A Novel Hybrid Feature Selection and Ensemble Learning Framework for Unbalanced Cancer Data Diagnosis With Transcriptome and Functional Proteomic

XIANFANG TANG\textsuperscript{1,1}, LIJUN CAI\textsuperscript{1}, YAJIE MENG\textsuperscript{1,1}, CHANGLONG GU\textsuperscript{1,1}, JIALIANG YANG\textsuperscript{2,2}, AND JIASHENG YANG\textsuperscript{3,3}

\textsuperscript{1}College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
\textsuperscript{2}Geneis Beijing Company Ltd., Beijing 100102, China
\textsuperscript{3}Academician Workstation, Changsha Medical University, Changsha 410219, China

Corresponding authors: Lijun Cai (ljcai@hnu.edu.cn) and Jiasheng Yang (yangjiashengnusemail@gmail.com)

This work was supported in part by the Hunan Provincial Innovation Foundation for Postgraduate under Grant CX20200434, in part by the National Nature Science Foundation of Hunan Province under Grant 2020JJ4209, in part by the Natural Science Foundation of Hunan Province under Grant 2018JJ2461, and in part by the Project to Introduce Intelligence from Oversea Experts to the Changsha City under Grant 2089901.

ABSTRACT The high dimension, high redundancy and class imbalance of cancer multiple omics data are the main challenges for cancer diagnosis. Existing studies have neglected the role of functional proteomics in the occurrence and development of cancer. In this study, a novel hybrid feature selection and ensemble learning framework, referred to as the three-stage feature selection and twice-competitive ensemble learning method (TSFS-TCEM), is proposed for cancer diagnosis. Firstly, we combine the transcriptome and functional proteomics data to construct a multi-omics data on breast cancer, which is the first time to apply these combined biological data for diagnosing breast cancer. Secondly, the proposed method introduces multiple models during the feature selection and diagnostic model construction. The three-stage feature selections integrate the features from different types of data and the twice-competitive ensemble learning framework resolves the data imbalance problem suffer from a single classifier. The TSFS-TCEM achieves a diagnostic accuracy of 99.64%, outperforming all compared methods. In addition, the 5-fold cross-validation sensitivity, specificity and F-Measure of the method are above 99.63%.

INDEX TERMS Functional proteomic, transcriptome profiles, the Cancer Genome Atlas (TCGA), ensemble method, hybrid feature selection, cancer diagnosis.

I. INTRODUCTION

Due to the low early detection rate and high mortality rate, cancer has become the main cause of human death. With the development of sequencing technology, cancer has been confirmed to be closely associated with genetics [1]. Besides, non-coding RNA (ncRNA), called “Dark matter”, also plays a vital role in the transcription process. This means that the abnormal expression of ncRNA may also cause disorders in gene expression [2]–[5]. Literature [2] merged mRNA and ncRNA transcriptomes of pancreatic cancer to obtain the altered miRNA regulation of mRNA expression. What’s more, functional proteins have extended the insight on the process and the associated regulatory mechanisms of cancer. With the emergence of Reverse Phase Protein Arrays (RPPA), it is possible to conduct large-scale proteomic research and analyze tumor functional proteomic [6]–[8]. Literature [8] developed the cancer proteome atlas portal (TCPA), which systematically revealed the functional proteomics expression in 32 cancer types. A recent research study reported that the transcriptome and proteome could result in new knowledge regarding novel functionally important gene products [9]. Hence, the existing methods that rely solely on transcriptome profiles or functional proteomic could lose some important biological information, such that they cannot reveal the formation process and regulatory mechanism of cancer properly. Analyzing and classifying the transcriptome profiles with high dimensions may be considered an NP-Hard...
problem [10]. Feature selection methods will not only get the optimal feature subsets and remove irrelevant information but also reduce computational cost. Numerous works adopt the feature selection methods on cancer diagnosis. These methods are categorized into three categories: filter, wrapper, and hybrid methods. The filter method reduces the dimensionality of transcriptome profiles by calculating the difference of samples’ features, the distance between the samples, and correlation measure between features and categories. These methods could quickly filter irrelevant feature, however, they are likely to ignore the relationship between the features. The wrapper method comprises a search algorithm and a learning model [11], [12]. It first uses the search algorithm to continuously generate candidate feature subsets, and then uses the learning model to evaluate the criterion feature subsets. This could effectively remove feature redundancy. However, the wrapper method suffers a large search space and low efficiency for high dimensional data, which require high computational cost. To solve the above limitations, hybrid feature selection methods have been applied in various fields with superior performance [13]–[15]. The hybrid feature selection achieves rapid dimensionality reduction and redundancy removal of high-dimensional data, but their overall performance is constrained by their packaging search strategies and classification learning algorithms. Besides, most of the hybrid feature selection ignore the feature differences between multiple biological datasets in the dimensionality reduction of transcriptome profiles. It may cause feature selection mainly from one dataset with different characteristics. Furthermore, the single model application in the hybrid methods also has a limitation in complexity to evaluate the multiple biological datasets, and may potentially cause the overfitting issue. Thus, considering the difference of data characteristics in transcriptome profiles and functional proteomic, it is of utmost importance to propose a suitable feature selection method to quickly remove irrelevant and redundant features with the requirement of the feature stability and diversity.

