Random Forest Algorithm in Unravelling Biomarkers of Breast Cancer Progression

Nadiia Kasianchuk[1,2], Dmytro Tsvyk[3], Eduard Siemens[4] and Halina Falfushynska[4,5]
[1]*Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego Str. 6, Poznań, Poland*
[2]*Faculty of Pharmacy, Bogomolets National Medical University, Taras Shevchenko Str. 13, Kyiv, Ukraine*
[3]*Institute of International Relations, Taras Shevchenko National University of Kyiv, Illyenka Str. 36, Kyiv, Ukraine*
[4]*Anhalt University of Applied Sciences, Bernburger Str. 55, Köthen, Germany*
[5]*Institute of Biological Sciences, University of Rostock, Albert Einstein Str. 3, Rostock, Germany*
nadkas2@st.amu.edu.pl, tsvykdima@gmail.com, halina.falfushynska@uni-rostock.de, eduard.siemens@hs-anhalt.de

Keywords: Breast Cancer, K-Means, Random Forest, Genes, Prognostic Factors, Classification, Biomarkers.

Abstract: Breast cancer is the leading cause of cancer death among women. As its development involves a multidimensional network of gene-environment interactions, advanced data analysis tools and bioinformatics are vital to uncover the nature of cancer. The initial database contained the expression values of 19737 genes in 1082 patients. Random Forest algorithm was used to distil the genes with the strongest influence on four substantial prognostic factors (survival period, tumour size, lymph node seizure, and metastasis). The obtained set consists of 230 potential biomarkers that facilitate the critical cancer-related pathways, such as p53, Wnt, VEGF, UPP, thereby influencing cell proliferation, tumouri- and angiogenesis. A considerable contrast in the expression was shown between the patients at different stages of cancer progression. The obtained set will simplify the diagnostics and prediction of tumour progression, enhance treatment outcomes and elaborate better strategies for curing breast cancer.

1 INTRODUCTION

Cancer is reported to be the second most common cause of death globally, accounting for an estimated 10 million deaths, or one in six in 2020[1] with breast cancer taking 2.26 million lives annually. As indicated by the World Health Organization (WHO), the current dramatic rise in the total number of diagnosed breast cancer cases is precipitated by alterations in human lifestyle and increased life expectancy. Though, the growing general awareness of the problem and enhanced screening technologies should not be overlooked as a contributing factor. With that, however, it is yet a long way to go, as people in low-middle-income countries and vulnerable strata in high-income states still face higher mortality rates. The reasons thereof include insufficient availability, affordability and accessibility of cancer care, as well as limited access to clinical innovations [1]. Regardless of the wealth, accumulated in a country, the implementation of cancer screening programmes is of the highest necessity, since it saves lives and also funds to be invested into treatment.

Cancer control is chiefly aimed at mitigating the prevalence and mortality rate of cancer [1, 2]. The disease screening programmes implement systematic evidence-based early diagnosis, the cornerstone whereof being a selection of suitable biomarkers of malignant neoplasm and a study of target pathways that aggravate cancer progress. Thus, the proper and in-time diagnosis and personalised therapy curb mortality and disability rates among persons affected by common cancers.

Such clinical biomarkers for breast cancer, as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) provide vital material for breast cancer prognosis and insights into the patient's probable reaction to treatment. That notwithstanding, other parameters are also deemed prospective to this end. *BRCA1*, *BRCA2*, *CXCR4*, caveolin, miRNA, and FOXP3, despite additional expenses, may come in handy to describe breast malignant neoplasm and tumour progression, to prevent metastasis and recurrences [3]. Yet, in step with the diversification of the raw data obtained, the analysis thereof becomes increasingly complex and demanding, however, a

[1]https://www.who.int/news-room/fact-sheets/detail/cancer

handsome reward for the work well performed constitutes the remarkably uplifted prognostic precision [4].

Hence, up-to-date risk assessment tools and novel approaches are rendered indispensable as modern science is broadening the range of applicable multi-gene tests and various biomarkers, including genomic, biochemical, and histopathological signatures [5]. Such a comprehensive data processing may be efficiently conducted only by means of transformative AI and machine learning solutions, so that the fruits thereof would provide valuable prognostic signals.

