Feature Selection for Breast Cancer Classification by Integrating Somatic Mutation and Gene Expression

Qin Jiang and Min Jin*

College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

Exploring the molecular mechanisms of breast cancer is essential for the early prediction, diagnosis, and treatment of cancer patients. The large scale of data obtained from the high-throughput sequencing technology makes it difficult to identify the driver mutations and a minimal optimal set of genes that are critical to the classification of cancer. In this study, we propose a novel method without any prior information to identify mutated genes associated with breast cancer. For the somatic mutation data, it is processed to a mutated matrix, from which the mutation frequency of each gene can be obtained. By setting a reasonable threshold for the mutation frequency, a mutated gene set is filtered from the mutated matrix. For the gene expression data, it is used to generate the gene expression matrix, while the mutated gene set is mapped onto the matrix to construct a co-expression profile. In the stage of feature selection, we propose a staged feature selection algorithm, using fold change, false discovery rate to select differentially expressed genes, mutual information to remove the irrelevant and redundant features, and the embedded method based on gradient boosting decision tree with Bayesian optimization to obtain an optimal model. In the stage of evaluation, we propose a weighted metric to modify the traditional accuracy to solve the sample imbalance problem. We apply the proposed method to The Cancer Genome Atlas breast cancer data and identify a mutated gene set, among which the implicated genes are oncogenes or tumor suppressors previously reported to be associated with carcinogenesis. As a comparison with the integrative network, we also perform the optimal model on the individual gene expression and the gold standard PMA50. The results show that the integrative network outperforms the gene expression and PMA50 in the average of most metrics, which indicate the effectiveness of our proposed method by integrating multiple data sources, and can discover the associated mutated genes in breast cancer.

Keywords: breast cancer, machine learning, classification, feature selection, gradient boosted decision tree

INTRODUCTION

Breast cancer is considered to be the most prevalent cancer among women and the second common cause of death in both developed and undeveloped countries. It is caused by multiple factors including genomic, transcriptomic, and epigenomic involvement in its formation and development. With the development of technology, understanding the pathogenesis of cancer from
the perspective of molecular contributes to effective diagnosis and treatment. The large-scale cancer genomics project, The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), has produced a large volume of data, providing ways to explore cancer formation and progression.

In general, the cancer transcriptome contains gene expression, including messenger RNA (mRNA), long non-coding RNA (lncRNA), and microRNA (miRNA). Previous studies focused on utilizing the gene expression profile to successfully diagnose individuals based on the differential gene expression (Li et al., 2017) and other clinically relevant phenotypes. Meanwhile, the cancer genome contains many mutations. Among them, one of the most important is somatic mutations, which include single-nucleotide variant (SNVs) and small insertions and deletions (indels). Some mutations that contribute to cancer progression from normal to malignant are called driver mutations, and others that accumulate in cells but do not contribute to cancer development are called passengers (Bozic et al., 2010).

Distinguishing driver mutations from the passengers that have no critical effect on cancer cells is a crucial step and challenging task in understanding the molecular mechanisms of cancer, which can guide effective treatment and prognosis for cancer patients and promote the development of targeted drugs. In earlier studies, researchers focused on detecting driver genes that cause tumors (Merid et al., 2014). A common approach is to identify driver genes by detecting positive signals in tumors. Because of the complexity of the cancer genome, driver genes contain not only driver mutations but also passenger mutations. This makes this kind of approach sometimes ineffective.

On the other hand, studies have shown that somatic mutations frequently perturb the expression level of affected genes and thus disrupt the pathways controlling normal growth (Kwong et al., 2020). For example, mRNAs carrying a premature stop codon, which can be introduced by truncation mutations, are typically eliminated by the process called nonsense-mediated mRNA decay, and thus, both the concentration of mRNA transcripts and protein products would be decreased owing to truncation mutations (Jia and Zhao, 2016). Considering the association between the somatic mutation and gene expression, several studies have emphasized the necessity of integrating both types of data to identify candidate driver genes (Masica and Karchin, 2011; Zhang and Wang, 2020). For cancer analysis, many researchers construct a co-expression network by integrating different types of data. He et al. (2017) and Wu et al. (2019) utilized the network by integrating somatic mutation with gene expression to identify the type of cancers and cancer subtypes. Mamidi et al. (2019) integrated germline and somatic mutation to discover biomarkers in triple-negative breast cancer and identified the molecular networks and biological pathways.

