.*" Two aspects of this statement are directly relevant for this treatise: firstly, the inclusion of universities (and therefore research-oriented education) in this list, and secondly, the fact that there is no particular emphasis on certification in relation to the other options, which is a noticeable difference when compared to the situation in the US (as discussed above in Section 2.3.1).

Regarding the proximity to research, ENFSI BPM FIT [ENFSI15] calls for even stronger ties between forensic practitioners and the corresponding research community: It distinguishes between "*technical (with administrative responsibilities)*" and "*administrative only*" personnel. For all technical personnel, it is stated that they "*have a direct responsibility to ensure they:*

- *Comply with national regulatory requirements;*
- *Are up to date with current technical developments and procedures;*
- *Understand the requirements of the criminal justice system;*
- *Maintain a portfolio of evidence demonstrating a participation in cases involving digital technology/digital evidence;*
- *Read journals, books and other literature containing pertinent information relating to forensic digital evidence examinations;*
- *Provide formal feedback to colleagues on problems encountered during analysis and the method that was employed to overcome it;*
- *Aid in the development of local procedures and standards and improve the technical advancement of examinations*"

Furthermore, "[t]*hey should also aid the quality management through development and critical peer review of proposed changes to local procedures and standards to improve the technical advancement of examinations within the forensic environment.*"

Up to this point, the statement is consistent with the other three ENFSI BPMs summarised in the context of this habilitation treatise. With the following requirement, however, the ENFSI BPM FIT [ENFSI15] significantly expands the expectations regarding the practitioners (case leaders as well as examiners):

> "*They should also take part in appropriate workshops, seminars, conferences, meetings and research and development projects.* [...] *Technical Experts should actively participate in casework examinations, and also participate annually in at least one of the following:*
>
> - *Publication of a technical paper in a recognised peer reviewed forensic journal related to digital technology/evidence;*
> - *Presentation of a paper or specific casework experience at a professional meeting/seminar;*
> - *Technical training events as a presenter/instructor;*
> - *Routinely communicate the relevance of selected forensic topics within the digital technology/evidence forensic community and the laboratory.*"

This direct request for close interaction with the (academic) research community is of significance for this habilitation treatise, because it shows that a significant expert group in this field is not only strongly motivated to interact with academics but also willing to involve themselves in the relevant activities of academic researchers (i.e., publication of papers and acquisition and execution of research projects).

[26] The role of third parties in the examination process is described differently in different ENFSI BPMs. In [ENFSI21], the third party is mostly involved in the initial assessment and in the handling of items, while in [ENFSI09], third-party subcontracting is provided for:

> "*In the event that no personnel within the laboratory are competent to be the Technical Specialist on a specific case or specific technical aspects relating to ENF, arrangements should be made for a qualified and competent consultant/contractor to be utilised from outside the laboratory to perform these duties.*"

2.3.4.2 Classes of Methods

The second item relevant here is the specification of classes of methods, whose modelling in the ENFSI BPMs differs from the BSI 'Leitfaden IT-Forensik' discussed in Section 2.3.3 above. Here, each BPM provides its own, context-dependent classification scheme, without a generalising common scheme. The closest thing to an operational model, as discussed in this habilitation project, is found in ENFSI BPM FIT [ENFSI15] with the high-level abstract analysis sub-process shown in Figure 3.3 of this BPM, sub-figure (a) and its split into 'human-based functions' and 'instrument-based functions'.

In relation to (academic) research on novel methods in the forensic sciences, the BPM FIT [ENFSI15] positions itself as follows:

> "*Research plays a fundamental role in both the development of new and novel techniques, and ensuring that existing techniques remain fit-for-purpose.*
> *Whilst it is true that in the area of forensic IT the technology and applications are constantly evolving and expanding, the foundations on which the discipline and its derivatives are based – computer science, physics (electronics) – are relatively fixed, and ultimately supported by mathematical theory.* [...] *Therefore, before going to the expense of developing and attempting to prove a new technique, time should be employed to research if the technique (or near equivalent) has already been suggested and perhaps even deployed by another laboratory.*
> *It is also important to critically cross-reference any source before attempting to use a reference as an axiom on which the laboratory bases its validated processes.* [...] *Best practice should always be to seek methods supported by axioms which have been rigorously tested* [...], *or even better rigorously proved.*"

While most of this statement obviously applies to IT forensics, especially the wariness regarding insufficiently tested approaches and 'junk science', it cannot be generalised for the much more recent and less mature sub-domain of media forensics, where proving the validity of an analysis method is much harder. The reasoning for this is provided by ENFSI BPM DIA [ENFSI21], with a statement about the (potential) uncertainty of measurements in media forensics analyses, and is cited in this document below on page 34.

[ENFSI21] contains the most complex classification scheme for methods of the BPMs considered here. It specifies four main categories of methods, each with a set of sub-categories. These four main categories are 'auxiliary data analysis' (described as "*methods based on auxiliary data (all data except the pixel data of an image)*"), 'image content analysis', 'strategy' ("*providing guidance on how to use these methods to perform typical authentication tasks*"), and 'peer review'. In the context of this treatise, it is relevant to mention that the listed set of methods in the sub-category 'image content analysis' contains many analysis methods based on pattern recognition.

The ENFSI BPM on ENF analysis [ENFSI09], which, in contrast to the other three ENFSI BPMs discussed here, focuses on exactly one analysis approach instead of a whole set of methods, emphasises that a clear distinction has to be made between a forensic method (e.g., ENF-based authenticity verification for digital evidence) and corresponding implementations.

2.3.4.3 Validation and Estimation of Uncertainty of Measurement

One main goal of the validation considerations is defined in [ENFSI09] as achieving precisely described, tool-driven and repeatable processes:

> "*For software tools that can be configured in a variety of ways and/or uses a number of different parameters, it is particularly important to document the set-up and individual parameter values in order to produce a process that can be repeated.*"

These reproducibility requirements are the same for 'manual analysis software' and for 'automated' (i.e., driven by pattern recognition) software solutions.

In FIT [ENFSI15], extensive considerations are made concerning the validation and estimation of uncertainty. An important aspect of these discussions is the distinction between verified and non-verified functions and tools, 'validated processes' and 'trustworthy processes'. The terminology used is specified as follows in [ENFSI15]:

- **Validation:** "*Validation relates to the ability of a process to meet the formal requirements agreed with the customer.*" The 'customer' in the ENFSI BPMs usually encompasses the judicial system and the investigating police officer, while [ENFSI15] explicitly points out that "*the requirements of the local judiciary shall ultimately take precedence*".
- **Verified functions in tools (as implementation of forensic methods):** "*Verification of functions within tools cannot, by themselves, be validated as the environment and ability of the user must be acknowledged as part of a process.*" Here, it is pointed out that this verification task is a shared burden in which many forensic practitioners are stakeholders:

  > "*International Standards and National Regulatory Codes of Practice promote flexible effective methods based on scientific proofs. They deliberately utilise abstract terminology in order that laboratories shall be able to create fit-for-purpose methods.*"

  Furthermore, it is not necessary to verify the entire tool/instrument, but only functions relevant to the corresponding type of investigation:

  > "*Due to the multiplicity of unused functionality that can exist within Forensic IT (FIT) instruments, and the complex ways in which FIT instruments may be combined to produce a result, validation shall be restricted to specific process (task or method) being undertaken.*"

- **Non-verified functions:** Regarding 'non-verified functions', the ENFSI BPM FIT recommends:

  > "*If no formally verified function is available to successfully complete an analysis stage then a non-verified function may be used. It is however, important to demonstrate that it provides results which exceed those capable from the verified functions available.*
  > *In cases where no verified equivalent functions are available to help make the comparison, then a far more detailed evaluation (with greater management overhead) will be required. In effect the analyst will need to verify the functionality used.*"

  To make sure that such use of non-verified functions does not turn the entire forensic investigation (respectively its result) invalid, [ENFSI15] states that "[i]*f a non-verified tool function is routinely used then it is expected that it should undergo formal verification, and be added to the laboratory approved list.*" Such use of non-verified functions might be very likely for the domain of media forensics, where so far no commercial and certified tools, but only lower TRL prototypes exist for many potentially relevant investigation approaches.
- **Validation techniques and procedures:** When handling complex (software) systems, the following guidance is provided:

> "*The overarching guidance around the development of validation techniques and procedures within this document is to sub-divide seemingly large (monolithic) complex systems into smaller, and hopefully simpler, (atomic) components through the use of black-box abstraction methodology.*"

This concept of component-based modelling is used in Chapter 3.2.3 as one of the core elements of the modelling approach discussed for media forensics processes.

- **Validated process:** A definition of the term 'validated process' is given as follows:

> "*A validated process is one which demonstrably conforms to its statement of requirements. A technical process may be considered to be validated for a particular purpose if, when tested, it meets the stated requirements for that purpose.*"

To make the creation/composition of validated processes easier, they "*can be constructed using a combination of smaller sub-processes and functions.*" For these, the "*verification of functions should be limited to those specific to the process, rather than attempting to verify all the functions available within a tool.*"

- **Trustworthy processes:** The following description is presented for 'trustworthy processes':

> "*In order to create trustworthy processes, verification* [...] *will be required to validate the developed process, and also demonstrate that the user and instrument functions used do actually operate within the bounds of known risks and their errors.*"

This verification should not be understood as a one-time effort. Re-verification has to be provided for the operational procedures, either on a routine basis (e.g., once a year) or event-based (e.g., when new technological advances occur that might have an impact on the reliability of established methods). Acknowledging the potential complexity of investigation processes, [ENFSI15] proposes to break down complex processes into sub-processes for verification purposes:

> "*To help reduce the cost, and effort, of verifying large processes it is recommended that processes are subdivided (atomised) into smaller encapsulated sub-processes.*"

A similar advice is given for the function verification of forensic tools:

> "*Function verification of forensic tools is concerned not with the verification of the functionality of the entire tool, but instead the verification of only those functions within forensic tools which are used within validated processes.*"

Furthermore it is highlighted that every verification of trustworthy processes must be "*conducted in-situ*" (i.e., in the lab environment in which the processes are supposed to be used and with the corresponding users that are expected to be the examiners using these processes in actual analyses) and that it "*is the responsibility of each laboratory to verify their specific methods and systems based on their formal local implementation.*" This last issue also includes the validation of requirements for the chain of custody as well as the documentation of an investigation.

Not only automated processes are within the scope of the verification and re-verification work to be performed. In [ENFSI15], human-based methods are also specifically included:

> "*Human-based functions are the pivotal elements within technical forensic processes, all forensic processes are likely to require user interaction, therefore an evaluation of user capability must be made as part of validated process within the laboratory. Even if an instrument-based function returns a valid result, it may still be reliant on the correct interpretation by the user associating the result.* [...] *Verification of human-based (user) functions are covered within proficiency testing* [...]"

Even though the availability of the required forensic practitioners with sufficient training and valid certification (if applicable) is such an important factor in every forensic investigation, addressing this issue is outside the scope of this habilitation treatise. Here, it is simply assumed that a lab has the required human resources available and that their competence is (re-)evaluated as required (e.g., in proficiency testing) or that it has the possibility to outsource such investigations that cannot be covered sufficiently well with its own personnel[27].
The whole issue of the validation of tools and processes is a necessary part of the risk assessment required for case handling. [ENFSI15] states on this issue:

> "*For the interpretation of evidential significance in the context of the case, a laboratory should always consider the use of techniques and equipment whose risks have been formally assessed; as part of the required functional verification, in preference to those which have not. This does not mean that a method or process that has not been formally evaluated cannot be used to aid the analysis; rather it means that if there is a wish to use such a solution, a formal justification as to why it has been chosen in preference to one that is part of a validated process must be made.*
> *When designing a validation process, five key elements of a successful validation policy are:*
>
> 1. *An understanding of known errors and uncertainty*
> 2. *The Statement of Requirements;*
> 3. *Risk Analysis and Assessments;*
> 4. *Effective validation test sets; and*
> 5. *Routine verification.*"

From this list, items 4 and 5 are more or less self-explanatory, while the discussion on items 1, 2 and 3 needs additional explanations, which are provided in the following for 'An understanding of known errors and uncertainty', 'Statement of Requirements' as well as 'Risk Analysis and Assessments':

The **'understanding of known errors and uncertainty'** requires an additional specification of the term 'uncertainty', which in [ENFSI15] is given as "*the unknown (random) difference (delta) between the measurement taken and its true value. It can never be completely defined, or eliminated, and is represented as a bounded region in which the true value exists within its given confidence level.*" In complex systems, uncertainty aggregates:

> "*Uncertainty within a system is additive in nature, and generally increases with the number of functions deployed within a process. The decision as to whether the uncertainty should be calculated at the function level or abstracted to the process level is at the discretion of each laboratory.* [...] *Software solutions will also contain additional uncertainty on top of the uncertainty associated with the physical systems, including the operating system, they are running on. This is especially true for software which relies on functions with no formal specification and/or calibrated standard. As a result, software uncertainty properties will also need to be acknowledged and accounted for.*"

Regarding the considerations on "*uncertainty within image authentication*", BPM DIA [ENFSI21] identifies three domain-specific potential factors as "*tool inaccuracies*", "*operator inaccuracies*" and "*data inconsistencies*". Acknowledging that these factors are interlinked, the BPM DIA elaborates: "*Given*

[27] In contrast to [ENFSI15], where it is implied that all investigations would have to be performed by internal personnel, [ENFSI09] also foresees the potential outsourcing of analyses to other laboratories, including non-governmental (i.e., commercial) contractors.

the intricate dependencies which could exist between uncertainties that arise at various points during the image authentication analysis procedures, the uncertainty attached to a specific measurement cannot always be quantified."

The **'Statement of Requirements'** is defined in [ENFSI15] as follows:

> "*The statement of requirements defines the problem to be solved by a technical process. It should provide explanatory text to set the scene for a lay reader, summarising the problem, noting the scope and acceptable risks or limits of any solution and acknowledging the relevant stakeholders. It should be created independently of and without regard to any particular implementation or solution.*"

Furthermore, the statement of requirements "*provides the interface (or formal bridge) between what the customer believes is achievable (customer requirements), and so desires, and what the laboratory can realistically achieve (laboratory capability) with the available staff, tools and the incurred time costs.*"
Ideally, this statement of requirements is not only a list with a set of needs and corresponding associated constraints and conditions but also includes a "*list of well-formed, testable requirements.*" [ENFSI15]
In the ENFSI BPM FIT, a list of types of such requirements is presented as an example, including functional and performance requirements as well as requirements focusing on the interfaces for the solution, its compliance with local laws and processes, etc. In addition, it is stated that, "[i]*f the risks are considered too great then either the statement of requirements will need to be amended, or alternate solutions sought, to reduce the risks to acceptable levels.*". This negotiation takes place between forensic practitioners (i.e., the laboratory represented by the case lead) and their 'customers'. It basically determines which methods are to be used in a forensic examination to be conducted, based on customer requirements such as "*agreed timeframe*", the type of methods to be used (only validated vs. validated and un-validated) or "*general risks associated with a case*" [ENFSI15]. Summarising this negotiation process, [ENFSI15] states that

> "[t]*he information described within the final 'Statement of Requirements' will form the basis on which the process being validated will ultimately be judged as either a pass or fail. Therefore, it is very important that the defined requirements are both accurate and realistic with respect to standard scientific principles and current available methodologies.*"

This generalising statement somewhat obscures the fact that this negotiation would have to take place for each and every forensic investigation to be conducted. A set of user requirements that is valid for one case might not apply at all in another investigation. As an example, the time frame allowed for an (initial) forensic assessment of some pieces of evidence might, in case of an ongoing kidnapping situation, be much shorter than usual.

For the **'Risk Analysis and Assessments'**, [ENFSI15] states that "*risk analysis and verification stages are paramount in creating a reliable validation method*", with the BPM providing a very general description of how to perform such a risk analysis and how to record/document the risk in a formal assessment process. Different examples of corresponding evaluation questions to be used within such an assessment process are provided, including method-specific questions, implementation-specific questions as well as questions regarding the lab's organisational procedures the use of methods within a process. An example for the first category would be: "*Does it operate correctly for its intended purpose?*" (This is equivalent to one of the Daubert criteria; see Section 2.3.1.) A good example for the second category would be: "*Does it operate correctly in its working environment(s)?*" (This is roughly equivalent to one of the FRE702 criteria; see Section 2.3.1.) An example for the third category would be a question such as: "*Is routine re-verification conducted?*"
Summarising the discussion on risk analysis, [ENFSI15] states:

> "*Risk analysis can not only be used to explain why a verified function has been used within a validated process, but also why in certain circumstances a formally unverified function has been chosen in preference.*"

Using the terminology of the BSI 'Leitfaden IT-Forensik' (see Section 2.3.3), the risk analysis would have to be performed in Strategic Preparation as well as Operational Preparation for each case.
The ENFSI BPM FIT [ENFSI15] explicitly integrates the competence of the forensic practitioner(s) available to handle a case in the risk analysis ("*The lower the level of knowledge* [of the analyst], *the greater will be the potential errors and risks.*"). But experienced analysts might also encounter challenges when interpreting the output of verified functions. In this case, the escalation procedure recommended is the following:

> "*If a new, unknown, discrepancy is detected then the evaluation will need to be highlighted for the peer review, and one or more of the verified tools may need to be reassessed, along with the existing validated process.*"

With regard to the use of non-verified functions, which is a very likely scenario for certain media forensics investigation methods that still lack maturity and for which only lower TRL solutions exist so far, the recommendation of [ENFSI15] for the corresponding risk assessment is the following:

> "*When using a non-verified function during analysis it is important that the analyst is competent enough to research the characteristics of the returned results, and can qualify them against standard validation methods employed within the laboratory* [...]"

The BPM DIA [ENFSI21] closely follows the validation principles established in FIT and expands them accordingly for the sub-domain of digital image authentication. A nice example-driven set of minimal requirements for performing a method validation in that domain is given in [ENFSI21] as:

- "*An outline of the applied methods and their use cases (e.g., for PRNU: a general description of PRNU-based source camera identification and when it is applicable).*
- *A detailed description of the process, such as in which order, which tools and functions are applied and with which settings (e.g., for PRNU: a description of how the camera's sensor pattern was extracted, how the correlation threshold was determined).*
- *A collection of rules to ensure that known restrictions, errors and flaws of the used tools do not adversely affect the results, and that the quality of results is optimised according to the given conditions (e.g., for PRNU: specifying the minimum number of reference images required, how to handle saturated images, details of limitations on the supported geometrical transformations, and potential issues related to multiple-camera devices, etc)*
- *A dataset with known source, recording conditions or processing operation should be used for (re)validation tests to check if the method gives the expected results (for instance to check that different software gives comparable results).*
- *A validation report.*"

2.3.4.4 Quality Assurance

Another item of relevance here is quality assurance, with considerations on proficiency testing and quality controls, which is very closely related to REQ2 and REQ4 of this treatise. Regarding the performance of individual examiners, BPM DIA [ENFSI21] recommends that quality controls including regular proficiency tests are put in place "*in order to mitigate against bias within the examination*". For safeguards on the processes and methods used, BPM DI [ENFSI21] recommends the use of a quality management system, defining its purpose as "[a]*ssuring the use of valid methods*".
The high significance of the availability of qualified personnel is underscored in [ENFSI15] with the following statement:

> "*Internal proficiency tests should be designed to provide useful feedback to the laboratory to help continually verify that the existing laboratory process human-based risks remain within acceptable bounds. If the user trend deteriorates, then either the risk assessment(s) must be adjusted accordingly or the process re-validated. If a proficiency test highlights a problem with a process, or a specific function within a process, then that may also indicate that there is a problem with the associated current validation or verification process.*"

2.3.4.5 Case Assessment / Initial Assessment

In the terminology of the BSI 'Leitfaden IT-Forensik' [BSI11], this is the stage of an investigation when the Operational Preparation (OP) is performed. It involves selecting the 'case leader' for the involved forensic laboratory and pre-scene preparation as well as the (potential) assessment at the scene. In FIT [ENFSI15], it is stated for the latter that the "*assessment at scene in this context also extends to the support and advice provided remotely to those that are at the scene so that submitted exhibits can later be effective processed within the laboratory.*" An important point raised by BPM DIA [ENFSI21] is that a vital part of the case-related documentation (including the chain of custody) starts at this point (i.e., the seizure of evidence by the authorities) and that it is of uttermost importance for the evidential value to ensure that this documentation is complete.

2.3.4.6 Evaluation and Interpretation

In ENFSI BPM FIT [ENFSI15], it is pointed out that

> "[a]*n understanding of how both the original application and the forensic tool interpret the data is necessary in order to scientifically evaluate and interpret the findings. The lower the level of knowledge* [of the examiner], *the greater will be the potential errors and risks.*"

Based on this statement and acknowledging aspects that impair performance, originating from different domains, the ENFSI BPM FIT identifies potential error sources (which have to be reflected upon in the result interpretation) as a "*combination of*:

- *The combined errors of the processes and measurements used;*
- *The time constraints to analyse the data;*
- *The analyst assigned to the case;*
- *The depth of detail in the case requirements; and*
- *The type and quantity of evidence located.*"

BPM DIA [ENFSI21] emphasises the impact of the analyst/examiner assigned to the case by pointing out that every forensic method applied in analysis usually involves "[result] *interpretation by the Examiner*", i.e., that in every case, the human examiner is in control of the method (even for approaches based on machine learning). Therefore, an effort has to be made to prevent problems caused by lack of training or personal bias introduced by the examiner to the case. Especially the latter is a non-trivial task, which is addressed in BPM DIA for the overall interpretation of findings and formulation of conclusions by relying on the organisational split between the case leader and the examiners on the case:

> "*During the evaluation stage all findings from the different elementary methods are evaluated by the Case Leader, resulting in a conclusion that states the evidential weight as a level of support for each one of the competing propositions. Some results of operations on images can be assessed independently, but many results have to be compared with other results to deliver evidential value. In this stage the Case Leader should also consider the background information* [...]"

The fact that the examiners usually do not have access to this background information about the case is intended to limit the bias they introduce during the analyses performed. To facilitate the conclusions drawn by the case leader, the examiners have to provide their results as a level of support for one of the competing propositions in a format that allows for a suitable combination. In BPM DIA [ENFSI21], the current practice is summarised as follows:

> "*Support levels are typically reported using a graded scale. Currently, there is no universally accepted scale for reporting* [image analysis] *conclusions and there is a wide range in scales used by different agencies. The ENFSI member laboratories are expected to comply with the ENFSI Guideline for Forensic Evaluative Reporting* [...] *which recommends both to*

> *use the likelihood ratio (LR) as an indication for the level of support (often referred to as the strength of evidence), and a graded scale to associate verbal expressions to numerical values, where required.*"

The main problem associated with this procedure is the computation or estimation of likelihood ratios and discriminating power. For the likelihood ratios, BPM DIA [ENFSI21] states: "*The assignment of a precise quantitative likelihood to any of the examination findings in* [image analysis] *is often impossible*", with reasons given including the nature of some methods (which might only admit qualitative evaluation), the lack of adequate reference data, and the fact that with some methods, the estimation of probabilities might be subjective, based on the experience of the examiner. For the discriminating power, ENFSI BPM DIA [ENFSI21] states:

> "*In image authentication, establishing the discriminating power of an elementary method is often challenging. While the performance of each elementary method is often evaluated and reported in the corresponding scientific paper, the testing conditions in such experimental evaluations are typically very different than those encountered in casework. Therefore, it is necessary for the Examiner to understand the discriminating power of an elementary method in the circumstances of the particular case. In order to accomplish this, an Examiner could:*
>
> - *Obtain or create reference items* [...] *which reflect* [...] *as close as possible the current examination, and establish performance of the method on such material.*
> - *Investigate the performance of the elementary method on available datasets and gather information on its discriminating power. This investigation should reveal the influencing conditions (e.g., parameter settings of this method or properties of the image) that may give rise to false negative and false positive results.*
> - *Examine the behaviour of the elementary methods with respect to findings from other similar features within the questioned image (e.g., for local analysis methods).*"

The first two items in this list focus on the benchmarking of forensic methods (respectively their implementation in tools integrated into forensic processes) and relate to REQ2 and REQ4. The third item concerns fusion / the combination of expert systems (relating to REQ1, REQ2 and REQ4). Both strategies are addressed in this habilitation treatise.

2.3.4.7 Presentation of Results

The forensic expert's role in this context is summarised in [ENFSI15] as follows:

> "*The overriding duty of those providing expert testimony is to the court and to the administration of justice. As such, evidence should be provided with honesty, integrity, objectivity and impartiality. Evidence can be presented to the court either orally or in writing. Only information which is supported by the examinations carried out should be presented. Presentation of evidence should clearly state the results of any evaluation and interpretation of the examination.*"

Regarding the form, [ENFSI15] specifies: "*The findings, and any expert opinion, are normally provided in the first instance in written form, as a statement of evidence or a report, for use by the investigator and/or the prosecutor/court.*"
In addition to this primary perspective (and its forms of expression), statements in other forms might also be requested from a forensic expert. Regarding actual casework, [ENFSI15] summarises such other statements as follows:

> "*Investigative reports and opinion, within this context, relates to officer specific applications where the information may not be designed to such stringent levels as those that are required for court review / use. This may be due to the requirement that information is needed urgently, such as in the case of a finding a missing person who is considered at risk, and where the time constraint is the most critical factor.*"

The BPM DIA [ENFSI21] is rather short on this issue, pointing out that

> "[t]*he way of reporting may vary depending on national legal stipulations or requirements. Nevertheless, the overall reporting process should still enable independent review or reproduction of the reported results.*"

This request for reproducibility is a strong driver for the modelling work discussed in Chapter 3 and relates to REQ3 of this habilitation treatise.

2.3.4.8 Tool Development

The ENFSI BPM FIT [ENFSI15] considers custom tool development for individual cases as well as the need for industrial-strength implementations at high technology readiness levels[28] (TRL) supplied by commercial software developers. For the case of custom tool development, it states:

> "*The development of scripts and software routines for use within a specific case shall also be classified as non-verified functions. In addition to including the software code within the case archive, it is also essential that a copy is retained within the laboratory software register* [...]. *If a non-verified tool function is routinely used then it is expected that it should undergo formal verification, and be added to the laboratory approved list.*"

For the topic of software development for forensic tools, the ENFSI BPM FIT [ENFSI15] has a separate appendix ('Appendix B – Custom (bespoke) development') that focuses on the question of "*whether to purchase a 3rd party product or develop a custom solution*", where the third-party option also includes the commissioning of an external developer with targeted tool development. Reasons and arguments are provided concerning cost efficiency of development options, software development practices (including software engineering, oversight and testing) as well as the required verification of such software.

2.3.4.9 Synopsis for the Discussion of Selected ENFSI BPMs as Background for this Treatise

The perspectives of the different ENFSI EWGs manifest in the corresponding Best Practice Manuals (BPMs) are fairly diverse. They do, of course, share a common perspective on many aspects, such as the relationship between the lab and the 'customer' and the basic understanding of the conduct of forensic processes, but they also differ in many significant points. This is not surprising, considering the varying scope, with some providing recommendations for a field as large as the forensic examination of digital technology and others looking only at exactly one analysis method (e.g., electric network frequency (ENF)). Furthermore, the maturity of the corresponding field is reflected in the document perspectives: FIT [ENFSI15] includes methods that have been in use for 50 years, while the whole field of digital images and methods for their authentication considered in DIA [ENFSI21] has emerged much more recently. In summary, the following main points from the discussed BPMs form the relevant background for this treatise (with the respective requirements REQ1 to REQ5, as defined in Section 1.3, identified where possible):

[28] The EU HORIZON 2020 – WORK PROGRAMME General Annexes define the Technology Readiness Level (TRL) as follows (see: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf):

- "*TRL 1 – basic principles observed*
- *TRL 2 – technology concept formulated*
- *TRL 3 – experimental proof of concept*
- *TRL 4 – technology validated in lab*
- *TRL 5 – technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)*
- *TRL 6 – technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)*
- *TRL 7 – system prototype demonstration in operational environment*
- *TRL 8 – system complete and qualified*
- *TRL 9 – actual system proven in operational environment*"

- The separation of roles as proposed in [ENFSI21]: The proposed split into task lead and examiners should significantly reduce bias in the investigations and is also an established best practice in many other fields of applied forensics, e.g., dactyloscopy. (REQ3)
- All BPMs emphasise the relevance of modelling - for conceptual models as basis for the discussion of types methods (as used extensively in ENFSI BPM DIA [ENFSI21]), for source model generation for authenticity and integrity verification (as done in the ENFSI BPM for Digital Audio Authenticity Analysis [ENFSI22a]) as well as for operational models (e.g., as basis of the classification of 'non-verified functions', 'verified functions', 'validated processes' and 'trustworthy processes' in ENFSI BPM FIT [ENFSI15], see Section 2.3.4.3). (REQ3)
- The discussions about validation (again, with the distinction between 'non-verified functions', 'verified functions', etc.) and uncertainty as well as quality assurance. (REQ1, REQ3)
- The (requested) close relationship between forensic practitioners and the corresponding scientific communities (see Section 2.3.4.1).
- Discussions about tool development (see Section 2.3.4.8). (REQ5)
- The issues concerning validation (benchmarking) of forensic methods (respectively their implementation in tools integrated into forensic processes) as well as the corresponding quality assurance (including proficiency testing). (REQ1-REQ4)
- The role of human experts (i.e., the examiners and case lead) in the forensic procedures, especially with regard to the evaluation, interpretation and presentation of results, with the corresponding responsibilities as well as the requested accountability (see Sections 2.3.4.6 and 2.3.4.7).
- The request for reproducibility of evaluation results as an important aspect of the presentation of investigation results (see Section 2.3.4.7). (REQ1, REQ2, REQ4)

2.3.5 The European Union (EU) Artificial Intelligence Act (AIA)

In addition to the efforts by practitioners in the field of law enforcement and forensics, there are other noteworthy activities in the EU that are of relevance for this field. Not all of them can be reflected upon in this habilitation treatise. One that has to be mentioned, however, is the regulatory push that results from EU legislation, and especially the EU Artificial Intelligence Act (AIA), also known as AI Act[29]) with its strong emphasis on the requirement of 'human-in-control' for critical AI applications (e.g., in the context of law enforcement and forensics as considered within this treatise).

The discussions about the regulation of AI-based solutions in the EU are, of course, accompanied by national discourses in the member states. On the national level in Germany, the BSI (as national cybersecurity authority) advances the discussion in this field through a series of whitepapers, including publications like 'Towards Auditable AI Systems - From Principles to Practice' [Berghoff21] and '*Sicherer, robuster und nachvollziehbarer Einsatz von KI - Probleme, Maßnahmen und Handlungsbedarfe*' [BSI21] (title translated into English as 'Secure, robust and traceable use of AI - problems, procedures and actions required'). The latter presents an in-depth discussion of problems, policies and procedures as well as open issues discussed by experts in this field. A brief summary of this discussion (with its focus on the three aspects of: a) development of standards, technical guidelines, test criteria and test methods, b) research on effective countermeasures against AI-specific attacks, and c) research into methods of transparency and explainability) is presented in feeder paper **[Kraetzer22]**, included as Chapter 12 of this cumulative habilitation treatise.

Synopsis: Modern-day media forensics is strongly driven by pattern recognition (respectively Artificial Intelligence (AI)), and with the (upcoming) AIA, new regulations will become effective in the EU and its

[29] Note: At the time of writing of this treatise, the final version of the AIA has not yet been published. The status reflected here is the provisional agreement on the Artificial Intelligence Act reached by the EU Parliament and EU Council negotiators on December 9th, 2023: `https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai`. The next steps would be the formal adoption of the agreed text by both Parliament and Council to become EU law. As part of this ongoing process, Parliament's Internal Market and Civil Liberties committees will still have to vote on the agreement.

member states that will significantly affect the design, implementation and use of AI-driven methods. The foreseeable changes initiated by the AIA will have to address the definition of standards that are suitable for assessing the security and reliability of AI systems, the design of security benchmarks to ensure a secure and robust operation of AI systems, and the research into methods for ensuring transparency and explainability of AI system decisions.

What holds true for every form of AI use is obviously also important when it comes to AI-driven processes that are (by regulation) restricted to decision support systems, e.g., in the case of forensics, where the internationally accepted standard is that investigation results have to be interpreted through expert testimony (see Section 2.3.4.7). Here, the corresponding expert has to be able to explain the investigation method as well as all aspects influencing the investigation outcome before a trier of fact (in most cases a judge, a group of judges or a jury). Besides other reasons, this human presentation and interpretation is considered necessary because the expert is also able to interpret contextual information to reason about the intention of an action (e.g., why a DeepFake video was created), which is a challenge where AI alone will fail. These issues relate to requirements REQ1, REQ2, REQ3, REQ4 and REQ5 as specified in Section 1.3.

2.3.6 Other Relevant Aspects of the European and German Situation

The feeder paper **[Kraetzer21]**, which is included as Chapter 9 of this cumulative habilitation treatise, provides a short summary of the discussions on requirements for media forensics methods in terms of scientific admissibility, briefly comparing the European and US perspectives. This short summary is to a large extend based on the publication ‘Scientific Evidence in Europe - Admissibility, Evaluation and Equality of Arms’ by C. Champod and J. Vuille. In their work, Champod and Vuille [Champod11] state that

> “[t]*he scientific admissibility of evidence, while subject to fairly precise rules in United States law,* [...], *is seldom addressed in European legal writings* [...]. *The question of scientific reliability is seen as intrinsically linked with the assessment of the actual evidence, that is with the determination of its probative value* [...].”

This implies that researchers in the fields of (media) forensics and applied pattern recognition have to rely on the verdict of the ‘customer’ (to use the ENFSI terminology), i.e., a judge or other legal expert for each individual case (and jurisdiction), defining the hurdles media forensics approaches have to take to achieve court admissibility. Summarising the discussions in [Champod11], it can be said that there is no EU-wide regulation on scientific admissibility questions, but that there are common principles that need to be considered. Some of them, expressing the perspective of one type of practitioner, are summarised above in Section 2.3.4, reflecting the forensic practitioners’ point of view as aggregated in the ENFSI BPMs.

In addition to the forensic experts, there is another prominent group of stakeholders involved in these considerations: the European law enforcement agenciess (LEAs). Organisations that are relevant in this context are Europol, INTERPOL, the national LEAs as well as integration actions between these, like the European Anti-Cybercrime Technology Development Association (EACTDA). The following selected examples from their work are relevant here:

- **Europol** (the European Union’s law enforcement agency with headquarters in The Hague, Netherlands): With its Innovation Lab hosting the Europol Tool Repository (ETR) as a LEA-exclusive online platform to share non-commercial, cost-free software developed by LEAs, and research and technology organisations as well as the Europol Platform for Experts (EPE) as an access-restricted expert forum.
- **INTERPOL** (or more precisely, the Office of the Special Representative of INTERPOL to the EU in Brussels): Focusing on the concerns of global law enforcement and the liaison with EU initiatives and policy decisions. An Interpol initiative that is especially relevant in the context of this habilitation treatise is the document series ‘Responsible AI Innovation in Law Enforcement’ (AI Toolkit). It defines in [INTERPOL23] “*five core Principles for responsible AI innovation*”

to "*provide the law enforcement community with a foundation for a principled approach to AI.*" These five core principles are identified as "*lawfulness, minimization of harm, human autonomy, fairness and good governance*". They are also relevant for the development and application of all solutions for (media) forensic tasks based on machine learning and therefore complement the recommendations given in the ENFSI BPMs on forensic tool development. What makes this document remarkable in the context of this treatise is the fact that it is very up-to-date in terms of AI-related issues, including multi-faceted considerations on technical aspects but also on legal, privacy, ethical and environmental aspects.

- **European Anti-Cybercrime Technology Development Association (EACTDA)**: A non-profit association for the development of technological solutions (i.e., tools) for European law enforcement agencies and forensic laboratories. It draws funding from the EU Security Research ecosystem and the EU H2020 programme and funds development initiatives like the Tools4LEAs project (https://www.eactda.eu/projects/Tools4LEAs/home.html) to take research project outcomes at low TRL and develop them further into tools fit for use by LEAs and forensic institutions. Members of the EACTDA collaboration framework are stakeholders such as the European Commission's department of Migration and Home Affairs (EC DG HOME), Europol, the European Union Agency for Law Enforcement Training (previously CEPOL) as well as national actors like ZITiS in Germany.

Synopsis: The situation in Europe might be significantly different from the situation in the US, but the requests for novel forensic methods from practitioners (here, LEA and forensic institutions) are currently meeting with a regulatory push for reliable and trustworthy AI methods (esp. in form of the EU AIA). A recent initiative illustrating that both should go well hand in hand is the document series "Responsible AI Innovation in Law Enforcement" (AI Toolkit; [INTERPOL23]) from UNICRI and INTERPOL.
As a consequence of these recent developments, trustworthy actors intending to take a leading role in the development of tools for LEAs as well as for forensic institutes have emerged over the last few years. One noticeable example on the EU level is EACTDA. On the national level, the Central Office for Information Technology in the Security Sector (ZITiS), founded in 2017, and (on a much smaller scale) Dataport, a state-owned institution under public law which develops, among other things, the digital case management systems for multiple German police agencies on the state level, take on a corresponding role in Germany.
This relates to REQ3, as specified in Section 1.3.

3 Work on Deriving Domain-Specific Forensic Process Models for Media Forensics

In this chapter, the approaches to conceptual and operational modelling for deriving domain-specific media forensics process models for selected application domains applied in the corresponding feeder papers by the author and his co-authors are summarised. These feeder papers are included in this cumulative habilitation treatise as chapters 5 ff. at the and of the document.

The work summarised here covers three application domains selected as examples: The first is **face morphing attack (FMA) detection** for digital images, which is covered in Section 3.1. For this application domain, the focus of the published research lies on conceptual modelling work. The second application domain is **DeepFake detection** for digital videos, covered in Section 3.2. For this domain, the focus shifts towards operational modelling. The third is **forensic steganalysis**, covered in Section 3.3, where the focus is again on conceptual modelling work.

The research presented here is heavily based on the Code of Practice for information technology (IT) forensics published by the German Federal Office for Information Security (BSI) in its '*Leitfaden IT-Forensik*' [BSI11] (see Section 2.3.3.1) and the additions to its process model published by others (see Section 2.3.3.2) as well as by the author himself. As established in Section 2.3.3.2, the BSI code of practice (and especially its phase-driven process model) is a suitable foundation for forensic modelling work and is (despite its age) still one of the most relevant documents in Germany to consider in this context. Nevertheless, various publications, like [Altschaffel20] have shown that it needs to be adapted in order to cover specific application contexts such as the media forensics context considered in this treatise.

A second strong influence of the work presented here are the Best Practice Manuals (BPMs) of the European Network of Forensic Science Institutes (ENFSI; see Section 2.3.4), especially the BPMs on Forensic Examination of Digital Technology (FIT) [ENFSI15] and on Digital Image Authentication (DIA) [ENFSI21].

3.1 Work Published in the Application Domain of Face Morph Attack Detection

In the context of this cumulative habilitation treatise, most of the conceptual modelling work by the author has been carried out in the context of the ANANAS project (see Section 1.2). This work supports the empirical work of the research group leader and colleagues on the topic of face morph attack detection for digital images. In this context, various abstractions (models) for media object (source) characteristics, morphing attack influences (of different morphing pipelines) on images, and detection models were provided by the author between 2016 and 2019. An overview of selected work on this topic is provided in this section, excerpting the work from the corresponding feeder papers and discussing it in the wider context of this treatise.

3.1.1 Modelling Media Generation Processes and Source Models

One of the most important tasks in media forensics is the modelling of the generative process that created the media object. The reason for this need is summarised perfectly in [ENFSI22a]:

> "*Authenticity analysis of digital audio recordings is based on traces left within the recording during the recording process, and by other subsequent editing operations. The first goal of the analysis is to detect and identify which of these traces can be retrieved from the audio recording, and to document their properties. In a second step, the properties of the retrievable traces are analysed, to determine if they support or oppose the hypothesis that the recording has been modified. It is not always obvious whether traces are due to recording or post-processing. A key objective of any authenticity analysis is therefore to determine whether observed features of a piece of audio evidence were introduced by the original recording process or by subsequent actions.*"

The structured analysis of the signal, the detection and identification of traces, and their documentation require a corresponding process model that fits the needs of such an investigation. In [ENFSI22a], an example of a simplified recording process flow model for audio recordings is presented, describing the typical traces left in an audio recording by the different components involved in the recording (including the environment, the microphone/transducer, A/D converter and encoder). This European Network of Forensic Science Institutes (ENFSI) BPM then discusses how traces of post-processing can be identified based on the modelled source characteristics: "*The main and most important characteristic of such traces is that they cannot be attributed to any part of the purported (claimed) recording process* [...]" [ENFSI22a]. Both kinds of traces (those attributed to the original recording process and those attributed to post-processing) are then used in a forensic process to support or refute "*a hypothesis that the evidence under analysis is an authentic recording, based on the characteristics of the traces within the recording and the available contextual information*" [ENFSI22a]. The work of the forensic practitioner on the case with regard to the results obtained is summarised in this ENFSI BPM as follows:

> "[I]*t should be clear that the goal of authenticity analysis is not to state which proposition is the correct one, but to evaluate which hypothesis is the more likely, and how strong (or weak) this support for that proposition is. It is possible to have no support for either case.*"

The simplified recording process flow model used in [ENFSI22a] for illustrative purposes is not fit for supporting in-depth analyses. For this purpose, a much more detailed process model for such a media generation process for audio material has been published by the author in his PhD thesis [Krätzer13]. Extended versions also modelling selected post-processing operations (like e.g., playback and re-recording in [Kraetzer12]) have been discussed by the author in further papers and a book chapter (see [Kraetzer15b]). Since this work on modelling media generation processes for audio recordings has already been used in parts in the author's dissertation project, it is excluded from further use in this habilitation treatise. Instead, the focus shifts towards conceptual modelling work started after the author's PhD thesis was published in 2013.

One feeder paper where such generation process and source modelling is carried out is **[Kraetzer17]** (included as Chapter 6 of this cumulative habilitation treatise). Based on the concept of a life-cycle model for photo-ID documents and well defined checks therein (see Section 3.1.2), the need for specific source models for these checks is argued. To address part of the need for such models, an image editing history model for face images is then introduced. This image editing history model formally describes the current state of a digital image by:

- describing the sequence of editing operations applied to the original camera image to obtain the current image, and
- aggregating knowledge about which traces are left behind in the image after applying each particular editing operation.

If the editing history of an image is known, the image can be analysed to determine which artefacts are produced by which sequence of editing operations. The aggregated knowledge gives a clue about which traces should be looked for in an image with an unknown editing history to reveal the presence or absence of particular image editing operations in its editing history. Here, the editing history model should provide support when deciding whether a face image is authentic or has been tampered with. This concept is consistent with the content analysis concepts later discussed in the ENFSI BPM for DIA [ENFSI21].

A visualisation of the descriptive image editing history model introduced in **[Kraetzer17]** is shown on page 90 (in Chapter 6). A specific editing history is represented in that visualisation by a path in the full-connected directed graph, with two specific nodes denoting the original (I_0) and the current state of an image (I_n). Other nodes correspond to image states after particular image editing operations. The current image in that history results from the propagation of the original image through the intermediate nodes, one in each layer. The set of image editing operations in each layer is the same. In addition to all relevant editing operations, the set includes a 'no op' operation to model the case of no image editing. An edge represents the parameters of the consecutive editing operations. It should be noted here that parameters can also include another image, e.g., for splicing operations. Formally, a current image is given by the following recursion: $I_n = (I_{n-1}, \text{editing operation: } E_n, \text{parameters: } p_n), I_0 = \text{original image}$.
Another important component of the model is the set of traces or artefacts that can be found in an image: $T = \{T_i, i = 1..k\}$.
In order to better describe the relation between an editing operation and traces in the image, the model is specified by introducing three attributes for each editing operation: *preserved*, *altered* and *acquired* characteristics or, more specifically, traces. Cropping, for instance, preserves camera-imposed fingerprints and the content of an image, changes the image dimensions and adds no new traces. A camera fingerprint as a trace is considered to be an element in T.

The set of image editing operations is divided into subsets of legitimate and illegitimate operations. An image is considered to be authentic if only legitimate image editing operations are present in its editing history. A single illegitimate operation in a path makes an image non-authentic. Detecting a non-authentic face image should raise an alarm in the Checks described in Section 3.1.2. For more detailed discussions on legitimate and illegitimate operations, the reader is referred to **[Kraetzer17]**, which is included as Chapter 6, pages 89 ff. in this cumulative habilitation treatise. In the paper, the image editing history model is integrated into a document life-cycle model (see Section 3.1.2) and used to provide illustrative editing graphs for two different types of face morphing pipelines for a precise description of the data generation pipelines used to generate the training and test material for the empirical evaluations in the paper.

Synopsis: The synopsis presented in 2017 in **[Kraetzer17]** and confirmed here for modelling media generation processes and providing corresponding source models is that researchers should be encouraged to use such formalism to gather knowledge about which traces are caused (or destroyed) by which editing operations (or sequences of editing operations). This perspective is coherent with the efforts discussed by ENFSI practitioners in the BPMs for image and audio material authentication ([ENFSI21] and [ENFSI22a] respectively), with the basics on media generation modelling provided there in 2021 and 2022.
The results presented for an image editing history model relate to REQ1 (as specified in Section 1.3). This conceptual model is then used in **[Kraetzer17]** to systematically describe the evaluation setups, which relates to REQ2. The classification of legitimate and illegitimate operations is part of an initial attacker model and relates to REQ3.

3.1.2 Media Life-Cycle (Usage) and Attack Modelling

In the feeder paper **[Kraetzer17]**, included as Chapter 6 of this cumulative habilitation treatise, a life-cycle model for photo-ID documents is introduced, and selected attacks on the documents themselves and on the digital images contained therein are discussed.

In **[Kraetzer17]**, the generalised process for the generation and use of an authentication token based on a face image is reduced to three core steps: Image data acquisition, document generation and document use

- In step 1, the face image is newly generated or acquired for re-use.
- In step 2, a person applies for a document, and all administrative and authentication steps that are necessary for commissioning the document creation are performed by the corresponding authority. The document is then created and picked up by or delivered to the applicant.
- In step 3, the document is used in authentication scenarios like border control, authentication at the desk of a car rental agency, etc.

All three steps are characterised in **[Kraetzer17]** by the following four-tuple data set: the acting entity (*AE*; a person, group of persons, or, in some cases, also automated processes) performing the necessary operations in the step, the identity (*ID*) assigned to the document (e.g., a passport number linked to a citizen of a country), the provided subject (*ProS*; equivalent to the intended validity set for the image or document under consideration; traditionally this set contains exactly one person, but in the case of the specific attacks discussed below several persons can be engaged), and the presented subject (*PreS*). The specification of subjects requires additional explanation: In the case of a morphing attack, there is a document that could be successfully used/provided by two (or more) subjects (*ProS*), but contains only one *ID* and can at one point of time (e.g., a check at a border control station) be presented by only one subject (*PreS*).

In addition to the steps and the entities, the model includes a third important component: The *Checks* that connect two steps in the pipeline (Check 1 as the assessment of a supplied face image for document creation purposes, and Check 2 as the validation of the document and the included face image during a border control event). In these Checks, various document-specific characteristics are evaluated, including, among others, authenticity (document as well as entity authenticity, the latter including a comparison of *ID*, *ProS* and *PreS*) and integrity checks as well as checks of the compliance with standards. **[Kraetzer17]** points out that at the time of writing of that paper, media forensic considerations were mostly neglected for these *Checks*, even though most of the example processes discussed in the paper already contained automated check components that could have benefited strongly from media forensic detectors. Addressing this gap and providing a strong motivation for the improvement of the *Checks* already in place by adding media forensic detectors was one of the goals of the ANANAS research project.
In **[Kraetzer17]**, the three misuse events selected as examples (presentation of a stolen document, document forgery and face morphing attack) are discussed and then compared using the life-cycle and attack modelling introduced. For details of this comparison, the reader is referred to pages 88 and 89 (in Chapter 6). In summary, only in the case of the face morphing attacks, the malicious operation happens before the document generation step and enables the (presumably criminal) attacker to obtain a 'clean', valid and un-tampered document issued by the official authority. This motivated the work on an image history model in **[Kraetzer17]**, which is summarised in this cumulative habilitation treatise as part of Section 3.1.1.
Which image manipulations have to be considered as belonging to the class of 'legitimate' operations or to the class of 'illegitimate' operations respectively depends on the nature of the corresponding Check and has to be specified accordingly in the life-cycle model. The basic principle for the decision was already introduced in Section 3.1.1: An image is considered to be authentic if only legitimate image editing operations are present in its editing history. A single illegitimate operation in a path makes an image non-authentic.

Synopsis: Media life-cycle models and corresponding attack models help to integrate different source models into the bigger context of a typical media life cycle. Source characteristics expected at specific points of time in a media life cycle (e.g., the camera image submitted for document generation, i.e., the input to Check 1 in that use case) might differ significantly from the source characteristics at other

points in the cycle (e.g., the ICAO-compliant passport image stored into a current eMRTD[30], i.e., the input to Check 2 in that use case). Corresponding authenticity and integrity checks therefore always have to consider the actual investigation context.
The results presented on the conceptual modelling of media generation processes and the corresponding source models in **[Kraetzer17]** relate to REQ1 (as specified in Section 1.3). The discussion of the different Checks and their constraints relates to the specification of different (targeted) forensic investigation processes to implement these Checks and therefore to REQ3.

3.1.3 Attack Detection Modelling

For the testing of hypotheses in media forensics, it is important to model the characteristics supporting the H0 hypothesis (i.e., the assumption that the evidence presents traces that the media object under investigation is authentic; see [ENFSI22a]) as well as the H1 hypothesis (i.e., the assumption that the evidence presents traces that the media object under investigation is not authentic). The first kind of traces is strongly related to the source model for the media object, and the second is often supported by specific attack models used in corresponding detectors. In the following sub-sections, detection modelling approaches from feeder papers are summarised, including considerations for single detectors, sequences of detectors, and multi-expert (fusion) detectors.

3.1.3.1 Attack Detection Modelling for Single Detectors

One description of an attack detection model (here used synonymously with 'manipulation detection model') for FMA is the one introduced in the feeder paper **[Neubert19]**. At its core lies a standard pattern recognition pipeline consisting of pre-processing (here a down-scaling to achieve compliance to ICAO requirements for face images to be used in Electronic Machine Readable Travel Documents (eMRTD)), feature extraction, and classification. The whole procedure is described in detail in the feeder paper included as Chapter 7 of this cumulative habilitation treatise (pages 101 ff.). The corresponding modelling of the three different kinds of face morphing attacks performed is presented in the same feeder paper on page 100.

One of the main reasons for attack modelling in ANANAS was the realisation that there are more than one possible technical implementions for the concept of face morphing attacks, and each of these potential realisations leaves characteristic artefacts in the generated media objects. So instead of a typical two-class problem ('genuine' vs. 'morphed'), a multi-class approach is necessary for FMA detection in practice (i.e., 'genuine' vs. 'morph type 1' vs. 'morph type 2', etc.).
In the feeder paper **[Neubert19]**, the differentiated attacks and the corresponding modelling of attack detection are used, among other research goals, to determine the impact of different morph generation pipelines on the detection performance when using two different feature spaces.

Synopsis: In many cases, the capabilities for providing traces supporting the hypothesis that the media object under investigation is not authentic rely on specific attack models used in corresponding detectors. As a consequence, the detection performance of such detectors in most cases strongly relies on the quality of these trained attack models, including the modelling of the classification problem (in the example above taken from the feeder paper **[Neubert19]** discussed as being either a 2-class problem ('genuine' vs. 'morphed') or an n-class model).
The results on attack detection modelling presented in **[Neubert19]** relate to REQ1 (necessary pre-processing methods for increasing the robustness of FMA detection adapted for an application scenario), REQ2 (evaluation of different potential influencing factors for the obtained error rates), REQ4 (re-evaluation of methods from **[Kraetzer17]**), and REQ5 (description of the feature space extensions performed and corresponding feature sub-space performance evaluations) as specified in Section 1.3.

[30] (Electronic) Machine Readable Travel Documents (eMRTD) are international travel documents (i.e., passports) compliant with the International Civil Aviation Organization (ICAO) Doc. 9303.

3.1.3.2 Attack Detection Modelling for Sequences of Detectors

One feeder paper significantly extending detection modelling from a single pattern recognition pipeline to a sequence of classification operators is **[Neubert18a]**. In this paper, a three-stage detection and verification sequence is introduced with the aim of reducing the false alarm rate (FAR) of the overall FMA detection. The sequence includes a state-of-the-art morph detector on the first level, which delivers a binary classification result ('morph' or 'authentic') for an input face image. The second-level component is the 3-class morph pipeline footprint detector newly introduced in that paper, which determines whether an image classified as 'morph' on the first level belongs to one of the three trained morphing pipelines ('complete morph', 'splicing morph' or 'combined morph'; see the feeder paper **[Neubert18a]** included as Chapter 8 of this cumulative habilitation treatise). The third-level component is a verification engine used to validate the classification result from the first level with the knowledge derived on the second level. It performs the final decision between the determined morph pipeline and the class 'authentic'.
The results of this conceptual extension from a single detector to a detection and verification sequence are empirically evaluated in feeder paper **[Neubert18a]** and show a significant reduction of the overall FAR by approximately 84% at the cost of additional run-time for the second and third classification and a slightly increased false missing rate (FMR), which rises by approximately 11%. In the feeder paper **[Neubert18a]**, it is argued that in the chosen application scenario (use of eMRTD at automated border control (ABC) gates), the decrease of the FAR is far more significant than the impact on the FAR (see page 111 in Chapter 8 of this cumulative habilitation treatise).

Synopsis: Sequences of detectors that incrementally generate and use knowledge about the media object under investigation can improve the overall performance of a media forensics investigation. The cost (increased run-time) is assumedly compensated by the benefits in terms of optimised performance and assumed better interpretability of a sequence of context-defined decisions.
The results discussed on attack detection modelling for sequences of detectors relate to REQ1 (a detailed image processing and feature extraction pipeline(s) description), REQ2 (a three-stage detection and verification sequence, with FAR and FMR discussions) and REQ4 (re-evaluating previously used FMA detectors), as specified in Section 1.3.

3.1.3.3 Attack Detection Modelling for Detector Fusion

There are many different approaches to defining fusion (or information fusion) in literature. In the context of this habilitation treatise, the term is best understood as the rule-based, automated combination of (independent) expert systems into one (media) forensic mechanism. In academic literature on various forms of security mechanisms, fusion has long been considered a significant means to increase the performance (in terms of decision accuracy) of pattern recognition systems. The rise of solutions driven by neural networks over the last few years has slightly changed the drive for fusion-based solutions. Nevertheless, even today, many established academic authors in media forensics consider fusion the only feasible solution to complex problems. A good example that supports this statement is the following quote taken from the recent 'Handbook of Digital Face Manipulation and Detection - From DeepFakes to Morphing Attacks' [Tolosana22] (as part of [Rathgeb22]):

> "*Recent studies suggest that no single feature/characteristic is adequate to build effective and robust detectors of face manipulations. On the other hand, many successful real-life machine learning solutions are based on ensemble models that fuse results from individual types of features or detectors* [...]."

In the same book, the detection problem and the need for a fusion-based solution are emphasised even further by including the issue of counter-forensic measures, intended to interfere with forensic analyses [Rathgeb22]:

> "[A] *skilled attacker, aware of the principles on which forensic tools work, may enact some counter-forensic measure on purpose* [...]. *Therefore, the integration of multiple tools, all designed to detect the same type of attack but under different approaches, may be expected*

> *to improve performance, and especially robustness with respect to both casual and malicious disturbances.*"

In [Kuncheva04], the following three types of reasons why a classifier ensemble might be better than a single classifier are identified:

- Statistical: Instead of picking a potentially inadequate single classifier, it would be a safer option to use a set of unrelated ones and consider all their outputs.
- Computational: Some training algorithms use hill-climbing or random methods, which might lead to different local optima when initialised differently.
- Representational: It is possible that the classifier space considered for a problem does not contain an optimal classifier.

Independently of the exact reason for choosing a fusion approach instead of a single classifier, the textbook [Kuncheva04] explicitly warns that "*an improvement on the single best classifier or on the group's average performance, for the general case, is not guaranteed*." By combining classifiers (or other expert systems), the users hope for a more accurate decision at the expense of increased complexity, but this cannot be guaranteed. This explains why forensic practitioners hesitate to rely on fusion. Here, the fact that a potentially negative impact on classification accuracy might occur due to incorrect use or parametrisation, and the increased complexity of fusion (including the inherently higher costs for plausibility validation[31]) are considered to be in conflict with the fundamental requirements for forensics, even though the potential benefits of fusion are acknowledged. In the feeder paper **[Kraetzer21]** (itself loosely based on a conference paper originally published as [Makrushin19]), included as Chapter 9 of this cumulative habilitation treatise, an in-depth discussion of the pros and cons of information fusion approaches in the context of media forensics is provided for the application scenario of digital image authenticity and integrity analysis for face morphing attack detection.

The findings of the feeder paper **[Kraetzer21]** are summarised in the following. The empirical evaluations performed for the application example of FMA detection for image authenticity and integrity verification compare the detection accuracy of five single face morphing attack detectors selected as examples and four fusion approaches (one at decision level (majority voting), and three at matching score level (weighted linear combination, Dempster-Shafer Theory of Evidence, and forensic likelihood ratio)). In the summary of the fusion experiments' results presented in **[Kraetzer21]** (see Chapter 9, pp. 133 ff. of this document), three main reasons are given why the fusion experiments fail to outperform the best individual classifier in the presented results:

1. "*Lack of diversity of the individual detectors*": Four of the five individual detectors were developed by the same research group and rely on training of Deep Convolutional Neural Networks (DCNN) with similar data sets but strong variances in data augmentation. Hence, it is very likely that these detectors make mistakes on the same samples in field application. Only the fifth detector relies on entirely different morphing detection clues and is developed by a different research group using a different data set for training. In theory, an assumed clustering of four apparently very similar detectors might prove a strong bias in fusion that should be avoided at any cost. In practice, the experiments on different ensembles of classifiers in **[Kraetzer21]** showed a better performance if only those four detectors were used instead of all five.
2. "*Lack of performance in individual detectors*": The results presented show that for one of the five individual detectors, significant generalisation problems occur for the estimated default decision threshold. Better calibration of the method (here, using more diverse datasets for the estimation of a suitable decision threshold) resulted in significantly lower error rates.
3. "*Lack of similarity between the training and test data*": Different proprietary data sets were used for training the individual classifiers, which is a very common method, but the datasets for

[31] As pointed out in the ENFSI documents analysed in Section 2.3.4, technical capabilities (such as accuracy or throughput) are by far not the most significant characteristics of forensic methods. In general, these are usually rated by forensic practitioners by their maturity, i.e., their scientific admissibility.

> adjusting fusion parameters (evaluation data set) and for actual testing are also very different from each other and from the training data set. It is doubtful whether it makes sense to use different data sources for adjusting fusion parameters and for testing, but this is the real-life situation: In practice, it is very difficult to precisely foresee and provide significant in-field data at the stage of system development or parameter adjustment. Moreover, there is no guarantee that the in-field data that will be obtained in the future is at all similar to the presented training data.

The main findings of the feeder paper **[Kraetzer21]** for the chosen application domain of face morphing attack detection for image authenticity and integrity verification are generalised in the paper itself as follows:

> "*The results presented in the empirical evaluations in this paper demonstrate that fusion can fail even with a set of relevant individual classifiers.* [...] *Summarizing the lessons learned from the approach of using fusion for* [face morphing attack detection] *detection as done in this paper and drawing some generalization toward other media forensics classification or decision problems, the following has to be said: The requirements for (media) forensic methods in terms of scientific admissibility (or Daubert compliance) are obviously important! Methods should indeed be published upon and peer reviewed, their error rates should be precisely known and standards for the application of methods should be known. But the threat that Champod and Vuille identify as a problem of ascertaining the error rates of a test 'can prove misleading if not all its complexities are understood'* [Champod11] *plays a very significant role as demonstrated in the evaluations performed here. Besides the requirements for individual expert systems to be used in forensic investigations (including its accurateness), if it comes to information fusion, additional constraints have to be observed. These are, at least:*
>
> - *The diversity of the detectors, which has to be ascertained either by knowledge about the precise means of decision generation and the diversity of those means or empirically.*
> - *An independent and thorough benchmarking of detectors to establish also an idea on the generalization power of performance claims made by their creators.*
> - *Considerations on the similarity/correlation between training data available (during training of the individual classifiers and the training of the fusion methods) and the data to be expected in field application are very important. If very precise assumptions are possible on the application data, weighting might be applicable in fusion. Elsewise, only unweighted fusion strategies like majority voting or the sum-rule should be employed, if any fusion is used in those cases at all.*
>
> *The diversity issue becomes very problematic if features (as the means to represent a decision problem in a feature space) are not hand crafted by experts but learned, e.g., by DCNN. In this paper, the diversity problem of the detectors used here as 'black boxes' has been established in direct contact with the developers of those methods, which is hardly an option in most field applications.*
> *Also, the recent trend to generate synthetic data sets for the training of pattern recognition methods (either traditional or neural network based) introduces another degree of freedom into the characteristics of datasets.* [T]*his approach is used to avoid tedious data collection tasks while creating sufficiently sized data sets for modern day data-greedy classifiers. The problem here is the influence of the synthesis process on its output (i.e., the synthesis-specific artifacts) that will become part of the model trained by each classifier. It is related to the questions of source characteristics imposing themselves into trained models but carries a different degree of relevance for forensic application scenarios.*
> *The general problem with training- and test data being mismatched in practice is hardly new. It hardly ever occurs in scientific papers on applied pattern recognition, because it can easily be prevented in lab tests. Nevertheless, it is a very good argument why media forensics methods should undergo rigorous testing and benchmarking by third parties, like*

> *it is done in the field of* [face morphing attack detection] *in the National Institute of Standards and Technology (NIST) FRVT MORPH challenge. Only such joint efforts can lead to methods that might become mature enough to aim at court admissibility.*"

Synopsis: In academic research on media forensics, (automated) information fusion is often seen as a valuable method to improve detection performances, bought at the cost of increased run-time complexity. Forensic practitioners on the other hand are reluctant to rely on fusion, due to the associated risk that the overall performance might be reduced (which can happen in practice, as shown with the results in **[Kraetzer21]**), and the more severe problem of explainability of the results.
The results discussed for attack detection modelling for detector fusion relate to REQ1 (in-depth discussion of the necessary conditions as well as potential pros and cons of using information fusion approaches in the context of media forensics), REQ2 (empirical evaluations using a set of FMA detectors with different fusion methods and fusion ensemble composition strategies), REQ3 (discussion on the need of benchmarking and proficiency testing for media forensics methods, especially in fusion setups), REQ4 (identification of diversity criteria for methods used in fusion), and REQ5 (systematic description of the used fusion methods and fusion ensemble composition strategies), as specified in Section 1.3.

3.1.4 Result Overview for the Application Domain of Face Morph Attack Detection

The work performed in this context focuses on finding means for specific application scenarios to combine image editing history models, attacker models and attack detection models into a unified picture. This work focuses primarily on research on the modelling of media generation processes and source models (here, the introduced descriptive image editing history model and life-cycle model for photo-identity (ID) documents). This pays respect to the significance placed on such source models, e.g., in [ENFSI22a]:

> "[T]*he properties of the retrievable traces are analysed, to determine if they support or oppose the hypothesis that the recording has been modified.* [...] *A key objective of any authenticity analysis is therefore to determine whether observed features of a piece of* [...] *evidence were introduced by the original recording process or by subsequent actions.*"

A second focus of the work in this application domain is placed on different attack detection models, including single detector approaches, sequences of detectors, and fusion approaches.
The results presented in the corresponding feeder papers contribute to all requirements REQ1-REQ5 derived in the problem outline of this treatise in Section 1.3.

3.2 Work Published in the Application Domain of DeepFake Detection

The operational modelling work done by the author in the context of the FAKE-ID project (2020-2024, see Section 1.2) supports the empirical work of the research group leader and of colleagues on the topic of DeepFake detection for digital videos. In the context of this ongoing research effort, various abstractions of forensic processes were presented in published work, refining an approach for operational modelling based on the German BSI 'Leitfaden IT-Forensik' [BSI11] and on publications extending these German national guidelines for IT forensic investigations, such as the Data Centric Examination Approach (DCEA) presented in [Kiltz20]. An overview of selected work on this topic is provided in this section, excerpting the work from the corresponding feeder papers and discussing them in the wider context of this cumulative habilitation treatise.

3.2.1 Initial Steps in Operational Modelling for DeepFake Detection

In the feeder paper **[Siegel21]** (included as Chapter 10 of this cumulative habilitation treatise), the first research paper co-authored by the author in the scope of the FAKE-ID research project, a first step

towards an operational model for DeepFake detection is made. The paper projects the needs identified for such an operational model onto the German BSI 'Leitfaden IT-Forensik' [BSI11] and the DCEA. The paper concludes with the synopsis that the DCEA is "*not yet perfectly capable to fit the needs of media forensics analyses*" and provides a justification of this claim as well as initial ideas for necessary process modelling expansions based on this established best practice.
The criticism (or rather identification of needs for expansion of this best practice model) in **[Siegel21]** focuses on two topics: on the one hand, the forensic data models available in the BSI 'Leitfaden IT-Forensik' and the DCEA, and on the other hand, the projection onto phases and forensic method classes.

Regarding the data models, **[Siegel21]** discusses two existing models from [Kiltz15] and [Kiltz20]: the original model for digital forensics and the adaptation thereof presented in [Kiltz15] for the field of digitized forensics (illustrated for the field of dactyloscopy[32]). The synopsis of **[Siegel21]** (see also pp. 152 and 153 in Chapter 10 of this cumulative habilitation treatise) on this discussion is:

> "*One important realization when trying to apply the DCEA data types for digital or digitized forensics* [...] *is that they do not perfectly match the media forensics task at hand. Using the original model for digital forensics, only four of the data types would be covered (raw data differentiated into different user data media streams (video, audio, network stream) and possibly hardware data (derived from the camera / microphone used) as well as details about data). If the model for digitized dactyloscopy is used, which is slightly better matching the characteristics of our application scenario, then eight of the ten data types would be directly relevant* [...] *while one other would very likely also to be of significance* [...]."

The authors reason that "*an adapted data type model for media forensics would be required to be able to make use of the full potential of the DCEA in this context*". Such a domain-adapted data model is presented in the follow-up publication **[Siegel22]**.

Regarding the phases and forensic method classes, a first projection of the different operational aspects of training, validating and applying the DeepFake detectors in the established process model DCEA is performed in **[Siegel21]** to show how such media forensics methods could be integrated into forensic procedures. The first of two items discussed in this context is the question of where a DeepFake detector is supposed to be placed in an operational model. The following answer is given in **[Siegel21]**:

> "*There exist two potential operation points in the phases described by DCEA: Either as a method of Explicit means of intrusion detection (EMID) as part of incident detection mechanisms, which would place the whole DeepFake detection with the training of the method and its application into the phase of Strategic Preparation (SP), or in Scaling of methods for evidence gathering (SMG) which would place DeepFake detection after an incident is detected or suspected and place corresponding components in the phases Operational Preparation (OP), Data Gathering (DG), Data Investigation (DI) and Data Analysis (DA). These two distinct operation points as a live detector or as means of post-mortem (or a posterior) analysis in data investigation have, amongst other effects, significant impact on the training scenario that can be assumed: In case of application as an live detector EMID in SP only pre-trained models can be applied. In case of a post-mortem SMG detector, in OP the material to be investigated can be analysed to design targeted training datasets perfectly matching the characteristics encountered. Using those sets (and own DeepFake algorithms to generate also specimen for this class) optimal models could be trained for each case.*"

The second item concerns questions of reproducibility, explainability and interpretability for media forensics methods based on machine learning. Here, the following summary is presented in **[Siegel21]**:

> "*The accompanying documentation in DCEA is meant to allow for interpretability and plausibility validation steps while compiling the case documentation in DO. For our work*

[32] Forensic fingerprint analysis and comparison

> *this implies not only documenting all details of the pattern recognition process at hand but also using this data to reason about the plausibility of decisions (e.g. by comparing the characteristics of training- and test sets to determine questions of generalisation power).*"

In addition to these initial steps on the path towards operational modelling, **[Siegel21]** also provides the conceptual model for a fusion-based DeepFake detection pipeline. This conceptual model is used in later publications, such as **[Siegel22]**, as the basis for the operational modelling of DeepFake detection processes.

Synopsis: All components of operational models for forensic processes require domain adaptation for specific application domains. In **[Siegel21]**, the needs for such a domain adaptation are discussed using the example of the forensic data model component of the DCEA as foundation of the corresponding research work.
The results presented on conceptual and operational modelling relate to REQ1 (modelling the investigation contexts for DeepFake detection), REQ2 (empirical evaluation with three sets of hand-crafted features and three different fusion strategies), REQ3 (first steps of a projection onto a pre-existing, data-centric examination approach for conceptual and operational forensics process modelling), and REQ5 (detailed descriptions of the modelling background and implementation of the individual detectors and used fusion operators), as specified in Section 1.3.

3.2.2 A Domain-Adapted Data Model for Media Forensics

In the feeder paper **[Siegel22]** (included as Chapter 11 of this cumulative habilitation treatise), the work of **[Siegel21]** is expanded by proposing a domain-adapted data model for media forensics and illustrating its applicability for the application scenario of DeepFake detection. The paper points out that "[p]*erforming abstract data modeling without precise knowledge about the context, in which the data type is supposed to be used, is a futile task*". As a consequence, a conceptual model for a generalised media forensics analysis process is first briefly discussed in the paper. Typical data streams within such a process are then identified, followed by a differentiation of the data streams into data types.

In summary, the following five data streams are identified in **[Siegel22]** (for details, see pp. 176 ff.):

- The *process description* is proposed as a sourceable or instantiable template, which is generated before starting the investigation.
- The *media data* contains all forms of media, such as images, videos, audio and/or network streams used and created in the course of the investigation process.
- The non-media output of the individual examination steps is gathered into the data stream *forensic process/pipeline internal data and reporting*.
- The *process control data* is the combination of all settings used in the investigation, including all parameters and models used.
- The *contextual data* contains all information regarding the context of a specific investigation.

The following summary is presented in **[Siegel22]** concerning these data streams:

> "*This subdivision of the data associated with an investigation is a functional classification paying respect on one hand to the characteristics of data objects involved and on the other hand to operational and security requirements. The media data stream of an investigation might easily contain terabytes of video data which would require a access to a private cloud for efficient handling, while the reporting data would assumed be much smaller in data size but be more frequent and have other constraints like reliable time-stamping. From the operational and security perspective also different protection levels (and as a consequence security mechanisms) would be required depending on the nature of the objects in a stream and the risks associated.*"

The Media Forensic Data Types (MFDTs) summarised in Table 3.1 are derived based on the data streams. For details on these MFDTs, the reader is referred to page 177 of this cumulative habilitation treatise. A projection between the five defined data streams, the data types for digitized forensics (DD) from [Kiltz20], and the domain-adapted data model is shown in Figure 3.1.

These MFDTs are then used in **[Siegel22]** for modelling forensic functions or procedural elements (operators) as well as forensic processes (see Sections 3.2.3 and 3.2.4).