Whereas some pioneering attempts have been taken in the field [4,6,7,8], meaningful and comprehensive options for integrative data-processing are still limited. The key purpose of this work is the development of a proper network of the most promising breast cancer biomarkers. The paper marks our starting point in shedding light on mighty, but feasible data analysis tools for biomedical research. The integration of the IT data-processing solutions into life science and medical institutions will, with regards to breast cancer, contribute to detecting biomarker discrepancies and minimalising misreported biomarker status.

2 MATERIALS AND METHODS

2.1 Data Preparation

The database presented by The CGA through cBioPortal website[2] contains clinical pictures of breast cancer patients supplemented by the expression values of 19737 genes. The latter is described as z-scores relative to normal samples (log RNA Seq V2 RSEM), thus showing the number of standard deviations below or above the population mean.

$$z = \frac{x - \mu}{\sigma}$$

Data cleaning and pre-processing are standardly applied to the obtained information. All indicators with missing values were excluded from further analyses to avoid inaccurate interpretations. Furthermore, genes with statistically outlying z-scores (e.g. *OR2B2*) were also justifiably removed.

Python environment version 3.9.13 (Python Software Foundation) was employed for the purposes of the present analyses, with NumPy, pandas and Matplotlib libraries being used for data manipulations. Sklearn package was of much help for the algorithms of classification, regression and clustering.

2.2 K-Means Data Clustering

All samples were clustered according to their survival rate, so that the workflow of the Decision Tree could be optimised.

K-means represents an unsupervised clustering algorithm, praised in data analysis mainly due to its simplicity, efficacy and high-speed data processing. Firstly, k objects are randomly chosen as centroids across the data, whereto other samples are then being assigned. Such process is based on the smallest distance between data sample and existing clusters, measured in the Euclidean distance metric.

$$d = \sqrt{((x_2 - x_1)^2 - (y_2 - y_1)^2)}$$

The number of clusters was drawn on the Silhouette score, calculated in the following way.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The coefficient has a range of [-1;1]. S=1 indicates perfect clusterisation (the distance between any elements in the cluster is significantly lower than that between any in neighbouring clusters). S=0 shows that some samples are located right upon or near the line, delimiting two adjacent clusters. S=-1 depicts poor clustering choice. The highest Silhouette coefficient is obtained for k = 2 and k = 3, however values for k = 4 and k = 5 are still acceptable (Figure 1).

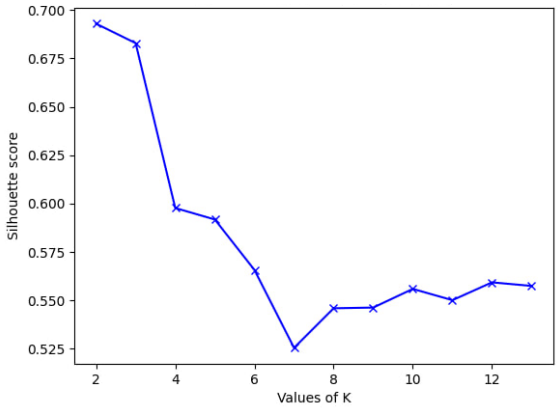

Figure 1: Silhouette analysis for Optimal k.

2.3 Random Forest Classification

For the purposes of classifying genes, Random Forest was trained on the prepared data, with gene expression values shown as z-scores relative to

[2]https://www.cbioportal.org

normal samples and survival expressed in months. The samples were clustered according to their survival. The further classification was performed based on the affiliation with a specific cluster. Categorical variables needed for the Random Forest (e.g. presence of metastasis), were transformed into numerical values (e.g. M0 – 0, M1 – 1).

Such hyperparameters as quantity of decision trees in the random forest and maximum depth of each tree were checked and chosen for the best performance and computing speed.

3 RESULTS AND DISCUSSION

3.1 Main Features of the Data Cohort

After cleaning and preparation, the analysed database contained the information about 1082 breast cancer patients, 1070 (98,89%) of whom were females and 12 were (1,11%) males (Table 1).

Table 1: Basic characteristics of the dataset.

