Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Machine learning-based prediction model for distant metastasis of breast cancer

Hao Duan [a,1], Yu Zhang [b,1], Haoye Qiu [a], Xiuhao Fu [a], Chunling Liu [a], Xiaofeng Zang [a], Anqi Xu [c], Ziyue Wu [a], Xingfeng Li [a], Qingchen Zhang [a], Zilong Zhang [a,**], Feifei Cui [a,*]

[a] *School of Computer Science and Technology, Hainan University, Haikou, 570228, China*
[b] *Beidahuang Industry Group General Hospital, Harbin, 150001, China*
[c] *The First School of Clinical Medicine, Shandong University of Traditional Chinese Medicine, Jinan, 250014, China*

ARTICLE INFO

Keywords:
Breast cancer distant metastasis
Predictive model
Biomarkers
Machine learning
Weighted correlation network analysis

ABSTRACT

Background: Breast cancer is the most prevalent malignancy in women. Advanced breast cancer can develop distant metastases, posing a severe threat to the life of patients. Because the clinical warning signs of distant metastasis are manifested in the late stage of the disease, there is a need for better methods of predicting metastasis.

Methods: First, we screened breast cancer distant metastasis target genes by performing difference analysis and weighted gene co-expression network analysis (WGCNA) on the selected datasets, and performed analyses such as GO enrichment analysis on these target genes. Secondly, we screened breast cancer distant metastasis target genes by LASSO regression analysis and performed correlation analysis and other analyses on these biomarkers. Finally, we constructed several breast cancer distant metastasis prediction models based on Logistic Regression (LR) model, Random Forest (RF) model, Support Vector Machine (SVM) model, Gradient Boosting Decision Tree (GBDT) model and eXtreme Gradient Boosting (XGBoost) model, and selected the optimal model from them.

Results: Several 21-gene breast cancer distant metastasis prediction models were constructed, with the best performance of the model constructed based on the random forest model. This model accurately predicted the emergence of distant metastases from breast cancer, with an accuracy of 93.6 %, an F1-score of 88.9 % and an AUC value of 91.3 % on the validation set.

Conclusion: Our findings have the potential to be translated into a point-of-care prognostic analysis to reduce breast cancer mortality.

1. Introduction

Breast cancer is caused by the uncontrolled proliferation of breast epithelial cells in response to various carcinogenic factors [1]. The early symptoms of breast cancer are usually the appearance of breast lumps etc. In advanced stages of breast cancer, distant metastasis of cancer cells can occur [2,3], leading to multiorgan lesions and threatening the lives of sufferers with breast cancer. Depending on the statistics of the International Cancer Organization, breast cancer has the highest morbidity of malignant tumors in women [4,5]. Although the overall treatment outcome of breast cancer sufferers is better with the improvement of the medical care level, the treatment outcome for breast cancer patients who develop distant metastasis remains unsatisfactory.

In addition to the level of medical care and treatment options, the timing of treatment for breast cancer with distant metastases is also an important factor affecting its treatment outcome. At present, the diagnosis of distant metastases in sufferers with breast cancer is still mostly based on the observation of clinical symptoms in the corresponding organs caused by metastases and the presence of imaging changes in certain areas. However, the above diagnostic methods cannot detect distant metastasis of breast cancer at an early stage and cannot provide better treatment timing and more treatment time for sufferers with deteriorating breast cancer.

In recent years, next-generation transcriptome sequencing

* Corresponding author.
** Corresponding author.
E-mail addresses: zhangzilong@hainanu.edu.cn (Z. Zhang), feifeicui@hainanu.edu.cn (F. Cui).
[1] The authors contributed equally.

https://doi.org/10.1016/j.compbiomed.2024.107943
Received 11 November 2023; Received in revised form 10 December 2023; Accepted 1 January 2024
Available online 6 January 2024

technologies have been rapidly developed, which can obtain the expression information of transcripts of specific cells or tissues in a certain state, and the analysis of transcriptome sequencing data can provide a theoretical basis for the diagnosis and prognosis of diseases such as cancer, and the massive sequencing data have opened up new possibilities for proposing better diagnostic and therapeutic tools for breast cancer [6–9]. At the same time, bioinformatics analysis of transcriptome data using machine learning methods has achieved excellent results [10–15].

Therefore, in this study, we identified potential biomarkers of distant metastasis in breast cancer by combining difference analysis, WGCNA [16,17] and LASSO regression analysis [18–21]. Based on these biomarkers, we trained several breast cancer distant metastasis prediction models based on logistic regression models [22], random forest models [23,24], SVM models [25,26], GBDT models [27] and XGBoost models [28] on the training set divided by the GSE9893 dataset to predict the presence of distant metastasis in breast cancer patients and selected the best prediction model from these prediction models as our predictive model for distant metastasis. Validation showed that our 21-gene predictive model for distant metastasis performed well and was able to predict the presence of distant breast cancer metastasis in breast cancer sufferers with excellent precision and effectiveness. In addition, we also performed functional analysis of some important genes and the obtained biomarkers of breast cancer distant metastasis, which can help us further understand the mechanism and factors affecting breast cancer distant metastasis.

2. Materials and methods

In this research, biomarkers of distant breast cancer metastasis were identified by combining differential analysis, WGCNA and LASSO regression analysis, and the expression data of these biomarkers were used to train a predictive model of distant breast cancer metastasis. The overall flow of this study is shown in Fig. 1. In the first step, we obtained the differentially expressed genes of breast cancer distant metastasis by performing differential analysis on the selected dataset; in the second step, we obtained the significant module gene set by performing WGCNA on the selected dataset; in the third step, the differential genes and significant module genes were intersected to obtain the breast cancer distant metastasis target genes; in the fourth step, the breast cancer distant metastasis target genes were analyzed by PPI, GO and KEGG, etc.; in the fifth step, LASSO regression analysis of breast cancer distant metastasis target genes to obtain distant metastasis of breast cancer biomarkers; in step six, correlation, expression characteristics and regulatory mechanism analysis of distant metastasis of breast cancer biomarkers; in step seven, extraction of expression data of distant metastasis of breast cancer biomarkers; in step eight, multiple breast cancer distant metastasis prediction models are constructed using expression data of breast cancer distant metastasis biomarkers based on logistic regression models, random forest models and other models; in the ninth step, the best performing breast cancer distant metastasis prediction model was selected based on the evaluation index.

2.1. Data preparation

Gene expression files for breast cancer were gathered from the GSE9893 dataset and the GSE43837 dataset in the Gene Expression Omnibus (GEO), both of which were derived from tumor tissues of breast cancer patients. Expression data from all samples in the GSE9893 dataset and GSE43837 dataset were retained. The GSE9893 dataset includes expression data from 155 breast cancer patients, of whom 48 developed distant metastasis and 107 did not. The GSE43837 dataset includes expression data from 38 breast cancer patients, of whom 19 developed distant metastasis and 19 did not.