As the molecular network has been verified to be effective for the biological discovery of cancers, current studies utilized the network across different types of cancer or cancer subtypes. However, the objective of most researches is the universality of the methods, which makes it difficult to be equally effective in all disease types. In this study, we aim to construct an efficient method of architecture for the diagnosis of breast cancer based on the network of somatic mutation and gene expression. We are focused not only on finding more biomarkers but also on the classification performance of the model. First, the somatic mutation is used to generate a binary mutation network; similarly, an expression network is obtained from the gene expression profiles. Then, for the expression network, we compute both the observed $p$-value and the adjusted $p$-value to correct for multiple-hypothesis testing (false discovery rate, FDR) and thus obtain the differential expression network. Meanwhile, an integrative network is constructed by combining the mutation network and the differential expression network. Thirdly, we rank the genes in the integrative network by mutual information (MI) and select the top 50 genes, which are highly correlated with breast cancer. Finally, we use the Bayesian optimization method to optimize the classification model, gradient boosting decision tree (GBDT), which is further applied to assess the features selected from the previous step. In terms of evaluation metrics, the traditional metric of accuracy does not consider the sample imbalance, so we propose a simple and effective metric, balanced accuracy, to reveal the ability of the different model to classify positive and negative samples.

### MATERIALS AND METHODS

We used statistical and machine learning methods to develop this novel method for feature selection and classification, including the preprocessing of data, filter method, and embedded method for feature selection, processing of imbalanced data, and the final classification model. Figure 1 shows the flowchart of the proposed method.

### Dataset Construction and Preprocessing

In this research, we use publicly available breast cancer datasets (BRCA) from TCGA, including transcriptome gene expression and somatic mutation. Considering the different structures of these two types of data, we used different methods to preprocess them. Table 1 shows the numbers of samples and features for the two datasets.

The BRCA gene expression dataset comprises 1222 samples and 57,063 genes. There are 113 normal samples and 1109 tumor samples. We used the edgeR package to filter the genes expressed in small amounts in most samples and normalized the data. The gene expression data was reduced from 57,063 to 34,465 by deleting the genes expressed in small amounts in most samples.

The somatic mutation data comes from the simple nucleotide variation (SNV) in the TCGA-BRCA project. The data file includes SNP, INS, and DEL, three types of mutations. The important fields in the data file are Hugo_Symbol (gene name), Variant_Type, and Tumor_Sample_Barcode (sample name). Statistically, the somatic mutation data contains 18,127 genes and 986 samples. To get the mutation frequency of each gene in all samples, we use a Perl script to process the data file. For example, if gene A is present in sample S, that means sample S has a mutation in gene A, then we code it as “1,” otherwise we code

---

**Table 1**

| Dataset Type       | Number of Samples | Number of Genes |
|--------------------|-------------------|-----------------|
| Gene Expression    | 1222              | 57,063          |
| Somatic Mutation   | 986               | 18,127          |

---
FIGURE 1 | The flowchart of the proposed method. From the somatic mutation matrix, the mutation frequency of each gene is obtained to select the highly mutated genes, which will be integrated by mapping into the gene expression matrix to get the integrative network. After FDR, FC, and mutual information ranking, the feature genes serve as the input to GBDT. Then, the optimal model is obtained by Bayesian optimization. During the training process, the optimal gene subset is obtained simultaneously.

it as “0.” Supplementary Table 1 shows the coding schedule of all genes in samples. Given the sample set \( S = \{s_1, s_2, \ldots, s_n\} \), \( n \) is the total number of samples, and \( s_i \) represents the sample \( i \). Gene set \( G = \{g_1, g_2, \ldots, g_m\} \), \( m \) is the total number of mutation genes, and \( g_j \) represents the gene \( j \). In the set of sample number \( C = \{c_1, c_2, \ldots, c_m\} \), \( c_k \) represents the number of samples with “1” in gene \( k \). The set \( C \) can be calculated by the number of “1” in each row in Supplementary Table 1.