Sex					
Female				98,89%	
Male				1,11%	
Diagnosis Age, years					
Mean	58,40 ± 13,19	Median	58	Mode	62
TNM (tumour–node–metastasis), %					
T1	25,51	N0	47,32	M0	82,62
T2	57,95	N1	32,81	cM0	0,55
T3	12,66	N2	11,00	M1	1,94
T4	3,60	N3	7,02	MX	14,88
TX	0,28	NX	1,85		
Tumour Type, %					
Infiltrating Ductal Carcinoma				71,53	
Infiltrating Lobular Carcinoma				18,58	
Medullary Carcinoma				0,55	
Metaplastic Carcinoma				0,74	
Mixed Histology (NOS)				2,68	
Mucinous Carcinoma				1,57	
Other				4,34	
Cancer subtype, %					
BRCA_LumA				46,12	
BRCA_LumB				18,21	
BRCA_Basal				15,80	
BRCA_Her2				7,21	
BRCA_Normal				3,33	
NA				9,33	

The mean diagnosis age was 58,40 ± 13,19 years. Median value for the same indicator equalled to 58 and the mode was 62 years. The most common type of tumour was Infiltrating Ductal Carcinoma (71,53%), followed by Infiltrating Lobular Carcinoma (18,58%). 2 samples with tumours, labelled as Breast Invasive Carcinoma and Infiltrating Carcinoma (NOS), were moved to group 'Other', for they did not meet the necessary criteria.

The American Joint Committee on Cancer introduced the tumour–node–metastasis (TNM) cancer staging system, which is used to indicate the severity of cancer. In such system, T-value characterises the size and the extent of a tumour, N shows the presence of cancer cells nearby or in the lymph nodes and M indicates whether cancer metastasised (spread to the distant parts of the body). The majority of patients from the database had T2 tumour stage (57,95%) and no metastases (M0 – 82,62%). 47,32% of the patients had no cancer cells near the lymph nodes (N0) and 32,81% had mild affection of lymph nodes (N1). Where TNM indicators mattered for the analyses, patients without TNM scores were not taken into consideration.

3.2 Clusterisation

The patients were clustered regarding their overall survival period. The number of clusters was selected due to the Silhouette coefficient. As $k = 2$ and $k = 3$ had the highest score, both were processed through the Random Forest algorithm. For $k = 3$, the high survival group comprised only 18 patients, medium and poor survival clusters numbered 254 and 810 entries respectively. For $k = 2$, the lower survival group was also considerably bigger (846 compared to 235 of higher survival). The noticeable imbalance in data distribution owes to the data's background.

3.3 Classification

Random Forests possess excellent prediction accuracy, ergo, they are employed particularly in forecasting treatment response in cancer cell lines, in localising tumour and in identifying its stage in patients.

In light of such massive amount of data, reliance solely on the Random Forest results might compromise the outcome of the analysis. The issue was handled by building several Random Forest Decision Trees trained on a set of key indicators of cancer severity (survival clusters for $k = 2$ and $k = 3$, metastasis, tumour size, nodes, etc.). The number of Decision Trees in the Random Forest varied from 50 to 100 and the depth of each tree was set at 5-7 to optimise the computing speed and to avoid overfitting of models.

Eventually, the number of genes was condensed from 19737 to several dozens, on which the further analyses were carried out (Table 2).

Table 2: Genetic prognostic markers classified by the Random Forest.

