Revista Argentina de Clínica Psicológica
2021, Vol. XXX, N°2, 420-432
DOI: 10.24205/03276716.2020.4040

GSEA–SDBE: A Gene Selection Method for Breast Cancer Classification Based on GSEA and Random Forest Based Dimension Reduction

Tailong Lei

Abstract

Selecting the most relevant genes for sample classification is a common process in gene expression studies. Moreover, determining the smallest set of relevant genes that can achieve the required classification performance is particularly important in diagnosing cancer and improving treatment. In this study, I propose a novel method to eliminate irrelevant and redundant genes, and thus determine the smallest set of relevant genes for breast cancer diagnosis. The method is based on random forest models, gene set enrichment analysis (GSEA), and our developed Sort Difference Backward Elimination (SDBE) algorithm; hence, the method is named GSEA–SDBE. Using this method, genes are filtered according to their importance following random forest training and GSEA is used to select genes by core enrichment of Kyoto Encyclopedia of Genes and Genomes pathways that are strongly related to breast cancer. Subsequently, the SDBE algorithm is applied to eliminate redundant genes and identify the most relevant genes for breast cancer diagnosis. In the SDBE algorithm, the differences in the Matthews correlation coefficients (MCCs) of performing random forest models are computed before and after the deletion of each gene to indicate the degree of redundancy of the corresponding deleted gene on the remaining genes during backward elimination. Next, the obtained MCC difference list is divided into two parts from a set position and each part is respectively sorted. By continuously iterating and changing the set position, the most relevant genes are stably assembled on the left side of the gene list, facilitating their identification, and the redundant genes are gathered on the right side of the gene list for easy elimination. A cross-comparison of the redundancy difference comparison elimination (RDCD) algorithm was performed by respectively computing differences between MCCs and ROC_AUC_score and then respectively using 10-fold classification models, e.g., RF, SVM, KNN, XGBoost, and ExtraTrees. Results showed that analyzing MCC differences and using random forest models was the optimal solution for the RDCD algorithm. Accordingly, three consistently relevant genes (i.e., *VEGFD*, *TSLP*, and *PKMYT1*) were selected for the diagnosis of breast cancer. The performance metrics (MCC and ROC_AUC_score, respectively) of the random forest models based on 10-fold verification reached 95.28% and 98.75%. In addition, survival analysis showed that *VEGFD* and *TSLP* could be used to predict the prognosis of patients with breast cancer.

Keywords: gene set enrichment analysis, random forest, backward elimination, redundant genes

1 Introduction

Selecting relevant genes to distinguish patients with or without cancer is a common task in gene expression research (Hartmaier *et al.*, 2017; Giovannantonio *et al.*, 2020). For genetic diagnosis in clinical practice, it is important to efficiently identify relevant genes and eliminate irrelevant and redundant genes to obtain the smallest possible gene set that can achieve good predictive performance (Dı´az-Uriarte *et al.*, 2006).

To this end, genetic selection methods are of great importance. These methods can be roughly divided into three categories: filters, wrappers, and mixers (Pok *et al.*, 2010). In a previous study, we focused on a hybrid approach that combines the advantages of filter and wrapper methods (Xie *et al.*, 2011). For cancer classification, previous hybrid

College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310058, China

approaches have utilized symmetrical uncertainty to analyze the relevance of genes based on support vector machines (Piao *et al.*, 2012), employed minimum redundancy and maximum relevance feature selection to select a subset of relevant genes (Elyasigomari *et al.*, 2017), and applied Cuckoo search to select genes from microarray technology (Sampathkumar *et al.*, 2020).

The hybrid approach essentially includes two processes, selecting relevant genes and eliminating redundant genes. To select relevant genes, previous research has utilized semantic similarity measurements of gene ontology terms based on definitions for similarity analysis of gene function (Pesaranghader *et al.*, 2016), applied the concept of global and local gene relevance to calculate the equivalent principal component analysis load of nonlinear low-dimensional embedding (Philipp *et al.*, 2020), and obtained relevant features from the TCGA transcriptome dataset by cooperative embedding (Shuzhen *et al.*, 2020). Because relevant genes often contain redundant genes, the process of gene elimination is important for obtaining the minimal number of relevant genes that can function effectively in a classification model. Many methods can be applied including feature similarity estimated by explicitly building a linear classifier on each gene (Zeng *et al.*, 2008), homology searching against a gene or protein database (Ono *et al.*, 2015), or the Cox-filter model (Suyan, 2018).

In the present study, I propose a novel hybrid method that can determine the smallest set of relevant genes required to achieve accurate classification of breast cancer diagnosis. Breast cancer transcriptome data can be downloaded from the TCGA database; this unbalanced data was used in the current analyses. Random forest and gene set enrichment analysis (GSEA) were applied to select relevant breast cancer genes and the proposed redundancy difference comparison elimination (RDCD) algorithm was then used to eliminate redundant genes from these relevant genes; hence, the proposed method was named GSEA–SDBE (where SDBE is Sort Difference Backward Elimination). First, a random forest model was constructed and trained with all the differential gene expression data and then the genes for which importance was almost zero were deleted. Subsequently, GSEA was applied to analyze the remaining differentially expressed genes (DEGs) according to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment and those genes that were strongly related to breast cancer were selected from the enriched KEGG pathways. Then, the RDCD algorithm was applied to identify the important relevant genes from the selected genes. The RDCD algorithm includes a process by which the difference in the Matthews correlation coefficients (MCCs) of random forest models is calculated before and after the deletion of a given gene, which indicates the degree of redundancy of the corresponding deleted gene on the remaining genes according to backward elimination. Using the RDCD algorithm, the most relevant genes are stably collected on the left side of the gene list while the redundant genes are gathered on the right side of the gene list. Through the GSEA–SDBE method, an optimal model was created that could determine the smallest set of relevant genes for breast cancer diagnosis. Results showed that this method could achieve excellent classification performance for breast cancer. Furthermore, some of the selected relevant genes could be used to predict prognosis in patients with breast cancer.

2 Materials and methods

2.1 Data preparation

2.1.1 Breast cancer transcriptome data

Transcriptome data from breast cancer samples and the clinical data of corresponding patients were downloaded from TCGA database (https://cancergenome.nih.gov/). A total of 1222 transcriptome samples, wherein each sample contained expression of 18584 genes, were obtained. This unbalanced dataset, which includes 113 normal and 1109 tumor tissues, was named BTC_1222 (113: 1109). In addition, the clinical data of 1109 patients with breast cancer were obtained.