Data type	Derived from DD	Description
MFDT1 - Digital input data	DD1	The initial media data considered for the investigation.
MFDT2 - Processed media data	DD2	Results of transformations to media data (e.g., grayscale conversion, cropping)
MFDT3 - Contextual data	DD3	Case-specific information (e.g., for fairness evaluation)
MFDT4 - Parameter data	DD4	Contains settings and other parameters used for acquisition, investigation and analysis
MFDT5 - Examination data	DD5, DD6, DD8	Includes the traces, patterns, anomalies, etc. that lead to an examination result
MFDT6 - Model data	DD7	Trained model data (e.g., face detection and model classification data)
MFDT7 - Log data	newly defined	Data relevant for the administration of the system (e.g., system logs)
MFDT8 - Chain of custody & report data	DD9, DD10	Data used to ensure integrity and authenticity (e.g., hashes and time stamps) as well as the accompanying documentation for the final report

Table 3.1: Media Forensic Data Types (MFDTs) proposed in **[Siegel22]**, adapted from **[Siegel22]**

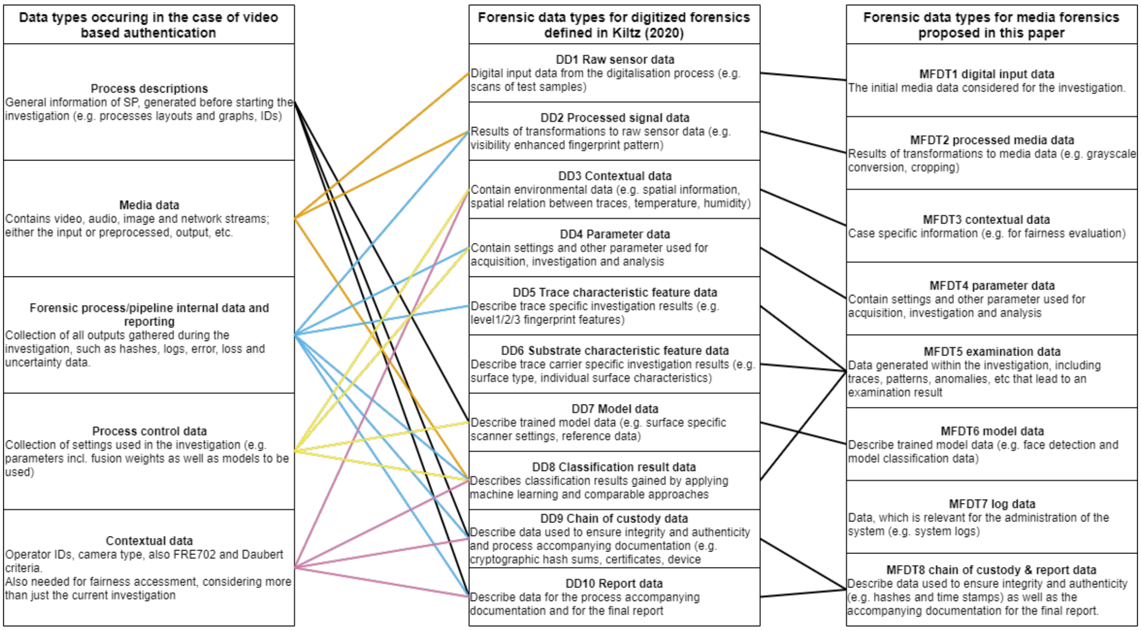

Data types occuring in the case of video based authentication
Process descriptions
General information of SP, generated before starting the investigation (e.g. processes layouts and graphs, IDs)
Media data
Contains video, audio, image and network streams; either the input or preprocessed, output, etc.
Forensic process/pipeline internal data and reporting
Collection of all outputs gathered during the investigation, such as hashes, logs, error, loss and uncertainty data.
Process control data
Collection of settings used in the investigation (e.g. parameters incl. fusion weights as well as models to be used)
Contextual data
Operator IDs, camera type, also FRE702 and Daubert criteria.
Also needed for fairness accessment, considering more than just the current investigation
Forensic data types for digitized forensics defined in Kiltz (2020)
DD1 Raw sensor data
Digital input data from the digitalisation process (e.g. scans of test samples)
DD2 Processed signal data
Results of transformations to raw sensor data (e.g. visibility enhanced fingerprint pattern)
DD3 Contextual data
Contain environmental data (e.g. spatial information, spatial relation between traces, temperature, humidity)
DD4 Parameter data
Contain settings and other parameter used for acquisition, investigation and analysis
DD5 Trace characteristic feature data
Describe trace specific investigation results (e.g. level1/2/3 fingerprint features)
DD6 Substrate characteristic feature data
Describe trace carrier specific investigation results (e.g. surface type, individual surface characteristics)
DD7 Model data
Describe trained model data (e.g. surface specific scanner settings, reference data)
DD8 Classification result data
Describes classification results gained by applying machine learning and comparable approaches
DD9 Chain of custody data
Describe data used to ensure integrity and authenticity and process accompanying documentation (e.g. cryptographic hash sums, certificates, device
DD10 Report data
Describe data for the process accompanying documentation and for the final report
Forensic data types for media forensics proposed in this paper
MFDT1 digital input data
The initial media data considered for the investigation.
MFDT2 processed media data
Results of transformations to media data (e.g. grayscale conversion, cropping)
MFDT3 contextual data
Case specific information (e.g. for fairness evaluation)
MFDT4 parameter data
Contain settings and other parameter used for acquisition, investigation and analysis
MFDT5 examination data
Data generated within the investigation, including traces, patterns, anomalies, etc that lead to an examination result
MFDT6 model data
Describe trained model data (e.g. face detection and model classification data)
MFDT7 log data
Data, which is relevant for the administration of the system (e.g. system logs)
MFDT8 chain of custody & report data
Describe data used to ensure integrity and authenticity (e.g. hashes and time stamps) as well as the accompanying documentation for the final report.

Figure 3.1: Mapping between the five data streams identified in **[Siegel22]**, the forensic data types (DD) for digitized forensics as presented in [Kiltz20] and [Kiltz15], and the domain-adapted data model for the media forensics application scenario of DeepFake detection used in this treatise (image based on **[Siegel22]**).

Synopsis: One of the unsolved issues identified in **[Siegel21]**, the lack of a suitable forensic data model, is addressed here by proposing MFDTs. This design of a modelling component relates to REQ3, as specified in Section 1.3.

Potential expansions of this modelling work would have to consider additional aspects, including privacy concerns[33] as well as material review data (see e.g., `https://forensicworkinggroup.com/MAT.pdf`) to provide even more structured (i.e., closer to the actual data) representations.

[33] [Vaughan20] states on 'Ethical governance and oversight':

> *"The increasing scope of digital forensic capabilities raises new ethical dilemmas for policing to consider. Some of these overlap with other digital investigation issues, such as using AI to analyse datasets and concerns about aggregating investigative data. Other concerns are particular to digital forensics (DF)*

3.2.3 Modelling of Procedural Elements

In the feeder paper **[Siegel22]** (included as Chapter 11 of this cumulative habilitation treatise), the core component of all operational models is defined as an 'operator' (or 'processing operation'). It is considered to be an atomar processing black box component with an identifier and (usually) a description of the processing performed in this operation. The following connectors are defined for an operator:

- *input*
- *output*
- *parameters*
- *log data*
- (*model data*)

The fifth connector is conditional. For this part, **[Siegel22]** states:

> "*To pay respects to the particularities of this field and make the following modeling task easier, a fifth connector is defined within this paper for a specific type of operator which requires a knowledge representation or a model for its processing operation. In that case, this fifth connector is labeled* model. *Depending on the nature of the operator this could be a rule set, signature set, statistical model, neural model, or any other form of knowledge representation.*"

In **[Siegel22]**, an example of a processing pipeline for face detection is described as a sequential combination of three atomar operators. This example is taken from the work presented on DeepFake detection and represents the sub-routine of face segmentation as a necessary step in DeepFake detection for videos. The first operator in this three-step processing sub-routine is loading the video from its *input*. The *parameters* need to be chosen based on the video format and the *output* is stored as a video stream. In the next operator, this video stream is then split into single frames as necessary pre-processing for an image-based face detection and segmentation algorithm. For the face detection and segmentation, a pre-trained model with 68 landmarks (here from [King09]) is loaded at the third operator's *model* connector. This is the only step in this example where *model* data is used.
Each step provides corresponding process documentation in the form of logs and chain of custody (CoC) data at its *log data* connector.

Synopsis: Atomar 'operators' are considered here as the core component of all operational modelling work. Their combination enables the creation of more complex processes. This process modelling issue relates to REQ3, as specified in Section 1.3.

3.2.4 Modelling of Forensic Processes

Combining the work on MFDTs and the 'operators' discussed in Sections 3.2.2 and 3.2.3 as well as the phase-driven modelling of forensic processes as discussed in the DCEA [Kiltz20], the feeder paper **[Kraetzer22]** updates the operator description, adding the 'model' connector as a conditional component (i.e., only available when a model-driven operation is considered).
The next step, taken in feeder paper **[Siegel22]** is to place the aforementioned components into an initial operational model for media forensics investigations for the application example of DeepFake detection. This operational model takes into account the role assigned in the BSI 'Leitfaden IT-Forensik' [BSI11] and the DCEA [Kiltz20] to the phase of SP. In this phase, all forensic methods need to be prepared (and evaluated, including proficiency testing). This also includes the training of models required for model-driven forensic methods. Later, as part of the OP or DI and DA phases, these models are then loaded and used. This design and preparation of a forensic investigation pipeline is called the **templating of a forensic process** in **[Siegel22]**. Its use is referred to as the **instantiation** of the corresponding forensic

science. For example, managing sensitive personal data coming from mobile phone analysis, which involves 'collateral intrusion' into others' privacy when reviewing messages, or friends and family photographs."

process. For illustrations of the templating and instantiation, the reader is referred to pages 177 to 179 of this cumulative habilitation treatise.

Based on these initial examples, the feeder paper **[Kraetzer22]** (included as Chapter 12 of this cumulative habilitation treatise) presents more detailed operational modelling considerations on the orchestration of operators (and sub-processes) in forensic investigation contexts (see pages 196 and 197). The synopsis is presented in Chapter 12 as follows:

> "*At the end of the process in SP, well specified templates exist that can easily be instantiated into practical investigations as soon as an event/incident triggers an investigation request.*"

Efforts put into the SP of a forensic process are assumed to prepare for an effective response in case of an incident or post-mortem investigation. They are intended to increase forensic readiness of response and investigation units as well as strengthen the whole field by providing standardised (and certified) methods and procedures, This is addressed in **[Kraetzer22]** as follows:

> "[In OP,] *a prepared (as well as benchmarked and potentially certified) template from SP is filled with life by invoking the corresponding orchestration of operators on the assigned processing nodes. Decision models pre-trained in SP are loaded* [...] *together with the used pre-processor and classifier parameters. Thus initialised, the operators are then applied to the input data to the process (MFDT1) to determine traces or information relevant for the investigation at hand.* [...] *When a template is then instantiated for a case in OP, the required documentation packages are marshalled together into the investigation accompanying documentation of the case.*"

What this hard split into SP and operations (OP, DG, DI, DA and Documentation phase (DO)) is supposed to provide are more precise process descriptions, which can more easily be verified by third parties. Furthermore, they make training, benchmarking and testing procedures more transparent and are thus intended to improve the identification of influence factors, training bias and potential error sources. To make this split less absolute and to facilitate learning from issues encountered in instantiation, the natural interaction between the SP and operations of forensic processes is reflected upon in an update of the phase model by including an explicit feedback loop (from OP) to the SP (see page 190 in Chapter 12).

Synopsis: Operational models for model-driven methods need to take into account the fact that these models have to be trained and evaluated prior to application. This is acknowledged here by splitting the proposed operational model into the templating in SP and the instantiation in a specific investigation in the OP phase. In a well-prepared forensic process, the instantiation of an investigation template in OP would (besides other things) trigger the initialisation of the accompanying documentation of the investigation by marshalling the corresponding documentation packages.
This process modelling issue relates to REQ3, as specified in Section 1.3.

3.2.5 Attack Detection Modelling for the Application Domain of DeepFake Detection

Even though the focus of the work in this application domain lies on operational modelling, all feeder papers for this application scenario also contain empirical studies on DeepFake detection, based on the corresponding attack detection modelling. The work in **[Siegel21]**, for example, focuses on three sets of hand-crafted features and three different fusion strategies to implement DeepFake detection. The results obtained with third-party reference databases show performances similar (peak AUC > 0.95) to those of methods using features learned by neural networks. In **[Siegel22]**, the work presented in **[Siegel21]** for a fusion-based DeepFake detection approach is re-structured and expanded using the newly introduced operational modelling components (see Section 3.2.3 above). In **[Kraetzer22]** (included as Chapter 12 in this cumulative habilitation treatise), the set of detectors and the operational modelling components from the previous papers are revisited, expanded and re-evaluated in a

benchmarking-driven approach for fusion weight determination for the DeepFake detection framework. In this context, two new feature spaces (both semantically analysing the blinking behaviour in a video) are integrated into the framework and then evaluated. For this evaluation, a discussion on benchmarking metrics (with a focus on Cohens Kappa) is presented.
In **[Siegel23b]**, a detector from **[Siegel21]** is re-evaluated using data minimisation. It is shown that the achieved accuracy is not significantly impaired. These empirical evaluations performed to establish the trade-off between detection performance and data minimisation for DeepFake detection are motivated by the European Union (EU) General Data Protection Regulation (GDPR).
[Kraetzer23] (included as Chapter 14 in this cumulative habilitation treatise) expands the DeepFake detection from a 2-class problem to an n-class decision problem, presenting results for the potential attribution/identification of the DeepFake generation method used. In the accompanying empirical evaluations, an estimation of the generalisation power (or lack thereof) of pre-existing DeepFake detectors in intra- and inter-data set benchmarking using different data selection strategies and classifiers is presented. The results presented in the context of the n-class DeepFake classification experiments imply significant problems with overfitting DeepFake detection models to specific DeepFake generation methods.

Despite the criticism presented in Section 3.1.3.3 regarding forensic solutions based on information fusion, they are considered to be one of the few existing options to overcome this problem with the generalisation power of DeepFake detection models.
While fusion is widely believed to be strongly beneficial to decision problem solution approaches like pattern recognition or anomaly detection, publications such as **[Kraetzer21]** and **[Kraetzer22]** point out that information fusion, which indeed has a huge potential to improve the accuracy of pattern recognition systems, is still applied with great hesitation in the forensic sciences. The reason given is that a potentially negative impact on the classification accuracy if wrongly used or parameterised as well as the increased complexity (and the inherently higher costs for plausibility validation) of fusion are in conflict with the fundamental requirements for forensics. To overcome this hesitation, the typical solution is the following (discussion expanded from **[Kraetzer22]**, page 197 of this cumulative habilitation treatise):

- Very thoroughly benchmark under different training and evaluation scenarios (see [Neubert18b] and (for DeepFake detection) **[Siegel21]** as well as **[Kraetzer22]**) the individual expert systems (here detectors) to be used in the fusion to precisely establish their requirements and capabilities as well as the error rates attached
- Benchmark different fusion schemes under different training and evaluation scenarios (see **[Kraetzer21]**), and establish the impact of different weighting strategies on the (detection) performance and error patterns
- Design systems as decision support systems instead of automated solutions to enable human-in-control principles
- Consider decision confidences (where available) for opinion forming
- Allow for auditability as well as human oversight for the entire process

Especially the third and the last item, the aspects of human-in-control and the required human oversight, are recent trends for critical AI applications (incl. forensics) which are, among other regulations, manifested in the current initiative towards an EU Artificial Intelligence Act (AIA, see Section 3.2.6).

Synopsis: Despite the fact that the focus of the work in this application domain lies on operational modelling, all papers published also contain conceptual modelling results, usually in the form of empirical experiments for DeepFake detection. The results of these, usually evaluations based on detector fusion, include, among other things, descriptions of feature sets suitable for detection (e.g., the eye-blinking semantics features from **[Kraetzer22]**). These DeepFake detection experiments are accompanied in the later papers by additional investigation goals including recent perspectives such as data minimisation (motivated by GDPR considerations) or the attribution or identification of the DeepFake generation method used. Especially the latter is of significance for the entire application domain, since the results obtained show a significant overfitting of the evaluated DeepFake detection models to specific DeepFake generation methods. This raises the question of which generalisation power DeepFake detection methods can actually achieve in unconstrained detection (i.e., without an applicable source model or an indication which DeepFake generation method might have been used).

The results presented for all of these empirical research efforts relate to REQ2 and REQ4 respectively, iteratively challenging previously achieved results for the evaluation of methods with new investigation procedures (e.g., the data minimisation in **[Siegel23b]**).

3.2.6 Selected GDPR and AIA Considerations for DeepFake Detection

Later feeder papers for this application scenario, like **[Siegel23b]** and **[Kraetzer23]**, also consider shifts in this research domain that are motivated not by technical innovations but rather by organisational and regulatory aspects.

In **[Siegel23b]**, the potential impact of data minimisation in DeepFake detection model generation is discussed with a series of experiments. One DeepFake detection approach is used as an example to show that data minimisation can be successfully applied in this context, without significant loss of detection accuracy.

For **[Kraetzer23]**, the operational modelling work discussed in Section 3.2.3 is expanded to include human-in-the-loop and human-in-control aspects as made necessary by changing requirements/legislation world-wide, esp. the (upcoming) EU Artificial Intelligence Act (AIA). This is consistent with the emphasis put on 'human-based functions' and the required qualifications and proficiency testing for the involved personnel in ENFSI BPM FIT [ENFSI15].

Additionally, **[Kraetzer23]** focuses on the need to separate duties in the evaluation, benchmarking and certification of model-driven forensic methods (see Chapter 14, page 227).

Synopsis: Even though the research community in media forensics is strongly focused on technical developments (usually more reliable detectors), changing environmental conditions, including changing regulatory requirements, also have to be considered by the researchers. One relevant act of legislation that will influence the development and use of methods based on machine learning (at least in Europe) is the (upcoming) EU AIA.

The results presented for these research efforts relate to REQ3 (in terms of updates of standardised procedures that might become necessary) and (to some extend) to REQ4 (existing methods might have to undergo re-evaluation under changed evaluation criteria), as specified in Section 1.3.

3.2.7 Reflection on the Conceptual and Operational Modelling in Relation to ENFSI BPM for Digital Image Authentication

In **[Siegel23a]**, conceptual modelling aspects from **[Kraetzer22]** and **[Kraetzer23]** are projected onto the conceptual model of the ENFSI Best Practice Manual (BPM) for Digital Image Authentication (DIA) [ENFSI21] to discuss the similarities and differences between both models, showing how the research work done in academia might expand the conceptual and operational models discussed in expert groups composed of forensic practitioners (such as the ENFSI Expert Working Group (EWG)).

The most important outcome of **[Siegel23a]** for this habilitation treatise is that the categorisation of forensic methods proposed in [ENFSI21] provides a starting point for modelling the classes of forensic methods in media forensics that seems to be much better suited for this domain than the original third pillar of the '*Leitfaden IT-Forensik*' model (the classes of forensic methods). The results of this modelling is shown on page 236 in Chapter 15 of this cumulative habilitation treatise.

Synopsis: Selected ENFSI BPMs (esp. [ENFSI15]) call for a closer cooperation of forensic practitioners and research communities in the corresponding fields. While this might at first glance seem unattractive to researchers, who are caught in their own mechanisms, including the race for funding summarised in Section 2.2.2, it would be hugely beneficial to the overall field. As a consequence, cutting-edge research would also have to take into account existing standardisation documents (like the ENFSI BPMs or other documents of similar nature like the BSI 'Leitfaden IT-Forensik') and indicate how to update such documents accordingly to reduce the gap between research and field application in this domain. The results presented for these research efforts relate to REQ3 (identify possibilities for updating standards), as specified in Section 1.3.

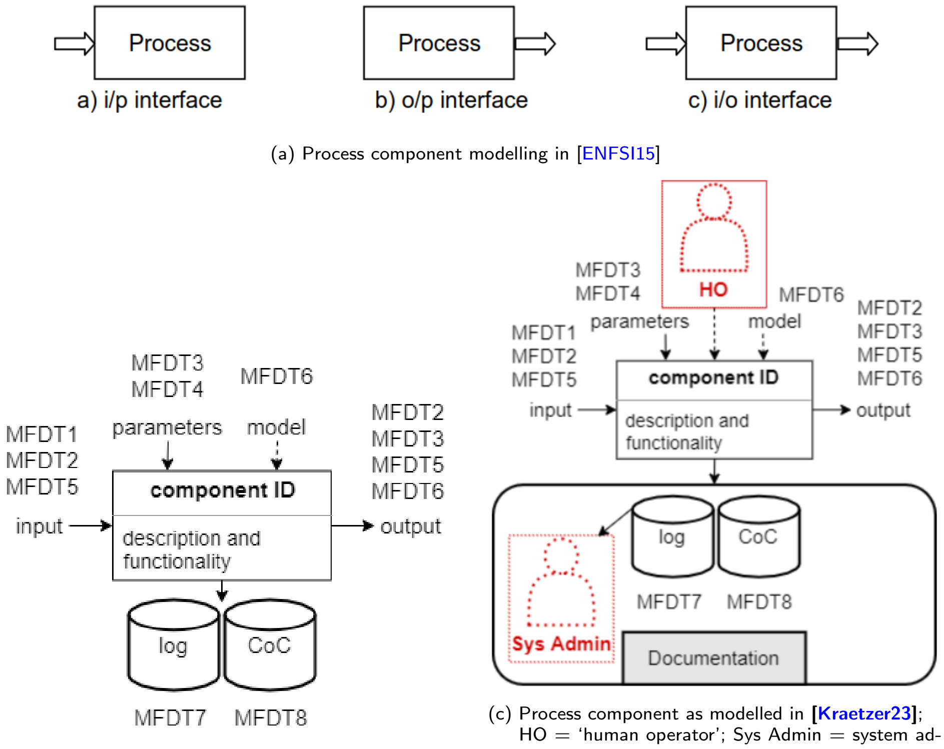
3.2.8 Overview of Results for the Application Domain of DeepFake Detection

The work performed in this context focuses on operational modelling aspects. On the basis of the original code of practice for IT forensics published by the German Federal Office for Information Security (BSI) in its '*Leitfaden IT-Forensik*' [BSI11] (see Section 2.3.3.1) and the additions to its process model published by others (see Section 2.3.3.2), corresponding expansions, mostly to the forensic data model and the phase-driven investigation model, are presented. In place of the third pillar of the '*Leitfaden IT-Forensik*' model (the classes of forensic methods), a different starting point for suitable modelling is identified for media forensics with the categorisation of forensic methods proposed in [ENFSI21].

Regarding the **forensic data model**, the feeder papers provide a domain-adapted version that is derived in **[Siegel22]** from the data model for digitized forensics as presented in [Kiltz20] and [Kiltz15] (see Section 3.2.2). In the papers following **[Siegel22]**, the MFDTs are used to describe the various inputs and outputs of operational blocks in forensic (sub-)processes (see below).

Regarding the work on the **phase-driven investigation model**, the following three aspects form the main items of the operational modelling work performed in the context of this habilitation project: the modelling of operational elements ('operators'), their orchestration into (sub-)processes, and the considerations on templating versus instantiation of forensic processes.
Figure 3.2 compares three different models of operational elements (called 'operators' in Section 3.2.3). The first sub-figure is taken from [ENFSI15] and shows the very rudimentary degree of granularity found there. The two sub-figures (b) and (c) show two different iterations of the author's own work (from **[Kraetzer22]** and **[Kraetzer23]**). The expansions illustrate first the combination of process and data modelling performed in **[Kraetzer22]**, and then (marked in red in the sub-figure (c)) the 'human-in-the-loop' and 'human-in-control' aspects discussed in **[Kraetzer23]**.

Regarding the orchestration of forensic (sub-)processes, Figure 3.3 compares two different examples, one from the ENFSI BPM FIT [ENFSI15] and one from **[Siegel22]**. The main difference is the level of detail used in the modelling (here the data streams and data flows), but the intention behind the split into sub-processes and functions is the same in both publications: to allow for a more precise description of process components and to make benchmarking, proficiency testing and (potential) certification of functions, sub-processes and processes easier.

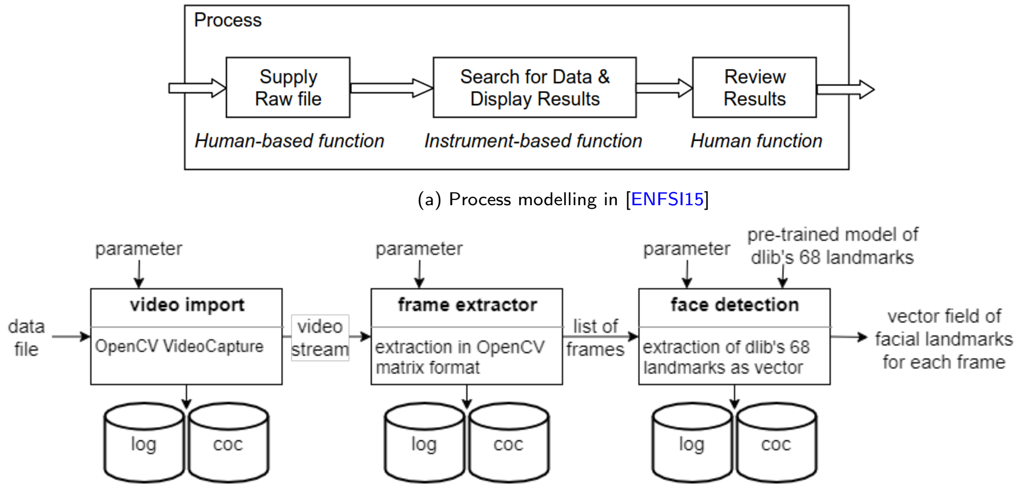


(a) Process component modelling in [ENFSI15]

(b) Process component (*operator*) in [Kraetzer22]

(c) Process component as modelled in [Kraetzer23]; HO = 'human operator'; Sys Admin = system administrator

Figure 3.2: Comparison of the process component modelling from ENFSI BPM FIT (sub-figure (a)), and two different iterations of the author's own work (sub-figures (b) and (c)).


Process
Supply Raw file
Search for Data & Display Results
Review Results
Human-based function
Instrument-based function
Human function
parameter
parameter
parameter
pre-trained model of dlib's 68 landmarks
data file
video import
OpenCV VideoCapture
video stream
frame extractor
extraction in OpenCV matrix format
list of frames
face detection
extraction of dlib's 68 landmarks as vector
vector field of facial landmarks for each frame
log
coc
log
coc
log
coc

(a) Process modelling in [ENFSI15]

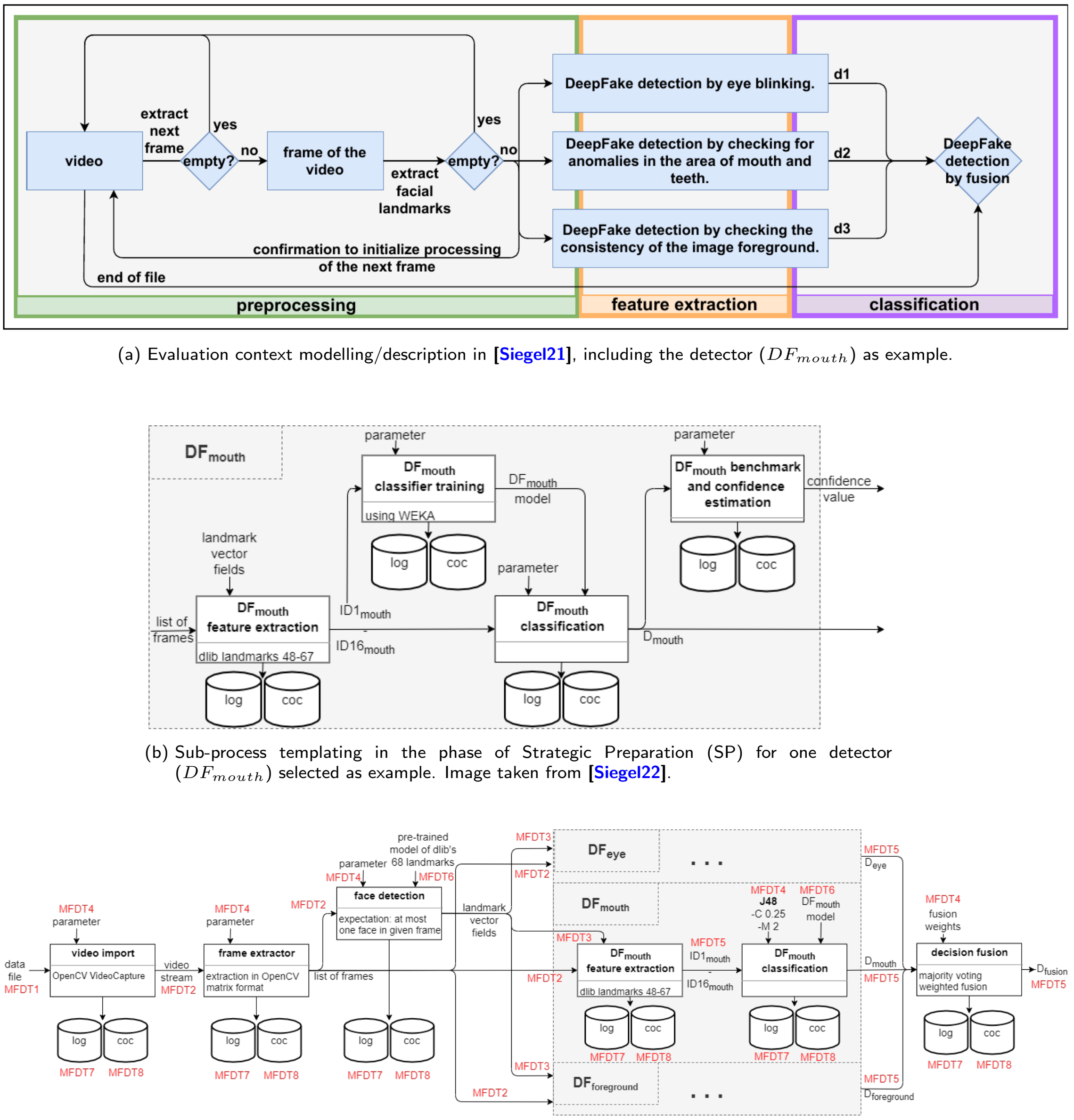
(b) Sub-process modelling from [Siegel22]

Figure 3.3: Comparison of the (sub-)process component modelling from ENFSI BPM FIT (sub-figure (a)) and one example of the author's own work (sub-figure (b)).

Regarding the considerations on templating versus instantiation of forensic processes, Figure 3.4 shows two different descriptions of the same investigation process. Sub-figure (a) represents an early conceptual model of the investigation as published in **[Siegel21]**. Sub-figures (b) and (c) represent the templating of a sub-process selected as an example in the phase of SP and the instantiation of the whole process in the phase of OP of an investigation. It should be noted here that later feeder papers like **[Kraetzer23]** expand this operational modelling even further, e.g., by including the different human operators required in the practical instantiation of a forensic investigation (see Chapter 14 of this cumulative habilitation treatise).
In addition to the much more precise description of the investigation process that is made possible by the concept of templating and instantiation, this concept also motivates the feedback loop from the investigations into SP that was first mentioned in **[Kraetzer22]**. It has to be assumed that during the investigation process itself, necessary improvements or updates to templates will be identified by the human operators/practitioners tasked with conducting the corresponding (sub-)processes. Usually, these would be communicated to the case leader who in turn (considering the investigation context) should be in a position to invoke updates of the template (and potentially initiate the re-certification of the updated template, etc.) if such a modification seems beneficial.

Only the feeder paper **[Siegel23a]** addresses the issue of the **classification of forensic methods**, the third main aspect in the BSI ‘Leitfaden IT-Forensik’ [BSI11]. As discussed in Section 3.2.7, the classification scheme introduced in the ENFSI BPM for Digital Image Authentication (DIA) [ENFSI21] seems more suitable for this application domain than the original scheme introduced in [BSI11].

In summary, the results presented here and in the corresponding feeder papers contribute to all requirements REQ1-REQ5 derived in the problem outline of this treatise in Section 1.3.



(a) Evaluation context modelling/description in [Siegel21], including the detector (DF_{mouth}) as example.

(b) Sub-process templating in the phase of Strategic Preparation (SP) for one detector (DF_{mouth}) selected as example. Image taken from [Siegel22].

(c) Process instantiation in the phase of Operational Preparation (OP) of the whole investigation process, including the sub-process for the detector DF_{mouth}. Image taken from [Siegel22].

Figure 3.4: Comparison of two different illustrations of the same investigation process: Sub-figure (a) shows a fusion-based detection approach as discussed in [Siegel21]; sub-figures (b) and (c) represent a later modelling of the same process (from [Siegel22]), with the templating (of the detector (DF_{mouth}) selected as example) in sub-figure (b) and instantiation in sub-figure (c).

3.3 Work Published in the Application Domain of Forensic Steganalysis

The conceptual modelling work done by the author in the context of the UNCOVER project (2021-2024, see Section 1.2) supports the empirical work of the research group leader and colleagues on the topic of forensic steganalysis. As explained in Section 1.2, parts of the research project UNCOVER have been classified by the EU as RESTREINT UE/EU RESTRICTED (R-UE/EU-R). As a consequence, publication activity on the results in the project is significantly reduced in comparison to other research projects with a significant number of academic partners. Only one paper addressing this research context is used as a feeder paper in this cumulative habilitation treatise. In the case of **[Kraetzer24]**, the publication underwent the specified procedure with the UNCOVER security advisory board (SAB) to prepare the paper for publication and verify that its content does not reveal any classified information. The following overviews are excerpts of the findings from this feeder paper and discuss them in the wider context of this cumulative habilitation treatise.

3.3.1 Conceptual Modelling for Forensic Steganalysis

In **[Kraetzer24]** (included as Chapter 16 of this cumulative habilitation treatise), the conceptual modelling work in forensic steganalysis is summarised based on [Provos02] and [Fridrich09] (see page 247). In [Provos02], the authors introduce an early multi-stage forensic steganalysis approach based on a multi-class steganalysis tool called Stegdetect to perform steganographic detection and tool attribution (i.e., tool identification), and a verification engine called Stegbreak to (try to) perform result verification by a message retrieval and interpretation attempt. The scope of the empirical evaluation in [Provos02] was limited to a set of three image steganography algorithms (JSteg, JPHide and OutGuess 0.13b, which were state-of-the-art algorithms in 2002) used for training the multi-class detector in Stegdetect and a multi-language dictionary of about 1,800,000 words used together with JPHide's retrieval function as core for the password brute-force tool Stegbreak.

In [Provos02], Stegdetect is then applied blindly (without knowledge of the true class) to two million images downloaded from eBay auctions and one million images obtained from USENET archives. As a result, Stegdetect classified over 1% of all images as apparently having been steganographically altered (mostly by JPHide) and therefore containing hidden messages. These images attributed to JPHide were then fed into Stegbreak under the assumption that at least some of the passwords used as embedding key for the steganographic embedding were weak passwords (i.e., words contained in the dictionary). To verify that their tools work correctly, Provos and Honeyman inserted tracer images into every Stegbreak job. As expected, the dictionary attack found the correct passwords for these tracer images. However, they did not find a single genuine hidden message. In their paper, the authors offer four possible interpretations of this result: a) there is no significant use of steganography on the internet; b) they have been analysing images from sources that are not used to carry steganographic content; c) nobody uses steganographic systems that can be found with their detector; or d) all users of steganographic systems carefully choose passwords that are not susceptible to dictionary attacks. Even though the result of this large-scale investigation was negative, the methodology and concepts for addressing the interpretability of the evaluations in [Provos02] are remarkable because they try to exploit an inherent weakness of steganography for this media forensics analysis approach: Actual steganography tools embed a message to be extracted again on the recipient's side. So if the key/password used by the covert sender in the embedding of the message is known or can be obtained (e.g., through a brute-force analysis of the key-space), the forensic practitioner performing the investigation can validate the success by extracting and decrypting/deciphering the message. In this regard, steganalysis shares a lot of similarities with cryptanalysis, with similar problems with the assumable analysis times for brute-forcing modern, well-designed schemes. In contrast to most encryption tools currently available, many existing steganography tools (especially older ones) use a short-cut to ensure successful synchronisation between covert sender and covert receiver: In these cases, the message to be embedded is expanded by a corresponding prefix header, usually containing a hard-coded synchronisation pattern and sometimes also additional information such as a checksum, etc.

Based on work like [Provos02], the text-book [Fridrich09] describes a conceptual model for forensic image steganalysis in more detail, addressing the following six steps:

1. Selection of investigation targets
2. Reliable 2-class detection that distinguishes steganographic images from cover images
3. Identification of the embedding method
4. Identification of the steganographic software
5. Searching for the stego key and extracting the embedded data
6. Decoding/deciphering the extracted data and obtaining the secret message (cryptanalysis)

Steps 2, 3 and 4 form a sequence of analyses based on machine learning / AI that usually needs to perform a reliable 2-class detection first to distinguish steganographic images from cover images. This step can be implemented by various means, including universal steganography detectors trained for a cover type, (banks of) trained targeted detectors or anomaly detectors trained for a specific source model. Without high accuracy in the methods used in this step, all consecutive analyses become much more difficult if not impossible to perform. The third and fourth step are closely related n-class classification/attribution problems, looking into the identification of the embedding method (e.g., least significant bit (LSB) replacement for images) and the specific implementation (i.e., software tool; e.g., one of the hundreds of different LSB replacement implementations for image steganography currently available on popular platforms like Github).
Publications like [Nissar10] and [Birnbaum23] show that in case the steganographic tool leaves detectable, tool-specific signatures in the meta-data or the content of a stego file, steps 2, 3 and 4 can potentially be combined into one operation performing multi-class detection and attribution.

Synopsis: Steganalysis on actual steganography tools[34] is one of the media forensics sub-domains where the verification of analysis results can be performed with intrinsic methods, here by using the corresponding message retrieval function of the steganographic tool and trying to obtain the key used to control the embedding. It has to be pointed out that if the steganography tool is designed and implemented well, this is no trivial task, comparable to the cryptoanalysis of modern-day encryption methods. Since the work of Provos and Honeyman in 2002 [Provos02], this basic assumption has defined the conceptual model for forensic steganalysis. More recent work (including [Birnbaum23]) shows that in case the steganography tool is not implemented well, signature-based detection and attribution can be used to improve forensic steganalysis in practice.
For other media forensics methods, similar forensic helper methods might be worth considering. In the case of the example of PRNU-based camera authentication discussed in Section 2.2.2, the EXIF metadata analysis and matching can be considered to be such a helper method, which would allow to perform a very fast camera model identification and can then be accompanied by the actual camera verification using the PRNU fingerprint. Another helper, this time relevant for the face morphing attack detection and DeepFake detection scenarios discussed in Sections 3.1 and 3.2, would be the wide-spread inclusion of digital watermarks in all digital cameras sold, as currently discussed by an industry alliance of global news organisations, technology companies, and the camera manufacturers dominating the marked for professional digital cameras (see `https://asia.nikkei.com/Business/Technology/Nikon-Sony-and-Canon-fight-AI-fakes-with-new-camera-tech`).
These considerations relate to requirements REQ1, REQ2, REQ3 and REQ4, as specified in Section 1.3.

3.3.2 Aspects of Explainability and Interpretability in Forensic Steganalysis

In **[Kraetzer24]**, a series of empirical evaluations is performed using Stegdetect, a 20-year-old steganalysis tool. Initially, the corresponding experiments were intended to evaluate how well trained models age, using the hard-coded detector models of Stegdetect, trained in 2002, on the output of steganographic tools supposedly supported by Stegdetect but used on a more recent image database.

[34] The main difference between academic research on steganographic methods and practical steganography is that most of the academic algorithms are only embedding simulations (without a corresponding retrieval function to extract the message) while actual steganography tools embed a message to be extracted again on the recipient's side.

This is done to counter a phenomenon that is known in the steganalysis community as cover-source mismatch (CSM) problem, the mismatch between the training data distribution and the unseen data used in testing, which is known to lead to a substantial loss of detection performance. CSM was first documented in [Goljan06], where it was observed that training a classifier on a dataset containing images only taken with a given camera and testing it on a second dataset built only using another camera led to far poorer performance than when the classifier was tested on a dataset built only with the first camera. This issue became even more evident during the competition 'Break Our Steganographic System' (BOSS) [Bas11], where the organisers added images to the testing set which were taken with a camera not present in the training set. This resulted in a significant drop in steganalysis performance on these very images. What is rarely highlighted is that these outliers were not only taken with an unknown camera, but that they had also all undergone double JPEG compression while the images used in training, had only been compressed once. This indicates that the processing pipeline might also play an important role in steganalysis performance.

The results presented in the feeder paper **[Kraetzer24]** (included as Chapter 16 of this cumulative habilitation treatise) show that this media forensics solution driven by machine lerning suffers significantly from CSM (i.e., ageing effects of the trained models): In the corresponding experiments, the performance obtained is much poorer than the original performance reported in 2002, even though the conditions of the tests are closely reconstructed (see pages 248 to 251).

Based on this first set of results, the black-box detector of Stegdetect is turned into a (more) transparent (here, gray-box) mechanism, which is achieved by using the raw feature vectors that could be obtained as output from Stegdetect for debugging, and training new detection models using a set of classification algorithms from Weka [Frank16]. For two of the evaluated algorithms, the newly trained models show a significant improvement in detection performance in comparison to the original detector models, confirming the need for re-training detectors to counter ageing effects (relevant for REQ4).

Additional experiments are conducted in **[Kraetzer24]**, aiming at understanding the feature space used in Stegdetect and thereby addressing explainability and interpretability issues raised (among others by [INTERPOL23] and the EU AIA, see Section 2.3.5 and the feeder paper **[Kraetzer22]**). These experiments show that Stegdetect apparently learns the statistical characteristics caused by the JPEG encoder used. In the case of steganography tools that embed in the JPEG transform domain (e.g., Jsteg, which directly modifies the discrete cosine transform (DCT) coefficients), these tools basically implement their own, non-standard JPEG encoder. Therefore, the attribution of the steganography tool as an attribution of the encoder will give reliable results in their case. In other cases, a high number of false positives will occur in applied steganalysis for all other tools that are using the same JPEG encoder library as the steganography tool under analysis.

Synopsis: The work presented in **[Kraetzer24]** includes multiple items that are relevant in the context of this treatise: Firstly, as discussed in Section 3.3.1, it shows that the intrinsic characteristics of an application domain (here, the retrieval function of a steganography tool) might aid the media forensics investigations. Secondly, it indicates that detector models for media forensics detectors age and might require re-training with novel, more representative content. Especially in media forensics, trained models are assumed to age considerably as the assumed source characteristics change significantly over time. This can be illustrated well with digital images, where the technical developments since the late 1990s have seen a steady increase in resolution as well as abrupt changes with new image formats. Thirdly, the evaluation (and re-evaluation) of methods requires evaluation setups that help to understand the learned classification boundary trained for ML-/AI-based media forensics methods. This was illustrated in **[Kraetzer24]** by showing that the detector learns to distinguish the characteristics caused by different JPEG encoders and not (as assumed) the characteristics of a steganographic method as implemented by a specific steganographic tool. In case these tools implement their own, non-standard JPEG encoder, the outcome might be reliable steganalysis. If the tools use standard JPEG libraries, the result of using the Stegdetect feature space will be a very high number of false positives for original images.

Concerning the explainability and interpretability of the outcomes of such forensic investigations (as requested, among others, in [INTERPOL23]), quality assurance[35] and proficiency testing need to ensure not only the quality of forensic methods included in trustworthy forensic processes but also the necessary technical expertise of the individual practitioners involved in an examination. Explainability and interpretability also need to enable the investigators to have sufficient understanding of the ML/AI systems used and to be able to ascertain and demonstrate the validity and integrity of evidence in the context of criminal proceedings. These issues relate to requirements REQ1, REQ2, REQ3 and REQ4, as specified in Section 1.3.

[35] This includes the initial evaluation of methods as well as a cyclic re-evaluation of the models trained and the feature spaces used to be performed during the operational life of such an ML-/AI-based forensic method.

4

Summary, Conclusions and Considerations on Potential Future Work

This chapter concludes the habilitation treatise and in Section 4.1 provides a comprehensive summary, reflecting the content of the treatise in relation to the requirements specified in Section 1.3. This is followed by a conclusion in Section 4.2 and considerations for possible future work in Section 4.3.

4.1 Summary of the Contributions

Sections 3.1.4, 3.2.8 and 3.3.2 contain short overviews of the results for the three selected application domains (face morphing attack (FMA) detection, DeepFake detection and forensic steganalysis) considered in this cumulative habilitation treatise.
Regarding requirements REQ1 to REQ5, as specified in Section 1.3, the synopsis of each section in Chapter 3, except for the summaries, contains an explicit projection onto the requirements concerned. In addition, these projections are also included in a condensed version in the feeder papers overview in Section 1.4.

What remains in terms of a summary is a reflection on the overall results of the work conducted by the author together with the working group leader and the various post-doc and PhD candidate co-authors. This reflection is again based on the requirements specified in Section 1.3:

- **REQ1: 'Describing necessary conditions for using a method'**
  In all papers involved, the intention was to provide precise and reproducible descriptions of methods and evaluation setups. The conceptual and operational modelling operations aimed at providing compact[36] yet clear descriptions to fit this purpose.
  The work on modelling media generation processes and source models, especially in the application scenario of face morph attach detection, focuses strongly on technical pre-requisites for the media forensics approach (see e.g., the discussions on passport-scaled images in Section 3.1.2, based on the work in **[Kraetzer17]**, included as Chapter 6). For the application scenario of DeepFake detection, the focus shifts towards operational modelling and in particular to the importance of the preparation (here termed 'templating') of forensic methods in the phase of Strategic Preparation (SP). This preparation includes the training of detector models to be used later (in Operational Preparation) in the instantiations in investigations/evaluations (see Section 3.2.4).
  For the application scenario of forensic steganalysis, the issue of model ageing (called the 'cover-source mismatch' in this domain, see Section 3.3.2) is a strongly limiting factor for the use of trained methods. Furthermore, the individual steps of the six-step conceptual model for forensic steganalysis summarised from existing literature have dependencies (i.e., they build on the results of the previous steps) which have to be taken into consideration (see Section 3.3.1).
  Regarding the work presented here for conceptual and operational modelling, it is hoped that this work might improve the communication between researchers and developers of forensic methods, and the corresponding forensic practitioners intending to use these methods.

[36] From an academic perspective, this also helps to adhere to the page number limits specified by the publishers of workshop, conference and journal papers in addition to the benefits for the documentation of forensic investigations.

- **REQ2: 'Evaluation of new methods'**
  The evaluation methodology and the evaluation results have been published for the evaluation of all methods. The tests were performed with a statistically significant number of samples and aimed at relevant domain coverage, using established third-party reference datasets wherever possible to enable comparability with the results of other researchers.
  The empirical results presented for all feeder papers can be summarised as follows:

  (a) Media forensic detectors are presented that obtain positive results, i.e., perform significantly better than random guessing in the corresponding tasks.
  (b) In no case are perfect detection results (i.e., 100% detection accuracy) obtained for relevant setups.
  (c) An overfitting to specific attack implementations is observed with all tested detectors, leading to generalisation problems for these detectors.
  (d) Detector sequences are successfully used to reduce error rates.
  (e) Fusion approaches are used to improve detection performance and generalisation power of the detection approaches.
  (f) The specific problems encountered with fusion approaches are discussed.
  (g) Issues regarding model ageing are discussed.

  The limitations of all tested methods are communicated clearly, and especially the tendency of overfitting towards specific attacks (item (c) above), the problems encountered with fusion-based approaches (item (f) above) and selected model ageing issues (item (g) above) are discussed in detail (see e.g., Section 3.2.5, Section 3.1.3.3 and Section 3.3.2 respectively).

- **REQ3: 'Standardisation of investigation processes'**
  It has to be conceded that successful standardisation work is outside the scope of what could be achieved with this habilitation project. Nevertheless, the work on operational modelling presented for the application scenario of DeepFake detection in Section 3.2 might assist standardisation efforts in this domain.

- **REQ4: 'Re-evaluation of methods'**
  In the course of this habilitation project, the same methods are evaluated iteratively with changing application contexts, data sets and other constraint variations. A good example of such a constraint variation are the empirical evaluations of the impact of data minimisation discussed in Section 3.2.6 (based on feeder paper **[Siegel23b]**, included as Chapter 13 of this cumulative habilitation treatise). This variation is not motivated by technical developments, but rather by organisational requirements, in this case the intention to find a method to train DeepFake detection models that better comply with the European Union (EU) GDPR.
  It is foreseeable that the current trend towards regulating AI development and use (e.g., with the upcoming EU Artificial Intelligence Act (AIA)) might trigger a number of re-evaluations of established procedures and methods, especially with a focus on explainable AI methods and decisions. In addition to changing external influences and new technological developments, specific characteristics of model-driven methods, especially the rate with which the corresponding (trained) models age, are also a factor to be considered as a driver for a cyclic re-evaluation of methods.

- **REQ5: 'Publication of methods and processes'**
  The original publication work was already achieved with the publication of the feeder papers and, in case of the workshop and conference papers, their presentation and discussion at the corresponding scientific events. This cumulative habilitation treatise places these individual publications in a wider context and at the same time integrates the feedback received from reviewers, workshop and conference audiences as well as the project partners in the corresponding research projects.

4.2 Conclusions

Using the common structure of the European Network of Forensic Science Institutes (ENFSI) Best Practice Manuals (BPMs) discussed in Section 2.3.4 as a projection surface, the following conclusions are drawn based on the work performed in the course of this habilitation project:

Personnel:
In addition to the 'customer' (i.e., the beneficiary of a forensic report), the [ENFSI21] defines the separate roles of the 'case leader' and the 'examiner'. This segmentation has administrative reasons, but is also a crucial requirement to prevent bias[37] in the investigation. In addition to human-in-control aspects, the feeder paper **[Kraetzer23]** focuses on such a separation of duties in the evaluation, benchmarking and certification of model-driven forensic methods (see Chapter 14, page 227).
A second important aspect is the need for a 'local quality management', which involves not only technical aspects (regarding forensic methods and processes; REQ2 and REQ4) but also personnel and corresponding questions of training and certification. This also relates to the aspects of human operators in forensic processes as well as to the corresponding discussions on human-in-control and human-in-the-loop in **[Kraetzer23]**.
A third important aspect that is of uttermost importance for this habilitation project are the requirements specified in ENFSI BPM Forensic Examination of Digital Technology (FIT) [ENFSI15] for all technical personnel involved in forensic processes, especially the aspects that ask for a close relationship with the corresponding research communities (including "[p]*ublication of a technical paper in a recognised peer reviewed forensic journal related to digital technology/evidence*" [ENFSI15], roughly equivalent to REQ5 of this treatise) and to "[a]*id in the development of local procedures and standards and improve the technical advancement of examinations*" (see Section 2.3.4.1). Here, the corresponding academic research communities would have to be open to assist the forensic practitioners accordingly.

Classes of methods:
As pointed out in Section 3.2.8, only the feeder paper **[Siegel23b]** addresses the issue of the classification of forensic methods (the third core concept in the German Federal Office for Information Security (BSI) 'Leitfaden IT-Forensik' [BSI11]). The work in this feeder paper relates to REQ3 and does not model the media forensic classes of methods required for the application domain of DeepFake detection based on the method classes from the BSI 'Leitfaden IT-Forensik' [BSI11] or the Data Centric Examination Approach (DCEA) [Kiltz20], but instead based on the work in the ENFSI BPM Digital Image Authentication (DIA) [ENFSI21], extending the modelling as necessary and projecting the Media Forensic Data Types (MFDTs) from **[Siegel22]** onto the updated scheme (see Section 3.2.7 and **[Siegel22]**, included as Chapter 15 of this cumulative habilitation treatise).

Validation and estimation of uncertainty of measurement:
Much research effort has been invested in the validation of methods and an estimation of their uncertainty of measurement (here: error behaviour). The results and their shortcomings (especially in terms of generalisation power) are summarised in Section 4.1 for items REQ1, REQ2 and REQ4. It has to be pointed out explicitly that all methods discussed as part of practical investigations in the context of this treatise have to be considered to be 'non-verified functions' in the terminology of the ENFSI BPM FIT [ENFSI15] (see Section 2.3.4.3 for the differentiation between 'non-verified functions', 'verified functions', 'validated processes', etc.). Everything above this level is entirely outside the possibilities of what can be achieved by a habilitation project like this one, but the work presented in the context of this cumulative habilitation treatise could act as a starting point for the (in-lab) validation of methods. This would have to be accompanied by many other efforts to allow for the creation of 'verified functions' as the basis of 'validated processes' or 'trustworthy processes' in forensic laboratories. These additional steps would need to include tasks like industrial-strength implementations of the methods as well as the creation of large, well-curated reference datasets for training, benchmarking and testing purposes.

[37] Various forms of contextual and operational bias are a crucial threat to police investigations as well as to forensic investigations, as illustrated for example by the misidentification by the FBI in the famous Brandon Mayfield case [DOJ11].

Quality assurance:
Even though the aspect of quality assurance in the ENFSI BPM is very much focused on verifying "*that* [in] *the existing laboratory process human-based risks remain within acceptable bounds*" [ENFSI15] (see Section 2.3.4.4), it is also aligned with REQ1, REQ2 and REQ4 of this treatise. Quality assurance requires in-house quality controls which are invoked regularly as well as event-based to safeguard the methods and processes used. The benchmarking work performed in the context of this treatise could help in designing the corresponding setups when establishing such quality controls.

Case assessment / initial assessment:
This item relates to REQ3 of this treatise and shows the relevance of the operational modelling performed for the application domain of DeepFake detection. For the involved forensic laboratory, the forensic investigation starts with case assignment, case assessment and the selection of the case leader. At this point, the case-related documentation in the lab begins (extending the documentation and chain-of-custody documents handed over to the lab), and templates for investigation processes prepared in the phase of SP are instantiated by the case leader or the examiners in Operational Preparation (OP); see Section 3.2.4).

Evaluation and interpretation:
These aspects again relate to REQ1, REQ2 and REQ4 of this treatise and address two relevant aspects: firstly, the "*performance of the elementary method on available datasets*" and "*information on its discriminating power*" [ENFSI21] - which is again a benchmarking aspect as already discussed above for the validation of methods and procedures as well as the quality assurance aspects -, and secondly, the importance of the experience of the examiner as well as of the case leader in estimating the discriminating power of elementary methods, resulting in conclusions that state the evidential weight as a level of support for each of the competing propositions (see Section 2.3.4.6).

Presentation of results:
The ENFSI statements on the presentation of results, as discussed in Section 2.3.4.7, point out that in forensics not only 'courtroom-ready' results are relevant, but other forms of reporting that are not "*designed to such stringent levels as those that are required for court review / use*" [ENFSI15] might also be of use. Nevertheless, even in these cases, all reporting "*should still enable independent review or reproduction of the reported results*" [ENFSI21]. This request for reproducibility is a strong driver for the conceptual and operational modelling discussed in Chapter 3 of this habilitation treatise and benefits from all progress made towards the five requirements (REQ1 to REQ5) specified in Section 1.3.

Tool development:
Performing tool development for forensic institutions is outside the scope of this habilitation project. In fact, most academic research is restricted by the funding instrument to specific technology readiness levels (TRL; see Section 2.3.4.8). In the case of the research projects ANANAS, FAKE-ID and UNCOVER, which provided the context for this habilitation project, the maximum TRL of the outcome was intended by the funding body to be TRL 6 ('technology demonstrated in relevant environment'). One result worth mentioning are two patent applications resulting from the ANANAS project that were filed with the European Patent Office and list the author as co-inventor. Both patent applications describe methods for face morph detection in the document life-cycle of eMRTD as discussed in Section 3.1.2 and Section 3.1.3. This relates to REQ1, REQ3 as well as to REQ5.

4.3 Considerations on Potential Future Work

The national digital forensic science strategy paper [Vaughan20] published by the National Police Chiefs' Council (NPCC) of the UK clearly states that the members of the NPCC assume that a change of current standard operational procedures (SOPs) in digital forensics (DF) is necessary (see Section 2.2.1). Even though the NPCC in its document obviously addresses the current situation in the United Kingdom of Great Britain and Northern Ireland (UK), it can be assumed that these observations can also be generalised to a large extent for other local contexts, including the situation in Germany (see Section 2.3.3).

An academic publication such as this habilitation treatise cannot directly propose SOPs for a forensic sub-domain. This would require standards published by the corresponding authority, like the German BSI or the National Institute of Standards and Technology (NIST) in the US. What a publication like this treatise might achieve is to provide stakeholders like members of ENFSI or policy makers in executive systems with arguments and recommendations for changing established best practices (like the ENFSI BPMs) or policy documents, which might then eventually result in updated SOPs.

The following two subsections present a possible roadmap from the perspective of this habilitation treatise for the technical (i.e., research and development) and organisational ('Improved Operations') aspects.

4.3.1 Research and Development

[Vaughan20] recommends that central national police bodies should "*coordinate and influence an R&D programme drawing in R&D effort nationally*". They should link "*casework-driven capabilities that practitioners develop*" with academic research and the research and development activities that are performed by vendors of forensics tools. Unfortunately, the latter group, the vendors of forensics tools, is becoming smaller, with [Vaughan20] attributing this to the fact that "*forensic science funding reductions have led to a substantial decrease in the size of the supplier market*". One way of compensating for this would be to opt for the 'improved commercial practices' recommended by [Vaughan20] to strengthen the supplier market (see Section 2.2.1). The author sees an alternative option in the practice of countries like the Netherlands or Canada, where governments heavily invest in technical solutions developed by security forces or forensic institutions themselves. In Germany, a new institution, the Central Office for Information Technology in the Security Sector (ZITiS), was founded to assist German law enforcement agencies with research and development activities. Such institutions would also be in an ideal position to influence the research agendas of national R&D programmes, build strategic relationships with academia and industry to develop and access new capabilities and host the academic advisory groups recommended in [Vaughan20].
In this context, it is interesting to see that EU-wide efforts are emerging in parallel with national efforts. While ENFSI tries to homogenise best practices, the European Anti-Cybercrime Technology Development Association (EACTDA) was founded in 2020 for the development of technological solutions (i.e., tools) for European law enforcement agencies (LEA) and forensic laboratories.

The conclusions drawn in this habilitation treatise for this aspect from a German academic perspective are the following:

- ENFSI is a highly esteemed entity in the field, and its Expert Working Group (EWG) massively support the whole field with their BPMs. Nevertheless, the interaction with academic research could be improved. The ENFSI Annual Meeting, for example, is open only to ENFSI members (see e.g., `https://enfsi.eu/agenda/enfsi-annual-meeting-2023-in-the-hague/`). Opening up part of this event for outsiders might significantly foster collaboration.
- As summarised in Section 2.3.6, the situation in (large parts) of Europe is significantly different from the situation in the US in terms of regulations and 'market' size for forensic R&D.[38] Even

[38] In the view of the author, the main difference lies in the strength of the private sector in US forensics, with a resulting preference for commercial tools with certification in comparison to the EU / ENFSI preference for verified tools (preferably

between European countries, the conditions for forensic practitioners differ significantly. As a result, a national policy for building strategic relationships with police forces, forensic practitioners, academia and industry to develop and access new capabilities as well as for the coordination of R&D funding in this field would have to be formulated in each country based on national legislation. For an EU country like Germany, these efforts would need to be closely aligned with the European stakeholders (e.g., Europol and EACTDA).

- The funding of academic research efforts in the forensics domain should always include practitioners (LEA or forensic institutions) as partners or associated partners to ensure that the research goals are chosen with practical application in mind. This requires law enforcement agencies and forensic laboratories willing to participate in funded research initiatives, even though the funding instruments usually do not foresee a progression above TRL 6, which means that the project outcomes will usually not be usable directly in the daily work of the participating LEA.[39]

- On the national or international (e.g., EU) level, funding and technology transfer schemes are required to help the research demonstrators, which are the usual outcome of publicly funded research as described in the previous item, overcome the so-called 'Valley of Death', i.e., raise them from TRL 6 to a TRL 9 solution. Potential actors have been founded in the last few years with ZITiS and EACTDA, which might be capable of providing this support. Part of the task of such transfer schemes might include identifying promising research at early stages (TRL 4 to TRL 6) and approaching the corresponding researchers.

- More and more of the tools developed for forensics are Open Source tools that range from large (funded) projects like the Assemblyline framework (https://cybercentrecanada.github.io/assemblyline4_docs/) to much smaller projects like 'dcfldd' (https://github.com/adulau/dcfldd). On the one hand, this reflects a trend with funding bodies, which in many cases currently aim for Open Source, Open Data and Open Access regarding the funded research. On the other hand, it also reflects the status quo in information technology (IT) communities, where 'doing it once for the benefit of many' has become an established paradigm of progress. This is especially beneficial for forensic software because in these cases forking and extension as well as source code analysis as part of tool validation are possible.
  At a first glance, the Open Source concept might seem to contradict the security interests of LEAs and forensic institutions, but a good example of how to integrate both is Hansken (the open digital forensics platform developed by the Netherlands Forensic Institute (NFI) as a Digital-Forensics-as-a-Service (DFaaS) solution): It is not Open Source in the sense of public Open Source but only provided to specific registered entities (mostly LEA and forensic institutes) together with a public Software Development Kit (SDK) for developing new plug-ins and components.

- Efforts are necessary to align the quality considerations between academics and forensic practitioners in many fields, including media forensics. The differences between academic 'lab condition' evaluations and robust forensic laboratory proficiency testing of methods often lead to different estimates of the maturity of methods. In academia, about 80% of the effort usually goes into the design and implementation of a method, and 20% into testing. For forensic labs, about 20% of the effort usually goes into the design and implementation/purchase, and 80% into testing, including validation by independent third parties, cyclic / event-based re-evaluation as well as proficiency testing (see Sections 2.3.4.3 and 2.3.4.4).

including source code review for Open Source tools) in validated processes. To wait for solutions required for (media) forensics investigation problems coming from the US seems not to be a viable solution for European practitioners because the market is very small in comparison - only a limited number of forensic laboratories (usually one in a smaller country, up to a few (dozen) in a large country) in contrast to the thousands of forensic labs active in the US (see Section 2.3.1). As a consequence of the very limited number of potential customers, the market is not very attractive for commercial software development or even for the adaptation of forensic software originally developed for the US market. For the sovereignty of European solutions, this issue will have to be addressed (as is currently seen with the example of Hansken in the Netherlands). The ENFSI BPMs discuss tool development by forensic laboratories as one item in their recommendations (see Section 2.3.4.8).

[39] In the experience of the author, gathered during various funded research projects, there is often a clash of expectations in joined research projects, where end-user partners hope for field-applicable methods/tools as an outcome, while participating academic partners intend to strictly stick to the boundaries imposed by the funding body - often a maximum TRL between 4 and 6 - in order to not jeopardise current and potential future funding.

4.3.2 Improved Operations

In [Vaughan20], the authors put an emphasis on 'improved operations', which for the NPCC means that

> "[s]*tandardising, industrialising and providing services centrally - 'doing it once for the benefit of many' are the foundations to transform DF science service. Standardising processes will allow forces to collaborate on casework, technology, R&D and quality assuring processes.*"

The current, fragmented system (which is described by the NPCC for the UK but is also found in many other countries world-wide) with a large number of national- and local-level forensics units, which individually develop their own methods, procure and deploy their own hardware and software and manage their individual quality accreditations "*involves substantial duplication of effort and inevitable waste of resources*" [Vaughan20].
The conclusions drawn in this habilitation treatise for this aspect from a German academic perspective are the following:

- Best practices documents, like the ENFSI BPMs, will have to be constantly updated to include novel methods and processes for forensic investigations. This requires a healthy dialogue between the forensics community (as represented by ENFSI) and academic communities in the corresponding fields. The ENFSI BPM FIT strongly recommends such a close interaction (see Section 2.3.4.1), but in practice there is still a lot of potential for improvement of the collaboration between practitioners and academia in this field.

- Existing national guidelines for forensic procedures (like the BSI guidelines document discussed in Section 2.3.3) need to be updated on a regular basis. Since ENFSI, due to the nature of the initiative, cannot provide SOPs but is restricted to best practice recommendations, the responsible national bodies have to turn these into standards that forensic practitioners and their 'customers' (police, prosecutors, judges, etc.) can rely on within the context of a national legislation.

- Forensic process models (as part of guidelines or SOPs) have been established as an important foundation for creating and maintaining trustworthy and validated forensic procedures. They enable reproducible procedures and precise investigation descriptions in documentation. It has to be acknowledged that they are also subject to an ageing effect: Technological as well as regulatory changes will make updates necessary. Continuously (or event-based) improving existing process models means adapting them to these changing circumstances. As the last update of the German BSI guidelines document was published in 2011 (see Section 2.3.3), an update would seem to be required, especially since new developments (e.g., the increasing role of ML-/AI-based investigation methods) have changed the field significantly since then.

- To better address the aspects of 'standardising' and 'industrialising' in the quote from [Vaughan20] given above, the guideline documents might shift from providing mostly conceptual models (e.g., the phase-driven investigation model of the BSI '*Leitfaden IT-Forensik*' [BSI11]) to more detailed operational models. A proposal for such operational models with their templating in SP and the instantiation in a specific investigation in OP is discussed in this treatise in the application context of DeepFake detection in Section 3.2.4. Work invested in this domain would also have to include efforts to define other required modelling components as part of a standardisation roadmap for forensic process models. These would have to include domain-adapted forensic data models, as discussed in [Kiltz20], [Altschaffel20] and Section 3.2.2 of this treatise.

- A drive towards more mature (media) forensics solutions has to take into account recent developments regarding ML-/AI-driven solutions (esp. the EU AIA and the Interpol initiative on 'Responsible AI Innovation in Law Enforcement' [INTERPOL23]). The current trend shows a shift towards 'human-in-control' (i.e., ML/AI as decision support systems), requiring explainability of AI systems and their decisions.[40] This is already taken into consideration to some extent[41] in the ENFSI BPMs but will assumedly have significant impact on R&D in this domain in the coming years.

- Data models are very important for industrialising forensics. Handling forensic data requires significant infrastructure. [Vaughan20] points out that "*digital forensic submissions make up the largest individual source of data in police forces*". This also poses additional challenges in the processing, analysing and sharing of (digital) forensic data, including the need "*to improve both the ability to review the output from DF examinations and how we coordinate disclosure with the* [Criminal justice system]" [Vaughan20]. In 2020, the NPCC requested the design and implementation of a national digital forensics (DF) data model for the UK:

  > "*Essential to improving how we handle DF data is to standardise the way we store this data, by developing a national DF data model. This model will define standard metadata for different types of digital forensic information – such as text messages, photos, documents or system log files – which will allow storage in a structured vendor-independent form.* [...] *This data model will support moving towards full interoperability between tools and infrastructure throughout the DF workflow, based on standard data formats and interfaces between tools to enable forces to share data between them.* [Vaughan20]"

  Obviously, such a national digital forensics data model would have to be built on an abstract data model like the one discussed in the BSI 'Leitfaden IT-Forensik' (see Section 2.3.3) and would need refinements for domain adaptation like the ones discussed in Section 3.2.2 before it would then have to be turned into precise data structure descriptions for format specification and implementation.

[40] [INTERPOL23] is very explicit on this item:

> "*In the context of criminal investigations, the explainability of AI systems used to obtain or analyze evidence is particularly important. In fact, in some jurisdictions, criminal evidence obtained with the support of AI systems has been challenged in courts on the basis of a lack of understanding of the way the systems function. While the requirements for evidence admissibility are different in each jurisdiction, a sufficient degree of explainability needs to be ensured for any AI system used to obtain and examine criminal evidence. This helps guaranteeing, alongside the necessary technical competencies, that law enforcement officers involved in investigations and forensic examinations have sufficient understanding of the AI systems used to be able to ascertain and demonstrate the validity and integrity of criminal evidence in the context of criminal proceedings.*"

. [41] ENFSI BPM FIT [ENFSI15] indicates that "*a set of minimal checks* [...] *should be applied when considering the verification of specific tool functions used within a full validation method or process.*" A list of such minimal checks provided as an example in [ENFSI15] includes questions aiming at the performance of the tool (e.g., "*Can you demonstrate that it does match its reported capability?*" and "*What are the known conditions under which it is known to fail?*") as well as questions aiming at its usability (e.g., "*What skill level is required by analysts to use the function?*").

5

[Kraetzer15a] Considerations on the Benchmarking of Media Forensics

This chapter of this cumulative habilitation treatise has originally been published as:

[Kraetzer15a] **Christian Kraetzer**, Jana Dittmann: *Considerations on the benchmarking of media forensics*. In Proceedings of the 23rd European Signal Processing Conference (EUSIPCO 2015), Nice, France, August 31 - September 4, 2015, IEEE, pp. 61-65, 2015.
https://doi.org/10.1109/EUSIPCO.2015.7362345

Shares/author contributions: For this paper, the following shares statement has been signed by all authors:
"The following paper submitted have been produced jointly with the following shares:
Christian Krätzer: Analysis of Daubert and FRE702 criteria as forensic compliance indicators; discussion of the maturity of an exemplarily selected media forensic application scenario; definition of relevance requirements for media forensic methods; results discussion
Jana Dittmann: Definition of relevance requirements for media forensic methods; results discussion
The work has been produced jointly with the co-authors mentioned. All authors have read and agreed to the published version of the manuscript."

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

6

[Kraetzer17] Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing

This chapter of this cumulative habilitation treatise has originally been published as:

[Kraetzer17] **Christian Kraetzer**, Andrey Makrushin, Tom Neubert, Mario Hildebrandt, Jana Dittmann: *Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing*. Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec 2017, Philadelphia, PA, USA, June 20-22, 2017, pp. 21-32, 2017.
https://dl.acm.org/doi/10.1145/3082031.3083244

Shares/author contributions: For this paper, the following shares statement has been signed by all authors:
"The following paper submitted have been produced jointly with the following shares:
Christian Krätzer: Designing the life-cycle model for documents for photo-based ID verification; work on attack comparison using the document life-cycle model; integration of document life-cycle and image editing model; design of the empirical evaluations; results discussion
Andrey Makrushin: Designing the image editing history model for identifying potential forensic traces for face morph attacks; specification of editing histories of a morphed images; design of the empirical evaluations; results discussion
Tom Neubert: Design and implementation of the morphing detector; data acquisition; design of the empirical evaluations; results discussion
Mario Hildebrandt: Design of the empirical evaluations; work on the benchmarking of the approaches using StirTrace; results discussion
Jana Dittmann: General concept; results discussion
The work has been produced jointly with the co-authors mentioned. All authors have read and agreed to the published version of the manuscript."

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

7

[Neubert19] A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on eMRTD

This chapter of this cumulative habilitation treatise has originally been published as:

[Neubert19] Tom Neubert, **Christian Kraetzer**, Jana Dittmann: *A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on eMRTD*. Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'19), July 2019, pp. 95–100. 2019.
https://doi.org/10.1145/3335203.3335721

The original publication is available as Open Access publication at the ACM Digital Library: https://dl.acm.org/doi/10.1145/3335203.3335721

Shares/author contributions: For this paper, the following shares statement has been signed by all authors:
"The following paper submitted have been produced jointly with the following shares:
Tom Neubert: Design and implementation of the new spatial domain feature space; Design of the ICAO-aligned pre-processing module; generation of the training data; definition of the evaluation goals; results discussion
Christian Krätzer: General concept; Design of the ICAO-aligned pre-processing module; definition of the evaluation goals; results discussion
Jana Dittmann: General concept; Design of the new spatial domain feature space; Design of the ICAO-aligned pre-processing module; results discussion
The work has been produced jointly with the co-authors mentioned. All authors have read and agreed to the published version of the manuscript."

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

8

[Neubert18a] Reducing the False Alarm Rate for Face Morph Detection by a Morph Pipeline Footprint Detector

This chapter of this cumulative habilitation treatise has originally been published as:

[Neubert18a] Tom Neubert, **Christian Kraetzer** and Jana Dittmann: *Reducing the False Alarm Rate for Face Morph Detection by a Morph Pipeline Footprint Detector*. Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 2018, pp. 1002-1006, 2018.
https://doi.org/10.23919/EUSIPCO.2018.8553067

Shares/author contributions: For this paper, the following shares statement has been signed by all authors:
"The following paper submitted have been produced jointly with the following shares:
Tom Neubert: Definition of the considered face morphing pipelines; implementation of the multi-level detection and validation process; data generation; definition of the evaluation goals; results discussion
Christian Krätzer: Design of the multi-level detection and validation process to reduce the FAR of state-of-the-art face morph detectors; definition of the evaluation goals; results discussion
Jana Dittmann: General concept; definition of the evaluation goals; results discussion
The work has been produced jointly with the co-authors mentioned. All authors have read and agreed to the published version of the manuscript."

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

9

[Kraetzer21] Potential Advantages and Limitations of Using Information Fusion in Media Forensics – A Discussion on the Example of Detecting Face Morphing Attacks

This chapter of this cumulative habilitation treatise has originally been published as:

[Kraetzer21] Christian Kraetzer, Andrey Makrushin, Jana Dittmann, Mario Hildebrandt: *Potential Advantages and Limitations of Using Information Fusion in Media Forensics - A Discussion on the Example of Detecting Face Morphing Attacks*. EURASIP Journal on Information Security 2021, 9 (2021). https://doi.org/10.1186/s13635-021-00123-4

Shares/author contributions (as given in the original paper - see page 138 of this cumulative habilitation treatise): “*CK works on the media forensic perspectives and fusion theory parts as well as the interpretation of results. AM work on the biometric perspective, the dataset creation, the conduction of the experiments (classifier selection and fusion operator implementation), and the interpretation of the experimental results. JD initial structuring of the work and definition of focus and scope of the work (incl. suggesting the two application scenarios as well as the usage of DST and LR based fusion). MH theoretical and practical work on likelihood based fusion and the interpretation of its results. The authors read and approved the final manuscript.*”

EURASIP Journal on
Information Security

RESEARCH Open Access

Potential advantages and limitations of using information fusion in media forensics—a discussion on the example of detecting face morphing attacks

Christian Kraetzer*, Andrey Makrushin, Jana Dittmann and Mario Hildebrandt

Abstract

Information fusion, i.e., the combination of expert systems, has a huge potential to improve the accuracy of pattern recognition systems. During the last decades, various application fields started to use different fusion concepts extensively. The forensic sciences are still hesitant if it comes to blindly applying information fusion. Here, a potentially negative impact on the classification accuracy, if wrongly used or parameterized, as well as the increased complexity (and the inherently higher costs for plausibility validation) of fusion is in conflict with the fundamental requirements for forensics.

The goals of this paper are to explain the reasons for this reluctance to accept such a potentially very beneficial technique and to illustrate the practical issues arising when applying fusion. For those practical discussions the exemplary application scenario of morphing attack detection (MAD) is selected with the goal to facilitate the understanding between the media forensics community and forensic practitioners.

As general contributions, it is illustrated why the naive assumption that fusion would make the detection more reliable can fail in practice, i.e., why fusion behaves in a field application sometimes differently than in the lab. As a result, the constraints and limitations of the application of fusion are discussed and its impact to (media) forensics is reflected upon.

As technical contributions, the current state of the art of MAD is expanded by:

a) The introduction of the likelihood-based fusion and an fusion ensemble composition experiment to extend the set of methods (majority voting, sum-rule, and Dempster-Shafer Theory of evidence) used previously

b) The direct comparison of the two evaluation scenarios “MAD in document issuing” and “MAD in identity verification” using a realistic and some less restrictive evaluation setups

c) A thorough analysis and discussion of the detection performance issues and the reasons why fusion in a majority of the test cases discussed here leads to worse classification accuracy than the best individual classifier

Keywords: Information fusion, Media forensics, Face morphing attacks, Morph attack detection (MAD), Fusion methods, Fusion ensemble composition

* Correspondence: christian.kraetzer@iti.cs.uni-magdeburg.de
Otto-von-Guericke University Magdeburg, Magdeburg, Germany

1 Introduction

Information fusion has a long research history and its core concept, the combination of outputs of different expert systems, has been rigorously studied and applied for at least two decades in various application domains. The concept of fusion has been studied under many different terminologies, e.g., classifier ensembles [1], combining pattern classifiers [2], or cooperative agents [3]. As a result of the growing popularity of machine learning at that point of time and practical problems arising from ever increasing feature space complexities, in 2002 [4] stated that "instead of looking for the best set of features and the best classifier, now we look for the best set of classifiers and then the best combination method." This statement was rephrased by [5] into "the role of information fusion […] is to determine the best set of experts in a given problem domain and devise an appropriate function that can optimally combine the decisions rendered by the individual experts [...]." In [2], the following three different types of reasons why a classifier ensemble might be better than a single classifier are identified: Statistical (instead of picking a potentially inadequate single classifier, it would be a safer option to use a set of unrelated ones and consider all their outputs), computational (some training algorithms use hill-climbing or random methods, which might lead to different local optima when initialized differently) and representational (it is possible that the classifier space considered for a problem does not contain an optimal classifier). Whatever the exact reason for choosing a fusion approach instead of a single classifier, [2] explicitly warns that "an improvement on the single best classifier or on the group's average performance, for the general case, is not guaranteed. What is exposed here are only 'clever heuristics' [...]". In summary, by combining classifiers (or other expert systems), the applicants hope for a more accurate decision at the expense of increased complexity.