Group	Number of genes
Survival (2 clusters)	50
Survival (3 clusters)	59
Tumour size	94
ZNF497, C9orf106, GPR89B, AVP, CUEDC2, PTGIR, MCOLN1, LEKR1, GUSBP19, ZNF696, PRELID2, MTA1, SH2D7, DCP1B, CMTM2, C6orf26, SGCG, GNA12, CLIN, C13orf38, PAQR5, ZFPM1, MED1 ,HAPLN3, SLC2A10, C13orf23, B3GALT4, GREM2, THG1L, FAM55B, CYP3A7, C2orf47, FDX1, AKR1D1, ATP1B4, PPIF, NOS1, COQ10B, TARBP1, EIF4A1, OPLAH, TMPRSS15, ZDHHC21, ACBD6, TUBA4A, BTN2A1, FAM181A, ENTPD5, CUZD1, POM121L1P, IL22, FAM71F1, C15orf52, RDH16, LTA, SLC5A10, EBAG9, MED8, IL18BP, C3orf38, ASB17, PE12L, KLHL32, C3orf10, EMILIN1, FAM189B, MLYCD, FRA, C12orf34, C16orf96, BEX5, RFXANK, USP38, WDR36, CDKN1B, FHOD1, CADM2, SLC2A4, DHTKD1, CD84, GPR171, VPS18, SELPLG, C1orf177, TMEM181, CLASP2, C6orf162, NKD1, LY9, DNMB, MFAP3L, PTMA, CCL15CCL14, WASF2	
Cancer in lymph nodes	112
CSDE1, MKI67IP, HIAT1, ZNF711, ADRA2A, GCHFR, ALDOA, PMS2CL, BLVRA, MARS, SMUG1, ZNF542, CLEC1A, EGR4, C14orf4, ZNF629, AGGF1, C12orf65, SLC37A3, LOC100271832, OR1J4, EYA4, C10orf105, PDHX, ECHDC3, RFPL2, CPSF4, ITCH, SEC61B, ITIH3, C12orf47, PMS2L3, IGFBP5, CPNE7, PRKAR2A, FNBP4, MLLT1, JMJD5, SLC22A9, GABRA3, ZNF384, WRAP53, REEP6, RBM39, CLEC12A, PYGL, CTDSP2, NMT1, C1orf130, TMEM127, DYTN, FBXO46, PYGO2, SPCS2P4, BOLA1, ITGB4, ARHGEF5, TPM3P9, TDRKH, AGFG1, ERMN, PPP2R5A, FRYL, ARSG, KIF14, HLAB, RSPH6A, IWS1, LOC90784, CCDC23, SMIM6, PRRC1, FBXO33, TMEFF2, CPAMD8, RBM7, MAPK8IP3, CXorf30, CCPG1, MMR, MRPL41, CDK1, PCDH17, PARVA, TTPA, RUVBL2, ABCA11P, OR4N4, COBL, OGDHL, RPL23P8, ZIM2, BAZ2B, SH2B2, FKBP6, TECPR1, VPS37A, XKR9, MREG, KLHL31, C1orf52, CCDC59, DDX4, SRD5A1, C2orf34, RANBP3, LRR, WFDC5, GOLGA9P, ENPP5, SOX15, MAD2L2	
Metastasis	24
DPY19L3, LONRF1, SETDB1, C2orf3, MRPL9, GFAP, ZNF516, C9orf11, FAM2B, LRRC37A16P, ERP27, RNF121, FAM22A, PM2D2, LGALS4, EIF4E1B, SNORA13, WDR67, SASH3, CAPN7, C6orf81, RASA4CP, NUTM2A-AS1, PM20D2	

3.4 Tumour Stage

Tumour size is a centrepiece in diagnostics and prognosis of cancer, hence, it underlies the widely used TNM staging system. The Random Forest algorithm identified 94 genes which affect UPP, TGFβ1 and TLR4-dependent pathways, as well as the coverage of the progestin and adipoQ receptor family.

As the present findings show, *WASF2* stiffly impacts tumour growth [9]. It is notably to blame for activating the actin-related protein 2/3 complex, which enables migration and invasion of cancer cells. The upregulation of *PTMA* exacerbates histological malignancy and heightens the possibility of cancer recurrence [10]. The properties alike are viewed as potential therapeutic targets for cancer treatment.

CUEDC2 provokes endocrine resistance in patients by inhibiting estrogen and progesterone receptors via the ubiquitin-proteasome pathway. Therefore, it undermines the effect of such first-line treatment as tamoxifen in estrogen receptor-positive breast cancer patients [11].

PAQR5 regulates cyclic adenosine monophosphate synthesis and manipulation of *MAPKs*. The downregulation of *PAQR5* has its hand in the increased methylation of the promoter DNA and in a poor survival outcome in renal cancer patients [12].

Such genes as *PAQR5* and *CUEDC2* still lack rightful scientific attention and thus have not yet been institutionalised as prognostic markers. As our study indicates, the exploratory horizons of the matter in question are much to be broadened.

3.5 Lymph Nodes

Distant metastatic activity of cancer poses a pervasive threat for human health and wanes the chances for survival. Given the onset of the tumour cells spread usually takes place in the lymph nodes, such developments are regarded as a proven prognostic factor [13].

112 genes were chosen by the Random Forest algorithm and subsequently structured by the expression values in N0 group (no cancer in lymph nodes) (Figure 2). Predictably, the highest divergence is shown between the group free of seizures and that with the highest number of the invaded nodes. 64 genes of 112 chosen are downregulated in patients (N0-N2 stages) with *CTDSP2*, *RMP7*, and *C12orf65* being the most altered. The selected genes engage in such pathways as p53, Wnt, PI3K/Akt, VEGF and mtRQC, thereby influencing cell proliferation, tumouri- and angiogenesis.

