RESEARCH ARTICLE

Identifying Diagnostic Biomarkers of Breast Cancer Based on Gene Expression Data and Ensemble Feature Selection

Lingyu Li[1], Yousif A. Algabri[1] and Zhi-Ping Liu[1,*]

[1]*Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China*

ARTICLE HISTORY

Received: July 08, 2022
Revised: October 30, 2022
Accepted: November 03, 2022

DOI:
10.2174/1574893618666230111153243

Abstract: ***Background***: In recent years, the identification of biomarkers or signatures based on gene expression profiling data has attracted much attention in bioinformatics. The successful discovery of breast cancer (BRCA) biomarkers will be beneficial in reducing the risk of BRCA among patients for early detection.

Methods: This paper proposes an Ensemble Feature Selection method to screen biomarkers (abbreviated as EFSmarker) for BRCA from publically available gene expression data. Firstly, we employ twelve filter feature selection methods, namely median, variance, Chi-square, Relief, Pearson and Spearman correlation, mutual information, minimal-redundancy-maximal-relevance criterion, ridge regression, decision tree and random forest with Gini index and accuracy index, to calculate the importance (weights or coefficients) of all features on the training dataset. Secondly, we apply the logistic regression classifier on the test dataset to calculate the classification AUC value of each feature subset individually selected by twelve methods. Thirdly, we provide an ensemble feature selection method by aggregating feature importance with classification AUC value. In particular, we establish a feature importance score (FIS) to evaluate the importance of each feature underlying all feature selection methods. Finally, the features with higher FIS are taken as identified biomarkers.

Results: With the direction of the FIS index induced by the EFSmarker method, 12 genes (COL10A1, COL11A1, MMP11, LOC728264, FIGF, GJB2, INHBA, CD300LG, IGFBP6, PAMR1, CXCL2 and FXYD1) are regarded as diagnostic biomarkers for BRCA. Especially, COL10A1, ranked first with a FIS value of 0.663, is identified as the most credible biomarker. The findings justified *via* gene and protein expression validation, functional enrichment analysis, literature checking and independent dataset validation verify the effectiveness and efficiency of these selected biomarkers.

Conclusion: Our proposed biomarker discovery strategy not only utilizes the feature contribution but also considers the prediction accuracy simultaneously, which may also serve as a model for identifying unknown biomarkers for other diseases from high-throughput gene expression data. The source code and data are available at https://github.com/zpliulab/EFSmarker.

Keywords: Biomarker, machine learning, ensemble feature selection, gene expression data, breast cancer, early detection.

1. INTRODUCTION

Breast cancer (BRCA) is one of the most common cancers and a leading cause of cancer mortality in over 3.5 million women globally [1]. Due to the high heterogeneity, initially asymptomatic and continuous drug resistance of BRCA, it is vital to find novel biomarkers for early diagnosis, assisting clinicians in making optimal treatment and prognostic throughput sequencing technologies and the opening of many publicly available databases have facilitated the discovery of decisions [2, 3]. The last decade has experienced the advent of high-potential BRCA biomarkers from genomic and transcriptomic data [4, 5]. This work aims to develop an effective and efficient framework for biomarker discovery from high-throughput data in order to provide an auxiliary means for the early diagnosis of BRCA or other complex diseases.

The biomarker discovery in biomedicine is essentially equivalent to the feature selection in machine learning [6]. Many feature selection methods are used to identify biomarkers from gene expression data and they can mainly be

*Address correspondence to this author at the Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China; Tel: +86-531-88392280; E-mail: zpliu@sdu.edu.cn

summarized into three categories: filter, wrapper and embedded methods [7]. The filter method first ranks all features by calculating feature relevance, screens the high-scoring features above a specified threshold, and then trains and validates these features using existing classification algorithms independently. The well-known statistically oriented filter-based feature selection methods include Median [8], Variance [9], Chi-square (Chi2) [10], Relief [11] and Correlation [12]. The information-based methods mainly contain mutual information (MI) [13] and minimal-redundancy-maximal-relevance criterion (mRMR) [14]. Besides, other model-based methods, for example, ridge regression (Ridge) [7], decision tree (DT) [15] and random forest (RF) [15], are also usually used to select features.

Feature subset selection in the filter method is conducted by a preprocessing step independently of the chosen predictor [16]. In contrast, the wrapper and embedded methods generally use a specific machine learning algorithm/model to evaluate a specific feature subset [16]. Specifically, the embedded method simultaneously perform feature selection and classification [7]. It is well known that different feature selection methods may generate different feature subsets. Amazingly, these features can often obtain higher prediction accuracy in both the internal datasets for selection and the external datasets for validation. However, reproducing a list of selected feature genes is crucial in various biomedical domains [17].

Recently, there has been a surge of interest in ensemble feature selection (EFS) algorithms, which have emerged as a viable alternative to integrate the advantages of a single feature selection algorithm and compensate for their disadvantages [18]. For instance, Chiew *et al.* [19] proposed a hybrid EFS framework for machine learning-based phishing detection systems. Wang *et al.* [20] developed an EFS method for some classification tasks. Abeel *et al.* [21] exploited multi-model EFS methods on the environmental sound data. Abeel *et al.* [21] conducted a large-scale analysis for EFS by data diversity to increase the robustness of the final selected feature subset. The EFS method has been an effective way to provide a unique and stable feature selection. It is also expected to improve the accuracy of classification after feature selection [16, 17, 22].

Currently, there are three major types of ensemble techniques in feature selection, namely data diversity, functional diversity and hybrid method, in the field of EFS from omics data. Specifically, data diversity uses the same feature selection method on different subsets of a dataset by repeated and random samplings [23]. Functional diversity applies different feature selection techniques to the same dataset [16]. In contrast, the hybrid method conducts multiple feature selection techniques on multiple datasets [19]. Many studies have suggested that functional diversity-based EFS is a promising and cutting-edge approach because of its effectiveness in achieving better learning results by combining different learning models [20, 23-25].

In this work, we aim to propose an ensemble feature selection method for biomarker discovery (abbreviated as EFSmarker) based on multiple independent feature selection methods to better approximate the optimal subset of features. Here, we first chose twelve filter feature selection methods. EFSmarker generates a feature score that is used to evaluate the importance of features by considering the feature weight and classification accuracy and integrating them into an aggregation-like framework. The results of expression validation, function enrichment analysis, literature checking and external validation demonstrate that the proposed biomarker discovery strategy is effective in identifying biomarkers in a case study of BRCA. The biomarkers allow non-invasive detection of patients with BRCA, which guides patients and doctors in selecting an alternative diagnostic approach.

The rest of this paper is organized as follows. In Section 2, we propose the general framework of EFSmarker for discovering biomarkers from gene expression data. We introduce twelve filter feature selection methods and the innovative aggregation strategy. We further establish the feature importance score (FIS) and prediction risk score (PRS) to evaluate the feature importance and disease risk. In Section 3, we verify the identified biomarkers not only *via* differential expression analysis for both gene and protein, functional enrichment analysis and literature check but also through external independent validations on ten real-world independent BRCA datasets. In Section 4, we present the discussion. Finally, we give conclusion and further research directions in Section 5.

2. MATERIALS AND METHODS

2.1. Dataset Collection and Data Preprocessing

The gene expression profiling data of BRCA used in our study are downloaded from The Cancer Genome Atlas (TCGA) database (https://cancergenome.nih.gov/) and Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). We perform the biomarker discovery experiments using the dataset from the TCGA database. The ten datasets from GEO are used for validation purposes. Table **1** lists the details of the datasets.

As mentioned, the BRCA dataset from TCGA consists of 112 adjacent normal breast tissue. In the experiments, we extract the 112 corresponding tumor samples for a fair sample matching. Namely, according to the first three elements of the sample name in the control group (separated by "-"), such as TCGA-A7-A0CE, we also extract disease samples with the same sample name in the tumor group. With this in place, it ensures that we choose the samples with the same tissue origin and participant. Thus, a balanced dataset with the same number of positive and negative samples is established.

Specifically, we first identify differentially expressed genes (DEGs) across the 224 samples (112 controls and 112 tumors) by DEseq2 [26], which results in 489 genes with adjusted *P*-value (P.adj) < 0.01 and |log(FC)| > 3.322. The full list of DEGs is listed in Supplementary Material Table **S1**. Next, combining 489 DEGs with 334 known BRCA-related disease genes or the genes with important roles in the

Table 1. The information of datasets used in this study, where the number in parentheses refers to the sample size.