2.2. Screening for differentially expressed genes in patients with distant metastases from breast cancer

Gene expression data in the GSE9893 and GSE43827 datasets were analyzed separately using “limma" in the R (4.1.0) package, and $|\text{logFC}| > 0.5$ and P-value <0.05 were used as selection criteria for differential genes in the breast cancer distant metastasis group and breast cancer non-distant metastasis control group in the GSE9893 and GSE43837 datasets, respectively.

2.3. Target gene screening based on weighted gene coexpression network analysis

First, all samples of GSE9893 were clustered, outliers were removed, and the sample clustering tree was reconstructed. Then, the trends of the scale-free topological fit index and average connectivity with the change of the “soft" threshold (β) are analyzed, and the “soft" threshold that makes the scale-free topological fit index >0.9 and the trend of the change of the average connectivity starts to level off is selected as the minimum “soft" threshold, and the hierarchical clustering dendrogram of the module identifiers is constructed based on the selected minimum “soft" threshold. Module membership (MM) can show the correlation between the expression values of the module genes and the module signature genes. Gene significance (GS) can show the correlation between module genes and samples. Based on the calculated MM and GS values, modules can be correlated with clinical features, and the genes of the modules with higher associations with clinical features are considered as important module genes. Finally, we took the intersection of the two differentially expressed gene sets with the three significant module genes derived from WGCNA to obtain potential breast cancer distant metastasis target genes.

2.4. Enrichment analysis of target gene

Gene Ontology (GO) analysis of breast cancer distant metastasis target genes was performed by setting the filtering criteria for functional analysis to a P-value <0.3 using the “clusterProfiler" software package, yielding molecular functions (MFs), biological processes (BPs) and cellular components (CCs). The filtering criterion for functional analysis using the “clusterPro filer" package was set to P-value <0.15, and enriched signalling pathways for potential distal breast cancer target genes were identified by performing Kyoto Encyclopaedia of Genes and Genomes (KEGG) functional enrichment analysis.

2.5. Protein–protein interaction network construction

Interaction networks of potential breast cancer distant metastasis target genes were constructed using the STRING database and the results were viewed using Cytoscape software which distinguish genes that interact more with other genes.

2.6. Biomarker screening for patients with distant metastases from breast cancer based on LASSO analysis

LASSO regression analysis of potential breast cancer distant metastasis target genes was performed using the “glmnet" software package. The best performing λ value in the LASSO model, i.e., the one with the smallest mean square error, was selected, and the genes with high predictive power corresponding to this λ value were screened. The genes obtained from their analysis were considered potential biomarkers for distant metastasis of breast cancer.

2.7. Correlation analysis of biomarkers for distant metastasis of breast cancer

Correlation analysis of biomarkers expression in distant breast

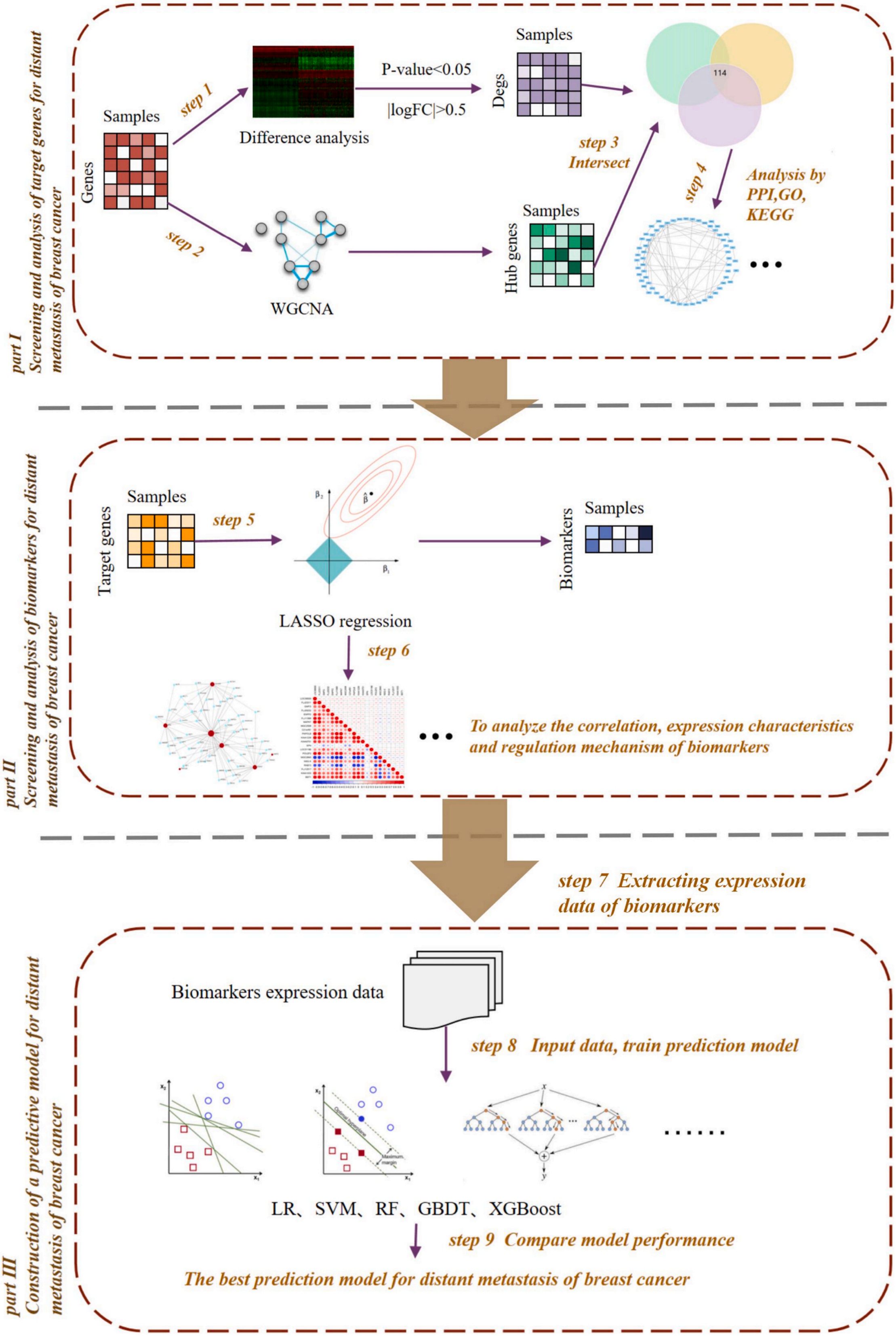
part I
Screening and analysis of target genes for distant metastasis of breast cancer
Samples
Genes
step 1
Difference analysis
P-value<0.05
|logFC|>0.5
Samples
Degs
114
step 2
WGCNA
Samples
Hub genes
step 3
Intersect
step 4
Analysis by PPI,GO, KEGG
part II
Screening and analysis of biomarkers for distant metastasis of breast cancer
Samples
Target genes
step 5
LASSO regression
Samples
Biomarkers
step 6
To analyze the correlation, expression characteristics and regulation mechanism of biomarkers
step 7 Extracting expression data of biomarkers
part III
Construction of a predictive model for distant metastasis of breast cancer
Biomarkers expression data
step 8 Input data, train prediction model
LR、SVM、RF、GBDT、XGBoost
step 9 Compare model performance
The best prediction model for distant metastasis of breast cancer