According to Supplementary Table 1 and set \( C \), we can obtain the frequency of mutations across patients to assess the percentage of patients carrying a particular mutation in each mutated gene. To further reduce the interference of genes with low mutation rates, we set the threshold \( p \) as the percentage of the total samples to select the genes with high mutation frequency. The selected gene set constitutes the mutation network. In the experiment, we compare the effects of different \( p \) on classification accuracy by the proposed model, and the result is shown in Supplementary Table 2. Due to the highest accuracy 97.31% obtained by setting the threshold \( p \) as 1%, we apply this value in the proposed method.
The Way to Combine Somatic Mutation and Gene Expression

Somatic mutations in cancer genomes frequently perturb the expression level of affected genes. Then, the pathways controlling normal growth are disrupted (Zhang et al., 2013). Similarly, the research by Ding et al. (2015) assessed the impact of mutations on gene expression as a means of quantifying potential phenotypic effects and for novel cancer gene discovery. Fleck et al. (2016) addressed the issue of cancer heterogeneity by using both somatic mutation and gene expression data and proposed a formulation to model the molecular progression of cancer. They discovered that the progression of the disease was reflected in both the accumulation of mutations and changes in gene expression levels.

Further study (Jia and Zhao, 2016) focused on the functional footprints of somatic mutations in 12 cancer types and grouped the mutations by mutation type, cluster, and status. This study unraveled the effects of somatic mutation features on mRNA and protein expression.

Our study is based on the assumption that mutations may cause changes in the cell's state, such as underexpression or overexpression of different genes. Then, we combine the somatic mutation network with the gene expression network to obtain an integrative network. In the integrative process of the two types of networks, we refer to the gene expression network to obtain the expression value of the somatic mutation genes in the mutation network. It is important to note that in the subsequent classification task, the normal samples in the expression network are added as the control group.

Fold Change and False Discovery Rate

Fold change (FC) is used to calculate the differential multiples of gene expression values between cancer samples and normal samples, which is the basic method for detecting differential genes, and represents the expression values of feature i and sample j in cancer samples and normal samples; FC is defined as:

\[ FC_i = \frac{\bar{X}_i}{\bar{Y}_i} \]  

(1)

When FC exceeds the initial set threshold, it can be considered that the feature is different, and it is generally considered that there is a significant difference when the difference multiple is more than 2. FC can directly obtain the differentially expressed values, but in the absence of false-positive control, the rate of false-positive results is relatively high.

According to statistical theory, in multiple-hypothesis testing, it is important to control the probability of making mistakes in multiple statistical inferences, called FDR. FDR can be used to analyze differentially expressed genes to control the proportion of false positives (Reiner-Benaim, 2010). Table 2 shows the confusion matrix for the statistical test. FDR can be defined as follows:

\[ FDR = E\left(\frac{V}{V + S}\right) = E\left(\frac{V}{R}\right) \quad (R > 0). \]

(2)

The number of false positives in multiple-hypothesis tests can be controlled by controlling that FDR is below the threshold \( q \). In general, keep FDR below 0.01, or ensure that there is at most one false positive for every 100 positive hypotheses. Feature genes with significant differences can be identified by FC and FDR, but these two methods do not evaluate the classification performance of these features.

Fold change and FDR are applied to integrative data to select the differentially expressed genes. By comparing the classification balanced accuracy under different FC and FDR thresholds shown in Supplementary Tables 3, 4, the optimal value of FC and FDR thresholds is obtained: log (FC) > 1.0, FDR < 0.05.