Purpose	Dataset	Platform	# of Genes	# of Samples (Normal-BRCA)
For discovery (1205)	TCGA	DCC	20501	1205 (112-1093)
For validation (688)	GSE10780	GPL570	21835	185 (143-42)
	GSE10797	GPL571	13701	33 (5-28)
	GSE10810	GPL570	11713	58 (27-31)
	GSE20437- GSE31519	GPL96	13749	109 (42-67)
	GSE21422	GPL570	21835	19 (5-14)
	GSE26910	GPL570	21835	12 (6-6)
	GSE38959	GPL4133	19902	43 (13-30)
	GSE42568	GPL570	21835	121 (17-104)
	GSE45827	GPL570	15064	52 (11-41)
	GSE61304	GPL570	21835	56 (4-52)

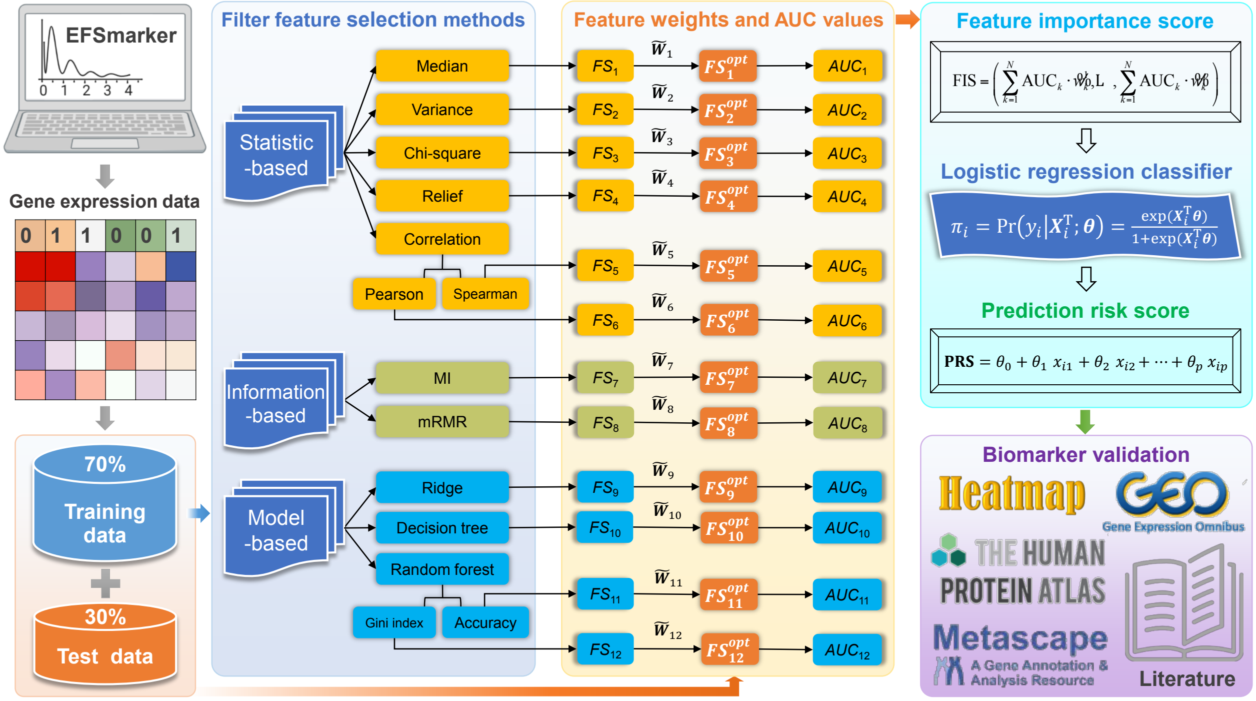
EFSmarker
0 1 2 3 4
Gene expression data
0 1 1 0 0 1
70%
Training data
30%
Test data
Filter feature selection methods
Statistic -based
Information -based
Model -based
Median
Variance
Chi-square
Relief
Correlation
Pearson
Spearman
MI
mRMR
Ridge
Decision tree
Random forest
Gini index
Accuracy
Feature weights and AUC values
$\widetilde{W}_1$ … $\widetilde{W}_{12}$
FS_1 … FS_{12}
FS_1^{opt} … FS_{12}^{opt}
AUC_1 … AUC_{12}
Feature importance score
$\mathrm{FIS} = \left(\sum_{k=1}^{N} \mathrm{AUC}_k \cdot w_k^1, \mathrm{L}, \sum_{k=1}^{N} \mathrm{AUC}_k \cdot w_k^p\right)$
Logistic regression classifier
$\pi_i = \Pr\left(y_i \middle| \boldsymbol{X}_i^{\mathrm{T}}; \boldsymbol{\theta}\right) = \frac{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta})}{1+\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\theta})}$
Prediction risk score
$\mathbf{PRS} = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip}$
Biomarker validation
Heatmap
GEO
Gene Expression Omnibus
THE HUMAN PROTEIN ATLAS
Metascape
A Gene Annotation & Analysis Resource
Literature

Fig. (1). The framework of identifying diagnostic biomarkers from gene expression data using EFSmarker method. The left top figure is adapted from "Icon Pack - Computer", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates. The right bottom figure is adapted from the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/), Human Protein Atlas database (https://www.proteinatlas.org/) and Metascape Portal (https://metascape.org/). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

BRCA knowledgebase (82 elite cancer genes from MalaCards [27], 47 genes from KEGG BRCA pathway [28], 70 genes used by Mammaprint [29] and 119 transcription factors (TFs) genes in RegNetwork [30]), we put these 823 genes into a whole candidate pool, based on which we conduct EFSmarker to select BRCA biomarkers.

2.2. Framework

Fig. (**1**) illustrates the framework for identifying biomarkers from gene expression data by the EFSmarker method. First, for training and testing purposes, we randomly divide all samples into two subsets, 70% for training and 30% for testing in the biomarker discovery section. Second, twelve filter feature selection methods are used to obtain the importance (weights or coefficients) of all features. They are summarized in three groups, *i.e.*, six statistic-based methods including Median, Variance, Chi2, Relief and Correlation (Person and Spearman), two information-based methods including MI and mRMR, and four model-based methods including Ridge, DT and RF (Gini index and Accuracy index). Third, we verify the classification performances (assessed by AUC) of the selected features on the test dataset using a logistic regression model. Fourth, we propose the ensemble feature selection strategy by aggregating weights and AUC values in order to establish a feature importance score (FIS) to evaluate the importance of each feature among all the twelve methods. Without loss of generality, we take the features with higher FIS as the identified biomarkers and validate them on the independent datasets using the trained logistic regression classifier. Particularly, we introduce the prediction risk score (PRS) to distinguish high-risk samples from low-risk ones. Lastly, we also conduct a series of validations, including differential expression analysis (gene and protein), function enrichment analysis, literature checking and independent dataset validation, to justify the validity and reliability of these biomarkers identified by our proposed EFSmarker method.

2.3. Filter Feature Selection Methods

The feature selection step is independent of the subsequent machine learning process for the filtering feature selection method. It first performs feature selection on the training dataset according to a specific rule and then uses the screened features to train the machine learning model. Here, we only consider its first step, i.e., feature selection "filtering". Based on that, we ensemble the obtained feature subsets according to some aggregation strategies.

We essentially focus on a binary classification problem with feature selection. Suppose we have *n* independently and identically distributed observations $(\boldsymbol{X}_i, y_i)$, i =1, 2, …, *n*. Taking these observations as a dataset D with

$$D = \left\{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \cdots, (\mathbf{X}_n, y_n)\right\}, \qquad (1)$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T \in \mathbb{R}^p$ represents the *p*-dimensional gene expression vector of the *i*-th sample, x_{ij} denotes the gene expression value of the *j*-th gene in the *i*-th sample, and y_i is a corresponding variable that takes a value of 0 or 1. For completeness, we introduce the twelve filter feature selection methods.

2.3.1. Statistic-based Methods

Mann-Whitney *U* Test proposed by another study [8] is also called Wilcoxon rank-sum test. It aims to test whether the means of normal samples with disease samples are significantly different. Suppose that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)^T \in \mathbb{R}^{n \times p}$, let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_m$ be the normal samples from $\boldsymbol{X}$, and $\boldsymbol{X}_{m+1}, \boldsymbol{X}_{m+2}, \cdots, \boldsymbol{X}_n$ be the disease samples from $\boldsymbol{X}$, then the corresponding Mann-Whitney U statistic is defined as

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n-m} S(\boldsymbol{X}_i, \boldsymbol{X}_j), \qquad (2)$$

where

$$S(X, Y) = \begin{cases} 1, & if\ X > Y, \\ 1/2, & if\ Y = X, \\ 0, & if\ X < Y. \end{cases} \qquad (3)$$

Here, we use the resulting *P*-value from Mann-Whitney U Test as the feature importance measure and call it as *Median* method, where the smaller *P*-value indicates the higher importance (Eqs. 2 and 3).

The variance selection method first calculates the variance of each feature $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$,

$$Var(\boldsymbol{x}_j) = E[(\boldsymbol{x}_j - \mu)^2], \quad j = 1, 2, \cdots, p, \qquad (4)$$

where μ is the mean of the variable $\boldsymbol{x}_j$ and $E[\cdot]$ represents the expectation. Then, it selects the feature with a variance greater than a threshold and finally chooses the top-ranked features with higher variances [9]. Feature with larger variance means a more critical role.

In bioinformatics, the Chi-square (Chi2) test [10] is used to test the correlation of certain independent variables/genes to qualitative dependent variables, i.e., different categories: normal or disease. Feature with larger Chi2 value indicates higher importance.

The Relief method designs a "relevant statistic" vector to measure the importance of features [11]. Each component of the vector corresponds to an initial feature, and the importance of a feature subset is determined by the sum of the components of the "correlation statistic" corresponding to each feature in the subset. The implementation of Relief mainly includes two steps: First, select the feature corresponding to the component of the "correlation statistic" larger than a specified threshold. Second, select the l features with the largest component of the "correlation statistic", where l is the number of features to be selected.

