A Breast Cancer Patients' Survival Prediction Model Using mRMR Feature Selection with Artificial Neural Network

Wei Xing
School of Artificial Intelligence, Jilin University, Changchun, China
double.xing@foxmail.com

Kangping Wang*
College of Computer Science and Technology, Jilin University, Changchun, China
wangkp@jlu.edu.cn

Huiyan Sun
School of Artificial Intelligence, Jilin University, Changchun, China
huiyansun@jlu.edu.cn

Yan Wang†
1College of Computer Science and Technology, Jilin University, Changchun, China; 2School of Artificial Intelligence, Jilin University, Changchun, China
wy6868@jlu.edu.cn

ABSTRACT

Breast cancer is one of the most common cancers among women in the world, and it poses a huge threat to women's health. Predicting the survival status of breast cancer patients is of great significance to the patients. At present, due to the low classification ability of traditional machine learning algorithms, it is not enough to assist clinical diagnosis. This study combined deep learning with more powerful classification performance with medical diagnosis to improve the accuracy of diagnosis. This study constructed a model (MA) combining min-redundancy and max-relevance algorithm(mRMR) and artificial neural network, which could not only predict whether breast cancer patients would survive for more than 5 years, but also selected the features (genes)set that made classification results optimal. In terms of accuracy, the MA model's classification effectiveness could achieve 72.38%. Through survival analysis to the optimal genes set, 11 genes highly correlated with cancer survival were obtained.

CCS CONCEPTS

• **Applied computing**; • **Life and medical science**; • **Bioinformatics**;

KEYWORDS

Artificial neural network, Min-redundancy and max-relevance algorithm, Survival prediction

ACM Reference Format:
Wei Xing, Kangping Wang, Huiyan Sun, and Yan Wang. 2021. A Breast Cancer Patients' Survival Prediction Model Using mRMR Feature Selection with Artificial Neural Network. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487131

*Corresponding authors
†Corresponding authors

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487131

1 INTRODUCTION

Breast cancer is a malignant tumor that originates from breast epithelial cells. Worldwide, there were about 2.1 million newly diagnosed female breast cancer cases and 640 thousand deaths in 2018 [1]. Despite the survival rates in breast cancer have increased due to improvements in the treatments, breast cancer is still one of the leading causes of cancer-related death among women all over the world [2, 3].

Cancer patients' survival evaluation has an important reference value for cancer patient's diagnosis and treatment. Accurate prediction for cancer patients' survival can not only help patients understand their life expectancy to keep their mental health, but also assist clinicians in formulating precise treatment plans to ensure treatment effects. At present, clinicians usually judge how long a patient can survive based on the pathological stage and clinical characteristics. Limited by the doctor's experience, the survival prediction effect is poor. With the development of DNA microarray technology, massive cancer gene expression data have been accumulated. Extracting relevant gene characteristics from cancer gene expression data provides new ideas for cancer patient survival prediction.

In recent years, machine learning, especially deep learning has developed rapidly [4-6]. As a new research direction in the field of machine learning, deep learning has better performance than traditional machine learning methods (such as SVM) and classical statistical methods in classification. Medical diagnosis based on deep learning provides a new way to predict the survival status of cancer patients. Shreyesh compared the performance across three of the most popular deep learning architectures (ANN, CNN, and RNN) and found the ANN model was the best performing model in lung cancer survival period prediction [7]. Khloud proposed a supervised convolutional neural network (CNN) model trained on images of 612 breast cancers to determined high- versus low-grade breast cancer cells [8]. Cheng combined systems biology feature selection with bimodal deep neural network to predict breast cancer patients' 5-year disease-specific survival [9]. Chen proposed a gene superset autoencoder (GSAE) to predict breast cancer subtypes [10].

Cancer gene expression data generally have the characteristics of high dimensionality, small sample size, and non-linearity. High-dimensional gene expression data contains many redundant features, which brings a lot of difficulties to the classification research, so reducing the number of features is particularly important. Wang used PCA to reduce the dimension of data and diagnosed cervical cancer with data after dimension reduction [11]. Lai applied systems biology methods to identify genes related to cancer to achieve the purpose of reducing the number of features [12]. Toaar used the min-redundancy and max-relevance method to reduce the dimension of the feature [13].

Our research mainly included two parts. On the one hand, cancer patients' survival was predicted; on the other hand, some genes related to cancer survival were found. In terms of discovering biomarkers related to cancer, most of the current researches were based on bioinformatics methods. Our research innovatively discovered genes related to cancer survival using neural network and bioinformatics methods.