2.1.2 Differential expression analysis and normalization

By performing the Mann–Whitney–Wilcoxon test in R software (wilcox.tes) with logFC > 1 and FDR < 0.05 as the thresholds, 4579 DEGs were screened between the normal samples and tumor samples from the BTC_1222 dataset. These samples were randomly shuffled and the expression values of each DEG in all samples were respectively standardized via min–max normalization.

2.2 Selecting genes by importance based on a random forest model

The random forest method can provide an assessment of variable importance to variable selection (Deng *et al.*, 2013; Alikovi *et al*., 2017). A random forest model was constructed and trained using Sklearn 0.22.2. post1 in python 3.6 with the dataset BTCD_1222. The model was used to calculate the importance of variables (genes) and the genes were sorted by their importance in

descending order. From these genes, a certain number of top genes were selected based on experience to reduce the burden of subsequent procedures.

2.3 Gene selection by GSEA

GSEA (Aravind *et al.*, 2007) can be used to determine whether a group of genes shows statistically significant and concordant differences between two biological states according to enrichment analysis; here, it was performed by the JAVA program. The KEGG database includes a collection of manually drawn graphical maps known as KEGG pathway maps (Ogata *et al.*, 1999). KEGG in the Molecular Signatures Database (MSigDB) (Liberzon *et al.*, 2011) was chosen as the back-end database of GSEA. GSEA was run and genes were selected through the core enrichment (Reimand *et al.*, 2011) of KEGG pathways strongly related to breast cancer. Therefore, it was possible to screen for DEGs that were closely associated with breast cancer. Genes that were weakly associated with or were unrelated to breast cancer were filtered out, even if they had high importance in a random forest model.

2.4 Metrics and benchmark methods

The performances of all classification models applied in this study were evaluated by 10-fold cross-validation. The models were trained and tested with 10-fold cross-validation. According to the prediction results and tested data, they were respectively merged in a given order. By comparing the prediction results with the tested data, true positives (*TP*), false positives (*FP*), false negatives (*FN*), and true negatives (*TN*) were obtained. Normal samples were negatives and tumor samples were positives. Tests were conducted on a real dataset with unbalanced data. Therefore, the effectiveness of the binary classification model was measured by several performance metrics (Robinson *et al.*, 2010) including accuracy (*Acc*), precision (*Pr*), sensitivity (*Se*), recall (*Re*), F1_score (*F1*), computed area under the receiver operating characteristic curve from prediction scores (*ROC_AUC_score*), and *MCCs*. The formulas and functions are as follows:

ROC_AUC_score = sklearn.metrics.roc_auc_score (1)

$$Acc = \frac{TN + TP}{TN + TP + FP + FN} \quad (2)$$

$$Re = \frac{TP}{TP + FN} \quad (3)$$

$$Pr = \frac{TP}{TP + FP} \quad (4)$$

$$F1 = \frac{2 \times (\mathrm{Pr} \times Re)}{\mathrm{Pr} + Re} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

$$Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (7)$$

2.5 SDBE algorithm

The training, testing, and calculation of various performance metrics for all classification models were based on 10-fold cross-validation. The focus was on finding a high-performance classification model with the fewest variables (genes); subsequently, a novel algorithm, namely SDBE, was proposed. The underlying principle of the SDBE algorithm is that the performance metrics of the classification model will not change significantly after a redundant gene is deleted. Therefore, the differences in the chosen performance metrics were computed before and after deletion of each gene to indicate the degree of redundancy of the corresponding deleted gene on the remaining genes in backward elimination based on the random forest method. These deleted genes were collected into a list in reverse order during backward elimination (John *et al.*, 1994).

From a set position, genes were sorted by their corresponding performance metric differences in descending order into the two parts and the two parts were then merged. Through continuously iterating and changing the set position, the important relevant genes were stably assembled on the left side of the gene list to facilitate their easy identification, whereas redundant genes were gathered on the right side of the gene list for easy elimination. The procedure underlying the SDBE algorithm is provided in Figure 1. The SDBE algorithm consists of seven stages as follows.

Stage 1

In each loop of backward elimination, 10-fold random forest models were trained and tested to calculate various performance metrics and the average importance of each variable, i.e., each gene. Next, these genes were sorted in descending order of average importance. After each loop of backward elimination, the deleted gene with the least importance and various metrics of the model were added to various dedicated lists. Thus, by

respectively transposing all the lists, a list of genes *G* (g_k, $0 \le k \le n$) in descending order of importance and various metric lists were obtained. These lists were provided to the stages that followed. Importantly, gene g_0 at the first position in the list of the genes was determined at this stage because the position of this gene would not change in subsequent steps.

Stage 2

One of model performance metrics, such as MCC or ROC_AUC_score, was chosen as the object of difference analysis for subsequent steps and the index variable *ST* was initialized to 0.

Stage 3

The following formula was used to compute the difference in the performance metric before and after gene deletion during backward elimination based on random forest modeling:

$$dm_i = m_i - m_{i-1}, 0 < i \le n,$$

where m_i and m_{i-1} respectively denote the metric before and after deleting gene g_i $(0 < i \le n)$ from sublist Gs (g_u, $0 \le u \le i$, $0 < i \le n$) of gene list G (g_k, $0 \le k \le n$) in backward elimination. Only one gene was deleted from the end of list Gs at each loop in backward elimination. The performance metric difference dmi $(0 < i \le n)$ could indicate the degree of redundancy of the corresponding deleted gene g_i $(0 < i \le n)$ on the remaining genes of sublist Gs.

Stage 4

The value of the variable *ST* was used as the index position to search forward in the metric difference list *DM* (dm_i, $0 < i \le n$) until an element <0 was encountered; the index of this element was used to update the variable *ST*.

Stage 5

The metric difference list *DM* was split into two parts, part1 and part2 (including the element at index *ST*) by index *ST*, and then the elements in part1 and part2 were respectively sorted in a descending order.

Stage 6

The elements of part1 and part2 were replaced with genes by the corresponding relationship between dm_i $(0 < i \le n)$ and g_i $(0 < i \le n)$, and then the two parts were merged into a new gene list *NG*. Subsequently, g_0 in the list *G* was added to the end of the new list *NG.* Then, the list *NG* was transposed.

Stage 7

The genes of the list *NG* were analyzed by backward elimination. At each step of backward elimination, the 10-fold classification mode, e.g., RF, SVM [], KNN, XGBoost, and ExtraTrees, was trained and tested to calculate various performance metrics. After each step of backward elimination, the performance metrics were respectively added to the corresponding metric lists. Then, the metrics lists, which were respectively transposed, and the list *NG* were sent to stage 3 to start a new iteration.