Fig. 1. The overall research process of the breast cancer distant metastasis prediction model. As shown in the figure, the overall research process was divided into three parts, firstly, the target genes of breast cancer distant metastasis were derived through differential analysis and weighted gene co-expression network analysis, secondly, the biomarkers of breast cancer distant metastasis were further derived through LASSO analysis and correlation analysis was carried out, and lastly, the predictive model of breast cancer distant metastasis was constructed based on the machine learning model and the resulting biomarkers, and its predictive effect was verified.

cancer metastases was carried out using the “corrplot" software package to further understand the interactions between the obtained biomarkers.

2.8. Expression of biomarkers for distant metastasis of breast cancer

The expression levels of potential biomarkers for distant metastasis of breast cancer were analyzed using *t*-test to derive the different expression levels of each biomarker in non-distant metastasis samples and distant metastasis samples of breast cancer in the GSE9893 dataset to further understand the approximate role played by each biomarker in distant metastasis of breast cancer.

2.9. Regulatory mechanisms of distant metastasis biomarkers in breast cancer

Based on the NetworkAnalyst platform, the transcription factors (TFs) associated with potential biomarkers of distant metastasis of breast cancer were enriched and analyzed, and a TF-biomarker network was constructed to further investigate their mechanisms of action in distant metastasis of breast cancer. The miRNA-biomarker network was constructed based on the NetworkAnalyst platform to predict the miRNAs associated with the regulation of potential biomarkers of distant metastasis of breast cancer.

2.10. Construction of a forecasting model for distant metastasis from breast cancer

We used potential biomarkers for distant breast cancer metastasis as features, and based on a grid optimization method and fivefold cross-validation on 80 % of the data randomly divided from the GSE9893 dataset, i.e., the training set, we performed hyperparameter combinations in the parameter space of each model to find the hyperparameters of the RF model, LR model, SVM model, GBDT model, XGboost model classification training and hyperparameter optimization to construct multiple breast cancer distant metastasis prediction models. Furthermore, the same validation dataset (the remaining 20 % of the GSE9893 dataset is the validation set) is used to compare several breast cancer metastasis prediction models using the same evaluation metrics, i.e., accuracy, F1-Score, and AUC, and the best performing model is selected as our breast cancer distant metastasis prediction model.

3. Results

3.1. Screening of differentially expressed genes in patients with distant metastases from breast cancer

After preprocessing the GSE9893 dataset and the GSE43827 dataset, differential analysis was carried out to initially screen differential genes for distant metastasis in breast cancer. A total of 6188 differential genes were screened from the GSE9893 dataset, with 3296 up-regulated genes and 2892 down-regulated genes (Fig. 2a). From the GSE43827 dataset, 2122 differential genes were identified, with 558 up-regulated and 1564 down-regulated genes (Fig. 2b).

3.2. Target gene screening based on weighted gene coexpression network analysis

After detecting outliers during the clustering of the GSE9893 dataset samples, cutHeight was set to 150 to remove 33 outlier samples (Fig. 2c). We then set the scale-free fit index to 0.9, resulting in a minimum ‘soft’ threshold of 6 for the construction of the scale-free network (Fig. 2d and e). Additionally, the minimum number of genes in the modules was set to 30, and the module fusion threshold (mergeCutHeight) was set to 0.5 to construct a hierarchical clustering dendrogram, which included 13 modules (Fig. 2f). The correlation coefficients of each module in the hierarchical clustering dendrogram with breast cancer metastasis were