2 MATERIALS AND METHODS

2.1 Data

As one of the most authoritative and comprehensive cancer databases in the world, the TCGA database (https://portal.gdc.cancer.gov//) contains 33 common cancer types' genomics data and cancer patients' clinical information, ensuring the credibility of these data [14]. Gene expression data and clinical information data (including cancer patients' survival time) of 1172 breast cancer samples (including 113 normal tissue samples and 1059 cancer tissue samples) were downloaded from the TCGA database.

2.2 Preprocessing

The preprocessing on gene expression data and clinical information data was as follows:

1) Low-expressed genes (gene expression level <10 in 80% of samples) were deleted from gene expression data.

2) Differentially expressed genes analysis (analysis of variance + fold change) between normal tissue samples and cancer tissue samples was performed. When the conditions (fdr<0.01 and $\log_2$|fold change|>1) were met, genes related to cancer were obtained.

3) Z-score was used to standardize gene expression data.

4) According to whether the patients survived for more than 5 years, the cancer patients were divided into two groups: long-survival patients and short-survival patients. 1 as the label of long-survival patients and 0 as the label of short-survival patients.

2.3 Differentially Expressed Genes Analysis

Differentially expressed genes analysis is a common analysis method in bioinformatics used to select genes related to cancer, which usually combines significance analysis and fold change to find differentially expressed genes in different groups.

Analysis of variance is one of the significance analysis methods, which is used to analyze the overall mean between different data sets to see if there are significant differences between them. As a classic significance test method, analysis of variance can consider the significance of multiple factors in one analysis.

Fold change is a method to determine the differentially expressed genes by calculating the ratio of gene expression levels under two conditions. In general, genes whose expression levels differ by 2 times or more between two groups are considered meaningful.

2.4 Min-Redundancy and Max-Relevance Algorithm (mRMR)

When high-throughput sequencing data was used to predict the survival time of cancer patients, a common problem was the "curse of dimensionality". In our research, the number of samples was limited, and the number of features was much larger than the number of samples, which brought certain difficulties to the learning and prediction of the model. Besides, data with multiple features and small sample size could lead to model over-fitting. Therefore, for problems involving a large number of features, reducing the dimensionality of the features became especially important.

mRMR obtained the best n features by maximizing the correlation between features and target variables and minimizing the redundancy between features. mRMR not only reduced the dimensionality of the data set without causing significant information loss but also selected the optimal combination of features among multiple genes. mRMR used mutual information (Eq.1) to measure the correlation between features and tags and the redundancy between features.

$$I(x, y) = \iint P(x, y) \log \frac{P(x, y)}{P(x) P(y)} dxdy \tag{1}$$

I () was the mutual information function. *P ()* was the probability density function. x and y were random variables.

As shown in Eq.2, mRMR used an incremental search method to find the feature in $X\text{-}S_{m-1}$ that could maximize the correlation between features and tags and minimize the redundancy between features based on S_{m-1}.

$$\max_{x_j \in X - S_{m-1}} [I(x_j, c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i)] \tag{2}$$

S_{m-1} was the selected feature set. $X\text{-}S_{m-1}$ was the remaining feature set. x_i, x_j were the features. *I ()* was the mutual information function. c was the target variables.

2.5 Artificial Neural Network

The artificial neural network is an important algorithm model in deep learning, and its performance is greatly improved compared with traditional machine learning models. The artificial neural network is composed of three parts: input layer, hidden layer, and output layer. Among them, the connection mode of two adjacent layers' network nodes is full connection. The artificial neural network can effectively handle complex nonlinear systems and use hidden layers in the network to abstract data at a high level, which improves the generalization ability of the entire model. Therefore, the artificial neural network has stronger learning ability, self-learning, and self-adaptability.