2.6 The entire pipeline of the GSEA–SDBE method

The gene selection procedure followed in the GSEA–SDBE method is provided in Figure 2.

3 Results

3.1 Differential expression analysis and normalization

From 4579 DEGs identified in from the BT_1222 dataset, 2702 were upregulated and 1877 were downregulated, respectively. These genes are represented in a volcano plot in Figure 3.

3.2 Random forest models

Having trained a random forest model with data on 4479 DEGs, the out-of-bag error was 0.01%. Genes were sorted by their importance in descending order, as shown in Figure 4. Selecting the top 2000 genes from the 4579 DEGs was optimal in the experiments; thus, the remaining 2579 genes, for which the importance was close to zero, were deleted.

3.3 GSEA

GSEA 3.0 was applied to analyze 2000 DEGs with KEGG pathways enrichment; the gene sets database was set to c2.cp.kegg.v7.1.symbols.gmt of the MSigDB. In enrichment results, 30 gene sets were obtained. These included five and 15 upregulated and downregulated gene sets in the phenotype "Tumor" (Supplementary Table S1), respectively. Four gene sets (Table 1) were selected that were strongly associated with breast cancer (Figure 5). Altogether, 60 genes were identified, including 20 upregulated genes and 40 downregulated genes, after deleting 12 repeated downregulated genes from 72 genes in the core enrichment of the four gene sets.

3.4 SDBE algorithm

In the SDBE algorithm, the training, testing, and calculation of various performance metrics for all classification models were based on 10-fold cross-validation. The expression data of 60 genes from

the GSEA enrichment analysis results were used in the SDBE algorithm. From stage 1 of the algorithm, 60 genes were listed in descending order of importance, as shown in Supplementary Table S2, and various metric lists (including acc, Pr, Se, Re, F1_score, ROC_AUC_score, and MCC) were illustrated using matplotlib in python 3.6 for comparison (see the red polylines in Figure 7). It was difficult to select the smallest gene set that could still achieve good predictive performance by sorting genes by their importance, although ranking genes by importance was vital to the process. The most important part of this step was determining the top gene in the list as this gene does not change in subsequent steps. From this stage, the gene and metric lists were passed to the stages that followed.

In stage 2 of the SDBE algorithm, the performance metrics ROC_AUC_score and MCC were respectively chosen as the objects of difference analysis for subsequent iterations; each iteration included stage 3–7 and the number of iterations was set at 19. To compare the influence of different classification models in the SDBE algorithm, the following were respectively chosen for use as the classification model: RF, SVM, KNN, XGBoost, and ExtraTrees. Therefore, the SDBE algorithm was cross-tested. Regardless of the object chosen for difference analysis (ROC_AUC_score or MCC) and the classification model (RF, SVM, KNN, XGBoost, or ExtraTrees) used, as the iteration progressed the most relevant genes were assembled in a stepwise manner on the left side of the gene list, whereas the redundant genes were gathered in a stepwise manner on the right side of the gene list (Figure 6). On the left side of the gene list, the identity and number of stable relevant genes differed depending on the analysis target and classification model, with three stable relevant genes being the maximum (Supplementary Table 2).

To cross-compare the SDBE algorithm, we used the 19th iterations of the algorithm and compared the same performance metrics of multiple classification models (RF, SVM, KNN, XGBoost, and ExtraTrees; Figure 6). As shown by the shapes of the polylines in Figure 6, using MCC as the object of difference analysis produced better results than using ROC_AUC_score. With MCC, the performance metrics of the RF model were better than the performance metrics of the other classification models; the blue polyline of the RF model was always above the other polylines. Therefore, we assessed the polyline of RF and found that the top three genes did not reach the peak or trough of the polyline but were close to each (Figure 6a). More importantly, the top three genes were stable and repeatable. Therefore, we extracted performance metrics of classification models trained and tested using the top three genes from Figure 6 for comparison (Tables 2 and 3). Except for FDR (1.77%), the relative performance metrics of the RF model in Table 2, showing MMC as the object, were superior to those in Table 3 (ROC_AUC_score as the object); moreover, the top three genes from the classification models RF, KNN, XGBoost, and ExtraTrees were identical when MMC was the object (Table 2) but typically differed among the models when ROC_AUC_score was the object (Table 3). Because the data used to train and test the classification models were unbalanced (113 vs. 1109 samples), the performance metrics MCC and ROC_AUC_score of the RF model were focused upon.

In summary, using MCC as the object of difference analysis and RF as the classification mode in the SDBE algorithm was optimal. In addition, three stable relevant genes, namely *VEGFD*, *TSLP*, and *PKMYT1*, were chosen for the diagnosis of breast cancer. Moreover, based on 10-fold verification, the performance metrics MCC and ROC_AUC_score for RF models were 95.28% and 98.75%, respectively.

3.5 Survival analysis of patients

First, patients were divided into two groups, high and low risk, based on the median expression of a certain gene. If the gene was downregulated, the patients whose expression of the gene was lower than the median expression were classified as high risk, whereas the remaining patients were low risk. If the gene was upregulated, the method of grouping was reversed.

Kaplan–Meier survival analysis and log-rank tests were used to determine the prognostic significance of expression of the three genes, *VEGFD*, *TSLP*, and *PKMYT1,* in patients with breast cancer. *VEGFD* and *TSLP* were downregulated genes, whereas *PKMYT1* was upregulated. A log-rank test revealed that patients with low *VEGFD* and *TSLP* expression had significantly shorter overall survival (OS) times than those patients with high expression of these genes (P = 0.0466 and P = 0.0003, respectively; Figure 8); the median OS times in months (with 95% confidence intervals) were 129 (114–142) and 116 (102–132), respectively; Figure 8 and Table 4). In contrast, the result of the log-rank test for *PKMYT1* was not significant (P = 0.2095) and the polylines of the high-risk and low-risk groups for this gene crossed at 120 months (Figure 8c). Therefore, *VEGFD* and *TSLP* could be used to predict prognosis in patients with breast cancer, whereas

PKMYT1 is not suitable for this purpose.

4 Discussion

In this study, DEGs were extracted from a breast cancer data set. Genes that are not significantly differentially expressed but have important biological significance for breast cancer could easily be missed in this process; however, even if these lost genes are retained, they may be deleted in subsequent processing. Indeed, such genes would be ignored by the classification model used in the GSEA–SDBE method described here. Nevertheless, this did not affect the ability of the method to identify some key genes for the diagnosis of breast cancer.