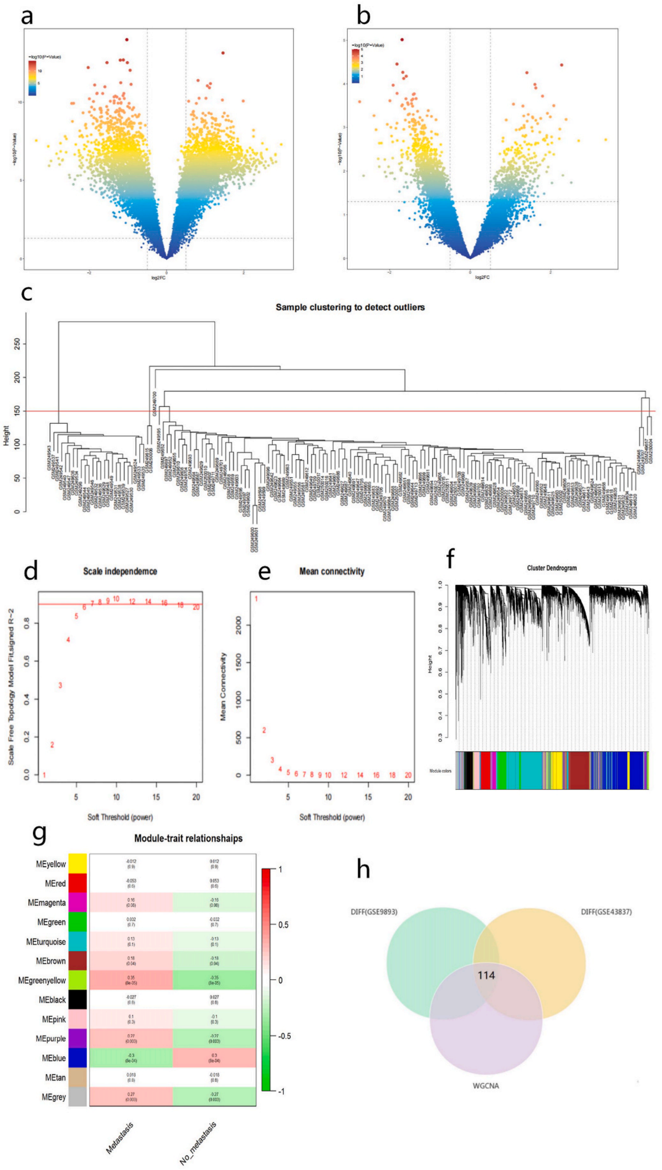

Fig. 2. Screening of target genes for distant metastasis of breast cancer. **a** The Volcano plot shows the expression signature of DEGs in dataset GSE9893, where the dots in the upper left represent down-regulated genes and the dots in the upper right represent up-regulated genes.**b** The Volcano plot shows the expression signature of DEGs in dataset GSE43827, where the upper left points represent down-regulated genes and the upper right points represent up-regulated genes. **c** Sample clustering dendrogram with cutHeight set to remove outliers. **d** The figure represents the scale-free fit index as a function of the soft threshold, with the x-axis being the soft threshold and the y-axis being the scale-free fit index. **e** The figure represents the mean connectivity as a function of the soft threshold, with the x-axis being the soft threshold and the y-axis representing the mean connectivity. **f** The figure shows a gene tree diagram of the clustering results, with the coloured rows below the tree diagram indicating the module assignment identified. **g** The figure shows the correlation between modules and functions, with the rows in the figure corresponding to MEs and the columns in the figure corresponding to clinical features. The figure contains correlation coefficients and p-values. **h** The graph represents the intersection of DEGs and target genes obtained by WGCNA.

obtained by correlating these modules with clinical features, and found that the yellow−green and blue modules were more correlated with distant breast cancer metastasis(Fig. 2g). Finally, the yellow−green module genes and blue module genes (a total of 3404 genes) were crossed with the two resulting differentially expressed gene sets to obtain 114 potential target genes for distant metastasis of breast cancer (Fig. 2h).

3.3. Enrichment analysis of target gene

GO analysis showed that these target genes have multiple functions, such as phospholipase activity and modification of amino acid binding. Additionally, they are associated with many CCs, such as myofibroblasts, lymphocytes and leukocytes, and are involved in several BPs, such as regulation of actin cytoskeleton organization, negative regulation of apoptotic processes in endothelial and epithelial cells, and regulation of actin filament depolymerization (Fig. 3a).KEGG enrichment analysis showed that these target genes were mainly enriched through signaling pathways, for example, the mitogen-activated protein kinase (MAPK) signaling pathway, aldosterone synthesis and secretion, and gonadotropin-releasing hormone signaling pathway (Fig. 3b).

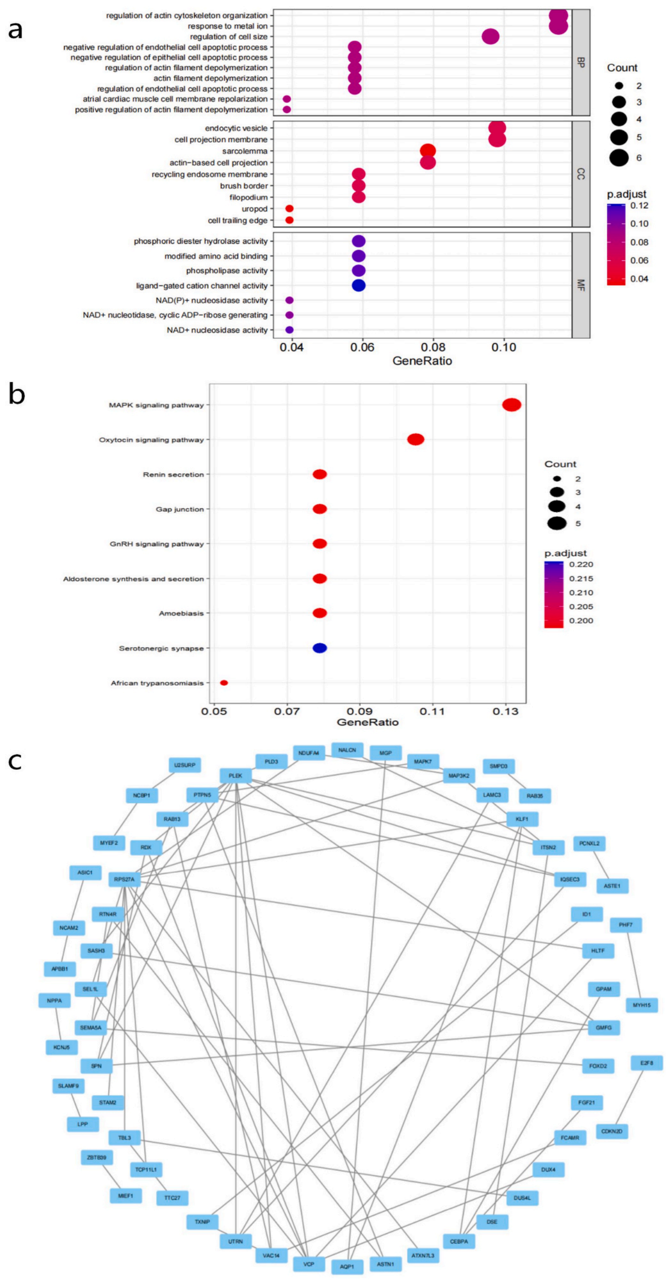

Fig. 3. Analysis of distant metastatic target genes in breast cancer. **a** The obtained GO enrichment analysis results of breast cancer distant metastasis target genes, from top to bottom, in terms of biological process, cellular composition and molecular function. **b** The obtained KEGG enrichment analysis results of distant metastasis target genes of breast cancer. **c** Protein interaction network of the obtained distant metastatic target genes of breast cancer.

3.4. Protein−protein interaction network construction

The interaction network of potential breast cancer distant metastasis target genes was constructed (Fig. 3c). The figure shows the PPI network derived after removing the noninterlinked genes from 114 key genes, in which PLEK, RPS27A, RTN4R and VCP interacted more with other genes.

3.5. LASSO-based analysis of biomarker screening in sufferers with distant metastases from breast cancer

In this study, LASSO analysis was used to further investigate potential biomarkers associated with distant breast cancer metastasis among 114 potential breast cancer distant metastasis target genes. LASSO regression model was constructed based on samples from breast cancer sufferers with distant metastasis and control samples, i.e., breast cancer sufferers without distant metastasis. The analysis showed that the LASSO model performed best when logλ was roughly −4.2, with the smallest mean square error (Fig. 4a and b). Therefore, the 21 genes corresponding to variables with coefficients not zero in the LASSO regression model were more predictive when the value of λ was taken such that logλ was approximately −4.2, and these 21 genes were further used as potential biomarkers for distant metastasis of breast cancer.