Mutual Information

Mutual information (Bonev et al., 2008) is a useful measure of information in information theory and is a kind of filter method. It refers to the correlation between two events set. The datasets consist of tens of thousands of gene columns and one label column. The gene column is defined as \( G_i \), and the label column is defined as \( L \). \( MI(G_i, L) \) is represented as the MI between the gene \( G_i \) and the label \( L \). The calculation equation is Eq. 3.

\[ MI (G_i, L) = H (G_i) + H (L) - H(G_i, L) \]  

(3)

\( H (G_i) \) is the information entropy of the gene column \( G_i \), \( H (L) \) is the information entropy of the label \( L \), and \( H(G_i, L) \) is the joint information entropy of \( G_i \) and \( L \). According to information theory, the information entropy is a measure of the uncertainty of a random variable. Suppose X is a random variable, and the range of possible values is \( S_x \), \( x \in S_x \) and the probability is \( p(x) \); the information entropy of X is defined as:

\[ H (X) = - \sum_{x \in S_x} p (x) \log p(x) \]  

(4)

| Parameter | p | FDR | |log(FC)| | M |
|-----------|----|-----|----------------|--------------|
| Threshold | 1% | 0.05 | 1 | 50 |

p is the percentage of the total samples, which represents the mutation frequency of a certain gene. FDR is the false discovery rate, and FC is the fold change. M is the number of genes that top ranking in mutual information.
$H(X, Y)$ is the joint information entropy, defined as:

$$H(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$ (5)

$p(x, y)$ is the joint probability density function. $MI(G_i, L)$ can be calculated according to Eqs 4 and 5. In our study, MI is used to measure the dependency between a feature and the classification type. In general, the greater value of MI indicates that the feature contains more information for classification. Therefore, we rank the MI values of each feature and selected the top M features from the integrative data, respectively. The final objective of this method is to remove irrelevant features to reduce the dimension of integrative data. We set different values of M to compare the classification-balanced accuracy and obtain the best value of M. The result in Supplementary Table 5 shows that the optimal M is 50. Table 3 shows the main parameters applied in the proposed method.

**GBDT With Bayesian Optimization**

The filter methods obtain a feature subset for which the discriminative capability is limited for classification purposes. Embedded methods can be used to search the optimal feature subset by a given classifier. In the training procedure, the features with high importance can be selected by ranking and the classification algorithm is optimized simultaneously. It is helpful to build a strong link between the feature subset and the classifier. The GBDT is an ensemble learning algorithm based on GBM, which is proposed by Friedman (Friedman, 2001). During the training process, multiple iterations are used to build multiple trees to make joint decisions. When the square error loss function is adopted, each regression tree learns the conclusions and residuals of all previous trees, and a current residual regression tree is obtained by the fitting. The meaning of residuals is as follows:

$$residuals = true \ value - predict \ value$$

The boosting tree (Galicia et al., 2018) is an accumulation of regression trees generated during the entire iteration process. The optimization process of learning is realized by using an additive model and a forward step algorithm. The GBDT was used in our study because of its flexibility for different types of data, excellent classification performance, and robustness for abnormal values.

However, it is tedious and important work to tune the hyperparameters when conducting the GBDT, because it greatly affects the performance of the algorithm. Manual tuning is time-consuming; grid and random searches (Bhat et al., 2018) require no human effort but a long-running time. Therefore, in this research, Bayesian optimization is adopted to find the optimal hyperparameters, which is first proposed by Snoek et al. (2012). Bayesian optimization seeks to minimize the value of the objective function by establishing an alternative function based on the objective function's past evaluation results. The Bayesian method is different from random or grid searches as they consider previous estimates when testing the next set of hyperparameters, thus saving a lot of effort.

Suppose hyperparameters set (represents a hyperparameter's value), the relationship between this set, and the loss function that need to be optimized, defined as $f(X)$. However, machine learning just likes a black box, which means we only know the input and output; $f$ is hard to be sure. So we should turn our attention to a function that can be solved. Assume function, we need to find in:

$$x^* = \arg \min_{x \in A} f(x)$$ (6)

Here, we chose Hyperopt in Python library, which adopted Tree Parzen Estimator (TPE), which used the Gaussian Mixture Model (Oh et al., 2019) to learn hyperparameters. First, we split the integrative dataset into 80% learning set and 20% test set then divided the learning set into 60% training set and 40% validation set. The performance of hyperparameters was evaluated on the validation set. The Bayesian optimization assigned a greater probability to the value of the hyperparameters set with a lower loss in the cross-validation. Finally, the best hyperparameters set was output.