A large number of genes are expressed in cancer patients. Genes are not isolated, they are connected, interact with each other, participate in or affect the same biological process. Considering the complex relationship network between genes, it is difficult for traditional machine learning models to fit the correlations, while neural

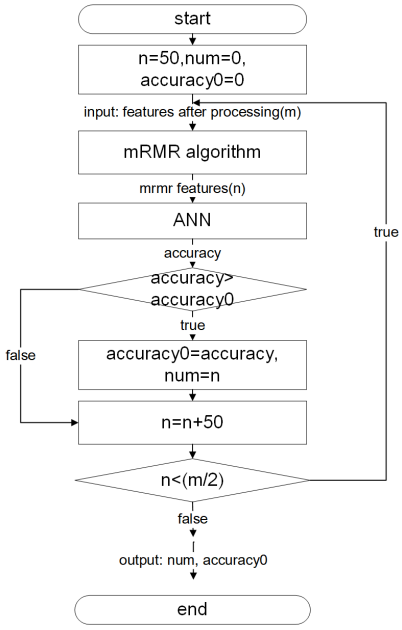

Figure 1: The Introduction to the MA Model.

networks can better simulate the nonlinear complex system between genes.

2.6 MA Model

MA model was combined by the min-redundancy and max-relevance algorithm and the artificial neural network. The artificial neural network was composed of 1 input layer, 2 hidden layers, and 1 output layer, with the rectified linear unit (ReLU) as the activation function, cross-entropy loss as loss function, and Adam as the optimizer. The number of neurons in the first hidden layer was the largest integer less than (the number of input features)/2. The number of neurons in the second hidden layer was the largest integer less than (the number of neurons in the first hidden layer)/2. The parameters were set as following, learning rate = 0.001, decay =0.0001.

Features after preprocessing were used as input. After the input features were processed by min-redundancy and max-relevance algorithm, n min-redundant and max-relevant features were obtained. Then, n min-redundant and max-relevant features were used for classification through an artificial network. The process was repeated i (i was a positive integer &0<50i< (the number of features after preprocessing)/2) times, and the number of min-redundant and max-relevant features(n) was 50*i. Finally, the features set that made the model's classification effectiveness best were obtained. The introduction to the MA classifier model was shown in Figure 1

2.7 Survival Analysis

Survival analysis is a statistical method that considers both results and survival time. Survival analysis can make full use of the incomplete information provided by the censored data, describe the distribution characteristics of survival time, and analyze the main

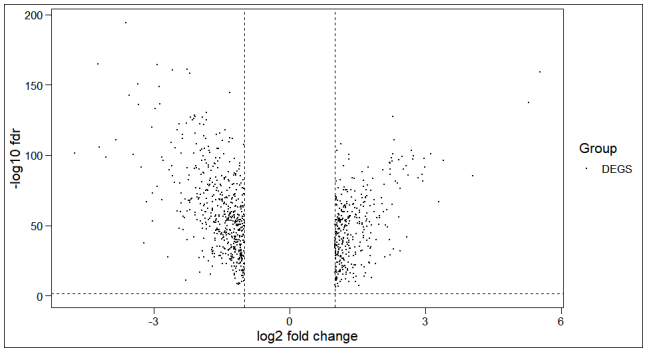

Figure 2: 889 Differentially Expressed Genes (DEGS).

factors affecting survival time. In brief, survival analysis is a common bioinformatics method used to find factors related to survival time.

3 RESULTS

3.1 Data after Preprocessing

As shown in Figure 2, after preprocessing on the gene expression data, 889 differentially expressed genes (DEGS) were obtained. Concerning clinical information data, a total of 138 samples (patients) had the information "survival time". According to the clinical information data, we found that among the 138 patients, the number of long-survival patients was 46 and the number of short-survival patients was 92. Finally, 138 patients' differentially expressed genes expression data (138*889) and label data (138*1) were used for subsequent analysis.

3.2 mRMR vs. PCA

PCA and mRMR have some similarities:

1) PCA and mRMR can reduce the dimensionality of data when facing high-dimensional data.

2) Any number of features can be obtained.

We were not sure which PCA or mRMR had better performance on feature reduction, so the classification effectiveness of artificial neural network was compared under the different number of features after being processed by PCA (or mRMR) (as shown in Figure 3). The artificial neural network was the same as the one in the MD model. According to Figure 3, when PCA (or mRMR) reduced the number of features (889 genes after preprocessing) to 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, after 5-fold cross validation was performed 5 times, in terms of the average accuracy, mRMR was better than PCA.

3.3 Optimal Feature Set

Features (889 genes) after preprocessing were used as input to find the feature set that made the MA model's classification effect the best by comparing the classification effectiveness of the MA model under different numbers of min-redundant and max-relevant features.