The prognostic potential of the chosen genes is corroborated by recent studies. *CPSF4* is linked to the unfavourable cancer outcome, since it tends to upregulate the key cancer development genes, such as *MDM4* and *VEGF* [14]. *REEP6* itself showed almost 4-fold increase in the z-score in N3 group in contrast to N0, which implies its prospects as a novel

biomarker, resonating with a freshly published study on the Triple-Negative Breast Cancer prognosis [15].

GCHFR is considerably upregulated in cancer tissues. Furthermore, its expression has been associated with the activation of ferroptosis which renders a tumour resistant to therapy [16].

Pygo2 is the most overexpressed gene in the chosen set. It gives rise to the sensitivity to common chemotherapeutics, as well as promotes dedifferentiation and progression of cancer cells [17]. Such effect may have roots in the gene's ability to target the Wnt/-catenin pathway, which is critical for the proper cell proliferation and development. Although *Pygo2* has already undergone some study as a therapeutic target for metastatic breast cancer, the data on the matter is still insufficient, hence, the supposed correlation is subject to further examination.

Therefore, the further *in silico* and *in vitro* research of these genes is expected to bring about new insights indeed into molecular machinery of cancer progression, thus laying ground for the intricated set of cancer biomarkers to be developed.

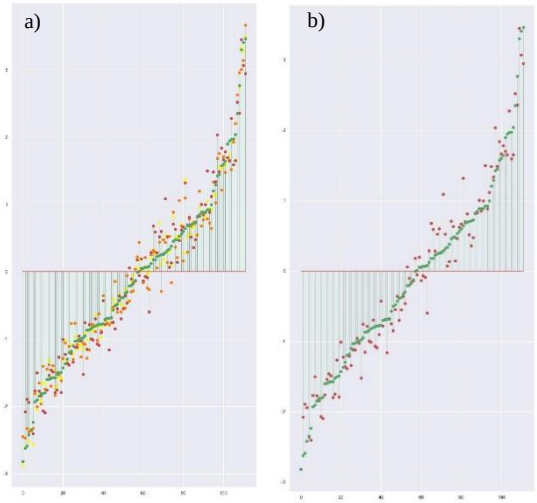

Figure 2: The expression of the classified genes a) in all groups; b) in N0 and N3 groups only. Horizontal and vertical axes represent genes and samples respectively.

3.6 Metastasis

Metastasis is defined as the spread of cancer cells from the initial tumour to the distant parts of the body via blood or lymph. It is a hallmark of cancer progression and is the most common reason for cancer-related deaths.

The present paper marks out 24 genes with divergent expression profiles in groups M0 (no metastasis) and M1 (metastasised cancer). Genes were categorised by z-score values in the metastasis-free group.

eIF4E is expressed almost seven-fold higher in M1 patients than in M0 within our database. There is evidence of *eIF4E* being employed to predict the rapamycin sensitivity in breast cancer cell lines. Furthermore, the cells might enhance resistance by altering the *eIF4E* activity [18].

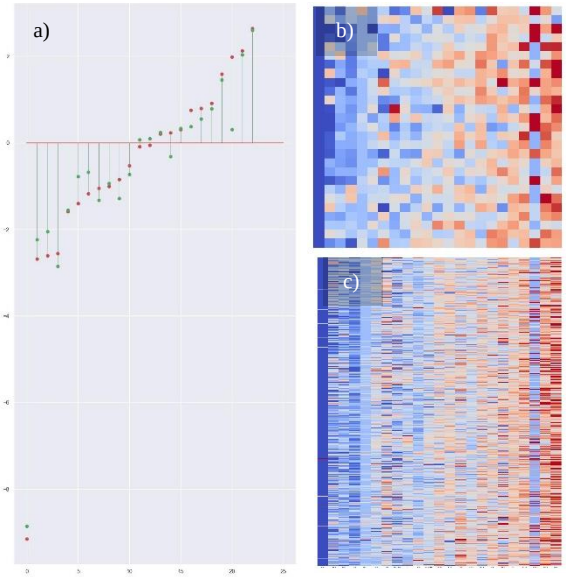

Figure 3: mRNA expression levels after *Z*-score normalization in BRCA tumour tissues (a) with M0 average values shown in green and M1 depicted in red. Heatmaps represents the expression values in all patients in (b) M0 and (c) M1 groups. Horizontal and vertical axes represent genes and samples respectively.