Correlation is usually used to calculate the correlation coefficient of each feature on the target variable and the *P*-value of the correlation coefficient [12]. The correlation-based filter can avoid multi-collinearity by selecting highly correlated features with the dependent variable but showing only a low correlation with other features. Here, we separately apply the Pearson correlation method (P_cor) and Spearman rank correlation method (S_cor) to calculate the correlation coefficients and adopt their *P*-values as the feature importance measure, where the importance value of eliminated feature is set to zero.

2.3.2. Information-based Methods

In information theory, the mutual information (MI) of two random variables measures the interdependence between the variables [13]. Let (X, Y) be a pair of random variables with values over the space $\mathrm{X} \times \mathrm{Y}$. Define the joint distribution be $P_{(X,Y)}$ and the marginal distributions be P_X and P_Y, respectively. Then, the mutual information is defined as

$$\boldsymbol{I}(X;Y) = \sum_{y \in \mathrm{Y}} \sum_{x \in \mathrm{X}} P_{(X,Y)}(x, y) \log\left(\frac{P_{(X,Y)}(x, y)}{P_X(x) P_Y(y)}\right). \qquad (5)$$

Unlike the correlation coefficient, mutual information is not limited to real-valued random variables. It is more gen-

Table 2. Overview of the filter methods: the name of the filter method (Method), a short description of the filter (Description), the type of the filter method (Category) and information about the R or Python from which the implementation is taken (Implementation).

Method	Description	Category	Implementation
Median	Mann-Whitney *U* Test	Statistic-based	wilcox.test in R
Variance	Feature variance	Statistic-based	sklearn.feature_selection in Pyhton
Chi2	Chi-square	Statistic-based	sklearn.feature_selection in Pyhton
Relief	Relevant feature	Statistic-based	numpy, random and sklearn.preprocessing in Python
P_cor	Pearson correlation	Statistic-based	cor in R, method = "pearson"
S_cor	Spearman rank correlation	Statistic-based	cor in R, method = "spearman"
MI	Mutual information	Information-based	sklearn.feature_selection in Pyhton
mRMR	Minimal-redundancy-maximal-relevance	Information-based	mRMRe in R
Ridge	Ridge regression	Model-based	glmnet package in R
DT	Decision tree	Model-based	sklearn.model_selection and sklearn.tree in Pyhton
RF_Gini	Random forest Gini importance	Model-based	Random forest package in R, Gini importance
RF_Acc	Random forest impurity importance	Model-based	Random forest package in R, accuracy importance

eral and determines how similar the joint distribution $P_{(X,Y)}$ and the product of the decomposed marginal distributions $P_X \cdot P_Y$.

The minimal-redundancy-maximal-relevance criterion (mRMR) is a particularly appealing and fast feature selection method. It is usually used to find a set of both relevant and complementary features with relatively low computational complexity [14]. The challenge of mRMR is that it produces highly variable results, where small changes in sample data often lead to significantly different sets of selected features.

2.3.3. Model-based Methods

Ridge regression is developed to address the overfitting problem of logistic regression by adding an L_2 norm penalty to the negative of the log-likelihood function [7]. The Ridge regression model is formula as

$$\boldsymbol{\theta} = \arg\min\left\{-\mathrm{L}(\boldsymbol{\theta} \mid \mathrm{D}) + \mathrm{P}\left(\boldsymbol{\theta};\lambda\right)\right\}, \tag{6}$$

where $\mathrm{L}(\boldsymbol{\theta} \mid \mathrm{D}) = \sum_{i=1}^{n}\left\{y_i \log\left[f(\boldsymbol{X}_i^{\text{ï}}\,\boldsymbol{\theta})\right] + \left(1-y_i\right)\log\left[1-f(\boldsymbol{X}_i^{\text{ï}}\,\boldsymbol{\theta})\right]\right\}$ is the log-likelihood function, $f(x) = \exp(x)/(1+\exp(x))$ is the sigmoid function, $\mathrm{P}(\boldsymbol{\theta};\lambda) = \lambda\sum_{j=1}^{p}\theta_j^2$ is the penalty term, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \mathrm{L}\ , \theta_p)^{\text{ï}}$ is the unknown coefficient vector with intercept term θ_0. In particular, we use the regression coefficient solved by Eq. (6) to measure the feature importance. The larger the absolute value of the coefficient, the more influential the feature.

Decision tree (DT) is a method based on the tree model, which can judge the relationship between features and predicted values, and adopt corresponding linear/non-linear algorithms according to the linear/non-linear relationship [15]. During feature selection, DT uses each feature individually for modeling and cross-validation and then selects a specified number of the highest-scoring features to form a feature subset.

Random Forest (RF) is an ensemble of multiple decision trees, which gain their randomness from randomly chosen starting features for each tree [15]. On the one hand, RF provides a measure evaluating importance based on the Gini index, which measures the node impurity in the trees. On the other hand, RF also provides an error-rate-based importance measure, *i.e.*, accuracy. If a variable is important enough, changing it will significantly increase the test error. On the contrary, changing it does not increase the test error means that the variable is not so important.

2.3.4. Implementation of Filter Methods

Table **2** provides an overview of the filter feature selection methods and the implementations used for the feature subset aggregation in this article.

2.4. Logistic Regression Classifier

For a given dataset D the logistic regression model is as follows

$$\pi_i = \Pr\left(y_i \mid \boldsymbol{X}_i;\boldsymbol{\theta}\right) = f\left(\boldsymbol{X}_i^{\text{ï}}\,\boldsymbol{\theta}\right) = \frac{\exp\left(\boldsymbol{X}_i^{\text{ï}}\,\boldsymbol{\theta}\right)}{1+\exp\left(\boldsymbol{X}_i^{\text{ï}}\,\boldsymbol{\theta}\right)}, \tag{7}$$

where π_i represents the predicted disease state of the *i*-th sample ($\pi_i = 1$ means disease, $\pi_i = 0$ means normal). Taking the logit transformation to Equation (7), we get the logistic regression classifier

$$\operatorname{logit}\left(\pi_i\right) = \log\frac{\pi_i}{1-\pi_i} = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip}. \tag{8}$$

Then, we establish the indicator of prediction risk score (PRS) based on Equation (8) to identify the BRCA diagnostic risk of a patient,

$$\mathrm{PRS} = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip}. \tag{9}$$

From Eq. (9), it can be known that the PRS index is a real number. The larger the PRS value, the greater the risk. The index has important guiding significance for clinicians to achieve early breast cancer diagnosis through a quantitative approach.

2.5. Ensemble Feature Selection Method

In the feature filter step, by applying all kinds of feature selection methods mentioned above to the same training data, we can always obtain the feature subset $\boldsymbol{FS}_k$,

$$\boldsymbol{FS}_k = (f_k^1, f_k^2 \cdots, f_k^p), \quad k = 1, 2, \cdots, K, \tag{10}$$

and corresponding weights $\boldsymbol{W}_k$,

$$\boldsymbol{W}_k = (w_k^1, w_k^2 \cdots, w_k^p), \quad k = 1, 2, \cdots, K, \tag{11}$$

where K is the number of used feature selection methods. We first normalize the weights by

$$\overline{\boldsymbol{W}}_k = (\overline{w}_k^1, \overline{w}_k^2 \cdots, \overline{w}_k^p), \tag{12}$$

where $\overline{w}_k^j = \dfrac{w_k^j}{\max_{1 \le j \le p} w_k^j}$. Thus, the highest weight of a feature is assigned as 1, and the lowest weight is 0 (Eq. 12). Then, we normalize the weight of each feature result across all feature selection methods to an interval from 0 to $1/K$, so they have a standard scale. Thus, we obtain the updated weights of features selected by K different feature selection methods,

$$\widetilde{\boldsymbol{W}}_k = (\tilde{w}_k^1, \tilde{w}_k^2, \cdots, \tilde{w}_k^p), \tag{13}$$

where $\tilde{w}_k^j = \overline{w}_k^j / K$. Equation (13) ensures the comparability of all feature selection methods and preserves the distance between the importance of one feature and another.

In the feature aggregation step, we propose the ensemble feature importance score (FIS) that can be formulated as

$$\mathrm{FIS} = \left(\sum_{k=1}^{K} \mathrm{AUC}_k \cdot \tilde{w}_k^1, \cdots, \sum_{k=1}^{K} \mathrm{AUC}_k \cdot \tilde{w}_k^p \right), \tag{14}$$

where $\mathrm{AUC}_k \; (k = 1, 2, \cdots, K)$ represents the prediction accuracy of the *k*-th feature subset using the logistic regression classifier (8) on the test dataset. Clearly, Eq. (14) no longer just summarizes the weights of all the features to create the ensemble result but also weighs the prediction accuracy of each feature selection. Especially, the weight assigned to a feature selection method is the AUC obtained by the logistic regression, trained on the training dataset and evaluated on the test dataset. The value of FIS is a real number between 0 and 1. The closer it is to 1, the more influential the feature.

2.6. Evaluation Measure

To evaluate the performance of EFSmarker, we employ the accuracy (Acc), precision (Pre), sensitivity (Sn), specificity (Sp) and F-measure [31]. Their formula is omitted because they are well known. Specifically, the receiver operating characteristic (ROC) curve takes a false positive rate (FPR), equal to 1-Sp, as the horizontal axis and a true positive rate (TPR), equal to Sn, as the vertical axis. Area Under Curve (AUC) refers to the area under the ROC curve [31]. The closer the AUC is to 1, the higher the accuracy of the model.