When the number of features after min-redundancy and max-relevance algorithm processing was 50, 100, 150, 200, 250, 300, after

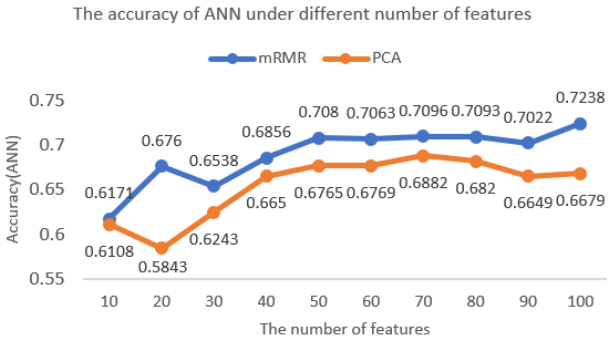

Figure 3: The Accuracy of ANN under the Different Number of Features (mRMR vs. PCA).

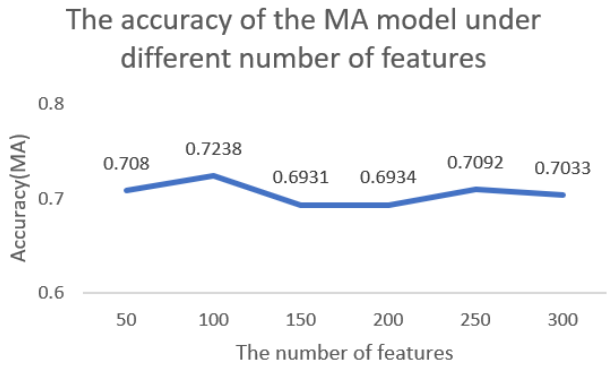

Figure 4: The Accuracy of the MA Model under the Different Number of Features.

5-fold cross-validation was performed 5 times, the average accuracy of the artificial neural network was 0.7080, 0.7238, 0.6931, 0.6934, 0.7092, 0.7033(Figure 4). Finally, the MA model's classification effectiveness got the highest accuracy at 72.38% when the number of min-redundant and max-relevant features was 100.

3.4 Model Comparison

To verify the effectiveness of the proposed cancer survival prediction model MA, three methods were adopted for comparison, respectively the model (SB-DNN) Cheng [9] proposed, support vector machines (SVM), and random forests (RF). When SVM and random forest were used for comparison, they have the same preprocessing as the MA model. The entire data set (889 genes expression data) was randomly divided into the training set and test set according to the ratio of 8:2. In order to ensure the fairness and robustness of the research methods, the data set was randomly divided into 5 times, and 5 experiments were carried out for each research method, and the final evaluation index was the average of the 5 experiments. In addition, in order to further verify the reliability of the proposed model, four performance evaluation indicators were calculated for each model-precision, accuracy, recall, and F1-score. Through the evaluation of four models' results, four models' classification effects

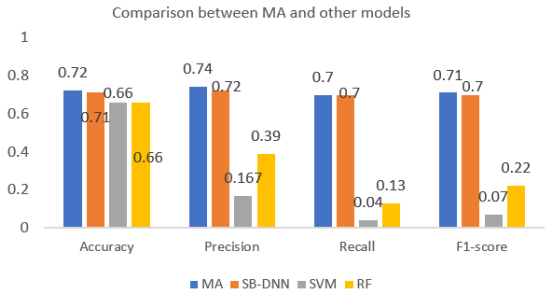

Figure 5: Comparison between the MA Model and Other Models.

were gotten. According to Figure 5, the accuracy, precision, recall, and F1-score of the MA model were 0.7238, 0.7367, 0.7040, and 0.7084. According to the comparison results with other models, we found that the MA model was better than SB-DNN, support vector machines, and random forests in terms of classification effect, so the MA model had good performance on cancer patient survival prediction. We implemented SVM and random forests using sklearn 0.0 (pycharm, python).

3.5 Genes Related to Survival Prognosis

We analyzed 100 min-redundant and max-relevant genes and found that 11 of them were highly correlated to the survival of cancer patients by use of cox regression model (p-value<0.05). As shown in Table 1, these 11 genes related to the patient's survival prognosis were ARNT2, CRTAP, ARHGAP23, TIMM17A, LMNB2, EZR, SIAH2, SLC37A1, FRMD4A, SOX12, and HMGB3.