The huge potential for accuracy improvement gained by applying fusion has been well illustrated in many fields of applied pattern recognition. A good example is the field of biometric user authentication where, e.g., [5] shows various benefits that this field can draw from fusion at different steps of the pattern recognition pipeline. When it comes to blindly applying information fusion, among the disciplines that are currently still hesitant are the forensic sciences. Here, the potentially negative impact to classification accuracy as well as the increased complexity (and the inherently higher cost for plausibility validation) of fusion are in conflict with fundamental requirements for (media) forensics (as is discussed in more detail in section 2.1). The goals of this paper are to explain the reasons for this reluctance to accept a potentially very beneficial technique such as information fusion and to illustrate the practical problems of applying fusion. To this end, an exemplary application scenario from media forensics called face morphing attack detection (MAD) is selected. This scenario is currently a hot research topic due to the fact that this kind of attack imposes a recent and currently unsolved threat to face image based authentication scenarios such as border crossing using travel documents (i.e., passports), see section 2.3.

By facilitating the understanding of the reluctance to blindly use fusion in (media) forensics as well as the potential pitfalls of practically applied fusion techniques, it is the hope to facilitate acceptance both in the media forensics community as well as the community of forensic practitioners. To achieve this, the paper provides the following contributions:

a) As general contributions, it is illustrated why (even with a set of classifiers relevant to a specific problem) the naive assumption that fusion would make the detection more reliable can fail in practice, i.e., why fusion behaves in a field application sometimes differently than in the lab and often delivers lower detection performances than single detectors. As a result, the constraints and limitations of the application of fusion are discussed and its impact to (media) forensics is reflected upon. The two main aspects addressed in this discussion are the generalization power of classification models and the relationship between training and test data sets. In the evaluations, it is shown that both aspects, despite being similar in nature, have to be considered separately for applied information fusion.

b) As technical contributions for face morphing attack detection (MAD), the current state of the art is expanded by:

- Introduction of likelihood ratio (LR) based fusion for face morphing attack detection (MAD) to extend the set of methods (majority voting, sum-rule, and Dempster-Shafer Theory (DST) of evidence [6]) used in [7].
- Direct comparison of the two evaluation scenarios: "MAD in document issuing" vs. "MAD in identity verification."
- Analysis and discussion of detection performance issues found with the fusion based detectors (note: questions of feature or classifier selection are out of scope for this paper), the results show that:
- Fusion can fail even when a set of accurate individual classifiers is available. The results presented for fusion detectors are in the vast majority of the cases worse than the results of the best individual classifier used.

- Trained thresholding and weighting strategies as well as sophisticated (context adapted) fusion methods (especially DST and LR based) can under specific circumstances perform significantly worse than unweighted, simplistic fusion approaches like the sum-rule or majority voting.
- Different fusion ensemble composition strategies (i.e., using all available detectors vs. selecting a subset of those) have an influence on the decision error rates.
- For the two evaluation scenarios "MAD in document issuing" (*SC1*) vs. "MAD in identity verification" (*SC2*) different detection and fusion trends are observed, resulting from differences in the inherent characteristics of the application scenario (esp. the amount and type of data available for investigations).

The rest of the paper is structured as follows: section 2 performs a discussion of related work on requirements for media forensic methods, the current state of the art in face morphing attacks detection (MAD) and information fusion approaches in MAD. In section 3, the investigation concept from [7] is summarized and extended into the concept for fusion-based face morphing attack detection used in this paper. Section 4 defines the evaluation setup (incl. the two application scenarios "MAD in document issuing" vs. "MAD in identity verification"). Section 5 presents the evaluation results and their discussion, while in section 6 the conclusions are drawn from the presented results.

2 Related work

Technical capabilities (such as accuracy) are by far not the most significant characteristics of forensic methods. In general, those are usually rated by practitioners in criminal investigations by their maturity, i.e., by their scientific admissibility. Section 2.1 discusses some issues of scientific admissibility in European contexts (where, due to the very nature of the EU and its member states, it is currently much less well regulated as for example in the USA) to establish an understanding on the requirements and limitations for forensic methods originating from this field.

Section 2.2 briefly summarizes the media forensics application domain selected for this paper, the face morphing attack detection (MAD). More detailed overviews over the research activities in this field, which is very active since 2014, can be found in the two survey papers [8, 9].

Several studies have demonstrated that both manually and automatically generated high-quality morphs cannot be recognized as such neither by algorithms nor by human examiners [10–13], and even low-quality morphs pose a threat to the identity verification process if it is completely automated. This explains the urgent need for automated face morphing detectors. At the time of writing this paper, none of the existing research initiatives working on this specific image manipulation detection problem has been able to present detectors that achieve sufficient detection accuracy on a wide range of morphed images (see the ongoing NIST FRVT MORPH challenge [14]). As a logical consequence fusion approaches are used to combine the existing detectors and thereby improve the overall performance. The state of the art approaches in information fusion for MAD are briefly discussed in section 2.3.

2.1 Requirements for media forensic methods in terms of scientific admissibility

When working in media forensics, the question of determining the maturity of methods arises. In lab tests analyzing data for which ground truth information exists, an answer to that question is easy. In that case, the degree of agreement between ground truth label and detector response can simply be used to express the accuracy of the method.

In field applications of forensics, there usually exists no ground truth information for an object under investigation. In these cases, other means of establishing the maturity or suitability of a forensic method have to be used. In forensics, the whole field of work looking into this aspect is termed "scientific admissibility." It is a very complex topic on which Champod and Vuille state in [15]: "The scientific admissibility of evidence, while subject to fairly precise rules in United States law, [...], is seldom addressed in European legal writings, [...]. The question of scientific reliability is seen as intrinsically linked with the assessment of the actual evidence, that is with the determination of its probative value [...]." Researchers in the fields of computer science and applied pattern recognition have to rely on the verdict of legal experts defining the hurdles media forensics approaches have to take before achieving the ultimate goal of court admissibility. Looking at [15], it can be stated that there is no EU wide regulation on scientific admissibility questions but that there are common principles that would have to be considered. In that in-depth analysis of the current legal situation in [15] a non-exhaustive list of such principles is presented, containing in its core the following aspects:

- Methods should be peer reviewed and accepted within the corresponding scientific community.
- Error rates associated with a method should be precisely known,
- Existence of standards for the application and maintenance of methods.

This list is very similar to the state-of-the-art criteria used by judges in the USA to address the questions of court admissibility for forensic (and other) methods, i.e., the so called Daubert and FRE702 criteria [15]. While pointing out the benefits of such selection principles, Champod and Vuille also provide some form of criticism into their application: for peer reviewed methods they point out that "this criterion does not indicate whether a technique accepted in scientific literature has been used properly in a given case" and regarding the issue of ascertaining the error rates of a test, they claim that those "can prove misleading if not all its complexities are understood" [15].

In the context of work presented in this paper, those statements imply two important things: First, that a very careful investigation of the precise constrains for the application of a method such as information fusion is required for any specific forensic application case. Second, that the associated complexities in practical application (such as the attempt to improve MAD detection used for illustration purposed within this paper) are clearly and openly discussed.

2.2 Face morphing attacks and their detection

Face images in documents are an established and well accepted means of identity verification. Current electronic machine readable travel documents (eMRTD) are equipped with digital portraits to automate the identity verification process. The automation saves manpower and enhances security due to switching from subjective (officers) to objective (automated face recognition systems) matching of faces. The benefit of automation is especially relevant in high-throughput applications like an airport border control. However, the automation entails the risk of face morphing attacks [16].

In publications such as [12, 16], it has been shown that the blending of face images (here called face morphing) of two or more persons can lead to a face image resembling the faces of all persons involved. Using such an image as a reference in a document is referred to as face morphing attack because it enables illicit document sharing among several users. Such morphing attacks have been shown to be effective in an automated border control (ABC) scenario giving a wanted criminal a chance to cross a border with a chosen (i.e., wrong) identity [10, 17, 18].

Document issuing procedures are different depending on the country and its national regulations. In many countries, the biometric face image can be (and often is) submitted as a hard copy. Here, the attack aims at fooling an officer at the document issuing office by submitting a morphed face image. As long as persons are allowed to submit images to the document issuing office during the document generation, face morphing attacks will remain a severe threat to photo-ID-based verification. Indeed, if an officer accepts a morphed face image, the issued document would pass all integrity checks, and if an automated face recognition (AFR) system matches a live face with a morphed document image, access will be granted to an impostor.

The risk of the morphing attack can be reduced by supporting both officers and AFR systems with a dedicated morph detector. The only way to completely remove the threat of such attacks would be to take the picture directly in the controlled environment of the issuing office and by ensuring that there is no malware-enabled morphing attack embedded into the digital part of the document issuing pipeline, too. The question whether to take the picture directly in place is a political issue, which has in the past lead to many controversial discussions (e.g., in France and Germany) between governmental regulation and the photo industry. But even if this problem would be solved for one country, there would still be the issues of legacy passports (which might still be valid for up to 10 years) as well as foreign documents.

Figure 1 depicts the document life-cycle of a document with a face morphing attack present. While publications such as [19] also discuss the role of forensics (and anti-forensics) in the quality assessment (QA) of the attacker during the morph generation process, in the scope of this paper, only the image forensic analysis of the images submitted into the document creation and the corresponding analysis in every document usage (e.g., in an ABC gate) are relevant. These two investigation points are representing the evaluation scenarios "MAD in document issuing" (*SC1*) and "MAD in identity verification" (*SC2*) considered in this paper. They are discussed in detail in section 4.

The face morphing attack detection (MAD) approaches are typically categorized into two groups regarding whether a trustworthy reference face image is presented or not. The first group is often referred to as single-image or no-reference MAD approaches. The second group is referred to as two-image differential or reference-based MAD approaches. Despite the fact that the reference-based MAD has more potential for robust operation, the non-reference MAD approaches are better represented in the literature.

Within the group of reference-based MAD approaches, as ponted out in [21] there are two subcategories: Reconstruction-based and reference-based MAD. The most prominent examples from the first subcategory try to reconstruct a likely original face (from the assumedly morphed face image provided) by making use of a trustworthy reference face image taken life from the person in front of a camera. This process is often referred to as de-morphing. The detection is done in this

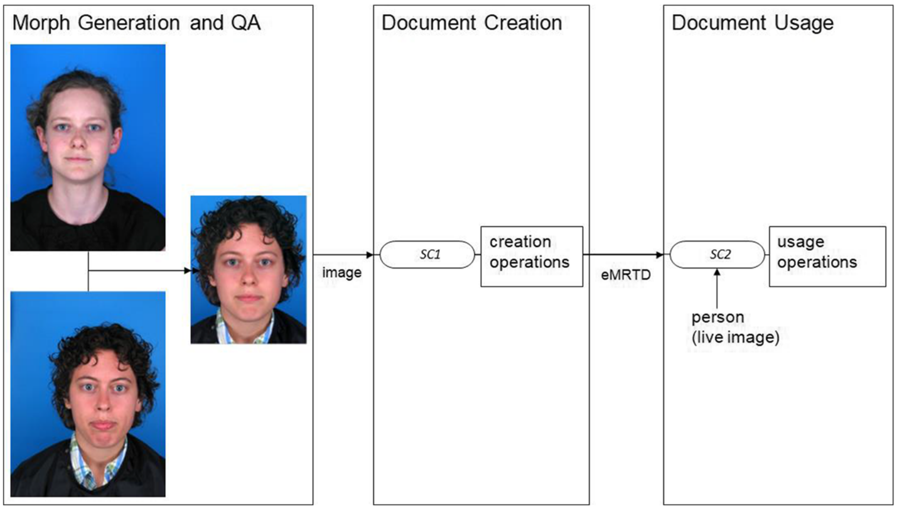


Fig. 1 Document life-cycle in case of a face morphing attack including the evaluation scenarios “MAD in document issuing” (*SC1*) and “MAD in identity verification” (*SC2*) (image derived from [19], combined morph generated based on [12] original face images taken from the ECVP face dataset [20]

case by comparing the reconstructed image and the reference one. The de-morphing is done either by inversion of the common morphing procedure [22] or by applying neural networks such as an autoencoder [23] or generative adversarial networks (GAN) [24]. Alternative approaches to implement reference-based MAD could also be relying on reference feature vectors instead of complete face images.

The approaches from the second subcategory extract features from both presented images (probe document image and trustworthy reference image) and either compare them to each other [13] or combine them for the further classification [25], or even train an additional classifier based on difference vectors [26]. The common problem of all single-image MAD approaches based on “hand-made” or "hand-crafted" features is that they do not detect morphing but rather traces of image manipulations. Since, there is a set of legitimate image manipulations such as in-plane rotation, cropping, scaling, and even some kinds of filtering the morphing characteristics can be easily simulated to prevent detection. The more sophisticated single-image MAD (like [27]) approaches make use of deep convolutional neural networks (DCNN) which are learned to automatically extract features characterizing morphing artifacts based on a large set of samples. If a training set is large and diverse enough covering all frequently used image manipulations, there is a chance that the network will learn not the characteristics of a special dataset, but actual characteristics of morphing. Training of different DCNN architectures for morphing detection was conducted in [17, 26, 28] applying transfer learning with pre-trained networks as well as learning from scratch. In [29], a feature-level fusion of two DCNNs (AlexNet and VGG19) trained by means of transfer learning is shown to outperform BSIF features.

The majority of the aforementioned detectors are learned with morphed face images created by the standard morphing approach which roughly includes three steps: alignment of faces, warping of face components given by polygons (usually triangles), and blending of color values [12, 17, 30]. However, the recent trend is the application of GAN to create realistic face images [31, 32]. The performance of MAD approaches to detect standard morphs and morphs produced by GAN are compared in [33, 34]. Several MAD approaches are compared within the framework of the ongoing NIST FRVT MORPH challenge [14].

2.3 Information fusion approaches in face morphing attack detection

Decision-making systems can be fused at four different levels [2]: data level, feature level, classifier level, and combination (or decision) level. The earlier the fusion is applied, the higher are implementation costs (esp. the computation power required), but also the higher accuracy is expected.

A huge number of different fusion approaches exist, ranging from simplistic methods, like the sum-rule (also known as average rule, meaning the linear combination of matching scores with equal weights) or majority voting to complex schemes like Dempster-Shafer Theory (DST) of evidence [35]. Since DST has a theoretical foundation for handling contradicting and missing decisions of expert systems, it has been successfully applied in a wide range of applications [36]. There, exist different ways on how to exactly implement fusion based on DST. For details of our own realization, we refer to section 4.3 accordingly.

For the question which fusion method should be chosen, there exists, to the best of the authors' knowledge, no universally agreed upon theory to answer this question. Some experts put a strong focus on one specific method, e.g., Kittler et al. in [37], where the authors claimed that the sum-rule is not only simple, intuitive, remarkably robust, but also outperforms in their experiments all other aggregation operators tested. Other experts, like Ho [4] and Kuncheva [38], explicitly refrain to give any generalized recommendation. Acknowledging the fact that, even when a critical mass of single classification models has been accumulated in a field of application, there are still open questions regarding their combination and the interpretation of the combination output.

If, within media forensics, the field of image manipulation detection is considered (which also contains MAD as a research question) the same wide range of methods are used in research papers, ranging from the simple to complex. A good example in this domain would be the work of Fontani et al. in [39, 40]. In those papers, the authors apply with DST a very sophisticated approach to image manipulation detection task and additionally use its benefits to counter anti-forensics.

A face morphing attack detector is in its nature a binary pattern classifier. The methods for combining such pattern classifiers have been thoroughly studied for a long time, e.g., in [38]. The paper [7] summarizes the state of the art in information fusion for MAD and extends it by introducing DST to this field. The test results presented do show that the error rates with the DST-based fusion are significantly lower compared to those of individual detectors as well as some simplistic fusion approaches applied previously (majority voting and average rule). Here, the work from [7] is used as basis for this paper, taking its fusion framework and extending it even further by including likelihood-based fusion. The reason to do so is the prominent role that the forensic sciences currently attribute to the usage of likelihood ratios in expert testimony, see, e.g., [41] for the example of footwear marks (and underlying forensic analyses, see, e.g., [42]).

While many scientific publications address applying fusion under lab conditions, only very few publications address the question of generalization as well as the applicability for forensic procedures within the context of criminal investigations. In [43], classical probabilities are replaced by Shafer belief functions and an analogy of the Bayes' rule is introduced that is capable to overcome the traditional inability to distinguish between lack of belief and disbelief. Besides mathematical modeling, the consequences of applying the fusion theory for legal practice are discussed. They conclude that there is still a lot of room for explaining the advantages and limitations of using information fusion to forensic researchers as well as the actual practitioners in criminal investigations. Here, the discussion of the advantages and disadvantages of information fusion is continued and its limitations, if applied in real-life conditions, are empirically demonstrated.

3 The concept of fusion-based face morphing attack detection

In theory, a necessary and sufficient condition for a combination or fusion of classifiers to be more accurate than any of its members is that the individual classifiers are accurate and diverse. An accurate classifier has a classification performance better than random guessing and two diverse classifiers make errors on different data points [44]. In practice, experimental evidence has been provided that, for the case of classifiers with a low level of dependence, a consensual decision is likely to be more accurate than any of individual decisions [45]. It has been also shown that lowering correlation among classifiers increases the accuracy of combination [46].

Application of fusion to MAD approaches and especially of the Dempster Shafer Theory (DST) is initially discussed in [7]. In the experiments performed there, the fusion always outperforms individual classifiers in terms of lower error rates. The evaluation concept from this paper is considered here as a reference. It is expanded and it is demonstrated that under certain conditions the superiority of fusion is not always the case. In particular, it is illustrated why the assumption that fusion would make the detection more reliable can nevertheless fail in practice. This enables a discussion on the constraints and limitations of the application of fusion and reflects upon the impact of generalization power of single classifiers as well as fusion methods and the relationship between training and test data sets. Figure 2 roughly depicts the initial evaluation concept.

The concept consists of five major components:

1. The set D of individual morphing attack detectors. Each individual morphing detector is considered as a black box (i.e., they are used as pre-trained methods implying that we have no influence on the training of the classification model). An input for an individual detector is a face image and an output is a score between 0 and 1. High scores indicate morphs and low scores genuine samples.
2. The set of approaches for establishing weights for individual decisions in the fused one. In the case of DST, the mass (belief) functions are required. The process of deriving such parameters is referred to as training in Fig. 2.

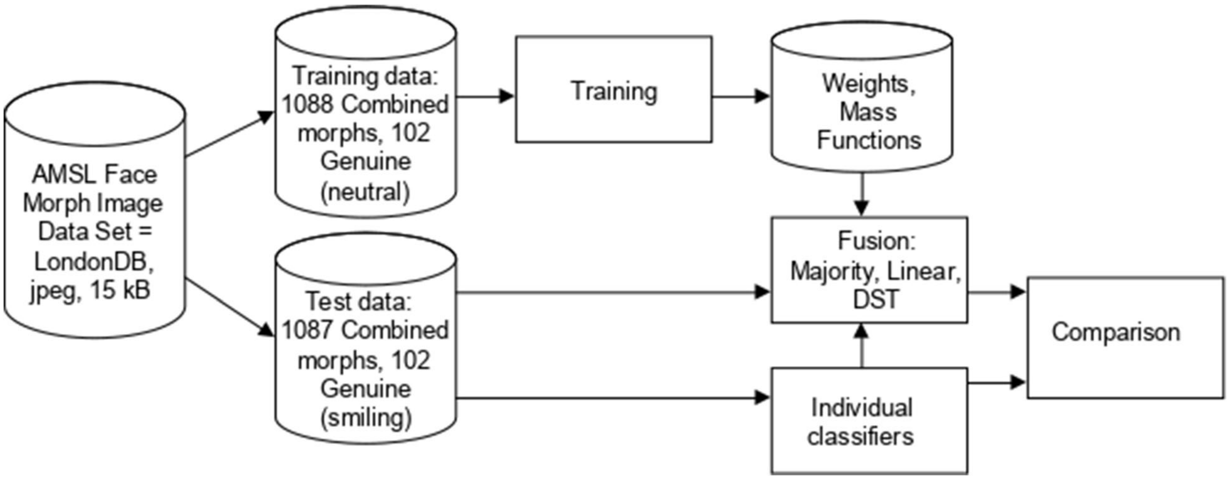

AMSL Face Morph Image Data Set = LondonDB, jpeg, 15 kB
Training data: 1088 Combined morphs, 102 Genuine (neutral)
Test data: 1087 Combined morphs, 102 Genuine (smiling)
Training
Weights, Mass Functions
Fusion: Majority, Linear, DST
Individual classifiers
Comparison

Fig. 2 Evaluation concept for MAD in identity verification from [7]

3. The set of fusion approaches *F*. A fusion approach gets a list of individual decisions and the “importance” of each decision and returns the consensual decision.
4. The evaluation data, which includes training data for establishing fusion parameters (e.g., weights or mass functions) and test data for estimation of error rates. The training and test datasets are created by splitting the AMSL Face Morph Image Data Set (made available via: https://omen.cs.uni-magdeburg.de/disclaimer/index.php). This dataset was initially created to simulate a border control scenario and includes cropped and JPEG-compressed face images which do not exceed 15 kByte and, therefore, fit onto a chip of an eMRTD. In the evaluation, this application scenario is referred to as “MAD in identity verification” (*SC2*). For creating morphed face images, the combined morphing approach from [30] is applied.
5. Comparison of individual detectors and fusion approaches. As a performance metric, we have chosen the error rates of classification approaches.

Here, this concept and its components are re-used and extended by the following: (1) providing a better separation between the training and test datasets by using completely different data sources, (2) adding a fusion approach based on forensic likelihood ratios, (3) adding two types of morphed face images: complete and splicing morphs [12], and (4) adding the application scenario “MAD in document issuing” (*SC1*).

For scientific rigor, it has been ensured in communication with the authors of the MAD approaches that the datasets used for training of the individual detectors do not overlap with the datasets used for training and testing of the fusion approaches.

4 Evaluation setup

Figure 3 depicts the evaluation concept for this paper. The components from [7] and the modifications and extensions summarized in section 3 are apparent in the comparison to Fig. 2.

The representation of the evaluation scenario is done by either using images in their native format and resolution (for application scenario “MAD in document issuing” *SC1*) or in the format specified for ICAO compliant eMRTD (for application scenario “MAD in identity verification” *SC2*). The evaluation scenarios are discussed in more detail in section 4.1. In section 4.2, the used single classifiers for MAD are discussed, while section 4.3 summarizes the fusion methods evaluated (including the strategies for determination of decision thresholds and score normalization). Section 4.4 introduces the performance metrics and 4.5 the databases that are used to create the evaluation data sets.

4.1 Detailed specification of two evaluation scenarios

So far, the evaluation of morphing attack detection (MAD) mechanisms has not been focused on the application scenario. The MAD approaches were rather classified in two groups regarding whether a trustworthy reference face image is presented or not (reference-based vs. single-image/no-reference approaches; see section 2.2). Here, two application scenarios “MAD in document issuing” (*SC1*) and “MAD in identity verification” (*SC2*), representing the two forensic checks required in the document life-cycle of a face image based identity document (see Fig. 1), are considered. Table 1 compares both application scenarios.

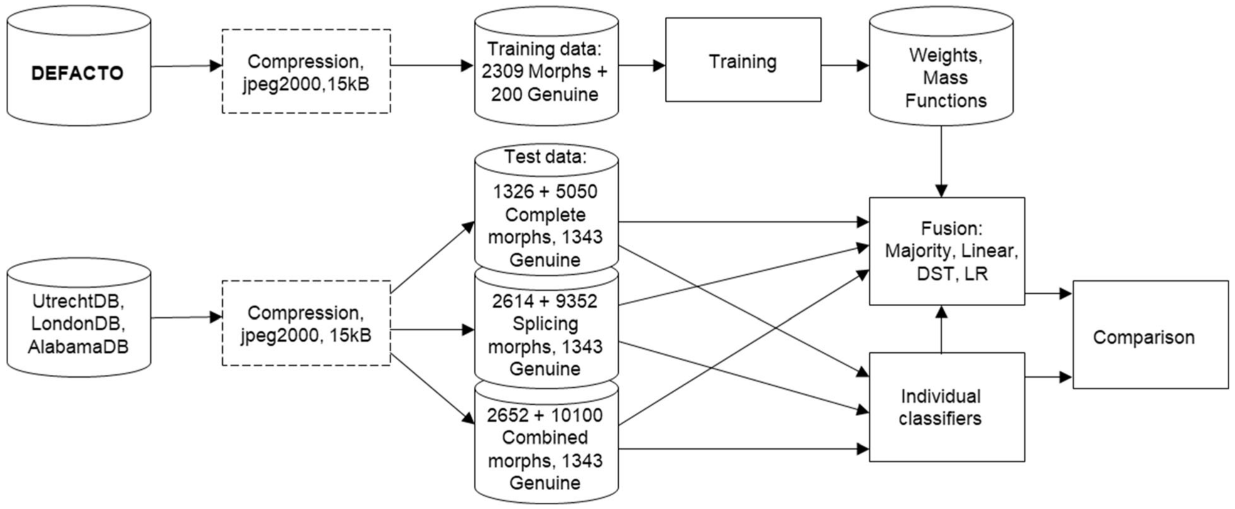

DEFACTO
Compression, jpeg2000,15kB
Training data: 2309 Morphs + 200 Genuine
Training
Weights, Mass Functions
Test data:
1326 + 5050 Complete morphs, 1343 Genuine
Fusion: Majority, Linear, DST, LR
UtrechtDB, LondonDB, AlabamaDB
Compression, jpeg2000, 15kB
2614 + 9352 Splicing morphs, 1343 Genuine
Comparison
Individual classifiers
2652 + 10100 Combined morphs, 1343 Genuine

Fig. 3 Our enhanced evaluation concept—switching from "MAD in identity verification" to "MAD in document issuing" is done by deactivating the dashed boxes

The most intuitive mapping would be to link single-image MAD approaches to *SC1* and reference-based MAD approaches to *SC2*. In fact, both application scenarios can be tuned in the way that the reference image is presented. For *SC2*, taking a "live" face image is an inherent part of the procedure. Note that this image could be used solely for face recognition and ignored by the MAD module. For the document issuing in *SC1*, a webcam could be installed next to the officer at the issuing authority, providing a possibility for capturing "live" face images of an applicant.

No-reference MAD approaches are limited to the search for content-independent statistical anomalies or content-dependent visual artifacts caused by the morphing process. Such methods often apply techniques developed within the context of digital image forensics (see section 2.2). Reference-based MAD algorithms try to reconstruct the morphing process aiming at predicting the face of an "accomplice" and comparing this face to the trustworthy "live" image. Hence, the presence of a reference face image rather gives additional options for the choice of detection mechanisms, but does not determine the application scenario.

In contrast, the face image format in *SC2* is very closely defined by national and international regulations, especially by the International Civil Aviation Organization (ICAO) standardization of eMRTD. As a result, the limitations to the digital image that should be stored in an eMRTD are caused by antiquated physical storage limitations. For instance, the current generation of German (and other countries) passports limits the free space for a digital face to 15 kB. During the application for a new document, an applicant submits a printed face photograph of the size of 35 × 45 mm. These images are scanned with the resolution of 300 dpi and undergo lossy compression before they are stored in the passport. The submission of printed face images is in fact the main vulnerability spot making the face morphing attack easy to execute. The reason is that the printing process destroys almost all traces of image manipulation so that human examiners are highly prone to

Table 1 Comparison of the document issuing (*SC1*) and identity verification (*SC2*) scenarios

	Document issuing (*SC1*)	Identity verification (*SC2*)
Attack's target	Officer at the document issuing authority	Identity verification system
Time constraints	Up to several minutes	Few seconds (< 2 s)
Face image format	- Low-size printed document image - High-resolution digital image from a certified photo-kiosk	- Low-resolution compressed digital document image - Low-size re-printed document image partially occluded by watermarks
Currently used morphing detection mechanisms	- Naked eye, comparison to the person in front of the desk	- No explicit mechanisms - AFR systems may be set to rejecting at low similarity
Proposed morphing detection mechanisms	- Primarily non-reference (blind) detection - Could be extended by reference-based detection	- Reference-based detection - Demorphing - Could be extended by non-reference (blind) detection

errors when categorizing such images [12]. The straightforward way to reduce the danger of the morphing attack is a prescription to submit high-resolution digital face photographs of a decent quality. Having done this, the image resolution would not be an issue any more for at least a document issuing scenario. As described in section 2.2, taking the picture directly in the controlled environment of the issuing office would limit the threat by morphing attacks. This is not only a political issue but would also require the elimination of further attack vectors.

The file format used in this paper to implement *SC2* is a face image compliant with ICAO specifications for eMRTD: 531 × 413 pixels (inter-eye distance of at least 120 pixels), in JPEG2000 format, compressed to fit the 15 kB size constraint. The file format to implement *SC1* is not that narrowly defined; here, the original file format of the reference databases (see section 4.5) is used.

4.2 Morph attack detection approaches

In this paper, five morph attack detection (MAD) approaches are examined. The first one ($D_{keypoints}$) is based on localization and counting of keypoints [19]. The keypoint-based morphing detector indirectly quantifies the blending effect as an indispensable part of the morphing process. Blending leads to a reduction of face details and therefore to a reduction of “significant corners” and edge pixels. The detector counts the relative number of keypoints in the face region detected by different approaches as well as the relative number of edge pixels. For classification within $D_{keypoints}$, a linear support vector machine (SVM) was trained based on 24-dimensional feature vectors with a dataset of 2000 genuine and 2000 morphed high-resolution passport images. These morphs were created using the approaches from [12, 30].

The other four MAD approaches are based on Deep Convolutional Neural Networks (DCNN). Two of them designated as $D_{ArXivNaive}$ and $D_{ArXivMC}$ are described in [26]. The other two designated as $D_{BIOSIGNaive}$ and $D_{BIOSIGMC}$ are described in [17]. All four of these detectors are based on the VGG19 network. Transfer learning is applied to build a binary classifier from the classification model originally trained for the ILSVRC challenge. The training dataset is comprised of approximately 2000 genuine images and the same number of morphs. Genuine images were collected from several public face databases and scraped from the internet. The major difference between classifiers is in the approach for generation of morphed face images for training. While the $D_{ArXivNaive}$ is an older detector trained with lower quality morphs and $D_{ArXivMC}$ is the same detector with an updated data augmentation strategy in the training, the $D_{BIOSIGNaive}$ and $D_{BIOSIGMC}$ detectors applied for the creation of the training data sophisticated morphing with artificially added high-frequencies to compensate the blurring effect of the blending operation. The differences between the *Naive* training and the *MC* (multiclass/complex morphs) versions lie in the composition of the training data: For *Naive* 50% genuine images and 50% complete morphs are used. For *MC* 50% genuine images and a mix of complete and partial morphs are used, with the aim of forcing the network to take all available information for its decision-making into account (i.e., prevent it from focus on selected face regions like the eyes to detect morphing attacks). The details on the training concept for *Naive* and *MC* versions of the detectors used here can be found in [17].

4.3 Fusion approaches

Here, each MAD approach operates as a “black box” returning a matching score for an input sample. As a consequence of the evaluation concept, fusion on signal level is out of scope for this paper and fusion on feature level (see section 2.3) is not feasible. Hence, the detection accuracy gain from one fusion approach at the decision level (majority voting) and three fusion approaches at the matching score level (weighted linear combination, Dempster-Shafer Theory (DST) of evidence, and forensic likelihood ratios (LR)) is explored. Below, the fusion operators F are described in detail:

4.3.1 Majority voting (F_M)

The naive consensus pattern of simple majority [38] is used for opinion combination. If the number of votes for every alternative is equal, the majority rule returns “no decision.”

4.3.2 Weighted linear combination (F_{WLC})

The sum-rule (or weighted linear combination) extends the average rule by assigning different weights to the output of the individual classifiers to be combined. For the case of the same weights, the fusion strategy is often referred to as average rule. Here, two different strategies are used: average rule as well as weighted linear combination with pre-determined weights (see section 5.1 for details on these two strategies).

4.3.3 Fusion based on Depster-Shafer Theory (F_{DST})

The Depster-Shafer Theory (DST) is based on two concepts: belief functions representing degrees of belief for one question from subjective probabilities for a related question and Dempster’s rule for combining such degrees of belief when they are based on independent items of evidence.

In our case, the frame of discernment is defined as $\Theta = \{\mathrm{mor}, \mathrm{gen}\}$, with m(mor)/m(gen) representing the basic beliefs that the face is morphed/genuine respectively, and m(Θ) is a mass of uncertainty. A degree of belief (mass) is assigned to each subset. As proposed in [7], we construct mass functions as cumulative distribution functions of matching scores obtained from an experiment. Let $\mathrm{p}_{\mathrm{mor}}(s)$ and $\mathrm{p}_{\mathrm{gen}}(s)$ be the approximations of probability density functions of scores for verification attempts with morphed and genuine images respectively. For a detector outcome s* ranging from 0 to 1, we define the mass m(mor) as an area under $p_{mor}(s)$ between 0 and s* and m(gen) as an area under $p_{gen}(s)$ between s* and 1, and the mass of uncertainty as a complement to the sum of both masses:

$$m(mor) = \int_{s=0}^{s^*} p_{mor}(s)ds, m(gen) = \int_{s=s^*}^{1} p_{gen}(s)ds \quad (1)$$

$$m(\Theta) = 1-(m(mor) + m(gen)) \quad (2)$$

Note that we interpret the detector outcome s* (also called matching score) as a decision confidence with 1 for 100% confidence that the image is morphed and 0 for 100% confidence that the image is genuine.

Technically, the three masses are calculated for each morphing detector based on the matching scores of training samples and stored as a parameter of our fusion engine. At the time of decision-making, for each outcome s_i* of the i_{th} detector, we obtain the values m_i(mor), m_i(gen), and m_i(Θ) as the nearest points on the corresponding discrete mass curves.

Dempster's rule of combination for two beliefs from independent sources is given by:

$$m(A{\neq}O) = \frac{1}{K} \sum_{A=A_1 \cap A_2} (m_1(A_1) \cdot m_2(A_2)) \quad (3)$$

$$K = 1- \sum_{A_1 \cap A_2 = 0} (m_1(A_1) \cdot m_2(A_2)) \quad (4)$$

where m(A) represents the combined mass on A (a given member of the power set), m_1 and m_2 represent the masses of first and second items of evidence respectively, and K represents the normalization constant. The second term in K describes the conflict between two items of evidence. If it is equal to 1 then K is equal to 0 implying that these two items contradict each other and cannot be combined by applying Dempster's rule.

The efficient application of the Dempster's rule for computation of combined belief can be found in [6]:

$$m(mor) = 1-\frac{1}{K}\prod_{i=1}^{n}(1-m_i(mor)) \quad (5)$$

$$m(gen) = 1-\frac{1}{K}\prod_{i=1}^{n}(1-m_i(gen)) \quad (6)$$

$$m(\Theta) = \frac{1}{K}\prod_{i=1}^{n} m_i(\Theta) \quad (7)$$

$$K = \prod_{i=1}^{n}(1-m_i(mor)) + \prod_{i=1}^{n}(1-m_i(gen))-\prod_{i=1}^{n} m_i(\Theta) \quad (8)$$

4.3.4 Fusion using likelihood ratios (F_{LR})

Likelihood ratios (LR) are used in forensics in order to express uncertainty [47]. The basic concept relies on the quotient of the probabilities of the correctness of two hypotheses with respect to an observation within binary decisions which are common in forensics. Semantically, the LR describe how much more probable one of the hypotheses is in comparison to a complementary one when specific observations can be made.

Within the scope of a forensic comparison of face images, LR are discussed, e.g., in [42] and is already used in some countries in the forensic practice as well, as shown, e.g., in [41] for a case involving footwear marks in the UK. Sometimes the observed LR are mapped to particular levels regarding the confidence in the hypothesis in order to make the result more accessible to forensic laymen as the requirements for particular LR differ between forensic domains, see, e.g., [48]. Generally, a likelihood ratio close to 1 indicates a weak decision as the probabilities for the two hypotheses are almost identical.

With the availability of multiple detection algorithms, a fusion using LR is also possible as suggested, e.g., in [49] for multiple biometric matchers. For each detection algorithm, a quality value needs to be determined as a weight in the fusion algorithm.

In our experiments, the LR for a single detector *D* providing confidence levels *c* in a two-class problem is determined by the quotient of the detectors confidence for a sample *s* toward a genuine sample—c_D(gen)—divided by the confidence toward a morphed sample—c_D(mor):

$$LR(s, D) = \frac{c_D(gen)}{c_D(mor)} \quad (9)$$

Note that the inverse of the LRs is used in the experiments performed here, in order to achieve a defined value of zero for a confident decision. Usually the tested hypothesis—in this case whether an image is a morph—would be used as the numerator. As a result, the F_{LR} shows the same behavior. In addition to that, it is possible to normalize F_{LR} using the number of detectors (in this paper 5). Otherwise, this number would have to be taken into account during the interpretation of fusion operator.

The LR-based fusion score F_{LR} of a sample image in question for the k = 5 detectors D = {$D_{keypoints}$,

$D_{ArXivNaive}$, $D_{arXivMC}$, $D_{BIOSIGNaive}$, $D_{BIOSIGMC}\}$ is determined as the quotient of weighted sum of LRs toward a genuine sample (LR_g) divided by the LRs toward a morph (LR_m) with $LR_g(s,D) = \frac{1}{LR_m(s,D)} = \frac{c_D(mor)}{c_D(gen)}$:

$$F_{LR}(s) = \frac{\sum_{i=1}^{k} LR_g(s,D_i) * w_i}{\sum_{j=1}^{k} LR_m(s,D_j) * w_j} \tag{10}$$

The factor w_i/w_j represents here the weighting factor for the LR fusion as described in section 5.1. A quotient $F_{LR}(s)$ closer to zero indicates a larger confidence of the decision toward a morph.

4.3.5 Normalization

In order to perform a reasonable fusion, the matching scores of the individual classifiers should be brought into the same range. The detectors $D_{ArXivNaive}$, $D_{arXivMC}$, $D_{BIOSIGNaive}$, and $D_{BIOSIGMC}$ return negative values for genuine faces and positive values for the morphed faces. The default decision threshold is 0. In contrast, the detector $D_{keypoints}$ returns values between 0 and 1. Lower values are for genuine faces and higher values for morphed faces. The default decision threshold is 0.5. Within the training phase performed in this paper using the DEFACTO dataset (see section 4.5), we perform min-max normalization of the matching scores and adapt the default decision thresholds. As a result, the normalized matching scores of all detectors range then from 0 to 1 and the new default decision threshold can be found in Table 3 (column τ_{fixed}). For each classifier, the MIN and MAX values of matching scores are stored to perform the min-max score normalization at the evaluation phase. The aforementioned decision thresholds are also stored as parameters of the fusion and are used in the evaluations in *SC1* and *SC2*.

4.4 Performance metrics

Morphing detection is a standard two class problem with two possible outcomes: “passport image is morphed” or “passport image is not morphed” and two types of errors: morphed image is recognized as non-morphed and vice versa. Driven by the idea that the morphing attack can be seen as a special case of the presentation attack, the detection performance metrics from the presentation attack detection testing standard [50] are adopted. Attack Presentation Classification Error Rate (APCER) describes the proportion of morphed face images incorrectly classified as genuine (bona fide) and Bona Fide Classification Error Rate (BPCER) describes the proportion of genuine (bona fide) face images incorrectly classified as morphed. MAD approaches are typically designed to report two values: a binary decision on whether the image is morphed or not and a confidence score for this decision from the interval [0; 1]. Higher values indicate higher confidence that the image is morphed. In fact, the binary decision is derived from the confidence score by comparing it to an algorithm-dependent predefined decision threshold. Hence, APCER and BPCER are the reciprocal functions of decision threshold. Formally, the BPCER is computed as the proportion of bona fide images over the threshold and the APCER as the proportion of morphed images below the threshold. At the stage of development, when an algorithm can be evaluated with different decision thresholds, the more informative way to compare algorithms is drawing the detection error trade-off (DET) curves (respectively the area under curve (AUC)) on the same plot. Traditionally, BPCER is seen as a convenience measure while APCER as a security measure. The DET curve represents BPCER as a function of APCE R. Here, also the half total error rate (HTER) is used as an average of BPCER and APCER with the fixed decision threshold to compare performances in an easier way.

4.5 Evaluation datasets

There are four databases used in the experiments in this paper: The DEFACTO database [51] containing morphs and genuine face images is used for the training of the fusion methods (see Fig. 2). This database is chosen as a neutral dataset for training because it ensured by the authors that it was not used in the creation (i.e., training) of any of the five used “black box” individual detectors and its used morphing method being unknown. By this choice, a realistic evaluation setup can be ensured, with training data (DEFACTO material) having an unknown similarity to test data (for *SC1* and *SC2*; see Fig. 1), reflecting the constraints that will be encountered in field application. The following datasets (and subsets) are used:

- The DEFACTO dataset contains 200 genuine face images and 39980 morphs. Since using the whole dataset would represent an extremely strong bias toward morphs, only a subset of 2309 randomly selected morphed images is used.
- Three other databases are used to simulate the evaluations conducted within the comparison between single classifiers and fusion methods performances:
  - For two of them (the ECVP (aka Utrecht) [20] and London Set [52] databases) morphed images are generated using the approaches from [12, 30]. The subsets of morphed images are

denoted as *complete*, *splicing*, and *combined* according to the generation method used.

- Additionally, as a source for further genuine face images, mugshots from the Alabama News Network [53] are taken.

Using the original sized images (and morphs based on those), the experiments simulate the passport issuing scenario (*SC1*). In order to simulate the verification scenario (*SC2*), the images are down-scaled (to 413 × 531 pixels) and compressed using the JPEG2000 format in a way that the image size does not exceed 15 kilobyte (kB) as described in section 4.1. Figure 3 shows the exact evaluation concept and Table 2 summarizes the information about the image (sub-)sets used in our experiments.

5 Evaluation results and discussion

This chapter contains a large number of results from different empirical evaluations as well as their interpretation. It is structured as follows:

- Section 5.1 summarizes the DEFACTO experiments, which serve as a baseline as well as an estimator for fusion weights (or mass functions).
- Section 5.2 evaluates the individual detectors and fusion methods (using the full ensemble of detectors) for the two simulated application scenarios *SC1* and *SC2*.
- Section 5.3 discusses the impact of the performed fusion to the field of MAD.
- Section 5.4 determines the impact of using smaller ensembles (i.e., subsets of the available detectors) for fusion.
- Section 5.5 determines the impact of less restrictive assumptions in the evaluation setup composition on the error rates achieved in fusion.
- Section 5.6 provides a final summary and generalization on the obtained results.

5.1 DEFACTO training and baseline experiments

The experiments with the DEFACTO dataset have two objectives:

1. Fair comparison of the MAD approaches to each other regarding their error rates with a disjunctive dataset. In fact, face images in the DEFACTO dataset do not overlap with those used for the training of MAD approaches. Moreover, the morphing procedure with the DEFACTO significantly differs from those with the individual MAD approaches.

2. Training of the fusion parameters including fusion weights and decision thresholds of the individual MAD approaches as well as mass curves for the DST-based fusion. An importance (or in other words a credibility) of one or another detector in the fusion is given by the fusion weight. Here, we consider two thresholding strategies "*fixed*" and "*adaptive*" to define at the same time the decision thresholds and weights (the latter only for F_{WLC} and F_{LR}):

For the "fixed" strategy, we rely on the default decision thresholds suggested by the developers of the MAD approaches and assign equal weights for fusion approaches that accept weights. This trivial strategy (which considers all available detectors as being equally important) is typically the only choice if no additional evaluation of classifiers can be performed, or if there is a suspicion that the evaluation dataset does not fit to the in-field data.

For the "adaptive" strategy, we set a new decision threshold at the point at which the EER of a MAD approach is reached. Additionally, we calculate the fusion weights for F_{WLC} and F_{LR} based on the EER values. To be more precise, the inverse of the EER values are used as weights of the individual MAD approaches in the fusion. Since the possible EER values for a binary

Table 2 Evaluation data sets

Database	Number of images	*SC1* (document issuing)	*SC2* (identity verification)
DEFACTO morphs	2309	tiff, 500 × 652	15kB, jpeg2000, 413 × 531
DEFACTO genuine	200	jpg, 500 × 652	15kB, jpeg2000, 413 × 531
ECVP complete	1326	png, 900 × 1200	15kB, jpeg2000, 413 × 531
London complete	5050	png, 1350 × 1350	15kB, jpeg2000, 413 × 531
ECVP splicing	2614	png, 900 × 1200	15kB, jpeg2000, 413 × 531
London splicing	9352	png, 1350 × 1350	15kB, jpeg2000, 413 × 531
EVCP combined	2652	png, 900 × 1200	15kB, jpeg2000, 413 × 531
Alabama genuine	1343	jpg, image resolution varies	15kB, jpeg2000, 413 × 531

classifier range from 0 (for a perfect classifier) to 0.5 (for a random guess) and the weight should spread over the interval [0, 1], an EER value is multiplied by 2, see Equation (11).

$$w_i = \max(0, 1\text{–}2 \cdot EER_i) \tag{11}$$

with i representing one of the five MAD approaches.

Figure 4 shows the DET curves of the five addressed MAD approaches on the original-sized DEFACTO images. Crossings with the dashed black line represent the EER of the detectors. Regarding the EER, three detectors $D_{ArXivNaive}$, $D_{BIOSIGMC}$, and $D_{BIOSIGNaive}$ demonstrate comparable performances, with $D_{BIOSIGMC}$ achieving the best performance by a small fraction. The $D_{ArXivMC}$ demonstrates slightly worse performance and the $D_{keypoints}$ is by far the worst detector.

Table 3 demonstrates the EER values of the individual MAD approaches, the decision thresholds τ at which the EER are reached, and the weights assigned to the approaches for fusion for both strategies "fixed" and "adaptive." If the fusion is done at the decision level, the decision thresholds are used to derive decisions from matching scores.

The mass functions for the DST fusion are demonstrated in Fig. 5. The mass curves for the "genuine" and "morphed" matching scores reproduce the classic error curves so that the crossing point indicates the EER.

What can be observed from the results in Table 3 is that $D_{BIOSIGMC}$ outperforms the other four detectors by presenting the smallest EER (resp. the highest AUC). As a result, it is assigned the highest weight for the fusion operations. The results for $D_{keypoints}$ confirm what was already indicated in Fig. 4: Despite its good performance on other image sets, this detector is here performing significantly worse than the other four. As a result, it gets with 0.42 the lowest weight assigned for the fusion.

If the EER locations (the projection of the EER onto the x-axis) and the uncertainty curves shown in Fig. 5 are analyzed, it can be seen that four of the five curves (resp. EER locations) are shifted from the center to the left (indicating a bias toward morphed images) and only $D_{keypoints}$ is shifted to the right with a strong bias toward genuine images. The amount of the shift correlates with the ranking of the detectors: $D_{BIOSIGMC}$ shows the smallest shift (a nearly centered uncertainty curve with a very small skew) while the other four show an increase in the shift (and skew) with their higher EER.

5.2 Experiments with individual detectors and fusion methods

The sections 5.2.1 and 5.2.2 summarize the results on the performance of the individual detectors and fusion methods evaluated with the two simulated application

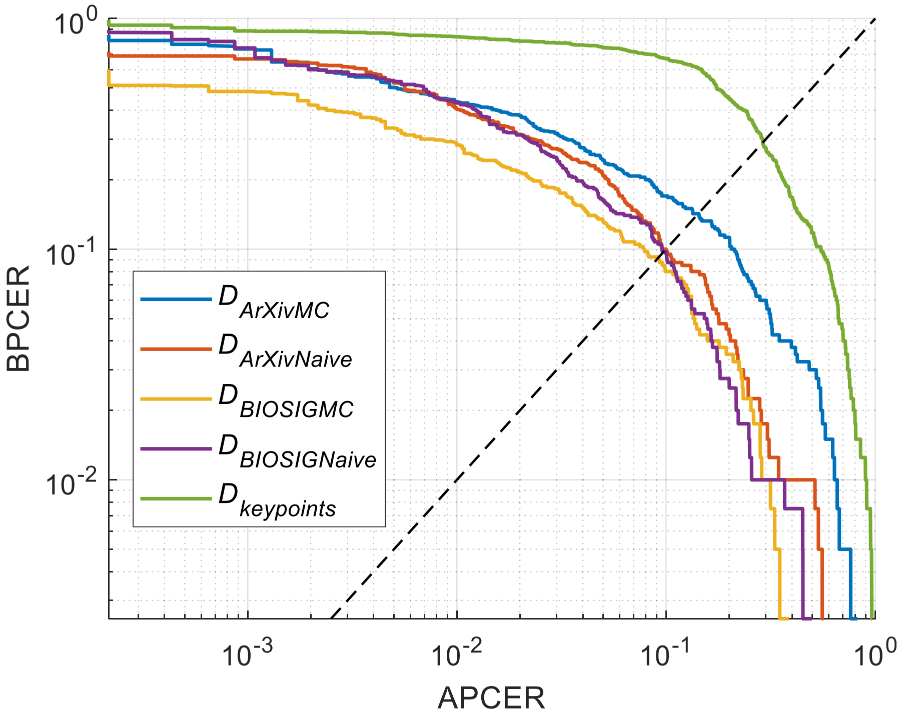


Fig. 4 DET curves of the individual detectors with the DEFACTO dataset (original-sized images)

Table 3 Evaluation of detectors with the DEFACTO dataset and associated weights

Detector	AUC	EER	$\tau_{adaptive}$	$w_{adaptive}$	τ_{fixed}	w_{fixed}
$D_{ArXivMC}$	0.94	0.14	0.35	0.72	0.47	1.00
$D_{ArXivNaive}$	0.97	0.10	0.40	0.80	0.59	1.00
$D_{BIOSIGMC}$	0.98	0.09	0.48	0.82	0.53	1.00
$D_{BIOSIGNaive}$	0.97	0.10	0.36	0.81	0.52	1.00
$D_{keypoints}$	0.77	0.29	0.87	0.42	0.50	1.00

scenarios *SC1* and *SC2*. All these tests use as data the combined images from the ECVP, London and Alabama datasets (see section 4.5). For *SC1* the original-sized images are used and for *SC2* the 15 kB versions.

5.2.1 Scenario SC1 ("MAD in document issuing")

Figure 6 shows the DET curves for the tests on complete, splicing, and combined morphs in *SC1*. The individual classifier performance is displayed by solid lines (with the same color coding as in Fig. 4), and the performance of the fusion methods is given as dashed lines (where a continuous space of operation points is possible) or symbols (in case only one operation point, either the "fixed" setting or the "adaptive," is possible).

For all three morphing types, the individual classifier $D_{arXivNaive}$ achieves the best performance for *SC1*, followed by the weighted linear combination (F_{WLC}). The three single classifiers $D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, and $D_{keypoints}$ show the lowest performance. F_M with "fixed" and "adaptive" thresholding strategy achieve the lowest performance of the fusion methods. The more sophisticated fusion operators (F_{DST} and F_{LR}) perform better than F_M, in some cases F_{DST} even outperforms F_{WLC}, but both show a significant bias toward morphed images. Especially for F_{DST}, this is apparent with an APCER close to 0 at a BPCER of roughly 0.2.

5.2.2 Scenario SC2 ("MAD in identity verification")

Figure 7 shows the DET curves for the tests on complete, splicing, and combined morphs in *SC2*. The same color coding and symbols are used as in Figs. 4 and 6.

The general performances of the individual and fusion based detectors in *SC2* are very similar to the *SC1* results shown in Fig. 6. A slight decrease in the detection performances can be observed for all tested methods. This decrease can be attributed to the fact that the 15 kB image format that is used in *SC2* leaves generally less room for media forensic investigations on image manipulation. What is remarkable in the results is that the results of the more sophisticated fusion operators (F_{DST} and F_{LR}), while also showing some performance decrease, loose some of their bias toward morphed images. Especially for the splicing morphs, it can be observed in Fig. 7 that F_{DST} shows an APCER larger than 0, even slightly outperforming at the corresponding APCER values all other detectors.

5.3 Discussion of the impact of fusion to face morphing attack detection

Tables 4, 5, and 6 summarize the results. Table 4 demonstrates a baseline using only the individual classifiers,

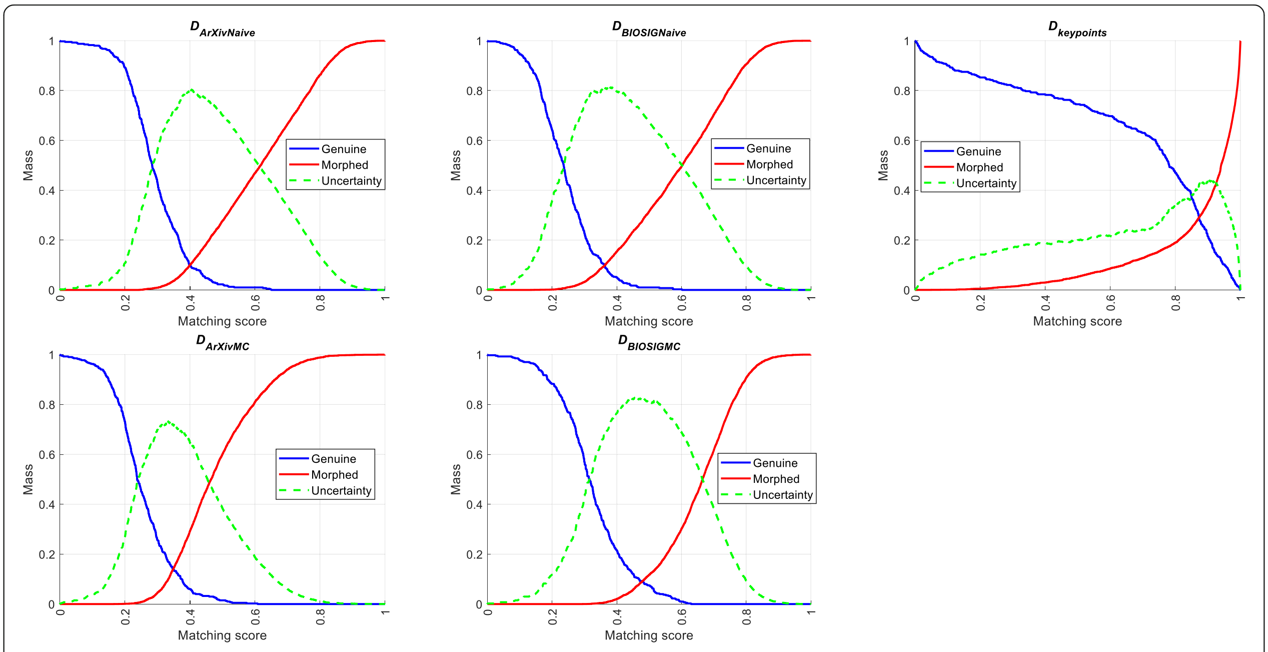


Fig. 5 Mass functions for the DST fusion resulting from the evaluation with the DEFACTO dataset (original-sized images)

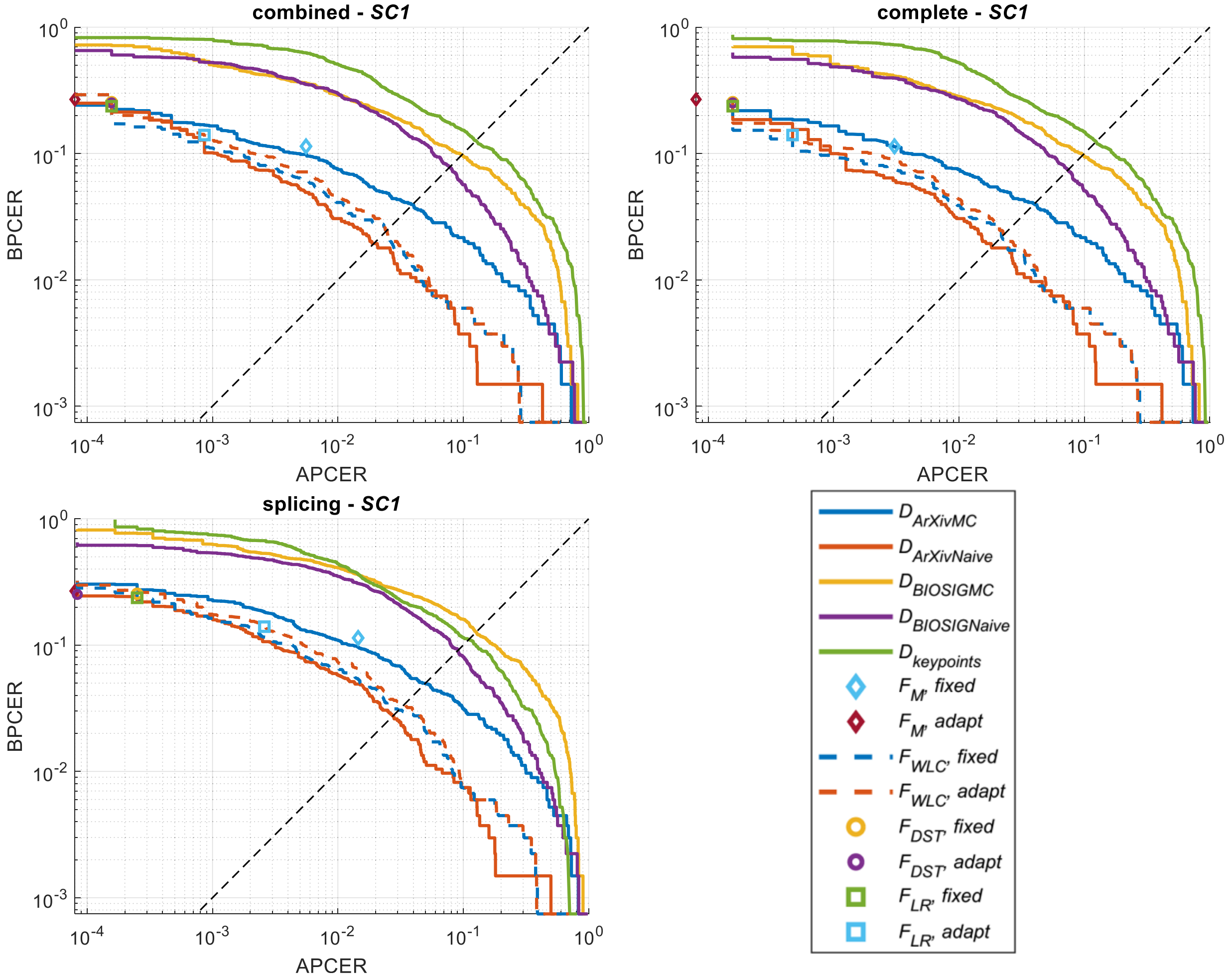


Fig. 6 DET curves for the tests on complete, splicing, and combined morphs in *SC1*

showing that $D_{ArXivNaive}$ performs best in testing in both application scenarios *SC1* and *SC2* on all three morph types.

Tables 5 and 6 present the single classifier and fusion results in the "fixed" (Table 5) and "adaptive" (Table 6) thresholding strategies. The difference lies in the basic assumption for the similarity of training data (here DEFACTO) and the material encountered in field application (here, the mix of ECVP, London, and Alabama material, either in original (for *SC1*) or the 15 kB version (*SC2*)). While the "adaptive" setting is the setting encountered in most lab experiments, the "fixed" one (which assumes a much lower similarity between training and test data) is a more realistic assumption, leading to more trustworthy error estimates in this media forensic analysis.

When focussing on the single classifier results obtained for both thresholding strategies ("fixed" decision threshold and fusion weights vs. "adaptive" decision threshold and fusion weights), it can be seen that $D_{BIOSIGMC}$, which performed best on the DEFACTO dataset (see Fig. 4 in section 5.1) demonstrates in the evaluations significantly worse performance in both application scenarios *SC1* and *SC2*. In Fig. 4, in two of the six tests (the two evaluations run on splicing morphs), it actually shows the lowest performance (i.e., highest HTER). When looking at Tables 5 and 6, these results are confirmed. For both thresholding strategies and all three different morphing types, $D_{BIOSIGMC}$ achieves the second lowest detection performances, followed only by $D_{keypoints}$. The best performance for a single classifier is in all cases achieved by $D_{arXivNaive}$ with the "fixed" decision threshold.

When comparing the single classifier and fusion results in Tables 5 and 6, the general picture established in section 5.2 is confirmed: In nearly all cases for *SC1* as well as *SC2*, the fusion approaches fail to outperform the best individual detector. Neither for selected morphing

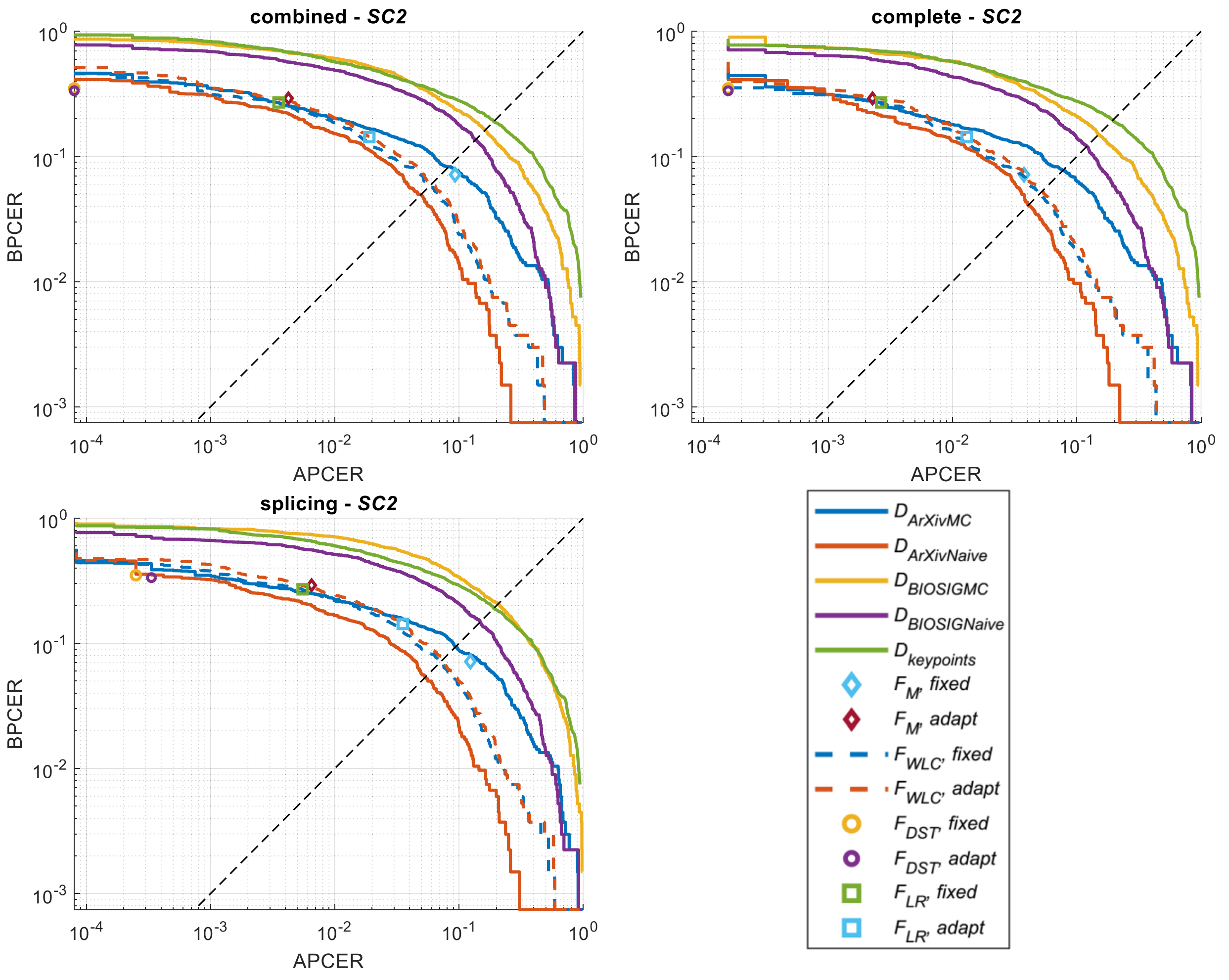


Fig. 7 DET curves for the tests on *complete*, *splicing*, and *combined* morphs in *SC2*

approaches nor for one of the two thresholding strategies, the fusion generally outperforms the best single classifier, even though in one case for *SC2* and splicing morphs it is close (best single is $D_{arXivNaive}$ with “fixed” at an HTER of 8.5% and the best fusion is F_{LR} with “adaptive” and an HTER of 8.92%). Most interestingly, the DST-based fusion, which is the most sophisticated fusion strategy and which is highly regarded in many other application fields, leads here in all cases to low performances.

For the thresholding strategies, it can be summarized that for the four classifiers $D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, $D_{arXivNaive}$, and $D_{arXivMC}$, there is a tendency that the best results are obtained with the “fixed” decision threshold while for $D_{keypoints}$ in the majority of the cases better results are obtained with the adaptive decision threshold. For the fusion, no clear tendency which thresholding strategy leads to better results can be observed.

When considering the differences in the detection performance for the three tested morph types (*combined*, *complete*, and *splicing*), it can be summarized that all detection approached discussed here yield very similar detection performances (both in *SC1* as well as *SC2*).

5.4 Variation of the fusion ensemble

During the review phase for this journal paper, the reviewers raised the question why it is assumed that a fusion using all five single classifiers is the optimal choice at hand. Alternative fusion ensembles using three or four classifiers might be capable to outperform the whole set of five, especially when removing the weakest candidate ($D_{keypoints}$). To address this issue, Table 7 compares the results of three different sets of fusion ensembles for the “fixed” decision thresholds. The results shown are for the complete set of 5 detectors as baseline, the best performing ensemble of 4 (here $D_{BIOSIGNaive}$, $D_{BIOSIGMC}$,

Table 4 Theoretical performance of the individual detectors with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type marked in bold)

		SC1		*SC2*	
Detector	**Morph type**	**EER**	**$\tau_{adaptive}$**	**EER**	**$\tau_{adaptive}$**
$D_{ArXivMC}$	Combined	3.95%	0.528241	8.27%	0.417467
$D_{ArXivNaive}$		**1.94%**	0.594687	**4.96%**	0.499938
$D_{BIOSIGMC}$		9.75%	0.617098	16.31%	0.561516
$D_{BIOSIGNaive}$		7.74%	0.558175	13.56%	0.478364
$D_{keypoints}$		12.65%	0.971509	19.08%	0.990942
$D_{ArXivMC}$	Complete	4.00%	0.526729	7.75%	0.424468
$D_{ArXivNaive}$		**1.82%**	0.600357	**4.06%**	0.507648
$D_{BIOSIGMC}$		9.75%	0.616997	14.98%	0.565297
$D_{BIOSIGNaive}$		7.45%	0.563476	12.07%	0.496052
keypoints		12.43%	0.972011	19.53%	0.990758
$D_{ArXivMC}$	Splicing	4.99%	0.501098	9.37%	0.406828
$D_{ArXivNaive}$		**2.76%**	0.566983	**5.37%**	0.492199
$D_{BIOSIGMC}$		12.67%	0.594189	20.64%	0.541497
$D_{BIOSIGNaive}$		9.08%	0.54235	14.84%	0.470685
$D_{keypoints}$		11.09%	0.976876	19.08%	0.990933

$D_{arXivNaive}$, and $D_{arXivMC}$; the evaluations performed in this case were a complete leave one out sequence but only the most relevant result is presented here) and the ensemble of three with the most disparate characteristics ($D_{arXivNaive}$, $D_{BIOSIGMC}$, $D_{keypoints}$; i.e., selection by limiting redundancy). The results show an apparent decrease of the HTER for *SC1* and *SC2* if switching from an ensemble of 5 (denoted as “5 det” in Table 7) to an ensemble of (the most suitable) 4 detectors (denoted as “4 det” in Table 7). When compared to the single detector performance reported in Table 5 above, it can be seen that the best ensemble of 4 also seems to outperform the individual detectors. Some of the figures presented have to be considered very carefully since they are hiding a problem in the scheme: This is absolutely no problem for cases where the individual weighting makes deadlocks neigh to impossible (e.g., in case of the F_{WLC}) but is especially relevant for the majority vote where significant numbers of “undecided” events occurred (e.g., cases where 2 detectors predicted one class and the other 2 the other) that are not reported in the table. These “undecided” events amount over the various tested ensembles to up to 10% of all majority vote cases.

In case of the chosen ensemble of 3 detectors (denoted as “3 det” in Table 7) all HTER values increased significantly, showing that this ensemble (which more strongly relies on the opinion of the rather weak $D_{keypoints}$) is outperformed by the bigger ensembles.

Similar to Table 7, Table 8 performs the same ensemble tests for the “adaptive” thresholding strategy. Here, the results also show better results for the best ensemble of 4 detectors when compared to the complete ensemble of 5. In contrast to the “fixed” thresholding strategy discussed above, the performance increase obtained by leaving $D_{keypoints}$ out seems smaller but also the number of “undecided” events is way smaller (less than 3%) so that here the gain has to be considered higher. This performance gain is also evident in the comparison to the single detector results discussed in Table 6.

Like in the case of the “fixed” thresholding strategy, the tested cases of 3 detector ensembles showed significantly worse results, increasing the HTER to 18% or even higher.

Summarizing the results on these detector ensemble selection experiments, it has to be said that the best performing set of 4 detectors outperformed for both thresholding strategies (“fixed” and “adaptive”) and *SC1* as well as *SC2* the complete ensemble of 5. For fusion methods that are prone to deadlock or “undecided” situations (esp. the majority vote), the even number of detectors in this cased caused a small issue, generating in the worst case up to 10% deadlock results that would have to be handled in application. All results for the chosen ensemble of the 3 most dissimilar detectors proved near fatal for the system performance since the HTER was significantly increased in all these cases.

5.5 Discussion on alternative evaluation setups

Another issue, raised during the review phase for this journal, is the choice of a realistic but rather challenging experimental scenario where the dataset used for training is disjoint from the ones used for testing. The question was how an overlap between training and testing set (i.e., more favorable conditions for the individual detectors) would influence the outcome of the experiments. To address this question, two different sets of less realistic experimental setups are discussed below: first, a tenfold stratified cross-validation with disjoint sets of genuine samples and morphs, and second an even less realistic (i.e., more lab-condition) test with a static percentage split on one a set containing genuine and morphs that are derived directly from these genuine images.

For the first of these alternative setups, additional tests are performed here to show how a deviation from rigorous evaluation routines reflects in the error rates obtained. Table 9 summarizes the results for the “fixed” as well as the “adaptive” thresholding strategy. If comparing the results in Table 9 to the results in Tables 4 and 5, then the single detector performances in the “fixed” thresholding remain nearly unchanged while the HTER values in case of the fusions decrease (e.g., from 11.85% to 2.6% in case of F_{LR} in *SC1* for *combined* morphs of from 13.70% to 5.9% in case of F_{LR} in *SC2* for *combined*

Table 5 Realistic performance of the individual detectors and fusion approaches with the fixed decision thresholds and equal fusion weights with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type marked in bold)

		SC1			*SC2*		
Detector	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
$D_{ArXivMC}$	Combined	7.45%	1.00%	4.22%	3.87%	18.56%	11.22%
$D_{ArXivNaive}$		2.01%	1.82%	**1.91%**	0.97%	13.32%	**7.14%**
$D_{BIOSIGMC}$		25.76%	1.35%	13.56%	23.10%	10.12%	16.61%
$D_{BIOSIGNaive}$		11.47%	5.25%	8.36%	9.54%	18.05%	13.80%
$D_{keypoints}$		87.86%	0.00%	43.93%	96.94%	0.00%	48.47%
F_M		11.39%	0.56%	5.97%	7.15%	9.30%	8.23%
F_{WLC}		18.09%	0.02%	9.05%	19.90%	0.84%	10.37%
F_{DST}		25.47%	0.02%	12.74%	35.02%	0.01%	17.52%
F_{LR}		23.68%	0.02%	11.85%	27.05%	0.35%	13.70%
$D_{ArXivMC}$	Complete	7.45%	1.00%	4.22%	3.87%	15.73%	9.80%
$D_{ArXivNaive}$		2.01%	1.60%	**1.81%**	0.97%	10.57%	5.77%
$D_{BIOSIGMC}$		25.76%	1.38%	13.57%	23.10%	8.38%	15.74%
$D_{BIOSIGNaive}$		11.47%	4.47%	7.97%	9.54%	14.31%	11.92%
$D_{keypoints}$		87.86%	0.00%	43.93%	96.94%	0.00%	48.47%
F_M		11.39%	0.31%	5.85%	7.15%	3.76%	**5.46%**
F_{WLC}		18.09%	0.02%	9.05%	19.90%	0.60%	10.25%
F_{DST}		25.47%	0.02%	12.74%	35.02%	0.02%	17.52%
F_{LR}		23.68%	0.02%	11.85%	27.05%	0.27%	13.66%
$D_{ArXivMC}$	Splicing	7.45%	2.57%	5.01%	3.87%	24.81%	14.34%
$D_{ArXivNaive}$		2.01%	3.54%	**2.77%**	0.97%	16.04%	**8.50%**
$D_{BIOSIGMC}$		25.76%	3.54%	14.65%	23.10%	17.39%	20.25%
$D_{BIOSIGNaive}$		11.47%	7.39%	9.43%	9.54%	21.22%	15.38%
$D_{keypoints}$		87.86%	0.02%	43.94%	96.94%	0.00%	48.47%
F_M		11.39%	1.45%	6.42%	7.15%	12.35%	9.75%
F_{WLC}		18.09%	0.07%	9.08%	19.90%	1.42%	10.66%
F_{DST}		25.47%	0.03%	12.75%	35.02%	0.03%	17.52%
F_{LR}		23.68%	0.03%	11.85%	27.05%	0.55%	13.80%

morphs). For the “adaptive” thresholding, the single detector HTER values reported significantly improve (e.g., from 9.62 to 2.2% for $D_{ArXivNaive}$ in *SC1* for *combined* morphs). In some cases, they are getting really close to the EER values for the corresponding experiment, which represents the best value that could be achieved in this test. The fusion results for this thresholding strategy see an even more significant drop in the HTER values presented (e.g., 13.41% to 2.8% for F_M in *SC1* for *combined* morphs).