Dimensionality reduction runs through the entire GSEA–SDBE method; each step in the method prepares for dimensionality reduction in the next step. According to experience, selecting too few genes leads to some important pathways not being enriched, whereas selecting too many genes overfills the core enrichment of pathways with genes that make subsequent gene elimination difficult and GSEA time consuming. Therefore, the list of DEGs was sorted in descending order by variable importance according to a random forest model; the top 2000 genes were selected for analysis and some genes with importance close to zero were removed based on experience.

Although the selection of KEGG pathways in GSEA based on experience is subjective, it does not prevent obvious DEGs with no important biological significance for breast cancer being filtered out. In addition, these genes may also enhance the performance of classification models and the selection of important genes would be compromised. However, redundant genes were filtered out during processing with the GSEA–SDBE method.

To eliminate redundant genes, the SDBE algorithm was applied. This algorithm computed the difference in performance metrics of the classification model before and after gene deletion during backward elimination, which indicated the degree of redundancy of the deleted gene on the remaining genes. When a gene was deleted from the gene list in this manner, the performance metrics of the classification model did not change significantly. Therefore, the deleted gene was similar to some remaining genes, and thus considered redundant.

Given the underlying principle of the SDBE algorithm, the top gene in the gene list would not participate in the sorting process and would not be recognized as redundant; additionally, the first gene in a similar gene group in the gene list would not be recognized as redundant or deleted. Therefore, stage 1 of the SDBE algorithm is particularly important because genes are sorted by their importance in RF during backward elimination at this stage.

At stage 5 of the SDBE algorithm, to speed up the sorting process and reduce the number of cycles, the metric difference list was divided into two parts from a set position and these two parts were respectively sorted in descending order. The change of the set position occurred at stage 4. From the set position in the metric difference list, a forward search was conducted until an element with a value less than the threshold, which was set at zero, was encountered; the index of this element was used to update the set position. If the threshold was set to a certain value greater than zero, this may be more conducive to sorting. However, from the 19 iterations shown Figures 2 and 3, the polylines of the performance metrics for the classification models, particularly RF with MCC as the object of difference analysis, met the requirements. Including many more iterations would have been more time consuming. However, setting ROC_AUC_score as the object of difference analysis was less effective compared with using MCC, which might be related to the complexity of the ROC_AUC_score formula.

References

[1] Alelyani S, Zhao Z, Liu H (2011) A dilemma in assessing stability of feature selection algorithms. IEEE International Conference onHPCC, pp 701–707

[2] Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C,Rudan I, Campbell H, Wright AF, Wilson JF, Agakov F, Navarro P, Haley CS (2015) Application of high-dimensional feature selection Evaluation for genomic prediction in man. Sci Rep 5:1–12

[3] Bol˙on-Canedo V, S˙anchez-Marono N, Alonso-Betanzos A,Benitez JM, Herrera F (2014) A review of microarray datasets and applied feature selection methods. Inf Sci 282:111–135

[4] Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, K˙uffner R, Zimmer R (2006) Reliable gene signatures for microarray classification Assessment of stability and performance. Bioinformatics 22(19):2356–2363

[5] Dunne K, Cunningham P, Azuaje F (2002) Solutions to instability problems with sequential wrapper-based approaches to feature selection. J Mach Learn Res. http://citeseerx.ist.psu.edu/viewdoc/

summary? doi=10.1.1.11.4109
[6] Goh WWB, Wong L (2016) Evaluating feature-selection stability in next-generation proteomics. J Bioinform Comput Biol 14(05):1–23
[7] Guzm˙ an-Martinez R, Alaiz-Rodriguez R (2011) Feature selection stability assessment based on the Jensen-Shannon divergence
[8] Kalousis A, Prados J, Hilario M (2005) Stability of feature selection algorithms. In: Fifth IEEE international conference on data mining (ICDM'05), pp 8
[9] Kamkar I, Gupta SK, Phung D, Venkatesh S (2015) Stable feature selection with support vector machines. In: Australasian joint conference on artificial intelligence. Springer, Cham, pp 298–308
[10] Kr ¨ A±zek P (2016) Improving stability of feature selection methods, Caip 2009, pp 865–872
[11] Kuncheva LI (2007) A stability index for feature selection. In: 25Th international multi-conference: artificial intelligence and applications. ACTA Press, pp 390–395
[12] Lausser L, M˙ ussel C, Maucher M, Kestler HA (2013) Measuring and visualizing the stability of biomarker selection techniques. Comput Stat 28(1):51–65
[13] Lustgarten JL, Gopalakrishnan V, Visweswaran S (2009) Mea-suring stability of feature selection in biomedical datasets. In American Medical Informatics Association Symposium. Ameri-can Medical Informatics Association, pp 406–410
[14] Nogueira S, Sechidis K, Brown G (2017) On the stability of feature selection algorithms. J Mach Learn Res 18(1):6345–6398
[15] Osanaiye O, Cai H, Choo KKR, Dehghantanha A, Xu Z, Dlodlo M (2016) Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. Eurasip Journal on Wireless Communications and Networking 2016(1)
[16] Sarah Nogueira B, Brown G (2016) Machine learning and knowledge discovery in databases. In: European conference on machine learning and principles and practice of knowledge discovery in databases, pp 442–457
[17] Sehhati M, Mehridehnavi A, Rabbani H, Pourhossein M (2015) Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information. IEEE/ACM Trans Comput Biol Bioinform 12(6):1440–1448
[18] Somol P, Novovi˘ cová J (2010) Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Trans Pattern Anal Mach Intell 32(11):1921–1939
[19] Turney P (1995) Technical Note: Bias and the quantification of stability. Mach Learn 20:23–33
[20] WaldR, Khoshgoftaar TM, NapolitanoA(2013)Stabilityoffilter-and Wrapper-Based feature subset selection. In: 25th international conference on tools with artificial intelligence. IEEE, pp 374–380
[21] Yu L, Ding C, Loscalzo S, Stable feature selection via dense feature groups. In: 14Th ACM SIGKDD International conference on Knowledge discovery and data mining - KDD 08. ACM Press NewYork pp 803–811 (2008)
[22] Zarkoob H, Mehrdad J (2015) Gangeh, and ali ghodsi. Fast and scalable feature selection for gene expression data using Hilbert-Schmidt independence criterion. IEEE Trans Comput Biol Bioinform 14(1):167–181
[23] ZhangM, ZhangL, ZouJ, YaoC, XiaoH,LiuQ,WangJ,WangD,Wang C, Guo Z (2009) Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinform 25(13):1662–1668
[24] Zhou DX (2013) On grouping effect of elastic net. Stat Probab Lett 83(9):2108–2112
[25] Zucknick M, Richardson Sa, Stronach EA (2008) Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. Statistical Applications in Genetics and Molecular Biology 7(1): Article7
[26] Merrill RM, Sloan A, Anderson AE, Ryker K. Unstaged cancer in the United States: a population-based study. BMC Cancer. 2011; 11:402.
[27] Roberts JC, Li G, Reitzel LR, Wei Q, Sturgis EM. No evidence of sex-related survival disparities among head and neck cancer patients receiving similar multidisciplinary care: a matched-pair analysis. Clin Cancer Res. 2010; 16:5019-2027.
[28] Adekolujo OS, Tadisina S, Koduru U, Gernand J, Smith SJ, Kakarala RR. Impact of marital status on tumor stage at diagnosis and on sur-vival in male breast cancer. Am J Mens Health. 2017; 11:1190-1199.
[29] Osborne C, Ostir GV, Du X, Peek MK, Goodwin JS. The influ-ence of marital status on the stage at diagnosis, treatment, and sur-vival of older women with breast cancer. Breast Cancer Res Treat. 2005; 93:41-47.