**A Weighted Metric for Imbalanced Dataset**

Class imbalance is a situation in which the number of training samples of different categories varies greatly in the classification task. There are many strategies to deal with the imbalance problem, such as undersampling and oversampling. EasyEnsemble is a method of undersampling, proposed by Li and Liu (2014). Multiple different training sets are generated by putting back the samples several times, and then multiple different classifiers are trained. The final result is obtained by combining the results of multiple classifiers. Another method is BalanceCascade (Liu et al., 2009), which adopts the idea of Boosting. It also uses undersampling to generate a training set, but those correctly classified samples are not put back. Undersampling is easy to lose information, and the way the final result is integrated also has an impact. The most common strategy for oversampling is SMOTE (Synthetic Minority Oversampling Technique) (Blagus and Lusa, 2013). In this method, the new samples are synthesized according to the nearest neighbor in the minority samples and then added into the dataset. However, there are two main problems in this algorithm: there is some blindness in the selection of the nearest neighbor and the problem of distribution marginalization is easy to occur. Additionally, undersampling and oversampling may change the distribution.
of data. For the task of cancer classification, the size of sample is small, more than a thousand at most, and these strategies do not seem appropriate. Therefore, in this study, we propose a weighted metric to modify the traditional accuracy metric instead of changing the distribution of the dataset. There are far more cancer samples than normal samples, which will lead to the high accuracy of the learning method if it returns a learning model that always predicts the new sample as a cancer category. To solve this problem, we separated the total sample set into a normal set and tumor set. The classification accuracy of the model in the two sample space embodies the model’s ability to correctly classify the positive and negative samples, named the weight for the two sample spaces. On the final test stage, we multiply this weight with the accuracy of two sample spaces on the test set.

Let \( N \) and \( T \) denote the sample set of normal class and that of tumor class, respectively. \( \bar{w}_n \) and \( \bar{w}_t \) are the accuracy of normal samples and tumor samples of classifier \( \text{clf} \) in the validation set, respectively. These two weights represent the different capacities of the given classifier for different types of samples. In the final testing stage, the optimized GBDT is conducted as the classifier to predict the independent test set. \( \bar{w}_n \) and \( \bar{w}_t \) will be considered in the final decision. As we split the dataset into 10 equal-sized datasets, \( \bar{w}_n \) and \( \bar{w}_t \) are the average accuracy of the 10 validation sets. Here, the average accuracy of normal samples and tumor samples on the 10 test sets are represented by \( \text{acc}_n \) and \( \text{acc}_t \). So the final balanced accuracy is defined as:

\[
\text{balanced acc} = \text{acc}_n \cdot \bar{w}_n + \text{acc}_t \cdot \bar{w}_t
\]

The core procedure of calculating the weighted metric for balanced accuracy is described in Figure 2. The weighted metric for the imbalanced dataset is easy to operate. It considers the classification ability of the classifier on samples of different categories and further revises the final test results by multiply weights, thus reducing the impact of class imbalance.

**Evaluation Criteria**

The following metrics are used to evaluate the performance of the classification model in this study:

- **Accuracy**: \( \text{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \)
- **Sensitivity**: \( \text{SES} = \frac{TP}{TP + FN} \)
- **Specificity**: \( \text{SPC} = \frac{TN}{TN + FP} \)
- **Precision**: \( \text{PRC} = \frac{TP}{TP + FP} \)
- **\( F_1 \) score**: \( F_1 = \frac{2TP}{2TP + FP + FN} \)

In this study, the tumor sample is positive, and the normal sample is the negative sample, where TP (true positive) is the number of tumor samples predicted as tumor, FP (false positive) is the number of tumor samples predicted as normal, TN (true negative) is the number of tumor samples normal and predicted as normal, and FN (false negative) is the number of normal samples and predicted as tumor. Meanwhile, the AUC is obtained.

Due to that the number of samples is much smaller than that of the features, in this study, first, we split the dataset into 10 equal-sized datasets. Then, we divide the datasets into 80% learning set and 20% test set and ensure that the test set does not participate in any training process (Meng et al., 2020). Finally, the independent test set is used to calculate the above evaluation metrics. This procedure is repeated on the 10 datasets. The average of the results generated on the 10 datasets is used as the final performance of the proposed model on the test set.