For the second, an even less realistic (i.e., more lab-condition) test no additional test has to be performed here. Instead results from an earlier publication on fusion in face morph attack detection are re-used here. As authors of [7], we used a static percentage split (50%: 50%) on one a set containing genuine (originating from exactly one public database) and morphs that are derived directly from these genuine images to perform initial tests with DST in this field. The results presented were astonishing HTER values of less that 1%. While the results did indicate the potential benefit of using fusion in MAD, the observed lack of realism in the setup made us question the actual extend of the performance increase we could realistically hope for. This realization motivated the research work on the empirical limitations of using information fusion and the constraints for its application that lead to this journal paper.

Summarizing the results obtained on alternative (i.e., less realistic) evaluation setups, it has to be said that the error rates obtained achieved when drawing training and test data from the same parent population are obviously lower than in a setup with disjoint populations used. In the experiments discussed above, the fusion approaches benefit more from the unrealistic lab-condition like

Table 6 Realistic performance of the individual detectors and fusion approaches with the adaptive decision thresholds and fusion weights based on the estimated EER with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type marked in bold)

		SC1			*SC2*		
Detector	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
$D_{ArXivMC}$	Combined	30.08%	0.01%	15.04%	20.79%	0.89%	10.84%
$D_{ArXivNaive}$		19.21%	0.03%	9.62%	20.34%	0.50%	10.42%
$D_{BIOSIGMC}$		39.76%	0.35%	20.05%	37.03%	4.51%	20.77%
$D_{BIOSIGNaive}$		34.18%	0.65%	17.41%	33.76%	3.29%	18.52%
$D_{keypoints}$		47.95%	1.15%	24.55%	73.17%	0.26%	36.72%
F_M		26.81%	0.01%	13.41%	29.14%	0.42%	14.78%
F_{WLC}		0.60%	10.87%	**5.73%**	0.30%	46.48%	23.39%
F_{DST}		25.17%	0.02%	12.59%	33.53%	0.01%	16.77%
F_{LR}		14.00%	0.09%	7.04%	14.31%	1.90%	**8.10%**
$D_{ArXivMC}$	Complete	30.08%	0.00%	15.04%	20.79%	0.63%	10.71%
$D_{ArXivNaive}$		19.21%	0.02%	9.61%	20.34%	0.35%	10.34%
$D_{BIOSIGMC}$		39.76%	0.38%	20.07%	37.03%	3.45%	20.24%
$D_{BIOSIGNaive}$		34.18%	0.44%	17.31%	33.76%	2.46%	18.11%
$D_{keypoints}$		47.95%	1.13%	24.54%	73.17%	0.09%	36.63%
F_M		26.81%	0.01%	13.41%	29.14%	0.23%	14.68%
F_{WLC}		0.60%	10.30%	**5.45%**	0.30%	41.15%	20.72%
F_{DST}		25.17%	0.02%	12.59%	33.53%	0.02%	16.77%
F_{LR}		14.00%	0.05%	7.02%	14.31%	1.30%	**7.80%**
$D_{ArXivMC}$	Splicing	30.08%	0.03%	15.05%	20.79%	1.38%	11.08%
$D_{ArXivNaive}$		19.21%	0.05%	9.63%	20.34%	0.64%	10.49%
$D_{BIOSIGMC}$		39.76%	1.11%	20.44%	37.03%	8.60%	22.82%
$D_{BIOSIGNaive}$		34.18%	1.07%	17.62%	33.76%	4.35%	19.05%
$D_{keypoints}$		47.95%	0.78%	24.36%	73.17%	0.25%	36.71%
F_M		26.81%	0.01%	13.41%	29.14%	0.65%	14.89%
F_{WLC}		0.60%	17.18%	8.89%	0.30%	56.84%	28.57%
F_{DST}		25.17%	0.01%	12.59%	33.53%	0.03%	16.78%
F_{LR}		14.00%	0.26%	**7.13%**	14.31%	3.53%	**8.92%**

evaluation setups than the single detectors and the "adaptive" thresholding strategy benefits more than the "fixed" one.

5.6 Summary on the fusion experiments results

There are three main reasons why fusion fails to outperform the best individual classifier in the results discussed in section 5.3:

1. *Lack of diversity of the individual detectors.* The detectors $D_{ArXivNaive}$, $D_{arXivMC}$, $D_{BIOSIGMC}$, and $D_{BIOSIGNaive}$ are developed by the same research group and rely on training of DCNN with similar data sets but strong variances in data augmentation. Hence, it is very likely that these detectors make in field application mistakes on the same samples. Only the $D_{keypoints}$ detector relies on entirely different morphing detection clues and is developed by another research group using a different data set for training. In theory, an assumed clustering of four apparently very similar detectors might prove a strong prejudice in fusion that should be avoided at any cost. In practice, our experiment on different ensembles of classifiers showed a better performance if only those four detectors are used instead of all five.

2. *Lack of performance in individual detectors.* It can be seen from the evaluation with the DEFACTO dataset, that $D_{keypoints}$ lacks generalization power.

Table 7 Comparing fusion ensembles consisting of all five, one set of four ($D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, $D_{arXivNaive}$, and $D_{arXivMC}$), and one set of three ($D_{arXivNaive}$, $D_{BIOSIGMC}$, $D_{keypoints}$) detectors with the fixed decision thresholds and equal fusion weights with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type and ensemble size marked in bold)

		SC1			*SC2*		
Fusion	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
F_M (5 det)	Combined	11.39%	0.56%	**5.97%**	7.15%	9.30%	**8.23%**
F_{WLC} (5 det)		18.09%	0.02%	9.05%	19.90%	0.84%	10.37%
F_{DST} (5 det)		25.47%	0.02%	12.74%	35.02%	0.01%	17.51%
F_{LR} (5 det)		23.68%	0.02%	11.85%	27.05%	0.35%	13.70%
F_M (4 det)		2.98%	0.56%	**1.77%**	1.56%	9.30%	**5.43%**
F_{WLC} (4 det)		5.29%	1.07%	3.18%	2.31%	12.17%	7.24%
F_{DST} (4 det)		22.34%	0.02%	11.18%	19.75%	0.49%	10.12%
F_{LR} (4 det)		7.67%	0.61%	4.14%	4.47%	8.53%	6.50%
F_M (3 det)		26.14%	0.19%	13.16%	23.25%	3.85%	**13.55%**
F_{WLC} (3 det)		88.38%	0.00%	44.19%	98.06%	0.00%	49.03%
F_{DST} (3 det)		25.91%	0.01%	**12.96%**	44.86%	0.01%	22.43%
F_{LR} (3 det)		60.46%	0.01%	30.23%	77.35%	0.01%	38.68%
F_M (5 det)	Complete	11.39%	0.31%	**5.85%**	7.15%	3.76%	**5.46%**
F_{WLC} (5 det)		18.09%	0.02%	9.05%	19.90%	0.60%	10.25%
F_{DST} (5 det)		25.47%	0.02%	12.74%	35.02%	0.02%	17.52%
F_{LR} (5 det)		23.68%	0.02%	11.85%	27.05%	0.27%	13.66%
F_M (4 det)		2.98%	0.30%	**1.64%**	1.56%	3.76%	**2.66%**
F_{WLC} (4 det)		5.29%	0.97%	3.13%	2.31%	9.57%	5.94%
F_{DST} (4 det)		22.34%	0.02%	11.18%	19.75%	0.64%	10.19%
F_{LR} (4 det)		7.67%	0.58%	4.12%	4.47%	6.51%	5.49%
F_M (3 det)		26.14%	0.09%	13.11%	23.25%	1.36%	**12.30%**
F_{WLC} (3 det)		88.38%	0.00%	44.19%	98.06%	0.00%	49.03%
F_{DST} (3 det)		25.91%	0.02%	**12.96%**	44.86%	0.00%	22.43%
F_{LR} (3 det)		60.46%	0.00%	30.23%	77.35%	0.00%	38.67%
F_M (5 det)	Splicing	11.39%	1.45%	**6.42%**	7.15%	12.35%	**9.75%**
F_{WLC} (5 det)		18.09%	0.07%	9.08%	19.90%	1.42%	10.66%
F_{DST} (5 det)		25.47%	0.03%	12.74%	35.02%	0.03%	17.52%
F_{LR} (5 det)		23.68%	0.03%	11.85%	27.05%	0.55%	13.80%
F_M (4 det)		2.98%	1.45%	**2.21%**	1.56%	12.35%	**6.96%**
F_{WLC} (4 det)		5.29%	2.33%	3.81%	2.31%	16.71%	9.51%
F_{DST} (4 det)		22.34%	0.05%	11.19%	19.75%	0.84%	10.29%
F_{LR} (4 det)		7.67%	1.42%	4.55%	4.47%	11.66%	8.06%
F_M (3 det)		26.14%	0.48%	13.31%	23.25%	6.01%	**14.63%**
F_{WLC} (3 det)		88.38%	0.00%	44.19%	98.06%	0.00%	49.03%
F_{DST} (3 det)		25.91%	0.03%	**12.97%**	44.86%	0.03%	22.44%
F_{LR} (3 det)		60.46%	0.01%	30.24%	77.35%	0.00%	38.67%

The default decision threshold of 0.5 is far away from the sub-optimal (i.e., containing an offset due to training data vs. test data mismatch) threshold of 0.87252 obtained from its evaluation. Even higher are the sub-optimal decision thresholds with the mixed test data set (London, ECVP , and Alabama images). The values of approximately 0.97 for the *SC1* and 0.99 for the *SC2* indicate a large discrepancy between the data used for the training of the classifier and for evaluation/testing. As a consequence, the APCER and BPCER values are imbalanced, both are on the margins of the [0, 1] interval

Table 8 Comparing fusion ensembles consisting of all 5, 4 ($D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, $D_{arXivNaive}$, and $D_{arXivMC}$), and 3 ($D_{arXivNaive}$, $D_{BIOSIGMC}$, $D_{keypoints}$) detectors with the adaptive decision thresholds and fusion weights based on the estimated EER with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type and ensemble size marked in bold)

		SC1			*SC2*		
Detector	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
F_M (5 det)	Combined	26.81%	0.01%	13.41%	29.14%	0.42%	14.78%
F_{WLC} (5 det)		0.60%	10.87%	**5.73%**	0.30%	46.48%	23.39%
F_{DST} (5 det)		25.17%	0.02%	12.59%	33.53%	0.00%	16.77%
F_{LR} (5 det)		14.00%	0.09%	7.04%	14.31%	1.90%	**8.10%**
F_M (4 det)		17.20%	0.00%	8.60%	16.10%	0.42%	8.26%
F_{WLC} (4 det)		6.40%	0.86%	3.63%	3.06%	10.31%	6.68%
F_{DST} (4 det)		23.90%	0.01%	11.95%	21.68%	0.68%	11.18%
F_{LR} (4 det)		7.89%	0.62%	**4.26%**	4.77%	8.36%	**6.56%**
F_M (3 det)		29.41%	0.01%	14.71%	36.36%	0.19%	**18.28%**
F_{WLC} (3 det)		0.00%	64.59%	32.29%	0.00%	93.25%	46.62%
F_{DST} (3 det)		25.69%	0.01%	**12.85%**	43.59%	0.00%	21.80%
F_{LR} (3 det)		33.88%	0.00%	16.94%	42.18%	0.03%	21.10%
F_M (5 det)	Complete	26.81%	0.01%	13.41%	29.14%	0.23%	14.68%
F_{WLC} (5 det)		0.60%	10.30%	**5.45%**	0.30%	41.15%	20.72%
F_{DST} (5 det)		25.17%	0.02%	12.59%	33.53%	0.02%	16.77%
F_{LR} (5 det)		14.00%	0.05%	7.02%	14.31%	1.30%	**7.80%**
F_M (4 det)		17.20%	0.00%	8.60%	16.10%	0.23%	8.16%
F_{WLC} (4 det)		6.40%	0.78%	**3.59%**	3.06%	8.03%	5.54%
F_{DST} (4 det)		23.90%	0.02%	11.96%	21.68%	0.77%	11.23%
F_{LR} (4 det)		7.89%	0.61%	4.25%	4.77%	6.29%	**5.53%**
F_M (3 det)		29.41%	0.01%	14.71%	36.36%	0.04%	**18.20%**
F_{WLC} (3 det)		0.00%	64.16%	32.08%	0.00%	91.78%	45.89%
F_{DST} (3 det)		25.69%	0.02%	**12.85%**	43.59%	0.00%	21.80%
F_{LR} (3 det)		33.88%	0.00%	16.94%	42.18%	0.02%	21.10%
F_M (5 det)	Splicing	26.81%	0.00%	13.40%	29.14%	0.65%	14.89%
F_{WLC} (5 det)		0.60%	17.18%	8.89%	0.30%	56.84%	28.57%
F_{DST} (5 det)		25.17%	0.01%	12.59%	33.53%	0.03%	16.78%
F_{LR} (5 det)		14.00%	0.26%	**7.13%**	14.31%	3.53%	**8.92%**
F_M (4 det)		17.20%	0.00%	8.60%	16.10%	0.65%	8.37%
F_{WLC} (4 det)		6.40%	2.01%	**4.21%**	3.06%	14.28%	8.67%
F_{DST} (4 det)		23.90%	0.05%	11.98%	21.68%	1.17%	11.43%
F_{LR} (4 det)		7.89%	1.45%	4.67%	4.77%	11.32%	**8.05%**
F_M (3 det)		29.41%	0.00%	14.71%	36.36%	0.24%	**18.30%**
F_{WLC} (3 det)		0.00%	75.02%	37.51%	0.00%	97.08%	48.54%
F_{DST} (3 det)		25.69%	0.01%	**12.85%**	43.59%	0.02%	21.80%
F_{LR} (3 det)		33.88%	0.00%	16.94%	42.18%	0.07%	21.12%

and the HTER values are close to 43% in *SC1* and 48% in *SC2* for the "fixed" thresholding strategy. If the decision threshold for $D_{keypoints}$ is readjusted, based on the training set (DEFACTO), the HTER values in testing become significantly lower, approximately 24% in *SC1* and 36% in *SC2*. However, the APCER and BPCER values are still imbalanced. The impact of one bad detector on the overall fusion is shown very well in the experiment on different ensembles of classifiers showed where a better performance was achieved when only an ensemble of four (all except $D_{keypoints}$) is used.

Table 9 Fusion under laboratory conditions: tenfold stratified cross-validation with 90% training/10% test split; genuine samples from the Alabama dataset [53]; morphs from LondonDB and UtrechtDB (best result per morph type and application scenario marked in bold)

	Combined				**Complete**				**Splicing**			
	SC1		***SC2***		***SC1***		***SC2***		***SC1***		***SC2***	
	EER	**HTER**	**EER**	**HTER**	**EER**	**HTER**	**EER**	**HTER**	**EER**	**HTER**	**EER**	**HTER**
Fixed												
$D_{ArXivMC}$	3.8%	4.3%	8.2%	11.2%	3.9%	4.2%	7.2%	9.8%	4.8%	5.0%	9.2%	14.3%
$D_{ArXivNaive}$	1.5%	**1.9%**	3.9%	7.1%	1.3%	**1.8%**	3.4%	**5.8%**	1.8%	**2.7%**	4.4%	8.5%
$D_{BIOSIGMC}$	9.3%	13.6%	15.7%	16.6%	9.3%	13.6%	14.7%	15.7%	12.8%	14.7%	20.2%	20.2%
$D_{BIOSIGNaive}$	7.1%	8.4%	13.4%	13.7%	7.0%	7.9%	11.8%	11.9%	8.4%	9.4%	14.2%	15.4%
$D_{keypoints}$	12.3%	43.9%	18.6%	48.8%	12.2%	43.9%	19.3%	48.8%	8.9%	43.9%	18.3%	48.8%
F_{M}		6.0%		8.2%		6.2%		5.9%		6.4%		9.7%
F_{WLC}		9.6%		10.9%		9.2%		10.6%		9.2%		10.6%
F_{DST}		2.6%		**5.9%**		3.0%		6.7%		2.9%		**7.3%**
F_{LR}		2.6%		**5.9%**		3.0%		6.7%		2.9%		**7.3%**
Adaptive												
$D_{ArXivMC}$	3.8%	3.9%	8.2%	8.3%	3.9%	4.0%	7.2%	7.9%	4.8%	4.9%	9.2%	9.4%
$D_{ArXivNaive}$	1.5%	**2.2%**	3.9%	**5.0%**	1.3%	**2.1%**	3.4%	**4.4%**	1.8%	**3.0%**	4.4%	**5.6%**
$D_{BIOSIGMC}$	9.3%	9.8%	15.7%	16.4%	9.3%	9.8%	14.7%	15.0%	12.8%	12.7%	20.2%	20.7%
$D_{BIOSIGNaive}$	7.1%	7.9%	13.4%	13.6%	7.0%	7.6%	11.8%	12.1%	8.4%	9.1%	14.2%	15.0%
$D_{keypoints}$	12.3%	12.7%	18.6%	19.0%	12.2%	12.5%	19.3%	19.3%	8.9%	11.4%	18.3%	19.2%
F_{M}		2.8%		6.0%		2.2%		4.5%		3.3%		6.8%
F_{WLC}		15.2%		39.2%		14.3%		35.5%		17.7%		45.8%
F_{DST}		2.8%		5.8%		3.3%		6.6%		3.1%		7.3%
F_{LR}		2.8%		5.8%		3.3%		6.6%		3.1%		7.3%

3. *Lack of similarity between the training and test data.* Different proprietary data sets are used for training individual classifiers, which is a very common case, but the datasets for adjusting fusion parameters (evaluation data set) and for actual testing are also very different from each other and the training data set. One can say that it makes absolutely no sense to use different data sources for adjusting fusion parameters and for testing, but this is the real-life situation. In practice, it is very difficult to precisely foresee and provide significant in-field data at the stage of system development or parameter adjustment. Moreover, there is no guarantee that the in-field data that will be obtained in the future is even similar to the presented training data.

The case study performed in this paper clearly demonstrates that if the training, evaluation, and test datasets lack similarity, the adaptation of the classifier parameters such as a decision threshold may lead to performance degradation. This can be well explained on the example of the classifier $D_{ArXivNaive}$ which in the tests performed shows the best generalization power. The classifier is well trained with the default decision threshold of 0.59072. An attempt to adapt the decision threshold based on the DEFACTO data set actually fails with shifting it to 0.39958, resulting in an EER of 10%. As a consequence, the APCER and BPCER values are imbalanced in the test leading to the HTER values of approximately 9.5% in *SC1* and 10.5% in *SC2* (see Table 6). However, if there is no adaptation of the decision threshold, the sub-optimal (i.e., offset) thresholds of 0.594687, 0.600357, and 0.566983 are close to the default one and the APCE R and BPCER values are well balanced in *SC1* leading to HTER values of 1.91%, 1.81%, and 2.77% for combined, complete, and splicing morphs respectively (see Table 5). In contrary, the sub-optimal thresholds in the *SC2* would be 0.499938, 0.507648, and 0.492199 for combined, complete, and splicing morphs respectively which are far away from the default value of 0.59072. Hence, in the test within *SC2* the APCER and BPCER values are imbalanced leading to the HTER values of 7.14%, 5.77%, and 8.50% for combined, complete, and splicing morphs respectively. The same situation can be observed with the detectors $D_{arXivMC}$, $D_{BIOSIGMC}$, and $D_{BIOSIGNaive}$.

Considering the results of different fusion strategies, it can be said that in almost all cases, the APCER and BPCER values are imbalanced in the case when training, evaluation, and test datasets lack similarity. This results in the conclusion that pre-determining the proper decision thresholds (as well as the fusion weights) in real-life conditions (where the training, evaluation, and in-field data might be dramatically different) is hardly possible.

When considering alternative (less strict) evaluation setups, where training and test data show and artificial similarity due to the fact that they have been drawn from the same parent distribution, we see in section 5.5 significantly lower HTER values not only for fusion results but in some cases also for the individual detectors.

The results presented more clear indicators that the similarity between the training and test data is the dominating factor for the error rates achieved. If this similarity is an artificial one (e.g., in an unrealistic setup where training, parameterization, and test data are drawn from the same parent population) instead of a natural one (i.e., the fusion as well as the individual detectors are suitably well trained) the low error rates obtained are meaningless.

The practical consequence of these three issues is that one of the individual detectors (obviously accurate but far from perfect in its performance) in all evaluations outperforms four different fusion approaches, ranging from simplistic to very sophisticated, in different parameterizations in the tests performed in 5.3 but becomes marginalized by fusion approaches as soon as either the ensemble of detectors used in the fusion is optimized (as done by removing one disturbing detector in section 5.4) or the similarity between training and test data is increased (as in section 5.5).

6 Conclusions

The results presented in the empirical evaluations in this paper demonstrate that fusion can fail even with a set of relevant individual classifiers. This can be seen in both application scenarios (“MAD in document issuing” and “MAD in identity verification”) evaluated in this paper. Here, the three reasons for this phenomenon discussed above are (a) low diversity of the detectors, (b) lack of performance in individual detectors, and (c) lack of similarity between the training and test data.

Summarizing the lessons learned from the approach of using fusion for MAD detection as done in this paper and drawing some generalization toward other media forensics classification or decision problems, the following has to be said: The requirements for (media) forensic methods in terms of scientific admissibility (or Daubert compliance) are obviously important! Methods should indeed be published upon and peer reviewed, their error rates should be precisely known and standards for the application of methods should be known. But the threat that Champod and Vuille identify as a problem of ascertaining the error rates of a test “can prove misleading if not all its complexities are understood” [15] plays a very significant role as demonstrated in the evaluations performed here.

Besides the requirements for individual expert systems to be used in forensic investigations (including its accurateness), if it comes to information fusion, additional constraints have to be observed. These are, at least:

- The diversity of the detectors, which has to be ascertained either by knowledge about the precise means of decision generation and the diversity of those means or empirically.
- An independent and thorough benchmarking of detectors to establish also an idea on the generalization power of performance claims made by their creators.
- Considerations on the similarity/correlation between training data available (during training of the individual classifiers and the training of the fusion methods) and the data to be expected in field application are very important. If very precise assumptions are possible on the application data, weighting might be applicable in fusion. Else-wise, only unweighted fusion strategies like majority voting or the sum-rule should be employed, if any fusion is used in those cases at all.

The diversity issue becomes very problematic if features (as the means to represent a decision problem in a feature space) are not hand crafted by experts but learned, e.g., by DCNN. In this paper, the diversity problem of the detectors used here as “black boxes” has been established in direct contact with the developers of those methods, which is hardly an option in most field applications.

Also, the recent trend to generate synthetic data sets for the training of pattern recognition methods (either traditional or neural network based) introduces another degree of freedom into the characteristics of datasets. In publications such as [54], this approach is used to avoid tedious data collection tasks while creating sufficiently sized data sets for modern day data-greedy classifiers. The problem here is the influence of the synthesis process on its output (i.e., the synthesis-specific artifacts) that will become part of the model trained by each classifier. It is related to the questions of source characteristics imposing themselves into trained models but carries a different degree of relevance for forensic application scenarios.

The general problem with training- and test data being mismatched in practice is hardly new. It hardly ever occurs in scientific papers on applied pattern recognition, because it can easily be prevented in lab tests. Nevertheless, it is a very good argument why media forensic methods should undergo rigorous testing and benchmarking by third parties, like it is done in the field of MAD in the NIST FRVT MORPH challenge. Only such joint efforts can lead to methods that might become mature enough to aim at court admissibility.

Abbreviations

ABC: Automated border control; AFR: Automated face recognition; APCE R: Attack Presentation Classification Error Rate (from [ISO/IEC 30107-3:2017]); AUC: Area under curve; BPCER: Bona Fide Classification Error Rate (from [ISO/IEC 30107-3:2017]); BSIF: Binarized Statistical Image Features; DB: Database; DCNN: Deep convolutional neural networks; DET: Detection error trade-off; dpi: Dots per inch; DST: Dempster-Shafer Theory; ECVP: European Conference on Visual Perception; EER: Equal error rate; HTER: Half total error rate; IEC: International Electrotechnical Commission; eMRTD: Electronic machine readable travel documents; EU: European Union; FRE: Federal Rules of Evidence; FRVT: (NIST) Face Recognition Vendor Test; GAN: Generative adversarial networks; HOG: Histogram of oriented gradients; ICAO: International Civil Aviation Organization; ID: Identity document; ILSV RC: ImageNet Large Scale Visual Recognition Challenge; ISO: International Organization for Standardization; JPEG: Joint Photographic Experts Group image file format; LBP: Local binary patterns; LR: Likelihood ratio; MAD: Morphing attack detection; MAX: Maximum; MIN: Minimum; NIST: National Institute of Standards and Technology; PNG: Portable Network Graphics; QA: Quality assessment; SC: Scenario; SIFT: Scale-invariant feature transform; SURF: Speeded up robust features; SVM: Support vector machine; TIFF: Tagged image file format; WLC: Weighted linear combination

Acknowledgements

The authors wish to thank Clemens Seibold (Fraunhofer HHI, Berlin, Germany) for collecting the version of the Alabama image set used here and for running our data through his four detectors as well as the important discussions about the training data used to generate those "black box" detectors. The authors wish to thank Tom Neubert for providing the keypoints detector used.

The experiments in sections 5.4 and 5.5 were inspired by the reviewers in the first round of reviews for this journal paper. We wish to express our gratitude to those experts unknown to us because we feel that their recommendations significantly helped to improve this paper.

Authors' contributions

CK works on the media forensic perspectives and fusion theory parts as well as the interpretation of results. AM work on the biometric perspective, the dataset creation, the conduction of the experiments (classifier selection and fusion operator implementation), and the interpretation of the experimental results. JD initial structuring of the work and definition of focus and scope of the work (incl. suggesting the two application scenarios as well as the usage of DST and LR based fusion). MH theoretical and practical work on likelihood based fusion and the interpretation of its results. The authors read and approved the final manuscript.

Funding

The work in this paper has been funded in part by the German Federal Ministry of Education and Science (BMBF) through the research program under the contract no. FKZ: 16KIS0509K (research project ANANAS). The work in this paper has been funded in part by the Deutsche Forschungsgemeinschaft (DFG) under contract no.: 421860277 (research project GENSYNTH). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The empirical work in this paper is based on the following publicly available datasets:

- The AMSL Face Morph Image Data Set (made available via: https://omen.cs.uni-magdeburg.de/disclaimer/index.php; last accessed Sept. 10, 2020)
- The Utrecht/ECVP as part of Psychological Image Collection at Stirling (PICS), (available at: http://pics.stir.ac.uk/2D_face_sets.htm; last accessed Sept. 10, 2020)
- The London DB has been made available by *L. DeBruine and B. Jones as: Face Research Lab London Set:* https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666 *(last accessed Sept. 10th, 2020)*
- The DEFACTO dataset (including the face morphing subset used in this paper) introduced in [51] is available at: https://defactodataset.github.io/ *(last accessed Sept. 10, 2020)*
- The used Alabama database is the collection of mugshots of the Alabama News Network (available at: https://www.alabamanews.net/mugshots/; last accessed Sept. 10, 2020)

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 21 September 2020 Accepted: 17 June 2021
Published online: 29 July 2021

References

1. H. Drucker, C. Cortes, L.D. Jackel, Y. LeCun, V. Vapnik, Boosting and other ensemble methods. Neural Comput. **6**(6), 1289–1301 (1994)
2. L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms* (Wiley-Interscience, 2004)
3. M. Tan, *Multi-agent reinforcement learning: independent vs. cooperative agents. Readings in agents* (Morgan Kaufmann Publishers Inc., San Francisco, 1997), pp. 487–494
4. T.K. Ho, in *Hybrid Methods in Pattern Recognition*, ed. by A. Kandel, H. Bunke. Multiple classifier combination: lessons and the next steps (World Scientific Publishing, 2002), pp. 171–198
5. A. Ross, K. Nandakumar, A.K. Jain, *Handbook of Multibiometrics* (Springer, 2006)
6. R.P. Srivastava, Alternative Form of Dempster's Rule for Binary Variables. Int. J. Intell. Syst. **20**(8), 789–797 (2005)
7. A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, P. Eisert, in *Proc. 27th European Signal Processing Conference (EUSIPCO)*. Dempster-Shafer Theory for Fusing Face Morphing Detectors (A Coruna, 2019), pp. 1–5
8. A. Makrushin, A. Wolf, in *Proc. 26th European Signal Processing Conference (EUSIPCO)*. An Overview of Recent Advances in Assessing and Mitigating the Face Morphing Attack (2018)
9. U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, C. Busch, Face Recognition Systems Under Morphing Attacks: A Survey. IEEE Access **7**(2019), 23012–23026 (2019)
10. M. Ferrara, A. Franco, D. Maltoni, in *Face Recognition Across the Electromagnetic Spectrum*, ed. by T. Bourlai. On the effects of image alterations on face recognition accuracy (Springer, Cham, 2016), pp. 195–222
11. R.S.S. Kramer, M.O. Mireku, T.R. Flack, K.L. Ritchie, Face morphing attacks: investigating detection with humans and computers. Cogn. Res. Princ. Implications **4**, 28 (2019)
12. A. Makrushin, T. Neubert, J. Dittmann, in *Proc. 12th Int. Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6*. Automatic generation and detection of visually faultless facial morphs (VISAPP, 2017), pp. 39–50
13. U. Scherhag, R. Raghavendra, K.B. Raja, M. Gomez-Barrero, C. Rathgeb, C. Busch, in *Proc. 5th International Workshop on Biometrics and Forensics (IWBF)*. On the Vulnerability of Face Recognition Systems: Towards Morphed Face Attacks (2017)
14. National Institute of Standards and Technology (NIST) FRVT MORPH, https://pages.nist.gov/frvt/html/frvt_morph.html
15. C. Champod, J. Vuille, in *International Commentary on Evidence. Vol. 9, Issue 1*. Scientific evidence in Europe – admissibility, evaluation and equality of arms (2011) Available at: https://core.ac.uk/reader/85212846 (Last accessed: 26 Aug 2020)
16. M. Ferrara, A. Franco, D. Maltoni, in *Proc. Int. Joint Conf. on Biometrics (IJCB)*. The magic passport (2014), pp. 1–7

17. C. Seibold, W. Samek, A. Hilsmann, P. Eisert, in Proc. Int. Workshop Digital Watermarking (IWDW2017). Detection of Face Morphing Attacks by Deep Learning (Springer, Berlin, 2017)
18. U. Scherhag, C. Rathgeb, C. Busch, in *Proc. 13th IAPR Workshop on Document Analysis Systems (DAS'18)*. Towards detection of morphed face images in electronic travel documents (2018)
19. C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, J. Dittmann, in *Proc. 5th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'17)*. Modeling attacks on photo-ID documents and applying media forensics for the detection of facial morphing (ACM, New York, 2017), pp. 21–32
20. Utrecht ECVP as part of Psychological Image Collection at Stirling (PICS), http://pics.stir.ac.uk/2D_face_sets.htm, last accessed: 31 Aug 2020.
21. R. Raghavendra, S. Venkatesh, K. Raja, C. Busch, in *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. Towards making morphing attack detection robust using hybrid scale-space colour texture features (2019), pp. 1–8
22. M. Ferrara, A. Franco, D. Maltoni, Face demorphing. Trans. Inf. Forensics Secur. **13**(4), 1008–1017 (2018)
23. D.O. del Campo, C. Conde, D. Palacios-Alonso, E. Cabello, Border control morphing attack detection with a convolutional neural network de-morphing approach. IEEE Access **8**, 92301–92313 (2020)
24. F. Peng, L. Zhang, M. Long, FD-GAN: Face De-Morphing Generative Adversarial Network for Restoring Accomplice's Facial Image. IEEE Access **7**, 75122–75131 (2019)
25. U. Scherhag, D. Budhrani, M. Gomez-Barrero, C. Busch, in *International Conference on Image and Signal Processing (ICISP 2018)*. Detecting Morphed Face Images Using Facial Landmarks (2018), pp. 444–452
26. C. Seibold, W. Samek, A. Hilsmann, P. Eisert, Accurate and robust neural networks for security related applications exampled by face morphing attacks. Arxiv/CoRR abs/1806.04265 (2018)
27. U. Scherhag, C. Rathgeb, J. Merkle, C. Busch, in *IEEE Transactions on Information Forensics and Security (TIFS)*. Deep Face Representations for Differential Morphing Attack Detection (2020)
28. L. Wandzik, G. Kaeding, R.V. Garcia, in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2018*. Morphing detection using a general- purpose face recognition system (2018), pp. 1012–1016
29. R. Raghavendra, K. Raja, S. Venkatesh, C. Busch, in *Proc. 30th Int. Conf. on Computer Vision and Pattern Recognition Workshop*. Transferable Deep-CNN features for detecting digital and print-scanned morphed face images (2017)
30. T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, J. Dittmann, Extended StirTrace Benchmarking of Biometric and Forensic Qualities of Morphed Face Images. IET Biometrics **7**(4), 325–332 (2018)
31. T. Karras, S. Laine, T. Aila, in *IEEE Conference on Computer Vision and Pattern Recognition*. A style-based generator architecture for generative adversarial networks (2019), pp. 4401–4410
32. T. Karras, T. Aila, S. Laine, J. Lehtinen, in *International Conference on Learning Representations*. Progressive growing of GANs for improved quality, stability, and variation (2018)
33. N. Damer, A.M. Saladie, A. Braun, A. Kuijper, in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS), Oct. 2018*. MorGAN: recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network (2018), pp. 1–10
34. S. Venkatesh, H. Zhang, R. Raghavendra, K. Raja, N. Damer, C. Busch, *Can GAN generated morphs threaten face recognition systems equally as landmark based morphs? -vulnerability and detection* (International Workshop on Biometrics and Forensics (IWBF), 2020)
35. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, 1976)
36. P. Smets, in *Proc. 15th Conf. On Uncertainty in Artificial Intelligence*. Practical uses of belief functions, vol 99 (1999), pp. 612–621
37. J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
38. L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd edn. (Wiley, New York, 2014)
39. M. Fontani, A. Bonchi, A. Piva, M. Barni, in *Proc. Media Watermarking, Security, and Forensics 2014, San Francisco, CA, USA, February 2, 2014*, ed. by A. M. Alattar, N. D. Memon, C. Heitzenrater. Countering anti-forensics by means of data fusion, vol 9028 (SPIE Proceedings, 2014), p. 90280Z SPIE
40. M. Fontani, T. Bianchi, A. De Rosa, A. Piva, M. Barni, A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence. IEEE Trans. Inf. Forensics Secur. **8**(4), 593–607 (2013)
41. Royal Courts of Justice, "R v T", [2010] EWCA Crim 2439, Redacted Judgment, 2011, Available at: http://www.bailii.org/ew/cases/EWCA/Crim/2010/2439.pdf (last accessed: 10 Mar 2021)
42. Y. Peng, L.J. Spreeuwers, R.N.J. Veldhuis, in *Proceedings of the 3rd International Workshop on Biometrics and Forensics, IWBF 2015*. Likelihood Ratio Based Mixed Resolution Facial Comparison (IEEE Computer Society, USA, 2015), pp. 1–5
43. T. Kerkvliet, R. Meester, Assessing forensic evidence by computing belief functions. Law Probability Risk **15**(2), 127–153 (2016)
44. T.G. Dietterich, in *Multiple classifier systems*. Ensemble methods in machine learning (Springer LNCS 1857, 2000), pp. 1–15
45. B. Quost, M.-H. Masson, T. Denœux, Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules. Int. J. Approx. Reason. **52**(3), 353–374 (2011)
46. K. Tumer, J. Ghosh, Error correlation and error reduction in ensemble classifiers. Connect. Sci. **3–4**(8), 385–404 (1996)
47. S.P. Lund, H. Iyer, Likelihood ratio as weight of forensic evidence: a closer look. J. Res. Nat. Instit. Stand. Technol. **122**, 122.027 (2017) 2017
48. A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence. Law Probability Risk **11**(1), 1–24 (2012)
49. K. Nandakumar, Y. Chen, S.C. Dass, A. Jain, Likelihood Ratio-Based Biometric Score Fusion. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 342–347 (2008)
50. ISO/IEC JTC1 SC37 Biometrics, ISO/IEC 30107-3:2017 Information technology- biometric presentation attack detection - Part3: Testing & reporting. ISO, 2017.
51. G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. L. Dugelay, M. Pic: DEFACTO: Image and Face Manipulation Dataset. 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5, (dataset: https://defactodataset.github.io/), 2019.
52. L. DeBruine, B. Jones: Face Research Lab London Set. May 30th, 2017. Available at: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666 (last accessed 10 Sept 2020).
53. Alabama News Network Mugshot database, online: https://www.alabamanews.net/mugshots/ (last accessed: 9 Sept 2020).
54. D. Maltoni, D. Maio, A.K. Jain, S. Prabhakar, in *Synthetic Fingerprint Generation*. Handbook of Fingerprint Recognition (Springer, London, 2009), pp. 271–302

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

10

[Siegel21] Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features

This chapter of this cumulative habilitation treatise has originally been published as:

[Siegel21] Dennis Siegel, **Christian Kraetzer**, Stefan Seidlitz, Jana Dittmann: *Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features*. Journal of Imaging 2023, no. 7, 2021. https://doi.org/10.3390/jimaging7070108

This article belongs to the Special Issue Image and Video Forensics: https://www.mdpi.com/journal/jimaging/special_issues/image_video_forensics

Shares/author contributions (as given in the original paper - see page 166 of this cumulative habilitation treatise): "*Conceptualization, Dennis Siegel (D.S.), Christian Kraetzer (C.K.) and Jana Dittmann (J.D.); data curation, D.S.; funding acquisition, C.K. and J.D.; investigation, D.S. and Stefan Seidlitz (S.S.); methodology, C.K. and J.D.; project administration, C.K.; software, D.S.; supervision, C.K. and J.D.; validation, C.K. and S.S.; visualization, D.S.; writing—original draft, D.S.; writing—review and editing, C.K., S.S. and J.D.*
All authors have read and agreed to the published version of the manuscript."

Journal of *Imaging*

MDPI

Article

Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features

Dennis Siegel *, Christian Kraetzer *, Stefan Seidlitz and Jana Dittmann

Department of Computer Science, Otto-von-Guericke University, 39106 Magdeburg, Germany; stefan.seidlitz@ovgu.de (S.S.); jana.dittmann@ovgu.de (J.D.)
* Correspondence: dennis.siegel@ovgu.de (D.S.); christian.kraetzer@ovgu.de (C.K.)

Abstract: DeepFake detection is a novel task for media forensics and is currently receiving a lot of research attention due to the threat these targeted video manipulations propose to the trust placed in video footage. The current trend in DeepFake detection is the application of neural networks to learn feature spaces that allow them to be distinguished from unmanipulated videos. In this paper, we discuss, with features hand-crafted by domain experts, an alternative to this trend. The main advantage that hand-crafted features have over learned features is their interpretability and the consequences this might have for plausibility validation for decisions made. Here, we discuss three sets of hand-crafted features and three different fusion strategies to implement DeepFake detection. Our tests on three pre-existing reference databases show detection performances that are under comparable test conditions (peak AUC > 0.95) to those of state-of-the-art methods using learned features. Furthermore, our approach shows a similar, if not better, generalization behavior than neural network-based methods in tests performed with different training and test sets. In addition to these pattern recognition considerations, first steps of a projection onto a data-centric examination approach for forensics process modeling are taken to increase the maturity of the present investigation.

Keywords: DeepFake detection; hand-crafted features; forensic process model; plausibility of decisions

check for updates

Citation: Siegel, D.; Kraetzer, C.; Seidlitz, S.; Dittmann, J. Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features. *J. Imaging* **2021**, *7*, 108. https://doi.org/10.3390/jimaging7070108

Academic Editors: Irene Amerini, Gianmarco Baldini and Francesco Leotta

Received: 31 May 2021
Accepted: 29 June 2021
Published: 1 July 2021

1. Introduction

DeepFakes (a neologism combining the terms "deep learning" and "fake") are synthetic videos (or images) in which a person's face (and optionally also voice) is replaced with someone else's likeness using deep learning technologies. Having emerged in late 2017, DeepFakes nowadays pose a serious threat to the trust placed in video footage. Papers such as [1,2] elaborate on the effect of DeepFakes on current politics, disinformation and trust.

Like countering any other form of image, audio or video manipulation, detecting DeepFakes is an important task for media forensics and is currently receiving a lot of research attention due to the significance of the threat.

According to a well established definition given in [3], information technology (IT) forensics is: *"The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal,* [...]".

This paper focuses on DeepFake detection as a novel challenge in the IT forensics subdiscipline of media forensics. In contrast to many other forensic subdisciplines, such as, e.g., the field of fingerprint analysis, this field is an especially young and immature research field, currently being far away from achieving the ultimate goal of courtroom readiness.

Regarding the basic methodology applied in the state-of-the-art work in DeepFake detection, it can be stated that most of the current research work is based on pattern recognition approaches using feature spaces learned with the help of neural networks. While this method achieves promising detection rates for small scale empirical evaluations with selected DeepFake datasets, it has the inherent drawback that it is extremely hard to validate the plausibility of decisions made by a neuronal network since the semantics of the features learned cannot easily be interpreted by humans. For other, more established, pattern recognition disciplines such as template matching or statistical pattern recognition, the issue of plausibility testing also exists, because the results generated by the application of machine learning strategies lack the intuitive verification that usually accompanies human decision-making processes. Nevertheless, for these disciplines, validation methods have been developed over the decades to establish whether the results of the learning and decision processes are reasonable. In practice, this means to establish that the patterns trained and detected are really the patterns that the user wants to distinguish between and that side-effects as well as external influence factors are known for the pattern recognition process. Such methods, which include, amongst others, feature selection strategies, as well as model analysis methods aimed at establishing the exact decision (or detection) performance and error behavior of an analysis method. The reason to do this is that this knowledge determines the plausibility of the result of the application of pattern recognition mechanisms in a forensic application scenario and should therefore be directly linked to the trust we place in their decisions.

In addition to the problems in estimating the plausibility of decisions of current (mostly neural network-driven) DeepFake detection methods, a second shortcoming in the current state of the art in this field has to be mentioned here: Apart from the considerations of efficiency (i.e., detection performance and plausibility), all forensic methods should aim at fulfilling some form of forensic conformity. Criteria for such conformity should address the admissibility of methods as a basis for expert witnesses' testimony as evidence in legal proceedings. For the United States of America (by far the most active legal system worldwide), those criteria are codified, amongst other regulations, by the so called Daubert standard (see e.g., [4] or [5] for a detailed discussion of this US case-law standard) in combination with the US Federal Rules of Evidence (FRE) [6]. In addition to those admission criteria for expert witnesses' testimony questions of evidence handling (i.e., chain of custody) also have to be looked into.

To address aspects of these two identified shortcomings (i.e., the explainability issues of feature spaces learned using a neural network on one hand and the lack of adherence to forensic process models on the other hand), this paper provides the following two main contributions :

- Using hand-crafted features for DeepFake detection and comparison with the performance of state-of-the-art deep learning-driven approaches, we discuss three sets of hand-crafted features and three different fusion strategies to implement DeepFake detection. Those features analyze the blinking behavior, the texture of the mouth region as well as the degree of texture found in the image foreground. Our tests on three pre-existing reference databases show detection performances that are under comparable test conditions to those of state-of-the-art methods using learned features (in our case obtaining a maximum AUC of 0.960 in comparison to a maximum AUC of 0.998 for a recent approach using convolutional neural networks). Furthermore, our approach shows a similar, if not better, generalization behavior (i.e., AUC drops from values larger than 0.9 to smaller than 0.7) than neural network based methods in tests performed with different training and test sets .
  In addition to those detection performance issues, we discuss at length that the main advantage which hand-crafted features have over learned features is their interpretability and the consequences this might have for plausibility validation for decisions made.
- Projection onto a forensic process model: With the aim to improve the maturity of pattern recognition-driven media forensics, we perform first steps of the projection of our work onto an established forensic process model. For this, a derivative of the forensic process model for IT forensics published in 2011 by the German Federal Office for Information Security (BSI) is used here. This derivative, or more precisely extension,

is called the Data-Centric Examination Approach (DCEA) and has seen its latest major overhaul in 2020 in [7]. While it is not yet perfectly capable of fitting the needs of media forensics analyses, our work shows first benefits of this modeling as well as points where DCEA would need to undergo further extension to fit those purposes.

The paper is structured as follows: In Section 2, the background and state of the art in DeepFake detection (Section 2.1), feature space design alternatives (Section 2.2) and the forensic process model chosen for this paper (Section 2.3) are discussed. Section 3 discusses the chosen solution concept for implementing DeepFake detection with hand-crafted features, while Section 4 focuses on implementation details.

Section 5 presents and discusses our evaluation results, structured into results for individual detectors (Section 5.1) and for fusion operators (Section 5.2). In Section 6, we provide a summary of the results and a comparison with other approaches from the state of the art (in Section 6.1) as well as our conclusion on the comparison between hand-crafted and learned features for DeepFake detection (in Section 6.2). Section 7 closes the paper with some indication for potential future work.

2. Background and State of the Art

By arguing that "Multimedia Forensics is not Computer Forensics", the authors of [8] point out that "*multimedia forensics and computer forensics belong to the class of digital forensics, but they differ notably in the underlying observer model that defines the forensic investigator's view on (parts of) reality,* [. . .] *while perfect concealment of traces is possible for computer forensics, this level of certainty cannot be expected for manipulations of sensor data*". Even though this statement dates back to 2009, before the rise of neural network-driven data generation methods, such as generative adversarial networks (GANs), it still holds true; additionally, modern-day targeted media manipulations such as DeepFake generation, either leave telltale traces of the manipulation (here, the synthesis and insertion of a face into a video) or violate the source characteristics (e.g., violating the noise pattern of the camera). Recent papers on DeepFake detection, such as [9], provide strong indication that, if applied correctly, targeted detection using pattern recognition methods might be a viable media forensics approach to counter DeepFakes.

In Section 2.1 of this chapter, the state of the art regarding recent DeepFake detection methods is briefly summarized. Following this survey, which points out that nearly all recent methods found in the literature are looking at learned feature spaces as a means of tackling this pattern recognition problem, Section 2.2 discusses the existing alternatives for feature space design and reflects upon their suitability in sensitive decision processes, such as e.g., medical image processing or (media) forensics. Additionally, in Section 2.3, a discussion on the needs for integration of pattern recognition-driven methods into a forensic process model is summarized.

2.1. DeepFake Detection

Usually, the detection of DeepFakes happens with various combined Convolutional Neural Network (CNN) architectures such as autoencoders (AEs). The reasons behind this are obvious: First, most DeepFakes are produced with AEs because internet platforms such as YouTube provide many video sources with different human faces which are usable for the training of DeepFake generators based on neural networks. FakeApp [10] is one example of an autoencoder–decoder structure which is able to swap the latent features of two different faces [11]. These architectures introduce several artifacts to the video while creating a DeepFake that are, in most cases, not visible for the human eye but are potential artifacts that could be utilized for DeepFake detection using image or video analysis methods. It stands to reason that neural networks are also useful for the detection of DeepFake videos, assuming that there is a sufficiently large set of representative data to train features, allowing for the localization of the aforementioned artifacts. Second, which is also a consequence of the first reason, several large and publicly available DeepFake databases (such as FaceForensic++ [12] or Celeb-DF [13]) already exist and provide huge datasets, which can easily be used for the training of CNN-based DeepFake detectors.

The survey paper from Nguyen et al. [11] summarizes different DeepFake detection approaches into the two main categories of *temporal features across video streams* (i.e., inter-frame analysis) and *visual artifacts within video frames* (i.e., intra-frame analysis). For example, the approach of Sabir et al. [14] extracts temporal features of video streams for the detection of DeepFake videos: The authors analyze a potential DeepFake video frame-by-frame for low level artifacts which are only present in single frames to class a video as a DeepFake. Then, they use a Recurrent Convolutional Network (RCN) model to detect and track the temporal artifacts across frames [11,14]. In Li et al.'s work [15], another CNN-based inter-frame analysis approach addresses the eye blinking of a person in a video under the assumption that many DeepFake generated videos are not able to reproduce a natural blinking behavior. The authors first extract the eye areas based on six eye landmarks from a segmented face region. After that, they use the extracted eye area of all video frames in a long-term recurrent convolutional network (LRCN) to detect temporal discrepancies in the blinking behavior [11,15]. An approach which should also be considered for these temporal features across video streams category is described in [16]. Here, the authors analyze (amongst other detection strategies) the lip movements with a combined neural network structure of Mel-Frequency Cepstral Coefficients (MFCCs), Principal Coefficients (PCAs) and an RNN-based (recurrent neural network) Long Short-Term Memory (LSTM) and check whether the lip movement is synchronized to the audio signal [16,17].

The second category for DeepFake detectors, defined by Nguyen et al. [11] (i.e., the intra-frame analyses), is divided into the subcategories of deep and shallow classifiers: During the DeepFake creation process, it is necessary to warp the face area by scaling, rotation and shearing. Deep classifiers address resolution inconsistencies between the warped face area and the surrounding context. These inconsistencies are represented in artifacts which are detectable by CNNs (see, e.g., [11,18]). In contrast, the so called shallow classifiers refer to different visual feature artifacts in head pose, eyes, teeth or in facial contours. In particular, the last three features are addressed in Matern et al.'s work [19]. They solve the DeepFake detection by analyzing the eye and teeth areas for missing reflections or details as well as the texture features from the facial region [11,19].

Other survey papers in this rapidly growing research field, such as the work of, e.g., Yu et al. [20], use the main structure of the DeepFake detection method to classify these methods into several detector categories. Similar to Nguyen et al., they distinguish broadly between inter- and intra-frame analyses. In their scheme, the first (i.e., temporal) features are covered by temporal consistency-based methods using mainly network structures such as recurrent CNNs which are able to detect temporal features frame-by-frame. The latter category is addressed by general network-based methods, which are divided into transfer learning methods and specially designed networks. The methods of transfer learning re-train detectors originally trained for a different recognition problem, while specially designed networks construct and train entirely novel architectures and detectors dedicated entirely to the task of detecting DeepFake videos.

In summary of the (survey) papers discussed above, it can be stated that most DeepFake detection approaches are based on (convolutional) neural networks to learn the feature space to be used. This approach usually requires big databases of real and DeepFake videos to generate detectors that usually perform with very high detection rates on test material that is similar to the used training material in terms of its characteristics.

Hand-crafted feature methods, as an alternative to features learned with neural networks, have the benefit that they (at least theoretically) could work without training. In addition to this and other potential benefits (see Section 2.2), hand-crafted feature spaces for the detection of DeepFake videos are much less common in the literature than neural network-based approaches. Most of the existing research papers relying on hand-crafted approaches (such as [21–23]) use Support Vector Machines (SVMs) for a fast and efficient detection of DeepFake videos.

For the DeepFake detection of persons of interest (POIs) such a Barack Obama, Hillary Clinton or Donald Trump, Gu et al. [23] analyzed speech in combination with face and head movements. They followed the assumption that a person has individual facial expressions and head movements while they are speaking. Their detection pipeline starts with a single video were they tracked facial and head movements first. These facial expressions are defined by 2D and 3D facial landmark positions and several facial action units which are then used for further evaluation steps. For the DeepFake detection, they trained and tested one-class SVMs only with extracted features from authentic videos of specific POIs.

Jung et al. [24] present a hand-crafted DeepFake detector called DeepVision [24], which evaluates eye blinking behavior. In their first step, they extract the face region from a potential DeepFake video. In the following, they use an eye tracker to detect the eye area of a person. After this step they check the eye area of each frame for closed or open eyes and calculate the eye blink elapsed times and eye blink periods.

Unfortunately, the authors of some survey papers, such as [25,26], refer to learned features using specially designed networks (such as those proposed in [15,18]) and also as being "hand-crafted". This is not our perspective of hand-crafted features because they only design the neural network architectures and not the actual features and their semantics. In the following section, we will provide working definitions for the terms hand-crafted and learned features to be used in this paper.

2.2. Feature Space Design Alternatives

In pattern recognition, feature extraction starts from an initial set of input data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps. It is generally seen to be one form of dimensionality reduction projecting the input into an easier to process and (optimally) less noisy representation. In applied pattern recognition, there generally exist two distinct approaches for feature design:

(a) Features are especially designed (so-called hand-crafted) by domain experts for an application scenario in a process, which, despite the fact that it is sometimes also called intuition-based feature design, usually requires strong domain knowledge. Here, the domain expert uses his/her own experience to construct the features to encode his/her own knowledge about the semantics (and internal as well as external influence factors) inherent to the different pattern classes in the problem at hand. As a result, usually rather low-dimensional feature spaces are designed, which require only small sets of training data (or none at all) for the training (i.e., adaptation/calibration) to a specific application scenario. The semantic characteristics intrinsic to these feature spaces can easily be exploited to validate decisions made using such a feature space.
Such features can also be the result of the transfer of features from other, related or similar pattern processing problems.

(b) Feature spaces are *g* by methods such as neural networks, where a structure (or architecture) for the feature space is designed (or chosen from a set of known goods) and then labelled training specimens are used to train the network from scratch or re-train an already existing network in transfer learning. The inherent characteristic of this process is that it requires very large sets of labelled, representative data for the training of the network (a little less so in case of transfer learning). The resulting feature spaces and trained models usually lack the encoding of easily interpretable semantics.

While neural network-based methods have seen a growing popularity in the field of media forensics in the last few years, they are still burdened by the problem that the plausibility of a decision made on the basis of such features is extremely hard to verify. One of the main reasons for this is the fact that the learned features as such hardly ever encode semantics that could be interpreted by a human expert. Instead, with the help of decision validation approaches such as the expert interpretation of heatmaps using methods such as

Layer-wise Relevance Propagation (LRP; [27]), it can be shown that these methods assign meaning to regions in the input (see e.g. [28]).

For this reason, i.e., problems with the interpretability of the feature space and corresponding decisions, many application fields with sensitive tasks are hesitant to rely on learned features. A good example of a very thorough discussion of the pros and cons of hand-crafted features in contrast to those learned using convolutional neural networks can be found in Lin et al.'s work [29]. In this paper, the authors discuss this issue for specific medical data analysis problems, which, similar to forensics, is another very sensitive research field applying pattern recognition. In their work, they highlight and demonstrate with their datasets three main drawbacks of neural network-based feature space learning:

1. In the case of only small amounts of training data being available (which seems to be a problem encountered often in medical data analysis problems, including clinical studies where "*the recruitment of a large number of patients or collection of large number of images is often impeded by patient privacy, limited number of disease cases, restricted resources, funding constraints or number of participating institutions*" [29]), the classification performance of hand-crafted features (which usually show persistent detection performances with small training datasets) outperformed their feature spaces learned by neural networks. This is hardly astonishing since it is well known that CNNs require a large amount of training data for reliable imaging classification. This situation changes with increasing training dataset sizes.
2. Another advantage of hand-crafted features is interpretability. Lin et al. summarize this issue as follows: "*Therefore, interpretability of* [hand-crafted] *features reveal why liver* [magnetic resonance] *images are classified as suboptimal or adequate*" [29], i.e., these features allow for expert reasoning on errors, loss or uncertainty in decision making.
3. Feature selection strategies help learning about significance and contextual relationship for hand-crafted features, while they fail to produce interpretable results for learned features.

For the more traditional feature space designs (i.e., using hand-crafted features), the question of plausibility verification is usually easier to address. A multitude of methods for feature space analysis have been discussed in the past, including feature space-driven plausibility validation as well as model-driven validation.

Initially, there existed two main approaches for feature selection: wrapper methods, in which the features are selected using the classifier, and filter methods, in which the selection of features is independent of the classifier used. Around 2001, both main approaches were joined into a so-called hybrid method (see, e.g., [30,31]), which are usually used nowadays to analyze hand-crafted feature spaces.

2.3. A Data-Centric Examination Approach for Incident Response and Forensic Process Modeling

Forensic process models are an important cornerstone in science and more importantly the practice of forensics. They guide investigations and make them comparable, reproducible as well as certifiable. Usually, the adherence to strict guidelines (i.e., process models) is regulated within any legal system (e.g., in the US by the fourth Daubert criterion ("*the existence and maintenance of standards and controls*" [4])). For mature forensic sciences, such as, for example, dactyloscopy, internationally accepted standards (such as the ACE-V process model for dactyloscopy) have been established over recent decades.

Due to the fact that IT forensics is a rather young discipline in this field (with media forensics being an even younger subdiscipline), it is hardly astonishing that here the forensic process models have not yet achieved the same degree of maturity as in other fields. Nevertheless, they would still be important to achieve universal court acceptability of methods. One well established forensic process model for IT forensics is the one proposed by the German Federal Office for Information Security (BSI). When it was originally published in 2011, its sole focus was on computer and network forensics, but since then it has evolved to also include suite and also some extend the needs of other subdisciplines such as digitized forensics. The latest major revision of this process model, which is used within this paper, can be found in [7] and is called the Data-Centric Examination Approach (DCEA). The core of the DCEA consists of three main aspects: a model of the *phases* of a phase-driven forensic process, a classification scheme for *forensically relevant data types* and *forensic method classes*.

The DCEA phases are briefly summarized in Table 1.

Table 1. Sets of examination steps for digital forensics, as defined in [7] (updated from [32]) .

Sets of Examination Steps	Description (According to [7])
Strategic preparation (SP)	Includes measures taken by the operator of an IT system and by the forensic examiners in order to support a forensic investigation prior to an incident
Operational preparation (OP)	Includes measures of preparation for a forensic investigation after the detection of a suspected incident
Data gathering (DG)	Includes measures to acquire and secure digital evidence
Data investigation (DI)	Includes measures to evaluate and extract data for further investigation
Data analysis (DA)	Includes measures for detailed analysis and correlation between digital evidence from various sources
Documentation (DO)	Includes measures for the detailed documentation of the proceedings, also for the transformation into a different form of description for the report of the incident

One important reason for this paper to use the DCEA to model our own work is the separation of preparation steps in an investigation into two distinct phases (the strategic preparation (SP) on one hand an the operational preparation (OP) on the other). In [7], the SP is generally defined as: "*The strategic preparation* [. . .] *includes all preparation procedures taken ahead of the actual occurrence of a specific incident*". Exemplary measures for SP in the context of digital forensics are given by [7] as: "*Documentation and extension of knowledge of IT systems specifics, tool testing for forensic data types and sets of methods determination for error loss and uncertainty estimation, setup of logging capabilities, performance of system landscape analysis, data protection considerations,* [. . .]". In contrast, the OP is specified to "[. . .] *include all preparation procedures taken after of the actual occurrence of a specific incident. Those procedures by definition do not alter any data on the targeted system*". These preparation phases are then followed by the actual application of forensic procedures, separated in DCEA into the triplet of data gathering (DG), data investigation (DI) and data analysis (DA). The whole process is, in every phase (including SP and OP), supported by accompanying documentation, which is in the last phase (documentation (DO)) used as basis for the generation of the official documents regarding the investigation (e.g. the evidence to be interpreted in expert testimony in a court case).

The second core aspect of DCEA is the classification scheme for forensically relevant data types, as summarized in Table 2. The categories in this scheme are not classes in a mathematical sense, since all other later data types are interpreted out of raw data. More recent publications, such as [33], have shown that this scheme needs to be extended accordingly if new investigation domains are considered.

Table 2. Forensic data types defined in [7] (updated from [34]).

Forensic Data Type	Description (According to [7])
Raw data	A sequence of bits or data streams of system components not (yet) classified
Hardware data	Data not or only in a limited way influenced by the OS and application
Details about data	Meta data describing other data
Configuration data	Modify the behavior of the system and applications
Communication protocol data	Modify the communication behavior of the system
Process data	Data about a running process
Session data	Data collected by a system during a session
User data	Content created, edited or consumed by the user

This original set of data types, which was designed with digital (IT) forensics in mind, needs to be adapted to every investigation domain. In [7,32], such an adaptation for the field of digitized forensics has been discussed for the field of dactyloscopy (forensic fingerprint analysis and comparison). This adaptation is summarized in Table 3 below because it is much closer to the requirements we face within this paper than the original data types summarized in Table 2.

Table 3. Forensic data types defined in [7] for an exemplary selected process in digitized forensics (here, digital dactyloscopy) (updated from [32]).

Forensic Data Type	Description (According to [7])
Raw sensor data (DD1)	Digital input data from the digitalization process (e.g., scans of test samples)
Processed signal data (DD2)	Results of transformations to raw sensor data (e.g., visibility enhanced fingerprint pattern)
Contextual data (DD3)	Contain environmental data (e.g., spatial information, spatial relation between traces, temperature, humidity)
Parameter data (DD4)	Contain settings and other parameters used for acquisition, investigation and analysis
Trace characteristic feature data (DD5)	Describe trace specific investigation results (e.g., level 1/2/3 fingerprint features)
Substrate characteristic feature data (DD6)	Describe trace carrier specific investigation results (e.g., surface type, individual surface characteristics)
Model data (DD7)	Describe trained model data (e.g., surface specific scanner settings, reference data)
Classification result data (DD8)	Describes classification results gained by applying machine learning and comparable approaches
Chain of custody data (DD9)	Describe data used to ensure integrity and authenticity and process accompanying documentation (e.g., cryptographic hash sums, certificates, device identification, time stamps)
Report data (DD10)	Describe data for the process accompanying documentation and for the final report

The third core aspect of DCEA is the definition of forensic method classes as presented in Table 4. For a detailed discussion on these method classes, including considerations on the estimation of availability in certain investigation contexts, practicalities of the forensic process, etc., we refer to [7].

Table 4. Grouping of sets of methods for the forensic process in digital forensics defined in [7] (updated from [32]).

Sets of Methods for the Forensic Process in Digital Forensics	Description (According to [7])
Operating system (OS)	Methods that provide forensically relevant data as well as serving their main purpose of distributing computing resources
File system (FS)	Methods that provide forensically relevant data as well as serving their main purpose of maintaining the file system
IT application (ITA)	Methods provided by IT applications that provide forensically relevant data as well as serving their main purpose
Explicit means of intrusion detection (EMID)	Methods that are executed autonomous on a routine basis and without a suspicion of an incident
Scaling of methods for evidence gathering (SMG)	Methods that are unsuited for routine usage in a production environment (e.g., due to false positives or high computation power requirements)
Data processing and evaluation (DPE)	Dedicated methods of the forensic process that display, process or document information

The DCEA is relevant for the work presented in this paper for two different reasons: On one hand, we will use it in Section 3 to provide a comparative description of the solution concept to address the issue of DeepFake detection in this paper. On the other hand, we will elaborate on the question related to how well this process model fits the needs of media forensics investigations and which changes or extensions would be required in DCEA to provide better support for this very young subdiscipline in IT forensics.

3. Solution Concept for DeepFake Detection with Hand-Crafted Features

The main findings considering the background and state of the art in Section 2 can be summarized as follows: DeepFake detection is a very active research field trying to address a significant recent threat. While many detection approaches have been published in the last few years (some reporting astonishing detection performances), only a small number of publications have been tackling the questions of interpretability and plausibility of results. We attribute this lack of studies mainly to the type of features used in the majority of the research published so far, which rely on neural networks to learn feature spaces used, a method that has inherent difficulties with interpretability (see Section 2.2). Additionally, this question of creating the feature spaces required for a pattern recognition-driven media forensics method such as DeepFake detection, a close integration of forensic procedures and "*the existence and maintenance of standards and controls*" [4] is an open issue. This can contribute to the comparative novelty of many media forensics methods, including DeepFake detection.

To address both of these apparent gaps (interpretability of feature spaces and projection into forensic procedures), our work in this paper focuses on the usage of hand-crafted features for this pattern recognition problem as well as discussions on the applicability of the Data-Centric Examination Approach (DCEA, see Section 2.3) to map out our work. Regarding the pattern recognition aspects, the concept in this paper focuses on four items:

- The design, implementation and empirical evaluation of features for DeepFake detection: Here, two feature spaces hand-crafted especially for DeepFake detection and a hand-crafted feature space derived from a different but similar pattern recognition problem domain (face morph detection) are implemented and evaluated. For the empirical evaluation, pre-existing reference databases containing DeepFake as well as benign ("original") face video sequences are used together with a pre-existing out of the box classification algorithm implementation. To facilitate the interpretation of results and the comparability with other detector performances reported in the state of the art, different well established metrics are used: detection accuracy, Cohen's kappa as well as (ROC) AUC (Area Under the Curve (of the Receiver Operating Characteristic)).

- The discussion of different information fusion techniques and the empirical comparison with detection performances of individual classifiers: Here, with feature-level fusion and decision-level fusion, two different concepts are applied. For the latter, with the majority voting and weighted linear combination, two popular choices are used and compared with single classifiers in terms of the classification performance achieved.
- The comparison of the detection performance of our hand-crafted features with performances of learned feature spaces from the state of the art in this field: Here, the results obtained by single classifiers as well as fusion approaches are compared in terms of detection accuracy with different approaches from the state of the art, relying on learned features.
- Attempts at validating the detectors' decisions on basis of the features and trained models: Some classifiers, such as the decision tree algorithm used in this paper, train models that can be read, interpreted and compared by humans. Here, we analyze the decision trees trained on different training sets to identify the most relevant features and see how much these trees have in common and where they differ.

In addition to these pattern recognition aspects, we project the different operational aspects in training, validating and applying the DeepFake detectors into the established process model DCEA to show how such media forensics methods would have to be integrated into forensic procedures. In this projection, the first question to be asked concerns where the detector is supposed to be used. There exist two potential operation points in the phases described by the DCEA: either as a method of explicit means of intrusion detection (EMID) as part of incident detection mechanisms, which would place the whole DeepFake detection with the training of the method and its application into the phase of strategic preparation (SP), or in scaling of methods for evidence gathering (SMG), which would place DeepFake detection after an incident is detected or suspected and place the corresponding components in the operational preparation (OP), data gathering (DG), data investigation (DI) and data analysis (DA) phases. These two distinct operation points as a live detector or as means of post-mortem (or a posterior) analysis in data investigation have, amongst other effects, significant impact on the training scenario that can be assumed: In the case of application as an live detector (EMID), in SP, only pre-trained models can be applied. In the case of a post-mortem (SMG) detector, in the OP the material to be investigated can be analyzed to design targeted training datasets perfectly matching the characteristics encountered. Using those sets (and own DeepFake algorithms to also generate a specimen for this class) optimal models could be trained for each case. In this paper, the conceptual choice made is that of a live detector, reserving considerations on targeted training for future work.

The concept of training brings us to a second issue where the principles of the DCEA can help structuring of the description of media forensics methods such as DeepFake detectors: The accompanying documentation in the DCEA is meant to allow for interpretability and plausibility validation steps while compiling the case documentation in DO. For our work, this implies not only documenting all details of the pattern recognition process at hand but also using these data to reason about the plausibility of decisions (e.g., by comparing the characteristics of training and test sets to determine questions of generalization power).

One important realization when trying to apply the DCEA data types for digital or digitized forensics, as summarized in Tables 2 and 3, is that they do not perfectly match the media forensics task at hand. Using the original model for digital forensics, only four of the data types would be covered (raw data differentiated into different user data media streams (video, audio, network stream) and possibly hardware data (derived from the camera/microphone used) as well as details about data). If the model for digitized dactyloscopy is used, which slightly better matches with the characteristics of our application scenario, then eight of the ten data types would be directly relevant (processed signal data (DD2), contextual data (DD3), parameter data (DD4), trace characteristic feature data (DD5), model data (DD7), classification result data (DD8), chain of custody data (DD9) and report data (DD10)), while one other would very likely also to be of significance (raw sensor data (DD1), which might be used to calibrate specific cameras or camera models, etc.).

It is apparent that an adapted data type model for media forensics would be required to be able to make use of the full potential of the DCEA in this context. Nevertheless, it is outside the scope of this paper to propose such an adapted data type model.

4. Implementation of the Individual Detectors and the Fusion Operators

For our DeepFake detection methods, the input video is evaluated frame-wise with the intention to analyze inter-frame patterns (e.g., the time between two blinks of one eye). In a pre-processing step, the presence of a face in a frame is determined, the face region is segmented and annotated frame-wise with a semantic model localizing 68 facial landmarks. This semantic model [35] is provided by the dlib library [36]. The output of this pre-processing is shown in Figure 1.

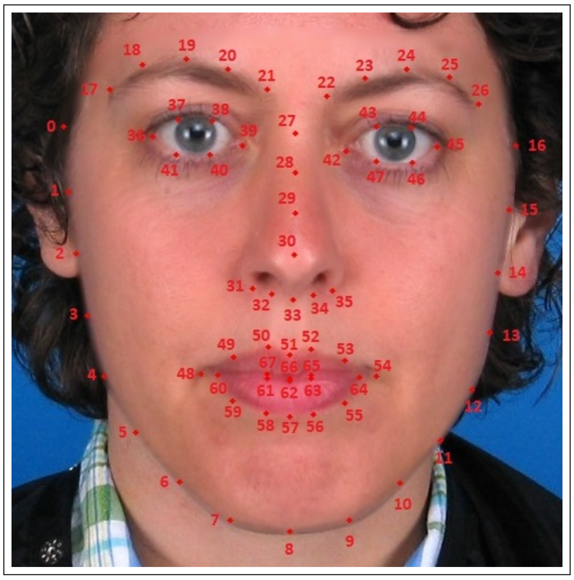


Figure 1. Visual representation of the 68 facial landmark model [35]. Image originates from Utrecht ECVP [37] with application of keypoints generation by dlib [36] followed by cropping.

In case no face can be localized in a frame, this event is logged, if a face is found, and the segmented face pixel matrix and the positions of these 68 facial landmarks are then forwarded to the feature extraction component of each individual detector as well as the concatenation operator for the feature-level fusion. This process is repeated frame-wise until the end of the video is reached, which initializes the detection operations performed. The entire processing sequence is shown in Figure 2. Due to the specific recording conditions of the datasets used in this paper (which all represent a single person in an ideal interview-like recording setting with perfectly illuminated faces and none of the facial key regions, such as eye and mouth, occluded), the pre-processing could be kept that simple. In case more realistic/averse videos have to be analyzed, this pre-processing would necessarily have to be extended.

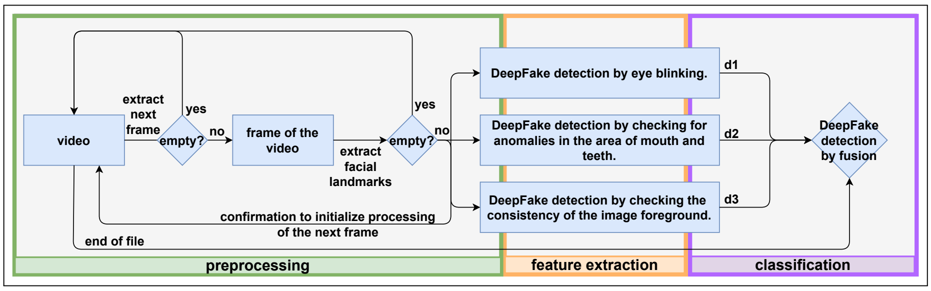

video
extract next frame
yes
empty?
no
frame of the video
extract facial landmarks
yes
empty?
no
DeepFake detection by eye blinking.
DeepFake detection by checking for anomalies in the area of mouth and teeth.
DeepFake detection by checking the consistency of the image foreground.
d1
d2
d3
DeepFake detection by fusion
confirmation to initialize processing of the next frame
end of file
preprocessing
feature extraction
classification

Figure 2. Concept pipeline considered in this paper.

The domain knowledge used here in hand-crafting features for DeepFake detection is based on the fact that DeepFake generators (similar to face morphing algorithms) rely on blending operations in the face region, which is a well established fact in the state-of-the-art research in this field [13]. Blending itself describes the process of a weighted combination of two or more faces to create a new identity [38]. This often goes hand in hand with a loss of local details in the face regions, while the background of a video or image is usually not affected, which is a fact also used in similar media forensics detectors such as, e.g., morphing attack detectors [39].

This knowledge is translated in Section 4.1 into three distinct hand-crafted feature spaces aiming at solving the following pattern recognition tasks to distinguish between DeepFake and genuine videos: (a) anomaly detection for eye blinking (Section 4.1.1), (b) anomaly detection in mouth and teeth region level of detail (Section 4.1.3), and (c) DeepFake detection based on image foreground texture (Section 4.1.3). In terms of the DCEA data type model, these features would make up the Trace characteristic feature data (DD5) from the data model discussed in [7] for digitized forensics. While the broad category actually fits, the extensive discussion on feature space design alternatives for DeepFake detection presented in Section 2.2 indicates that more detailed modeling would be required to sufficiently address this aspect.

To implement the actual classification, we decided not to design or implement our own but instead rely on a proven classification algorithm detection which does facilitate feature space as well as model-driven plausibility considerations. The actual algorithm that we use here is the WEKAs [40] J48 decision tree, which is an open source implementation of Ross Quinlans C4.5 decision tree algorithm [41]. The classifier is used here in its default parameterization, i.e., without parameter optimization being applied.

To further increase the performance and robustness of DeepFake detection, different fusion operators for feature-level fusion and decision-level fusion are implemented, as shown in Section 4.2.