3.6. Correlation analysis of biomarkers of distant metastasis in breast cancer

Correlation analysis of the expression of distant metastasis biomarkers of breast cancer showed that the expression of three genes, SPN, MGC2840 and RAB13, were in negative correlation with the expression of other genes in the distant metastasis biomarkers of breast cancer in the samples of breast cancer sufferers. The expression of other genes in the distant metastasis biomarkers, except SPN, MGC2840 and RAB13, were basically positively correlated(Fig. 5a).

3.7. Expression characteristics of distant metastasis biomarkers of breast cancer

We further researched the effect of each potential biomarker in the distant metastasis of breast cancer and examined their expression profiles in samples from breast cancer sufferers. The expression status of SPN, RAB13 and MGC2840 were higher in the group of breast cancer sufferers with distant metastasis, while these three biomarkers were the same as those whose expression was negatively associated with the expression of other genes in the correlation analysis. In contrast, the expression levels of the remaining genes except for three genes, SPN, RAB13 and MGC2840, were higher in the group of breast cancer sufferers without distant metastasis (Fig. 5b).

3.8. Regulatory mechanisms of distant metastasis biomarkers in breast cancer

To investigate the regulatory mechanisms of distant metastasis biomarkers in breast cancer, we predicted their TFs and constructed a network consisting of TFs and biomarkers (Fig. 5c). SPN, RAB13 and C21orf91 may be regulated by KLF6, MLLT1, ZFX and ELF1. MAPK7 may be regulated by KLF7, DEK, PHF8 and SP1; GMFG may be adjusted by KLF8, KLF7 and CDC5L. In a subsequent step, a miRNA-biomarker network (Fig. 5d) was structured to obtain miRNAs, for example, hsa-miR-802, hsa-miR-103 and other miRNAs that may play key

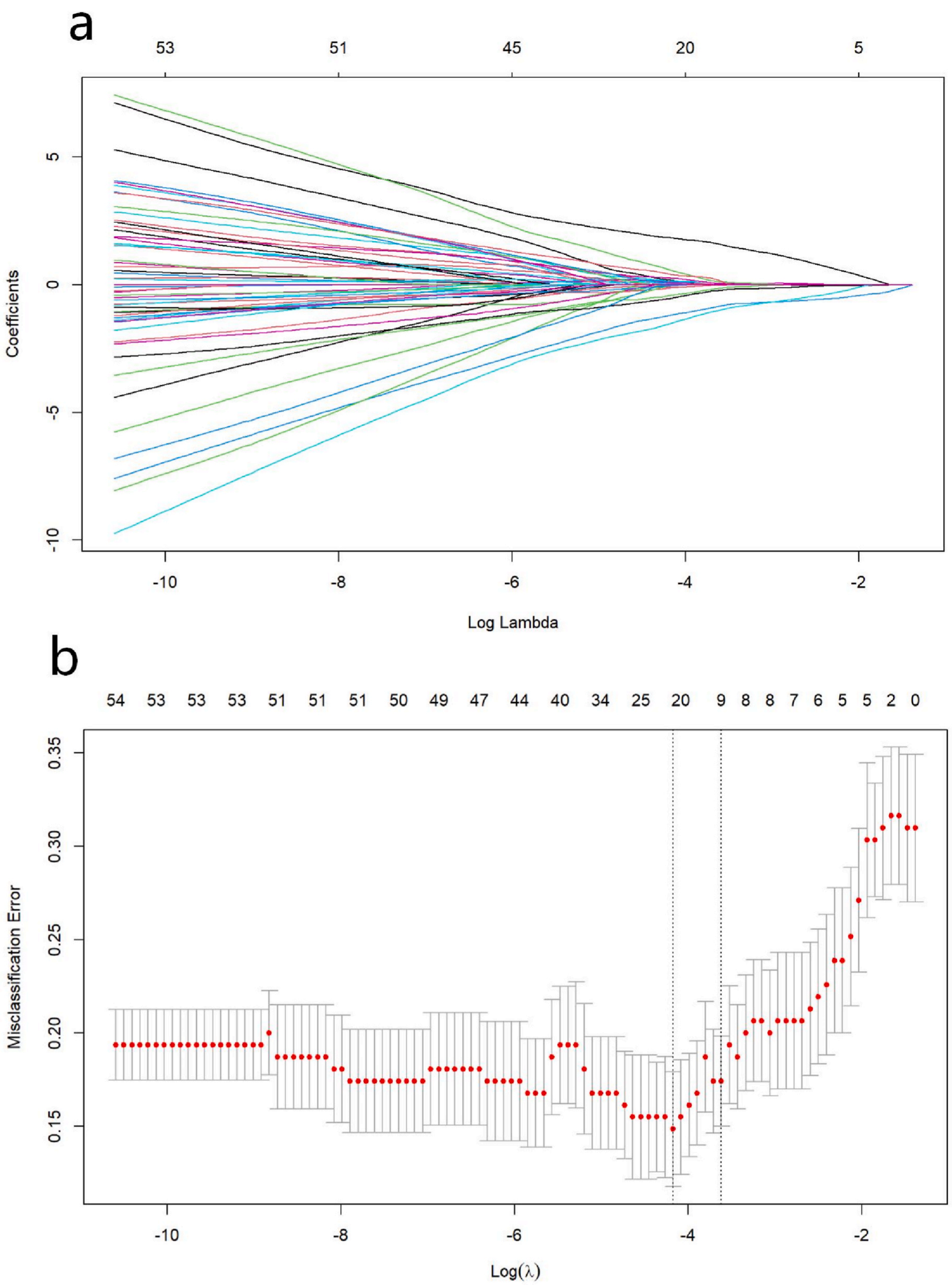

Fig. 4. Biomarker screening for distant metastases of breast cancer based on LASSO analysis. **a** Trend of absolute values of biomarker coefficients for distant metastases of breast cancer in the LASSO regression model with changes in λ values. **b** Trend of mean square error with λ for the LASSO regression model.

regulatory roles.