3. RESULTS

3.1. Filtered Features and Classification Results

First of all, twelve filter feature selection algorithms are performed on the training dataset. In particular, the correlation method induces two feature subsets based on two kinds of measure index of feature importance; we mark them as P_cor and S_cor, respectively. Similarly, the RF method generates two feature subsets as well based on two approaches, which are noted as RF_Gini and RF_Acc. Each method obtains the feature attribute ranking in the feature filter process, *i.e.*, the weight of feature importance based on coefficient or other parameters. Consequently, we show the defined feature importance weight, the screening threshold value and the number of filtered features on the training dataset in Table **3**. In the internal validation, the logistic regression classifier shown in Equation (8) is applied respectively to train and test on the training and test datasets. The classification and prediction results characterized by AUC values are also shown in Table **3**.

As shown in Table **3**, Median, Variance, Chi2, Ridge, DT and RF_Acc get better prediction results in the internal validations, all AUC values of these six methods are higher than 0.7. While the other six methods, Relief, P_cor, S_cor, MI, mRMR and RF_Gini, perform worse, their AUC values are all lower than 0.7. In terms of average classification performance that AUC = 0.755 for statistic-based methods and AUC = 0.613 for information-based methods and AUC=0.767 for model-based methods, which shows that statistics-based and model-based are relatively better than the information-based ones. Especially in statistic-based filter methods, Median, Variance and Chi2 show advantages over Relief, P_cor and S_cor. In model-based filter methods, the performance of RF_Gini is not as good as the other three ones. However, both information-based filter methods have more disadvantages than the other two kinds of methods.

3.2. Feature Aggregation and Biomarker Discovery

In this section, we first investigate their intersection genes with nonzero weights among twelve selected feature subsets and show the results in Fig. (**2**). As mentioned

Table 3. The feature selection results on the training dataset and the classification performances on the test dataset in twelve methods.

Method	Weight	Threshold Value	# of Features	AUC
Median	1-pvalue+min(pvalue)	1	302	0.852
Variance	variannce	2.5	234	0.875
Chi2	chi2/max(chi2)	0.3	92	1.000
Relief	weight	0.8	210	0.579
P_cor	Pearson correlation coefficient	0.7	457	0.678
S_cor	Spearman correlation coefficient	0.7	457	0.666
MI	mi/max(mi)	0.5	272	0.570
mRMR	feature_count	200	200	0.655
Ridge	abs(coefficient)/max(abs(coefficient))	0.5	325	0.738
DT	score/max(score)	0.5	157	0.820
RF_Gini	abs(Gini)/max(abs(Gini))	mean	625	0.628
RF_Acc	abs(accuracy)/max(abs(accuracy))	mean	708	0.880

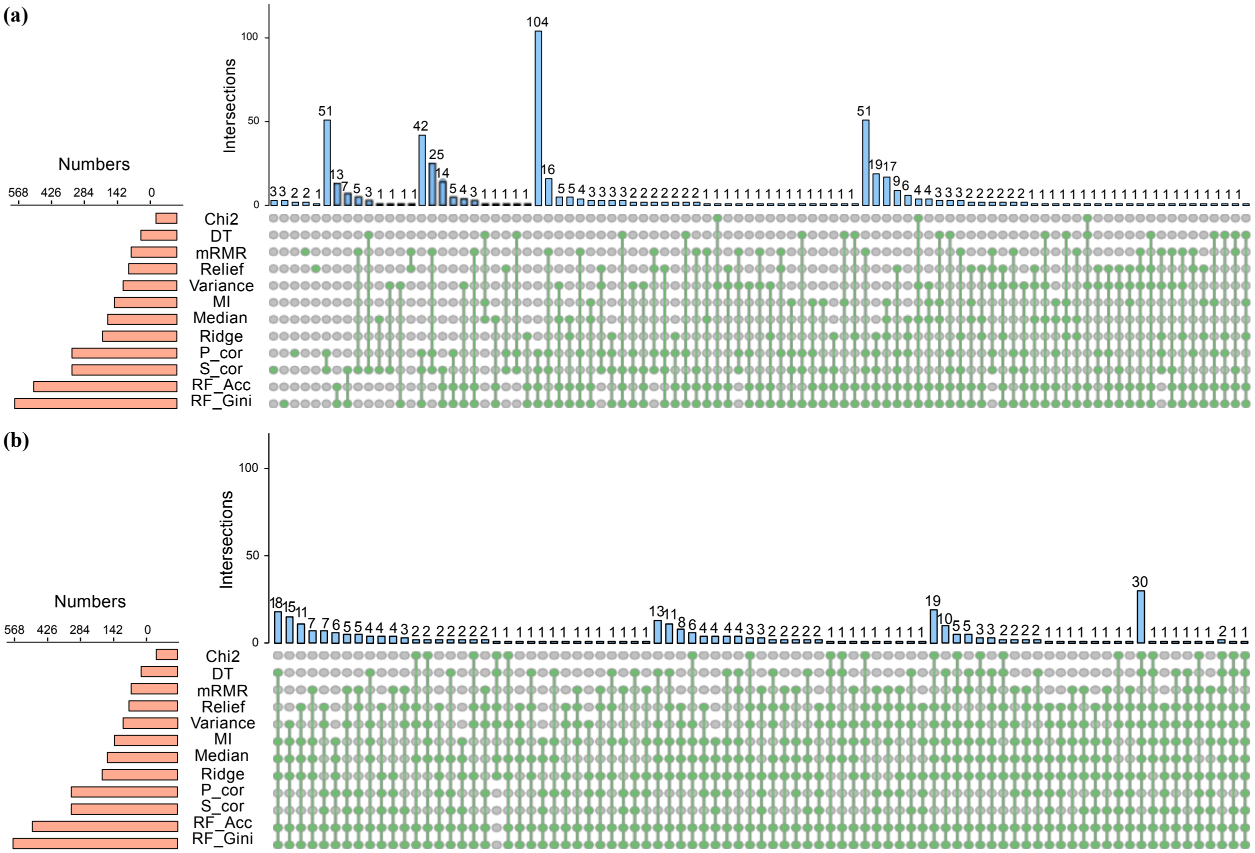

Fig. (2). The overlaps of the twelve feature sets from twelve different filter feature selection methods. (**a**). The number of overlapping genes is less than and equal to five. (**b**). The number of overlapping genes is more than and equal to six. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

before, there is a huge difference in various feature subsets filtered by different methods. It is worth noting that Fig. (**2a**) shows that there are 104 overlap genes in the feature subsets selected by the four methods P_cor, S_cor, RF_Acc and RF_Gini. If mRMR is added to these four methods, there are 51 common genes in the five feature subsets. Even so, only RF_Acc achieved satisfactory classification accuracy among these five methods. Moreover, as shown in Fig. (**2b**), the intersection genes of the nine feature subsets except for mRMR, P_cor and S_cor are up to 30, which in turn seems to indicate the consistency/correlation of the capabilities of their corresponding nine feature selection methods. So, using the EFS method to select genes containing essential and concise information by aggregating different feature subsets is necessary.

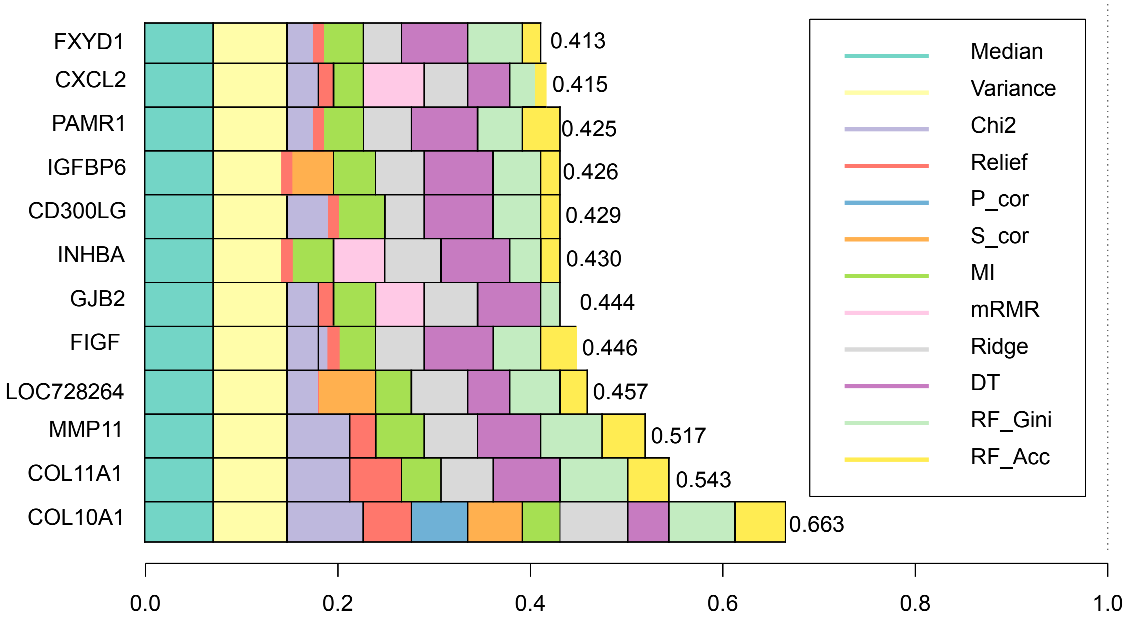

Fig. (3). The cumulative barplot of 12 BRCA biomarkers, where each feature selection method is given with a different color. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

As Equation (14) proposed, FIS archives the aggregation of weight and AUC by taking the classification accuracy as the weight of feature weight to generate a new evaluation indicator to select more important and credible features. We calculate the final FIS of all features according to Equation (14) and show the results in Supplementary Materials Table **S2**. The 12 features whose FIS was larger than 0.400 are selected as BRCA biomarkers. Fig. **(3)** shows the 12 gene symbols with corresponding FIS values, where the total length of the stacked bars represents the feature importance. For the cumulative barplot retrieved from FIS, the different colors represent different feature selection methods, and the different lengths represent the contribution ratios of varying feature selection methods. As shown in Fig. **(3)**, COL10A1 is selected by eleven methods except for mRMR and obtains the largest FIS value among all features. So, COL10A1 becomes one of the most potential biomarkers we have identified by absolute advantage.