In terms of datasets (i.e., processed signal data (DD2)), the pre-existing, publicly available and widely accepted reference datasets TIMIT-DF [16,42], FaceForensics++ [12,43,44] and Celeb-DF [13] are used in our evaluations. VidTIMIT [42], which was used to create TIMIT-DF [16], is a long-established reference database for various video processing tasks. It represents recording criteria that are ideal for face recognition and similar tasks: uniform lighting, the presence of exactly one person in each video, a frontal position to the camera, an average duration of 3 to 5 s and the speaking of ten different, pre-defined sentences. A total of 430 videos are included in the set, recorded using 43 volunteers. The resulting DeepFake videos were generated for TIMIT-DF by face swapping in two different resolutions with the autoencoder resolutions 64 × 64 and 128 × 128, respectively. Through prior selection, 16 suitable pairs of faces were selected for the generation, resulting in 32 DeepFake entities. This yields a total of 640 DeepFakes, which were taken into account in the TIMIT-DF dataset [16].

The second dataset considered is called DeepFakeDetection (DFD) [44], which originates from the FaceForensics++ [12] dataset. It contains a total of 363 source videos based on 28 actors (*DFD-source*). DeepFake synthesis was performed with an autoencoder resolution of 256 × 256 pixels and a total of 3068 DeepFake videos (*DFD-DF*) were generated. All videos considered were compressed with H.264 at CRF 23. Due to time constraints, only a subset of the DFD dataset, containing 55 DFD-source and 55 DFD-DF videos, were used. Video selection was carried out manually, selecting videos in which only a single person can be found speaking towards the camera. In the DFD dataset, this was carried out by searching for the keyword *talking* in conjunction with *against wall* or *outside*.

The third dataset is Celeb-DF [13], which includes videos (harvested from YouTube) of celebrities being interviewed. These source videos were divided in [13] into the two datasets *Celeb-YouTube* and *Celeb-real*, whereby only *Celeb-real* was considered for the DeepFake synthesis. The synthesis method is more advanced than the one from TIMIT-DF in terms of quality, using an autoencoder resolution of 256 × 256. Due to an average video duration of about 13 to 15 s, only a subset of this dataset is used in our own paper. For our evaluations, 120 source and 120 DeepFake videos were chosen. For simplification, the entire dataset is subsequently also referred to as *Celeb-DF*.

Those three datasets, summarized in Table 5 were used to design different training and testing scenarios to be able to establish facts about the generalization power of the detectors trained, which is an important aspect of the quality assessment for every method. Such evaluations would have to be performed as part of quality assurance in the strategic preparation (SP) phase of each forensic process.

Table 5. Collection of datasets used for this paper.

Dataset	Number of Videos	Reference
VidTIMIT	430 *	[42]
TIMIT-DF	640	[16,42]
DFD-source	55 *	[12,44]
DFD-DF	55 *	[12,44]
Celeb-YouTube	60 *	[13]
Celeb-real	60 *	[13]
Celeb-DF (v2)	120 *	[13]

*: Numbers do not reflect the total but rather the number of videos used in the context of this work.

4.1. Individual Detectors Using Hand-Crafted Features

In general, the 68 facial landmark model [35] used in this paper (see Section 4) can be structured into different facial areas, as shown in Figure 1. Here, the following segmentation alternatives are used to derive the features for our individual detectors: The first set of keypoints, numbers 0 to 26, describes the edges of the face along the chin and eyebrows. These keypoints are used to segment the image foreground, as explained in Section 4.1.3. Keypoints 27 to 35 describe the nose, which is neglected in this work. The eyes are described with the help of keypoints 36 to 47 and form the basis for the detection of blinking behavior considered in Section 4.1.1. The final keypoints, 48 to 67, are used to model the mouth, which is examined in more detail in Section 4.1.2.

In the following subsections, our three distinct detectors relying on different hand-crafted features spaces are described. A summarizing overview over all features extracted is presented in Table A1 at the end of the document in Appendix A.

4.1.1. DeepFake Detection Based on Eye Blinking

The first implemented detector is based on the biometric modality eye and acts on the behavior of eye blinking. Using the 68 facial landmark model [35], each eye is described by six keypoints (keypoints 36 to 41 and 42 to 47, respectively). The process of blinking itself occurs subconsciously about 10 to 15 times per minute. On average one blink takes 0.3 to 0.4 s between closing and reopening the eyes. It should be noted that blinking behavior is also influenced by gender, age, time of day and how tired the person is [24]. In some publications, the minimum duration of human blinking is noted as 0.1 s [45]. To enable the detection of blinking, the eyes are modeled to two possible states—*open* and *closed*. To distinguish between these two states, the degree of aperture for each eye is determined individually by the formula:

$$AspectRatio = \frac{y_{Max} - y_{Min}}{x_{Max} - x_{Min}}$$

The parameters of this bounding box are determined from the six keypoints of the 68 facial landmark model, which describe the respective eye. The main idea of the feature design here is strong likeliness of DeepFake synthesis artifacts leading to lower average AspectRatio values, due to the inherent impact of the blending operation. Considering diversity in eye shapes and the inclusion of emotions, as shown in Figure 3, results on the use of a dynamic threshold (determined empirically on the training data used) were used to distinguish the eye states.

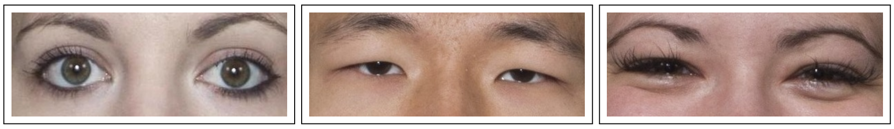


Figure 3. Illustration of the challenges of correctly detecting the aperture of eye opening as widely open (**left**), based on ethnicity (**center**) and inclusion of emotions (**right**). Images originate from LondonDB [46] dataset with application of cropping.

The eye state classification is carried out as binary decision, under the assumption that the aspect ratio always represents exactly one of two values, representing the two eye states (open and closed). The threshold under consideration was implemented as a bimodal distance function. Here, both states are described by a value which corresponds to the most frequent value of the upper and lower thirds of values found in the training data. The *closed* state is described by the most frequent value of the lower third of the value range. Conversely, *open* is described by the most frequent value, which is found in the upper third of the value range. Subsequently, the state for each eye and frame is determined via smaller distance to one of the two values representing the states.

For DeepFake classification based on eye blinking, a feature vector of fixed size of 13 dimensions was designed. Seven out of these 13 features are directly based on the AspectRatio, one is based on the difference between the two eyes and the other six are based on eyelid movements. This eyelid movement is detected as a rate of change on a frame-by-frame basis. Features 8 to 13 are based on the given eye state modeling. One feature introduces a new metric of anomaly, hereinafter referred to as *noise*. This noise is described as a frequent change in eye states below the expected frequency. In detail, this timespan is set to 0.05 s and thus corresponds to half the duration of a blink to detect anomalies only. Another feature describes the percentage of time in the video that the person has their eyes open. The last four features considered refer to the extreme values given the duration in each eye state.

In the summarizing overview of all features in this paper, given in Table A1 at the end of the document in Appendix A, these eye blinking features are the first 13 feature vector elements.

4.1.2. DeepFake Detection Based on Mouth Region

The second implemented detector is based on the biometric modality lip-movement. The focus of this approach is on analyzing the highly detailed teeth region. Under the assumption of blending as part of DeepFake creation, a blurred, less detailed image of the teeth is expected. The 68 facial landmark model is also used to localize the mouth region by using keypoints 48 to 67. These keypoints allow the mouth to be displayed as two separate images, one of which represents the entire mouth described by keypoints 48 to 59. This representation is henceforth called the OuterBoundRegion (OBR). The other keypoints (60 to 67) can be used show another representation considered in this work. This, in the following, is called the InnerBoundRegion (IBR) and represents the mouth area with the exception of the lips. The IBR is used to determine whether the mouth is open, since a closed mouth can be represented by a non-existent IBR. The third and last representation considered to describe the mouth region is the so-called TeethRegion (TR). The TR is created by segmenting the OBR to preserve potential teeth found in the image. An example of the representations can be found in Figure 4. In addition, the degree of aperture of the mouth is determined as an additional parameter based on the OBR. Here, the x and y dimensions are considered separately in order to act independently of the spoken phoneme. The respective values are determined by the bounding box of the OBR using $Aperture_x = x_{Max} - x_{Min}$ and $Aperture_y = y_{Max} - y_{Min}$ for each frame.

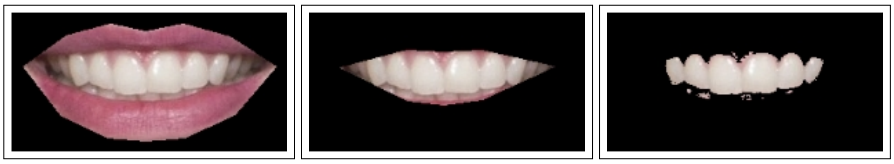


Figure 4. Illustration of the proposed representations for the mouth region OBR (**left**), IBR (**center**) and TR (**right**). Mouth image originates from LondonDB [46] dataset with application of keypoint generation by dlib [36], segmentation and cropping.

Based on these representations, a total of three states are conceived to describe the mouth. These states are: *closed mouth*, *open mouth without detectable teeth* and *open mouth with detectable teeth*. The subdivision of the states is made by two binary decisions. The first decision is based on the IBR and describes whether the mouth is open. The metric used for the decision is the number of pixels found in the IBR. Here, a conscious decision is made against cropping and scaling of the representations in order to prevent distortion of the image when viewing different visemes [47]. As a consequence, the number of pixels of the OBR is taken as a reference. Thus, the decision threshold is determined empirically on training data as: $\frac{PixelCount_{IBR}}{PixelCount_{OBR}} > 0.211137$, for criteria for an open mouth. The second decision, if the mouth is classified as open, is made with the help of the number of pixels in the TR, once again using the OBR as a reference. The threshold considered (after empirical determination from training data) is $\frac{PixelCount_{TR}}{PixelCount_{OBR}} > 0.11455$ for detectable teeth. An example of each state considered can be found in Figure 5.

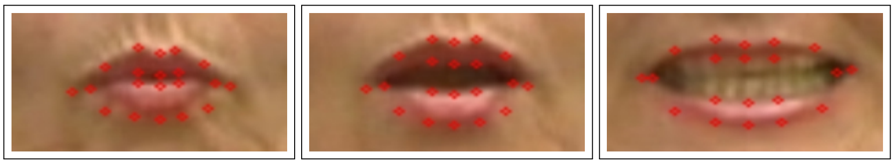


Figure 5. Illustration of the proposed mouth states: closed (**left**), open without detectable teeth (**center**) and open with detectable teeth (**right**). Image originates from VidTIMIT [42] dataset with application of keypoint generation by dlib [36] and cropping.

For the detection of DeepFakes based on mouth region, a feature vector of dimensionality 16 is designed. Six of these features are based on mouth movements. This mouth movement is recognized image-wise as the rate of change, which corresponds to the extreme values for the x- and y-dimensions, respectively. The other 10 features are based on the detected mouth states, leaving out the *closed mouth* state. Thus, the focus of this review is based on the description of the level of detail in the mouth region. For this purpose, FAST and SIFT keypoint detectors as well as Sobel edge detection and the number of closed regions are considered. All of them are implemented by OpenCV [48] and used with default parameters. For the *open without teeth* state, the maximum of each feature, and for state *open with teeth*, the minimum of each feature are determined over all frames. Lastly, the percentage of time in both states is considered. The expectation for this approach is a low level of detail in the *open with teeth* state for DeepFakes or even a wrong assignment to the *open without teeth* state, although teeth are recognisable, due to blending of artifacts.

In the summarizing overview over all features in this paper, given in Table A1 at the end of the document in Appendix A, these mouth movement features are elements 14 to 29 of the feature vector.

4.1.3. DeepFake Detection Based on Image Foreground

The third and last proposed detector is based on domain transfer of hand-crafted features from a similar media forensics task. As shown by Kraetzer et al. [39], such a domain transfer seems plausible to detect blending anomalies in face morph attack detection. This requires an image foreground, which is characterized by a uniform distance towards the camera. Image foreground $Img_{Foreground}$ is designed as an extension of the facial region Img_{Face}, which is determined based on the 68 facial landmark model—more precisely, keypoints 0 to 26. The extension of the facial region is carried out by widening along the vertical axis to include the upper body, which is potentially shown in the image. A third representation, called Img_{ROI}, is conceived as the differential image of the previous two, formally described as $Img_{ROI} = Img_{Foreground} - Img_{Face}$. A visual example of each representation can be found in Figure 6.

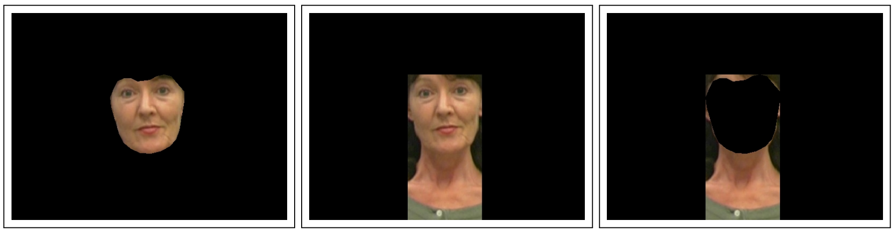


Figure 6. Illustration of the proposed representations for anomaly detection based on image foreground Img_{Face} (**right**). Image originates from VidTIMIT [42] dataset with application of keypoint generation by dlib [36] and segmentation.

For the detection of DeepFakes based on the image foreground, a feature vector of fixed size of eight elements was designed. The first subset of features is based on face detection itself, counting the number of frames and sequences where no face can be found. Here, it is assumed that a failure is due to anomalies of the DeepFake synthesis. The second set of features is based on the level of detail in Img_{Face} relative to Img_{ROI}. For each frame and representation, the characteristics of FAST and SIFT keypoints as well as the Sobel edge image are determined. The implementation of these metrics is carried out using the default parameters given by OpenCV [48] and the scoring for each frame corresponds to $\frac{Img_{Face}}{Img_{ROI}}$. Lower values for DeepFakes are expected here. Lastly, the respective extreme values of all frames are extracted as features.

In the summarizing overview of all the features in this paper, given in Table A1 at the end of the document in Appendix A, these features are elements 30 to 37 of the feature vector.

4.2. Fusion Operators

To further increase the performance as well as robustness of the detection, different methods of fusion were implemented for our evaluation. The fusion itself is considered here both at *feature level* and *decision level* [49]. At the feature level, the feature spaces of the individual detectors are concatenated, without additional pre-processing such as weighting or filtering. At the decision level, a total of four operators are applied: The first operator makes an unbiased decision using *simple majority voting* [50]. In contrast, the other three operators implement weighted linear combinations and derive the weights for each individual detector based on its classification performance on the training set. Considering the different training scenarios, there are two sets of weights, each based on the training using dataset TIMIT-DF [16,42], DFD ([12,44]) or Celeb-DF [13]. The explicit weights determined this way can be found in Section 5.2. In summary, the following five fusion operators are considered:

1. Feature-level fusion: concatenation of all features;
2. Decision-level fusion: simple majority voting;
3. Decision-level fusion: weighted, based on accuracy using TIMIT-DF for training;
4. Decision-level fusion: weighted, based on accuracy using DFD for training;
5. Decision-level fusion: weighted, based on accuracy using Celeb-DF for training.

5. Evaluation Results

The evaluation of the created approaches (i.e., our three feature spaces used in training and testing with the used J48 classifier) for DeepFake detection is looking into aspects of *performance*, *generalizability* and *plausibility* of the decisions made (i.e., the kind of information summarized in the DCEA data type model for digitized forensics as Classification result data (DD8)). To address *performance* and *generalizability*, the three datasets used for training and testing are presented as different scenarios (as shown in Table 6). Scenarios S1, S5 and S9 , which represent evaluations in a simplistic (i.e., very naive) setup, split one dataset in disjointed training and test subsets. These three scenarios are used to validate the *performance* of detectors under optimal conditions.

In contrast, for evaluations on the *generalizability*, separate training and testing datasets are used in scenarios S2, S3, S4, S6, S7 and S8. Since the individual detectors classify binary according to *{DeepFake, OK}*, the evaluation is carried out using the metrics' true positive rate (TPR; a true positive (TP) in our case being a DeepFake detected as a *DeepFake*), true negative rate (TNR; a true negative (TN) being an unmodified video classified as *OK*), accuracy and Cohen's kappa (κ).

Table 6. Representation of the considered training and testing scenarios, given by differentiation of the training and testing datasets used.

↓ **Training/Testing** →	**TIMIT-DF**	**DFD**	**Celeb-DF**
TIMIT-DF	scenario 1 (S1)	scenario 2 (S2)	scenario 3 (S3)
DFD	scenario 4 (S4)	scenario 5 (S5)	scenario 6 (S6)
Celeb-DF	scenario 7 (S7)	scenario 8 (S8)	scenario 9 (S9)

In addition, the hand-crafted features are evaluated in terms of *interpretability* and *relevance*. This is carried out by manually evaluating the trained decision trees in model-driven decision validation, looking at the individual features used to make the decision, the threshold used, and their distance from the root node. To support this analysis, the complete list of all features and experts' assumptions about their content behavior can be found in Table A1 at the end of the document in Appendix A. To extend the initial model-driven decision validation, a comparison of the three decision trees trained on the different datasets, TIMIT-DF, DFD and Celeb-DF, is made.

5.1. Results for Individual Detectors

The detection approach based on *blink behavior* has a generally higher TPR than TNR, regardless of the scenario considered. For S1, it has a TNR of 70.47% and a TPR of 90.94%, resulting in an accuracy of 82.15% and κ of 0.6306. In comparison, S9 shows a TNR of 63.33% and TPR of 75.00%, resulting in an accuracy of 69.17% and κ of 0.3833. It is assumed that the Celeb-DF dataset also represents an improvement of the DeepFake synthesis over the older TIMIT-DF by incorporating more realistic blinking behavior. Considering the generalizability, a drastic decrease in detection rates can be seen in S7, S8 and S9, with a tendency to label all videos as DeepFake. In numbers, S3 indicates a TNR of 33.33% and TPR of 75.00%, with an accuracy of 54.17% and κ of 0.0833. In comparison, S7 shows a TNR of 6.05% and TPR of 99.53%, resulting in an accuracy of 61.96% and κ of 0.0659. By performing feature selection on the 13 features considered, only the eyelid movement-based features ($\text{ID2}_{\text{blink}}$ to $\text{ID7}_{\text{blink}}$) seem suitable. In addition, looking at the two eyes separately shows added value. As a result of the model-driven comparison of both trained decision trees, a DeepFake can be described by a higher difference between opening and closing speeds, relative to a non-manipulated video. However, the ranges of the values found as well as the associated thresholds are different for the TIMIT-DF and Celeb-DF datasets, explaining the drastic performance decrease for S3 and S7. Training on the DFD dataset shows only the use of features $\text{ID9}_{\text{blink}}$ and $\text{ID10}_{\text{blink}}$ for decision making.

The second detection approach considered, based on the *mouth region*, has the highest individual classification performances. For S1, a TNR of 88.84%, TPR of 97.81%, accuracy of 94.21% and κ of 0.8779 was achieved. In contrast, S9 resulted in a TNR of 91.67%, TPR of 97.50%, accuracy of 94.58% and κ of 0.8917, thus showing better results in direct comparison. Based on this result, it is suspected that newer DeepFake generators, such as the one used to create Celeb-DF, also exhibit said blending artifacts. Once again, there are clear losses in generalizability for S3 and S7: For S3, a TNR of 40.83%, TPR of 72.50%, accuracy of 56.67% and κ of 0.1333 were observed. S7 shows slightly better results with a TNR of 63.02%, TPR of 71.09%, accuracy of 67.85% and κ of 0.3378, which are justified by more general inclusion conditions of the Celeb-DF data and more general classification model. Based on the 16 features considered in feature selection, the set of features describing the grade of detail, excluding the ones using Sobel operator, are used to classify a DeepFake. This clearly shows that blending results in a loss of detail in the facial region, which can be found for both states *open without teeth* and *open with teeth*. Additionally, the assumption that the state *open with teeth* is found less frequently for DeepFakes is correct. However, it should be noted here that the approach only works if an open mouth can be found—for example, if a person is speaking.

The trend of high TPR at the expense of TNR is also emerging for the *detector based on the image foreground*. For S1, a TNR of 52.33%, TPR of 87.50%, accuracy of 73.36% and κ of 0.4182 were observed. For S9, the results look similar, with a TNR of 56.67%, TPR of 85.00%, accuracy of 70.83% and κ of 0.4167. This approach also shows poor generalizability, with a TNR of 43.33%, TPR of 70.00%, accuracy of 56.67% and κ of 0.1333 for S3. Lastly, S7 shows a TNR of 32.79%, TPR of 79.38%, accuracy of 60.83% and κ of 0.1297. For the decision making itself, the features based on the level of detail except for the Sobel operator, as well as the number of frames without a found face, are used. However, the $\text{ID1}_{\text{foreground}}$ shows a different classification strategy depending on the dataset considered, when at least one frame without a face is found. While for TIMIT-DF and DFD it is interpreted as a DeepFake, for Celeb-DF it serves the classification OK. It is suspected that for TIMIT-DF and DFD, the synthesis may result in artifacts, making the face undetectable. On the other hand, less strict recording conditions in Celeb-DF do not exclude side shots that cannot be detected by the facial landmark model. The use of features $\text{ID3}_{\text{foreground}}$ to $\text{ID6}_{\text{foreground}}$ corresponds to the assumptions about blending, whereby lower levels of detail are taken as an indication of a DeepFake.

In conclusion, regardless of the detection approach considered, in all cases, a value for Cohen's kappa > 0 was obtained, implying for all cases a detector performance better than chance agreement (i.e., better than guessing). Nevertheless, it has to be admitted that the differences between the more naive setups (S1 and S9 with $\kappa > 0.35$) and the more realistic setups (S3 and S7 with $\kappa < 0.15$ for all but one case) indicate a very limited generalization power of the trained detectors.

Analyzing the trained models in more detail, it has to be highlighted that the decision tree trained on Celeb-DF is shown to be smaller and more compact. This is justified by a lower number of suitable features for the detection of higher quality DeepFakes. In addition, S3 generalizes better than S7, which goes hand in hand with the preceding statement. Here, Celeb-DF represents a more general dataset, with fewer indicators of DeepFakes, where the trained model applies better to TIMIT-DF than vice versa.

5.2. Results for Fusion Operators

For all fusion operators considered, the metrics TPR, TNR, accuracy and Cohen's kappa are used to allow comparability between fusion and individual detectors. In addition, the *receiver operating characteristic* (ROC) for all scenarios considered, based on the different approaches of fusion at the decision level, are determined. The resulting graphs can be found in Figure 7. Based on the ROC, the *area under curve* (AUC) is determined in order to realize a better comparison with research results in the state of the art in the literature.

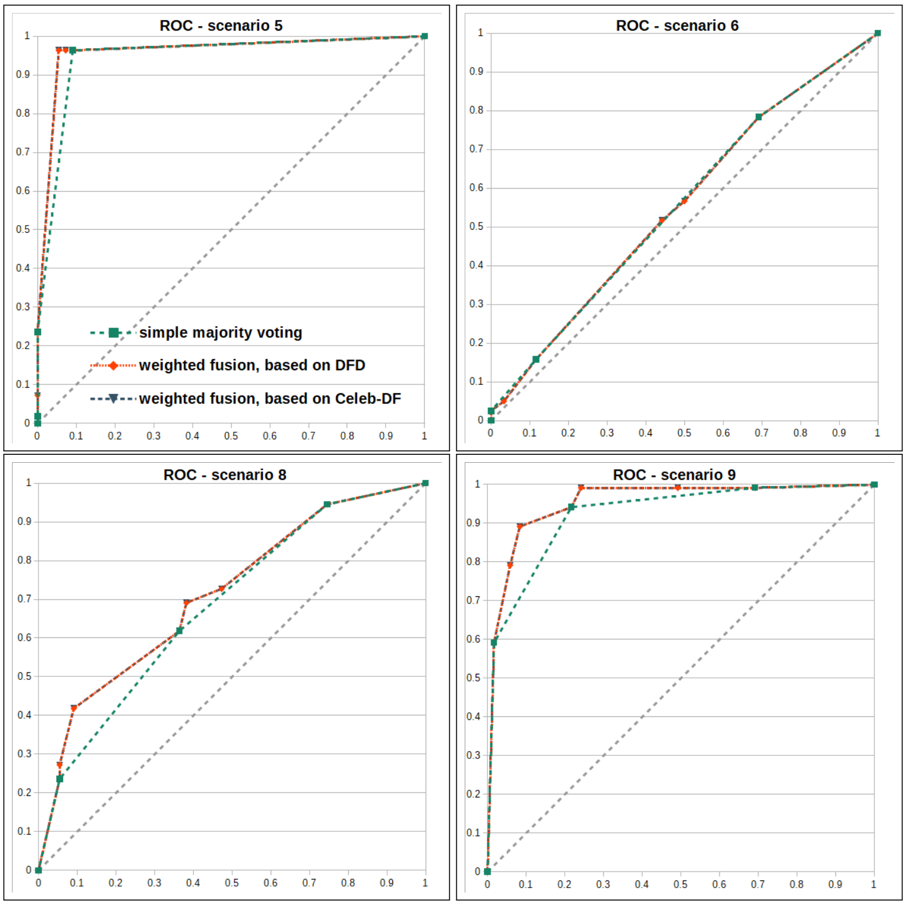


Figure 7. Receiver operation curves (ROCs) for the decision-level fusion methods simple majority voting and weighted fusion, based on DFD and Celeb-DF. Scenarios S5, S6, S8 and S9, which consider both the DFD and Celeb-DF datasets, are presented here. The false alarm rate (false positive rate) is plotted on the x-axis. The sensitivity (true positive rate) is plotted on the y-axis.

The first fusion approach considered is carried out at the *feature level* by concatenating all features without prior adjustments or filtering. A descriptor of this vector can be found in Table A1. For S1, a TNR of 96.74%, TPR of 98.13%, accuracy of 97.57% and κ of 0.9494 and for S9 a TNR of 92.50%, TPR of 95.83%, accuracy of 94.17% and κ of 0.8833 are achieved. This outperforms the best individual detector from Section 5.1. However, this performance is accompanied by even more significant losses for generalizability seen for S3 and S7: A TNR of 70.83%, TPR of 38.33%, accuracy of 54.58% and κ of 0.0917 are achieved for S3 and a TNR of 63.02%, TPR of 64.22%, accuracy of 63.74% and κ of 0.2653 are achieved for S7.

The model-driven feature selection shows that mainly features of the mouth region are used here. From the other two feature spaces, only $\text{ID2}_{\text{blink}}$ and $\text{ID6}_{\text{foreground}}$ are considered (the latter is found in the root of the respective decision trees). This again implies that the individual features based on blinking and image foreground appear more unsuitable than the features based on the mouth region. In addition, the differences between the performances on the datasets and corresponding differences in threshold determination described at the end of Section 5.1 are again apparent.

The second approach of the fusion operators takes place at *decision-level* in the form of *simple majority voting*. Here, detection rates of TNR of 79.53%, TPR of 98.75%, accuracy of 91.03% and κ of 0.8075 for S1 and TNR of 78.33%, TPR of 94.17%, accuracy of 86.25% and κ of 0.7250 for S9 are determined. Furthermore, simple majority voting shows the best generalizability of all approaches for S3, with a TNR of 53.33%, TPR of 64.17%, accuracy of 58.75% and κ of 0.1750. A TNR of 26.74%, TPR of 91.41%, accuracy of 65.42% and κ of 0.2015 are determined for S7.

For the considered weighted *decision-level* fusion approaches, the weight combinations $w_{\text{blink}} = 0.328967$, $w_{\text{mouth}} = 0.377246$ and $w_{\text{foreground}} = 0.293787$ based on the use of TIMIT-DF for training, $w_{\text{blink}} = 0.257934$, $w_{\text{mouth}} = 0.420621$ and $w_{\text{foreground}} = 0.321445$ based on the use of DFD as well as $w_{\text{blink}} = 0.294849$, $w_{\text{mouth}} = 0.403197$ and $w_{\text{foreground}} = 0.301954$ based on the use of Celeb-DF for training are derived based on the determined detection performances in training. In addition, the optimal threshold value for the classification is determined manually. For both cases, the ideal threshold can be described as:

$$w_{\text{blink}} + w_{\text{foreground}} < \text{threshold} < w_{\text{mouth}} + w_{\text{blink}\,|\,\text{foreground}}$$

It is therefore necessary that both the detector based on the mouth region and another one arrive at the classification result *DeepFake* so that the fusion also arrives at that conclusion. In the following, a threshold value of 0.65 is used. Considering the results, these resemble the detector based on the mouth region and show S1 with a TNR of 91.40%, TPR of 97.03%, accuracy of 94.77% and κ of 0.8904, as well as a TNR of 91.67%, TPR of 94.17%, accuracy of 90.42% and κ of 0.8083 for S9. In the context of generalizability, this fusion approach for S3 shows a TNR of 59.17%, TPR of 55.83%, accuracy of 57.50% and κ of 0.1500. Scenario S7 has a TNR of 63.72%, TPR of 70.94%, accuracy of 68.04% and κ of 0.3427 are determined, representing the best results of all considered implementations for S7. A marginal improvement of the weights based on the Celeb-DF can be found in consideration of the ROC AUC, as shown in Figure 7.

In conclusion, previous trends are confirmed showing that S7 has a higher performance than S3 and thus more refined DeepFakes and less limiting factors of acquisition are necessary for a more accurate classifier.

Table 7 summarizes and compares the performances of the individual and fusion detectors. While the best performances are very similar, the fusion-based approaches show a much smaller range in their results, which implies that the strongest of the three single detectors (using the mouth region features) has a dominating impact out of all three fusion operators tested. By switching from single classifiers to fusion approaches, here no gain could be made in terms of increasing generalization power. The reason has to be sought in the different thresholds that were derived for both training sets (see the corresponding discussion at the end Section 5.1).

Table 7. Classification results based on accuracy in percent, followed by Cohen's kappa in parenthesis, for the different methods proposed in this paper. Best result for each combination of training and test data is highlighted bold.

Training Dataset →	**TIMIT-DF** [16,42]			**DFD** [12,44]			**Celeb-DF** [13]		
↓ proposed method test dataset →	TIMIT-DF	DFD	Celeb-DF	TIMIT-DF	DFD	Celeb-DF	TIMIT-DF	DFD	Celeb-DF
DeepFake detection based on eye blinking	82.15% (0.63)	50.00% (0.00)	**57.50% (0.15)**	58.32% (0.15)	59.09% (0.18)	52.92% (0.06)	62.06% (0.07)	58.18% (0.16)	69.17% (0.38)
DeepFake detection based on mouth region	**94.21% (0.88)**	**76.36% (0.53)**	56.67% (0.13)	**64.95% (0.29)**	**96.36% (0.93)**	53.75% (0.08)	**67.85% (0.34)**	**69.09% (0.38)**	**94.58% (0.89)**
DeepFake detection based on image foreground	73.36% (0.42)	53.64% (0.07)	56.67% (0.13)	58.33% (0.17)	73.64% (0.47)	**54.02% (0.11)**	60.83% (0.13)	54.55% (0.09)	70.83% (0.42)
Feature-level fusion	**97.57% (0.95)**	66.36% (0.33)	54.58% (0.09)	65.05% (0.30)	**97.27% (0.95)**	**56.25% (0.13)**	63.74% (0.27)	60.00% (0.20)	**94.17% (0.88)**
Decision-level fusion: simple majority voting	91.03% (0.81)	69.09% (0.38)	**58.75% (0.18)**	59.72% (0.24)	61.18% (0.24)	52.08% (0.04)	65.42% (0.20)	62.73% (0.25)	86.25% (0.73)
Decision-level fusion: weighted (threshold=0.65)	94.77% (0.89)	**70.91% (0.42)**	57.50% (0.15)	**67.00% (0.33)**	95.45% (0.91)	53.75% (0.08)	**68.04% (0.34)**	**65.45% (0.31)**	90.42% (0.81)

6. Summary and Conclusions

To allow for a direct comparison of hand-crafted and learned features, Section 6.1 discusses our obtained performances and the generalization behavior observed in direct comparison with a state-of-the-art paper using deep learning under comparable evaluation conditions. Furthermore, we compare our feature concept implementations for *eye blinking*, *mouth region* and *foreground texture analysis* with other hand-crafted and learned features considering the same facial regions.

In Section 6.2, we summarize our conclusions on the comparison of hand-crafted and learned features for DeepFake detection.

6.1. Summary of the Results and Comparison with other Approaches from the State of the Art

In the sections below, we provide a comparison of the results obtained in our experiments with selected work from the state of the art in this fast growing research field. Section 2.1 shows that there exists a wide range of different approaches to distinguish DeepFake from real videos, with a strong tendency towards relying on features learned by using neural networks. In subsection 6.1.1, we compare our results with selected detection performances and generalization behaviors observed in the state of the art. In Section 6.1.2, we compare our concepts for feature designs (looking at hand-crafted features, especially for eye blinking, mouth region and image foreground) with similar approaches by other authors.

6.1.1. Performances and Generalization Power

Table 8 consists of two parts, the upper half represents our results on fusion-based detectors trained on the DFD and Celeb-DF dataset and tested on TIMIT-DF, DFD and Celeb-DF. The values given above are the results taken from Table 7 translated into area under curve (AUC).

The second half are the results resented by Bondi et al. in [9], where the authors performed very similar experiments like us only with a feature space learned with a convolutional neural network (CNN). In their paper, they also used a total of four sets to design training and test setups as we did with our S1 to S9. Two of the sets are Celeb-DF and DFD, which are also used by us. Comparing our work and the AUC results from Bondi et al. on the sets that are used in both papers, we can state that our approach with hand-crafted features performs only slightly worse (maximum AUC = 0.960) than their method relying on learned features (maximum AUC = 0.998). Furthermore, we can point out that their experiments with training and testing on different sets of DeepFakes results in very similar, if not worse problems in terms of generalization power (i.e., AUC drops from values larger than 0.9 to smaller than 0.7).

Table 8. Comparison (in terms of AUC) of different state-of-the-art DeepFake detectors with the presented methods. Further separation based on differentiating training and test dataset.

Training Dataset →		**DeepFakeDetection (DFD)** [12,44]			**Celeb-DF** [13]		
↓ fusion method	test dataset →	TIMIT-DF	DFD	Celeb-DF	TIMIT-DF	DFD	Celeb-DF
Ours: simple majority		0.668	0.947	0.556	0.690	0.685	0.925
Ours: weighted based on accuracy using DFD for training		0.685	0.960	0.556	0.682	0.712	0.954
Ours: weighted based on accuracy using Celeb-DF for training		0.685	0.960	0.556	0.698	0.712	0.955
[9]: Baseline		-	0.987	0.754	-	0.708	0.998
[9]: Triplet Training		-	0.882	0.759	-	0.554	0.995
[9]: EfficientNetB4. Binary Cross Entropy with augmentation		-	0.990	0.842	-	0.795	0.998
[9]: EfficientNetB4. Triplet Loss with augmentation		-	0.982	0.809	-	0.604	0.995

6.1.2. Comparison of Feature Concepts

In the case of DeepFake detection, *eye blinking* is a feature which is used for hand-crafted as well as learned feature space approaches. Section 2.1 also recaps the main functionality of DeepVision by Jung et al. [24] where they describe a hand-crafted detection method of the eye blinking behavior of persons in potential DeepFake videos. This approach is similar to our proposed feature detector for the eye blinking behavior. After the face detection happens in both cases, the detection of both eyes frame-by-frame. In our work, for every detected eye the AspectRatio changes are tracked over time. Jung et al. [24] evaluate only the amount of blinking events in a video and also the blink elapsed time as well as the blinking period time, which would correspond to the features $ID8_{blink}$ to $ID13_{blink}$ of our work. Implementation differences are visible in handling the threshold for state (open vs. closed) determination.

Li et al. [15] used a CNN for the segmentation of the eyes after they located the face area in a video. For their inter-frame blinking analysis they use an RNN with LSTM cells. The output of each RNN neuron is connected to a fully connected network, which estimate the output of the LSTM cells if an eye is open or closed.

Unfortunately, a direct comparison with these other publications in terms of performances is not possible here, since entirely different datasets were used.

To our knowledge, there is currently in the literature no similar DeepFake detection approach analyzing only the visible *mouth region* in the video with hand-crafted features. Currently, our approach only analyzes the mouth region in the video stream but does not consider of the spoken speech in the audio stream combined with the lip movements. Extending it with methods for fake voice detection, as in [51], would be an interesting next step for this method.

Considering neural network-based approaches for analysing the mouth region, Agarwal et al. [47] present the hypothesis that DeepFake videos are not able to reproduce spoken phoneme such as "M", "B" or "P", where the mouth is normally completely closed for the pronunciation. Their detection pipeline starts with the extraction of all phoneme locations. The phoneme generation is managed by the transcribing API Speech-To-Text of Google and then manually reduced to six phoneme groups ({OY,UH,UW}, {AA}, {M,B,P}, {L}, {F,V}, {CH,JH,SH}). The video stream is then aligned to these phonemes. After that, they measure the visemes for several evaluation tests in three different ways (manual, profile, CNN) [47]. This approach corresponds to a simplified lip-sync approach for a DeepFake detection, which is realized in [16] (see Section 2.1).

To the best of our knowledge, in the current literature, no hand-crafted approach analyzing only the *image foreground* to detect DeepFakes using image foreground can be found.

Looking for neural network-based approaches implementing such a feature space, the papers of Zhang et al. [52,53] have to be mentioned here. In contrast to our approach, they developed an automatic approach using a CNN. The idea behind their approach is that the image compression ratio of the face and background is different between the DeepFake and original. The reason behind this issue is that the resolution all current DeepFake algorithms is very limited. In addition, the generated fake faces are modified by affine transformations such as scaling, rotating and shearing. Based on this hypothesis, Zhang et al. try to detect the resulting artifacts of these affine transformations. The detection of the compressing distortions happens in their case with the well known error level analysis (ELA) method [54]. It follows that the training of a CNN with these ELA images which extracts the counterfeit features of the ELA images. If the CNN is able to extract these counterfeit features, then the input image of the CNN is a DeepFake. Even though the detection in [52,53] uses only DeepFake images in its tests, it would be possible to upgrade this approach for a DeepFake detection of videos.

6.2. Comparison of Hand-Crafted and Learned Features for DeepFake Detection and Conclusions

Our proposed hand-crafted features as well as hand-crafted features from other sources such as [21–24] have shown that also such expert knowledge-driven approaches are able to distinguish real from DeepFake videos. The detection rates are usually high but in most cases slightly lower than the performance achieved with learned feature spaces. The main advantage that hand-crafted features have over learned features is their interpretability and the consequences this might have for plausibility validation for decisions made.

All current approaches for DeepFake detection in the literature show error rates which are far from perfect. In particular, when DeepFake detectors are evaluated in a realistic setting, i.e., with independent training and test sets, then current hand-crafted as well as learned feature space approaches suffer generalization problems if the characteristics of training and test data are different. This has been demonstrated in our results but also in papers performing similar tests with learned feature spaces, such as Bondi et al. in [9].

Obviously, the problems of individual detectors could be increased if the DeepFake generators would include active mechanisms (counter-forensics) into the generation process to enforce false results with known detectors. Various strategies could and should be applied to address these performance and reliability issues. In this paper, we performed fusion operations to improve detection performances of hand-crafted feature spaces. In their work, Lin et al. [29] propose to extend fusion even further by combining hand-crafted features and CNN features. By doing so, they imply that it would enable us to find a solution that combines the interpretability of hand-crafted features with the potentially higher classification accuracy of learned features. The main benefit of such fusion approaches is that they generate complexer decision constructs that could compensate the problems of individual detectors in the set and might be more resilient against counter-forensics. However, these benefits would be bought at the cost of throughput/runtime behavior and a much more difficult interpretability of decisions.

In most cases, hand-crafted approaches do not need much data for model training, which may also result in lower process costs for memory or calculation time. Additionally, approaches which are including neuronal networks and specially convolutional neuronal networks need much more memory (mostly graphic memory) and CPU or GPU power for the training of the detection networks. In particular, the analyzing process of whole videos and specially a recurrent network structure have a huge impact to the needed memory. These learned approaches are also expensive in purchase costs for (new) hardware architectures. However, when the networks are finally trained, the networks are able to detect DeepFake videos in a very short time, similar to models created/trained with hand-crafted features. Therefore, neither choice would limit the application in incident response procedures (EMID), where fast (close to real time) detector responses would be required for live detectors.

7. Future Work

Our proposed hand-crafted features reach acceptable detection rates for DeepFake videos. However, not every video was classified correctly. Some DeepFake videos were detected as real video and vice versa. It is necessary to detect, analyze and find the reasons for a misclassification to improve our proposed approaches for DeepFake detection. A further improvement can be achieved by investigating different feature selection methods to strengthen the suitability of the proposed features. Possible improvements would also affect approaches from other sources, as it is extremely unlikely that any detection method can correctly classify every video, especially considering potential counter-forensics methods included in the DeepFake generation. Different detection approaches should be analyzed and the benefits of these approaches should be finally combined into a single detection method with a better detection rate and higher robustness against counter-forensics. This also concerns the fusion of hand-crafted and learned features whereat also the integration of hand-crafted methods into learned approaches are meant. In this context, the evaluation of our approaches should expand to other DeepFake databases to create a wider base for training or construct more evaluation scenarios to validate the generalizability of the approach.

A DeepFake video usually consists of two media types: the visible video and the underlying audio. These different media types should be analyzed in combination at the same time. For example, our handcrafted detector for the mouth region should be expanded to include a lip synchronization detector. It is also possible to extract the current emotion of a person in a video. Here, it is imaginable to analyze the emotion of one area (e.g., the left eye) and compare it to another (e.g., the right eye and/or the mouth). Possible aspects to determine emotions are facial expression (e.g., gesture of mouth and eyes), as well as the way of speaking.

In this paper, we started with trying to project the media forensics method of DeepFake detection onto a forensic process model (here, the data-centric examination approach (DCEA) introduced in Section 2.3). In future work, more effort is required to extend this projection, including a required extension of the DCEA data type model to make it suitable for the media data characteristics encountered here. As discussed in Section 3, the most significant change would be the design of a new, domain specific data type model for this media forensics task. While many components (such as the Processed signal data (DD2), Contextual data (DD3), Classification result data (DD8), Chain of custody data (DD9) and Report data (DD10)) could be re-used with only minor modifications, others (esp. Parameter data (DD4), Trace characteristic feature data (DD5) as well as Model data (DD7)) would need a major overhaul. The updated data modeling would also have to reflect that, in this media forensics task, different correlated (media) data streams such as video, audio, network, meta and synchronization data would have to be analyzed in parallel to substantiate the findings.

In addition to the data-driven nature of DCEA, a second reason for its choice as a forensic process model here is that it explicitly requests of modeling the error, (information) loss and (decision) uncertainty of forensic methods [7]. These considerations have to by extended for media forensics from closed set tests (where the ground truth class label in a pattern recognition problem is known) to field applicability (where only the detector response is available and the true class of a specimen encountered will remain unknown).

Author Contributions: Conceptualization, D.S., C.K. and J.D.; data curation, D.S.; funding acquisition, C.K. and J.D.; investigation, D.S. and S.S.; methodology, C.K. and J.D.; project administration,

C.K.; software, D.S.; supervision, C.K. and J.D.; validation, C.K. and S.S.; visualization, D.S.; writing—original draft, D.S.; writing—review and editing, C.K., S.S. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: The work in this paper is funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number FKZ: 13N15736.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this work, the following the pre-existing reference databases have been used for our evaluations: TIMIT-DF ([16,42]), Celeb-DF ([13]) and FaceForensics++ ([12,43,44]). They are publicly available at: https://www.idiap.ch/en/dataset/deepfaketimit (last accessed: 30 June 2021); https://github.com/yuezunli/celeb-deepfakeforensics (last accessed: 30 June 2021); https://github.com/ondyari/FaceForensics (last accessed: 30 June 2021).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Collection of Features Proposed in this Paper

Table A1. Collection of all features and their expected behaviors proposed in this paper.

ID	Feature	Description
$\text{ID1}_{\text{fusion}}$ $\text{ID1}_{\text{blink}}$	Maximum AspectRatio difference between both eyes.	The expected difference is close to 0, whereby a larger distance is suspected as an indication of a DeepFake. Additionally, the absence of winking is required for this feature.
$\text{ID2}_{\text{fusion}}$ $\text{ID2}_{\text{blink}}$	Absolute maximum AspectRatio rate of change for the left eye.	Based on several studies the eyelid movement varies based on different aspects, e.g., age and gender [24,45]. Nevertheless, the maximum speeds, as well as the relation of opening and closing speeds, could be an indication for DeepFake detection. This rate of change for each frame is determined by the difference between previous and following frame. Normalization is carried out by multiplying the rate of change by the frame rate of the video. This results in the AspectRatio change every 3 seconds, described as $\frac{\Delta AspectRatio}{3s}$. The suitability of these features is based on the disregard of blink behavior in DeepFake synthesis.
$\text{ID3}_{\text{fusion}}$ $\text{ID3}_{\text{blink}}$	Maximum AspectRatio rate of change for the left eye. Maximum opening speed of the left eye.	
$\text{ID4}_{\text{fusion}}$ $\text{ID4}_{\text{blink}}$	Minimum AspectRatio rate of change for the left eye. Maximum closing speed of the left eye.	
$\text{ID5}_{\text{fusion}}$ $\text{ID5}_{\text{blink}}$	Absolute maximum AspectRatio rate of change for the right eye.	
$\text{ID6}_{\text{fusion}}$ $\text{ID6}_{\text{blink}}$	Maximum AspectRatio rate of change for the right eye. Maximum opening speed of the right eye.	
$\text{ID7}_{\text{fusion}}$ $\text{ID7}_{\text{blink}}$	Minimum AspectRatio rate of change for the right eye. Maximum closing speed of the right eye.	
$\text{ID8}_{\text{fusion}}$ $\text{ID8}_{\text{blink}}$	Noise count in the eye state signal.	*Noise* is defined as a rapid change of eye state, where one state lasts for a maximum of 0.08 seconds. A higher number of these noises is expected for DeepFakes.
$\text{ID9}_{\text{fusion}}$ $\text{ID9}_{\text{blink}}$	Percentage of video time at which the state *open* is classified.	Another feature that can be justified by studies about human blinking behavior [24,45]. Assuming a healthy person in a non-manipulated video, on average a value of about 0.9 should be expected.
$\text{ID10}_{\text{fusion}}$ $\text{ID10}_{\text{blink}}$	Minimum duration detected for the eye state *open* in seconds.	Features based on the durations of the states are again based on the knowledge of human blinking behavior. It is assumed that the eyes are open longer than they are closed. As a conclusion $\text{ID12}_{\text{blink}} < \text{ID10}_{\text{blink}}$ and $\text{ID13}_{\text{blink}} < \text{ID11}_{\text{blink}}$ are expected.
$\text{ID11}_{\text{fusion}}$ $\text{ID11}_{\text{blink}}$	Maximum duration detected for the eye state *open* in seconds.	
$\text{ID12}_{\text{fusion}}$ $\text{ID12}_{\text{blink}}$	Minimum duration detected for the eye state *closed* in seconds.	
$\text{ID13}_{\text{fusion}}$ $\text{ID13}_{\text{blink}}$	Maximum duration detected for the eye state *closed* in seconds.	

Table A1. *Cont.*

ID	Feature	Description
ID14$_{\text{fusion}}$ ID1$_{\text{mouth}}$	Absolute maximum rate of change in y-dimension.	This rate of change for each frame is determined by the difference between previous and following frame. Normalization is carried out by multiplying the rate of change by the frame rate of the video. This results in the AspectRatio change every 3 s, described as $\frac{\Delta AspectRatio}{3s}$. For these features, a maximum speed is assumed, which is determined by training the model. Exceeding this maximum speed is assumed to be an indication for the classification DeepFake. Limitation: only works with videos where the person moves their lips during the video, e.g., when speaking.
ID15$_{\text{fusion}}$ ID2$_{\text{mouth}}$	Maximum rate of change in y-dimension. Lip opening movement in y-dimension.	
ID16$_{\text{fusion}}$ ID3$_{\text{mouth}}$	Minimum rate of change in y-dimension. Lip closing movement in y-dimension.	
ID17$_{\text{fusion}}$ ID4$_{\text{mouth}}$	Absolute maximum rate of change in x-dimension.	
ID18$_{\text{fusion}}$ ID5$_{\text{mouth}}$	Maximum rate of change in x-dimension. Lip opening movement in x-dimension.	
ID19$_{\text{fusion}}$ ID6$_{\text{mouth}}$	Minimum rate of change in x-dimension. Lip closing movement in x-dimension.	
ID20$_{\text{fusion}}$ ID7$_{\text{mouth}}$	Percentage of video time at which the state *open without teeth* is classified.	The assumption for feature ID7$_{\text{mouth}}$ is that DeepFakes are more often classified in this state compared to non-manipulated videos. The cause is the blending subprocess in the creation of DeepFakes, which leads to a loss of information and detail in the mouth region due to smoothing. As a consequence, DeepFakes are assumed to have both a comparatively low level of detail due to said blending and a comparatively high level of detail due to possible misclassification of *open with teeth* as *open without teeth*. Normalization takes place relative to the number of pixels in the TR (see Figure 4). Default value is set to -1 to be outside the considered range.
ID21$_{\text{fusion}}$ ID8$_{\text{mouth}}$	Maximum number of regions based on all frames of the video for state *open without teeth*.	
ID22$_{\text{fusion}}$ ID9$_{\text{mouth}}$	Maximum number of FAST keypoints based on all frames of the video for state *open without teeth*.	
ID23$_{\text{fusion}}$ ID10$_{\text{mouth}}$	Maximum number of SIFT keypoints based on all frames of the video for state *open without teeth*.	
ID24$_{\text{fusion}}$ ID11$_{\text{mouth}}$	Maximum number of Sobel edge pixels based on all frames of the video for state *open without teeth*.	
ID25$_{\text{fusion}}$ ID12$_{\text{mouth}}$	Percentage of video time at which the state *open with teeth* is classified.	The assumption for feature ID12$_{\text{mouth}}$ is that non-manipulated videos are more often classified in this state compared to DeepFakes. The cause is the blending subprocess in the creation of DeepFakes, which leads to a loss of information and detail in the mouth region due to smoothing. As a consequence, DeepFakes are assumed to have a comparatively low level of detail due to said blending. Normalization takes place relative to the number of pixels in the TR (see Figure 4). Default value is set to -1 to be outside the considered range.
ID26$_{\text{fusion}}$ ID13$_{\text{mouth}}$	Minimum number of regions based on all frames of the video for state *open with teeth*.	
ID27$_{\text{fusion}}$ ID14$_{\text{mouth}}$	Minimum number of FAST keypoints based on all frames of the video for state *open with teeth*.	
ID28$_{\text{fusion}}$ ID15$_{\text{mouth}}$	Minimum number of SIFT keypoints based on all frames of the video for state *open with teeth*.	
ID29$_{\text{fusion}}$ ID16$_{\text{mouth}}$	Minimum number of Sobel edge pixels based on all frames of the video for state *open with teeth*.	
ID30$_{\text{fusion}}$ ID1$_{\text{foreground}}$	Total number of frames in the video without a detectable face.	The consideration of these features is made under the assumption that DeepFake synthesis could result in artifacts, causing the face detection to fail. Normalization is relative to the number of frames of the video to ensure comparability regardless of the video length.
ID31$_{\text{fusion}}$ ID2$_{\text{foreground}}$	Total number of segments in the video without a detectable face.	

Table A1. *Cont.*

ID	Feature	Description
$ID32_{fusion}$ $ID3_{foreground}$	Maximum number of FAST keypoints based on all frames of the video for the image foreground.	The assumption for this set of features is that an almost constant value can be found throughout the course of the video. As a result, no significant differences between minimum and maximum of each feature are expected. Greater distances are seen as an indication of DeepFakes. Normalization is carried out on the basis of the two representations *Face* and *ROI* (see Figure 6 for reference) based on the level of detail as well as the number of pixels. Formally, this takes the form of $\frac{Feature_{Face}}{Feature_{ROI}}$, where $Feature_{Face\,\vert\,ROI} = \frac{FeatureCount_{Face\,\vert\,ROI}}{Pixelcount_{Face\,\vert\,ROI}}$. In order to prevent division by 0, the default value is set to -1 to be outside the considered range.
$ID33_{fusion}$ $ID4_{foreground}$	Minimum number of FAST keypoints based on all frames of the video for the image foreground.	
$ID34_{fusion}$ $ID5_{foreground}$	Maximum number of SIFT keypoints based on all frames of the video for the image foreground.	
$ID35_{fusion}$ $ID6_{foreground}$	Minimum number of SIFT keypoints based on all frames of the video for the image foreground.	
$ID36_{fusion}$ $ID7_{foreground}$	Maximum number of Sobel edge pixel based on all frames of the video for the image foreground.	
$ID37_{fusion}$ $ID8_{foreground}$	Minimum number of Sobel edge pixel based on all frames of the video for the image foreground.	

References

1. Chesney, R.; Citron, D. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.* **2019**, *98*, 147.
2. Vaccari, C.; Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media Soc.* **2020**, *6*. [CrossRef]
3. Palmer, G.L. *A Road Map for Digital Forensics Research—Report from the First Digital Forensics Research Workshop (DFRWS) (Technical Report DTR-T001-01 Final)*; Technical Report; Air Force Research Laboratory, Rome Research Site: Utica, NY, USA, 2001.
4. Champod, C.; Vuille, J. Scientific Evidence in Europe—Admissibility, Evaluation and Equality of Arms. *Int. Comment. Evid.* **2011**, *9*. [CrossRef]
5. Krätzer, C. Statistical Pattern Recognition for Audio-forensics—Empirical Investigations on the Application Scenarios Audio Steganalysis and Microphone Forensics. Ph.D. Thesis, Otto-von-Guericke-University, Magdeburg, Germany, 2013.
6. U.S. Congress. *Federal Rules of Evidence*; Amended by the United States Supreme Court in 2021; Supreme Court of the United States: Washington, DC, USA, 2021.
7. Kiltz, S. Data-Centric Examination Approach (DCEA) for a Qualitative Determination of Error, Loss and Uncertainty in Digital and Digitised Forensics. Ph.D. Thesis, Otto-von-Guericke-University, Magdeburg, Germany, 2020.
8. Böhme, R.; Freiling, F.C.; Gloe, T.; Kirchner, M. Multimedia forensics is not computer forensics. In *Computational Forensics*; Geradts, Z.J.M.H., Franke, K.Y., Veenman, C.J., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 90–103.
9. Bondi, L.; Cannas, E.D.; Bestagini, P.; Tubaro, S. Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection. *arXiv* **2020**, arXiv:2011.07792.
10. FakeApp 2.2.0. Available online: https://www.malavida.com/en/soft/fakeapp (accessed on 30 June 2021).
11. Nguyen, T.T.; Nguyen, C.M.; Nguyen, D.T.; Nguyen, D.T.; Nahavandi, S. Deep Learning for Deepfakes Creation and Detection. *arXiv* **2021**, arXiv:1909.11573.
12. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
13. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, Seattle, WA, USA, 14–19 June 2020; pp. 3204–3213.
14. Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; Natarajan, P. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. *arXiv* **2019**, arXiv:1905.00582.
15. Li, Y.; Chang, M.; Lyu, S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv* **2018**, arXiv:1806.02877.
16. Korshunov, P.; Marcel, S. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *arXiv* **2018**, arXiv:1812.08685.
17. Tolosana, R.; Vera-Rodríguez, R.; Fiérrez, J.; Morales, A.; Ortega-Garcia, J. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *arXiv* **2020**, arXiv:2001.00179.

18. Li, Y.; Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. *arXiv* **2018**, arXiv:1811.00656.
19. Matern, F.; Riess, C.; Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 83–92. [CrossRef]
20. Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A Survey on Deepfake Video Detection. *IET Biom.* **2021**. [CrossRef]
21. Yang, X.; Li, Y.; Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. *arXiv* **2018**, arXiv:1811.00661.
22. McCloskey, S.; Albright, M. Detecting GAN-generated Imagery using Color Cues. *arXiv* **2018**, arXiv:1812.08247.
23. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019.
24. Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
25. Ciftci, U.A.; Demir, I. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *arXiv* **2019**, arXiv:1901.02212.
26. Verdoliva, L. Media Forensics and DeepFakes: An overview. *arXiv* **2020**, arXiv:2001.06564.
27. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef]
28. Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.R. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2660–2673. [CrossRef]
29. Lin, W.; Hasenstab, K.; Cunha, G.M.; Schwartzman, A. Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. *Sci. Rep.* **2020**, *10*, 20336. [CrossRef] [PubMed]
30. Sánchez-Maroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter Methods for Feature Selection—A Comparative Study. In *Intelligent Data Engineering and Automated Learning—IDEAL 2007*; Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 178–187.
31. Law, M.; Figueiredo, M.; Jain, A. Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1154–1166. [CrossRef] [PubMed]
32. Kiltz, S.; Dittmann, J.; Vielhauer, C. Supporting Forensic Design—A Course Profile to Teach Forensics. In Proceedings of the 2015 Ninth International Conference on IT Security Incident Management and IT Forensics, Magdeburg, Germany, 18–20 May 2015; pp. 85–95.
33. Altschaffel, R. Computer Forensics in Cyber-Physical Systems: Applying Existing Forensic Knowledge and Procedures from Classical IT to Automation and Automotive. Ph.D. Thesis, Otto-von-Guericke-University, Magdeburg, Germany, 2020.
34. Kiltz, S.; Hoppe, T.; Dittmann, J. A New Forensic Model and Its Application to the Collection, Extraction and Long Term Storage of Screen Content off a Memory Dump. In Proceedings of the 16th International Conference on Digital Signal Processing, DSP'09, Santorini, Greece, 5–7 July 2009; IEEE Press: New York, NY, USA , 2009; pp. 1135–1140.
35. Sagonas, C.; Antonakos, E.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 Faces In-The-Wild Challenge. *Image Vis. Comput.* **2016**, *47*, 3–18. [CrossRef]
36. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
37. 2d Face Sets—Utrecht ECVP. Available online: http://pics.stir.ac.uk/2D_face_sets.htm (accessed on 19 May 2021).
38. Makrushin., A.; Neubert., T.; Dittmann., J. Automatic generation and detection of visually faultless facial morphs. In Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications—Volume 6: VISAPP, (VISIGRAPP 2017), INSTICC, Porto, Portugal, 27 February–1 March 2017; SciTePress: Setubal, Portugal, 2017; pp. 39–50. [CrossRef]
39. Kraetzer, C.; Makrushin, A.; Neubert, T.; Hildebrandt, M.; Dittmann, J. Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec '17, Philadelphia, PA, USA, 20–22 June 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 21–32. [CrossRef]
40. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *SIGKDD Explor.* **2009**, *11*, 10–18. [CrossRef]
41. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
42. Sanderson, C.; Lovell, B. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. *LNCS* **2009**, *5558*, 199–208. [CrossRef]
43. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv* **2018**, arXiv:1803.09179.
44. Dufour, N.; Gully, A.; Karlsson, P.; Vorbyov, A.V.; Leung, T.; Childs, J.; Bregler, C. DeepFakes Detection Dataset by Google & JigSaw. Available online: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html (accessed on 19 May 2021).
45. Wubet, W.M. The Deepfake Challenges and Deepfake Video Detection. *Int. J. Innov. Technol. Explor. Eng.* **2020**, *9*. [CrossRef]
46. DeBruine, L.; Jones, B. Face Research Lab London Set. Available online: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666/1 (accessed on 19 May 2021).
47. Agarwal, S.; Farid, H.; Fried, O.; Agrawala, M. Detecting Deep-Fake Videos From Phoneme-Viseme Mismatches. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 14–19 June 2020.

48. Bradski, G. The OpenCV Library. *Dobb J. Softw. Tools*, **2000**, *120*, 122–125.
49. Ross, A.A.; Nandakumar, K.; Jain, A.K., Levels of Fusion in Biometrics. In *Handbook of Multibiometrics*; Springer: Boston, MA, USA, 2006; pp. 59–90. [CrossRef]
50. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; Wiley: Hoboken, NJ, USA, 2004. [CrossRef]
51. Rana, S.; Ridwanul, M. DeepFake Audio Detection. *GitHub Repos.* Available online: https://github.com/dessa-oss/fake-voice-detection (accessed on 30 May 2021).
52. Zhang, W.; Zhao, C.; Li, Y. A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis. *Entropy* **2020**, *22*, 249. [CrossRef] [PubMed]
53. Zhang, W.; Zhao, C. Exposing Face-Swap Images Based on Deep Learning and ELA Detection. *Proceedings* **2020**, *46*. [CrossRef]
54. Krawetz, N. A Picture ' s Worth . . . Digital Image Analysis and Forensics. In Proceedings of the Black Hat Briefings 2007, Las Vegas, NV, USA, 28 July–2 August 2007; pp. 1–31.

11

[Siegel22] Forensic Data Model for Artificial Intelligence based Media Forensics - Illustrated on the Example of DeepFake Detection

This chapter of this cumulative habilitation treatise has originally been published as:

[Siegel22] Dennis Siegel, **Christian Krätzer**, Stefan Seidlitz, Jana Dittmann: *Forensic Data Model for Artificial Intelligence based Media Forensics - Illustrated on the Example of DeepFake Detection.* Proc. Electronic Imaging. Springfield, VA: Society for Imaging Sciences and Technology, Bd. 34 (2022), pp. 324-1–324-1, 2022.
https://doi.org/10.2352/EI.2022.34.4.MWSF-324

Shares/author contributions (as given in the original paper - see page 180 of this cumulative habilitation treatise): "*Initial idea & methodology: Jana Dittmann (JD) and Christian Kraetzer (CK); Conceptualization: Dennis Siegel (DS), Stefan Seidlitz (StS), CK and JD; Modelling of the new data structure: CK, DS, StS; Modelling of the templating approach for media forensics in SP/OP: CK, DS, StS; Writing – original draft: DS; Writing – review & editing: CK, StS and JD.*
All authors have read and agreed to the published version of the manuscript."

https://doi.org/10.2352/EI.2022.34.4.MWSF-324

Forensic Data Model for Artificial Intelligence based Media Forensics - Illustrated on the Example of DeepFake Detection

Dennis Siegel[1]***, Stefan Seidlitz***[1]***, Christian Kraetzer***[1]***, Jana Dittmann***[1]

[1] ***Otto-von-Guericke University, Magdeburg, Germany***

Abstract

The recent development of AI systems and their frequent use for classification problems poses a challenge from a forensic perspective. In many application fields like DeepFake detection, black box approaches such as neural networks are commonly used. As a result, the underlying classification models usually lack explainability and interpretability.

In order to increase traceability of AI decisions and move a crucial step further towards precise & reproducible analysis descriptions and certifiable investigation procedures, in this paper a domain adapted forensic data model is introduced for media forensic investigations focusing on media forensic object manipulation detection, such as DeepFake detection.

Introduction

IT-forensics is a domain that, due to its novelty and the fast changes experienced in the threat landscape that has to be considered, still sees a lot of research activity. Many of the corresponding research initiatives unfortunately remain on a purely academic level, lacking the degree of maturity required for field application of analysis methods.

In this context the existence of standardized process models plays an important role on the path to mature solutions, because to achieve the ultimate benchmark for a forensic method (which would be its admissibility in court proceedings), it would require a standardization and certification of the tool(s) and procedures as well as training and certification of the practitioners / forensic experts. While much work exists on forensic process models (including crucial components such as data models) for older sub-disciplines of IT forensics, for the younger sub-discipline of media forensics domain adapted solutions are still amiss.

As main contribution of this paper, a domain adapted forensic data model is introduced for media forensic investigations focusing on media forensic object manipulation detection. The new data model is derived by domain transfer from established best practices. Furthermore, its applicability is demonstrated by using the new model to completely rework an analysis pipeline description form an earlier paper on DeepFake detection.

These results are considered important to move a crucial step further towards precise & reproducible analysis descriptions and certifiable investigation procedures. In addition they constitute an important step towards explainable artificial intelligence (XAI), fair AI and human oversight concepts who are major aspects of the upcoming EU Artificial Intelligence Act (AIA).

The paper is structured as follows: In section a short summary on the state of the art on forensic process models and corresponding data models is presented. In section a new domain adapted data model is derived from the existing state-of-the-art, which is then used in section to rework an existing investigation pipeline description for DeepFake detection to improve this description. At the end of the paper, section presents a short summary and presents starting points for potential future work.

State-of-the-art on Forensic Process Models

Since the legislative and administrative process governing the usage of evidence in court (including expert testimony) is different for every country, it always has to be reflected in the light of the national regulations. In the German situation (which is relevant for the authors of this paper) one of the most important guidelines for IT forensics (and sub-disciplines) is the “Leitfaden IT-Forensik” [2] of the German Federal Office for Information Security (BSI; the national cyber security authority). It provides various means for modeling forensic processes, including the definition of a phase-driven investigation & reporting model, a basic data model and a classification of methods and tools. Since its last official update in 2011, it has been reflected upon and extended in many publications, such as [6] and [1].

What is currently amiss in this line of research is a domain specific adaptation to media forensics. This became apparent to the authors when analysis work performed in a previous publication (here: [12], where an analysis of video data with the aim of DeepFake detection is performed using three individual detection operators and alternative fusion operators) turned out to be hard (if not entirely impracticable) to project onto the pre-existing data models.

The following section elaborates more on this research gap while section briefly summarizes with the Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling (DCEA) the latest extension to the BSI guidelines from [2], which is used here as starting point for the extension work.

The work in the following chapters is than focused primarily on extending the data model and secondarily on the impact to aspects of the investigation & reporting mode.

Media forensic processes

Textbooks on media forensics such as [5] as well as relevant research work like [9] agree upon the fact that at the core of modern media forensics pipelines looking into questions of integrity one or more pattern recognition or anomaly detection mechanisms are to be found. After data collection and pre-processing operations either sequences or parallel networks of such operators (in the latter case followed by fusion operators) are used to implement a set of analysis tasks. The output of the analyses will then have to be interpreted by an human expert, e.g., in form of an expert testimony in court.

While agreement exists in the community on the fundamental outline of analysis pipelines, the existing state-of-the-art lacks domain specific data models. Those are required to: a) facilitate efficient requirement engineering, design specification, implementation, certification and deployment of media forensic analysis pipelines, b) enable error, loss and uncertainty estimations in individual forensic analyses performed (see [6]) and c) ease processes aiming at the explainability and fairness in forensic investigations (novel factors that have to receive increased attention due to the current changes in legislation governing the application of AI, such as the upcoming EU Artificial Intelligence Act).

Due to the lack of such domain specific data models, this paper focuses on proposing such a model, suitable to the task at hand. This is done by performing a domain transfer on an established data model for digitized forensics (see section).

A Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling

Forensic process models are an important cornerstone in the science and more importantly the practice of forensics. They guide investigations and make them comparable, reproducible as well as certifiable. Usually, the adherence to strict guidelines (i.e. process models) are regulated within any legal system (e.g. in the US by the fourth of the Daubert criteria ("*the existence and maintenance of standards and controls*" [3])). For mature forensic sciences, like for example dactyloscopy, internationally accepted standards (like the ACE-V process model for dactyloscopy) have been established over the last decades.

Due to the fact that IT forensics is a rather young discipline in this field (with media forensics being an even younger sub-discipline) it is hardly astonishing that here the forensic process models have not yet achieved the same degree of maturity as in other fields. Nevertheless, they would still be important to achieve universal court acceptability of methods. One well established forensic process model for IT forensics is the one proposed by the German Federal Office for Information Security (BSI). When it was originally published in 2011, its sole focus was on computer and network forensics but since then it has evolved to suite also to some extend the needs of other sub-disciplines such as digitized forensics. The latest major revision of this process model, which is used within this paper, can be found in [6] and is called the Data-Centric Examination Approach (DCEA). The core of DCEA consists of three main aspects: a model of the *phases* of a phase driven forensic process, a classification scheme for *forensic method classes* and *forensically relevant data types*.

The six DCEA *phases* are briefly summarized as: Strategic preparation (SP), Operational preparation (OP), Data gathering (DG), Data investigation (DI), Data analysis (DA) and Documentation (DO). While the first two (SP and OP) contain generic (SP) and case-specific (OP) preparation steps, the three phases represent the core of any forensic investigation. The phase DO is split in [6] into two aspects: case accompanying documentation (Chain-of-Custody, etc) as well as final documentation (e.g. the expert opinion statement presented in court). For details on the phase model the reader is referred, e.g. to [6] or [1].

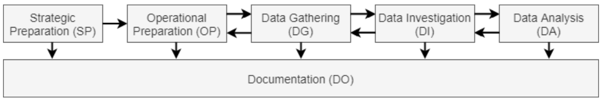


Figure 1. *Phase model (based on [2])*

The second core aspect of DCEA is the definition of *forensic method classes* as presented in [6]. They consist of methods of: the Operating system (OS), the File system(s) (FS), IT applications (ITA), Explicit means of intrusion detection (EMID), Scaling of methods for evidence gathering (SMG) and Data processing and evaluation (DPE). Like the phases, this aspect is of limited relevance for this paper. For details on this classification scheme for investigation methods the reader is referred to [6].

The third (and in the context of this paper most relevant) aspect is the specification of *forensically relevant data types*. More recent publications, such as [1], have shown that this scheme needs to be extended accordingly if new investigation domains are considered.

The original set of data types, which was designed with digital IT forensics in mind, needs to be adapted towards every investigation domain. In [7] and [6] such an adaptation for the field of digitized forensics has been discussed for the field of dactyloscopy (forensic fingerprint analysis and comparison). This adaptation is summarized in Table 1. Because it is much closer to the requirements faced within this paper than the original data model, it is used as starting point for the modeling work performed here.

Deriving a Forensic Data Model for Artificial Intelligence based Media Forensic Investigations focusing on Integrity

Performing abstract data modeling without precise knowledge about the context, in which the data type is supposed to be used, is a futile task. Therefore, first a generalized media forensic analysis process is briefly discussed in section . This is followed in section by an identification of the typical data streams within such a process. As the last step in the data modeling, the data streams are further differentiated into data types in section .

Modeling a generalized media forensic analysis process

In general, each processing operation (or operator) is considered here as an atomar processing black box component with an identifier and (usually) a description of the processing performed in this operation. Each component has four well defined connectors: *input*, *output*, *parameters* and *log data*. To pay respects to the particularities of this field and make the following modeling task easier, a fifth connector is defined within this paper for a specific type of operator which requires a knowledge representation or a model for its processing operation. In that case, this fifth connector is labeled *model*. Depending on the nature of the operator this could be a rule set, signature set, statistical model, neural model, or any other form of knowledge representation.

Figure 2 shows the modeling for a small, exemplary selected processing sub-routine within a bigger media forensic investigation process (here the sub-routine of face segmentation as necessary step in DeepFake detection for videos). The first operator in this three step processing sub-routine is loading the video from its *in-*

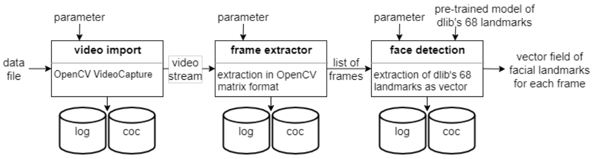

parameter
parameter
parameter
pre-trained model of dlib's 68 landmarks
data file
video import
OpenCV VideoCapture
video stream
frame extractor
extraction in OpenCV matrix format
list of frames
face detection
extraction of dlib's 68 landmarks as vector
vector field of facial landmarks for each frame
log
coc
log
coc
log
coc

Figure 2. *Exemplary modeling of the process for face detection.*

put. The *parameters* need to be chosen based on the video format and the *output* is stored as video stream. This video stream is then in the next operator split into single frames as necessary preprocessing for an image based face detection and segmentation algorithm. For the face detection and segmentation, a pre-trained model with 68 landmarks (here from [8]) is loaded at the third operators *model* connector. This is the only step in this small example where model data is used.

Each step provides corresponding process documentation in the form of logs and chain of custody (CoC) data at its *log data* connector.

Identifying typical data streams

Based on the atomar operator description above and generalizing media forensic (i.e., passive) investigations focusing on analyzing the integrity of media objects, here five typical data streams are identified: The *process description* is proposed as a sourceable or instantiable template, which is generated before starting the investigation. It is supposed to be generated in the phase of Strategic preparation (SP) and contains general information (such as process layouts/graphs, interfaces and operators involved) independent from a specific investigation. Besides the actual process layout this stream inherits also information from DD7, DD9 and DD10 of the data types form digitized forensics (see table 1).

The second data stream *media data* contains all forms of media such as images, videos, audio and/or network streams used and created within the investigation process. Media data could be found both on input and output connectors of a component and would in case of an investigation in digitized forensics contain information from DD1, DD2 and DD8.

The non-media output of the individual examination steps is combined into the data stream *forensic process/pipeline internal data and reporting*. It contains actual (intermediate) investigation results and CoC data such as hashes and logs as well as error, loss and uncertainty indicators, meta data and traceability/explainability information (such as a risk and circumstantial evidence map (RCEM)). This output is gathered in the phases OP, DG, DI and DA and would in case of an digitized forensics investigation be described by DD2, DD3, DD8, DD9 and DD10.

Another important aspect is the combination of all settings used in the investigation, including all parameters and models used. This combination is defined as *process control data* and contains in digitized forensics DD3, DD4, DD7 and DD8.

The last data stream is *contextual data*, which contains all information regarding the context of a specific investigation. In general it contains information such as operator IDs, data source descriptors (e.g., camera types) and the results of a content analysis of the media objects required for plausibility and fairness evaluation. In case of an digitized forensics investigation contextual data would be found in DD3, DD8, DD9 and DD10.