3.9. Construction of breast cancer distant metastasis prediction model

We constructed several breast cancer distant metastasis prediction models based on LR model, RF model, SVM model, GBDT model, and XGBoost model on the training set using potential biomarkers of breast cancer distant metastasis as features. The same validation set (the remaining 20 % of the GSE9893 dataset) was used to validate several breast cancer metastasis forecasting models, and the specific predictive

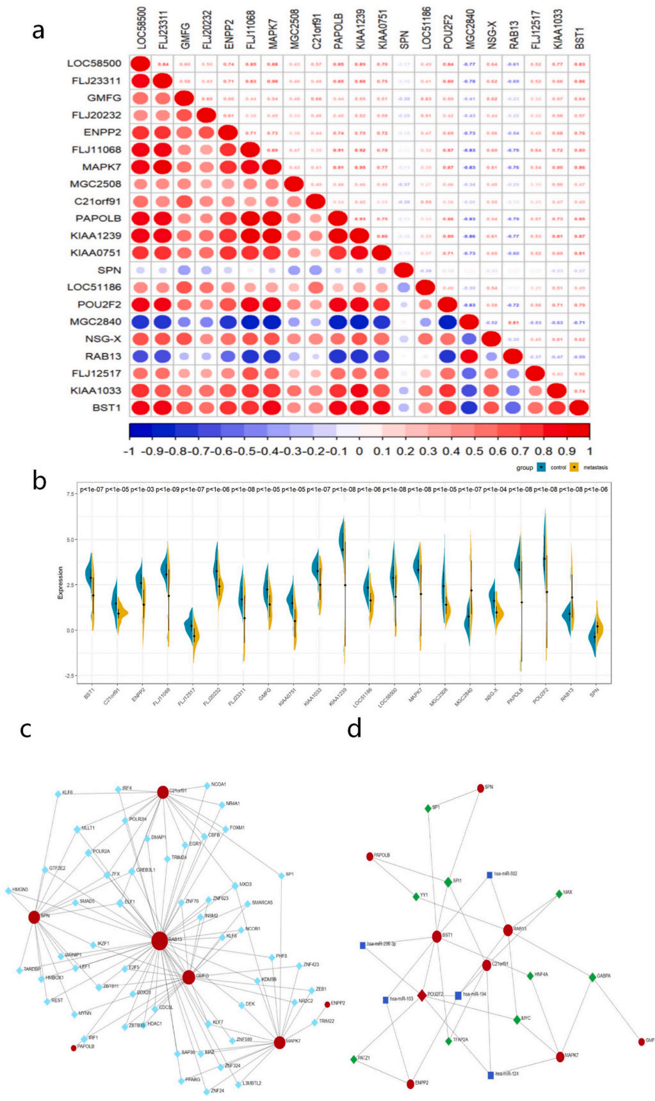

Fig. 5. Analysis of biomarkers for distant metastasis from breast cancer. **a** Correlation analysis between distant metastasis biomarkers of breast cancer. **b** Expression characteristics of distant metastasis biomarkers of breast cancer in the GSE9893 dataset. **c** TF-biomarker network, where red nodes indicate biomarkers and light blue nodes indicate TFs. **d** miRNA-biomarker network. In this network where red circular nodes represent biomarkers, green nodes and red square nodes represent TFs, and blue nodes represent miRNAs.

capabilities of several breast cancer distant metastasis prediction models based on different seed models was obtained from the validation set (Table 1). Based on the comparison of the Accuracy, F1-Score and AUC values of the different seed models in the validation set, we found that the random forest model-based breast cancer distant metastasis prediction model had the strongest prediction ability, with an accuracy of 93.6 %, F1-Score of 88.9 % and AUC of 91.3 % for the samples in the validation set. It was therefore used as our final predictive model for distant metastasis from breast cancer.

Table 1
Comparison of breast cancer distant metastasis prediction models on the validation set (20 % of the GSE9893 dataset that did not participate in model training).

Predictive models	Accuracy	F1-score	AUC
Model based on LR	0.894	0.815	0.863
Model based on RF	**0.936**	**0.889**	**0.913**
Model based on SVM	0.872	0.786	0.847
Model based on GBDT	0.915	0.857	0.898
Model based on XGBoost	0.872	0.769	0.827

4. Discussion

Breast cancer is the disease with the highest morbidity rate among female malignant tumors in the world. With significant advances in medical technology, the prognosis for breast cancer has improved significantly. However, the prognosis for breast cancer sufferers who develop distant metastasis remains poor, posing a serious threat to the lives of breast cancer patients. In this research, we used a comprehensive bioinformatics approach, including WGCNA, differential analysis and enrichment analysis, to analyze and screen for target genes that may be related with the emergence of distant metastases in breast cancer and to research the effect of these target genes in distant metastases based on the outcomes of enrichment analysis, which demonstrated that these target genes were associated with phospholipase activity and modified amino acid binding. Phospholipase D (PLD) is known to cause multiplication of breast cancer cell lines, phospholipase activity, particularly phospholipase D, is related with breast cancer metastasis [29]. The amino acid asparagine (ASN) has been shown to promote the spread and proliferation of breast cancer cells, and reducing the intake of ASN can lead to a reduction in the size of breast cancer cells and reduce the probability of breast cancer cells metastasizing to the lungs [30]. KEGG enrichment analysis has shown that these target genes are associated with the MAPK signaling pathway, among others. Dysregulation of the MAPK signaling pathway, one of the most frequently dysregulated pathways in cancer, promotes uncontrolled cell multiplication and metastasis and inhibits apoptosis [31], and aberrant activation of the MAPK signaling pathway has now been identified in some clinical diagnoses of breast cancer patients [32,33]. Therefore, interfering with the MAPK signaling pathway with appropriate inhibitors may be essential in the fight against cancers such as breast cancer.

In this study, LASSO regression analysis was further performed on the obtained target genes that might be associated with the occurrence of distant metastasis of breast cancer, and 21 biomarkers of distant metastasis from breast cancer, including SPN, MGC2840 and RAB13, were obtained. Correlation analysis of breast cancer distant metastasis biomarker expression showed that the expression of three genes, SPN, MGC2840 and RAB13, was negatively correlated with the expression of other genes in breast cancer distant metastasis biomarkers. The expression levels of SPN, RAB13 and MGC2840 were higher in samples from breast cancer victims who occurred distant metastases, and therefore these three genes may have a facilitative effect in the metastasis of breast cancer cells to other organs. We also predicted the transcription factors of distant metastasis biomarkers of breast cancer and constructed a TF-biomarker network, from which we concluded that SPN, RAB13 and C21orf91 might be regulated by KLF6 and others, and MAPK7 might be regulated by SP1 and others. The transcription factor KLF6 missplicing actually generates a protein that can cause the spread or metastasis of cancer cells; thus, SPN and RAB13 may have a facilitative effect in the metastasis of breast cancer cells to other organs. Research shows that the transcription factor Sp1 plays a regulatory effect in biological processes such as cell differentiation, tumour progression and metastasis [34–36]. Moreover, the transcription factor Sp1 is highly expressed in some victims with cancers such as breast cancer [37,38] and plays an essential part in the invasion and metastasis of malignancies like colon, gastric, pancreatic and breast cancers and is one of the poor prognostic contributors for cancers such as breast cancer [39]. From the above analyses, it can be seen that the obtained biomarkers are closely related to distant metastasis of breast cancer, so the construction of breast cancer distant metastasis prediction model based on the obtained biomarkers of breast cancer distant metastasis has a reliable biological basis and good interpretability.