**RESULTS**

**Classification Results of Proposed Method SFS**

In our experiments, the training set is used to train the classifier. The obtained parameters are verified on the validation set. In addition, we calculate \( \bar{w}_n \) and \( \bar{w}_t \) (normal samples’ accuracy and tumor samples’ accuracy in the validation set). Moreover, balanced accuracy was calculated by Eq. 6. The proposed method adopts FC, FDR, MI, and GBDT with Bayesian optimization. The parameters are applied as follows:

1. FC: \(|\log(FC)| > 1.0\)
2. FDR: FDR 0.05
3. MI: select the top 50 features of MI value ranking
4. Bayesian optimization: tuning the parameters of GBDT with Bayesian optimization using the 50 features to get the optimal model.

These methods are combined in the ways shown in Table 4. Case 1: None of the above methods are used. Case 2: FC and FDR are used to obtain the differentially expressed genes. Case 3: FC + FDR, MI are used to select informative features. Case 4: FC + FDR, MI, and Bayesian optimization are adopted to optimize GBDT, and this case is the proposed method.

The testing accuracy is obtained by the classifier GBDT on the independent test set. The results shown in case 1 are the classification accuracy using GBDT without any feature selection. It can be observed that the GBDT without any feature selection obtains a testing accuracy of about 97.96%, but the testing balanced accuracy is only about 85.47%, which implied the learning efficiency of the GBDT without feature selection is not much high. In case 2, although FC and FDR effectively reduce the running time, it does not improve the accuracy significantly, because they ignore the correlation between features. In case 3, we add MI to further select key features, and the results show that there is an improvement (1.58%) in balance accuracy and a significant reduction in running time. In case 4, we use Bayesian optimization to optimize GBDT to obtain the optimal model. According to the results, we conclude that the accuracy and balanced accuracy are improved by 0.74 and 5.14%, which were compared with case 3. Particularly, the proposed method shown in case 4 obtains the highest testing accuracy and balanced accuracy. The performance of testing balanced accuracy is
The calculation procedure of balanced accuracy. The raw dataset is split into 10 equal datasets. The diagram shows the procedure on one of the 10 datasets. First, the integrative dataset is divided into a learning set and an independent test set. The learning set is further divided into a training set and a validation set. The training set is used to train the GBDT model and the validation set is used to obtain the weight for normal and tumor space ($\omega_n$ and $\omega_t$), which is represented by the accuracy of normal and tumor space. Finally, when the optimal model is tested on the test set, $\omega_n$ and $\omega_t$ will be used to modify the final accuracy to obtain the balanced accuracy.

### TABLE 4 | The mean values of seven evaluation metrics obtained from four methods on integrative dataset.

| Classifier | B_ACC | ACC | SES | SPC | PRC | F1 | AUC |
|-----------|-------|-----|-----|-----|-----|----|-----|
| SVM       | 0.9413 | 0.9865 | 0.9910 | 0.9435 | 0.9941 | 0.9926 | 0.9672 |
| RF        | 0.9208 | 0.9902 | 0.9968 | 0.9261 | 0.9924 | 0.9946 | 0.9615 |
| KNN       | 0.9480 | 0.9914 | 0.9965 | 0.9522 | 0.9950 | 0.9953 | 0.9738 |
| Proposed  | 0.9731 | 0.9951 | 0.9964 | 0.9626 | 0.9962 | 0.9973 | 0.9895 |

In the experiments, we randomly split the dataset into 10 equal-sized datasets. The mean values of the seven metrics are obtained on the 10 test sets. The proposed method outperforms other methods in balanced accuracy, accuracy, specificity, precision, F1 score, and AUC. The bold values are the best results.

improved by 13.85%, compared with the method in case 1. From the perspective of vertical comparison, the features selected by the proposed method have better classification performance. From the perspective of horizontal comparison, balanced accuracy improves more than traditional accuracy, which indicates that the proposed model shows greater advantages when the sample balance is considered.