This subdivision of the data associated with an investigation is a functional classification paying respect on one hand to the characteristics of data objects involved and on the other hand to operational and security requirements. The media data stream of an investigation might easily contain terabytes of video data which would require a access to a private cloud for efficient handling, while the reporting data would assumed be much smaller in data size but be more frequent and have other constraints like reliable time-stamping. From the operational and security perspective also different protection levels (and as a consequence security mechanisms) would be required depending on the nature of the objects in a stream and the risks associated.

Deriving the domain specific data model

Taking the data streams identified above for media forensics into account, it is necessary to adapt the existing data models. As starting point, here the data types from digitized forensics are chosen because they require a less wide-ranging re-modeling. The objective of deriving a domain specific data model for integrity

Forensic data type	**Description** (according to [6])
DD1 Raw sensor data	Digital input data from the digitalization process (e.g. scans of test samples)
DD2 Processed signal data	Results of transformations to raw sensor data (e.g. visibility enhanced fingerprint pattern)
DD3 Contextual data	Contain environmental data (e.g. spatial information, spatial relation between traces, temperature, humidity)
DD4 Parameter data	Contain settings and other parameter used for acquisition, investigation and analysis
DD5 Trace characteristic feature data	Describe trace specific investigation results (e.g. level1/2/3 fingerprint features)
DD6 Substrate characteristic feature data	Describe trace carrier specific investigation results (e.g. surface type, individual surface characteristics)
DD7 Model data	Describe trained model data (e.g. surface specific scanner settings, reference data)
DD8 Classification result data	Describes classification results gained by applying machine learning and comparable approaches
DD9 Chain of custody data	Describe data used to ensure integrity and authenticity and process accompanying documentation (e.g. cryptographic hash sums, certificates, device identification, time stamps)
DD10 Report data	Describe data for the process accompanying documentation and for the final report

Forensic data types defined in [6] for an exemplary selected process in digitized forensics (here digital dactyloscopy) (updated from [7])

focused media forensics is a specification and overlap-free representation of data types. As a result of the modeling performed here, eight media forensic data types (MFDT, see table 2) are defined, which are loosely derived from the ten data types of digitized forensics. *Digital input data* (MFDT1) is a re-definition based on DD1 and considers now any kind of media data as it is initially taken as input to the investigation. *Processed media data* (MFDT2) is derived from DD2 and contains all operator output which are media data. *Contextual data* (MFDT3) is derived from DD3 and includes case specific information regarding the investigation process and -objects. Contextual data can also be used to control targeted parametrization and thus allow case or objects specific parameter optimization. They also allow for plausibility and fairness evaluations as part of the assessment of an investigation performed. *Parameter data* (MFDT4) is similar to DD4 from digitized forensics and contains all configurations and parametrizations for operators in an investigation (except for model data, see MFDT6 below), including those who are used for training of classifiers and models before the actual investigation. *Examination data* (MFDT5) combines and extends the data types DD5, DD6 and DD8 from digitized forensics. It comprises all occurring non-media outputs (e.g., trace information, patterns and anomalies identified) of the investigation. *Model data* (MFDT6) corresponds to DD7 from digitized forensics. It includes trained models of machine learning algorithms like rule based approaches or decision trees as well as models of neural networks (incl. their network architecture). *Log data* (MFDT7) is an component of the documentation which is here newly added to the data model and is used for administration and maintenance (including Syslogs and information about the memory usage). Data in MFDT7 are not relevant for the specific case in the investigation, but are necessary for the administration of the system (e.g., to notice that the memory allocated for the task is not sufficient). *Chain of custody & report data* (MFDT8) is a combination of DD9 and DD10 from digitized forensics. They characterize the case relevant documentation for integrity and authenticity assurance as well as the accompanying documentation for the final report. For admissibility in court the final report would be required following the corresponding chain of custody guidelines.

Chain of custody & report data (MFDT8) also have to address the description of the deployed (process) modeling with regard to origin and provenance of decision (AI) models used. Especially in the context of neural networks a detailed specification of the network structure(s) (MFDT4, MFDT6) as well as the used parameters for training, (potential transfer-learning), testing and validation phases (MFDT4) would be required to allow for the necessary reproducibility of setups and corresponding error, loss and uncertainty as well as explainability considerations for explainable AI. But not only classifier designs and parameterizations have to be reported upon: Another aspect for the documentation refers to the data used in the process(es) of model generation, focusing on the training and validation sets taken from the content of data types MFDT1 and MFDT2. The decisive factors in this respect are origin, diversity and quantity of data (summarized within MFDT3). It is also significant for the documentation to characterize the differences between training and test/evaluation/validation phases of each mechanism. For example the consideration of disjoint data sets for training and testing yields a more generalizable and trustworthy result than a cross-validation would obtain. Furthermore, the documentation of initial control parameters (MFDT4: e.g., learning rate, optimizer, loss function) as well as information about the training process (MFDT7 & MFDT8: training duration, used hardware, etc.) are very important for traceability as well as interpretability.

Also important is the run-time of the detection process, which needs to be evaluated and documented in relation to the hardware used. Another documentation criteria refers to the type of result data (MFDT2 or MFDT5) calculated by methods such as neural network. In decision-based classification, the result is often represented by a classification/prediction label (MFDT5) and/or confidence estimate (MFDT5). In some cases it can also be an image or other media object (MFDT2) that represents relevant information such as a map of anomalies found, to be interpreted by a human investigator.

In field application, because of the typical black box usage of mostly Neural Networks, with an unknown internal behaviour in the hidden layers between in- and output, it might be possible that there exist no process data or feature vectors/data (MFDT5). But for a mature forensic method aiming for court admissibility such kind of black box behavior would not sufficient, because result data of forensic operators must be comprehensible. Because of that, methods focusing on explainability (e.g., LIME [11] or

Data type	Derived from DD	Description
MFDT1 Digital input data	DD1	The initial media data considered for the investigation.
MFDT2 Processed media data	DD2	Results of transformations to media data (e.g. grayscale conversion, cropping)
MFDT3 Contextual data	DD3	Case specific information (e.g. for fairness evaluation)
MFDT4 Parameter data	DD4	Contain settings and other parameter used for acquisition, investigation and analysis
MFDT5 Examination data	DD5, DD6, DD8	Including the traces, patterns, anomalies, etc that lead to an examination result
MFDT6 Model data	DD7	Describe trained model data (e.g. face detection and model classification data)
MFDT7 Log data	newly defined	Data, which is relevant for the administration of the system (e.g. system logs)
MFDT8 Chain of custody & report data	DD9, DD10	Describe data used to ensure integrity and authenticity (e.g. hashes and time stamps) as well as the accompanying documentation for the final report.

Media Forensic Data Types (MFDT) proposed in this work

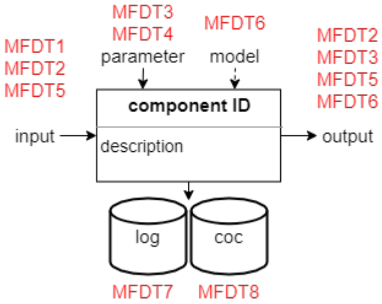


Figure 3. Template structure for a single component

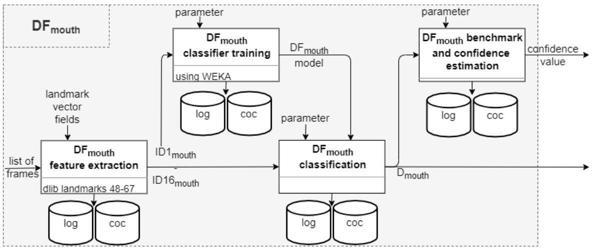


Figure 4. *Illustration of the DeepFake detection based on mouth region* ***modeled as a template*** *in the proposed context model in the phase of* ***Strategic preparation (SP)***

LRP[10]) have to be included in the investigation. Moreover, the network structure could be expanded between hidden layers with more output layers to allow obtaining processed data (MFDT2) or feature vectors (MFDT5). As a necessary result, a neural network would become more transparent, interpretable and explainable.

Figure 3 shows the link between media forensic data types (MFDT) for the operator description presented above. As discussed in section , depending on whether a model is used in an operator or not, each component has four or five well defined connectors. The operator (i.e., process step itself; here shown as a box) has an unique identifier and a description of the process. This description should increase traceability as well as explainablity. The input of a component has a form of media data, the court exhibits itself (MFDT1) or after previously done preprocessing steps (MFDT2) or examination data (MFDT5). Depending of the processing step, the generated output could be media data (MFDT2), a derived information on the investigation context (MFDT3) or investigation results (MFDT5). It is also possible during the phase of Strategic preparation (SP) that a model is trained (MFDT6). The process control is done by parameters (MFDT4). Furthermore, the gathered contextual data (MFDT3) can be used for optimization of the parameters in the specific investigation. MFDT3 could for example be information about the recording device, resolution or lighting conditions, which might be useful to estimate decision uncertainty and thereby allowing to estimate the fairness of an investigation. The loading of a model (MFDT6) is limited to model-driven operators, which why it is shown by a dashed line. Process accompanying documentation will be divided and separately saved in log data (MFDT7) and chain of custody data (MFDT8) based on the modeled data types.

Illustration of the practicability of applying the proposed new data model

As indicated in section , one motivation for this paper were apparent problems when projecting an exemplary selected media forensics processing pipeline designed for DeepFake from a previous paper (here [12]) of the authors onto existing data models. In this section it is shown, how the adapted data model from chapter can be successfully used for the pipeline in that publication. The modeling work is done in two separate steps. The first instantiation is focusing on training models for the operators in Strategical preparation (SP). This initialization is done using well established DeepFake reference data sets for the training. First, the original videos and corresponding DeepFakes are imported and pre-processed in a suitable format so that they can be further processed as a video stream. The video stream is then divided into individual frames (single images). The resulting list of images is used for both face detection and subsequent DeepFake detection (see figure 2). Assuming one face per frame, a pre-trained 68 landmark model is used for face detection. It locates the position of each of those facial landmarks and stores them in a vector field. The detection algorithm itself consists of the components feature extraction, classifier training, classification and benchmarking. In the feature extraction each frame is evaluated based on the corresponding landmarks relevant to the region it focuses on and generates a feature vector relevant to the classification. Exemplary for the detector DF_{mouth}, the classifier DF_{mouth} model is created using the J48 classifier from Weka [4], testing different models and parameter settings. The optimal model then gets integrated into the classification, which then returns the decision (e.g. D_{mouth}). Afterwards a second instance of validation is done by benchmarking and confidence estimation. Based on the confidences the weights for a consecutive fusion step are determined. The same procedure is done for the algorithms DF_{eye} and $DF_{foreground}$ with differences in the considered landmarks and generated features (for details see [12]). During the whole process each step gets documented and stored in the log and chain of custody databases respectively.

The second instantiation of the modeling corresponds to the actual investigation determining whether a DeepFake manipulation occurred in the presented videos. Considering the pipeline presented in figure 1, it covers all phases from OP to Documentation. The first processing steps are identical to those performed in the SP instantiation. This is to be expected, because both training and testing of an operator should be done under the same conditions (i.e., after identical pre-processing). Changes can be found in the application of the detection operators. Here the parts regarding model training are left out because the models pre-trained in SP are loaded instead, together with the used classifier parameters. Thus initialized the operators are applied to video material to determine traces of DeepFake manipulations. The respective individual decisions D_{eye}, D_{mouth} and $D_{foreground}$ are then merged into the fusion module to determine a final decision D_{fusion}. The required fusion weights used for this purpose also come from the SP. A complete mapping of this process, including a labeling of the Media Forensic Data Types communicated at each connector, can be found in Figure 5.

Conclusion and Future Work

In this paper a domain adapted forensic data model is introduced for media forensic investigations focusing on media

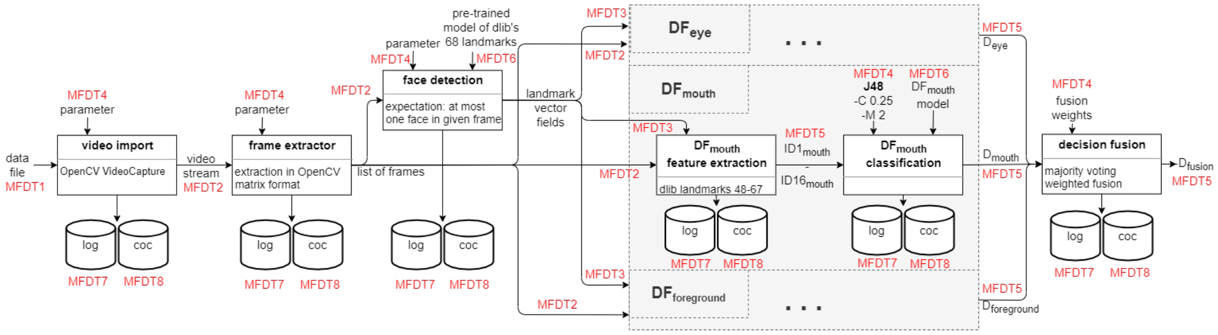


Figure 5. Illustration of the DeepFake detection pipeline ***instantiated*** *in the forensic process model phase of* ***Operational preparation (OP)****, with the inclusion of occurring data types*

forensic object manipulation detection. The new data model is derived by domain transfer from established best practices. Furthermore, its applicability is demonstrated by using the new model to completely rework an analysis pipeline description form an earlier paper.

The work performed here motivates future work on the following aspects: First, on extending the considerations on templating and instantiation works in Strategic preparation (SP) and Operational preparation (OP) phases to move a further step towards precise and reproducible analysis descriptions and thereby towards certifiable investigation procedures.
Second, on expanding the modeling with regard to knowledge data generation and representation to be better able to include also more complex operations (e.g. modern training scenarios for neural network based detectors) as well as context dependent pipeline alternatives into forensic workflows.
Third, on extending the work on error, loss and uncertainty (on basis of [6]) as well as explainability and fairness in AI-driven forensics.

Acknowledgements

The work in this paper is funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number FKZ: 13N15736 (project "Fake-ID"). The authors wish to thank Dr. Stefan Kiltz for the discussions on the DCEA and its applicability as well as Dr. Robert Altschaffel and Dr. Mario Hildebrandt for discussions on domain adaptations required for forensic data models.
Author Contributions: Initial idea & methodology: Jana Dittmann (JD) and Christian Kraetzer (CK); Conceptualization: Dennis Siegel (DS), Stefan Seidlitz (StS), CK and JD; Modelling of the new data structure: CK, DS, StS; Modelling of the templating approach for media forensics in SP/OP: CK, DS, StS; Writing – original draft: DS; Writing – review & editing: CK, StS and JD.

References

[1] Robert Altschaffel. *Computer forensics in cyber-physical systems : applying existing forensic knowledge and procedures from classical IT to automation and automotive*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[2] BSI. *Leitfaden IT-Forensik*. German Federal Office for Information Security, 2011.

[3] Christophe Champod and Joëlle Vuille. Scientific evidence in europe - admissibility, evaluation and equality of arms. *International Commentary on Evidence*, 9(1), 2011.

[4] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor.*, 11(1):10–18, 2009.

[5] Anthony T. S. Ho. *Handbook of digital forensics of multimedia data and devices / edited by Anthony T.S. Ho and Shujun Li, Department of Computing and Surrey Centre for Cyber Security (SCCS), University of Surrey, UK.* Wiley/IEEE Press, Hoboken, 2015.

[6] Stefan Kiltz. *Data-Centric Examination Approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[7] Stefan Kiltz, J. Dittmann, and C. Vielhauer. Supporting forensic design - a course profile to teach forensics. *2015 Ninth International Conference on IT Security Incident Management and IT Forensics*, pages 85–95, 2015.

[8] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, 2009.

[9] Christian Krätzer. *Statistical pattern recognition for audio-forensics*. PhD thesis, University of Magdeburg, 2013.

[10] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. The lrp toolbox for artificial neural networks. *Journal of Machine Learning Research*, 17(114):1–5, 2016.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[12] Dennis Siegel, Christian Kraetzer, Stefan Seidlitz, and Jana Dittmann. Media forensics considerations on deepfake detection with hand-crafted features. *Journal of Imaging*, 7(7), 2021.

Author Biography

__Jana Dittmann__ is a Professor on multimedia and security at the University of Otto-von-Guericke University Magdeburg (OvGU). She is the leader of the Advanced Multimedia and Security Lab (AMSL) at OvGU, which is partner in national and international research projects and has a wide variety of well recognized publications in IT security. __Christian Kraetzer__ is a post-doc researcher and __Dennis Siegel__ as well as __Stefan Seidlitz__ are PhD students at AMSL.

12

[Kraetzer22] Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection

This chapter of this cumulative habilitation treatise has originally been published as:

[Kraetzer22] **Christian Kraetzer**, Dennis Siegel, Stefan Seidlitz, Jana Dittmann: *Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection*. MDPI Sensors 2022 (Special Issue Detecting and Preventing Deepfake Attacks), 22(9), 3137; 2022.
https://doi.org/10.3390/s22093137

The original publication is available as Open Access publication distributed under the terms of the Creative Commons 4.0 CC BY license at MDPI Sensors:
https://www.mdpi.com/1424-8220/22/9/3137/html

Shares/author contributions (as given in the original paper - see page 211 of this cumulative habilitation treatise): "*Conceptualisation, Christian Kraetzer (C.K.); data curation, Dennis Siegel (D.S.) and Stefan Seidlitz (S.S.); formal analysis, C.K.; funding acquisition, C.K. and Jana Dittmann (J.D.); investigation, C.K., D.S. and S.S.; methodology, C.K. and J.D.; software, D.S. and S.S.; supervision, C.K. and J.D.; validation, C.K., S.S. and J.D.; visualisation, D.S.; writing—original draft, C.K.; writing—review and editing, C.K., D.S., S.S. and J.D.*
All authors have read and agreed to the published version of the manuscript."

sensors

MDPI

Article

Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection

Christian Kraetzer *, **Dennis Siegel**, **Stefan Seidlitz** and **Jana Dittmann**

Department of Computer Science, Otto-von-Guericke University, 39106 Magdeburg, Germany;
dennis.siegel@ovgu.de (D.S.); stefan.seidlitz@ovgu.de (S.S.); jana.dittmann@iti.cs.uni-magdeburg.de (J.D.)
* Correspondence: christian.kraetzer@ovgu.de

Abstract: Academic research in media forensics mainly focuses on methods for the detection of the traces or artefacts left by media manipulations in media objects. While the resulting detectors often achieve quite impressive detection performances, when tested under lab conditions, hardly any of those have yet come close to the ultimate benchmark for any forensic method, which would be courtroom readiness. This paper tries first to facilitate the different stakeholder perspectives in this field and then to partly address the apparent gap between the academic research community and the requirements imposed onto forensic practitioners. The intention is to facilitate the mutual understanding of these two classes of stakeholders and assist with first steps intended at closing this gap. To do so, first a concept for modelling media forensic investigation pipelines is derived from established guidelines. Then, the applicability of such modelling is illustrated on the example of a fusion-based media forensic investigation pipeline aimed at the detection of DeepFake videos using five exemplary detectors (hand-crafted, in one case neural network supported) and testing two different fusion operators. At the end of the paper, the benefits of such a planned realisation of AI-based investigation methods are discussed and generalising effects are mapped out.

Keywords: media forensics; forensic process model; certifiable investigation methods; DeepFake detection

check for updates

Citation: Kraetzer, C.; Siegel, D.; Seidlitz, S.; Dittmann, J. Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection. *Sensors* **2022**, *22*, 3137. https://doi.org/10.3390/s22093137

Academic Editor: Ahmed Bouridane

Received: 1 March 2022
Accepted: 16 April 2022
Published: 20 April 2022

1. Introduction

Modern day media forensics is a strongly pattern recognition, respectively, artificial intelligence (AI) driven domain. In a recent white paper titled "Secure, robust and traceable use of AI-problems, procedures and actions required" [1] (translated from the German title "*Sicherer, robuster und nachvollziehbarer Einsatz von KI-Probleme, Maßnahmen und Handlungsbedarfe*"), the German Federal Office for Information Security (BSI, the German national cybersecurity authority; Since forensics, as part of legal proceedings, is regulated on basis of national legislation, research in forensics also has to acknowledge national legal and statutory requirements – in the case of this paper, therefore, besides internationally accepted best practices, like the Daubert standard (see Section 2), the German national situation is reflected, due to the fact that all authors are working at a German research institution and the work is funded in part by the German Federal Ministry of Education and Research (BMBF)) summarises the current situation with regards to trustworthy and reliable AI applications as follows: There is currently an urgent need for further research into the security of AI systems, in order to be able to make reliable statements about the security and confidence of such systems. According to the BSI, there are three specific aspects on which research needs to focus:

1. Development of standards, technical guidelines, test criteria and test methods: Currently, there exist no such standards that are sufficiently suitable for assessing the security and reliability of AI systems for critical contexts (such as health care, finance

health care, finance, etc.). There is also a lack of security benchmarks for less critical applications (with a few exceptions).

2. Research effective countermeasures against AI-specific attacks: The existing measures for such attacks are often insufficient. In order to ensure a secure and robust operation of AI systems, further countermeasures must be researched.
3. Research into methods of transparency and explainability: The often inadequate explainability of AI systems has a significant influence on their Information Technology (IT) security and causes a lack of acceptance of the systems.

What holds true for every form of AI usage is even more important if it comes down to AI-driven processes that are (by regulation) restricted to decision support systems, e.g., in the case of forensics, where it internationally accepted standard that investigation results have to be interpreted in expert testimony. Here, the corresponding expert has to be able to explain the investigation method as well as all aspects influencing an investigation outcome in front of a trier of fact (in most cases a single judge, a group of judges or a jury). Besides other reasons, this human presentation and interpretation is considered necessary because the expert can also interpret contextual information to reason about the intention of an action (e.g., why a DeepFake video has been created—see Section 2.2 for a list of white hat application scenarios for this dual-goods media manipulation method), which is a challenge where the AI alone will still fail.

As **contributions of this paper**, the following items are addressed:

- The need for modelling forensic processes is reasoned upon.
- A concept for modelling media forensic investigation pipelines is derived from established guidelines.
- The applicability of such modelling is illustrated on the example of a media forensic investigation pipeline focusing on the detection of DeepFake videos. It is important to already mention at this point, that the DeepFake detectors, test criteria and test methods used in this paper are used for illustrative purposes on the processes and are **not** claiming to represent the state-of-the-art in detector research.
- The benefits of such a planned realisation of AI-based investigation methods are discussed.

Regarding the first of these items (the reasoning on the need for modelling forensic processes) it is shown that forensic process models are an important cornerstone in the science and more importantly the practice of forensics. They guide investigations and make them comparable, reproducible as well as certifiable. Usually, the adherence to strict guidelines (i.e., process models) are regulated within any legal system (e.g., in the US by the fourth of the Daubert criteria (“*the existence and maintenance of standards and controls*” [2])). For mature forensic sciences, like for example fingerprint analysis, internationally accepted standards (like the Analysis, Comparison, Evaluation and Verification methodology (ACE-V) process model for dactyloscopy) have been established over the last decades. Due to the fact that IT forensics is a rather young discipline in this field (with media forensics being an even younger sub-discipline), it is hardly astonishing that here the forensic process models have not yet achieved the same degree of maturity as in other fields. For this reason, an effort is made here to move this field forward by presenting a concept for modelling media forensic investigation pipelines, which is derived from well-established guidelines. Since all the authors are working at a German research institution, here an extension of the guidelines on IT forensics [3] by the German Federal Office for Information Security (BSI) is used as the basis for this work.

Regarding the third item from the list of contributions identified above, the applicability of the proposed modelling work is illustrated on the example of a media forensic investigation pipeline focusing on the detection of DeepFake videos. This application scenario it chosen because it is a recent threat scenario that currently achieves a lot of research attention due to the potential implications it has for the trust assumptions in video material used (amongst other scenarios) in political debates. Here, an already complex investigation pipeline taken from previous work [4] consisting of three detectors plus a fusion operator (with two alternative fusion methods tested) is extended by adding two additional detectors. Despite the fact that both new detectors are performing in benchmarking significantly better than guessing (with a Kappa value of $\kappa \sim 0.4$), the following empirical evaluations show a slight drop in the average detection performance (at least by $\kappa = 0.025$). This drop is neither the expected nor an intuitive outcome, but illustrates how important an extensive benchmarking of methods prior to field application (also in a fusion setup) is. While the detection methods used here are admittedly not amongst the most sophisticated detectors currently available, their general performance, especially the mentioned problems with the generalisation power, are representative for the current situation in this field of applied pattern recognition.

Following the discussions on this empirical work, the benefits of such a planned realisation of AI-based investigation methods are discussed in the contexts of development of standards, technical guidelines, test criteria and test methods on one hand and research into methods of transparency and explainability of AI methods on the other hand.

The rest of the paper is structured as follows: In Section 2, a brief overview on the state of the art on forensic process modelling for media forensics and DeepFake detection is presented. This is followed in Section 3 by a summary of related work aiming at advancing the basic forensics guidelines used in this paper (here, the German BSI guidelines on IT forensics). Based upon these foundations, Section 4 introduces the modelling work in this paper (based on previous work in Siegel et al. [5]). This chapter also summarises known evaluation best practices, metrics as well as DeepFake data sets. In Section 5, an application example using components from the introduced process modelling is given for the description of a fusion-based DeepFake detector pipeline. The descriptions are divided into a planning/templating phase and the instantiation of the pipeline for all evaluations in this paper. Section 6 provides a brief summary of the results, before the following Section 7 projects the conclusions onto the contributions identified in Section 1. The paper is concluded by a short view into potential future work in Section 8.

2. State of the Art on Forensic Process Modelling for Media Forensics and DeepFake Detection

In a very recent textbook on media forensics targeting digital face manipulations [6], the authors reflect the current academic perspective on media forensics as: "*In case manipulation detection methods are used by public authorities competent for preventing, investigating, detecting, or prosecuting criminal offences this shall be done in a lawful and fair manner. While these are broad concepts, case law further explains how to apply these concepts.*" Those mentioned characteristics are further specified in [6] as:

- Lawfulness: "*refers to the need* [. . .] *to adopt adequate, accessible, and foreseeable laws with sufficient precision and sufficient safeguards whenever the use of the detection technology,* [. . .], *could interfere with fundamental rights and freedoms*".
- Fairness: "*points to the need for being transparent about the use of the technology. Furthermore, it is obvious that the use of the detection methods should be restricted to well-defined legitimate purposes*, [. . .]".

Regarding the fairness, the authors in [6] point out that when intended for court usage, explainability of the forensic algorithms used is a strong requirement. In addition, they state that: "*From an organizational point, one should also know that decisions purely and solely based on automated processing, producing adverse legal effects or significantly effecting subjects, are prohibited, unless authorized by law, and subject to appropriate safeguards, including at least human oversight and intervention.*"

In accordance with other well established works originating in the academic parts of media forensics research (like [7]), the synopsis presented in [6] is that "[*t*]*he absence of a unified approach, common regulatory framework, and commonly accepted practices has resulted in a situation where different initiatives emerge across countries which share some common elements but also numerous differences that can lead to challenges related to interoperability.*"

An important step towards more mature forensics are forensic process models. They guide investigations and are supposed to make them comparable, reproducible, as well as certifiable. Usually, the adherence to strict guidelines (i.e., process models) are regulated within any legal system (e.g., in the US by the fourth of the Daubert criteria ("*the existence and maintenance of standards and controls*" [2])).

Due to the fact that IT forensics is a rather young discipline in this field (with media forensics being an even younger sub-discipline) it is hardly astonishing that here the forensic process models (if they exist at all) have not yet achieved the same degree of maturity as in other fields. Nevertheless, they would still be important to achieve universal court acceptability of methods.

To pay respect to the difficulties in this domain, the following two subsections provide the following: A brief overview over forensic process modelling requirements and best practices for media forensics are presented in Section 2.1, starting with an international perspective and then narrowing down for the German perspective relevant for the authors of this paper. These discussions are then followed in Section 2.2 by a brief summary on the current state of the art in the application domain of DeepFake detection, which is the chosen application scenario within this paper.

2.1. Forensic Process Modelling for Media Forensics

In contrast to the international perspective of academic research on media forensics, its field application is governed by national legislation. Undeniably the most active judicial system worldwide, with a high demand for forensic and media forensic investigations, is found in the USA. Naturally, a well-established set of best practices is the result. In Section 2.1.1, a very brief overview on these best practices is presented. In the following Section 2.1.2, the German situation, relevant to the authors, is reflected.

As a preamble to this section, it has to be highlighted that all authors are computer scientists and possess absolutely no legal training. All statements and interpretations presented below on legal considerations are therefore layman's interpretation of freely available material, which are made to the best of the authors' knowledge.

2.1.1. Forensic Process Modelling Requirements and Best Practices (US Perspective)

On the U.S. federal level, strict rules for the integration of the results of forensic investigations were established in 1975. These rules, the Federal Rules of Evidence (FRE [8]), define the framework within which evidence can be admitted into court. Even if these rules are in their original form only applicable on U.S. federal level, their concepts for handling forensic data have influenced many other judicial systems worldwide and are also considered with interest in many European legal systems (see [2]).

In general, under the FRE, forensic results have to be interpreted by experts to the court. The reason for this lies in the assumption that any judge (or jury) will lack the expert knowledge to completely interpret the findings of a forensic investigation on his/her own and that therefore expert testimony is strictly required in court proceedings. If the expert's opinion helps the fact finder in understanding the significance of factual data, then the expert witness is essential for the case and its opinion evidence is admissible.

Using the terminology of U.S. jurisdiction, the trial judge acts as a form of 'gatekeeper', assuring that scientific expert testimony truly proceeds from reliable (or scientific) knowledge. Considerations on relevance and reliability require the trial judge to ensure that the expert's testimony is 'relevant to the task at hand' and that it rests 'on a reliable foundation'. According to [9], the primary rules that are relevant for the presentation of forensic evidence in court (i.e., that apply to expert witnesses) in the FRE are FRE rule 702 ("*Testimony by Experts*") and FRE rule 703 ("*Bases of Opinion Testimony by Experts*").

In the year 2011, FRE rule 702 ("*Testimony by Experts*") was amended to: "*A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if: (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony* *is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case*".

When analysing this rule, it can be seen that, in regarding the admissibility of an expert, the judge has to establish whether the following four points are met:

- **Qualification of a witness as expert:** First, a witness has to qualify as an expert. The conclusion of this process is that the presiding judge decides whether the witness may offer opinion testimony as an expert.
- **Type of knowledge considered:** The first seven words of FRE rule 702 specify different types of knowledge (e.g., scientific, technical or other specialised knowledge) that an expert can offer.
- **Who is addressed by the expert:** Basically, there are two entities the expert has to convince. First, the judge, to get admitted in pre-trial hearings, and second the 'fact finder' (the "*trier of fact*" in FRE rule 702 [10], either a jury in normal cases or a judge in non-jury trials) at the trial itself.
- **Qualification:** Any expert has to testify upon the five criteria listed in FRE rule 702 "*knowledge, skill, experience, training, or education*" [10]. This information helps the judge to decide whether an expert can be admitted to trial in a specific case and helps the 'fact finder' (i.e., usually the jury) to assign corresponding weights to each expert's testimony in the decision process.

If these four points are established, the judge determines for the case whether an expert is qualified to testify under FRE rule 702. The April 2000 (effective December 2000) amendment of FRE rule 702 includes three further requirements, which must also be met. The goal of these additional requirements is to make it easier to present effective scientific and technical expert testimony whenever such evidence is warranted and provide a basis for the exclusion of opinion testimony that is not based on reliable or mature methodology. These additional requirements are [10]: "[. . .] *if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.*" In April 2011, another requirement was added to this list [8] "[. . .] *the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue* [. . .]".

In the notes on FRE rule 702 published by the Legal Information Institute at Cornell Law School in December 2010 [11], the current regulations regarding the interpretation of this rule for U.S. federal courts are summarised as follows: "*Rule 702 has been amended in response to* Daubert v. Merrell Dow Pharmaceuticals, Inc., *509 U.S. 579 (1993), and to the many cases applying Daubert, including* Kumho Tire Co. v. Carmichael, *119 S.Ct. 1167 (1999). In Daubert the Court charged trial judges with the responsibility of acting as gatekeepers to exclude unreliable expert testimony,* [. . .]". The main result of this amendment are the so called Daubert hearings where the judge(s) are supposed to use the so called Daubert criteria (see below) to assess the admissibility of methods and investigation results to legal proceedings.

The other FRE regarding opinions and expert testimony (rule 701 "*Opinion Testimony by Lay Witnesses*", rule 703 "*Bases of an Expert's Opinion Testimonies*", rule 704 "*Opinion on an Ultimate Issue*", rule 705 "*Disclosing the Facts or Data Underlying an Expert's Opinion*" and rule 706 "*Court-Appointed Expert Witnesses*"; see [8]) are further regulating the usage of forensic investigation results in court, but are of little relevance to this paper. For a more detailed analysis, see [12].

Regarding the second and third point of the list given above in the analysis of FRE rule 702 ('Type of knowledge considered' and 'Who is addressed by the expert'), it has to be summarised that if something is declared to be 'science' in regard to FRE rule 702, then the criteria for the evaluation of scientific methods introduced in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993) [13], ref. [14] have to be applied by the judge to make the expert prove this declaration.

In 1923, the court in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923) made a first suggestion how to proceed with the admission of expert testimony based on novel forensic techniques. The court in Frye suggested [15]: "*Just when a scientific principle or discovery* *crosses the line between the experimental and demonstrable stages is difficult to define.* [. . .], *the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs."* In Frye (or the Frye standard as it is also referred to) the court concluded that the polygraph test that was intended to be used in this case could not be admitted because it lacked the required general acceptance in the corresponding research fields. Prior to this seminal ruling in Frye, according to [9], the competence of an expert was equivalent to his success in real life. In [9] it is summarised as: *"If a person earned a living selling his or her knowledge in the marketplace, then that person would be considered an expert who could testify at trial."*

The Frye standard was in 1975 partially replaced by the FRE. Initially, they contained no special rule that, when dealing with 'scientific' evidence, novel or otherwise, ensured that science-based testimony is reliable and, therefore, admissible. Therefore, all evidence was considered admissible if relevant, provided its use in court was not outweighed by "*unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time or needlessly presenting cumulative evidence*", as stated in FRE rule 402 [8].

The next relevant step in legal developments on expert testimony (and therefore the means of introducing forensic sciences into court) occurred in 1993, when the U.S. Supreme Court made another ground-breaking decision on expert testimony in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993) [13]. Daubert was in 1999 followed by another important court case, *Kumho Tire Co. v. Carmichael*, 119 S.Ct. 1167 (1999). Both Daubert and Kumho Tire arose out of civil lawsuits. An extensive and intelligible summary of the proceedings in the Daubert cases (original and the affirmation in the U.S. Court of Appeals) is presented in [9]. The main point of interest for this paper is that the court unanimously held that Frye did not survive the enactment of the FRE. In interpreting FRE rule 702, the court in Daubert stated that if the admissibility of scientific evidence is challenged, it is the function of the trial court to act as 'gatekeeper' to determine whether proffered opinion evidence is relevant and reliable. The U.S. Supreme Court specified several flexible and non-exclusive criteria (the so-called Daubert criteria or Daubert standard) to guide other courts when they have to consider in deciding whether a scientific field is sufficiently reliable to warrant admission of opinion evidence. As a further important milestone, in 1999 in *Kumho Tire Co. v. Carmichael*, 119 S.Ct. 1167 (1999), the U.S. Supreme Court applied the Daubert criteria of proof of reliability to all forms of expert opinion testimony (i.e., scientific, applied science, technological, skill and experience). Additionally, the court in Kumho Tire made it clear that the list of Daubert criteria was meant to be helpful and is not a definitive checklist, but rather a flexible, non-exclusive recommendation. As a result, no attempt has been made in US law to 'codify' these specific criteria. Other U.S. law cases have established that not all of the specific Daubert criteria can apply to every type of expert testimony. The specific criteria, explicated by the Daubert court, are [11]:

> "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*";
> "*whether the technique or theory has been subject to peer review and publication*";
> "*the known or potential rate of error of the technique or theory when applied*";
> "*the existence and maintenance of standards and controls*";
> "*whether the technique or theory has been generally accepted in the scientific community*".

While the criteria DC2 to DC5 are self-explanatory (including the fact that publication in DC2 means 'open publication'), DC1 is summarised more precisely in [13] as "*the theory or technique (method) must be empirically testable, falsifiable and refutable*".

The Daubert criteria are widely accepted in the classical fields, like medical forensics. It can also be, and is, applied in the much younger field of IT-forensics (see e.g., [16,17]). It has to be admitted that the field of media forensics, which is the focus of this thesis, is still lacking maturity in this regard. Here, only very specific methods applied in this field already fulfil the Daubert criteria sufficiently. Overviews over the more mature techniques in this field are given in [18,19].

A well-established reference in this field is the document Forensic Examination of Digital Evidence: A Guide for Law Enforcement [20] of the U.S. Department of Justice-National Institute of Justice. Unfortunately, this document has not received any update since 2004. Its place has been taken over in past years by publications of well-established (and court-trained) forensic experts, such as [21,22] or [23]. Homogenising the different individual views, expert bodies, like the Organization of Scientific Area Committees (OSAC) Task Group (TG) on Digital and Multimedia Evidence have become normative institutions arguing for harmonisation of procedures: "[. . .] *digital/multimedia evidence, and other forensic disciplines, would be in a much stronger position to demonstrate their scientific basis if they were considered as belonging to a harmonized forensic science rather than as mere disciplines at the intersection of forensic specialties and other sciences.*" [24]. As a reason, the following is given: "*Like many other specializations within forensic science, the digital/multimedia discipline has been challenged with respect to demonstrating that the processes, activities, and techniques used are sufficiently scientific.*" This OSAC TG aims at advancing digital/multimedia evidence, and forensic science as a whole by (amongst other aspects):

> "*Strengthen scientific foundations of digital/multimedia evidence by developing systematic and coherent methods for studying the principles of digital/multimedia evidence to assess the causes and meaning of traces in the context of forensic questions, as well as any associated probabilities.*"
> "*Assess ways to mitigate cognitive bias in cases that require an understanding of the context of traces in order to analyze digital/multimedia evidence,* [. . .]"
> "*Establish effective ways to evaluate and express probative value of digital/multimedia traces for source level and activity level conclusions. This includes studying how quantitative evaluation of digital/multimedia evidence can be constructed for different forensic questions,* [. . .] *as well as studying how such evaluative results can be communicated to decision-makers.*"

As a consequence, generalisable and standardised forensic process models are currently sought for to bridge the gap between the strict legal requirements (see the FRE 702 and Daubert requirements discussed above) and the current degree of (or rather lack of) maturity of many media forensic approaches originating form academic research.

2.1.2. The German Perspective

As discussed in detail in [2], the situation in the U.S. can not be directly projected onto the European situation. One of the main reason is that forensics are still entirely governed by national legislation.

For the authors the German situation is relevant. Here, the currently most relevant official guideline is the BSI code of practice for IT forensics ("Leitfaden IT-Forensik" [3]) of the German Federal Office for Information Security (BSI). One of the intentions of this document was to try to homogenise forensic proceedings in the highly fragmented system with 35 different police agencies independent from each other on federal- and state level. In this regard, it is very similar in its intention to the document Forensic Examination of Digital Evidence: A Guide for Law Enforcement [20] (2004) of the U.S. Department of Justice-National Institute of Justice and similar to its U.S. pendant, it is outdated with the last updated version of the "Leitfaden" (German for guidelines) having been published in 2011. Nevertheless, it is still a valuable starting point and has been used as such for more recent work on forensic process modelling, see Section 3 below.

In its core, the BSI guidelines for IT forensics define a phases driven process model model, tool categories and a forensic data model. In the phase driven process model, which is for this paper the most relevant component of these guidelines, six different phases are described: Strategic preparation (SP), Operational preparation (OP), Data gathering (DG), Data investigation (DI), Data analysis (DA) and Documentation (DO). These phases, which are outlining the process itself, are briefly summarised in Table 1 the interaction pattern of these phases is shown in Figure 1. The actual passing of data and results between the phases is taking place in the horizontal transitions, shown as horizontal arrows in the figure.

It has to be admitted here, that this paper somewhat diminishes the role the Documentation receives in [3]. Originally, the DO is considered to have two distinguishable aspects: the accompanying documentation of the process (which can be seen as a combination of complete logs as well as a tamper-proof (hence the uni-directional, solid-lined vertical arrows in the figure), digital chain-of-custody) and the final documentation (e.g., as the written expert report intended to be used in court as basis for an expert testimony). In the present context, it is important to point out that the latter (i.e., the drafting of the final documentation for a case) should be used to reflect upon potential improvements of the processes and their implementation, acting as a feedback loop into SP. This is shown in Figure 1 by adding the dashed arrow from DO into SP.

Table 1. Sets of examination steps for digital forensics as defined in [25] (updated from [3,26]).

Phases	Description (According to [25])
Strategic preparation (SP)	Includes measures taken by the operator of an IT system and by the forensic examiners in order to support a forensic investigation prior to an incident
Operational preparation (OP)	Includes measures of preparation for a forensic investigation after the detection of a suspected incident
Data gathering (DG)	Includes measures to acquire and secure digital evidence
Data investigation (DI)	Includes measures to evaluate and extract data for further investigation
Data analysis (DA)	Includes measures for detailed analysis and correlation between digital evidence from various sources
Documentation (DO)	Includes measures for the detailed documentation of the proceedings, also for the transformation into a different form of description for the report of the incident

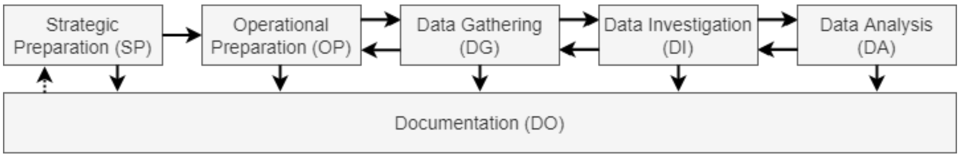


Figure 1. Phase model (based on [3]), extended to include an optional feedback loop from the Documentation (DO) into the strategic preparation (SP).

One important aspect here is the separation of preparation steps in an investigation into two distinct phases (the strategic preparation (SP) on one hand, and the operational preparation (OP) on the other). In recent work on this model (e.g., [25], which is available in English), the SP is generally defined as: "*The strategic preparation* [. . .] *includes all preparation procedures taken ahead of the actual occurrence of a specific incident*". Exemplary measures for SP in the context of digital forensics are given by [25] as: "*Documentation and extension of knowledge of IT systems specifics, tool testing for forensic data types and sets of methods determination for error loss and uncertainty estimation, setup of logging capabilities, performance of system landscape analysis, data protection considerations,* [. . .]". In contrast, the OP is specified to "[. . .] *include all preparation procedures taken after of the actual occurrence of a specific incident. Those procedures by definition do not alter any data on the targeted system*". These preparation phases are then followed by the actual application of forensic procedures, which can be separated into the triplet of data gathering (DG), data investigation (DI) and data analysis (DA). The whole process is in every phase (including SP and OP) supported by accompanying documentation, which is in the last phase (documentation (DO)) used as the basis for the generation of the official documents regarding the investigation (e.g., the evidence to be interpreted in expert testimony in a court case). It has to be acknowledged here that these BSI guidelines on outlining a forensic process, while acknowledging established best practices in this field, significantly differ from other national guidelines, even in other EU states. This can be illustrated by comparing it, for example, with the model described in [27], which very well reflects the Norwegian approach. It also builds upon a phase-driven model, but with a different established phases layout: (1) Identification Phase, (2) Collection Phase, (3) Examination Phase, (4) Analysis Phase and (5) Presentation Phase. This is much closer to long-time established best practices in traditional (analogue world) forensic sciences and requires then explicit activities to achieve and maintain “Digital Forensic Readiness” [27] (an equivalent to the Strategic Preparation phase in the BSI guidelines) to successfully cope with modern day digital and digitised forensics tasks.

The second core aspect of the BSI guidelines is the classification scheme for forensically relevant data types. More recent publications (see Section 3 below) have shown that the original scheme as proposed by the BSI in 2011 needs to be extended accordingly if investigation domains other than hard-disk, RAM or network forensics are considered.

The third core aspect of the BSI guidelines is the definition of forensic method classes. For a detailed discussion on these method classes, including considerations on the availability in certain investigation contexts, practicalities of their application in a forensic process, etc., we refer to [25].

2.2. *(Brief) Summary on the Domains of DeepFake Generation and Detection*

The methodologies and solution concepts for the generation of DeepFake material are manifold. Due to this reason, these generative processes (which are mostly outside the scope of this paper) are covered extensively in survey publications, like [28] or the corresponding chapters in [6]. Generally, they are divided into the classes of *facial re-enactment*, *facial replacement* (or face swapping), *face editing* and *face synthesis*.

If a target persons facial expression corresponds to the expression of another person, presented as controlling input, then the generation process is called as facial re-enactment. For the case of face replacement, a source face is transferred to a face in a target media object where the facial expression of the target person has not changed. Face editing addresses the same face in an image or video. Only the facial expressions or some face parts are modified. In contrast to the methodologies described above, face synthesis refers to newly created faces that are not linked to real persons [28].

The usage of DeepFakes does not automatically imply black hat (i.e., malicious) applications, but also a large number of white hat (i.e., benign or non-malicious) application scenarios exist. The following subsection 2.2.1 briefly summarises examples for both types of application scenarios, using the four different classes of generative processes mentioned above.

The different generation strategies also create class-related artefacts in the output media objects. In Section 2.2.2, these are very briefly summarised.

2.2.1. DeepFake Use Cases

The use of DeepFakes has a wide range of possible application scenarios, where their impact can be on an individual or societal level. While mainly the negative aspects are highlighted in existing literature, there are also positive examples of the use of DeepFakes. As stated in [6]: “[. . .] *it is important to note that face manipulation techniques are also expected to have positive impact on society and economy.* [. . .] *can help to address privacy issues through privacy-enhancing techniques, they facilitate the training of machine learning models with synthetic data* [. . .], *they can help with sustainability by facilitating virtual fitting rooms for the beauty and fashion industries and drive economic development with (high added value) mobile e-commerce, entertainment, and social media applications*”.

This non-exhaustive list can easily be extended, with most use cases having both positive and negative aspects to be considered. DeepFakes have received first news coverage due to usage in pornographic contexts using face-swaps, where primarily women became victims of targeted defamation. Face-swaps are also used for white hat applications, e.g., showing the user wearing certain clothes (‘magic mirror’ scenarios for online shopping).

In the context of lip synchronisation techniques used in DeepFakes, the most prevalent examples show the manipulation of video footage and the spoken word of well known politicians (e.g., former US president Barack H. Obama or Nancy Pelosi in her time as Speaker of the United States House of Representatives), to spread misinformation. On the other hand, the same technique can be used to break language barriers, in the example of the "Malaria Must Die" campaign, where the famous football star David Beckham addresses the audience in this health campaign in nine languages, due to the help of DeepFake technology. In addition to the use of real voices, the use of synthetic voices is also an application scenario to be discussed here. In white hat applications, this can increase the accessibility of content (e.g., in text-to-speech systems). A possible threat of this synthetic voice (or rather imitation of an existing voice) are the so-called Vishing attacks [29]. In terms of face editing, the main purpose is fun applications, e.g., to simulate ageing, different hair styles and makeup. Although the authors are not aware of any attack scenario based on face editing, its use for rejuvenation and artificial ageing could pose a challenge for youth protection. Another well-used application is the fictitious resurrection of deceased people, which is often used to retain established actors for cinematic productions (e.g., Peter Wilton Cushing in the film "Rogue One"). It can also be used for a more immersive experience in education or to provide a more immersive experience in education (e.g., historical facts presented by a contemporary witness). While the intents are positive, the use results both in ethical and legal questions.

Finally, it is important to note that AI cannot decide whether a DeepFake is used positively or negatively. A human observer/expert is always needed here to decide between black hat and white hat application, based on the context of the usage of DeepFake technologies as summarised above.

2.2.2. DeepFake Detection

Because of their creation process, most DeepFakes are inherently compromised with artefacts or traces which might unmask fake media. The amount and type of those artefacts are versatile and depend on the used creation method. Artefacts are divided primary into visual artefacts within single video frames (intra-frame) and temporal artefacts across several video frames (inter-frame) [30]. Furthermore, Mirski et al. [28] subdivide both those categories in smaller artefact categories. In case of visual artefacts, they differ between blending, environment and forensics. Blending refers to "*generated content* [which] *is blended back into the frame*" [28]. Blending artefacts are marked by edges. Environment artefacts are specified by the content which differs from the rest of the frame (e.g., different lightning conditions). Forensic artefacts are special fingerprints which are created by DeepFake generation models (e.g., Convolutional Neural Networks (CNN)). Additionally, imperfections like unnatural head poses are mentioned in this context. Temporal artefacts are distinguishing between behaviour, physiology, synchronisation and coherence. Regarding behaviour artefacts, it is easier to replace one face by another than to copy the (gestical) behaviour of the person. The investigations of similar but also different behaviours could be a hint to those DeepFake artefacts. With a specific video camera setting, it is possible to detect physiological signals like the heart rate of a person. Currently, DeepFake videos are not able to reproduce these physiological signals before those physiological artefacts are indications for interferences in DeepFake videos. Synchronisation artefacts address inconsistencies between lip movements and the corresponding voice. Coherence artefacts describe, e.g., flickers and jitters which may be present in DeepFakes [28].

Many different approaches detect those different artefacts with varying detection methods, which can be ordered into the following two main groups: Hand-crafted and learned feature methods. Most approaches detect DeepFake artefacts with neural networks and a huge amount of example data. After many training iterations, they analyse the example data and produce learned features which are needed for further classification steps. Convolutional Neural Networks (CNNs) are able to detect spatial features whereas Recurrent Neural Networks (RNNs) are preferable for the detection of temporal features. Li et al. [31]

detect eye blinking with a Long-term Recurrent Convolutional Network (LRCN) model, which consists mainly of three parts: feature extracting, sequence learning and state prediction. They also suggest this approach for a DeepFake detection. However, the paper of Li et al. [31] does not goes into detail in case of the evaluation of DeepFake detection.

In contrast to the neural networks-based methods, the alternative approach is learning to identify DeepFake material with pre-defined, hand-crafted features defined by domain experts. Hand-crafted feature methods are in DeepFake detection less common and of these few existing papers using hand-crafted methods, like [32–34], most detect DeepFake videos using Support Vector Machines (SVMs), typical 2-class classifiers. Other hand-crafted feature methods (e.g., [4]) are implemented by decision trees. Jung et al. [35] created a detector called DeepVision based on the Eye-Aspect-Ratio (EAR) of Soukupov et al. [36], which combines Machine Learning techniques with heuristic methods based on results of medical-, biology- and brain engineering research. They used the knowledge of the behaviour of human eye blinking for the detection approach of their DeepFake detector. Nevertheless, they tested their approach in [36] only on a statistically insignificant number of different DeepFake videos (without any attempt to also determine the amount of false positive errors on benign material).

Three of the five detectors used for empirical experiments in this paper are re-used from previous work, published in [4]. All three are relying on hand-crafted features. These three detectors are combined in information fusion with two newly implemented detectors (see Sections 5.1.2 and 5.1.3 for details). This usage of ensembles of detectors for a complex decision forming has been established as best practice for DeepFake detection. Regarding detection pipelines intended for (forensic) field usage, in [6] the need for fusion-based approaches is strongly argued for as follows: "[. . .], *a skilled attacker, aware of the principles on which forensic tools work, may enact some counter-forensic measure on purpose* [. . .]. *Therefore, the integration of multiple tools, all designed to detect the same type of attack but under different approaches, may be expected to improve performance, and especially robustness with respect to both casual and malicious disturbances*".

3. Related Work and the Derived Challenge for This Paper

Modern day science means reaching out while standing on the shoulders of giants. In this paper, pre-existing work already extending the German BSI guidelines for IT forensics [3] is used to advance towards a comprehensive concept for modelling media forensic investigation pipelines. Two different branches-related work are considered here: On one hand, the works of Kiltz et al. on evolving the BSI guidelines into the so-called Data-Centric Examination Approach (DCEA) for modern IT forensics (see Section 3.1), and on the other hand, the authors own previous work on a domain adaptation for media forensics (see Section 3.2). At the end of this chapter, in Section 3.3, the challenge addressed in this paper is briefly summarised.

3.1. The Data-Centric Examination Approach (DCEA)

As discussed in Section 2.1.2 above, the last published official revision of BSI guidelines dates back to 2011. Since then, it has been used and extended. Significantly updated version, which is also used within this paper, can be found in [25] and is called by its authors the Data-Centric Examination Approach (DCEA). The DCEA re-uses and extends the three core aspects already present in the BSI guidelines from 2011: a model of the *phases* of a phase driven forensic process, a classification scheme for *forensically relevant data types* and *forensic method classes*.

The majority of the extensions done in recent publications focus on domain adaptation for further investigation domains. While the original guidelines focused on hard-disk, RAM and network traffic analysis, [25] extends this scope to also include aspects relevant for digitised forensics (exemplary discussed for the field of dactyloscopy (forensic fingerprint analysis and comparison)). Other publications, like, e.g., [37], adapt to domains with specific constraints like Internet of Things (IoT) forensics.

As a preparatory work for this journal paper, the authors already presented an domain specific adaptation for media forensics, which is discussed in the following section.

3.2. Model Adaptation for Media Forensic Tasks

As pointed out above, modelling of media forensics processes is nothing new. In the past, it has mainly been used in academia to provide understandable and reproducible description of media analysis pipelines (see, e.g., [12]). To move forward and address the crucial challenges of development of standards, technical guidelines and certifiable test criteria and test methods as well as research into transparency and explainability of AI driven forensic methods, more elaborate modelling is required.

In [5], a first step for a concept for modelling media forensic investigation pipelines is derived from established guidelines has been done by modelling a corresponding domain adapted data (types) model, derived from DCEA. This new data model, called Media Forensic Data Types (MFDT) is summarised in Table 2.

Table 2. Media Forensic Data Types (MFDT) proposed in [5].

Data Type	Description
MFDT1 Digital input data	The initial media data considered for the investigation.
MFDT2 Processed media data	Results of transformations to media data (e.g., greyscale conversion, cropping)
MFDT3 Contextual data	Case specific information (e.g., for fairness evaluation)
MFDT4 Parameter data	Contain settings and other parameter used for acquisition, investigation and analysis
MFDT5 Examination data	Including the traces, patterns, anomalies, etc that lead to an examination result
MFDT6 Model data	Describe trained model data (e.g., face detection and model classification data)
MFDT7 Log data	Data, which is relevant for the administration of the system (e.g., system logs)
MFDT8 Chain of custody & report data	Describe data used to ensure integrity and authenticity (e.g., hashes and time stamps) as well as the accompanying documentation for the final report.

Taking the typical data streams in media forensics into account, in [5] an adaptation of the existing data models was performed. As starting point the data types from digitised forensics were chosen because they required a less wide-ranging re-modelling than any other previously defined model. The objective for the modelling was (besides the domain adaptation) a specification and overlap-free representation of data types. As a result the following eight media forensic data types (MFDT) were defined: *Digital input data* (MFDT1) considers any kind of media data as it is initially taken as input to the investigation. *Processed media data* (MFDT2) contains all operator output which are media data. *Contextual data* (MFDT3) includes case specific information regarding the investigation process and objects. Contextual data can also be used to control targeted parametrisation, and thus allow case or objects specific parameter optimisation. They also allow for plausibility and fairness evaluations as part of the assessment of an investigation performed. *Parameter data* (MFDT4) contains all configurations and parametrisations for operators in an investigation (except for model data, see MFDT6 below), including those who are used for training of classifiers and models before the actual investigation. *Examination data* (MFDT5) comprises all occurring non-media outputs (e.g., trace information, patterns and anomalies identified) of the investigation. *Model data* (MFDT6) is made up by trained models of machine learning algorithms like rule-based approaches or decision trees as well as models of neural networks (including their network architecture). *Log data* (MFDT7) is a component of the documentation and is used for administration and maintenance (including Syslogs and information about the memory usage). Data in MFDT7 are not relevant for the specific case in the investigation, but are necessary for the administration of the system (e.g., to notice that the memory allocated for the task is not sufficient). *Chain of custody and report data* (MFDT8) characterise the case relevant documentation for integrity and authenticity assurance, as well as the accompanying documentation for the final report. For admissibility in court, the final report would be required following the corresponding chain of custody guidelines. This data model is re-used as it is within this paper as one component in the concept for modelling media forensic investigation pipelines.

3.3. The Challenge Addressed in This Paper

The discussions above illustrate the apparent gap between the academic research community (as potential solution providers for forensic methods) on one hand and the requirements imposed onto forensic practitioners on the other hand. The intention of this paper is to facilitate the mutual understanding of these two classes of stakeholders and assist with first steps intended at closing this gap. To do so, first a concept for modelling media forensic investigation pipelines is derived from established guidelines. Then, the applicability of such modelling is illustrated on the example of a media forensic investigation pipeline focusing on the detection of DeepFake videos. At the end of the paper, the benefits of such a planned realisation of AI-based investigation methods are discussed and generalising effects are mapped out.

4. Materials & Methods for the Design of a Process-Driven Investigation Model for DeepFake Detection

Even in the most recent academic publications in this field (like [6]), DeepFake detectors are only evaluated in lab tests without any concerns about integration into operational procedures. This might be sufficient for rapid prototyping and academic research, but does not suffice for field applicable forensic methods. In this section, a perspective for the path forward, towards more mature investigations, is presented. Its starts with the necessary methodology and concepts, which are followed by the a discussion on suitable metrics and materials (here specifically an overview over publicly available data sets that exist for benchmarking purposes).

The main methodology for modelling media forensic investigation pipelines was outlined already briefly in [5] (where a domain specific forensic data model was derived) and is significantly extended here.

The ultimate benchmark for any forensic method, which is its applicability in court, can only be achieved on a national level. It is acknowledged here that, due to the fact that all authors are living and working in Germany, the work presented (despite being written in English and presented in an international Journal context) is focused on the German situation and corresponding technical guidelines. Furthermore, at this point it has to be emphasised again that the authors are computer scientists and possesses absolutely no legal training.

Any integration into an operational context would have to focus on various aspects. These would include, among other issues:

- **Organisational:** Specifying the method (as an investigation workflow) and establishing its constraints, limitations and potential errors attached to the method and/or its application.
- **Technical:** Buying and installation of the investigation environment (e.g., forensic workstations) and all required infrastructure (including software such as police casework systems as well as a suitable chain of custody realisation for digital assets).
- **Personnel:** Hiring, training and (re-)certification of experts for applying the method.

Within this paper, the focus lies on the organisational aspects of operationalising investigation methods. It basically follows the BSI guidelines on IT forensics [3] with its split into the separate contexts (the preparation in the forensic process model phase of

Strategic preparation (SP), here called *templating*, and the actual usage of a method in the other phases, starting with the Operational preparation (OP), here called *instantiation*).

In the following Section 4.1 the operational units (operators) that are supposed to form parts of an investigation pipeline are modelled. This is followed in Section 4.2 by considerations on the orchestration of operators into an investigation pipeline. Section 4.3 then discusses evaluation best practices and publicly available benchmarking data sets.

The work in this chapter is intended to prepare an illustration of an investigation pipeline for DeepFake detection in Section 5, using own prior work (i.e., detectors).

4.1. Modelling of Operator Units

Each operation (or operator) in a forensic process is considered in the approach used here as an atomic processing (black box) component with an identifier, a well defined and documented functionality and (usually) a description of the processing performed in this operation. Each component is modelled here as having four well-defined connectors (see Figure 2): *input*, *output*, *parameters* and *log data*. To pay respects to the particularities of this field and make the following modelling task easier, a fifth connector is defined within this paper for a specific type of operator which requires a knowledge representation or a model for its processing operation. In their case, this fifth connector is labelled *model*. Depending on the nature of the operator this could be a rule set, signature set, statistical model, neural model, or any other form of knowledge representation. Each of these decision-forming approaches has individual advantages and disadvantages. Often, a comparison of these methods and their trained models is solely done based on detection and generalisation performance (e.g., by means of accuracy or area under curve). In addition, sometimes other performance criteria determined are representing feature space dimensionality and number of modelled classes (see, e.g., [38]).

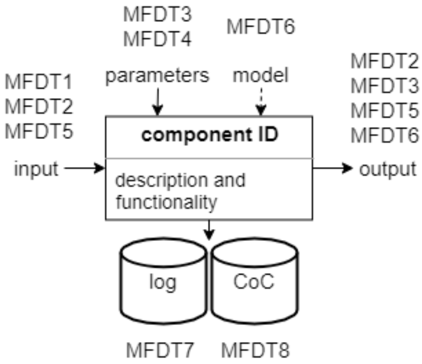


Figure 2. Template structure for a single component (adapted from [5]).

Figure 2 shows the link between media forensic data types (MFDT; see Section 3.2) for the operator description presented above. The input of a component has a form of media data, the court exhibits itself (MFDT1) or after previously done pre-processing steps (MFDT2) or examination data (MFDT5). Depending on the processing step, the generated output could be media data (MFDT2), a derived information on the investigation context (MFDT3) or investigation results (MFDT5). It is also possible during the phase of strategic preparation (SP) that a model is trained (MFDT6). The process control is done by parameters (MFDT4). Furthermore, the gathered contextual data (MFDT3) can be used for optimisation of the parameters in the specific investigation. MFDT3 could, for example, be information about the recording device, resolution or lighting conditions, which might be useful to estimate decision uncertainty and thereby allowing us to estimate the fairness of an investigation. The loading of a model (MFDT6) is limited to model-driven operators, which is why it is shown by a dashed line. Process accompanying documentation will be divided and separately saved in log data (MFDT7) and chain of custody data (MFDT8) based on the modelled data types.

4.2. Orchestration of Operators into an Investigation Context

For mature media forensics approaches, the integration (or orchestration) of individual operators into an investigation context has to be done in two distinct episodes: First, in the planning and preparation of a type of investigation in the phase of strategical preparation (SP), and second, in the initialisation of a forensic pipeline for a case-specific investigation in the phase of operational preparation (OP).

First, in SP, the work is focusing on crucial tasks of organisational, personnel and technical nature. Aspects of organisational are, e.g., defining (hereafter called *templating*) workflows and procedures and getting these procedures certified (if necessary). Examples of aspects of personnel nature would be the training of investigators (including their certification if necessary) as well as the assignment of responsibilities. Technical aspects include the hardware and software to be used, i.e., installation of the investigation systems and all required infrastructure (log servers, chain of custody (CoC) infrastructure, etc.) as well as the training of decision models for model driven operators and the benchmarking of trained operator to assess their reliability.

At the end of the process in SP, well-specified templates exist that can easily be instantiated into practical investigations as soon as an event/incident triggers an investigation request. Figure 3 shows an example for such a templating, derived from the description of a DeepFake detection pipeline in [4].

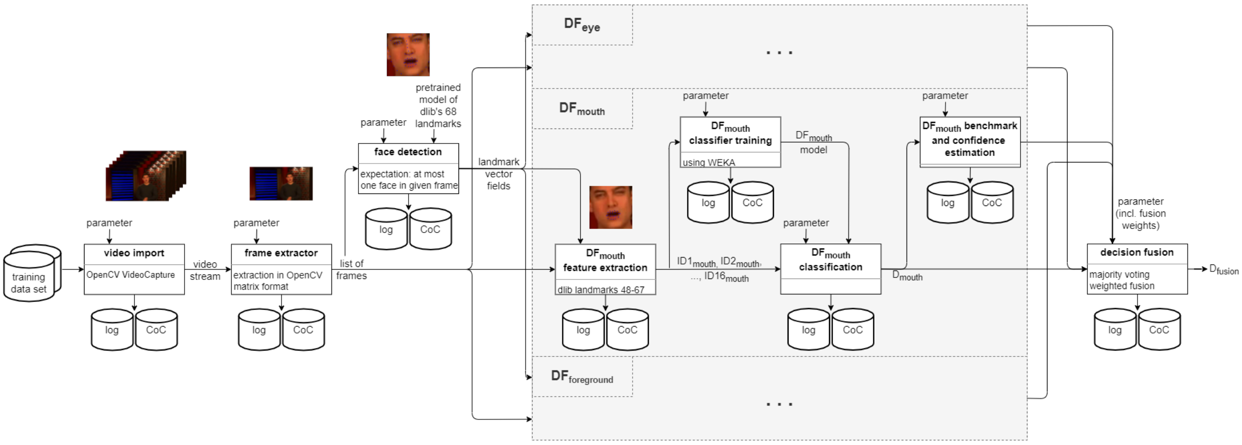

DF$_{eye}$
. . .
DF$_{mouth}$
pretrained model of dlib's 68 landmarks
parameter
face detection
expectation: at most one face in given frame
log
CoC
landmark vector fields
parameter
video import
OpenCV VideoCapture
training data set
video stream
parameter
frame extractor
extraction in OpenCV matrix format
list of frames
parameter
DF$_{mouth}$ classifier training
using WEKA
DF$_{mouth}$ model
DF$_{mouth}$ feature extraction
dlib landmarks 48-67
ID1$_{mouth}$, ID2$_{mouth}$, ..., ID16$_{mouth}$
parameter
DF$_{mouth}$ classification
D$_{mouth}$
parameter
DF$_{mouth}$ benchmark and confidence estimation
parameter (incl. fusion weights)
decision fusion
majority voting weighted fusion
D$_{fusion}$
DF$_{foreground}$
. . .

Figure 3. Illustration of the DeepFake detection pipeline described in [4] (exemplified using the first frame of file Celeb-real ID0_0000 [39]) in its **templating** in the forensic process model phase of **Strategical Preparation (SP)**.

The second episode (hereafter called *instantiation*) corresponds to a set of actual investigations, e.g., determining whether a DeepFake manipulation occurred in a video file or stream. Considering the pipeline presented in Figure 1, these investigations cover all phases from OP to Documentation.

Here, a prepared (as well as benchmarked and potentially certified) template from SP is filled with life by invoking the corresponding orchestration of operators on the assigned processing nodes. Decision models pre-trained in SP are loaded (as shown in Figure 4), together with the used pre-processor and classifier parameters. Thus initialised, the operators are then applied to the input data to the process (MFDT1) to determine traces or information relevant for the investigation at hand.

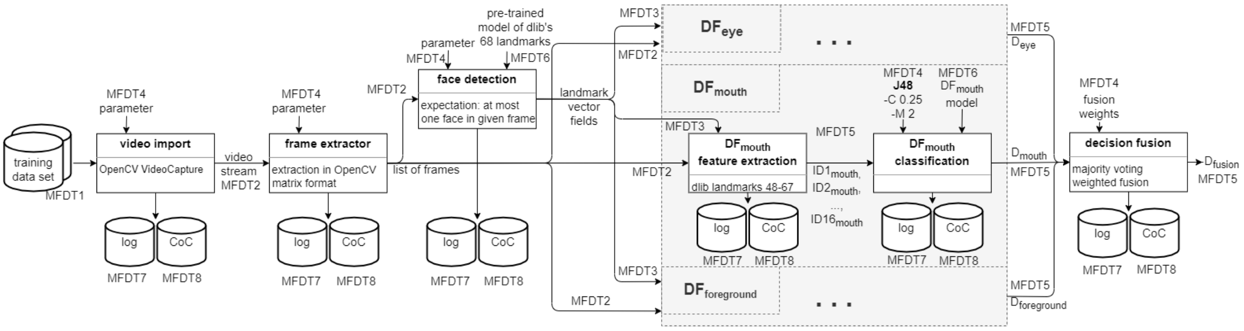


Figure 4. Illustration of the DeepFake detection pipeline from [4], **instantiated** in the forensic process model phase of **Operational Preparation (OP)**, with the inclusion of occurring data types described in [5].

Since sophisticated analysis pipelines will have to rely on information fusion (i.e., the combination of multiple expert systems; see Section 2.2.2), additionally the required fusion weights required for this purpose have to be loaded from the materials prepared in SP.

In addition to the preparation of the templates for actual investigation pipelines in the SP, corresponding documentation packages are also prepared and constantly updated. When a template is then instantiated for a case in OP, the required documentation packages are marshalled together into the investigation accompanying documentation of the case.

4.3. Evaluation Best Practices and Publicly Available Benchmarking Data Sets

As final part of the materials and methods for this paper, in the following sub-sections, first some evaluation best practices are summarised, together with a discussion on suitable metrics, followed by an survey on existing data sets for DeepFake detector benchmarking.

4.3.1. Evaluation Best Practices

As correctly summarised in [6], fusion-based detection methods are required for forensic DeepFake detection to acknowledge the fact that "[...], *a skilled attacker, aware of the principles on which forensic tools work, may enact some counter-forensic measure on purpose* [...]". The fusion is therefore not only intended to boost the overall detection performance (at the cost of an higher run-time complexity), but also to improve [...] "*robustness with respect to both casual and malicious disturbances*" [6]. While fusion is widely believed to be strongly beneficial to decision problem solution approaches like pattern recognition or anomaly detection, publications like [40] point out that information fusion, which indeed has an huge potential to improve the accuracy of pattern recognition systems, is still very hesitantly applied in the forensic sciences. The reason given is, that a potentially negative impact on the classification accuracy, if wrongly used or parameterised, as well as the increased complexity (and the inherently higher costs for plausibility validation) of fusion are in conflict with the fundamental requirements for forensics. To overcome this hesitation, the typical solution is to:

- Very thoroughly benchmark under different training and evaluation scenarios (see [4]) the individual expert systems (here detectors) to be used in the fusion to precisely establish their requirements and capabilities as well as the error rates attached.
- Benchmark different fusion schemes under different training and evaluation scenarios (see [40]) and establish the impact of different weighting strategies onto the (detection) performance and error patterns.
- Consider decision confidences (where available) into the opinion forming.
- Allow for auditability as well as human oversight for the entire process.

Especially the last item, the aspect of required human oversight is a recent trend for critical AI applications (including forensics) which is, among other regulations, manifested in the current initiative towards an Artificial Intelligence Act (AIA), see [41,42].

A very important issue regarding the benchmarking of pattern recognition-based expert systems is the usage of a fair performance evaluation metric. Here, it is proposed to use the Kappa statistics κ instead of the accuracy. It is basically a single-rater version of Cohen's Kappa (see [43,44]) in the range $[-1, 1]$. Therefore, the Kappa statistic measures the agreement of prediction with the true class (i.e., the agreement normalised for chance agreement). The following equation shows the computation of the Kappa statistics κ for an n-class problem:

$$\kappa = \frac{1}{n} \sum_{a=1}^{n} \frac{P_a - P_{chance}}{1 - P_{chance}} \tag{1}$$

For each of the n classes, P_a is the corresponding percentage agreement (e.g., between the classifier and ground truth) and P_{chance} is the probability of chance agreement. Therefore, $\kappa = 1$ indicates perfect agreement and $\kappa = 0$ indicates chance agreement for the overall classification. Only in rare cases negative κ values are achieved, i.e., the classification performance of a system is worse than simple guessing at the class. This is most likely the case when the model was trained to distinguish between patterns completely different than the ones actually presented in the evaluations.

For equally distributed classes, P_{chance} for all classes is simply $\frac{1}{n}$. For differently distributed classes, [44] describes different methods of how to calculate estimate P_{chance}. For the computation of the Kappa statistics within this paper, the WEKA implementation [45] is used, estimating Kappa from the distribution of the classes in the supplied test set.

By using Kappa statistics, it is possible to construct for classification-based investigations a degree of closeness of measurements of a quantity to its actual (true) value that is exempt from the influence of the probability of guessing correctly. Such a metric does allow for direct comparison between the classification performances of classifiers on problems of different numbers of classes.

Regarding the interpretability of Kappa κ, ref. [46] presents a mapping between the Kappa value and the agreements of the different raters (see Table 3). Within this paper, the fact is used that it is actually known in the benchmarking performed in SP to which class an input belongs in the evaluations performed. Based on this realisation, here the Kappa values are mapped onto statistical confidence using the mapping defined in Table 3.

Table 3. Kappa values, agreements according to [46] and the statistical confidence mapping used in this paper.

Kappa Value κ	Agreement According to [46]	Confidence Mapping Used Here
$\kappa < 0$	No agreement	Poor
$0 \leq \kappa < 0.2$ $0.2 \leq \kappa < 0.4$	Slight agreement Fair agreement	Poor to fair
$0.4 \leq \kappa < 0.6$ $0.6 \leq \kappa < 0.8$	Moderate agreement Substantial agreement	Fair to good
$0.8 \leq \kappa \leq 1.0$	Almost perfect agreement	Good

The usage of Kappa in research is not without controversy. Authors like Sim et al. [47] argue that: "[. . .], *the magnitude of kappa is influenced by factors such as* [. . .] *the number of categories* [. . .]". Furthermore, Kappa is generally not easy to interpret in terms of the precision of a single observation, because according to [48], the standard error of the measurements would be required to interpret its statistical significance. To address this problem, Sim et al. propose in [47] multiple evaluations as the basis for the construction of a confidence interval around the obtained value of Kappa, to reflect sampling errors.

Both facts (implicit influence of the number of classes as well as the standard error in the measurement) are also considered here. In the statistical confidence mapping introduced for this paper, the first fact should be negligible for the practical investigations, because the number of classes considered (and therefore assumedly also their implicit influence) is exactly the same (i.e., two). Regarding the second fact, here the actual classes in the investigations are actually known in the benchmarking performed in SP, which solves part of this problem. Regarding the precision, it is assumed here (based on the achieved evaluation results in initial tests) that it is high enough to allow for meaningful investigations (i.e., the corresponding confidence interval would be suitably small).

In spite of the drawbacks that might be attached to the usage of Kappa, Sim et al. [47] argue that: "*If used and interpreted appropriately, the kappa coefficient provides valuable information on the reliability of data obtained with diagnostic and other procedures* [...]."—which is exactly the motivation why Kappa is used in this paper for benchmarking and weight estimation purposes instead of the mere classification accuracy.