In this research, we present a machine learning-based breast cancer distant metastasis forecasting model that can predict whether a breast cancer patient has distant metastasis based on the expression data of breast cancer distant metastasis biomarkers in breast cancer patient samples. The validation results showed that our proposed breast cancer distant metastasis forecasting model could predict whether breast cancer patients had distant metastasis more accurately, with an accuracy of 93.6 % on the validation set, an F1-score of 88.9 % and an AUC value of 91.3 %. Previously, Mahendran Botlagunta et al. used extensive data analysis to conclude that blood analysis data could be used as a basis for early identification of breast cancer metastasis and structured a machine learning algorithm-based model for classification and diagnosis prediction of breast cancer metastasis, and the model had a prediction accuracy of 83 % [40]. The accuracy of our model for predicting distant breast cancer metastasis on the basis of the obtained biomarkers of distant breast cancer metastasis is approximately 10 % higher than that of the breast cancer metastasis diagnosis prediction model proposed by Mahendran Botlagunta et al. Our prediction model is an improvement over previous studies in terms of accuracy in predicting the emergence of distant metastases in breast cancer victims.

There are also some limitations of this study. Firstly, this study only focused on the occurrence of distant metastasis of breast cancer and did not refine the study to the specific organs that breast cancer metastasises to. Secondly, this study discovered breast cancer distant metastasis biomarkers through data mining, and actual clinical trials and so on are needed to further validate these findings. In future studies, we believe that the accuracy of breast cancer distant metastasis prediction can be further improved by using advanced pre-training models such as BERT, and with the further accumulation of breast cancer-related sequencing data, the prediction of breast cancer distant metastasis can be refined to specific organs, which can provide a reference for the development of more personalised treatment plans.

5. Conclusion

In summary, we identified 21 biomarkers of distant metastases in breast cancer, including SPN and RAB13, by means of difference analysis, WGCNA and LASSO analysis, and constructed several prediction models using these biomarkers as features, from which we selected the best-performing random forest model-based breast cancer distant metastasis forecasting model. Our proposed prediction model (https://github.com/dw666666/Prediction-model-of-distant-metastasis-of-breast-cancer.git) can accurately predict the emergence of distant metastasis in breast cancer victims, which can provide a more accurate early identification of metastatic breast cancer patients and provide better timing and more treatment time for metastatic breast cancer patients.

Funding

The work was supported by the National Natural Science Foundation of China (No. 62262015, No. 62102064, No.62101100) and the National Natural Science Foundation of Sichuan Province of China (No. 2022NSFSC0943).

Availability of data and materials

Code used in this study is available from https://github.com/dw666666/Prediction-model-of-distant-metastasis-of-breast-cancer.git. Publicly available datasets were analyzed in this study. GEO data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors expressed great pleasure to publish in this journal.

CRediT authorship contribution statement

Hao Duan: Conceptualization, Formal analysis, Methodology, Writing – original draft. **Yu Zhang:** Conceptualization, Methodology, Writing – original draft. **Haoye Qiu:** Resources, Writing – original draft. **Xiuhao Fu:** Formal analysis, Writing – original draft. **Chunling Liu:** Writing – original draft, Investigation. **Xiaofeng Zang:** Writing – original draft, Investigation. **Anqi Xu:** Writing – original draft, Investigation. **Ziyue Wu:** Resources, Writing – original draft. **Xingfeng Li:** Resources, Writing – review & editing. **Qingchen Zhang:** Resources, Writing – review & editing. **Zilong Zhang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Feifei Cui:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

Not applicable.

References

[1] D. Su, Q. Lu, Y. Pan, Y. Yu, S. Wang, Y. Zuo, L. Yang, Immune-related gene-based prognostic signature for the risk stratifica-tion analysis of breast cancer, Curr. Bioinf. 17 (2) (2022) 196–205.
[2] M.T. Chen, H.F. Sun, Y. Zhao, W.Y. Fu, L.P. Yang, S.P. Gao, L.D. Li, H.L. Jiang, W. Jin, Comparison of patterns and prognosis among distant metastatic breast cancer patients by age groups: a SEER population-based analysis, Sci. Rep. 7 (1) (2017) 9254.
[3] S. Lin, Y. Lin, K. Wu, Y. Wang, Z. Feng, M. Duan, S. Liu, Y. Fan, L. Huang, F. Zhou, Construction of network biomarkers using inter-feature correlation Co-efficients (FeCO3) and their application in detecting high-order breast cancer biomarkers, Curr. Bioinf. 17 (4) (2022) 310–326.
[4] H. Zhao, X. Yin, L. Wang, K. Liu, W. Liu, L. Bo, L. Wang, Identifying tumour microenvironment-related signature that correlates with prognosis and immunotherapy response in breast cancer, Sci. Data 10 (1) (2023) 119.
[5] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2020, CA A Cancer J. Clin. 70 (1) (2020) 7–30.
[6] Z. Zhang, F. Cui, C. Wang, L. Zhao, Q. Zou, Goals and approaches for each processing step for single-cell RNA sequencing data, Briefings Bioinf. 22 (4) (2021) 1–10.
[7] C. Cao, J. Wang, D. Kwok, F. Cui, Z. Zhang, D. Zhao, M.J. Li, Q. Zou, webTWAS: a resource for disease candidate susceptibility genes identified by transcriptome-wide association study, Nucleic Acids Res. 50 (D1) (2022) D1123–d1130.
[8] W. Tang, S. Wan, Z. Yang, A.E. Teschendorff, Q. Zou, Tumor origin detection with tissue-specific miRNA and DNA methylation markers, Bioinformatics 34 (3) (2018) 398–406.
[9] Z. Zhang, F. Cui, C. Lin, L. Zhao, C. Wang, Q. Zou, Critical downstream analysis steps for single-cell RNA sequencing data, Briefings Bioinf. 22 (5) (2021) 1–11.
[10] Z. Zhang, F. Cui, C. Cao, Q. Wang, Q. Zou, Single-cell RNA analysis reveals the potential risk of organ-specific cell types vulnerable to SARS-CoV-2 infections, Comput. Biol. Med. 140 (2021) 105092.
[11] Z. Zhang, F. Cui, M. Zhou, S. Wu, Q. Zou, B. Gao, Single-cell RNA sequencing analysis identifies key genes in brain metastasis from lung adenocarcinoma, Curr. Gene Ther. 21 (4) (2021) 338–348.
[12] X. Chen, Y. Lin, Q. Qu, B. Ning, H. Chen, B. Liao, X. Li, Analyzing association between expression quantitative trait and CNV for breast cancer based on gene interaction network clustering and group sparse learning, Curr. Bioinf. 17 (4) (2022) 358–368.
[13] R. Qi, F. Guo, Q. Zou, String kernels construction and fusion: a survey with bioinformatics application, Front. Comput. Sci. 16 (6) (2022) 166904.
[14] Q.W. Li, L.C. Zhang, L. Xu, Q. Zou, J. Wu, Q.Y. Li, Identification and classification of promoters using the attention mechanism based on long short-term memory, Front. Comput. Sci. 16 (4) (2022) 164348.
[15] L. Jiang, C. Liu, Y. Fan, Q. Wu, X. Ye, Q. Li, Y. Wan, Y. Sun, L. Zou, D. Xiang, et al., Dynamic transcriptome analysis suggests the key genes regulating seed development and filling in Tartary buckwheat (Fagopyrum tataricum Garetn.), Front. Genet. 13 (2022) 990412.
[16] Z. Liu, L. Liu, S. Weng, C. Guo, Q. Dang, H. Xu, L. Wang, T. Lu, Y. Zhang, Z. Sun, et al., Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer, Nat. Commun. 13 (1) (2022) 816.