The *ZNF516* expression demonstrates dramatic increase in cancer tissues comparing to the healthy ones. The gene strongly influences the cell growth, proliferation and motility by repressing the transcription of the epidermal growth factor receptor (*EGFR*) [19]. *ZNF516* belongs to the protein transduction cascade which facilitates key processes in cancer development (e.g. angiogenesis, metastasis and immunosuppression). However, the detailed mechanism of such actions in breast cancer are still elusive [20].

The drastically low z-score (-8,8542 and -9,1529 for M0 and M1, respectively) in *SNORA13* gene could be precipitated by the errors to have taken place in performing the clinical analyses or curation of the results.

4 CONCLUSIONS

Machine learning algorithms like Random Forest are brilliant tools for processing vast amounts of data. The set of cancer-related genes was distilled to simplify the establishment of prognostic biomarkers.

The TNM cancer staging system includes 3 principal progression parameters – tumour diameter, lymph nodes invasion and metastasis, which were used for the classification. 230 genes were distinguished as prognostic factors, among which 94 impact tumour growth, 112 affect lymph nodes and 24 influence metastasising.

The present findings reveal the facilitating effects of such genes as *WASF2, CUEDC2, CPSF4, Pygo2, ZNF516* on the extension of cancerous tissues. These genes also engage in such pathways as p53, Wnt, PI3K/Akt, VEGF, UPP and TGFβ1, thereby influencing cell proliferation, tumouri- and angiogenesis.

The theoretical value of the present paper is expressed in employing the IT machinery in the analyses of yet unsystematised biomedical data; and in laying out the genetic canvas which is to be extended in our further works. The applicatory novelties of the research comprise setting up a basis for advanced diagnostic model to be utilised by breast cancer practitioners, as well as molecular biologists and scientists alike.

For the purposes of this study the initial version of the algorithm was employed, thus it has its limitations to be resolved in the offing. The prospective avenues of improvement include, *inter alia*, increasing credibility of the algorithm and implementing more sophisticated multifunctional analyses. With some fine-tuning, this model will be useful in guiding choices on how to treat patients with breast cancer.

ACKNOWLEDGMENTS

This work was partly supported by EMBO IG 4728-2020, Jacek Arct and ‘New Technologies for Women’ scholarships for NK and Alexander von Humboldt Stiftung (Philipp Schwartz-Initiative) for HF.

REFERENCES

[1] S.C. Shah, V. Kayamba, R.M. Jr. Peek, and D. Heimburger, “Cancer Control in Low- and Middle-Income Countries: Is It Time to Consider Screening?” vol. 5, pp. 1-8, 2019, doi: 10.1200/JGO.18.00200.

[2] J. Li, X. Guan, and et al., “Non-Invasive Biomarkers for Early Detection of Breast Cancer,” Cancers (Basel), vol. 12(10), pp. 2767, 2020, doi: 10.3390/cancers12102767.

[3] B.K. Banin Hirata, J.M. Oda, R. Losi Guembarovski, C.B. Ariza, C.E. de Oliveira, and M.A. Watanabe, “Molecular markers for breast cancer: prediction on tumor behavior,” Dis. Markers, vol. 2014, pp. 513158, 2014, doi: 10.1155/2014/513158.

[4] A.N. Richter and T.M. Khoshgoftaar, “A review of statistical and machine learning methods for modeling cancer risk using structured clinical data,” Artif. Intell. Med., vol. 90, pp. 1-14, 2018, doi: 10.1016/j.artmed. 2018.06.002.

[5] A. Zaremba, P. Zaremba, S. Siry, Y. Shermolovich, and S. Zagorodnya, “In vitro and in silico study of anti-influenza activity of 2-dioxopyrimidin-5-trifluoromethyl-tetrahydrothiophene with subsequent increase in its affinity for the target protein,” Proceedings of the 7th International Electronic Conference on Medicinal Chemistry, 1-30 November 2021, MDPI: Basel, Switzerland, doi:10.3390/ ECMC2021-11439.

[6] L. Peng, W. Chen, W. Zhou, F. Li, J. Yang, and J. Zhang, “An immune-inspired semi-supervised algorithm for breast cancer diagnosis,” Comput. Methods Programs Biomed., vol. 134, pp. 259-265, 2016, doi: 10.1016/j.cmpb.2016.07.020.