4.3.2. Publicly Available (Benchmarking) DeepFake Data Sets

The rapidly growing research efforts for the detection of DeepFakes result in the creation of DeepFake data sets, which are, on the one hand, usable for the implementation and training of new DeepFake detectors. On the other hand, they are needed for benchmarking approaches for detectors. Recent work has shown that data sets are necessary, which include specimen generated with more then one of the generation approaches for DeepFake videos.

In this scope, TIMIT-DF [49] and UADFV [34] were the first publicly available data sets. The number of identities (TIMIT-DF: 43, UADFV: 49) in those databases are very low, as are the perceptual qualities of these DeepFakes. These are the reasons why they have become less relevant during the last years. The Face-Forensics++ [50] data set uses four different creation methodologies (Face2Face, DeepFake, FaceSwap, NeuralTextures). It also increases the amount of DeepFake videos to 1000 per generation algorithm. Note that it is not known to the authors of this paper how many identities have been used to build those 1000 videos.

Not really a part, but provided by the same creators as Face-Forensic++ is the data set Google-DFD [51], which contains 28 identities. It is one more example for a set that includes more than one DeepFake generation approach. Li et al. [39] introduced Celeb-DF with 59 identities. It consists of three parts: Celeb-real (590 videos), YouTube-real (300 videos) and Celeb-synthesis (5639 videos), whereas Celeb-real is used for the DeepFake videos in Celeb-synthesis. Furthermore, the authors in [39] proposed the grouping of several DeepFake data sets into different generations (using the number of frames of the videos in a set): Generation 1 consist of the data sets TIMIT-DF, UADFV and Face-Forensics++. Generation 2 consists of Google-DFD and Celeb-DF. Additionally, the DFDC-Preview data set of Dolhansky et al. [52], which contains of 66 identities, is classified in the second generation of DeepFake data sets, later. This data set is the first part of the DeepFake Detection Challenge (DFDC), which will introduce a newer generation (generation 3) of DeepFakes in [53]. The authors adopt the classification attributes of Li et al. [39]. Every data set which has a total frame amount of 10,000,000 or more as well as 10,000 videos or more is grouped into this third generation. The DFDC data set is another example for a data set that is build with more than one DeepFake generation method. In the time of the publication of Dolhansky et al. [53], there were only two data sets which belonged to this generation: DeeperForensics-1.0 [54] with 100 identities and the DFDC data set of Dolhansky et al. [53] with 960 identities. DeeperForensics-1.0 [54] includes adversarial attacks in DeepFake videos (e.g., added noise, blur, compression), aiming at making detection attempts more realistic (i.e., considering an attacker that tries to hide the traces of the DeepFake attack).

Newer data sets are more and more specific for a defined use case which complicates the grouping of those new databases into the old generations. For example FakeAVCeleb [55] has the size for a generation 3. However, this data set has a different structure. It combines real and fake media as well as image/video and audio data in different ways for 490 identities. It contains video DeepFakes with real audio, audio DeepFakes with real video and also DeepFakes consisting of fake audio and fake video.

The real part is reused from VoxCeleb2 [56]. Furthermore, its authors try to increase diversity (in terms of ethnic backgrounds, ages, and gender). DeepFakeMnist+ [57] is a small DeepFake data set which tries to reproduce different emotions or face movements. While in the previously named data sets the DeepFakes are (mostly) created by the data set authors themselves, WildDeepfake [58] consisted of 707 collected DeepFake videos (plus corresponding benign counterpart video) from the internet (i.e., representing an 'in the wild' set of mixed/heterogeneous generation methods). The advantage of this set creation strategy is the diversity of different forgery techniques, which are also examined by Zi et al. [58]. Kwon et al. [59] generates 175,776 fake clips from 62,166 real clips with 403 (mostly Koreans) identities with different generation models. They also labelled their data set using categories for *age*, *sex* and *location*. Jain et al. [60] used for the data set DF-Mobio 72 identities. It is also divided into 31,950 real videos and 14,546 DeepFake videos. Note that the real videos are taken from the Mobio data set in McCool et al. [61]. This data set contains videos which are taken by the cameras of mobile devices (i.e., smartphones). Table 4 summarises those data sets regarding to the amount of identities and the real and DeepFake video size.

Table 4. Overview on existing publicly available video reference data sets for DeepFake detection (in case of an '?' in the table, the number of individuals has not been documented for this data set). The TIMIT-DF and Celeb-DF data sets used in this paper are marked in bold.

Data Set	# Individuals	# Real Video	# DeepFake Video
UADFV [34]	49	49	49
TIMIT-DF [62,63]	43	559	640
FaceForensics++ [50,64]	?	1000	4000
DFD [51]	28	363	3068
Celeb-DF [39]	59	890	5639
DFDC [53]	960	23,654	104,500
DeeperForensics [54]	100	50,000	10,000
WildDeepfake [58,65]	?	3805	3509
DeepFakeMnist+ [57]	?	10,000	10,000
FakeAVCeleb [55]	490	20,000+	20,000+
KoDF [59]	403	62,166	175,776
DF-Mobio [60]	72	31,950	14,546

In addition to these dedicated DeepFake databases, a huge number of publicly available face video databases have also been created in other research domains, which can be used to represent the other class (here, non-DeepFake or genuine face videos). Those data sets can be used to design different training and testing scenarios, to be able to establish facts about the generalisation power of the detectors trained, which is an important aspect of the quality assessment for every method. Such evaluations would have to be performed as part of quality assurance in the strategic preparation (SP) phase of each forensic process.

5. Application of the Updated Process Modelling to Describe a Fusion-Based DeepFake Detector

The following two sub-sections summarise the work performed, split into the part done in Strategic Preparation (SP, see Section 5.1) and the one started in Operational Preparation (OP) and then conducted throughout the gathering, investigation and analysis phases of a media forensic process (see Section 5.2).

The test procedures and criteria used here are measurements on the detection performance using Kappa statistics and a discussion of the impact of similarity or dissimilarity of training and test data on the detection performance.

The detection methods used here (and discussed in detail in Sections 5.1.2 and 5.1.3) are admittedly not amongst the most sophisticated detectors currently available, but the general performance shown, including problems with the generalisation power are representative for the current situation in this field of applied pattern recognition.

5.1. Templating (In SP) the Empirical Investigations for This Paper

The research performed in [4] is here extended by adding two additional detectors, which have to be included into the benchmarking and fusion accordingly. Figure 5 visualises the additions to the template made in contrast to the original template presented in Figure 3 above.

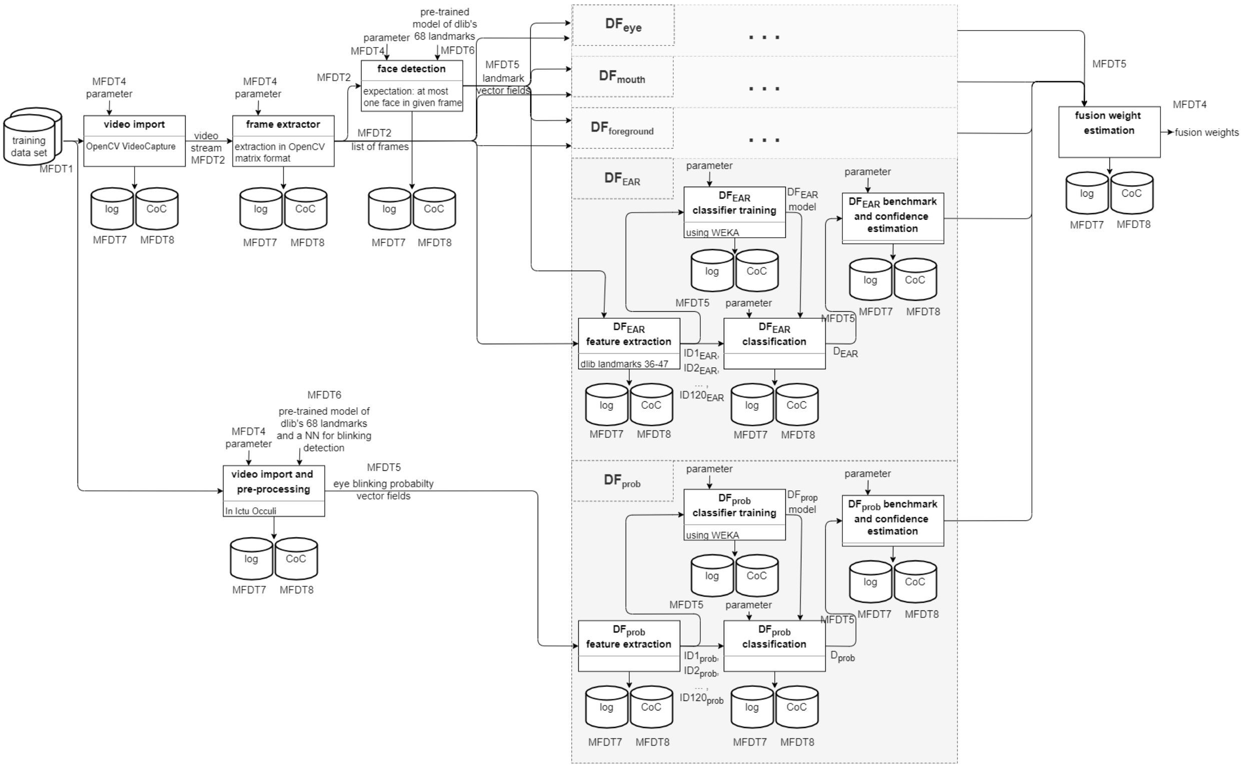


Figure 5. Illustration of the DeepFake detection pipeline **template** (based on Figure 3 above and [4]) created in **Strategic Preparation (SP)** for the usage for the experiments in this paper.

The detector DF_{prob} has an separate pre-processing pipeline (described in Section 5.1.3 below) while all four other detectors share the same pre-processing pipeline.

5.1.1. Data Sets Used for Training and Benchmarking

Here, from the long list of available data sets (as summarised in Table 4 in Section 4.3.2 above), the same ones are re-used here as in [4] for the necessary training operations of detectors, their benchmarking, the determination of fusion weights and the evaluation of fusion approaches. This is done to keep the results comparable (and is not intended to imply a specific fitness/quality of these sets): The Celeb-DF data set is split into disjointed subsets labelled Celeb-real, Celeb-synthesis and YouTube-real. For training purposes, hereafter referred to as $Celeb_{train}$, Celeb-real and a subset of Celeb-synthesis of 590 videos (which represents the number of samples in Celeb-real) are taken for the training of models.

The rest is reserved for the evaluations performed in Section 5.2. In addition, a second benchmarking round is performed using the whole TIMIT-DF data set, consisting of 559 real and 640 DeepFake videos.

5.1.2. Pre-Existing Detectors Re-Used in This Paper

In [4], a total of three different detectors for DeepFake detection were presented, analysing different video areas (eyes, mouth and image foreground). While the detector based on the mouth region showed the best individual results ($\kappa = 0.89$) in evaluations, the detectors based on eyes and image foreground ($\kappa = 0.42$) also revealed potential for the detection of DeepFakes with ($\kappa = 0.38$ and 0.42, respectively).

5.1.3. Detectors Newly Implemented for This Paper

Besides the three detectors from previous work (see Section 5.1.2), two new detectors (one hand-crafted and one neural network based) aiming at DeepFake detection based on eye blinking inconsistencies are proposed here. For both the pre-processing is performed frame by frame. First, the face is detected using dlibs 68 landmarks [66], with each eye represented by 6 key points [4]. For comparison purposes, the facial region is resized to an area of 256 × 256 pixels. For the hand-crafted approach, the so-called eye-aspect ratio (EAR), given by $EAR = \frac{|p_2 - p_6|| + ||p_3 - p_5||}{2||p_1 - p_4||}$ [35], is calculated as the first representation for each eye (EAR_l and EAR_r).

For the neural network-based approach, the model presented in [31] by Li et al. is used. To define the degree of aperture of each eye, the bounding box given by the corresponding 6 landmarks is determined. This is followed by probability estimation of the eye blinking based on a Long-term Recurrent Convolutional Network (LRCN), using the bounding box of each eye as input. As a result of both pre-processing approaches, two vectors (one for each eye) are given, representing the EAR and probability of eye blinking, respectively.

For the following classification based on these pre-processed signals, only a hand-crafted approach is taken due to time constraints. In addition to EAR_l and EAR_r, two additional representations of the blinking behaviour as difference quotient ($diff_l(i) = EAR_l(i) - EAR_l(i-1)$ and $diff_r(i) = EAR_r(i) - EAR_r(i-1)$) for each eye are computed using consecutive video frames (i) and $(i-1)$. Based on [35], the values for the following statistical descriptors μ (arithmetic means), $\mu - \sigma$, $\mu - 2\sigma$, $\mu - 3\sigma$, $\mu + \sigma$, $\mu + 2\sigma$, $\mu + 3\sigma$ are calculated, where σ is an a posterior determined standard deviation. In Figure 6, those descriptors are plotted as horizontal lines. Using these seven different descriptors, two different types of features are derived: noise estimators (here, a set of crossing rates) and energy estimators (here, equivalent to the area under a curve, representing the aperture of the eye). Additional features are derived trying to iteratively estimate the skew of the distribution (which has to be assumed since the eye is usually opened for a longer time than it is closed do to blinking). Including (μ) and (σ), this results in a set of 30 features that are gathered for each of the four representations. Then, the four sub-vectors are concatenated in a fixed sequence, which in turn results in a vector of 120 features length.

What can be derived from Figure 6 is the fact that hand-crafted feature designs, like the ones discussed above, can very well represent semantic characteristics of the signal (here, the blinking of an eye in a video stream) with easily interpretable features.

The classification models used are acquired by training five different machine learning algorithms, specifically NaiveBayes [67,68], LibSVM [69], SimpleLogistic [70–72], JRip [73,74] and J48 [75,76] (the WEKA implementations of these pre-existing algorithms in their standard configurations; note: these classifiers were selected by the authors on basis of previous work, for future work, more sophisticated classifier selection and parameter optimisation could be required), of which the best one in terms of κ value is selected for the investigation process template created. In the comparison of these five different machine learning methods, no significant differences in run-time for the model training was detectable. Only with respect to detection performance, given using Kappa statistics κ, differences are noticeable. Table 5 shows the performances of each model and highlights the best one per detector, which then is used in the evaluations performed.

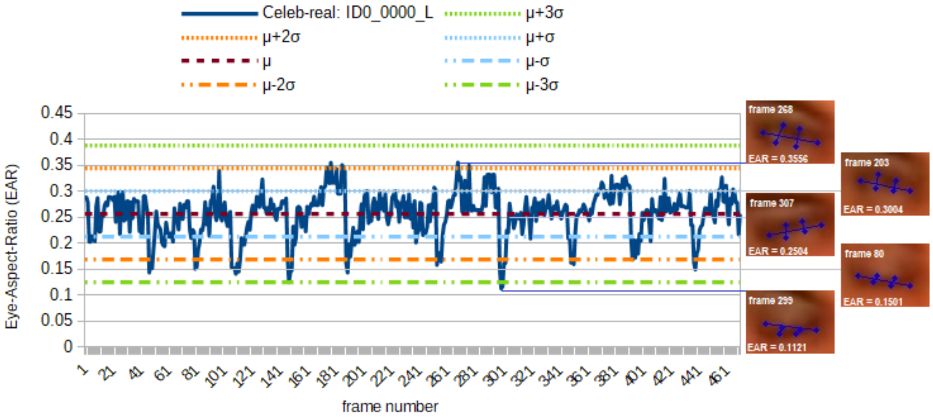


Figure 6. Illustration of the eye-aspect ratio (EAR) and the proposed descriptors (eye aperture analysed for the left eye of file Celeb-real ID0_0000 [39], the small sub-figures on the right hand side showing the corresponding aperture using segments of selected frames of that video).

Table 5. Achieved κ values for each classification model using $Celeb_{Train}$ in 10-fold cross validation.

Detector	NaiveBayes	LibSVM	Simple Logistics	JRip	J48
DF_{EAR}	0.0695	0.3254	0.3508	**0.3678**	0.2966
DF_{prob}	0.2162	0.0063	**0.3275**	0.2480	0.2273

5.1.4. Fusion Operators and Weight Estimation

In [4] a total of five different fusion strategies (feature-level fusion as well as decision-level in forms of simple majority voting and weighted-average fusion) have been shown to increase the performance over single detectors. In line with this approach, the strategies simple majority voting and weighted fusion are re-used here, but in contrast to [4] the weights are determined in SP phase using κ instead of the decision accuracy. As discussed in Section 4.3.1, Kappa is a more fair approach regarding unevenly split data sets and, as shown in Section 4.3.2, most public available data sets have a higher amount of DeepFake videos. Since there are five detectors, the weights have to be adjusted based on the results on the training data set $Celeb_{train}$ resulting in $w_{eye} = 0.1590$, $w_{mouth} = 0.3724$, $w_{foreground} = 0.1757$, $w_{EAR} = 0.1548$ and $w_{prob} = 0.1381$. A detailed comparison of the individual Kappa values for each individual detector and its derived weights can be found in Table 6.

Table 6. Overview of the results of each individual detector and the derived fusion weights.

Detector	κ	Fusion Weight
DF_{eye} [4]	0.38	$w_{eye} = 0.1590$
DF_{mouth} [4]	0.89	$w_{mouth} = 0.3724$
$DF_{foreground}$ [4]	0.42	$w_{foreground} = 0.1757$
DF_{EAR}	0.37	$w_{EAR} = 0.1548$
DF_{prob}	0.33	$w_{prob} = 0.1381$

As proposed in Section 4.3.1, a detector should only be included if at least a significant confidence mapping is achieved. For forensic field applications this value should obviously be well above $\kappa = 0.95$, in academic research it should be at least in the range of "fair to good" with $\kappa \geq 0.4$.

Here, for the sake of illustrating the effect of this decision, this rule is violated on purpose. As a result the detector DF_{eye} as well as the two new detectors DF_{EAR} and DF_{prob} are included in the template as inputs to the fusion despite the fact that their performances in benchmarking are below $\kappa = 0.4$.

5.2. Instantiation of the Pipeline for the Evaluations in This Paper

Figure 7 shows the instantiation of the pipeline in OP for the evaluations performed in this paper. In contrast to Figure 4, it shows the extension made since [4] (i.e., the addition of two additional detectors). Since the whole pre-processing is equivalent to the one performed in SP (see Figure 5), this part is omitted here.

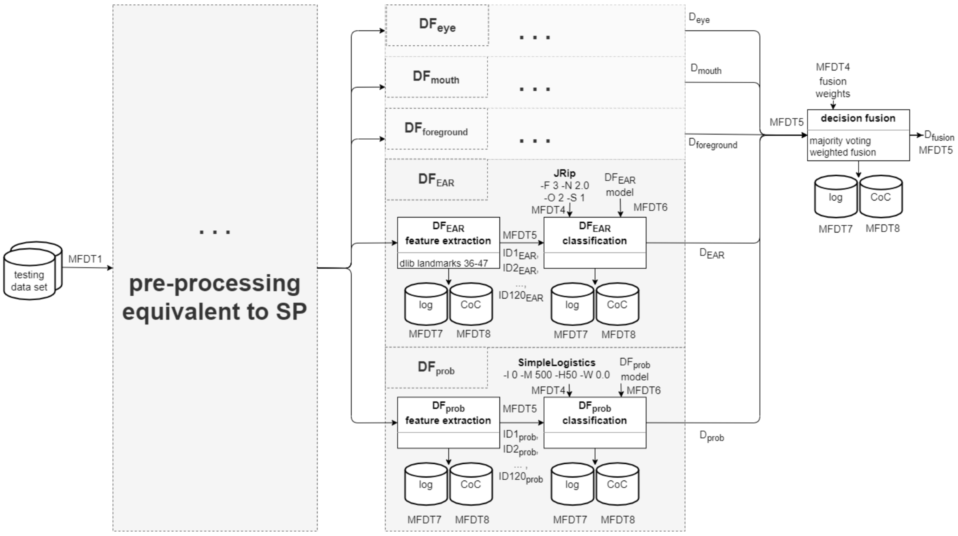


Figure 7. Illustration of the DeepFake detection pipeline in this paper, based on [4], **instantiated** in the forensic process model phase of **Operational Preparation (OP)**, with the inclusion of occurring data types described in [5].

5.2.1. Data Sets Used for Evaluation

Here, from the long list of available data sets (as summarised in Section 4.3.2 above), the following are reserved for the evaluation operations of the individual detector, YouTube-real (300 real videos) and a part of Celeb-synthesis (5049 DeepFake videos), this set $Celeb_{Test}$ is disjointed from the data used for training, benchmarking and weight estimation purposes in Section 5.1 above. In addition, the benchmarking for the fusion is done on the same data set as in [4] and contains a total of 120 real and 120 DeepFake videos of Celeb-DF.

5.2.2. Single Detector Evaluation Results

The evaluation of the $Celeb_{Test}$ shows in Table 7 a drastic decrease in Kappa for both DF_{EAR} and DF_{prob}, reducing it to almost zero. To further investigate the causes, the corresponding curves for a DeepFake video and its corresponding source are compared. As shown in Figure 8 (top row), the curves for both detectors appear to be almost identical, showing that the synthesis method used in Celeb-DF is able to reproduce the natural blink behaviour.

Table 7. Overview of the results for the proposed individual detectors on the specified testing data set.

Detector	κ on $Celeb_{Test}$	κ on TIMIT-DF
DF_{EAR}	0.0191	0.552
DF_{prob}	0.0408	0.433

The tests on TIMIT-DF show a different result, both in terms of generated curves as well as Kappa values. Both DF_{EAR} and DF_{prob} show Kappa above 0.4, with 0.55 and 0.43, respectively. In addition, there are also major differences in the curves to be seen in Figure 8, and thus showing that the DeepFake blinking behaviour in this data set is clearly distinguishable from real blinking. The comparison of higher and lower quality DeepFakes of TIMIT-DF further shows that the higher quality is closer to a real blink. Taking into account the generations in which both data sets are placed, it can either be seen as a potential flaw in earlier generations or it could be caused by the generation method (which is not considered in the generations specified).

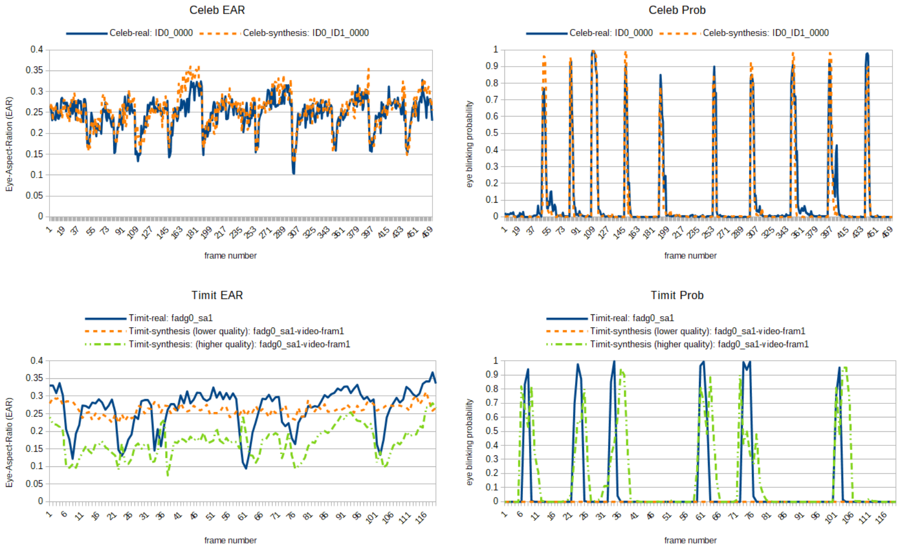


Figure 8. Comparison of the acquired blinking curves for a DeepFake and its source based on EAR (**left**) and probabilities (**right**) for both considered data sets Celeb-DF (**top**) and TIMIT-DF (**bottom**).

5.2.3. Results of the Fusion-Based Detection

As shown in Table 8, the inclusion of the new detectors results in a slight drop in the average detection performance. For the majority voting, $\kappa = 0.542$ is determined, in contrast to the previously achieved performance of $\kappa = 0.725$, a drop in performance of 0.18. Presumably, this can be explained by the use three of the five detectors individually showing $\kappa < 0.4$ in DF_{eye}, DF_{EAR} and DF_{prob}, all based on eye features outvoting the other two (performance-wise better) detectors.

In the case of the weighted fusion, the drop is smaller, but still recognisable (by 0.02 from 0.808 to 0.783). The optimal decision threshold for the fusion operator is determined iteratively here. It is noticed in these experiments that a shift in optimal threshold value for the classification shifts from 0.65 to 0.5 occurs. This new threshold could be equivalent to DF_{mouth} (with a weight of 0.3724) agreeing with at least one other detector, which is very similar to the fusion outcome shown in [4].

Table 8. Overview of the results for the fusion strategies in comparison to previously achieved results given in [4].

Fusion Strategy	κ
majority voting (old detectors) [4]	0.725
weighted fusion (old detectors) [4] threshold = 0.65	0.808
majority voting (5 detectors)	0.542
weighted fusion (5 detectors) threshold = 0.5	0.783

6. Results

This chapter provides a brief summary of the results, before the following Section 7 projects the conclusions onto the contributions identified in Section 1.

6.1. Experimental Evaluation Results and Comparison with the SOTA/Related Work

It is apparent that the detection performances of the detectors used in this paper are not fit to compete with the best detection performances presented in the state-of-the-art publications on detector designs. However, it can be assumed that the findings presented here generalise as follows:

- The comparison of the investigation results and the differences experienced when looking at the performances on the TIMIT-DF and Celeb-DF data sets indicate a sensitivity of trained detection approaches to specific DeepFake generation methods. In consequence two alternative strategies for compensating this sensitivity should be explored: Generalisation or specialisation of the training scenario for detectors. For the first alternative, training sets with large heterogeneous DeepFake parts would be required, potentially resulting in models with a high false positive rate due to the fact that the model component(s) characterising the DeepFake class are very dispersed in the feature space. For the second alternative, targeted training for the different DeepFake generation would be required, effectively transforming the task into an n-class problem.
- Extensive benchmarking of detectors is required for any application of forensic methods. What is true for single detectors, becomes even more relevant when combining single expert systems into a fusion approach. The practical evaluations summarised in Section 5.2 above show how adding two detectors, which are performing individually better than the probability of guessing correctly (which would be $\kappa = 0$), negatively impairs a fusion outcome. What has not been reflected upon in the discussions made in Section 5 is that the question of fairness and bias are also becoming much more complex in the context of fusion: Out of the five detectors used within this paper, three are concentrating on the eye regions. This effectively leverages the weight estimation for fusion weights, which were made under the implicit assumption of the independence of involved detectors.

6.2. Lessons Learned during the Templating and Instantiating of the Pipeline in SP/OP

When reflecting the work presented in this paper on the three specific aspects for current research needs according to the whitepaper "Secure, robust and traceable use of AI-problems, procedures and actions required" [1] as discussed in Section 1, the following can be summarised: Instead of focusing on research on effective countermeasures (i.e., DeepFake detectors), like most of the scientific papers currently published, the work presented is focusing on the other two aspects: first, supporting the development of standards, technical guidelines, test criteria and test methods as well as, second, the research into methods of transparency and explainability.

The efforts invested in the Strategic Preparation (SP) of a forensic process are assumed to prepare for effective response in case of an incident. They are intended to increase forensic readiness of response and investigation units as well as strengthening the whole field by providing standardised (and certified) methods and procedures.

Reflecting this basic principle of a hard split into SP and operations (OP, DG, DI, DA and DO) into academic research might seem weird at first, but is, in the opinion of the authors, a step that might in the long run help to bridge the gap between academic research in media forensics and the practitioners requiring court admissible methods.

What this split is supposed to provide are more precise process descriptions, which can easier be verified by third parties. Furthermore, they make training, benchmarking and testing procedures more transparent and are thereby supposed to better allow the identification of influence factors, training bias and potential error sources.

In addition to the templating and instantiation considerations regarding the design and implementation of processes, the modelling work presented has a second relevant aspect: the domain-adapted media forensic data types presented in Table 2. With their help, the actual data flows in complex systems, such as police case management systems, should become manageable. More details regarding this data model, its usage and benefits are presented in [5].

7. Conclusions and Discussion

Drawing the conclusions from the work presented and projecting them to the contributions identified in Section 1, it has to be said that:

The **need for modelling forensic processes** is reasoned upon, with a brief overview over forensic process modelling requirements and some best practices for media forensics (in Section 2.1). It can be (and is, e.g., in [1]) summarised as: (a) the development of standards, technical guidelines, test criteria and test methods; (b) research into effective detectors/countermeasures and (c) research into methods of transparency and explainability of AI-driven methods. Out of these three well-grounded needs, current research (esp. academic research) in media forensics focuses mainly on (b), ignoring the fact that without also achieving the other two, the required degree of maturity for court room acceptability will not be achieved. This paper tries to highlight this gap and facilitate the understanding between the media forensic research community and practitioners in the field of applied forensics.

A **concept for modelling media forensic investigation pipelines** is derived from established guidelines. Due to the nationality of the authors, this concept is derived from long standing German guidelines on IT forensics, which are extended here to better fit the specifics in the field of media forensics. By doing so, the authors do not claim that this starting point, published by the German Federal Office for Information Security (BSI) in 2011, is the most suitable choice (which it obviously is not), but acknowledge the fact that regulation concerning the admissibility of procedures and methods happens on a national level. The introduced approach to modelling investigation pipelines focuses on a two-step procedure: First, in a preparatory step called here Strategic Preparation (SP), the planning or *templating* of an investigation pipeline, combined with all organisational, technical and personnel steps required for implementing one or multiple pipelines of this nature (see Section 4.2) happens. The operations in this phase would include among other things the certification or investigation methods and procedures as well as the training of corresponding experts. In the following step, here called Operational Preparation (OP), the previously prepared pipelines are *instantiated* as required, i.e., used in a standardised way to perform specific investigations.

Despite the fact that the work presented in this paper is still a rough sketch on the actual work required to get methods 'court ready', it gives an idea on the required next steps after a technical solution (e.g., detector) has been found fit for publication by its authors.

The **applicability of the introduced modelling is illustrated** on the example of a media forensic investigation pipeline focusing on the detection of DeepFake videos, extending previous work of the authors on possible fusion-driven detection pipelines. The results show after adding two further detectors, which where in benchmarking in SP on purpose wrongfully determined to be suitable ($\kappa \sim 0.4$), a drop of the detection results in the experiments. This implies that the benchmarking strategies used here still leave significant room for improvement.

The **benefits of such a planned realisation of AI-based investigation methods** are discussed to some extend. Here, it is apparent that these discussions only cover the tip of the iceberg! One recent trend of how to counter the issue of manipulations is well summarised in [6] by the following statement: "*Face manipulation brings an array of complex legal issues. There is no comprehensive legislation on the use of manipulated images, yet several aspects are already regulated in various countries. It should hence not surprise that the development of new manipulation technology and the detection thereof also leads to new issues and questions from a legal perspective which deserve further research. If it is used to mislead, manipulated images can cause significant harm* [. . .] *In some countries, altered (body) images used for commercial purposes (such as the fashion industry) need to be labelled. More generally, legislative proposals in several countries try to tackle the transparency issue by imposing an obligation to inform users that they interact with AI-generated content (such as DeepFakes)*". However, this implicitly only white hat application of methods like DeepFakes. No (criminal or other) threat actor will adhere to such an obligation when spreading fake news or other media-related manipulations. As a consequence, entities such as news agencies strongly relying on media objects submitted from external sources would also require mature manipulation detection mechanisms that would have to be integrated into their already established source (material) verification routines. The exact extent and scope of such analysis methods and 'filters', their transparency and fairness, as well as their potential impact to public and politic debates are currently a hot debate especially in Europe (see for example [77] for the discussion of free speech implications of Article 17 (regulating upload filters) of the EU 'Directive on copyright and related rights in the Digital Single Market' as adopted in 2020).

8. Future Work

Instead of focusing on research on effective DeepFake detectors, the work presented is concentrating on the two aspects of supporting the development of standards, technical guidelines, test criteria and test methods as well as the research into methods of transparency and explainability. Despite the fact that the detectors in this paper were mostly used for illustrative purposes, their quality of course also has to be enhanced, either by improving the existing detectors or including better ones into the fusion-driven decision system. The next step along these empirical lines would then be the design of best-practices for evaluations, focusing on data sets, first allowing for more realistic error rate estimates (e.g., 'ìn-the-wild' sets (eventually also including counter-forensics), like [53,54,58]) and second for fairness considerations (e.g., looking at challenging data sets like [55,59] to determine racial bias). As shown in Table 4, a wide range of suitable data sets is currently available for such a purpose.

In addition to those single-classifier benchmarking aspects, also bias evaluations regarding the fusion would be necessary to be performed: In our illustrative example pipeline, the five detectors were chosen and combined in a way that is overvaluing/biased towards parts of the signal (here, the eye regions the video-which is significantly over-represented with three out of the five detectors focusing on this small part of the video) effectively counteracting the actual weighting done for the fusion. This might first seem a unlikely situation in practice but with many neural network driven detection methods, it remains unclear what exactly the features learned are actually representing. Therefore, such situations are a threat that is, in the opinion of the authors, likely to occur with learned feature spaces and that is so far going mostly unnoticed by many practitioners. As a consequence it would be required to understand such learned feature spaces to avoid such kind of bias.

Returning to the infrastructure considerations dominating the work in this paper, two separate aspects are discussed below: first, perspectives for extending the presenting modelling and evaluation work, and second, the big issue of demystifying modern machine learning and AI systems.

8.1. Extending the Presented Modelling and Evaluation Work

As indicated by the results discussed in Section 7 regarding the perceived drop in the fusion performance after the adding of two assumed good new detectors, one of the most important next steps regarding the introduced approach is the design of a **practical benchmarking framework for single detectors in Strategic Preparation (SP)**. This would allow for a more fine-granular detector evaluation and corresponding fusion operator design and parameterisation (including the fusion weights). Such a benchmarking framework would have to consider a wide range of data sets, classified using the four different concepts for DeepFake generation (facial reenactment, facial replacement (or face swapping), face editing and face synthesis) as well as clearly identified types of traces imposed to the media objects by the corresponding modifications. Only with such a framework, necessary reasoning regarding performance influencing factors as well as bias and/or fairness issues (see e.g., [55,59]) can be performed.

In addition to the escalation of the benchmarking extent, also the basic strategies should be revised. Here, in accordance with the work presented in [60], **a re-modelling of the detection/classification problem as an n-class problem** (where n corresponds to the number of different DeepFake creation strategies, see Section 2.2) might become necessary. This paradigm shift is assumedly strongly beneficial to the detection performance as well as the interpretation of error behaviours (i.e., the decision plausibility and transparency). Conceptually, it allows us to handle each of the different creation strategies as what they are: different manipulation pipelines leaving clearly distinguishable artefacts or traces. Technically, it would allow to train much more precise decision models for each DeepFake creation strategy, instead of representing them all as different sub-spaces of the class 'DeepFake' in the currently trained models.

A second important issue for future work is the extension of **investigations into error, loss and uncertainty in the forensic processes** as motivated in previous work (esp. [25]). This requires research efforts especially in the field of demystifying AI system decisions (see Section 8.2 below), not only for classical decision methods with hand-crafted features, but also for the more recent approaches relying of neural network to learn feature spaces that lack intuitive interpretation.

Third, but most important, increasing the maturity of approaches requires the **extension of the work from modelling into practical frameworks**. Here, joint efforts with system developers (e.g., for police case management systems) as well as certification bodies would be required to achieve this goal. An very interesting success story in this regard is the following: Regarding digital camera forensics a major breakthrough can be seen in the law case *United States of America v. Nathan Allen Railey* (United States District Court for the Southern District of Alabama (for a short summary of the relevant part of the court case see [78])). In the Daubert hearings of this case, the method of digital camera authentication based on intrinsic characteristics of its image acquisition sensory developed by Jessica Fridrich and her university research group (see e.g., [79]) got accepted for the first time as forensic evidence. The FBIs Forensic Audio, Video, and Image Analysis Unit (FAVIAU) established in the Daubert hearings that this approach (and the corresponding tool 'FindCamera' developed and evaluated in a public private partnership effort lead by the FBI and the US Airforce Research Labs) meets all necessary Daubert criteria and the presiding judge furthermore decided that this evidence (or more precisely the FBI expert testimony based on this media forensic analysis) also meets the FRE rule 702 criteria.

As discussed above in Section 2.1.2 forensics are entirely governed by national legislation. Therefore, this requires nation specific efforts to get such methods and procedures court-ready.

8.2. Demystifying Machine Learning and AI Systems

To demystify machine learning, a comparison of the advantages and disadvantages of each individual decision forming method as well as each trained model is required. Besides the previously mentioned aspects of detection and generalisation performances and the dimensionality and composition of the feature space, further aspects such as training duration (including training success estimates) and model complexity (including its impact on explainability) have to be considered. Modern neural network based analysis and detection methods show impressive results regarding detection results on data similar to the training data used to create the network. What are still issues for research are the generalisation power of such methods (i.e., how well the systems perform on previously unseen data) as well as the transparency, correctness and fairness of their decisions.

Recent research is looking especially into these questions of understanding the inner workings of black box neural networks, e.g., by determining the most important neurons in a network and deducing knowledge from those analyses. Some papers, like [80], even extend into automated linguistic annotation methods to provide better understandable descriptions of internal workings.

Author Contributions: Conceptualisation, C.K.; data curation, D.S. and S.S.; formal analysis, C.K.; funding acquisition, C.K. and J.D.; investigation, C.K., D.S. and S.S.; methodology, C.K. and J.D.; software, D.S. and S.S.; supervision, C.K. and J.D.; validation, C.K., S.S. and J.D.; visualisation, D.S.; writing—original draft, C.K.; writing—review and editing, C.K., D.S., S.S. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: The work in this paper is funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number FKZ: 13N15736.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this work the following the pre-existing reference databases have been used for our evaluations: TIMIT-DF [49,62] and Celeb-DF [39]. They are publicly available at: https://www.idiap.ch/en/dataset/deepfaketimit (accessed on 15 April 2022), https://github.com/yuezunli/celeb-deepfakeforensics (accessed on 15 April 2022).

Acknowledgments: The authors wish to thank Stefan Kiltz for the discussions on the DCEA and its applicability as well as Robert Altschaffel and Mario Hildebrandt for discussions on domain adaptations required for forensic data models.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bundesamt für Sicherheit in der Informationstechnik (BSI). *Sicherer, Robuster und nachvollziehbarer Einsatz von KI-Probleme, Maßnahmen und Handlungs-Bedarfe*; BSI: Bonn, Germany, 2021.
2. Champod, C.; Vuille, J. Scientific Evidence in Europe-Admissibility, Evaluation and Equality of Arms. *Int. Comment. Evid.* **2011**, *9*. [CrossRef]
3. BSI. *Leitfaden IT-Forensik*; German Federal Office for Information Security: Bonn, Germany, 2011.
4. Siegel, D.; Kraetzer, C.; Seidlitz, S.; Dittmann, J. Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features. *J. Imaging* **2021**, *7*, 108. [CrossRef]
5. Siegel, D.; Krätzer, C.; Seidlitz, S.; Dittmann, J. Forensic Data Model for Artificial Intelligence based Media Forensics-Illustrated on the Example of DeepFake Detection. In *Media Watermarking, Security, and Forensics 2022-Electronic Imaging 2022*; Alattar, A., Nasir Memon, G.S., Ed.; Society for Imaging Science and Technology IS&T: Springfield, VA, USA, 2022.
6. Rathgeb, C.; Tolosana, R.; Vera-Rodriguez, R.; Busch, C. (Eds.) *Handbook of Digital Face Manipulation and Detection From DeepFakes to Morphing Attacks*; Springer: Berlin/Heidelberg, Germany, 2022.
7. Ho, A.T.S.; Li, S. *Handbook of Digital Forensics of Multimedia Data and Devices*; Anthony, T.S., Li, S., Eds.; Department of Computing and Surrey Centre for Cyber Security (SCCS), University of Surrey: Guildford, UK; Wiley/IEEE Press: Hoboken, NJ, USA, 2015.
8. U.S. Congress. *Federal Rules of Evidence*; Amended by the United States Supreme Court Apr. 26, 2011, eff. Dec. 1, 2011; U.S. Congress: Washington, DC, USA, 2011.

9. SWGFAST. *The Fingerprint Sourcebook*; Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST); National Institute of Justice, U.S. Department of Justice: Gaithersburg, MD, USA, 2011.
10. LLI. *Federal Rules of Evidence-FRE702*; Legal Information Institute, Cornell Law School (LLI): Ithaca, NY, USA, 2010.
11. LLI. *Federal Rules of Evidence-Notes on FRE702*; Legal Information Institute, Cornell Law School (LLI): Ithaca, NY, USA, 2010.
12. Krätzer, C. Statistical Pattern Recognition for Audio-Forensics-Empirical Investigations on the Application Scenarios Audio Steganalysis and Microphone Forensics. Ph.D. Thesis, Otto-von-Guericke-Universität Magdeburg, Magdeburg , Germany, 2013.
13. USC. *United States Court (USC) 509 U.S. 579*; Daubert v. Merrell Dow Pharmaceuticals, Inc.: Washington, DC, USA, 1993.
14. USCA. *United States Court of Appeals (USCA), Ninth Circuit. No. 90–55397*; Argued and Submitted March 22, 1994. Decided January 4, 1995, 1995. Daubert, William and Joyce Daubert, individually and as Guardians Ad Litem for Jason Daubert, (a minor); Anita De Young, individually, and as Guardian Ad Litem for Eric Schuller, Plaintiffs-Appellants, vs. Merrell Dow Pharmaceuticals, Inc., a Delaware corporation, Defendant-Appellee; USCA: San Francisco, CA, USA, 1995.
15. U.S. Congress. *Frye v. United States, 293 F. 1013 (D.C. Cir.)*; U.S. Congress: Washington, DC, USA, 1923.
16. Meyers, M.; Rogers, M. Computer Forensics: The Need for Standardization and Certification. *Int. J. Digit. Evid.* **2004**, *3*, 1–11.
17. Nelson, B.; Phillips, A.; Steuart, C. *Guide to Computer Forensics and Investigations*, 4th ed.; Course Technology: Boston, MA, USA, 2010.
18. Bijhold, J.; Ruifrok, A.; Jessen, M.; Geradts, Z.; Ehrhardt, S.; Alberink, I. Forensic audio and Visual Evidence 2004–2007: A Review. In Proceedings of the 15th INTERPOL Forensic Science Symposium, Lyon, France, 23–26 October 2007.
19. Daeid, N.N.; Houck, M. (Eds.) *Interpol's Forensic Science Review*; Taylor & Francis Inc.: Abingdon, UK, 2010.
20. Ashcroft, J.; Daniels, D.J.; Hart, S.V. *Forensic Examination of Digital Evidence: A Guide for Law Enforcement*; U.S. Department of Justice-National Institute of Justice: Washington, DC, USA, 2004.
21. Casey, E. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*; Academic Press: Cambridge, MA, USA, 2011.
22. Bartholomew, P. Seize First, Search Later: The Hunt for Digital Evidence. *Touro Law Rev.* **2014**, *30*, 1027–1052.
23. Daniel, L.E.; Daniel, L.E. *Digital Forensics for Legal Professionals: Understanding Digital Evidence from the Warrant to the Courtroom*; Syngress: Washington, DC, USA, 2015.
24. Pollit, M.; Casey, E.; Jaquet-Chiffelle, D.O.; Gladyshev, P. *A Framework for Harmonizing Forensic Science Practices and Digital/Multimedia (OSAC Technical Series Publication 0002R1)*; Organization of Scientific Area Committees (OSAC): Gaithersburg, Maryland, DC, USA, 2019.
25. Kiltz, S. Data-Centric Examination Approach (DCEA) for a Qualitative Determination of Error, Loss and Uncertainty in Digital and Digitised Forensics. Ph.D. Thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, Magdeburg, Deutschland, 2020.
26. Kiltz, S.; Dittmann, J.; Vielhauer, C. Supporting Forensic Design-A Course Profile to Teach Forensics. In Proceedings of the 2015 Ninth International Conference on IT Security Incident Management and IT Forensics, Magdeburg, Germany, 18–20 May 2015; pp. 85–95.
27. Flaglien, A.; Sunde, I.M.; Dilijonaite, A.; Hamm, J.; Sandvik, J.P.; Bjelland, P.; Franke, K.; Axelsson, S. *Digital Forensics*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2017.
28. Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
29. Di Filippo, M.; Froede, S. *Deep(C)Phishing: Next Level Vishing & Phishing*; 18. Deutscher IT-Sicherheitskongress 1.-2. Februar 2022-DIGITAL; Bundesamt für Sicherheit in der Informationstechnik (BSI): Bonn, Germany, 2022.
30. Nguyen, T.T.; Nguyen, C.M.; Nguyen, D.T.; Nguyen, D.T.; Nahavandi, S. Deep Learning for Deepfakes Creation and Detection: A Survey. *arXiv* **2019**, arXiv:1909.11573.
31. Li, Y.; Chang, M.; Lyu, S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv* **2018**, arXiv:1806.02877.
32. McCloskey, S.; Albright, M. Detecting GAN-generated Imagery using Color Cues. *arXiv* **2018**, arXiv:1812.08247.
33. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 38–45.
34. Yang, X.; Li, Y.; Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, UK, 12–17 May 2019; pp. 8261–8265. [CrossRef]
35. Jung, T.; Kim, S.; Kim, K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access* **2020**, *8*, 83144–83154. [CrossRef]
36. Soukupová, T.; Cech, J. *Real-Time Eye Blink Detection Using Facial Landmarks*; In Proceedings of the 21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia, 3–5 February 2016.
37. Altschaffel, R. Computer Forensics in Cyber-Physical Systems: Applying Existing Forensic Knowledge and Procedures from Classical IT to Automation and Automotive. Ph.D. Thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, Magdeburg, Germany, 2020.
38. Zhang, C.; Liu, C.; Zhang, X.; Almpanidis, G. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Syst. Appl.* **2017**, *82*, 128–150. [CrossRef]

39. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 3204–3213. [CrossRef]
40. Kraetzer, C.; Makrushin, A.; Dittmann, J.; Hildebrandt, M. Potential advantages and limitations of using information fusion in media forensics a discussion on the example of detecting face morphing attacks. *EURASIP J. Inf. Secur.* **2021**, *2021*, 9. [CrossRef]
41. European Commission. On Artificial Intelligence-A European Approach to Excellence and Trust. COM(2020) 65 Final. 2020. Available Online: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed on 14 September 2021).
42. European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM(2021) 206 Final. 2021. Available Online: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 (accessed on 14 September 2021).
43. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. 02000104. [CrossRef]
44. Eugenio, B.D.; Glass, M. The Kappa Statistic: A Second Look. *Comput. Linguist.* **2004**, *30*, 95–101. [CrossRef]
45. Frank, E.; Hall, M.A.; Witten, I.H. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2016.
46. Landis, J.; Koch, G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
47. Sim, J.; Wright, C.C. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Phys. Ther.* **2005**, *85*, 257–268. [CrossRef] [PubMed]
48. Reichow, B. *Evidence-Based Practices and Treatments for Children with Autism*; Springer: Berlin/Heidelberg, Germany, 2011.
49. Korshunov, P.; Marcel, S. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *arXiv* **2018**, arXiv:1812.08685.
50. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 1–11. [CrossRef]
51. Dufour, N.; Gully, A. Contributing Data to Deepfake Detection Research. 2019. Available online: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html (accessed on 9 September 2021).
52. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Canton Ferrer, C. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv* **2019**, arXiv:1910.08854.
53. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Canton Ferrer, C. The DeepFake Detection Challenge Dataset. *arXiv* **2020**, arXiv:2006.07397.
54. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; pp. 2886–2895. [CrossRef]
55. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. *arXiv* **2021**, arXiv:2108.05080.
56. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. *arXiv* **2018**, arXiv:1806.05622.
57. Huang, J.; Wang, X.; Du, B.; Du, P.; Xu, C. DeepFake MNIST+: A DeepFake Facial Animation Dataset. *arXiv* **2021**, arXiv:2108.07949.
58. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.G., WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020 Association for Computing Machinery: New York, NY, USA, 2020; pp. 2382–2390.
59. Kwon, P.; You, J.; Nam, G.; Park, S.; Chae, G. KoDF: A Large-Scale Korean DeepFake Detection Dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10744–10753.
60. Jain, A.; Korshunov, P.; Marcel, S. Improving Generalization of Deepfake Detection by Training for Attribution. In Proceedings of the International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 6–8 October 2021.
61. McCool, C.; Marcel, S.; Hadid, A.; Pietikäinen, M.; Matejka, P.; Cernocký, J.H.; Poh, N.; Kittler, J.; Larcher, A.; Lévy, C.; et al. Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, Melbourne, VIC, Australia, 9–13 July 2012; pp. 635–640.
62. Sanderson, C.; Lovell, B. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. *Lect. Notes Comput. Sci.* **2009**, *5558*, 199–208. [CrossRef]
63. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast Face-Swap Using Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 3697–3705. [CrossRef]
64. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv* **2018**, arXiv:1803.09179.
65. Zi, B.; Ma, X.; Chang, M.; Chen, J.; Jiang, Y.G. Deepfake In The Wild. Available online: https://github.com/KnightofDawn/deepfake_in_the_wild (accessed on 29 March 2022).
66. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
67. John, G.H.; Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*; Morgan Kaufmann: San Mateo, CA, USA, 1995; pp. 338–345.

68. Weka documentation for NaiveBayes. Available online: https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/NaiveBayes.html (accessed on 25 March 2022).
69. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27:1–27:27. [CrossRef]
70. Landwehr, N.; Hall, M.; Frank, E. Logistic Model Trees. *Mach. Learn.* **2005**, *59*, 161–205. [CrossRef]
71. Sumner, M.; Frank, E.; Hall, M. Speeding up Logistic Model Tree Induction. In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, 3–7 October 2005; pp. 675–683.
72. Weka Documentation for SimpleLogistic. Available online: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/SimpleLogistic.html (accessed on 25 March 2022).
73. Cohen, W.W. Fast Effective Rule Induction. In Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufmann, Tahoe City, CA, USA, 9–12 July 1995; pp. 115–123.
74. Weka Documentation for JRip. Available online: https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html (accessed on 25 March 2022).
75. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
76. Weka Documentation for J48. Available online: https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/J48.html (accessed on 25 March 2022).
77. Øe, H.S. *Opinion of Advocate General-Case C-401/19 Republic of Poland vs European Parliament*; Council of the European Union: Luxembourg, 2021.
78. Kirby, B. *Expert Witnesses Link Camera to Child Porn Found on Defendant Nathan Railey's Laptop*; Advance Local: New York, NY, USA, 2011.
79. Goljan, M.; Fridrich, J.J.; Filler, T. Large scale test of sensor fingerprint camera identification. In Proceedings of the Media Forensics and Security I, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, 19 January 2009; Delp, E.J., Dittmann, J., Memon, N.D., Wong, P.W., Eds.; Volume 7254, p. 72540. [CrossRef]
80. Hernandez, E.; Schwettmann, S.; Bau, D.; Bagashvili, T.; Torralba, A.; Andreas, J. Natural Language Descriptions of Deep Visual Features. *arXiv* **2022**, arXiv:2201.11114.

13

[Siegel23b] Pros and Cons of Comparing and Combining Hand-Crafted and Neural Network based DeepFake Detection based on Eye Blinking Behavior

This chapter of this cumulative habilitation treatise has originally been published as:

[Siegel23b] Dennis Siegel, Stefan Seidlitz, **Christian Kraetzer**, Jana Dittmann: *Pros and Cons of Comparing and Combining Hand-Crafted and Neural Network based DeepFake Detection based on Eye Blinking Behavior*. Proc. Electronic Imaging, vol. 35(4):pp. 378-1–378-1, 2023. doi: 10.2352/EI.2023.35.4.MWSF-378
https://library.imaging.org/ei/articles/35/4/MWSF-378

Shares/author contributions (as given in the original paper - see page 221 of this cumulative habilitation treatise): "*Initial idea & methodology: Jana Dittmann (JD), Stefan Seidlitz (StS) and Dennis Siegel (DS); Conceptualization: Stefan Seidlitz (StS), Dennis Siegel (DS) and Christian Kraetzer (CK); Empirical work on hand-crafted approaches: DS; Empirical work on deep learning driven approaches: StS; Writing – original draft: StS; Writing – review & editing: DS, CK and JD.
All authors have read and agreed to the published version of the manuscript.*"

https://doi.org/10.2352/EI.2023.35.4.MWSF-378

Pros and cons of comparing and combining hand-crafted and neural network based DeepFake detection based on eye blinking behavior

***Dennis Siegel*[1]*, Stefan Seidlitz*[1]*, Christian Kraetzer*[1]*, Jana Dittmann*[1]

[1] ***Otto-von-Guericke University, Magdeburg, Germany***

Abstract

Temporal feature spaces are a promising approach for DeepFake detection, since DeepFake synthesis is most often done on a frame-by-frame basis. With the existing and upcoming regulations on European level, the EU General Data Protection Regulation (EU GDPR) and Artificial Intelligence Act (AIA) in particular, data minimization and decision transparency are of concern also for such media forensic methods. In order to bring these aspects together, this paper utilizes two different algorithms both analyzing the eye blinking in the videos. The first one is implemented using deep learning to predict blinking behavior. It shows challenges of hyper-parameter tuning for the training of such a model. The second detector uses an existing hand-crafted approach to identify a suitable number of frames (i.e., video duration) required to reliably detect DeepFakes. Considering GDPR concerns, an optimal trade-off between detection performance and data minimization is found in the range of 35 to 40 seconds of video, giving a detection accuracy of 96.88% for the DeepFakes tested.

Introduction and Motivation

DeepFakes present a recent advancement in technology enabling manipulations in digital media that focus on the replacement of a face in a video by another face. They have a wide area of use cases and their intent is not always clear, as they may also have positive aspects that need to be considered [19]. In particular the usage as a privacy enhancement technique (PET) has to be named here [6]. Regardless of their use case, DeepFakes should be identifiable, to detect and prevent their misuse, which requires suitable detection approaches. In general, these can be categorized according to temporal and spatial methods. This division goes hand in hand with image or video DeepFakes. Spatial methods utilize image manipulation detection techniques. In contrast, temporal methods have stricter requirements of inputting a video and potentially higher computational costs. Their suitability is given due to flaws / restrictions in current DeepFake synthesis methods. This is due to the fact that most DeepFake synthesis methods are working frame by frame, creating temporal anomalies in video streams. [24]

In this paper the focus is on temporal methods. It contains the following contributions: First, the evaluation of a deep learning based eye blinking predictor. Second, the identification of medical concerns regarding blinking and development of privacy enhancement strategies. Third, the identification of suitable video duration thresholds for DeepFake detection using eye blinking.

State of the art in DeepFake detection

A wide variety of different approaches for DeepFake detection has been introduced in literature. Mirsky and Lee [24] categorize detection approaches based on spatial and temporal features. Furthermore, the approaches are divided by them into hand-crafted and deep features. A similar survey overview can be found in Nguyen et al. [26], where the separation is done based on image- and video-based techniques, without further splitting based on the used machine learning techniques. In Yu et al. [41] the separation is solely done for DeepFake videos. Again, the categories are similar, including approaches for both spatial and temporal features. Although spatial approaches are also important (especially forensic approaches focusing on individual images), they are outside the scope of this paper. Instead, the following sections present selected approaches to DeepFake detection using hand-crafted and deep learning based temporal approaches.

DeepFake detection using hand-crafted feature spaces

In general, it is difficult to separate approaches based on the categories of ‘hand-crafted’ and ‘deep learning’. There are various combinations of both modalities by introducing hand-crafted feature spaces, which are classified by deep learning [3, 7]. In terms of traditional machine learning classification, most hand-crafted detectors utilize support vector machines (SVM) [23, 39].

Agarwal et al. present DeepFake detection based on lip synchronization, by comparing the spoken word sounds (phonemes) with mouth movements in video (viseme) [3]. The evaluation is done both manually, by introducing a human operator labeling frames and automated using a convolutional neural network (CNN). In addition, the detection performance is evaluated based on video duration.

In [34] three hand-crafted detectors are proposed based on eye, mouth and the comparison of foreground and background to detect DeepFakes. While these detectors did not yield acceptable detection performances individually, a decision-level fusion increased the performance. In [19] both an hand-crafted and deep learning based feature extractor are used to detect DeepFakes based on inconsistencies in eye blinking behavior.

DeepFake detection using deep learning feature spaces

Established images based DeepFake detectors are by reason of the video compression not always applicable for video data, because video compression results in strong degradation within the video frames [2]. Furthermore, most neural networks based detectors (e.g. [21], [25] or [32]) solely detect DeepFakes based on individual frames. In consequence, it is possible that contiguous frames of DeepFake videos results in inconsistencies between the frames which are in certain circumstances not visible by the human eye. In the area of neural networks those temporal artifacts are detectable by recurrent network structures. For example, Korshunov et al. [18] used a Long Short-Term Memory (LSTM) architecture to detect inconsistencies between the audio and video stream. For the audio stream they used Mel frequency cepstral coefficients (MFCC) and for the video stream they calculate 42 distances between mouth keypoints of the 68 landmarks dlib model [17]. Güera et al. [13] combine in a convolutional LSTM the spatial dimension using Convolutional Neural Networks (CNNs) and the temporal dimension using LSTM to analyze coherence inconsistencies between the frames. Another recurrent network structure is the Gated Recurrent Unit (GRU) network which is used by Sabir et al. [31]. They first cropped the frames to the facial area which are then compared by the GRU network to detect temporal discrepancies across the frames.

Motivated by the fact that the human blinking behavior was not or less present in first DeepFake videos, Li et al. [20] proposed a LSTM based blinking detector. They combined the LSTM layer with a convolutional layer to detect closed or opened eye states in the faces of all video frames. Newer DeepFake generation approaches solved the problem of missing blinking events within the video. The detector of Li et al., also known as In Ictu Oculi, is not able to differentiate between a real and a fake eye blink event. Only videos without blinking events allows the detector to classify those videos as fake. Further, the detector was not tested on DeepFakes which are not generated by the DeepFake tool used by its authors. An implementation of In Ictu Oculi is provided by its authors on GitHub[1] but this version only works as a blinking detector, not being able to differentiate between real and fake videos (it only returns open or closed eye states for the frames within a video with a probability between 0 and 1 but no indication of whether this implies a DeepFake or not).

Regulatory Requirements and their Impact to Feature Space Design

Additional requirements for AI applications (such as media forensics methods and frameworks) conditions are established by legislation at the European level. One such regulation was introduced with the EU General Data Protection Regulation (EU GDPR, [10]). It addresses general principles of data protection in terms of data collection and processing. In particular, the following three (out of seven) principles are of importance ([10]):

- Lawfulness, fairness and transparency: "*Processing must be lawful, fair, and transparent to the data subject.*"
- Purpose limitation: "*You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.*"
- Data minimization: "*You should collect and process only as much data as absolutely necessary for the purposes specified.*"

In addition, Article 9 of the GDPR states: "*Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.*" [10]

Another regulation relevant in the context of this paper is the upcoming EU Artificial Intelligence Act (AIA) [11], addressing the usage of AI systems. One aspect of particular importance is the criterion of human oversight in using AI systems (Article 14). This is supposed to lead to a reduction of black-box algorithms and enforces human-in-the-loop and human-in-control aspects for AI systems. In addition, Article 52 Paragraph 3 of the current AIA draft states, that DeepFakes must be marked as such [11].

In this paper, as underlying forensic process model, the principles established in the best practice guidelines on IT forensics of the German BSI (German Federal Office for Information Security) [5] (German: "*Leitfaden IT-Forensik*") are used. This best practice document provides various means for modeling forensic processes, including the definition of a generic phase-driven investigation & reporting model, a basic data model and a classification of methods and tools. Like many other best practice documents in this field it covers basic investigation principles, process models, forensic data types, etc. but does not provide domain specific process models and guidelines for specific media forensic investigations such as DeepFake detection. Here, existing research, such as the latest extension to the BSI guidelines [5] described as the Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling (DCEA) summarized in [16] and [35], is used as basis for extending the scope of these guidelines to achieve a higher degree of maturity for the state of the art in taylor-made models for media forensics (incl. DeepFake detection).

The core of DCEA has three main components: a model of the *phases* of a forensic process, a classification scheme for *forensic method classes* and *forensically relevant data types*. The six DCEA *phases* are briefly summarized as: Strategic preparation (SP), Operational preparation (OP), Data gathering (DG), Data investigation (DI), Data analysis (DA) and Documentation (DO). At this point only the importance of the SP has to be pointed out, since it is the phase that also includes all research and evaluation activities considered in this paper. For further details on the phase model as well as the method classes and data types, the reader is referred, e.g. to [16].

Privacy concerns in the evaluation of biometric data

The human face is an often used biometric trait, that besides the ID also reveals other information about the person. Even pictures of parts of the face allow to derive personal attributes like the gender, age, ethnical background, etc. as well as certain health issues [37]. The work presented in that paper indicates that it is possible to identify illnesses such as glaucoma and cataracts based even on single images. Furthermore, there are various studies addressing the aspect of spontaneous eye blinking. On average, a human blinks around 10 to 15 times a minute (i.e., once every 4 to 6 seconds [1]). In a study by Sforza et al. [33] it was identified,

[1] https://github.com/yuezunli/WIFS2018_In_Ictu_Oculi

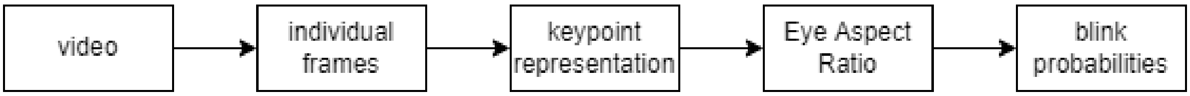


Figure 1. *Reduction steps taken for data acquisition*

that woman blink more frequently than men and it further differentiates based on age. In contrast, babies and children blink less frequent with around 2 times a minute. In addition, the blinking frequency can be affected by external influences, such as talking [36]. Another study by Jung et al. [15] identified a correlation between children frequently blinking and tic disorders.

Based on that, in conjunction with the previously discussed Article 9 of the GDPR, it is important to protect these personal attributes and prevent misuse or (unintended) information leakage. In general, there are three different possibilities to handle privacy concerns:

- all critical aspects are available to all
- features are overwritten by default parameters
- critical aspects are excluded, removed or overwritten

The first approach does not exclude personal attributes, instead it makes them available to all entities with access to the data set. This might require extensive labeling and also the agreement of the subjects of each sample. The privacy enhancement can be done either on feature or image level. On feature level, one possibility would be to overwrite features by default values. Relevant features have to be identified, that enable deriving personal attributes. As stated by Angwin et al. [4] personal attributes do not rely on individual features, but rather a correlation of multiple features. Lastly, critical aspects could be excluded, removed or overwritten. One possible approach for this is by using semantic image inpainting [40]. By now there are various existing privacy preserving methods, such as de-identification of facial images [8]. Othman and Ross [27] use morphing techniques to change the appearance of an face image. By using both a male and female image in the morph process they preserve the identity, but change the gender. Also DeepFake synthesis can be used for this purpose. In [6] its usage on social media is discussed, to anonymize faces in online media. For this purpose, the faces are replaced selectively based on the degree of acquantance, so to the user unknown faces are anonymized.

Although using image inpaiting or DeepFake to secure privacy in the video database seems most appropriate, it is not currently possible to use these techniques for the task of DeepFake detection. One reason for this is that the methods cause a change in the data and thus real training data might been changed by this method and then have to be regarded as DeepFake. To mitigate the downside for DeepFake detection, it is necessary to restore the original media of the synthesis. Based on a recent DeepFake challenge by Guarnera et al. [12], one question was to recreate the source image of DeepFake synthesis. Unfortunately no algorithms were submitted for this subtask.

In contrast to the possibilities discussed above, this paper presents an approach of information reduction based on a multi-level representation minimization. As shown in figure 1 a total of five different representations were considered for data extraction. Each reduction step also reduces the amount of information in the corresponding representation. So by changing from frames to keypoint representation for example, the requirement of storing the data as image is removed and replaced by keypoint graphs.

Development of a LSTM Network to predict blinking behavior

The development of the LSTM network based blinking predictor would occur within the strategical preparation (SP) phase of a forensic framework. The proposed forensic pipeline is illustrated in figure 2. The State-of-the-Art section above gives a small overview about existing LSTM approaches, but many more LSTM approaches exists. In consequence it is important to decide which approach is applicable for an eye blinking prediction which come with many different training iterations. A stacked LSTM network consisting of more than one LSTM layer seems the best strategy to train human eye blinking behavior.

For this paper, the training data for the LSTM network is the Celeb-real part of the Celeb-DF [22] data set. In preparation, the eye aspect ratio (EAR) for both eyes in each video is generated, according to the proposed method in [19]. All curves were normalized in the range of 0 and 1, calculated by the lowest and highest eye aspect ratio (EAR) value of all training samples. The prediction utilizes a sliding window approach, where two consecutive windows are taken, the first one for model training and the second for prediction. The window size is calculated as 5 seconds multiplied by 30 frames, which is the median frame rate of all Celeb-real videos. In other words the LSTM network was trained on 150 frames to create a prediction for the next 150 frames.

After the LSTM training the aim was to compare the predicted EAR curves with the calculated EAR curve from all videos of the Celeb-DF data set, divided into the three classes Celeb-real, YouTube-real and Celeb-synthesis. The difference between both the calculated and predicted curves is determined by $\sum_{x=s}^{n}(max(calc_x, pred_x) - min(calc_x, pred_x))/n$, with s being the index of the first predicted frame and n the total number of predicted frames. The calculated distance can then be used as feature for DeepFake detection.

Evaluation setup

As indicated above, for the training the Celeb-real part of the Celeb-DF [22] is used. The dlib face detector [17] analyzes all videos to detect faces in every frame of all included videos. In the training phase, a total of 56 videos had to be removed, because the face detection was not successful in several frames. Furthermore, the videos of Celeb-real do not have the same video length and some even had less than 300 frames. Due to the selected window size for the LSTM network of 300 frames, these videos were unusable. Additional 51 videos have been removed from the training data set because they were to short.

Addressing the hyper-parameter tuning for the LSTM network training, different training strategies were carried out. Different LSTM unit amounts were tested from ranging from 100 to 300, different counts of LSTM layers were tested from 2 to 4 layers and also dropout in different strengths from $p = 0.1$ to $p = 0.9$

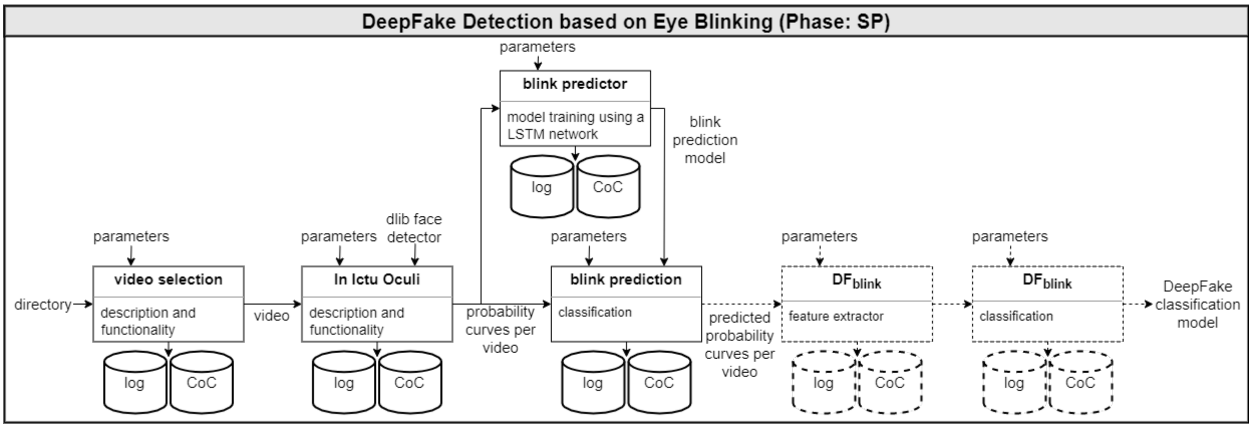


Figure 2. Illustration of the DeepFake detection pipeline used in this paper in its templating in the forensic process model phase of Strategical Preparation (SP). Components outside the scope of this paper are marked by dashed lines.

was inserted after every LSTM layer. The maximal training iterations was adjusted after full convergence of the training loss, which in most cases was after 500 epochs.

Evaluation results

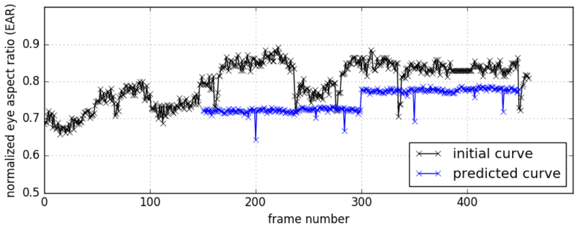


Figure 3. Predicted eye aspect ratio and initial EAR curve on the example of the left eye for the video id13_0008 [22]

Figure 3 shows an initial EAR curve calculated from a Celeb-real video of the Celeb-DF data set and the predicted blinking curve with estimated blinking events of a trained LSTM network. For the model a stacked LSTM network with two LSTM layers followed by a dropout layer with $p = 0.2$ and a final dense layer was trained for 500 epochs. The model by itself is not well trained, but a tendency of the predicted curve is visible and promising. The model predict approximately every 50 to 100 frames a blinking event, considering 30fps approximately every 1.66 to 3.33 seconds. Although the blinking appears to be slightly too often, it can be explained by the fact that the person in the video was talking (which results usually in a slightly increased blinking behavior). The sudden change in predicted values on frame 300 occurs because of a new segment, which is predicted with the real EAR data between frame number 150 and 300. The consequence of the results shown in figure 3 is the insight that further hyper-parameter tuning of the LSTM network is needed, which will also increase the computational cost that has to be invested into this detector in the strategical preparation phase. This highlights an important difference between hand-crafted and deep learning based approaches, namely the scope and depth of hyper-parameter tuning. At the current state of this blink predictor, further hyper-parameter tuning is required, to make the results more reliable.

The second evaluation goal is to identify a suitable video duration to detect DeepFakes based on eye blinking. For this purpose, the DeepFake detector DF_{eye} [34] is used on an in-house data set aggregating data from FaceForensics++ [29, 30], Celeb-DF [22], DFD [9] and HiFiFace [38] (2904 samples in total). The model is trained using the J48 [28] classification algorithm provided by WEKA [14] in its default parameterization and with 10-fold stratified cross-validation. Afterwards, the samples used are analyzed for the impact of the duration on the achieved accuracy. Due to the different frame rates, the optimal length is determined based on the video duration instead of the number of frames.

frames per second	15	18	24	25	29	29-30	30	60
# samples	6	2	413	852	5	19	1596	11

Framerate distribution in the considered data set.

Figure 4 shows the results categorized in 5 second video duration spans and the corresponding number of samples (=videos in the used set) per duration. The peak performance of 96./8% accuracy is achieved for video durations between 35 and 40 seconds (containing 96 samples in the used set). Longer samples first result in slight decrease in accuracy. A perfect classification is then again achieved for samples with an duration of at least 55 seconds, however the amount of samples of this duration (33) is too small to be relevant in the larger picture. The results obtained here suggest, that there is both a minimum (for accurate detection) and maximum (for privacy enhancement purposes) length for videos in the range of 35 to 40 seconds.

Summary, Conclusions and Future Work

This paper shows possibilities and challenges of deep learning approaches for the purpose of DeepFake detection, especially focusing on the relevance of suitable hyper-parameter tuning. The current state of the blinking predictor enables future work to extend the existing approach towards a full blown blinking-based DeepFake detector. This can be used to integrate both hand-crafted and neural network-based methods and evaluate and compare them against each other. Furthermore, the possibility to use both blinking probability curves generated by [20] as well as eye aspect ratios as baseline, allows to consider different representations of data. This enables the comparison of different training

duration (in s)	5-10	10-15	15-20	20-25	25-30	30-35	**35-40**	40-45	45-50	50-55	55-60	60-65	65-70	70-75
# samples	121	1179	707	343	158	107	**96**	49	57	18	17	8	7	1
accuracy (in %)	66.12	73.20	75.53	83.38	84.81	91.59	**96.88**	89.80	92.98	88.89	100	100	100	100

Evaluation results based on an inhouse data set for the detector DF_{eye}.

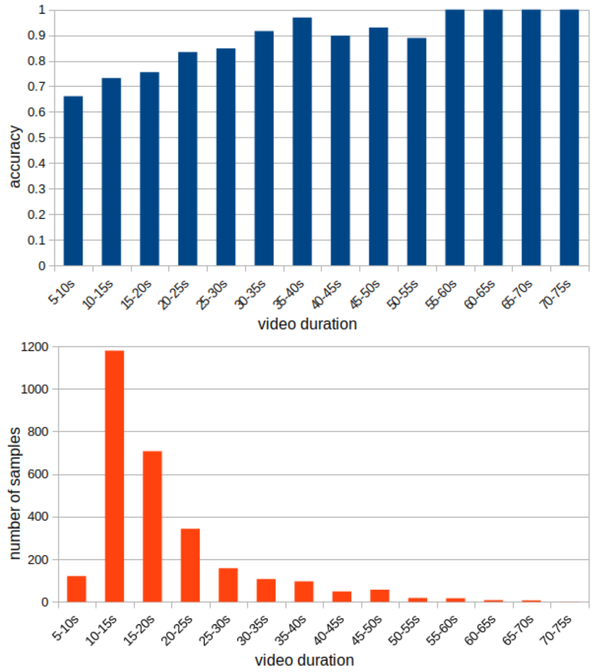


Figure 4. *Evaluation results based on an inhouse data set for the detector* DF_{eye}.

data representations, to evaluate the usage for privacy enhancement against the detection performance.