### The Hyperparameters of GBDT Adjusted by Bayesian Optimization

Bayesian optimization aims to find the minimum value of the objective function by establishing a proxy function (probabilistic model). The proxy function is easier to optimize than the objective function (Victoria and Maragatham, 2020), so the next input value to be evaluated is selected by applying some criterion. For hyperparameter optimization, the objective function is the validation error of the machine learning model using a set of hyperparameters. Its goal is to find the hyperparameters that produce the minimum error on the validation set and to generalize these results to the test set. The cost of evaluating an objective function is significant because it requires the training of a machine learning model with a specific set of hyperparameters. Bayesian hyperparameter tuning uses a constantly updated probabilistic model to "focus" the search process on the hyperparameters that are likely to be optimal by reasoning from past results. In this study, for the objective function, the input was a set of hyperparameters, and the output was the fivefold cross-validation loss with classifier GBDT. We chose Tree Parzen Estimation (TPE) as the optimization algorithm. Figure 3 shows the best sets of hyperparameters obtained by Bayesian optimization and random search with 300 iterations. The balanced accuracy gained on the test set by using the best two sets of hyperparameters in GBDT was 97.31 and 96.8%, respectively. The results indicated that Bayesian optimization outperforms random search in the respect of hyperparameter tuning.

In the comparative experiments, we select three other classifiers, SVM, KNN, and RF. Supplementary Tables 6, 7 and Supplementary Figure 2 show the procedure of tuning parameters for the three classifiers. According to the balanced accuracy obtained in those models, the optimal parameters are as follows:

1. SVM: $C = 1$, kernel = “linear”
2. KNN: n_neighbor = 7, metric = “manhattan”
3. RF: max_depth: 46, min_sample_leaf: 2, min_sample_split: 94, n_estimators: 75
Table 5 shows the mean values of seven evaluation metrics obtained from four methods on the integrative dataset. The results indicate that the proposed method outperforms SVM, KNN, and RF by 3.4, 5.7, and 2.6% with balanced accuracy. Particularly, the AUC obtained by the proposed method is 2.3, 2.9, and 1.6% higher than the above three classifiers, respectively. We can conclude that the proposed method achieves the best performance on the integrative dataset in terms of balanced accuracy (97.31%), accuracy (99.51%), specificity (98.26%), precision (99.82%), F1 score (99.73%), and AUC (98.95%).

The Effect of Integrative Dataset
To explore the effect of the integrative dataset, we apply the proposed method to individual gene expression and integrative dataset, respectively. Besides, we choose PMA50 as the control model. PMA50 refers to a set of 50 genes selected by Parker et al. (2009), which are reported to be highly related to breast cancer. In Table 6, for the gene expression and PMA50, the proposed method achieves the best testing accuracy. The blue and orange bars in Figures 4A,B intuitively reflect the results. However, for the integrative dataset, the proposed method obtains 99.51% testing accuracy and 97.31% balanced accuracy, which outperforms the gene expression model and PMA50 model. This fact indicates that the features selected by the proposed model have better classification performance.

The results in Table 6 and Figure 4 also show the results obtained by the other classifiers. The SVM classifier gives the accuracy of 98.78% on the gene expression dataset, which is higher than that on the integrative dataset. However, the balanced accuracy is higher on the integrative dataset (94.93%). On the other hand, RF and KNN give a higher testing accuracy on the integrative dataset than that on the gene expression dataset, which is illustrated by the blue bars in Figure 4A. However, in Figure 4A, the proposed model obtains the highest three bars, which reveals that the proposed method performs better than other classifiers in all three types of datasets. For a balanced accuracy in Figure 4B, SVM and the proposed model obtain the best results on the integrative dataset, and RF and KNN obtain the best ones on gene expression and PMA50, respectively. The reason for this difference lies in the sensitivity of different classifiers to data distribution. The feature genes in the PMA50 model and the integrative model obtain higher balanced accuracy 97.4% (KNN) and 97.3% (proposed method) than that in the gene expression model, which illustrates that KNN and the proposed method provide the better capability to classify the minority sample class.