[30] Rendall MS, Weden MM, Favreault MM, Waldron H. The protec-tive effect of marriage for survival: a review and update. Demography. 2011; 48:481-506.
[31] Parker L, Levin DC, Frangos A, Rao VM. Geographic variation in the utilization of noninvasive diagnostic imaging: national Medicare data, 1998-2007. AJR Am J Roentgenol. 2010; 194:1034-1039.
[32] Patel MK, Cote ML, Ali-Fehmi R, Buekers T, Munkarah AR, Elshaikh MA. Trends in the utilization of adjuvant vaginal cuff brachytherapy and/or external beam radiation treatment in stage I and II endometrial cancer: A Surveillance, Epidemiology, and End-Results Study. Int J Radiat Oncol Biol Phys. 2012; 83:178-184.
[33] Zhou AH, Chung SY, Patel VR, et al. Do geographic differences or socioeconomic disparities affect survival in sinonasal squamous cell carcinoma? Int Forum Allergy Rhinol. 2017; 7:1195-1200.
[34] Vyas A, Madhavan SS, Sambamoorthi U, et al. Healthcare utilization and costs during the initial phase of care among elderly women with breast cancer. J Natl Compr Canc Netw. 2017; 15:1401-1409.
[35] Quaglia A, Tavilla A, Shack L, et al; EUROCARE Working Group. The cancer survival gap between elderly and middle-aged patients in Europe is widening. Eur J Cancer. 2009; 45:1006-1016.
[36] De Angelis R, Sant M, Coleman MP, et al; EUROCARE-5 Working Group. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE-5—a population-based study. Lancet Oncol. 2014; 15:23-34.

Tables and Figures

Table 1. **Gene sets (pathways) that were strongly related to breast cancer.**

GENE SET NAME	ES	NES	NOM P value	FDR Q value	Gene number (core enrichment)
KEGG_CELL_CYCLE	0.60	1.37	0.201	0.319	20
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	−0.29	−0.96	0.496	0.726	17
KEGG_JAK_STAT_SIGNALING_PATHWAY	−0.48	−1.34	0.143	1.000	11
KEGG_PATHWAYS_IN_CANCER	−0.23	−0.84	0.720	0.790	24

Table 2. **MCC as the object of difference analysis: 10-fold cross-validation classification metrics (%) of the top three genes.**

Modes	ROC_AUC_score	MCC	Recall	FPR	F1_score	Accuracy	Top three genes
RF	0.9875	0.9528	0.9928	0.0177	0.9955	0.9918	*VEGFD, TSLP, PKMYT1*
SVM	0.9684	0.8832	0.9810	0.0442	0.9882	0.9787	*VEGFD, PKMYT1, BUB1B**
XGBoost	0.9861	0.9396	0.9900	0.0177	0.9941	0.9893	*VEGFD, TSLP, PKMYT1*
KNN	0.9653	0.8897	0.9837	0.0531	0.9891	0.9803	*VEGFD, TSLP, PKMYT1*
ExtraTrees	0.9818	0.9345	0.9900	0.0265	0.9937	0.9885	*VEGFD, TSLP, PKMYT1*

Genes marked with * are unstable genes in the SDBE algorithm.

Table 3. **ROC_AUC_score as the object of difference analysis: 10-fold cross-validation classification metrics (%) of the top three genes.**

Modes	ROC_AUC_score	MCC	Recall	FPR	F1_score	Accuracy	Top three genes
RF	0.9799	0.8840	0.9774	0.0177	0.9877	0.9779	*VEGFD, SPRY2, BUB1B**
SVM	0.9828	0.8501	0.9657	0.0	0.9825	0.9689	*VEGFD, CCNB1*, TSLP**
XGBoost	0.9812	0.8952	0.9801	0.0177	0.9890	0.9803	*VEGFD, CCL14, TSLP*
KNN	0.9771	0.8627	0.9720	0.0177	0.9849	0.9710	*VEGFD, TSLP, CCL14*
ExtraTrees	0.9809	0.9260	0.9883	0.0265	0.9927	0.9869	*VEGFD, TSLP, CDC25C*

Genes marked with * are unstable genes in the SDBE algorithm.

Table 4. Results of survival analysis for high-risk and low-risk groups according to three genes.

Gene Name	Expression in tumor	P value	High risk			Low risk		
			SP (5 y)	M-OS [95% CI]	n	SP (5 y)	M-OS [95% CI]	n
VEGFD	Downregulated	0.0466	0.8088	129 [114–142]	846	0.8552	149 [122–inf]	262
TSLP	Downregulated	0.0003	0.7896	116 [102–132]	786	0.8837	248 [122–inf]	322
PKMYT1	Upregulated	0.2095	0.7743	149 [102–inf]	419	0.8494	131 [115–215]	689

P value: comparison between high risk and low risk; Inf: data points not obtained; SP (5 y): 5-year survival probability; M-OS (95% CI): median overall survival time in months with 95% confidence intervals.