ARNT2 was positively associated with the prognosis of breast cancer [15]. After being affected by increased expression of the ARPC1B gene, CRTAP might exert a positive effect on treated patients [16]. ARHGAP23's genetic alterations led to urothelial carcinoma [17]. TIMM17A had potential in the prognosis and treatment of breast cancer [18]. LMNB2 depletion suppressed the proliferation and induced the apoptosis of triple-negative breast cancer cells [19]. EZR-AS1 knockdown significantly suppressed the proliferation and cell cycle progression of breast cancer cells [20]. The genetic deficiency in SIAH2 resulted in vascular normalization and delayed tumor growth [21]. The up-regulation of SLC37A1 gene expression played a well-known stimulatory effect in breast cancer cells [22]. SOX12's knockdown significantly inhibited the proliferation, migration, and invasion of breast cancer cells [23]. Regulation of HMGB3 by miR-205 reduced both proliferation and invasion of breast cancer cells and there was an indirect correlation between the expression of HMGB3 mRNA and patient survival [24]. To summarize, 11 genes we identified were related to cancer, and most of the 11 genes were related to the survival and prognosis of patients, which further validated our experimental results.

4 DISCUSSION

In the field of cancer medicine, cancer is closely related to the expression of multiple genes, and the study of genes can discover the biological mechanism behind cancer. As a classic feature reduction

Table 1: Genes Related to the Patient's Survival Prognosis.

Id	Gene	P-value
9915	ARNT2	0.0220
10491	CRTAP	0.0413
57636	ARHGAP23	0.0158
10440	TIMM17A	0.0172
84823	LMNB2	0.0238
7430	EZR	0.0063
6478	SIAH2	0.0452
54020	SLC37A1	0.0396
55691	FRMD4A	0.0306
6666	SOX12	0.0109
3149	HMGB3	0.0028

method, PCA can play a good role in dimensionality reduction. PCA forms new orthogonal features by mapping n-dimensional features to K-dimensional. Although the new features can retain useful information from the original features, the new features have no biological significance. mRMR used to reduce features not only achieves the effect of dimensionality reduction but also the features obtained by screening can also be used for biological analysis. Toaar [13] and Wang [25] used mRMR in their research, but they didn't make a comparison between PCA and mRMR. Due to similarities between PCA and mRMR, a comparison was needed.

In predicting breast cancer patients' survival, Cheng [9] used a different data set from ours. When we repeated Cheng's experiment based on TCGA data, we found that SB-DNN didn't perform better than the MA model. The clinical information provided by TCGA wasn't complete. In terms of breast cancer, there were a total of 1172 samples in TCGA, but only 138 samples (patients) had the information "survival time". Perhaps this was the reason few people used TCGA data for survival prediction. Besides, due to the small sample size and the impact of other uncertain factors on patient survival (such as patients' mentality), it was difficult to improve the accuracy of classification.

Comparing the MA model with other models was only part of our research. The ultimate goal of our research was to find genes related to breast cancer survival. Discovering genes related to cancer survival using neural network and bioinformatics methods was meaningful to targeted therapy and precision medicine.

5 CONCLUSION

In terms of dimensionality reduction effectiveness, compared with accuracy, mRMR was better than the traditional dimensionality reduction method PCA. Through the combination with the artificial neural network, we found that when the number of features after being processed by mRMR was 100, the classification effectiveness of the MA model was the best (72.38%). In addition, after the survival analysis of these 100 features which made the MA model's performance the best, we found that 11 of the 100 genes were related to the patient's survival prognosis. As for the specific biological mechanisms of these 11 genes that affected cancer patients' survival prognosis, further biological research was needed. Through comparison, the proposed MA method had a higher average accuracy value, so the MA model showed a more excellent classification effect relative to other classification methods.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Nos. 62072212, 61902144), the Development Project of Jilin Province of China (Nos. 20200401083GX, 2020C003). This work was also supported by Jilin Province Key Laboratory of Big Data Intelligent Computing (No. 20180622002JC).

REFERENCES

[1] BSc, F. B. and ME, J. F. Global cancer statistics (2018). GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.

[2] Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L. and Ngom, A (2019). A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. *Front Genet*, 10, 256.

[3] Jubair, S., Alkhateeb, A., Tabl, A. A., Rueda, L. and Ngom, A (2020). A novel approach to identify subtype-specific network biomarkers of breast cancer survivability. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 9, 1.

[4] Jansen, T., Geleijnse, G., Van Maaren, M., Hendriks, M. P., Ten Teije, A. and Moncada-Torres, A (2020). Machine Learning Explainability in Breast Cancer Survival. *Stud Health Technol Inform*, 270, 307-311.