Biomarkers and GO/Pathway Analysis
The 50 genes (listed in the Supplementary Table 8) discovered by the proposed model include 16 genes, IQGAP3 (Hu et al., 2019), KIF4A (Xue et al., 2018), TSHZ2 (Yamamoto et al., 2011), MKI67 (Schmidt et al., 2007), TNXB (Hu et al., 2009), KIFC1 (Ogden et al., 2017), KDM5B (Catchpole et al., 2011), PPEF1 (Ye et al., 2020), RYR3 (Shrestha et al., 2012), TMEM132C (Zhang et al., 2020), FANCD2 (Barroso et al., 2006), ATAD2 (Kalashnikova et al., 2010), KIF26B (Wang et al., 2013), BRCA2 (Wooster et al., 1995), BLM (Arora et al., 2015), and ARFGGEF (Kim et al., 2011), which are reported to be directly associated with breast cancer by previous researches. Although the other 14 genes have not been verified by biological experiments, we further analyze the Gene Ontology and pathway enrichment to explore their impact on the tumor formation and progression.

Gene Ontology and pathway analysis produces biological function and pathway enriched for mutation genes. The result reveals that BRCA2, KDM5B, and IQGAP3 are associated with mammary gland epithelial cell proliferation and gland
**DISCUSSION**

This research presents a Staged Feature Selection method for breast cancer classification based on gene expression and somatic mutation datasets. In the proposed method, FC and FDR were used to select differentially expressed genes, MI was adopted to remove the irrelevant and redundant features, and an embedded method based on GBDT with Bayesian optimization was presented to obtain the informative features. Besides, the weighted metric was proposed to evaluate the classification accuracy, which could avoid the impact of sample imbalance on classification. The experiment results showed that the proposed method selected 50 feature genes and achieved the accuracy of...
99.51%, the balanced accuracy of 97.31% and the sensitivity of 99.64%, the specificity of 98.26%, the precision of 99.82%, the F1 score of above 99.73%, and the AUC of 98.95%, which was superior to the other three classifiers. It was verified that the proposed method was an efficient tool for feature selection in breast cancer classification.

The results presented the effectiveness of integration with gene expression and somatic mutation data for breast cancer classification, which indicated that it could provide more useful information for cancer classification by integrating multiple information. However, this study only focused on breast cancer, and the scalability of the proposed method on other types of cancers remained to be further explored, which will provide helpful information for cancer prevention and treatment. Therefore, in future work, we will apply the approach to classify other types of cancer, explore ways to incorporate more relevant data, and introduce other techniques to boost our method. Besides, the pathogenesis of some biomarkers discovered by the proposed model still has to be verified by biological experiments.

**DATA AVAILABILITY STATEMENT**

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/repository.

**AUTHOR CONTRIBUTIONS**

QJ processed the data, designed the algorithm and the programming codes, and wrote the manuscript. MJ supervised the project and revised the manuscript. Both authors contributed to the article and approved the submitted version.

**FUNDING**

This work was supported in part by the National Natural Science Foundation of China under Grant 61773157.

**ACKNOWLEDGMENTS**

We thank the members for their support and guidance.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.629946/full#supplementary-material.
Yamamoto, M., Cid, E., Bru, S., and Yamamoto, F. (2011). Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. *PLoS One* 6:e17149. doi: 10.1371/journal.pone.0017149

Ye, T., Wan, X., Li, J., Feng, J., Guo, J., Li, G., et al. (2020). The clinical significance of PPEF1 as a promising biomarker and its potential mechanism in breast cancer. *Onco Targets Ther.* 13, 199–214. doi: 10.2147/ott.s229432

Zhang, J., Zhang, S., Wang, Y., and Zhang, X.-S. (2013). Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst. Biol.* 7(Suppl. 2):S4.

Zhang, W., and Wang, S. L. (2020). A novel method for identifying the potential cancer driver genes based on molecular data integration. *Biochem. Genet* 58, 16–39. doi: 10.1007/s10528-019-09924-2

Zhang, X., Kang, X., Jin, L., Bai, J., Zhang, H., Liu, W., et al. (2020). ABCC9, NKAPL, and TMEM132C are potential diagnostic and prognostic markers in triple-negative breast cancer. *Cell Biol. Int.* 44, 2002–2010. doi: 10.1002/cbin.11406

**Conflict of 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.

**Copyright © 2021 Jiang and Jin.** This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.