3.3. Gene Expression Validation

We further explore the gene expression characteristics of the 12 identified biomarkers from the perspective of cluster and DEGs, according to the raw data of gene expression profiles in the discovery dataset. Fig. (**4a**) demonstrates the heatmap of 12 BRCA biomarkers, where the two-group clustering in these biomarker genes is pronounced. Clearly, all samples can be divided into two categories from left to right and the ratio of positive and negative samples is about 1:1. The results are entirely consistent with the former classification results. Fig. (**4b**) shows the boxplots of each biomarker attribute of BRCA and normal groups, where the *P*-value is calculated by Welch's *T*-test and encoded by significance signatures. Fig. (**4b**) indicates that all biomarkers are significantly differentially expressed in BRCA samples compared with those in normal samples. Besides this, Zhang *et al.* [32] found that COL10A1 is up-regulated in different breast cancer subtypes. Jia *et al.* [33] identified that FXYD1 is down-regulated in breast tumors. They are consistent with our DEGs analysis results, which provide direct evidence of the effectiveness of the proposed EFSmarker method in discovering diagnostic biomarkers of BRCA.

In addition, an online immunohistochemical study of three selected biomarkers (MMP11, GJB2 and IGFBP6) is conducted using The Human Protein Atlas database (https://www.proteinatlas.org/) and shown in Fig. (**5**). The result shows that the protein expression of MMP11 and GJB2 is up-regulated, and the protein expression of IGFBP6 is down-regulated in BRCA, which was consistent with the outcomes of DEGs expression.

3.4. Functional Enrichment Analysis

In this section, we perform the functional enrichment analysis using their gene ontology (GO) annotations to analyze the underlying functions implicated in these biomarkers. This will verify the biomarkers from the functional perspective and in turn, prove the effectiveness of the EFSmarker. Fig. (**6**) shows the top 6 enriched GO terms of biological process (BP) ordered by log *P*-value. The rest GO terms details obtained by Metascape can be found in Supplementary Materials Table **S3**. As shown in Fig. (**6**), we find these BRCA biomarkers are mainly related to extracellular matrix organization (GO:0030198), extracellular structure organization (GO:0043062), and external encapsulating structure organization (GO:0045229), all of which belong to the cellular processes. They are consistent with the prior knowledge of tumorigenesis about BRCA. For example, Lochter *et al.* [34] have established that the extracellular matrix is required

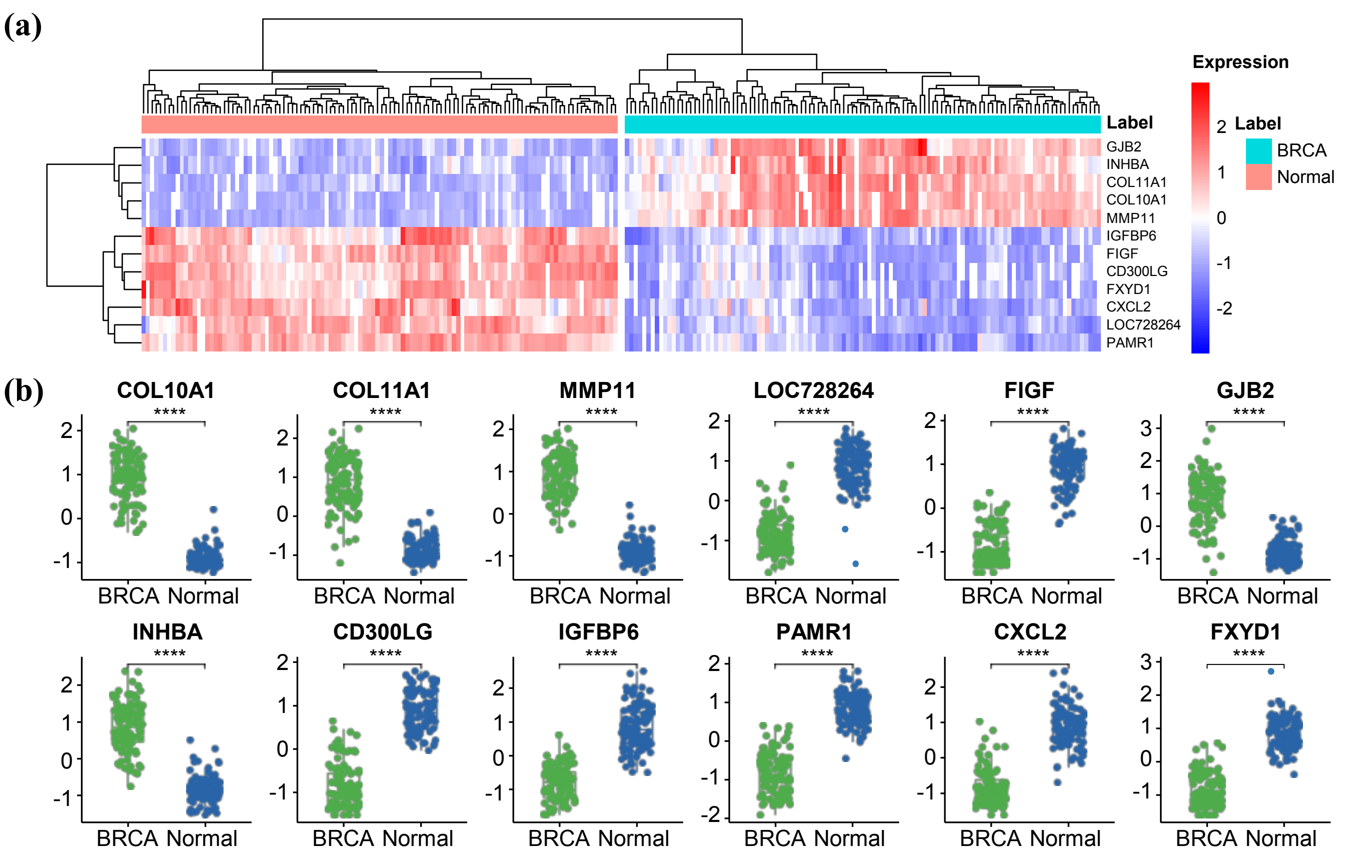

Fig. (4). The expression validation of 12 identified biomarkers of BRCA. (**a**) The heatmap of the 12 biomarkers. (**b**) The expression of the 12 biomarkers, where the significance is calculated by Welch's *T*-test with '*': $p < 0.05$, '**': $p < 0.01$, '***': $p < 0.001$ and '****': $p < 0.0001$. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

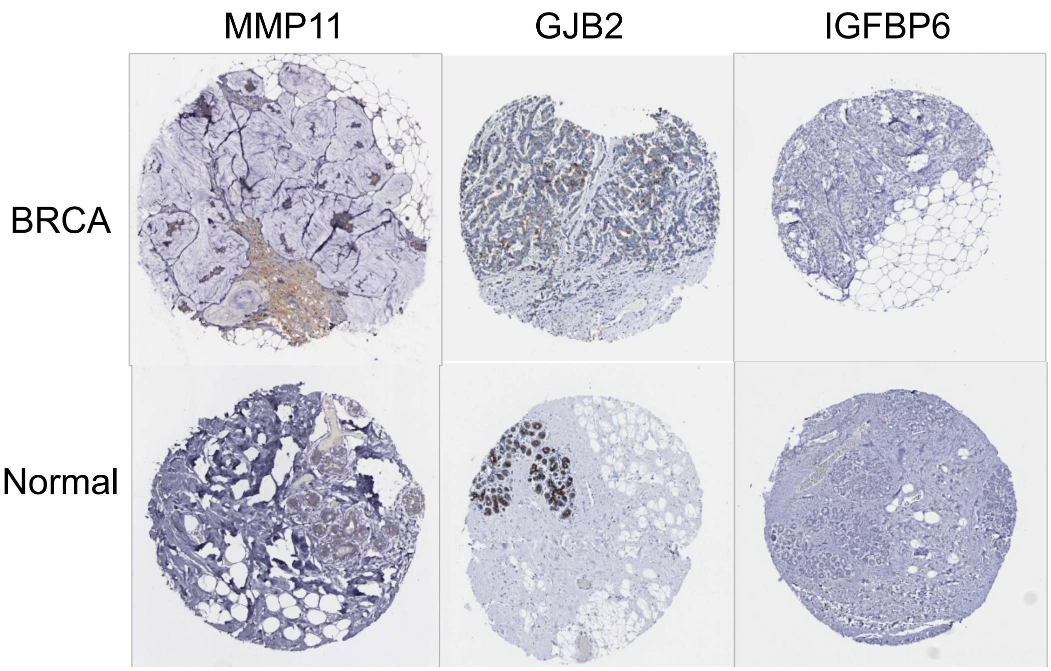

Fig. (5). Protein expression validation of selected biomarker genes in The Human Protein Atlas database. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