[17] M. Wang, J. Liao, J. Wang, M. Qi, K. Wang, W. Wu, TAF1A and ZBTB41 serve as novel key genes in cervical cancer identified by integrated approaches, Cancer Gene Ther. 28 (12) (2021) 1298–1311.

[18] Y. Gao, G.Y. Cai, W. Fang, H.Y. Li, S.Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P.F. Cui, et al., Machine learning based early warning system enables accurate mortality risk prediction for COVID-19, Nat. Commun. 11 (1) (2020) 5033.

[19] A.J. McEligot, V. Poynor, R. Sharma, A. Panangadan, Logistic LASSO regression for dietary intakes and breast cancer, Nutrients 12 (9) (2020) 2652.

[20] L. Meng, T. Zheng, Y. Wang, Z. Li, Q. Xiao, J. He, J. Tan, Development of a prediction model based on LASSO regression to evaluate the risk of non-sentinel lymph node metastasis in Chinese breast cancer patients with 1-2 positive sentinel lymph nodes, Sci. Rep. 11 (1) (2021) 19972.

[21] M. Mohammed, H. Mwambi, I.B. Mboya, M.K. Elbashir, B. Omolo, A stacking ensemble deep learning approach to cancer type classification based on TCGA data, Sci. Rep. 11 (1) (2021) 15626.

[22] S. Sperandei, Understanding logistic regression analysis, Biochem. Med. 24 (1) (2014) 12–18.

[23] M. Schonlau, R.Y. Zou, The random forest algorithm for statistical learning, STATA J.: Promoting communications on statistics and Stata 20 (1) (2020) 3–29.

[24] Z. Lv, J. Zhang, H. Ding, Q. Zou, Rf-PseU, A random forest predictor for RNA pseudouridine sites, Front. Bioeng. Biotechnol. 8 (2020) 134.

[25] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of Support vector machine (SVM) learning in cancer genomics, Cancer Genomics Proteomics 15 (1) (2018) 41–51.

[26] Y. Wang, Y. Zhai, Y. Ding, Q. Zou, SBSM-pro: Support Bio-Sequence Machine for Proteins, 2023 *arXiv preprint arXiv:230810275*.

[27] C. Zhang, C. Liu, X. Zhang, G. Almpanidis, An up-to-date comparison of state-of-the-art classification algorithms, Expert Syst. Appl. 82 (2017) 128–150.

[28] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artif. Intell. Rev. 54 (3) (2020) 1937–1967.

[29] M. Lingrand, S. Lalonde, A. Jutras-Carignan, K.F. Bergeron, E. Rassart, C. Mounier, SCD1 activity promotes cell migration via a PLD-mTOR pathway in the MDA-MB-231 triple-negative breast cancer cell line, Breast Cancer 27 (4) (2020) 594–606.

[30] S.R.V. Knott, E. Wagenblast, S. Khan, S.Y. Kim, M. Soto, M. Wagner, M.O. Turgeon, L. Fish, N. Erard, A.L. Gable, et al., Asparagine bioavailability governs metastasis in a model of breast cancer, Nature 554 (7692) (2018) 378–381.

[31] M. Mutlu, O. Saatci, S.A. Ansari, E. Yurdusev, H. Shehwana, O. Konu, U. Raza, O. Sahin, miR-564 acts as a dual inhibitor of PI3K and MAPK signaling networks and inhibits proliferation and invasion in breast cancer, Sci. Rep. 6 (2016) 32541.

[32] S.P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, et al., The clonal and mutational evolution spectrum of primary triple-negative breast cancers, Nature 486 (7403) (2012) 395–399.

[33] N. Cancer Genome Atlas, Comprehensive molecular portraits of human breast tumours, Nature 490 (7418) (2012) 61–70.

[34] S. Banerjee, V. Sangwan, O. McGinn, R. Chugh, V. Dudeja, S.M. Vickers, A. K. Saluja, Triptolide-induced cell death in pancreatic cancer is mediated by O-GlcNAc modification of transcription factor Sp1, J. Biol. Chem. 288 (47) (2013) 33927–33938.

[35] K. Beishline, J. Azizkhan-Clifford, Sp1 and the 'hallmarks of cancer', FEBS J. 282 (2) (2015) 224–258.

[36] E. Deniaud, J. Baguet, R. Chalard, B. Blanquier, L. Brinza, J. Meunier, M. C. Michallet, A. Laugraud, C. Ah-Soon, A. Wierinckx, et al., Overexpression of transcription factor Sp1 leads to gene expression perturbations and cell cycle inhibition, PLoS One 4 (9) (2009) e7035.

[37] N.Y. Jiang, B.A. Woda, B.F. Banner, G.F. Whalen, K.A. Dresser, D. Lu, Sp1, a new biomarker that identifies a subset of aggressive pancreatic ductal adenocarcinoma, Cancer Epidemiol. Biomarkers Prev. 17 (7) (2008) 1648–1652.

[38] M. Kanai, D. Wei, Q. Li, Z. Jia, J. Ajani, X. Le, J. Yao, K. Xie, Loss of Kruppel-like factor 4 expression contributes to Sp1 overexpression and human gastric cancer development and progression, Clin. Cancer Res. 12 (21) (2006) 6395–6402.

[39] S. Maor, D. Mayer, R.I. Yarden, A.V. Lee, R. Sarfstein, H. Werner, M.Z. Papa, Estrogen receptor regulates insulin-like growth factor-I receptor gene expression in breast tumor cells: involvement of transcription factor Sp1, J. Endocrinol. 191 (3) (2006) 605–612.

[40] M. Botlagunta, M.D. Botlagunta, M.B. Myneni, D. Lakshmi, A. Nayyar, J. S. Gullapalli, M.A. Shah, Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms, Sci. Rep. 13 (1) (2023) 485.