In addition to the work on blinking prediction, a video duration analysis based on this approach is possible. The experiments performed within this paper established an optimal minimum of 35 seconds and maximum of 40 seconds duration for this particular data set. In general, the human eye blinking and the evaluation itself is influenced by various external factors, such as distance of the person towards the camera and the fact that the person was talking in most samples used. Because of that, more training data is required to also increase the necessary diversity of training and testing material.

Acknowledgements

Funded in parts by the German Federal Ministry of Education and Research (BMBF) under grant number (FKZ): 13N15736 (project "Fake-ID") and in particular for the privacy aspects on the example of medical concerns by the European Union Project "CyberSec LSA OVGU-AMSL" under grant number (FKZ): ZS/2018/12/96222.

Author Contributions: Initial idea & methodology: Jana Dittmann (JD), Stefan Seidlitz (StS) and Dennis Siegel (DS); Conceptualization: Stefan Seidlitz (StS), Dennis Siegel (DS) and Christian Kraetzer (CK); Empirical work on hand-crafted approaches: DS; Empirical work on deep learning driven approaches: StS; Writing – original draft: StS; Writing – review & editing: DS, CK and JD.

References

[1] F.H. Adler and R.A. Moses. *Adler's Physiology of the Eye: Clinical Application*. C.V. Mosby Company, 1981.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018*, pages 1–7. IEEE, 2018.

[3] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 2814–2822. IEEE, 2020.

[4] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, and ProPublika. Machine Bias - ProPublica. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, May 23, 2016. Accessed: 23/06/2021.

[5] BSI. *Leitfaden IT-Forensik*. German Federal Office for Information Security, 2011.

[6] Umur A. Ciftci, Gokturk Yuksek, and Ilke Demir. My face my choice: Privacy enhancing deepfakes for social media anonymization. *CoRR*, abs/2211.01361, 2022.

[7] Umur Aybars Ciftci and Ilke Demir. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. January 2019.

[8] Benedikt Driessen and Markus Dürmuth. Achieving anonymity against major face recognition algorithms. In Bart De Decker, Jana Dittmann, Christian Kraetzer, and Claus Vielhauer, editors, *Communications and Multimedia Security*, pages 18–33, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[9] Nick Dufour and Andrew Gully. Contributing Data to Deepfake Detection Research. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`, September, 24 2019. Accessed: 09/09/2021.

[10] European Commission. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). April, 27 2016. [Online]. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504` [Last retrieved: 12.01.2023].

[11] European Commission. Proposal for a Regulation of the european parliament and of the council Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *COM(2021) 206 final*, April, 21 2021. [Online]. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206` [Last retrieved: 14.09.2021].

[12] Luca Guarnera, Oliver Giudice, Francesco Guarnera, Alessandro Ortis, Giovanni Puglisi, Antonino Paratore, Linh MQ Bui, Marco

Fontani, Davide Alessandro Coccomini, Roberto Caldelli, et al. The face deepfake detection challenge. *Journal of Imaging*, 8(10):263, 2022.

[13] David Güera and Edward J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018.

[14] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor.*, 11(1):10–18, 2009.

[15] Hee-Yeon Jung, Sun-Ju Chung, and Jeong-Min Hwang. Tic disorders in children with frequent eye blinking. *J. AAPOS*, 8(2):171–174, April 2004.

[16] Stefan Kiltz. *Data-Centric Examination Approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[17] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, 2009.

[18] Pavel Korshunov and Sébastien Marcel. Speaker inconsistency detection in tampered video. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2375–2379, 2018.

[19] Christian Kraetzer, Dennis Siegel, Stefan Seidlitz, and Jana Dittmann. Process-driven modelling of media forensic investigations-considerations on the example of deepfake detection. *Sensors*, 22(9), 2022.

[20] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *CoRR*, abs/1806.02877, 2018.

[21] Yuezun Li and Siwei Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts, 2019.

[22] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3204–3213. IEEE, 2020.

[23] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *CoRR*, abs/1812.08247, 2018.

[24] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), January 2021.

[25] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. *CoRR*, abs/1810.11215, 2018.

[26] Thanh Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.*, 223:103525, 2022.

[27] Asem Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. pages 682–696, 09 2014.

[28] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[29] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[30] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1–11. IEEE, 2019.

[31] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *CoRR*, abs/1905.00582, 2019.

[32] Selim Seferbekov. Deepfake detection (dfdc) solution by @selimsef, Jun 2029.

[33] Chiarella Sforza, Mario Rango, Domenico Galante, Nereo Bresolin, and Virgilio Ferrario. Spontaneous blinking in healthy persons: An optoelectronic study of eyelid motion. *Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists)*, 28:345–53, 08 2008.

[34] Dennis Siegel, Christian Kraetzer, Stefan Seidlitz, and Jana Dittmann. Media forensics considerations on deepfake detection with hand-crafted features. *Journal of Imaging*, 7(7), 2021.

[35] Dennis Siegel, Christian Krätzer, Stefan Seidlitz, and Jana Dittmann. Forensic data model for artificial intelligence based media forensics-illustrated on the example of deepfake detection. *Electronic Imaging*, 34:1–6, 2022.

[36] The Healthline Editorial Team. Excessive eye blinking: Causes, diagnosis, treatment and more, Aug 2019. [Online]. Available at: `https://www.healthline.com/health/eye-health/eye-blinking` [Last retrieved: 12.01.2023].

[37] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. *Iris Recognition in Cases of Eye Pathology*. Springer Singapore, Singapore, 2019.

[38] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1136–1142. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[39] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 8261–8265. IEEE, 2019.

[40] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6882–6890. IEEE Computer Society, 2017.

[41] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *IET Biometrics*, n/a(n/a), 02 2021.

Author Biography

__Jana Dittmann__ is a Professor on multimedia and security at the University of Otto-von-Guericke University Magdeburg (OvGU). She is the leader of the Advanced Multimedia and Security Lab (AMSL) at OvGU, which is partner in national and international research projects and has a wide variety of well recognized publications in IT security. __Christian Kraetzer__ is a post-doc researcher and __Dennis Siegel__ as well as __Stefan Seidlitz__ are PhD students at AMSL.

14

[Kraetzer23] Human-in-control and Quality Assurance Aspects for a Benchmarking Framework for DeepFake Detection Models

This chapter of this cumulative habilitation treatise has originally been published as:

[Kraetzer23] **Christian Kraetzer**, Dennis Siegel, Stefan Seidlitz, Jana Dittmann. *Human-in-control and Quality Assurance Aspects for a Benchmarking Framework for DeepFake Detection Models*. Electronic Imaging, vol. 35(4):pp. 379–1–379–1, 2023. doi:10.2352/EI.2023.35.4.MWSF-379.
https://library.imaging.org/ei/articles/35/4/MWSF-379

Shares/author contributions (as given in the original paper - see page 229 of this cumulative habilitation treatise): "*Author Contributions: Initial idea & methodology: Jana Dittmann (JD) and Christian Kraetzer (CK); Conceptualization: Christian Kraetzer (CK), Dennis Siegel (DS), Stefan Seidlitz (StS) and JD; Re-modelling of the process model components: CK, DS, StS; Empirical evaluations: DS; Writing – original draft: CK; Writing – review & editing: DS, StS and JD.*
All authors have read and agreed to the published version of the manuscript."

https://doi.org/10.2352/EI.2023.35.4.MWSF-379

Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models

***Christian Kraetzer*[1]*, Dennis Siegel*[1]*, Stefan Seidlitz*[1]*, Jana Dittmann*[1]**

[1] ***Otto-von-Guericke University, Magdeburg, Germany***

Abstract

Human-in-control is a principle that has long been established in forensics as a strict requirement and is nowadays also receiving more and more attention in many other fields of application where artificial intelligence (AI) is used. This renewed interest is due to the fact that many regulations (among others the the EU Artificial Intelligence Act (AIA)) emphasize it as a necessity for any critical AI application scenario. In this paper, human-in-control and quality assurance aspects for a benchmarking framework to be used in media forensics are discussed and their usage is illustrated in the context of the media forensics sub-discipline of DeepFake detection.

Introduction and Motivation

Intended courtroom usage of forensic methods requires standardized investigation and analysis procedures that underwent quality assurance as well as standardization prior to application to case work. Internationally accepted best practices governing this field are e.g. the United States Federal Rules of Evidence (FRE; esp. FRE 702, see [18]) and the Daubert standard in the US.

Authors like Champod et al. point out that, even if the Daubert standard is only directly legally binding for court proceedings on US federal level, they are also in many other countries worldwide considered as a best practice for evaluation of the degree of maturity of forensic methods as basis for expert testimonies intended to be used in court (see e.g. [3], where the influence of the Daubert standard on the evaluation and admissibility of scientific evidence in Europe is discussed).

Within this paper focusing on the benchmarking of media forensic methods, especially the following three (out of five) Daubert criteria are relevant ([3]):

- "*whether the expert's technique or theory can be or has been tested – that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability*"
- "*the known or potential rate of error of the technique or theory when applied*"
- "*the existence and maintenance of standards and controls*"

Especially the second and the last of the criteria quoted above are of importance within this context, because they imply on one hand a strong need for process modeling as foundation of work in standardization and on the other hand require benchmarking work to allow to suitably measure or estimate the potential rate of error of the method when applied in practice.

Many process models for forensic processes exist for 'traditional' forensic sub-disciplines (e.g. dactyloskopy), with the intended purpose of making corresponding investigations fit for courtroom usage. What they usually have in common is that they define standards for application of methods and requirements for the certification of practitioners, strictly putting an expert operator in control of the investigation, leading to an expert testimony in court. Most media forensic approaches today still lack maturity in this regard because the focus here currently lies mostly only on proposing AI detectors for specific forensic tasks, like image manipulation detection or DeepFake detection, neglecting most of the necessary modeling, benchmarking and standardization work required to make such approached mature enough for court room appearance.

This gap (i.e., the lack of required domain specific process modeling and benchmarking work) is addressed in this paper in part by the following contributions in this paper:

- An extension of existing modeling work on domain specific process models for media forensic investigations (here illustrated on the example of DeepFake detection), to include human-in-the-loop and human-in-control aspects as requested by changing requirements/legislation worldwide, esp. the upcoming EU Artificial Inteligence Act (AIA).
- An empirical estimation of the generalization power (or lack there-off) of pre-existing DeepFake detectors in intra and inter data set benchmarking, using different data selection strategies and classifiers.
- Initial tests on 2- vs. multi-class modeling of the decision problem, showing interesting results for the potential attribution / identification of the used DeepFake synthesis method.

The rest of the paper is structured as follows: First, a very brief overview over the current state of the art on domain specific process modeling for media forensics in Europe and Germany is given. The following section presents the modeling work in this paper, extending an existing Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling (DCEA) by including quality assurance aspects for a benchmarking framework for DeepFake detection models. Based on this modeling work, the core part of this paper presents empirical benchmarking activities on the example case of DeepFake detection, describing the setup and results for performance benchmarking for various DeepFake detection models compared in the same framework. The paper closes with conclusions and a summary of perspectives for potential future work.

Domain specific process modeling for media forensics in Europe and Germany

The most recent best practice document for media forensics in Europe is, at the time of writing this paper, the European Network of Forensic Science Institutes (ENFSI) Best Practice Manual (BPM) for Digital Image Authentication [8]. In its own words it "*aims to provide a framework for procedures, quality principles, training processes and approaches to the forensic examination*" and is intended "*to establish and maintain working practices in the field of forensic Image Authentication* (IA) *that will: deliver reliable results, maximize the quality of the information obtained and produce robust evidence. The use of consistent methodology and the production of more comparable results will facilitate interchange of data between laboratories*." It generalizes a workflow for an image authentications examination and provides a classification scheme for methods for digital image authentication but insists that it "*is not a standard operating procedure (SOP) and addresses the requirements of the judicial systems in general terms only*" [8].
The reason why the ENFSI BPM does not intend to be a standard operating procedure or a forensic process model as basis for standardization purposes is, that such processes are governed by national law and ENFSI has no directive authority in Europe. Here, national regulation would be required to define the precise legal context for any media forensic investigation and the usage of the corresponding results in court.

Regarding the German situation, which is relevant for the authors of this paper, the most relevant best practice document regarding IT forensics in general (incl. media forensics) is the BSI (German Federal Office for Information Security) guide on IT forensics [2] (German: "*Leitfaden IT-Forensik*"). It provides various means for modeling forensic processes, including the definition of a generic phase-driven investigation & reporting model, a basic data model and a classification of methods and tools. Like many other best practice documents in this field it covers basic investigation principles, process models, forensic data types, etc. but does not provide domain specific process models and guidelines for specific media forensic investigations such as DeepFake detection. Here, existing research, such as the latest extension to the BSI guidelines [2] described as the Data-Centric Examination Approach for Incident Response- and Forensics Process Modeling (DCEA) summarized in [14] and [25], is used as basis for extending the scope of these guidelines to achieve a higher degree of maturity for the state of the art in taylor-made models for media forensics (incl. DeepFake detection).

The core of DCEA has three main components: a model of the *phases* of a forensic process, a classification scheme for *forensic method classes* and *forensically relevant data types*.
The six DCEA *phases* are briefly summarized as: Strategic preparation (SP), Operational preparation (OP), Data gathering (DG), Data investigation (DI), Data analysis (DA) and Documentation (DO). While the first two (SP and OP) contain generic (SP) and case-specific (OP) preparation steps, the three phases DG, DI and DA represent the core of any forensic investigation. At this point the importance of the SP has to be pointed out, since it is the phase that also includes all standardization, benchmarking, certification and training activities considered in this paper. For details on the phase model the reader is referred, e.g. to [14] or [1].

The second core aspect of DCEA is the definition of *forensic method classes* as presented in [14]. The third aspect is the specification of *forensically relevant data types*. They can be summarized as: MFDT1 "digital input data" (the initial media data considered for the investigation), MFDT2 "processed media data" (results of transformations to media data), MFDT3 "contextual data" (case specific information (e.g. for fairness evaluation)), MFDT4 "parameter data" (contain settings and other parameter used for acquisition, investigation and analysis), MFDT5 "examination data" (including the traces, patterns, anomalies, etc that lead to an examination result), MFDT6 "model data" (describe trained model data (e.g. face detection and model classification data)), MFDT7 "log data" (data, which is relevant for the administration of the system (e.g. system logs)), and MFDT8 "chain of custody & report data" (describe data used to ensure integrity and authenticity (e.g. hashes and time stamps) as well as the accompanying documentation for the final report).

In general, each processing operation (or operator) in an DCEA process pipeline is considered here as an atomar processing black box component with an identifier and (usually) a description of the processing performed in this operation. Each component has four well defined connectors: *input*, *output*, *parameters* and *log data* (see figure 1). To pay respects to the particularities of this field and make the following modeling task easier, a fifth connector is defined within this paper for a specific type of operator which requires a knowledge representation or a model for its processing operation. In that case, this fifth connector is labeled *model*. A detailed description of the modeling of these operators is given in [25].

The focus of the proposed extensions of the DCEA lies in this paper on the integration of the human operator into the procedures. Human-in-control is an principle that has long been established in forensics as a strict requirement and is nowadays also receiving more and more attention in any field of application where artificial intelligence (AI) is used. Among other regulations, the EU Artificial Intelligence Act (AIA, [7]) emphasizes it as a necessity for any critical application scenario. This extension is shown in figure 1 where two human operators are added to the component: One (labeled 'HO') as operator in control of the functionalities of the component and another one (labeled 'Sys admin') in the loop on the infrastructure, analyzing the system logs (MFDT7) and reacting to potential technical events such as an hard disc failure, etc.

Example case: Quality assurance aspects for a benchmarking framework for DeepFake detection models

Depending on the actual position of the component in a forensic investigation pipeline, the human operator (HO) in control could be a someone defining in-house quality assurance strategies (e.g. human operator 'HO1 ' in figure 2), a media forensics expert performing explainable AI (xAI) tasks in the used feature space ('HO 2' in figure 2) or a data scientist at a standardiza-

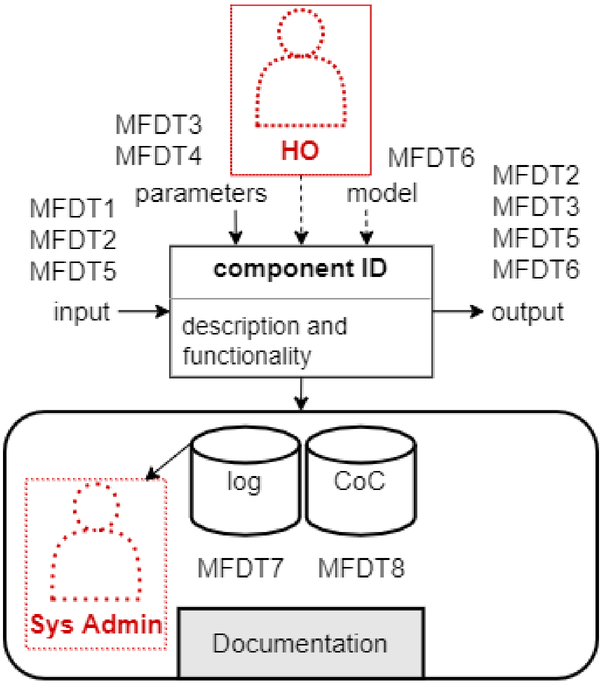


Figure 1. Template structure integrating the human operator(s) (HO) highlighted in red.

tion body like NIST running a benchmark and performing certification of the model trained ('HO 3' in figure 2). Obviously, all these different example HO would need different expertise and might have conflicting intentions.

The empirical evaluations performed in this paper focus on the interplay between 'HO 1', 'HO 2' and 'HO 3' in figure 2. Their interaction represents the cycle of decision model development (or training), its benchmarking and reasoning on the obtained results. For the performed evaluations in DeepFake detection, the following evaluation goals are defined:

- Estimation of generalization power (or lack there-off) in intra and inter data set evaluations, using different data selection strategies and classifiers
- Initial discussion on 2- vs. n-class classification (where n is the number of DeepFake synthesis methods plus one class for original, non-modified videos)
- Impact of data augmentation in training (model robustness)
- First considerations on video post-processing operations as potential counter-forensics

Evaluation setup

The evaluation setup is build according to the process model and the evaluation goals discussed in the previous section. Its general purpose is to provide an evaluation framework for DeepFake detection models, based on suitable DeepFake data sets. The video selection is done for each data set, where the selected number of videos corresponds to the minimal size of all data sets given. The extracted source data is augmented using different augmentation methods. All videos are processed in feature extractors introduced in [24] to classify DeepFakes based on eye (DF_{eye}), mouth (DF_{mouth}) and image foreground ($DF_{foreground}$) regions respectively. In addition, meta data is gathered to enable a human operator (here 'HO 1') to further curate the data. The extracted feature lists are then split into distinct training and test data for all model generation and benchmark strategies. This separation is further used to enable different evaluation scenarios, such as intra and inter data set evaluations.

Benchmarking data set selection

Previous experiments given in [16] have shown that early DeepFake video data sets, such as TIMIT-DF [23, 15], show visible flaws in the videos, making them unsuitable for a fair benchmarking of detectors. Therefore, a manual curation and evaluation of data sets to be used is performed. FaceForensics++ [21, 22] is another early data set, that includes various DeepFake synthesis methods, but also got a recent extension in HiFiFace [27]. Initially, DeeperForensics [11] was included into the data pool to be used in this paper as an augmented data set based on FaceForensics++, but was then replaced by in-house augmentation for comparability reasons. The DeepFake Detection data set (DFD) by Google and JigSaw [6] is available as a part of FaceForensics++, providing both additional real videos as well as the output of a DeepFake synthesis method. Celeb-DF [19] is large data set, using an Autoencoder for synthesis. Furthermore, FakeAVCeleb [13] was originally considered for usage in this paper, due to the fact that it also includes audio data and a labeling of ethical background and gender, but it was dropped due to a low resolution of 224x224. In table 1 a summary of selected data sets can be found.

Data augmentation

Based on the selected data sets an equal amount of 363[1] videos per subset of each data set are taken for evaluation. The selection occurs pseudo-random based on a seed (here, the randomly chosen value 7 is taken as seed). To further augment the data sets and simulate a less optimal training scenario, the selected videos undergo two different post-processing operations: One additional data set is generated by re-sampling the videos to 15 frames per second, a second data set is created by resizing them to a width of 480 pixels while keeping the aspect ratio to prevent distortion. This augmentation is done using the FFmpeg library [9]. In total, 7986 videos (363 + 7*3*363) are used in this benchmark.

For classification, a total of five different classifiers from WEKA [10] are selected to represent a variety of different algorithms. These are NaiveBayes [12], LibSVM [4], Simple Logistics [17, 26], JRIP [5] and J48 [20].

To ensure the distinct split of training and test data two different approaches are taken. The first one utilizes methods built into WEKA, which includes a 66% training 34% testing percentage split, as well as 3-, 5- and 10-fold stratified cross-validation. The second approach involves manual pre-processing and dividing of the samples in fixed splits. This allows for more precise grouping of the data and thus enables addressing of specific evaluation questions. For reproducibility, the splits occur pseudo-randomly by using a deterministic script with a seed (again the value 7 is

[1] The number of files in the smallest set used (here 'DFD-actors') defines the size of the subsets drawn from all other data sets used, to ensure equally sized representations in training and testing.

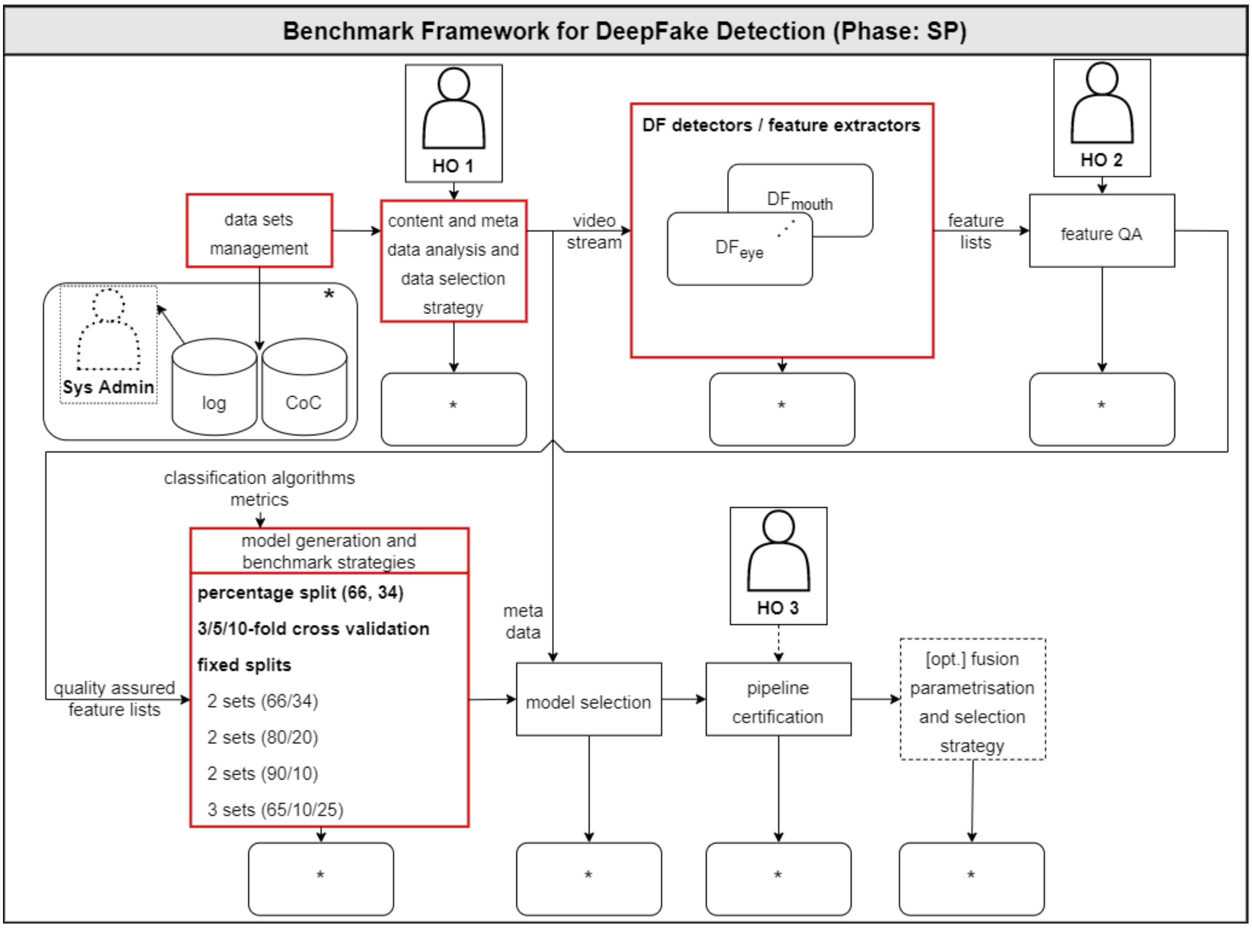

Benchmark Framework for DeepFake Detection (Phase: SP)
HO 1
HO 2
HO 3
data sets management
content and meta data analysis and data selection strategy
video stream
DF detectors / feature extractors
DF mouth
DF eye
feature lists
feature QA
Sys Admin
log
CoC
*
classification algorithms metrics
model generation and benchmark strategies
percentage split (66, 34)
3/5/10-fold cross validation
fixed splits
2 sets (66/34)
2 sets (80/20)
2 sets (90/10)
3 sets (65/10/25)
quality assured feature lists
meta data
model selection
pipeline certification
[opt.] fusion parametrisation and selection strategy

Figure 2. Illustration of the DeepFake detection pipeline created as a template in the forensic process model phase of Strategic Preparation (SP), with the inclusion of human operators (HO) implementing human-in-control as well as human-in-the-loop (for 'Sys Admin'). Contribution is highlighted in red.

data set	# individuals	# real video	# DeepFake video	subset	# selected videos
FaceForensics++ [21, 22]	?[2]	1 000	4 000	youtube-real	363
				Face2Face	363
				FaceShifter	363
				NeuralTexture	363
DFD [6]	28	363	3 068	DFD-actors (real)	363
HiFiFace [27][1]	?[2]	0	1 000	FaceSwap	363
Celeb-DF [19]	59	890	5 639	Celeb-real	363
				Celeb-synthesis	363

Overview of the data sets used in this paper for benchmarking of DeepFake detection models.

[1]**: Based on the youtube-real subset of FaceForensics++.**

[2]**: Numbers correspond, but unfortunately the exact number have not been disclosed by the original authors.**

taken). Using this script, disjointed training and testing splits of 66%/34%, 80%/20% and 90%/10% are generated automatically.

Evaluation results

As discussed previously, the evaluation is done in multiple individual experiments. In the first experiment the evaluation aims at different model generation and benchmark strategies, using the non-augmented data for evaluation. With the consideration of all three detectors DF_{eye}, DF_{mouth} and $DF_{foreground}$ the same tendencies of classification can be found, with some small exceptions. Figure 3 displays the results on the example of DF_{eye}. In general, it can be said, that there are almost no differences between 3-, 5- and 10-fold cross validation in this benchmark. In terms of predefined splits, an increase in detection performances can be found with increasing training data set size. This comes with an exception for the J48 classifier on the detectors DF_{eye}, where smaller training splits yield higher detection performance on the test set, indicating generalization problems (here in the handling of outliers in the test data) for this setup. Besides this mall glitch in the performance of J48, none of the tested classifiers is significantly better than the others. Each of the detection approaches had a different classifier scoring best, in all cases achieved on the 90/10 fixed split. LibSVM for DF_{eye} (Kappa=0.4991), J48 for DF_{mouth} (0.4113) and Simple Logistic for $DF_{foreground}$ (0.3620). The Kappa statistics of DF_{mouth} are in the range of 0.2544-0.4113, showing a significant drop in performance compared to previous results in [24]. This might suggest that anomalies in the mouth region are data set specific and do not occur to the same extent as in the previous experiment. Same can be said for $DF_{foreground}$ ranging Kappa values from 0.1810 to 0.3620, showing lower but less fluctuating performances.

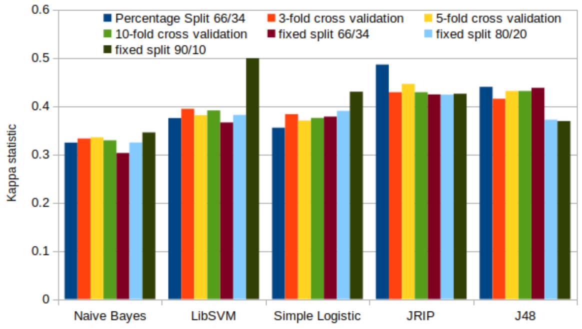


Figure 3. *Detection performances (Kappa values) for different model generation and benchmark strategies, on the example of* DF_{eye}

The second experiment addresses the usage of augmentation strategies in both training and testing. For this purpose, the data set is divided into native and augmented videos. Independently of the detector, augmentation usage solely for training or testing, results in a drop of detection performance. But it also has to be noted, that the integration of augmentation strategies in both training and testing did not impact the detectors negatively, and even increased the performance of DF_{mouth} and $DF_{foreground}$ (see the corresponding table).

The third experiment focuses on the impact of different DeepFake synthesis methods and considers every method as an individual class. Based on the considered data sets this results in a 6-class classification problem, which is then back projected to 2-class (‘original’ vs. ‘DeepFake’) for direct comparison. In terms of individual synthesis methods, it turned out that HiFiFace is clearly different from the others, especially for DF_{eye}. Here, none of the other types is classified as HiFiFace and also the HiFiFace subset is solely classified as ‘real‘ or ‘HiFiFace‘. This suggests that more recent DeepFakes show less flaws in creation, here on the case of eye region and blinking specifically. This distinction is not found for DF_{mouth} and $DF_{foreground}$. However, considering the results, the separation does not show an improvement in detection performance in any detector compared to a 2-class classification. Nevertheless, it allows for an attribution / identification of the used synthesis method and therefore for a better justification of the decision made by using this model.

detector	2-class	6-class
DF_{eye}	73.55% (0.4312)	72.73% (0.4348)
DF_{mouth}	69.90% (0.3347)	67.60% (0.3186)
$DF_{foreground}$	71.83% (0.3199)	62.19% (0.2228)

Comparison of 2- and 6-class DeepFake detection.

Summary, Conclusions and Future Work

Summarizing the empirical results presented in this paper, it is shown that the promising results previously shown in [24] are not reliable (i.e., not generalizing well) when properly benchmarked: The extension of the data considered (in different evaluation scenarios) shows challenges in generalization power, an important lesson learned regarding human-in-control and QA aspects, highlighting the relevance of benchmarking for data selection as well as feature and decision model quality assurance.

First tests on 2- vs. multi-class modeling of the decision problem show interesting initial results for the potential attribution / identification of the used DeepFake synthesis method.

Important future work would be to extend the introduced benchmarking framework to include additional datasets to cover an even wider range of DeepFake synthesis methods and also more different sets of ‘genuine’ (non-DeepFake) samples with different preprocessing histories. In this regard, the first data augmentation tests discussed here could be a suitable starting point for creating more robust detector models. Extensions along this line could e.g. use the DeeperForensics data set (with its augmentations) as an extension of FaceForensics++.

Besides the generalization issue, also the closely related question of training bias and fairness has to be considered in future work, potentially with evaluations using the FakeAVCeleb data set with its metadata annotations (incl. among other characteristics an indication on the ethical background of the person in the video).

From the perspective of potential courtroom fitness, an important future step would be to find a independent and trustworthy third party like NIST in the US or the BSI in Germany that could be motivated to perform independent benchmarking (and potentially also certification) of methods and trained models.

Acknowledgements

The work in this paper is funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number FKZ: 13N15736 (project “Fake-ID”).

Author Contributions: Initial idea & methodology: Jana Dittmann (JD) and Christian Kraetzer (CK); Conceptualization: Christian Kraetzer (CK), Dennis Siegel (DS), Stefan Seidlitz (StS) and JD; Re-modelling of the process model components: CK, DS, StS; Empirical evaluations: DS; Writing – original draft: CK; Writing – review & editing: DS, StS and JD.

training data set →	no augmentation (no aug)		with augmentation (w aug)		combination of both for train and test
↓ detector test data set →	no aug	w aug	no aug	w aug	
DF_{eye}	73.55% (0.4312)	57.72% (0.1155)	65.25% (0.1931)	70.42% (0.1919)	72.38% (0.3643)
DF_{mouth}	69.90% (0.3347)	69.86% (0.1089)	71.25% (0.3239)	70.09% (0.1558)	70.94% (0.2489)
$DF_{foreground}$	71.83% (0.3199)	70.58% (0.1109)	71.11% (0.2814)	70.11% (0.0611)	71.50% (0.2080)

Evaluation results for augmentation strategies. All values are determined using J48 under default parameters.

References

[1] Robert Altschaffel. *Computer forensics in cyber-physical systems : applying existing forensic knowledge and procedures from classical IT to automation and automotive*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[2] BSI. *Leitfaden IT-Forensik*. German Federal Office for Information Security, 2011.

[3] Christophe Champod and Joëlle Vuille. Scientific evidence in europe - admissibility, evaluation and equality of arms. *International Commentary on Evidence*, 9(1), 2011.

[4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] William W. Cohen. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

[6] Nick Dufour and Andrew Gully. Contributing Data to Deepfake Detection Research. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`, September, 24 2019. Accessed: 09/09/2021.

[7] European Commission. Proposal for a Regulation of the european parliament and of the council Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *COM(2021) 206 final*, April, 21 2021. [Online]. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX\%3A52021PC0206` [Last retrieved: 14.09.2021].

[8] European Network of Forensic Science Institutes. Best practice manual for digital image authentication. *ENFSI-BPM-DI-03*, October 2021. [Online]. Available at: `https://enfsi.eu/wp-content/uploads/2022/12/1.-BPM_Image-Authentication_ENFSI-BPM-DI-03-1.pdf` [Last retrieved: 12.01.2023].

[9] FFmpeg. FFmpeg, 2018. [Online]. Available at: `https://ffmpeg.org/` [Last retrieved: 12.01.2023].

[10] Mark A. Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor.*, 11(1):10–18, 2009.

[11] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2886–2895. Computer Vision Foundation / IEEE, 2020.

[12] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[13] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2021.

[14] Stefan Kiltz. *Data-Centric Examination Approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020.

[15] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR*, abs/1812.08685, 2018.

[16] Christian Kraetzer, Dennis Siegel, Stefan Seidlitz, and Jana Dittmann. Process-driven modelling of media forensic investigations-considerations on the example of deepfake detection. *Sensors*, 22(9), 2022.

[17] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. 95(1-2):161–205, 2005.

[18] Legal Information Institute. Rule 702. testimony by expert witnesses, Dec 2019.

[19] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3204–3213. IEEE, 2020.

[20] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[21] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[22] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1–11. IEEE, 2019.

[23] Conrad Sanderson and Brian Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. *LNCS*, 5558:199–208, 2009.

[24] Dennis Siegel, Christian Kraetzer, Stefan Seidlitz, and Jana Dittmann. Media forensics considerations on deepfake detection with hand-crafted features. *Journal of Imaging*, 7(7), 2021.

[25] Dennis Siegel, Christian Krätzer, Stefan Seidlitz, and Jana Dittmann. Forensic data model for artificial intelligence based media forensics-illustrated on the example of deepfake detection. *Electronic Imaging*, 34:1–6, 2022.

[26] Marc Sumner, Eibe Frank, and Mark Hall. Speeding up logistic model tree induction. In *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 675–683. Springer, 2005.

[27] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1136–1142. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

Author Biography

__Jana Dittmann__ is a Professor on multimedia and security at the University of Otto-von-Guericke University Magdeburg (OvGU). She is the leader of the Advanced Multimedia and Security Lab (AMSL) at OvGU, which is partner in national and international research projects and has a wide variety of well recognized publications in IT security. __Christian Kraetzer__ is a post-doc researcher and __Dennis Siegel__ as well as __Stefan Seidlitz__ are PhD students at AMSL.

15

[Siegel23a] Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data

[Siegel23a] Dennis Siegel, **Christian Kraetzer**, Jana Dittmann: *Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data*. Proc. The Seventeenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2023), Porto, Portugal, September, 2023, IARIA, ISBN: 978-1-68558-092-6, pp. 43-51, 2023.

The original publication is available as Open Access publication at ThinkMind:
https://www.thinkmind.org/index.php?view=article&articleid=securware_2023_1_80_30054

This paper received a **Best Paper Award**[42] form SECURWARE 2023: https://www.iaria.org/conferences2023/AwardsSECURWARE23.html

Shares/author contributions (as given in the original paper - see page 241 of this cumulative habilitation treatise): "*Initial idea & methodology: Jana Dittmann (JD), Christian Kraetzer (CK); Conceptualization: Dennis Siegel (DS); Modeling & application in the context of DeepFake: DS; Writing – original draft: DS; Writing – review & editing: CK, JD and DS.*
All authors have read and agreed to the published version of the manuscript."

[42] Based on the reviews of the original submission, the camera-ready version, and the presentation during the conference. The authors also received an invitation to submit an expanded article version to one of the IARIA Journals.

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

16

[Kraetzer24] Explainability and Interpretability for Media Forensic Methods: Illustrated on the Example of the Steganalysis Tool Stegdetect

[Kraetzer24] Christian Kraetzer, Mario Hildebrandt: *Explainability and Interpretability for Media Forensic Methods: Illustrated on the Example of the Steganalysis Tool Stegdetect*. In P. Radeva, A. Furnari, K. Bouatouch, and A. A. de Sousa (eds.), Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2024, Volume 4: VISAPP, Rome, Italy, February 27-29, 2024, pp. 585–592. SCITEPRESS, 2024.

The original publication is available at Scitepress:
https://www.scitepress.org/Papers/2024/124238/124238.pdf

Shares/author contributions (as given in the original paper - see page 252 of this cumulative habilitation treatise): "*Author Contributions: Initial idea & methodology: Christian Kraetzer (CK); Conceptualization: CK, Mario Hildebrandt (MH); Discussion on Explainability and Interpretability in ML/AI for Law Enforcement and Forensics: CK, Empirical evaluations - design: CK, MH; Empirical evaluations - realisation: MH; Writing – original draft: CK; Writing – review & editing: MH.*
All authors have read and agreed to the published version of the manuscript."

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

This page was intentionally left blank in this version of the thesis for copyright reasons!

Appendix A: Complete List of Publications that have been Published during the Time of the Habilitation Project

This appendix provides a list of all scientific publications of the author published during the time of this habilitation project (2013 to 2024). This list is separated into the categories of (editorship of) books (1), book chapters (2), journal papers (7) and conference & workshop papers (33).
A complete list of all publications by the author, also including material that has been published prior to the author's PhD thesis is available at: https://omen.cs.uni-magdeburg.de/itiamsl/deutsch/mitarbeiter/christian-kraetzer/publications.html

Books

- Christian Kraetzer, Yun-Qing Shi, Jana Dittmann and Hyoung Joong Kim (Eds.): Proceedings of the 16th International Workshop on Digital Forensics and Watermarking (IWDW 2017), Magdeburg, Germany, August 23-25, 2017, Springer International Publishing, LNCS 10431, ISBN: 978-3-319-64184-3, DOI: 10.1007/978-3-319-64185-0, 2017.

Book chapters

- Jana Dittmann, Christian Kraetzer: Privacy concepts in biometrics: lessons learned from forensics. In Claus Vielhauer (Ed.): User-Centric Privacy and Security in Biometrics. Institution of Engineering and Technology (IET), ISBN: 978-1-78561-207-7, Book DOI: 10.1049/PBSE004E, 2017.
- Christian Kraetzer and Jana Dittmann: Microphone Forensics. In Handbook of Digital Forensics of Multimedia Data and Devices. Anthony T. S. Ho, Shujun Li (Eds.). Wiley-IEEE Press (John Wiley & Sons, Chichester, United Kingdom), ISBN 978-1-118-64050-0, September 2015.

Journal papers

- Milan Tahraoui, Christian Krätzer, Jana Dittmann, Hartmut Aden: Defending Informational Sovereignty by Detecting Deepfakes? Opportunities and Risks of an AI-Based Detector for Deepfakes-Based Disinformation and Illegal Activities. Weizenbaum Journal of the Digital Society, 3(2), (2023).
- Christian Krätzer, Dennis Siegel, Stefan Seidlitz, Jana Dittmann: Process-driven modelling of media forensic investigations-considerations on the example of DeepFake detection. J. Sensors - Basel: MDPI, Bd. 22 (2022), 9, 2022.
- Kevin Lamshöft, Jonas Hielscher, Christian Krätzer, Jana Dittmann: The threat of covert channels in network time synchronisation protocols. Journal of cyber security and mobility, Gistrup: River Publishers, Bd. 11 (2022), 2, pp. 165-204, 2022.
- Christian Kraetzer, Andrey Makrushin, Jana Dittmann, Mario Hildebrandt: Potential advantages and limitations of using information fusion in media forensics - A discussion on the example of detecting face morphing attacks. EURASIP Journal on Information Security, 2021, Bd. 2021, Heft 1, Springer, 2021.
- D. Siegel, C. Kraetzer, S. Seidlitz und J. Dittmann. Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features. Journal of Imaging, 7(7), 2021.
- Tom Neubert, Andrey Makrushin, Mario Hildebrandt, Christian Kraetzer, and Jana Dittmann: Extended StirTrace Benchmarking of Biometric and Forensic Qualities of Morphed Face Images. Journal IET Biometrics. 10.1049/iet-bmt.2017.0147, 2018.
- Veit Köppen, Christian Krätzer, Jana Dittmann, Gunter Saake, and Claus Vielhauer: Impacts on Database Performance in a Privacy-Preserving Biometric Authentication Scenario. In International Journal on Advances in Security, 8(1+2), IARIA, pages 99-108, 2015.

Conference and workshop papers

- Christian Kraetzer, Mario Hildebrandt: Explainability and Interpretability for Media Forensic Methods: Illustrated on the Example of the Steganalysis Tool Stegdetect. Accepted for the 19th International Conference on Computer Vision Theory and Applications (VISAPP2024), Rome, Italy, February 27-29, 2024.
- Bernhard Birnbaum, Christian Krätzer, Jana Dittmann: Stego-Malware Attribution: Simple Signature and Content-based Features Derived and Validated from Classical Image Steganalysis on Five Exemplary Chosen Algorithms. Proc. 17. International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2023), September 25, 2023 to September 29, 2023 - Porto, Portugal, 2023.
- Dennis Siegel, Christian Kraetzer, Jana Dittmann: Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data. Proc. 17. International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2023), September 25, 2023 to September 29, 2023 - Porto, Portugal, 2023.
- Dennis Siegel, Stefan Seidlitz, Christian Krätzer, Jana Dittmann: Pros and cons of comparing and combining hand-crafted and neural network based DeepFake detection based on eye blinking behavior. Proc. Electronic Imaging. Springfield, VA : Society for Imaging Sciences and Technology, Bd. 35 (2023), S. 378-1-378-6, 2023.
- Christian Krätzer, Dennis Siegel, Stefan Seidlitz, Jana Dittmann: Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models. Proc. Electronic Imaging. Springfield, VA : Society for Imaging Sciences and Technology, Bd. 35 (2023), S. 379-1-379-6, 2023.
- Milan Tahraoui, Christian Krätzer, Jana Dittmann: Defending Informational Sovereignty by Detecting Deepfakes: Risks and Opportunities of an AI-Based Detector for Deepfake-Based Disinformation and Illegal Activities. Eds.: Bianca Herlo, Daniel Irrgang. Proceedings of the 4th Weizenbaum Conference 2022: Practicing Sovereignty - Interventions for Open Digital Futures. Berlin. ISSN 2510-7666, pp. 142-161, 2022.
- Dennis Siegel, Christian Krätzer, Stefan Seidlitz, Jana Dittmann: Forensic data model for artificial intelligence based media forensics - Illustrated on the example of DeepFake detection. Proc. Electronic Imaging, Springfield, VA: Society for Imaging Sciences and Technology, Bd. 34 (2022), 4, 2022.
- Anna Louban, Milan Tahraoui, Hartmut Aden, Jan Fährmann, Christian Krätzer, Jana Dittmann: Das Phänomen Deepfakes - Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung. In Künstliche Intelligenz, Demokratie und Privatheit/ Auswirkungen der Künstlichen Intelligenz auf Demokratie und Privatheit - Baden-Baden: Nomos Verlagsgesellschaft; Friedewald, Michael *1965-* . - 2022, S. 265-288, 2022.
- Tom Neubert, Claus Vielhauer, Christian Kraetzer: Artificial Steganographic Network Data Generation Concept and Evaluation of Detection Approaches to secure Industrial Control Systems against Steganographic Attacks. In Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES 21). Association for Computing Machinery, New York, NY, USA, Article 70, 1–9, 2021.
- Christian Kraetzer, Andrey Makrushin, Jana Dittmann, Mario Hildebrandt: Potential advantages and limitations of using information fusion in media forensics - A discussion on the example of detecting face morphing attacks. EURASIP Journal on Information Security, 2021, Bd. 2021, Heft 1, Springer, 2021.
- S. Ezennaya-Gomez, S. Kiltz, C. Kraetzer und J. Dittmann: A Semi-Automated HTTP Traffic Analysis for Online Payments for Empowering Security, Forensics and Privacy Analysis. In: The 16th International Conference on Availability, Reliability and Security, ARES 2021, New York, NY, USA, 2021. Association for Computing Machinery.

- J. Hielscher, K. Lamshöft, C. Krätzer und J. Dittmann: A Systematic Analysis of Covert Channels in the Network Time Protocol. In: The 16th International Conference on Availability, Reliability and Security, ARES 2021, New York, NY, USA, 2021. Association for Computing Machinery.
- M. Hildebrandt, A. Shakir, A. Ziemke, M. Abdelrazek, H. Stuetzer, D. Blut, K. Lamshoeft, S. Ezennaya-Gomez, C. Kraetzer und J. Dittmann: AiroIdent User identification based on analyzing WPA2 encrypted traffic containing search engine interactions. Electronic Imaging, 2021(4):344–1–344–8, 2021.
- K. Lamshöft, T. Neubert, C. Kraetzer, C. Vielhauer und J. Dittmann: Information Hiding in Cyber Physical Systems: Challenges for Embedding, Retrieval and Detection Using Sensor Data of the SWAT Dataset. In: Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, IHMMSec '21, S. 113–124, New York, NY, USA, 2021. Association for Computing Machinery.
- A. Makrushin, C. Kauba, S. Kirchgasser, S. Seidlitz, C. Kraetzer, A. Uhl und J. Dittmann: General Requirements on Synthetic Fingerprint Images for Biometric Authentication and Forensic Investigations. In: Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, IHMMSec '21, S. 93–104, New York, NY, USA, 2021. Association for Computing Machinery.
- S. Seidlitz, K. Jürgens., A. Makrushin, C. Kraetzer und J. Dittmann: Generation of Privacy-friendly Datasets of Latent Fingerprint Images using Generative Adversarial Networks. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,, S. 345–352. INSTICC, SciTePress, 2021.
- S. Wendzel, L. Caviglione, W. Mazurczyk, A. Mileva, J. Dittmann, C. Kraetzer, K. Lamshöft, C. Vielhauer, L. Hartmann, J. Keller und T. Neubert: A Revised Taxonomy of Steganography Embedding Patterns. In: The 16th International Conference on Availability, Reliability and Security, ARES 2021, New York, NY, USA, 2021. Association for Computing Machinery.
- R. Altschaffel, J. Hielscher, C. Kraetzer, K. Lamshöft und J. Dittmann: Forensic behavior analysis in video conferencing based on the metadata of encrypted audio and video streams - considerations and possibilities. Proc. SECURWARE 2020: the Fourteenth International Conference on Emerging Security Information, Systems and Technologies, November 21-25, S. 82-89, 2020.
- A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann und P. Eisert: Dempster-shafer Theory for Fusing Face Morphing Detectors. 27th European Signal Processing Conference (EUSIPCO) - [Piscataway, NJ]: IEEE, S. 1-5, 2019.
- A. Makrushin, C. Kraetzer, G. Mittag, H. Birkholz, U. Rabeler, A. Wolf, C. Seibold, A. Hilsmann, P. Eisert, L. Wandzik, R. Vicente Garcia und J. Dittmann: Distributed and GDPR/IPR Compliant Benchmarking of Facial Morphing Attack Detection Services. Proceedings of the International Conference on Biometrics for Borders 2019: Morphing and Morphing Attack Detection Methods, 2019, Warsaw, Poland, 2019.
- T. Neubert, C. Kraetzer und J. Dittmann: A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on eMRTD. Proceedings of the ACM Workshop on Information Hiding and Multimedia Security - New York, NY: The Association for Computing Machinery, S. 95-100, 2019.
- Andrey Makrushin, Christian Kraetzer, Tom Neubert, Jana Dittmann: Generalized Benford's Law for Blind Detection of Morphed Face Images. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '18). ACM, New York, NY, USA, 49-54.
- Tom Neubert, Christian Kraetzer, Jana Dittmann: Reducing the False Alarm Rate for Face Morph Detection by a Morph Pipeline Footprint Detector. Proc. 26th European Signal Processing Conference (EUSIPCO 2018), Rome, Italy, September 3rd - 7th, 2018.
- Christian Kraetzer, Jana Dittmann: Steganography by synthesis: Can commonplace image manipulations like face morphing create plausible steganographic channels?. In Proceedings of the

13th International Conference on Availability, Reliability and Security (ARES 2018). ACM, New York, NY, USA, Article 11, 8 pages. DOI: https://doi.org/10.1145/3230833.3233263, 2018.

- Andrey Makrushin, Christian Kraetzer, Tom Neubert and Jana Dittmann: Generalized Benford's Law for Blind Detection of Morphed Face Images. In IH&MMSec '18: 6th ACM Workshop on Information Hiding and Multimedia Security, June 20–22, 2018, Innsbruck, Austria. ACM, Innsbruck, Austria, https://doi.org/10.1145/3206004.3206018
- Christian Kraetzer, Andrey Makrushin, Tom Neubert, Mario Hildebrandt, Jana Dittmann: Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '17). ACM, New York, NY, USA, 21-32.
DOI: https://doi.org/10.1145/3082031.3083244, 2017.
- Christian Kraetzer, Robert Altschaffel, Jana Dittmann: Tendenzen zum Profiling von verschlüsselten Netzwerkverkehren - Möglichkeiten und Grenzen. In: 6th International Symposium "New Technologies": Stuttgart, Germany, 05./06.10.2016. - Herrausgeber: Bundeskriminalamt, Wiesbaden, 2016.
- Ronny Merkel, Christian Kraetzer, Mario Hildebrandt, Stefan Kiltz, Sven Kuhlmann and Jana Dittmann: A Semantic Framework for a better Understanding, Investigation and Prevention of Organized Financial Crime. Proc. GI Sicherheit 2016.
- Christian Kraetzer and Jana Dittmann: Considerations on the Benchmarking of Media Forensics. In: Proc. 23rd European Signal Processing Conf. (EUSIPCO), Nice, France, IEEE, ISBN: 978-0-9928626-3-3, 2015.
- Maik Schott, Claus Vielhauer, Christian Krätzer: Using Different Encryption Schemes for Secure Deletion While Supporting Queries. In Datenbanksysteme für Business, Technologie und Web (BTW 2015) - Workshopband. Second Workshop on Databases in Biometrics, Forensics and Security Applications (DBforBFS), March 2nd, 2015 - Hamburg, Germany, Lecture Notes in Informatics (LNI) - Proceedings Series of the Gesellschaft für Informatik (GI), Volume P-242, ISBN 978-3-88579-636-7, ISSN 1617-5468, pages 37-46, 2015.
- Jana Dittmann, Veit Köppen, Christian Krätzer, Martin Leuckert, Gunter Saake, Claus Vielhauer: Performance Impacts in Database Privacy-Preserving Biometric Authentication. In Proc. The Eighth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2014). November 16 - 20, 2014 - Lisbon, Portugal, 2014.
- Christian Krätzer, Mario Hildebrandt, Andreas Dobbert, Jana Dittmann: Digitalisierte Forensik: Sensorbildfusion und Benchmarking. Proc. D-A-CH Security 2014. Graz, Austria, 16.-17. September 2014.
- Robert Altschaffel, Christian Kraetzer, Jana Dittmann, Stefan Kiltz: A hierarchical model for the description of internet-based communication. In: Proc. of 8th International Conference on IT Security Incident Management and IT Forensics, IMF 2014., IEEE, Piscataway, NJ, IEEE, pp. 85-94, Conference: Münster, Germany, 2014.05.12-14, 2014.

Bibliography

[Altschaffel20] R. Altschaffel. *Computer forensics in cyber-physical systems : applying existing forensic knowledge and procedures from classical IT to automation and automotive*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020. URL http://dx.doi.org/10.25673/35364. 26, 43, 73

[Apostel60] L. Apostel. *Towards the formal study of models in the non-formal sciences*. Synthese, vol. 12:pp. 125–161, 1960. URL https://api.semanticscholar.org/CorpusID:46978054. 14

[Ashcroft04] J. Ashcroft, D. J. Daniels, and S. V. Hart. *Forensic Examination of Digital Evidence: A Guide for Law Enforcement*. U.S. Department of Justice - National Institute of Justice, 2004. URL https://www.ncjrs.gov/pdffiles1/nij/199408.pdf. 23, 25

[Bas11] P. Bas, T. Filler, and T. Pevný. *"Break Our Steganographic System": The Ins and Outs of Organizing BOSS*. In *Proceedings of the 13th International Conference on Information Hiding*, IH'11, p. 59–70. Springer-Verlag, Berlin, Heidelberg, 2011. ISBN 9783642241772. 65

[Berghoff21] C. Berghoff, J. Böddinghaus, V. Danos, G. Davelaar, T. Doms, H. Ehrich, A. Forrai, R. Grosu, R. Hamon, H. Junklewitz, M. Neu, S. Romanski, W. Samek, D. Schlesinger, J.-E. Stavesand, S. Steinbach, A. von Twickel, R. Walter, J. Weissenböck, M. Wenzel, and T. Wiegand. *Whitepaper - Towards Auditable AI Systems - From Principles to Practice*. Whitepaper, Federal Office for Information Security Germany (BSI), Bonn, Germany, 2021. 39

[Birnbaum23] B. Birnbaum, C. Kraetzer, and J. Dittmann. *Stego-Malware Attribution: Simple Signature and Content-based Features Derived and Validated from Classical Image Steganalysis on Five Exemplary Chosen Algorithms*. In *The Seventeenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2023, Porto, Portugal, 25-29 September 2023)*, pp. 33–42. IARIA, 2023. ISBN 978-1-68558-092-6. 64

[Böhme09] R. Böhme, F. C. Freiling, T. Gloe, and M. Kirchner. *Multimedia Forensics Is Not Computer Forensics*. In Z. J. M. H. Geradts, K. Y. Franke, and C. J. Veenman (eds.), *Computational Forensics*, pp. 90–103. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-03521-0. 1

[BSI11] BSI. *Leitfaden IT-Forensik*. German Federal Office for Information Security, 2011. URL https://www.bsi.bund.de/DE/Themen/Oeffentliche-Verwaltung/Sicherheitspruefungen/IT-Forensik/forensik_node.html. 4, 13, 25, 36, 43, 51, 52, 55, 59, 61, 69, 73

[BSI21] BSI. *Sicherer, robuster und nachvollziehbarer Einsatz von KI - Probleme, Maßnahmen und Handlungsbedarfe*. Bundesamt für Sicherheit in der Informationstechnik (BSI), Bonn, Germany, 2021. URL https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=6. 39

[BSI23a] BSI. *BSI Grundschutz webpage*, 2023. URL https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/it-grundschutz_node.html. German Federal Office for Information Security (BSI). 24

[BSI23b] BSI. *IT-Grunschutz-Baustein DER.2.2 Vorsorge für die IT-Forensik*. German Federal Office for Information Security (BSI), 2023. URL https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/IT-GS-

Kompendium_Einzel_PDFs_2023/05_DER_Detektion_und_Reaktion/DER_2_2_Vorsorge_fuer_die_IT_Forensik_2023.pdf. 13

[Champod11] C. Champod and J. Vuille. *Scientific Evidence in Europe – Admissibility, Evaluation and Equality of Arms*. International Commentary on Evidence, vol. 9, 2011. URL https://api.semanticscholar.org/CorpusID:147528378. 20, 23, 40, 50

[Cozzolino22] D. Cozzolino and L. Verdoliva. *Multimedia Forensics Before the Deep Learning Era*, pp. 45–67. Springer International Publishing, Cham, 2022. ISBN 978-3-030-87664-7. doi:10.1007/978-3-030-87664-7_3. URL https://doi.org/10.1007/978-3-030-87664-7_3. 14

[DOJ11] DOJ. *A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case*. Technical Report, U.S. Department of Justice, Office of the Inspector General, Washington, DC, USA, 2011. URL https://www.oversight.gov/sites/default/files/oig-reports/s1105.pdf. 69

[ENFSI09] ENFSI. *Best Practice Guidelines for ENF Analysis in Forensic Authentication of Digital Evidence*. Technical Report BPM-ENF-001, European Network of Forensic Science Institutes (ENFSI), Wiesbaden, Germany, 2009. 27, 28, 29, 30, 33

[ENFSI15] ENFSI. *Best Practice Manual for the Forensic Examination of Digital Technology*. Technical Report BPM-FIT-01-2015, European Network of Forensic Science Institutes (ENFSI), Wiesbaden, Germany, 2015. 2, 15, 16, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 43, 58, 59, 60, 69, 70, 74

[ENFSI21] ENFSI. *Best Practice Manual for Digital Image Authentication*. Technical Report BPM-DI-03-2021, European Network of Forensic Science Institutes (ENFSI), Wiesbaden, Germany, 2021. 11, 15, 16, 20, 27, 28, 29, 30, 33, 35, 36, 37, 38, 39, 43, 45, 58, 59, 61, 69, 70

[ENFSI22a] ENFSI. *Best Practice Manual for Digital Audio Authenticity Analysis*. Technical Report FSA-BPM-002, European Network of Forensic Science Institutes (ENFSI), Wiesbaden, Germany, 2022. 14, 28, 29, 39, 44, 45, 47, 51

[ENFSI22b] ENFSI. *Framework for Template for Field Specific Best Practice Manual (BPM)*. Guidance QCC-FWK-003, European Network of Forensic Science Institutes (ENFSI), Wiesbaden, Germany, 2022. 28

[Ferrara14] M. Ferrara, A. Franco, and D. Maltoni. *The magic passport*. In *IEEE International Joint Conference on Biometrics*, pp. 1–7. 2014. doi:10.1109/BTAS.2014.6996240. 3

[FHNW21] FHNW. *Welche Vorgehensmodelle für die IT-Forensik gibt es und welche Phasen beinhalten die Vorgehensmodelle?*, 2021. URL https://www.fhnw.ch/plattformen/iwi/2021/11/18/welche-vorgehensmodelle-fuer-die-it-forensik-gibt-es-und-welche-phasen-beinhalten-die-vorgehensmodelle/. Fachhochschule Nordwestschweiz. 26

[Flaglien17] A. Flaglien, I. M. Sunde, A. Dilijonaite, J. Hamm, J. P. Sandvik, P. Bjelland, K. Franke, and S. Axelsson. *Digital Forensics*. John Wiley & Sons, Ltd., 2017. ISBN 9781119262381. 26

[Frank16] E. Frank, M. A. Hall, and I. H. Witten. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, MA, 4th edn., 2016. 65

[Fridrich09] J. Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, New York, NY, USA, 1st edn., 2009. ISBN 0521190193, 9780521190190. 12, 63, 64

[Goljan06] M. Goljan, J. Fridrich, and T. Holotyak. *New blind steganalysis and its implications*. In E. J. D. III and P. W. Wong (eds.), *Security, Steganography, and Watermarking of Multimedia Contents VIII*, vol. 6072, p. 607201. International Society for Optics and Photonics, SPIE, 2006. doi:10.1117/12.643254. URL https://doi.org/10.1117/12.643254. 65

[Goljan09] M. Goljan, J. J. Fridrich, and T. Filler. *Large scale test of sensor fingerprint camera identification*. In E. J. Delp, J. Dittmann, N. D. Memon, and P. W. Wong (eds.), *Media Forensics and Security I, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 19, 2009, Proceedings*, vol. 7254 of *SPIE Proceedings*, p. 72540. SPIE, 2009. doi:http://dx.doi.org/10.1117/12.805701. 21

[Grancay17] M. Grancay, J. Vveinhardt, and E. Sumilo. *Publish or Perish: How Central and Eastern European Economists Have Dealt with the Ever-Increasing Academic Publishing Requirements 2000—2015*. Scientometrics, vol. 111(3):p. 1813–1837, 2017. ISSN 0138-9130. doi:10.1007/s11192-017-2332-z. URL https://doi.org/10.1007/s11192-017-2332-z. 19

[Hildebrandt20] M. Hildebrandt. *On digitized forensics: novel acquisition and analysis techniques for latent fingerprints based on signal porcessing and pattern recognition*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020. URL http://dx.doi.org/10.25673/34885. 26

[Ho15] A. T. S. Ho. *Handbook of digital forensics of multimedia data and devices / edited by Anthony T.S. Ho and Shujun Li, Department of Computing and Surrey Centre for Cyber Security (SCCS), University of Surrey, UK*. Wiley/IEEE Press, Hoboken, 2015. ISBN 9781118705797. 14, 20

[Hoppe14] T. Hoppe. *Prävention, Detektion und Reaktion gegen drei Ausprägungsformen automotiver Malware - eine methodische Analyse im Spektrum von Manipulationen und Schutzkonzepten*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2014. URL http://dx.doi.org/10.25673/4169. 26

[INTERPOL23] INTERPOL. *Toolkit for Responsible AI Innovation in Law Enforcement: Principles for Responsible AI Innovation*. Guidelines, United Nations Interregional Crime and Justice Research Institute (UNICRI) and International Criminal Police Organization (INTERPOL), Brussels, 2023. 40, 41, 65, 66, 73, 74

[ISO94] ISO. *Information technology – Open Systems Interconnection – Basic Reference Model: The Basic Model*. Standard ISO/IEC 7498-1:1994, International Organization for Standardization, Geneva, CH, 1994. URL https://www.iso.org/standard/20269.html. 14

[Kent06] K. Kent, S. Chevalier, T. Grance, and H. Dang. *SP 800-86. Guide to Integrating Forensic Techniques into Incident Response*. Tech. rep., Gaithersburg, MD, USA, 2006. 23

[Ker13] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevný. *Moving Steganography and Steganalysis from the Laboratory into the Real World*. IH&MMSec '13, p. 45–58. Association for Computing Machinery, New York, NY, USA, 2013. ISBN 9781450320818. doi:10.1145/2482513.2482965. URL https://doi.org/10.1145/2482513.2482965. 19

[Kiltz15] S. Kiltz, J. Dittmann, and C. Vielhauer. *Supporting Forensic Design - A Course Profile to Teach Forensics*. 2015 Ninth International Conference on IT Security Incident Management and IT Forensics (IMF2015), pp. 85–95, 2015. xv, 25, 26, 52, 54, 59

[Kiltz20] S. Kiltz. *Data-Centric Examination Approach (DCEA) for a qualitative determination of error, loss and uncertainty in digital and digitised forensics*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2020. URL http://dx.doi.org/10.25673/34647. xv, 4, 9, 25, 26, 51, 52, 54, 55, 59, 69, 73

[King09] D. E. King. *Dlib-ml: A Machine Learning Toolkit*. J. Mach. Learn. Res., vol. 10:pp. 1755–1758, 2009. URL https://dl.acm.org/citation.cfm?id=1755843. 55

[Kraetzer12] C. Kraetzer, K. Qian, and J. Dittmann. *Extending a context model for microphone forensics*. In N. D. Memon, A. M. Alattar, and E. J. D. III (eds.), *Media Watermarking, Security, and Forensics 2012*, vol. 8303, p. 83030S. International Society for Optics and Photonics, SPIE, 2012. doi:10.1117/12.906569. URL https://doi.org/10.1117/12.906569. 44

[Kraetzer15a] C. Kraetzer and J. Dittmann. *Considerations on the benchmarking of media forensics*. In *23rd European Signal Processing Conference, EUSIPCO 2015, Nice, France, August 31 - September 4, 2015*, pp. 61–65. IEEE, 2015. doi:10.1109/EUSIPCO.2015.7362345. URL https://doi.org/10.1109/EUSIPCO.2015.7362345. x, 5, 6, 21, 22, 75, 78, 80

[Kraetzer15b] C. Kraetzer and J. Dittmann. *Microphone Forensics*, chap. 11, pp. 411–441. John Wiley & Sons, Ltd, 2015. ISBN 9781118705773. doi:https://doi.org/10.1002/9781118705773.ch11. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118705773.ch11. 44

[Kraetzer17] C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, and J. Dittmann. *Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing*. In M. C. Stamm, M. Kirchner, and S. Voloshynovskiy (eds.), *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec 2017, Philadelphia, PA, USA, June 20-22, 2017*, pp. 21–32. ACM, 2017. doi:10.1145/3082031.3083244. URL https://doi.org/10.1145/3082031.3083244. x, 6, 7, 44, 45, 46, 47, 67, 83, 86, 88, 90, 92, 94, 96

[Kraetzer21] C. Kraetzer, A. Makrushin, J. Dittmann, and M. Hildebrandt. *Potential advantages and limitations of using information fusion in media forensics - a discussion on the example of detecting face morphing attacks*. EURASIP J. Inf. Secur., vol. 2021(1):p. 9, 2021. doi:10.1186/s13635-021-00123-4. URL https://doi.org/10.1186/s13635-021-00123-4. x, 8, 40, 49, 50, 51, 57, 113, 116, 118, 120, 122, 124, 126, 128, 130, 132, 134, 136, 138

[Kraetzer22] C. Kraetzer, D. Siegel, S. Seidlitz, and J. Dittmann. *Process-Driven Modelling of Media Forensic Investigations-Considerations on the Example of DeepFake Detection*. Sensors, vol. 22(9), 2022. ISSN 1424-8220. doi:10.3390/s22093137. URL https://www.mdpi.com/1424-8220/22/9/3137. x, 4, 10, 11, 22, 26, 39, 55, 56, 57, 58, 59, 60, 61, 65, 181, 184, 186, 188, 190, 192, 194, 196, 198, 200, 202, 204, 206, 208, 210, 212, 214

[Kraetzer23] C. Kraetzer, D. Siegel, S. Seidlitz, and J. Dittmann. *Human-in-control and quality assurance aspects for a benchmarking framework for DeepFake detection models*. Electronic Imaging, vol. 35(4):pp. 379–1–379–1, 2023. doi:10.2352/EI.2023.35.4.MWSF-379. URL https://library.imaging.org/ei/articles/35/4/MWSF-379. x, 11, 57, 58, 59, 60, 61, 69, 223, 226, 228, 230

[Kraetzer24] C. Kraetzer and M. Hildebrandt. *Explainability and Interpretability for Media Forensic Methods: Illustrated on the Example of the Steganalysis Tool Stegdetect*. In P. Radeva, A. Furnari, K. Bouatouch, and A. A. de Sousa (eds.), *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory* *and Applications, VISIGRAPP 2024, Volume 4: VISAPP, Rome, Italy, February 27-29, 2024*, pp. 585–592. SCITEPRESS, 2024. doi:10.5220/0012423800003660. URL https://doi.org/10.5220/0012423800003660. x, 12, 63, 64, 65, 243, 246, 248, 250, 252

[Krätzer13] C. Krätzer. *Statistical pattern recognition for audio-forensics - empirical investigations on the application scenarios audio steganalysis and microphone forensics*. Ph.D. thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik, 2013. URL http://dx.doi.org/10.25673/3967. 22, 26, 44

[Kuncheva04] L. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 1st edn., 2004. 49

[Kung86] C. H. Kung and A. Sölvberg. *Activity Modeling and Behavior Modeling*. In *Proc. of the IFIP WG 8.1 Working Conference on Information Systems Design Methodologies: Improving the Practice*, p. 145–171. North-Holland Publishing Co., NLD, 1986. ISBN 0444700145. 14

[Lukas06] J. Lukas, J. Fridrich, and M. Goljan. *Digital camera identification from sensor pattern noise*. IEEE Transactions on Information Forensics and Security, vol. 1(2):pp. 205–214, 2006. doi:10.1109/TIFS.2006.873602. 21

[Lyle22] J. Lyle, B. Guttman, J. Butler, K. Sauerwein, C. Reed, and C. Lloyd. *Digital Investigation Techniques: A NIST Scientific Foundation Review*. Tech. rep., 2022. doi:10.6028/NIST.IR.8354-draft. 23

[Makrushin19] A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, and P. Eisert. *Dempster-Shafer Theory for Fusing Face Morphing Detectors*. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. 2019. doi:10.23919/EUSIPCO.2019.8902533. 49

[Mylopoulos92] J. Mylopoulos. *Conceptual Modelling and Telos*. John Wiley & Sons, Inc., New York, NY, United States, 1992. ISBN 978-0-471-55462-2. 15

[Neubert18a] T. Neubert, C. Kraetzer, and J. Dittmann. *Reducing the False Alarm Rate for Face Morph Detection by a Morph Pipeline Footprint Detector*. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1002–1006. 2018. doi:10.23919/EUSIPCO.2018.8553067. x, 7, 48, 105, 108, 110

[Neubert18b] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann. *Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images*. IET Biometrics, vol. 7(4):pp. 325–332, 2018. doi:https://doi.org/10.1049/iet-bmt.2017.0147. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2017.0147. 57

[Neubert19] T. Neubert, C. Kraetzer, and J. Dittmann. *A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on EMRTD*. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec'19, p. 95–100. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450368216. doi:10.1145/3335203.3335721. URL https://doi.org/10.1145/3335203.3335721. x, 3, 7, 47, 97, 100, 102, 104

[Nissar10] A. Nissar and A. Mir. *Classification of steganalysis techniques: A study*. Digital Signal Processing, vol. 20(6):pp. 1758–1770, 2010. ISSN 1051-2004. doi:https://doi.org/10.1016/j.dsp.2010.02.003. 64

[Pollitt19] M. Pollitt, E. Casey, D.-O. Jaquet-Chiffelle, and P. Gladyshev. *A Framework for Harmonizing Forensic Science Practices and Digital/Multimedia Evidence - OSAC Technical Series 0002R1*. Tech. rep., 2019. doi:/10.29325/OSAC.TS.0002. 1, 2

[Provos02] N. Provos and P. Honeyman. *Detecting Steganographic Content on the Internet*. In *NDSS*. The Internet Society, 2002. ISBN 1-891562-14-2, 1-891562-13-4. 12, 63, 64

[Rathgeb22] C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch (eds.). *Handbook of Digital Face Manipulation and Detection From DeepFakes to Morphing Attacks*. Springer, 2022. 4, 14, 20, 48

[Siegel21] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann. *Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features*. Journal of Imaging, vol. 7(7), 2021. ISSN 2313-433X. doi:10.3390/jimaging7070108. URL https://www.mdpi.com/2313-433X/7/7/108. x, xv, 4, 8, 9, 10, 26, 51, 52, 53, 54, 56, 57, 61, 62, 141, 144, 146, 148, 150, 152, 154, 156, 158, 160, 162, 164, 166, 168, 170

[Siegel22] D. Siegel, C. Krätzer, S. Seidlitz, and J. Dittmann. *Forensic data model for artificial intelligence based media forensics - Illustrated on the example of DeepFake detection*. Electronic Imaging, vol. 34(4):pp. 324–1–324–1, 2022. doi:10.2352/EI.2022.34.4.MWSF-324. URL https://library.imaging.org/ei/articles/34/4/MWSF-324. x, xv, 9, 10, 11, 52, 53, 54, 55, 56, 59, 60, 62, 69, 173, 176, 178, 180

[Siegel23a] D. Siegel, C. Krätzer, and J. Dittmann. *Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data*. In *Proc. The Seventeenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE 2023)*, pp. 43–51. IARIA, 2023. ISBN 978-1-68558-092-6. x, 11, 58, 61, 231, 234, 236, 238, 240

[Siegel23b] D. Siegel, S. Seidlitz, C. Krätzer, and J. Dittmann. *Pros and cons of comparing and combining hand-crafted and neural network based DeepFake detection based on eye blinking behavior*. In *Media Watermarking, Security, and Forensics*. 2023. x, 10, 57, 58, 68, 69, 215, 218, 220, 222

[Tahraoui23] M. Tahraoui, C. Krätzer, J. Dittmann, and H. Aden. *Defending Informational Sovereignty by Detecting Deepfakes - Opportunities and Risks of an AI-Based Detector for Deepfakes-Based Disinformation and Illegal Activities*. Weizenbaum Journal of the Digital Society, vol. 3(2), 2023. doi:10.34669/WI.WJDS/3.2.3. URL https://ojs.weizenbaum-institut.de/index.php/wjds/article/view/95. 4

[Tolosana22] R. Tolosana, C. Rathgeb, R. Vera-Rodriguez, C. Busch, L. Verdoliva, S. Lyu, H. H. Nguyen, J. Yamagishi, I. Echizen, P. Rot, K. Grm, V. Štruc, A. Dantcheva, Z. Akhtar, S. Romero-Tapiador, J. Fierrez, A. Morales, J. Ortega-Garcia, E. Kindt, C. Jasserand, T. Kalvet, and M. Tiits. *Future Trends in Digital Face Manipulation and Detection*, pp. 463–482. Springer International Publishing, Cham, 2022. ISBN 978-3-030-87664-7. doi:10.1007/978-3-030-87664-7_21. URL https://doi.org/10.1007/978-3-030-87664-7_21. 48

[Vaughan20] J. Vaughan, N. Baker, and M. Underhill. *Digital Forensic Science Strategy*. Policy, National Police Chiefs' Council (NPCC), London, UK, 2020. 1, 2, 3, 17, 18, 54, 71, 73, 74

[Øe21] H. S. Øe. *OPINION OF ADVOCATE GENERAL - Case C-401/19 Republic of Poland vs European Parliament, Council of the European Union*, 2021. URL https://curia.europa.eu/juris/document/document.jsf;jsessionid=F18703A7435613FFFCE6AAAC77D5CDD8?text=&docid=244201&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=1687830. 4