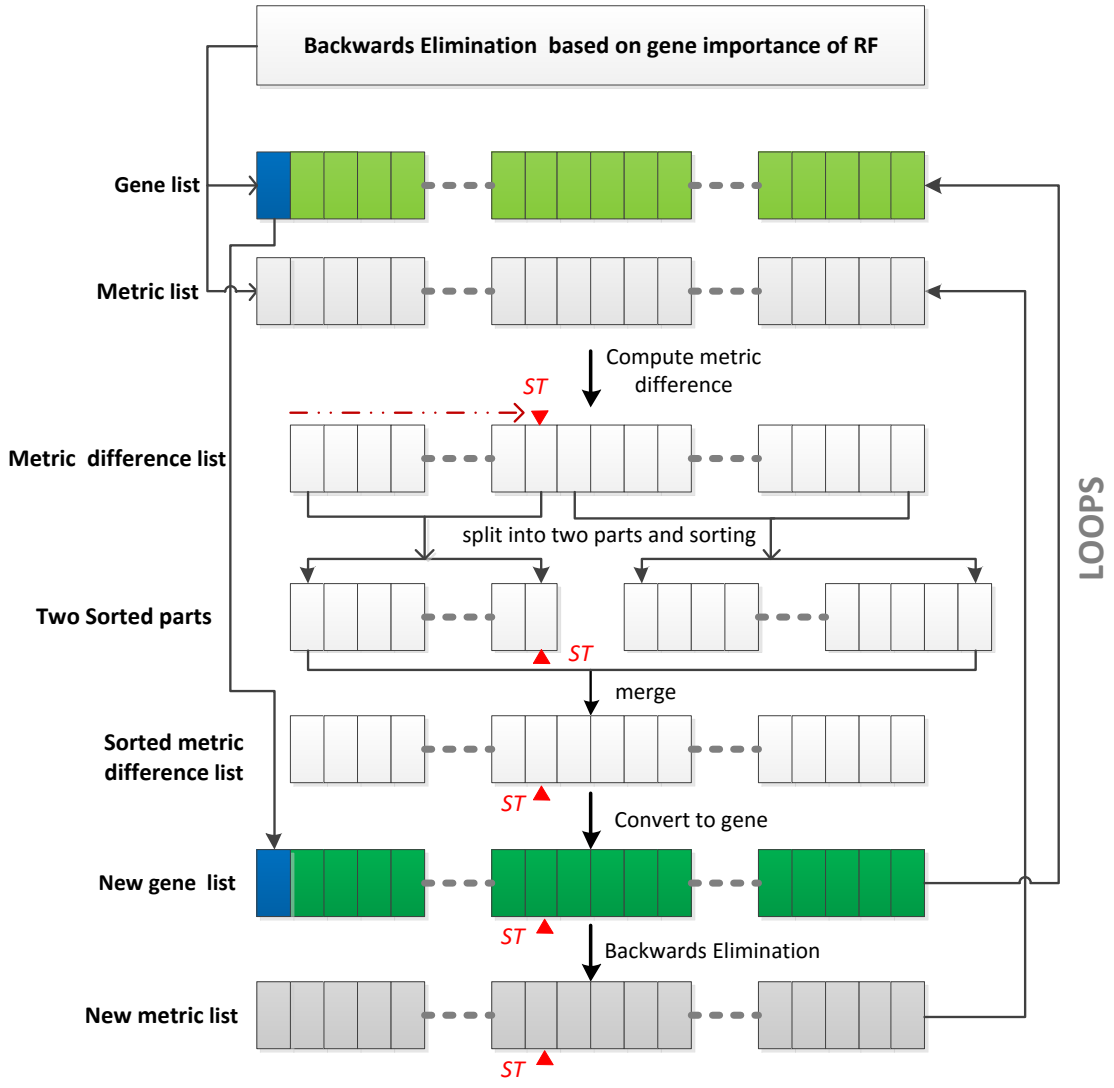


Figure 1. **Procedure of the Sort Difference Backward Elimination (SDBE) algorithm.**