[5] Li, R., Shinde, A. and Liu, A (2020). Machine Learning–Based Interpretation and Visualization of Nonlinear Interactions in Prostate Cancer Survival. *JCO Clin Cancer Inform*.

[6] Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. and Geleijnse, G (2021). Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci Rep*, 11(1): 6968.

[7] Doppalapudi, S., G.Qiu, R. and Badr, Y (2020). Lung cancer survival period prediction and understanding: Deep learning approaches. *International Journal of Medical Informatics*.

[8] Elsharawy, K. A. and Gerds, T. A (2021). Artificial intelligence grading of breast cancer: a promising method to refine prognostic classification for management precision. *Histopathology*.

[9] Cheng, L.-H., Hsu, T.-C. and Lin, C (2021). Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction. *Scientific Reports*.

[10] Chen, H.-I. H., Zhang, T. and Zhang, S (2018). GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. *Bmc Systems Biology*.

[11] Wang, J., Li, L. and Yang, P (2018). Identification of cervical cancer using laser-induced breakdown spectroscopy coupled with principal component analysis and support vector machine. *Lasers in Medical Science*, 33, 1381-1386.

[12] Lai, Y.-H., Chen, W.-N. and Hsu, T.-C (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific Reports*, 10(1): 4679.

[13] Toaar, M., Ergen, B. and Cmert, Z (2019). A Deep Feature Learning Model for Pneumonia Detection Applying a Combination of mRMR Feature Selection and Machine Learning Models. *IRBM*.

[14] Tomczak, K., Czerwińska, P. and Wiznerowicz, M (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology*, 19, 1A, A68-77.

[15] Xian-Yang, Q., Feifei, W., Jun, Y. and Junzo, Y (2011). siRNA-mediated knockdown of aryl hydrocarbon receptor nuclear translocator 2 affects hypoxia-inducible factor-1 regulatory signaling and metabolism in human breast cancer cells. *Febs Letters*, 585(20): 3310-3315.

[16] Belickova, M., Cermak, J., Merkerova, M. D. and Vesela, J (2012). Changes Associated With Lenalidomide Treatment in the Gene Expression Profiles of Patients With Del(5q). *Clin Lymphoma Myeloma Leuk*, 12, 5.

[17] Park, H. S. and Park, J(2020). Seminal Vesicle Involvement by Carcinoma In Situ of the Bladder: Clonal Analysis Using Next-Generation Sequencing to Elucidate the Mechanism of Tumor Spread. *Cancer Research and Treatment*.

[18] Yang, X., Si, Y. and Tao, T (2016). The Impact of TIMM17A on Aggressiveness of Human Breast Cancer Cells. *Anticancer Research*, 36(7):1237.

[19] Zhao, C.-C. and Chen, J (2021). Lamin B2 promotes the progression of triple negative breast cancer via mediating cell proliferation and apoptosis. *Bioscience Reports*.

[20] Bai, Y., Zhou, X., Huang, L., Wan, Y. and Li, X (2018). Long noncoding RNA EZRAS1 promotes tumor growth and metastasis by modulating Wnt/βcatenin pathway in breast cancer. *Experimental and Therapeutic Medicine*.

[21] Jansen, M. P. H. M., Ruigrok-Ritstier, K., Dorssers, L. C. J. and Staveren, I. L. v (2009). Downregulation of SIAH2, an ubiquitin E3 ligase, is associated with resistance to endocrine therapy in breast cancer. *Breast Cancer Research&Treatment*, 116(2): 263-271.

[22] Domenico, I., Rosamaria, L. and Anna Rita, C (2010). SLC37A1 Gene expression is up-regulated by epidermal growth factor in breast cancer cells. *Breast Cancer Research&Treatment*, 122(3): 755-764.

[23] Ding, H., Quan, H., Yan, W. and Han, J (2016). Silencing of SOX12 by shRNA suppresses migration, invasion and proliferation of breast cancer cells. *Bioscience Reports*, 36, 5, e00389-e00389.

[24] Elgamal, O. A., Park, J.-K., Gusev, Y. and Azevedo-Pouly, A. C. P (2013). Tumor Suppressive Function of mir-205 in Breast Cancer Is Linked to HMGB3 Regulation. *Plos One*, 8, 10, e76402.

[25] Wang, C., Guo, J. and Zhao, N (2019). A Cancer Survival Prediction Method Based on Graph Convolutional Network. *IEEE Transactions on NanoBioscience*, PP, 99, 1-1.