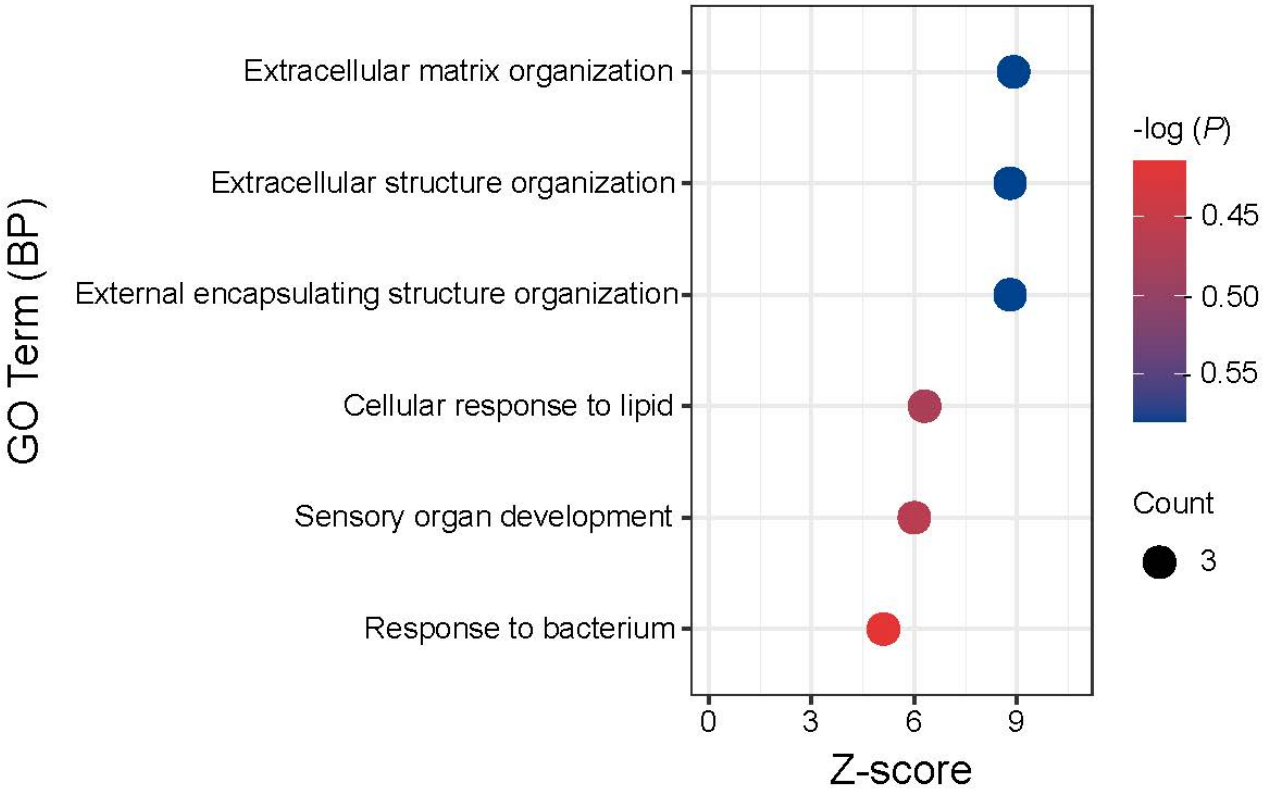

Fig. (6). The enriched GO biological processes of identified biomarkers. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

for normal functional differentiation of mammary epithelial and the extracellular matrix undergoes gross alterations during carcinogenesis. They also suggested that the extracellular matrix and extracellular matrix-receptors might participate in the control of most of the successive stages of breast tumors, from appearance to progression and metastasis [36-43]. The function enrichment results indicate that the approach of identifying biomarkers of BRCA by the EFSmarker method is effective.

3.5. Literature Checking

To further confirm the validity of the 12 BRCA biomarkers, we verify and summarize the existing literature reports, and the detailed results are illustrated in Table **4**. It demonstrates that eleven genes have been confirmed in the literature to be associated with the prevalence and prognosis of BRCA. These results prove the efficiency and creditability of the detected biomarkers by EFSmarker. Although FIGF is a gene that can not be retrieved in GeneCards, Mamoor *et al.* [35] have found and demonstrated its important role in humans with BRCA. These results prove the efficiency and creditability of discovering biomarkers by EFSmarker. Although the remaining gene LOC728264 has not been confirmed, it is likely to be used as the potential BRCA biomarker that needs further validation.

3.6. Independent Dataset Validation

At last, to verify the identified biomarkers of BRCA, we validate them in ten independent datasets from the GEO database. In each validation, we train the logistic regression classifier formulated by Equation (8) using the biomarker genes in the discovery dataset, namely TCGA BRCA data. Then, the trained classifier is used to predict the phenotypic states by the corresponding expression values of these biomarkers in the independent validation datasets, respectively. In particular, the ROC curves and classification AUC values with these 12 identified biomarkers in ten independent datasets are shown in Fig. (**7**). Totally, 688 samples have been tested. It follows that all of the AUC values are larger than 0.9. The findings prove the efficiency of our identified biomarkers in classifying BRCA samples from normal ones.

Moreover, we calculate the PRS values by Equation (9) for BRCA on ten independent validation datasets and show the results in Fig. (**8**). Clearly, there is a significant difference in the PRS indices for the BRCA samples when compared to the normal samples, in which the *P*-value is also calculated by Welch's *T*-test. Therefore, PRS provides a valuable approach to identify samples at high risk of disease from a vast number of people. It encourages a person with a higher risk value to perform further tests and interventions for an early, accurate diagnosis and treatment.

Table 4. The literature supports for the 12 screened biomarker genes of BRCA.

Gene Symbol	Entrez ID	Gene Name	Literature Verification
COL10A1	1300	Collagen Type X Alpha 1 Chain	Zhang *et al.* revealed that COL10A1 might be a predictive biomarker for breast cancer prognosis [32].
COL11A1	1301	Collagen Type XI Alpha 1 Chain	Karaglani *et al.* explored the possibility of COL11A1 variants n breast cancer tissue specimens as prognostic biomarkers [36].
MMP11	4320	Matrix Metallopeptidase 11	Eiro *et al.* proposed the evaluation of MMP11 expression by intratumoral mononuclear inflammatory cells (MICs) as a useful biological marker for breast cancer prognosis [37].
GJB2	2706	Gap Junction Protein Beta 2	Liu *et al.* indicated that GJB2 plays a vital role in the progression of ductal carcinoma *in situ* to invasive ductal carcinoma and may serve as a potential prognosis marker [38].
INHBA	3624	Inhibin Subunit Beta A	Wang *et al.* demonstrated that INHBA might be a predictor of treatment effect and prognosis of breast cancer patients in the early stage [39].
CD300LG	146894	CD300 Molecule Like Family Member G	Mamoor *et al.* revealed that CD300LG is most significantly different in primary tumors of the breast when compared to normal breast tissues and is part of the transcriptional signature of human metastatic breast cancer [40].
IGFBP6	3489	Insulin Like Growth Factor Binding Protein 6	Longhitano *et al.* highlighted that the activation of GPR81/IGFBP6 crosstalk might represent a new therapeutic target in breast cancer [41].
PAMR1	25891	Peptidase Domain Containing Associated With Muscle Regeneration 1	Lo *et al.* identified PAMR1 as a putative tumor suppressor which was frequently inactivated by promoter hypermethylation in breast cancer tissues [42].
CXCL2	2920	C-X-C Motif Chemokine Ligand 2	Pan *et al.* demonstrated that CXCL2 combined with HVJ-E suppresses tumor growth and lung metastasis in breast cancer [43].
FXYD1	5348	FXYD Domain Containing Ion Transport Regulator 1	Jia *et al.* suggested that FXYD1 is downregulated in breast tumors and may play an important role in the development of luminal A breast cancer [33].
FIGF	--	--	Mamoor *et al.* proved that molecular functions and down-regulation of FIGF may be important for the metastasis of primary tumor-derived cancer cells to the lymph nodes and the brain in humans with metastatic breast cancer [35].
LOC728264	--	--	--

Note: * No literature has been reported so far.