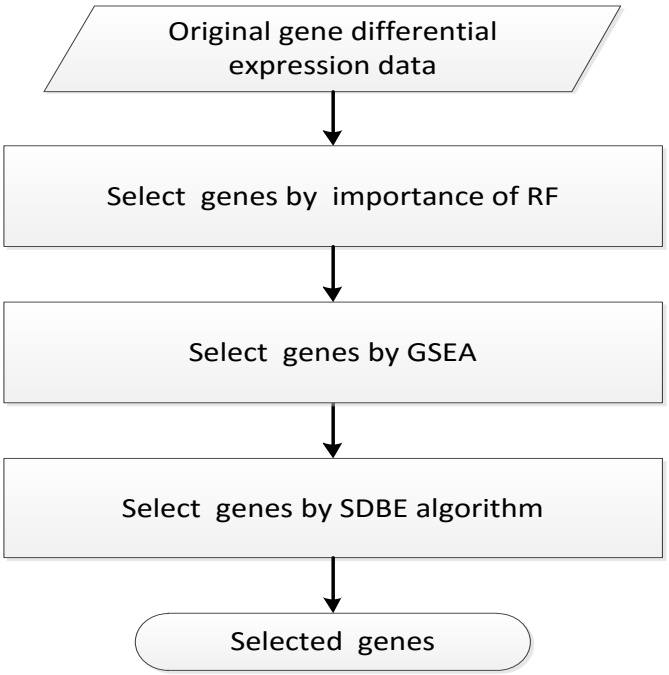


Figure 2. **Gene selection procedure in the GSEA–SDBE method.**

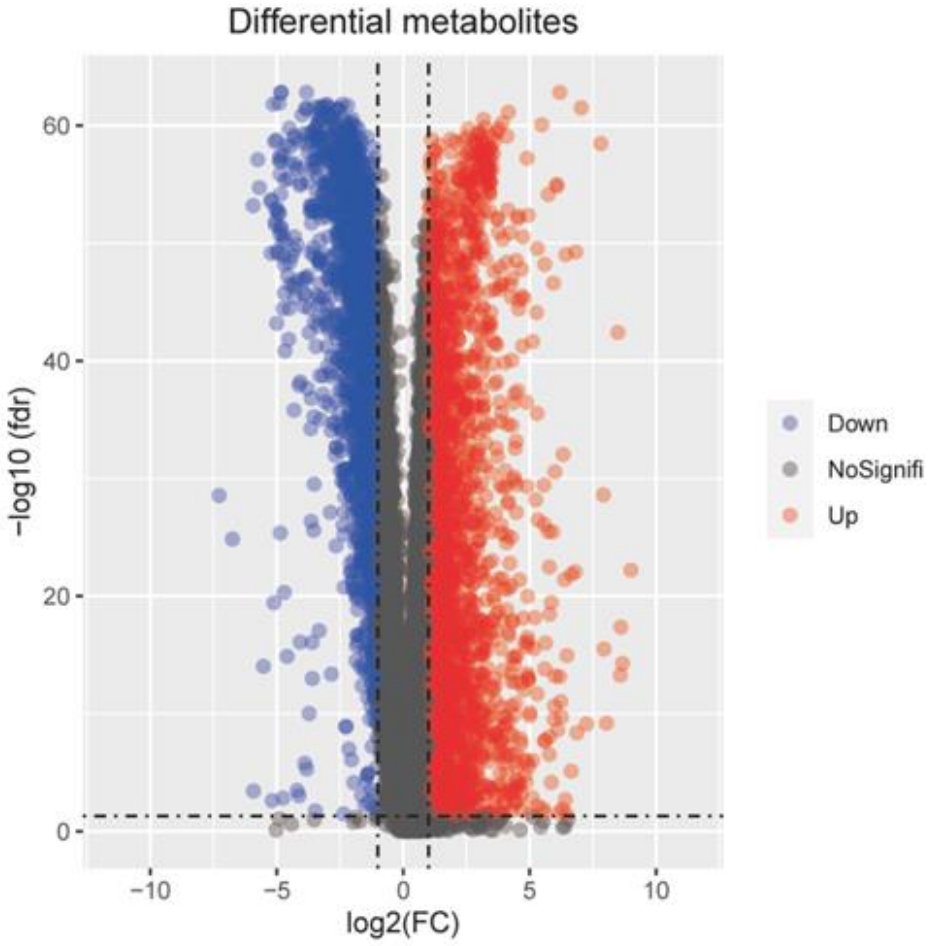


Figure 3. **Volcano plot of differentially expressed genes. The red and blue dots represent upregulated and downregulated genes, respectively.**

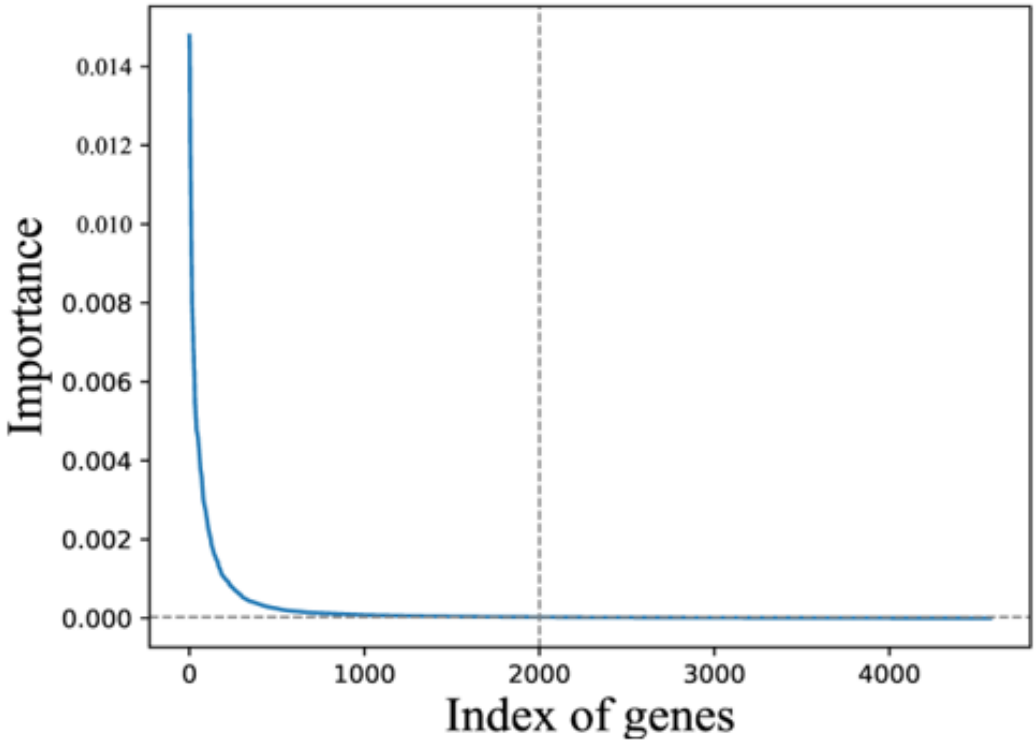


Figure 4. **Genes sorted by importance in descending order.**

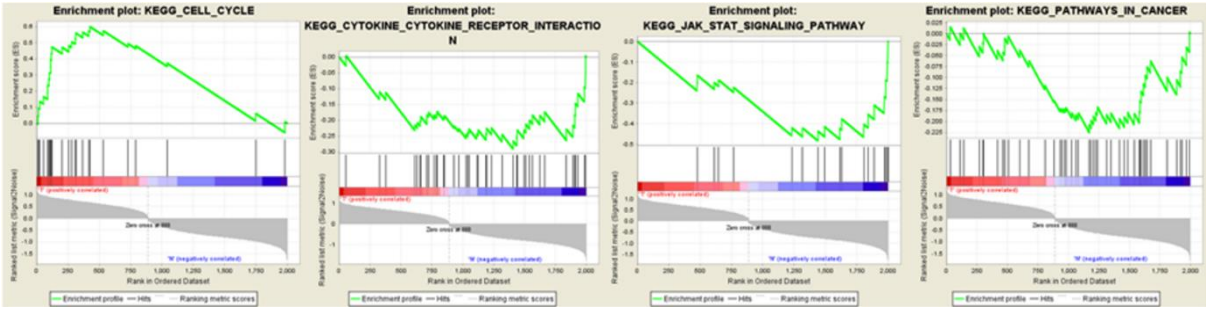
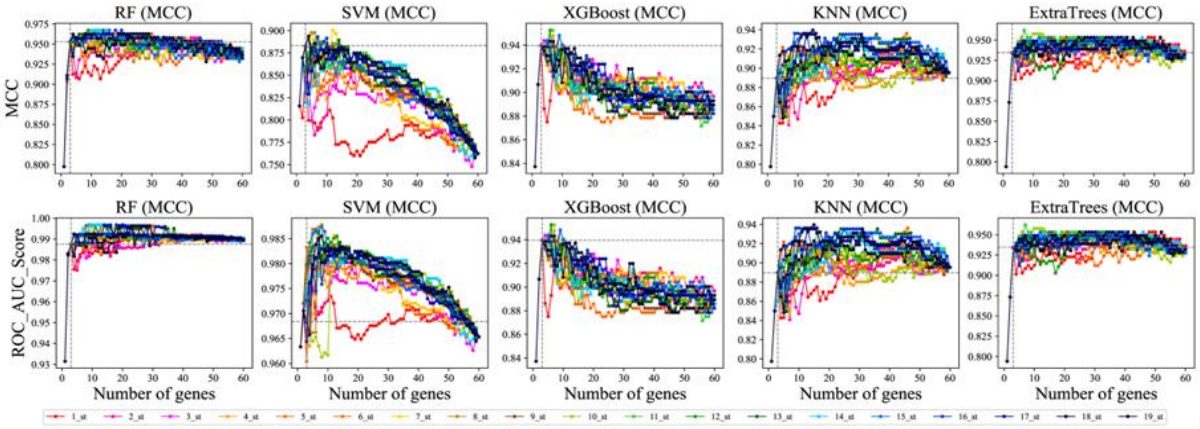
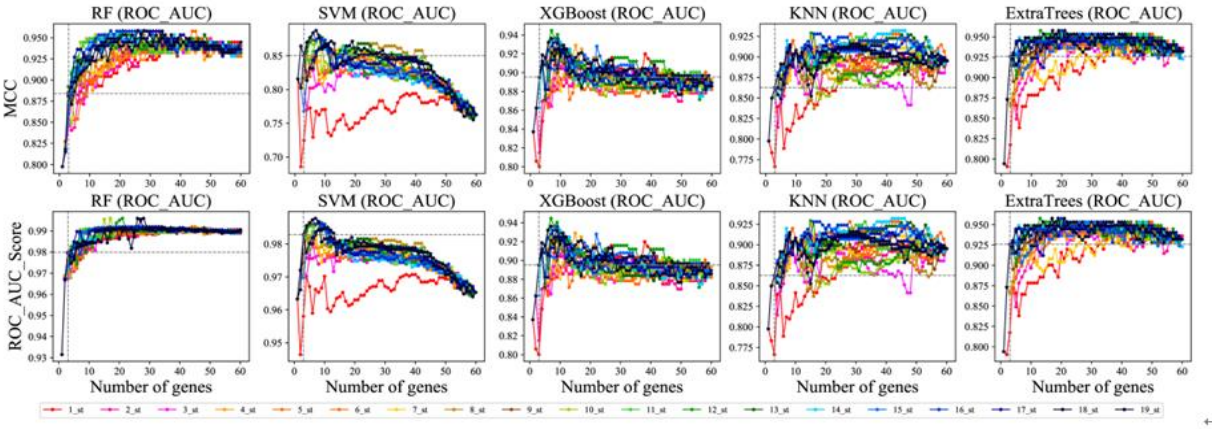