[7] H. Falfushynska, O. Lushchak, and E. Siemens, “The Application of Multivariate Statistical Methods in Ecotoxicology and Environmental Biochemistry,” Proceedings of International Conference on Applied Innovation in IT, vol. 10 (1), pp. 99-104, 2022.

[8] P. Rzymski, N. Kasianchuk, D. Sikora, and B. Poniedziałek, “COVID-19 Vaccinations and Rates of Infections, Hospitalizations, ICU Admissions, and Deaths in Europe during SARS-CoV-2 Omicron wave in the first quarter of 2022,” Journal of Medical Virology, vol. 95(14), 2022, doi: 10.1002/jmv.28131.

[9] P.S. Rana, A. Alkrekshi, W. Wang, V. Markovic, and K. Sossey-Alaoui, “The Role of WAVE2 Signaling in Cancer”, Biomedicines, vol. 9(9), pp. 1217, 2021, doi: 10.3390/biomedicines9091217.

[10] Y.H. Kuo, A.L. Shiau, and et al., “Expression of prothymosin α in lung cancer is associated with squamous cell carcinoma and smoking”, Oncol. Lett., vol. 17(6), pp. 5740-5746, 2019, doi: 10.3892/ol.2019.10248.

[11] S. Roy, S. Saha, and et al., “Molecular crosstalk between CUEDC2 and ERα influences the clinical outcome by regulating mitosis in breast cancer”, Cancer Gene Ther., vol. 29(11), pp. 1697-1706, 2022, doi: 10.1038/s41417-022-00494-x.

[12] C. Tao, W. Liu, and et al., “PAQR5 Expression Is Suppressed by TGFβ1 and Associated With a Poor Survival Outcome in Renal Clear Cell Carcinoma”, Front. Oncol., vol. 11, pp. 827344, 2022, doi: 10.3389/fonc.2021.827344.

[13] X. Chen and H. Ishwaran, “Random forests for genomic data analysis,” Genomics, vol. 99(6), pp. 323-9, 2012, doi: 10.1016/j.ygeno.2012.04.003.

[14] N.E. Reticker-Flynn, W. Zhang, and et al., “Lymph node colonization induces tumor-immune tolerance to promote distant metastasis”, Cell, vol. 185(11), pp. 1924-1942.e23, 2022, doi: 10.1016/j.cell.2022. 04.019.

[15] Y. Song, K. Sun, and et al., "CPSF4 promotes tumor-initiating phenotype by enhancing VEGF/NRP2/TAZ signaling in lung cancer", Med. Oncol., vol. 40(1), pp. 62, 2022, doi: 10.1007/s12032-022-01919-1.

[16] Y. Zhou, Y. Che, Z. Fu, H. Zhang, and H. Wu, "Triple-Negative Breast Cancer Analysis Based on Metabolic Gene Classification and Immunotherapy", Front. Public Health, vol. 10, pp. 902378, 2022, doi: 10.3389/fpubh.2022.902378.

[17] V.A. Kraft, C.T. Bezjian, and et al., "GTP Cyclohydrolase 1/Tetrahydrobiopterin Counteract Ferroptosis through Lipid Remodeling", ACS Cent Sci. vol. 6(1), pp. 41-53, 2020, doi: 10.1021/acscentsci.9b01063.

[18] M. Saxena, R.K.R. Kalathur, and et al., "2-Histone Interaction Is Critical for Cancer Cell Dedifferentiation and Progression in Malignant Breast Cancer", Cancer Res. vol. 80(17), pp. 3631-3648, 2020, doi: 10.1158/0008-5472.CAN-19-2910.

[19] S. Satheesha, V.J. Cookson, and et al., "Response to mTOR inhibition: activity of eIF4E predicts sensitivity in cell lines and acquired changes in eIF4E regulation in breast cancer", Mol. Cancer, vol. 10, pp. 19, 2011, doi: 10.1186/1476-4598-10-19.

[20] L. Li, X. Liu, L. He, and et al., "ZNF516 suppresses EGFR by targeting the CtBP/LSD1/CoREST complex to chromatin", Nat. Commun., vol. 8(1), pp. 691, 2017, doi: 10.1038/s41467-017-00702-5.