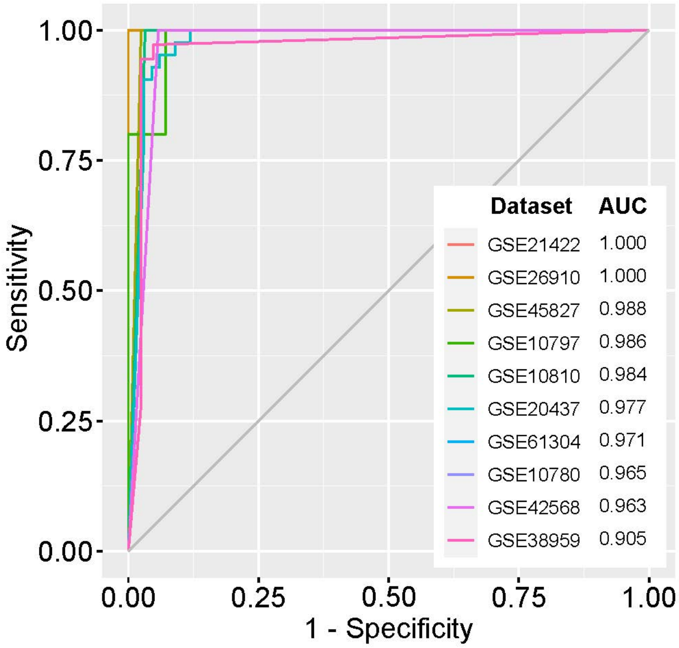

Fig. (7). The ROC curves of identified BRCA biomarkers in ten independent datasets.

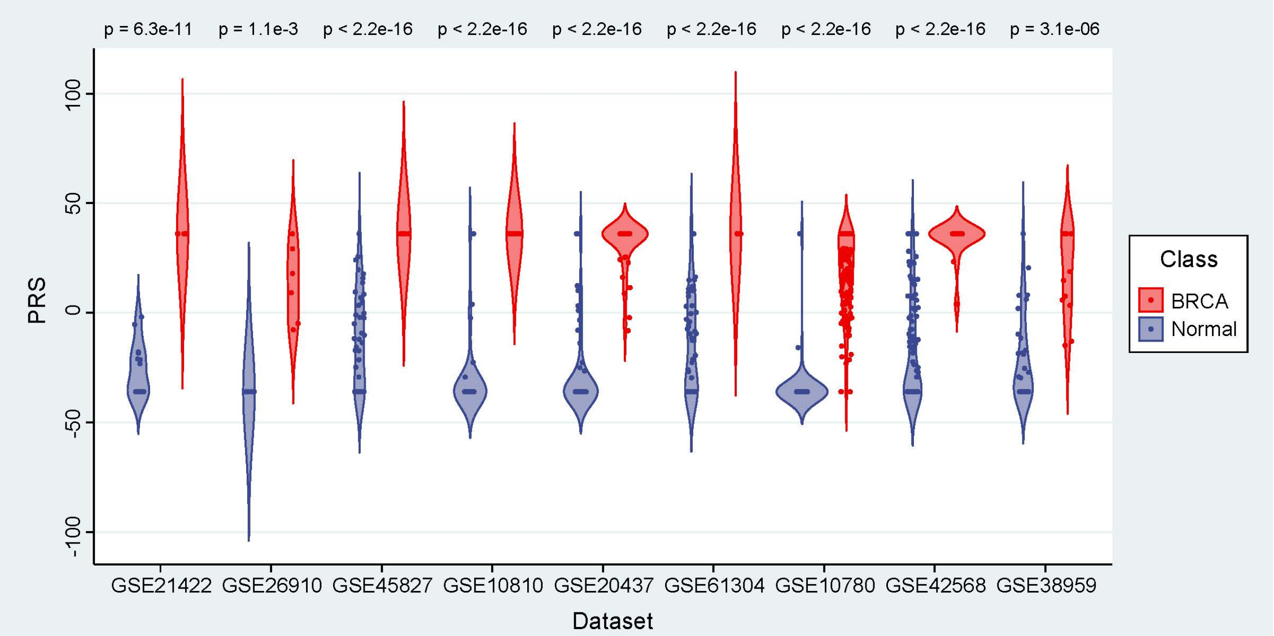

Fig. (8). The PRS significance of identified biomarkers in independent datasets for BRCA, where the *P*-values are obtained by Welch's *T*-test. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

4. DISCUSSION

In the former sections, we have reviewed some related studies employing feature selection methods to discover novel BRCA biomarkers. In some cases, researchers have identified some potential biomarkers for BRCA by using various feature selection methods. It is possible to formulate principles to assure maximum stability of the feature selection results. However, there are still some limitations, such as what criteria feature genes screened by different methods must meet to serve as accurate biomarkers of specific diseases.

We have previously accomplished related work for the first problem by introducing a stability metric [17]. In this paper, we mainly focus on the second problem. To address this problem, we proposed a bioinformatics method called EFSmarker to identify more credible and vital biomarkers of BRCA based on gene expression data and ensemble feature selection strategy. Unlike other feature aggregation methods, we do not simply take the intersection or union of multiple feature subsets to select the final expected best feature subset. Instead, we proposed the FIS index, which fully considers the classification performance of different subsets of features and the different weights of all features in each subset.

We identified the potential biomarkers of 12 genes by aggregating twelve different filter feature selection methods based on feature weight and prediction AUC value. The clustering analysis and differential expression analysis show some certain gene expression (up-regulated or down-regulated) status of these selected genes in BRCA patients, functional enrichment analysis illustrates that these biomarker genes are mainly involved in various dysfunctional cellular processes, and literature checking results clearly indicate our identified BRCA biomarkers are indeed extremely reliable and potentially valuable. Moreover, we also verified the 12 identified biomarkers on ten external independent datasets, a total of 688 samples, for classification and prediction. These biomarkers not only have potential significance for screening patients at high risk of developing BRCA but also significantly impact the early non-invasive diagnosis of BRCA in clinical practice. Obviously, the proposed EFSmarker pipeline conveniently serves as a general model for identifying diagnostic biomarkers for other complex diseases from omics data.

CONCLUSION

In this work, we proposed an ensemble feature selection method, called EFSmarker, to identify BRCA biomarkers from gene expression data. We implemented the feature selection process by combining twelve filter feature selection methods based on the differentially expressed genes and known prior gene information. Based on the discovery dataset downloaded from TCGA, we conducted the EFSmarker method and established the FIS index to select the features that were identified to the greatest extent by different methods with higher weights as potential biomarker genes. As a result, 12 genes were discovered and taken as BRCA diagnostic biomarkers, in which COL10A1 is regarded as the most potential biomarker by the highest FIS value. Finally, we validated these biomarkers through internal validations, online immunohistochemical study, functional analysis, literature checking and external verifications. In particular, existing research results showed that 11 out of 12 genes had been confirmed to be associated with the prevalence and prognosis of BRCA and the independent validation results in ten datasets where the predictive AUC values were higher than 0.9. All these justifications proved the effectiveness and efficacy of the identified biomarkers.

The biomarkers of BRCA revealed by our machine learning method can provide quantitative reference and a decision-making basis for early diagnosis. As far as we know, this work is one of the few that uses such large-scale experimental samples as the control group and the case group, which integrates multiple filtering feature selection methods to identify biomarkers of BRCA. However, this study still has limitations. For example, in real experiments, we only validated the performance of the EFSmarker on the BRCA dataset but did not test its ability to discover biomarkers for other cancers. In the feature selection part, we only used the most common filter selection methods and did not add other wrapper and embedded methods. In this paper, we only present a general framework for identifying biomarkers. Actually, our proposed EFSmarker framework can accommodate various feature selection methods, which users can replace with desired methods according to their needs. In future work, we will further improve the framework to identify more precise gene subsets as diagnostic biomarkers.

LIST OF ABBREVIATIONS

BRCA	=	Breast Cancer
EFS	=	Ensemble Feature Selection
GEO	=	Gene Expression Omnibus
DEGs	=	Differentially Expressed Genes
FIS	=	Feature Importance Score
PRS	=	Prediction Risk Score
MI	=	Mutual Information
RF	=	Random Forest
DT	=	Decision Tree
RPR	=	True Positive Rate
FPR	=	False Positive Rate
ROC	=	Receiver Operating Characteristic
AUC	=	Area Under Curve

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The authors confirm that the data supporting the findings of this study are available within the article.

FUNDING

This work was partially supported by the National Key Research and Development Program of China under grant number 2020YFA0712402; the National Natural Science Foundation of China (NSFC) under grant number 61973190, Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) under grant number 2019JZZY010423, the Natural Science Foundation of Shandong Province of China (Grant no. ZR2020ZD25), the Fundamental Research Funds for the Central Universities (Grant no. 2022JC008), the Program of Qilu Young Scholars of Shandong University and the Scholarship under Shandong University's Exchange Program.

CONFLICT OF INTEREST

Dr. Zhi-Ping Liu is the Editorial Advisory Board Member for the journal CBIO.

ACKNOWLEDGEMENTS

The authors would like to thank the editors and anonymous reviewers for their valuable comments and suggestions which greatly improved our paper. The authors also thank the members of our lab at the Shandong University and the AMACL lab members at the University of Hong Kong for their assistance in the project.

SUPPLEMENTARY MATERIAL

The codes and Supplementary Materials related to this article can be found online at https://github.com/zpliulab/EFSmarker.