The class imbalance distribution is an essential problem in cancer diagnosis. It makes the diagnosis model pay more attention to the majority class, such that wrong diagnosis of the minority class, regarding as more cost sensitive than the major class, become a major concern in the classification modeling. In the literature, lots of contributions have been made to solve this issue. The existing methods could be divided into two ways. In the first way, resampling techniques, like under-sampling and over-sampling, were used to modify data distribution. It converted imbalanced cancer diagnosis into a balanced classification problem. These techniques are effective to overcome the imbalance issues, but some disadvantages could obviously be found in the resampling procedures. For example, the under-sampling method may cause the loss of important information because of the random sampling of maximum classes. The over-sampling methods include synthetic minority over-sampling technique (SMOTE) [16], adversarial generation network [17]. Although over-sampling can solve data imbalance, the validity of the generated data is questionable due to the complexity of cancer pathogenesis. To overcome the above disadvantages, in the other way, diagnosis models were utilized internally to solve data imbalance issue. Literature [18] pointed out that ensemble learning with diversity diagnosis models could achieve better performance than single diagnosis models. Most bagging methods integrate multiple homogeneous models based on the same algorithm, but they don’t integrate heterogeneous models based on the different algorithms. What’s more, introducing multiple algorithms multiple models into the bagging method may not improve the diagnostic performance. Sometimes too many combinations lead to negative effects. Therefore, the construction of an ensemble model for cancer diagnosis needs further research.

Based on the above-mentioned findings, we proposed a novel hybrid feature selection and ensemble learning framework, named three-stage feature selection and twice competitive ensemble method (TSFS-TCEM), for cancer diagnosis.

(1) We combined transcriptome profiles and functional proteomic to construct the multi-omics data for breast cancer diagnosis, which is the first time to apply these combined biological data for diagnosing breast cancer.

(2) The TSFS-TCEM introduced multiple algorithms multiple models into the feature selection level and diagnostic model level. The TSFS was divided into three stages for ensuring diversity of multi-omics. In the first stage, fold change-false discovery rate(FC-FDR) and information gain could quickly filter the features of high-dimensional transcriptome profiles, and functional proteomic was inappropriate for the FC-FDR method due to the data characteristics. The second stage was to reduce the dimensionality and eliminate redundancy of the multi-omics data. The third stage feature selection used multiple repeatability tests to select the most stable and optimal subset. We proposed the TCEM to classify the imbalanced multi-omics of breast cancer datasets. The twice competition entailed the competition of the prediction model sets and the competition of the optimal model, which could improve the performance of the bagging method.

Finally, 5-fold cross-validation was used to evaluate the performance of TSFS-TCEM.

II. RELATED WORKS
In view of the particularity of cancer diagnosis, it faces the problem of data imbalance. Machine learning algorithms are used in various fields [19]–[21], especially in biological field. High dimensional biological data has caused an increase in computational cost. There are some researches to adopt the hybrid feature selection and ensemble methods.

A. FEATURE SELECTION
Literature [22] adopted t-statistic and support vector machine (SVM) to overcome the limitation of SVM-RFE. The result showed that the proposed method selected the differentially significant features on the 6 microarray datasets, and the performance was superior to the three other feature
methods to perform feature dimensionality reduction and improve the performance of bagging method. It needs to construct a suitable mechanism to enhance the performance of the whole algorithm. Besides, it was easier to influence the different processes, and it was harder to deal with the feature ablation problem of the multi-omics data. In the third stage feature selection uses multiple repeatability tests to select the most stable and optimal chromosome.