Figure 5. **Enrichment plots for the four gene sets (pathways) that were strongly related to breast cancer.**

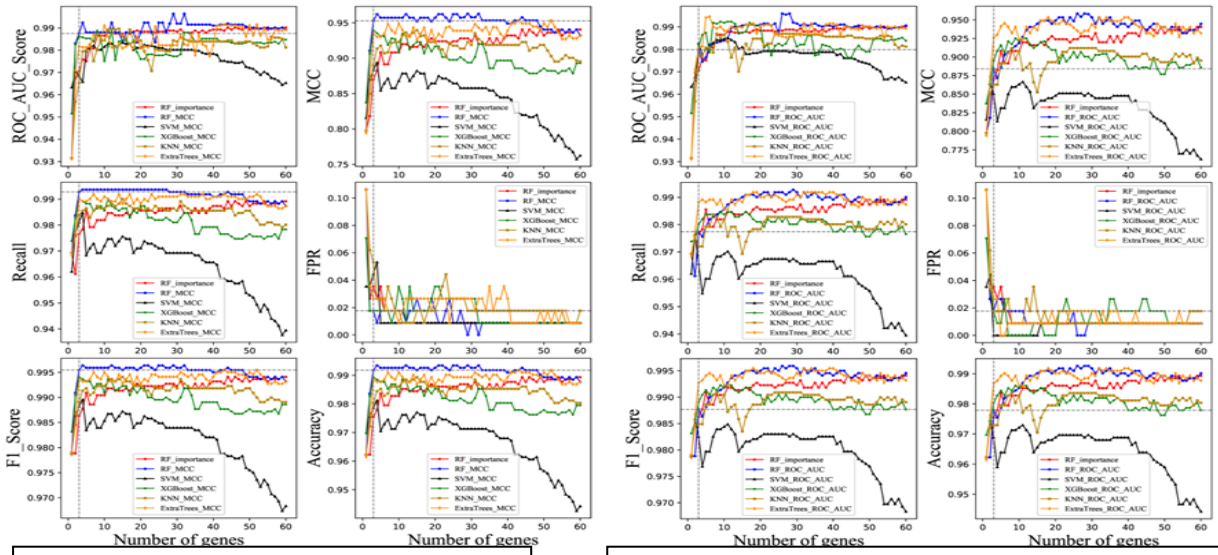


(a) MCC as the object of difference analysis.



(b) ROC_AUC_score as the object of difference analysis.

Figure 6. **Polylines of classification metrics, MCC, and ROC_AUC_score in 19 iterations.**



(a) MCC as the object of difference analysis

(b) ROC_AUC_score as the object of difference analysis.

Figure 7. **Polylines of classification metrics at the 19th iteration of the Sort Difference Backward Elimination (SDBE) algorithm.**

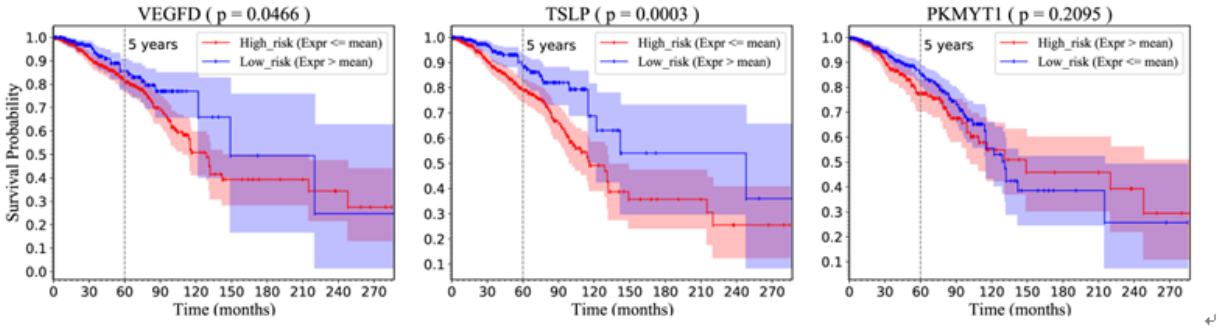


Figure 8. **Kaplan–Meier survival graphs for expression of *VEGFD* (a), *TSLP* (b), and *PKMYT1* (c). Red and blue curves denote high-risk and low-risk groups, respectively.**