REFERENCES

[1] Huang H, Hu J, Maryam A, *et al.* Defining super-enhancer landscape in triple-negative breast cancer by multiomic profiling. Nat Commun 2021; 12(1): 2242. http://dx.doi.org/10.1038/s41467-021-22445-0 PMID: 33854062

[2] Zarotti C, Papassotiropoulos B, Elfgen C, *et al.* Biomarker dynamics and prognosis in breast cancer after neoadjuvant chemotherapy. Sci Rep 2022; 12(1): 91. http://dx.doi.org/10.1038/s41598-021-04032-x PMID: 34997055

[3] Li L, Liu ZP. Detecting prognostic biomarkers of breast cancer by regularized Cox proportional hazards models. J Transl Med 2021; 19(1): 514. http://dx.doi.org/10.1186/s12967-021-03180-y PMID: 34930307

[4] Rajkumar T, Amritha S, Sridevi V, *et al.* Identification and validation of plasma biomarkers for diagnosis of breast cancer in South Asian women. Sci Rep 2022; 12(1): 100. http://dx.doi.org/10.1038/s41598-021-04176-w PMID: 34997107

[5] El Bairi K, Haynes HR, Blackley E, *et al.* The tale of TILs in breast cancer: A report from the international immuno-oncology biomarker working group. NPJ Breast Cancer 2021; 7(1): 150. http://dx.doi.org/10.1038/s41523-021-00346-1 PMID: 34853355

[6] Li L, Liu Z. A connected network-regularized logistic regression model for feature selection. Appl Intell 2022; 52: 1-31. http://dx.doi.org/10.1007/s10489-021-02377-4

[7] Li L, Liu ZP. Biomarker discovery for predicting spontaneous preterm birth from gene expression data by regularized logistic regression. Comput Struct Biotechnol J 2020; 18: 3434-46. http://dx.doi.org/10.1016/j.csbj.2020.10.028 PMID: 33294138

[8] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947; 18(1): 50-60.
http://dx.doi.org/10.1214/aoms/1177730491

[9] Dai YH, Wang YF, Shen PC, *et al.* Radiosensitivity index emerges as a potential biomarker for combined radiotherapy and immunotherapy. NPJ Genom Med 2021; 6(1): 40.
http://dx.doi.org/10.1038/s41525-021-00200-0 PMID: 34078917

[10] Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dublin Philos Mag J Sci 1900; 50(302): 157-75.
http://dx.doi.org/10.1080/14786440009463897

[11] Kononenko I. Estimating attributes: Analysis and extensions of relief. European conference on machine learning. In. European conference on machine learning. Berlin: Springer, 1994, p.171-82.

[12] Zuber V, Strimmer K. Gene ranking and biomarker discovery under correlation. Bioinformatics 2009; 25(20): 2700-7.
http://dx.doi.org/10.1093/bioinformatics/btp460 PMID: 19648135

[13] Wang Y, Liu ZP. Identifying biomarkers for breast cancer by gene regulatory network rewiring. BMC Bioinformatics 2022; 22(12): 308.
PMID: 35045805

[14] De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: An R package for parallelized mRMR ensemble feature selection. Bioinformatics 2013; 29(18): 2365-8.
http://dx.doi.org/10.1093/bioinformatics/btt383 PMID: 23825369

[15] Zhang Z, Liu ZP. Robust biomarker discovery for hepatocellular carcinoma from high-throughput data by multiple feature selection methods. BMC Med Genomics 2021; 14(S1): 112.
http://dx.doi.org/10.1186/s12920-021-00957-4 PMID: 34433487

[16] Ben Brahim A, Limam M. Ensemble feature selection for high dimensional data: A new method and a comparative study. Adv Data Anal Classif 2018; 12(4): 937-52.
http://dx.doi.org/10.1007/s11634-017-0285-y

[17] Li L, Ching WK, Liu ZP. Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods. Comput Biol Chem 2022; 100: 107747.
http://dx.doi.org/10.1016/j.compbiolchem.2022.107747 PMID: 35932551

[18] Mera-Gaona M, López DM, Vargas-Canas R, Neumann U. Framework for the ensemble of feature selection methods. Appl Sci 2021; 11(17): 8122.
http://dx.doi.org/10.3390/app11178122

[19] Chiew KL, Tan CL, Wong K, Yong KSC, Tiong WK. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Inf Sci 2019; 484: 153-66.
http://dx.doi.org/10.1016/j.ins.2019.01.064

[20] Wang J, Xu J, Zhao C, Peng Y, Wang H. An ensemble feature selection method for high-dimensional data based on sort aggregation. Syst Sci Control Eng 2019; 7(2): 32-9.
http://dx.doi.org/10.1080/21642583.2019.1620658

[21] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 2010; 26(3): 392-8.
http://dx.doi.org/10.1093/bioinformatics/btp630 PMID: 19942583

[22] Zhao S, Zhang Y, Xu H, Han T. Ensemble classification based on feature selection for environmental sound recognition. Math Prob Eng 2019; 2019
http://dx.doi.org/10.1155/2019/4318463

[23] Awada W, Khoshgoftaar TM, Dittman D, Wald R, Napolitano A. A review of the stability of feature selection techniques for bioinformatics data. In. IEEE 13th International Conference on Information Reuse & Integration (IRI) 2013. p.356-63.

[24] Cheng LH, Hsu TC, Lin C. Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction. Sci Rep 2021; 11(1): 14914.
http://dx.doi.org/10.1038/s41598-021-92864-y PMID: 34290286

[25] Dittman DJ, Khoshgoftaar TM, Wald R, Napolitano A. Comparing two new gene selection ensemble approaches with the commonly-used approach. In. 11th International Conference on Machine Learning and Applications. Boca Raton, FL, USA: IEEE 2012; p. 184-91.

[26] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15(12): 550.
http://dx.doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

[27] Rappaport N, Twik M, Plaschkes I, *et al.* MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 2017; 45(D1): D877-87.
http://dx.doi.org/10.1093/nar/gkw1012 PMID: 27899610

[28] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 2021; 49(D1): D545-51.
http://dx.doi.org/10.1093/nar/gkaa970 PMID: 33125081

[29] Cardoso F, van't Veer LJ, Bogaerts J, *et al.* 70-gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med 2016; 375(8): 717-29.
http://dx.doi.org/10.1056/NEJMoa1602253 PMID: 27557300

[30] Liu Z, Wu C, Miao H, Wu H. RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015; 2015
http://dx.doi.org/10.1093/database/bav095

[31] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997; 30(7): 1145-59.
http://dx.doi.org/10.1016/S0031-3203(96)00142-2

[32] Zhang M, Chen H, Wang M, Bai F, Wu K. Bioinformatics analysis of prognostic significance of COL10A1 in breast cancer. Biosci Rep 2020; 40(2): BSR20193286.
http://dx.doi.org/10.1042/BSR20193286 PMID: 32043519

[33] Jia X, Lei H, Jiang X, *et al.* Identification of crucial lncRNAs for Luminal A breast cancer through RNA sequencing. Int J Endocrinol 2022; 2022: 6577942.
http://dx.doi.org/10.1155/2022/6577942

[34] Lochter A, Bissell MJ. Involvement of extracellular matrix constituents in breast cancer. Semin Cancer Biol 1995; 6(3): 165-73.
http://dx.doi.org/10.1006/scbi.1995.0017 PMID: 7495985

[35] Mamoor S. Vascular endothelial growth factor D, VEGF-D, encoded by FIGF is differentially expressed in metastatic breast cancer, both in metastases to the brain and to the lymph nodes. OSF Preprint 2020.

[36] Karaglani M, Toumpoulis I, Goutas N, *et al.* Development of novel real-time PCR methodology for quantification of COL11A1 mRNA variants and evaluation in breast cancer tissue specimens. BMC Cancer 2015; 15(1): 694.
http://dx.doi.org/10.1186/s12885-015-1725-8 PMID: 26466668

[37] Eiro N, Cid S, Fernández B, *et al.* MMP11 expression in intratumoral inflammatory cells in breast cancer. Histopathology 2019; 75(6): 916-30.
http://dx.doi.org/10.1111/his.13956 PMID: 31342542

[38] Liu Y, Pandey PR, Sharma S, *et al.* ID2 and GJB2 promote early-stage breast cancer progression by regulating cancer stemness. Breast Cancer Res Treat 2019; 175(1): 77-90.
http://dx.doi.org/10.1007/s10549-018-05126-3 PMID: 30725231

[39] Wang XQ, Liu B, Li BY, Wang T, Chen DQ. Effect of CTCs and INHBA level on the effect and prognosis of different treatment methods for patients with early breast cancer. Eur Rev Med Pharmacol Sci 2020; 24(24): 12735-40.
PMID: 33378021

[40] Mamoor S. CD300LG (Nepmucin) is differentially expressed in brain metastatic breast cancer. OSF Preprint 2020.

[41] Longhitano L, Forte S, Orlando L, *et al.* The crosstalk between GPR81/IGFBP6 promotes breast cancer progression by modulating lactate metabolism and oxidative stress. Antioxidants 2022; 11(2): 275.
http://dx.doi.org/10.3390/antiox11020275 PMID: 35204157

[42] Lo PHY, Tanikawa C, Katagiri T, Nakamura Y, Matsuda K. Identification of novel epigenetically inactivated gene PAMR1 in breast carcinoma. Oncol Rep 2015; 33(1): 267-73.
http://dx.doi.org/10.3892/or.2014.3581 PMID: 25370079

[43] Pan YC, Nishikawa T, Chang CY, Tai JA, Kaneda Y. CXCL2 combined with HVJ-E suppresses tumor growth and lung metastasis in breast cancer and enhances anti-PD-1 antibody therapy. Mol Ther Oncolytics 2021; 20: 175-86.
http://dx.doi.org/10.1016/j.omto.2020.12.011 PMID: 33575480

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.