B. ENSEMBLE METHOD

Literature [26] adopted the boosting method and developed three computational methods of SVM: boosted support vector machines negative, boosted support vector machines plus, and boosted support vector machines plus negative. The experiment based on the gene expression data of breast cancer, oral cancer, and the result indicated that the proposed method is superior to SVM. Literature [27] based on Bagging method to obtain SVM integrated model to classify gene expression profile data of lung adenocarcinoma, and achieved the highest accuracy rate of 94%, and the results show that this method is more stable than a single SVM model. Literature [28] obtained the minimum features through the johnson method and a combined classifier of K-nearest neighbor (KNN) to classify the Wisconsin breast cancer dataset. The combined rules based on the maximum voting. The result showed that the ensemble classifier was superior than the single classifier. Literature [29] proposed maximum relevance minimum redundancy and artificial neural network as the hybrid method, the ensemble method based on decision tree (DT) with the bagging method for the classification of brain tumor tissues. The result showed that it was more efficient to combine the hybrid feature selection and ensemble method, but their ensemble method only considered decision tree. Literature [30] combined three feature selection methods as the hybrid feature selection, which included chi-square, information gain and principal component analysis as the hybrid method. This study used three ensemble method to evaluate the six base classifiers, such as Gaussian Naïve Bayesian. The result showed that the predictive accuracy of bagging method was 95.94%, which is lower than boosting and staking methods.

All in all, these methods adopted different indicators to evaluate the different processes, and it was easier to influence the performance of the whole algorithm. Besides, introducing multiple algorithms multiple models into the bagging method, it needs to construct a suitable mechanism to improve the performance of bagging method.

III. METHODS

A. THREE-STAGE FEATURE SELECTION

Most researches use the single or hybrid feature selection methods to perform feature dimensionality reduction and de-redundancy. However, the transcriptome profile has the characteristics of high dimensionality and high redundancy. The functional protein datasets have low dimensionality and low abundance. Feature selection for these two types of data is prone to the ablation of the particular data types. The overall framework of TSFS is shown in Figure. 1, the three stage feature selection method is divided into three stages. The first stage is the feature filtering. In this stage, FC-FDR and Information gain are used to quickly filter the features of high-dimensional transcriptome profiles, and functional proteomic is inappropriate for the FC-FDR method due to the data characteristics. The second stage is to reduce the dimensionality, eliminate redundancy as well as deal with the feature ablation problem of the multi-omics data. In the third stage feature selection uses multiple repeatability tests to select the most stable and optimal chromosome.

1) FIRST STAGE FEATURE SELECTION

In the first stage, the FC-FDR method is used to filter the unrelated feature of transcriptome profile, and then the information gain extracts features to obtain the 0.5L-dimensional features of transcriptome profiles. Protein data has few dimensions, and so such 0.5L-dimensional protein data is obtained only by information gain.

FC-FDR: Fold change (FC), the basic method for detecting the differentially expression genes, is used to calculate the multiple of the difference in expression value between positive and negative samples. Define the cancer datasets as \( C \in R^{n \times w} \). It can be divided into two parts, that is, \( X \in R^{n \times w1} \) and \( Y \in R^{n \times w2} \) belongs to the tumor (Positive) and normal (Negative) samples. \( X_i \) and \( Y_i \) are the average expression value of the feature \( i \) on the tumor and normal samples, respectively, and then the value of FC can be obtained by formula (1):

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

FDR hypothesis test can effectively solve the false positive problem in FC method. Using a statistical test, we reject the null hypothesis if the test is declared significant. On the contrary, we do not reject the null hypothesis if the test is non-significant. Define \( V \) is the number of false positives, \( S \) is the number of true positives. We can get the formula (2):

\[
FDR = \frac{V}{V+S}
\]

Generally, the threshold of FDR is set 0.05 to control the number of false positives in multiple hypotheses. If FDR is less than 0.05, it can demonstrate that the hypotheses are believable.

Information Gain: Information gain can be used as an indicator to select features [31]. Given a discrete random feature \( A \) from transcriptome profiles \( C \), its probability distribution is \( p(A = A_i) = p_i, i \in R^{n \times w} \). Then the entropy of the feature \( C \) is defined as:

\[
H(C) = -\sum_{i=1}^{n} p_i \log p_i
\]
The conditional entropy $H(C|A)$ represents the uncertainty based on the feature $A$.

$$H(C|A) = \sum_{i=1}^{n} p_i H(C|A = A_i)$$ (4)

$g(C, A)$ represents the differences between entropy and conditional entropy.

$$g(C, A) = H(C) - H(C|A)$$ (5)

The value of $g(C, A)$ is bigger, it means that the feature $A$ is more important to transcriptome profiles $C$.

2) SECOND STAGE FEATURE SELECTION

The second stage feature selection is used to the population algorithm randomly generates $R$ chromosomes and each chromosome owns $O$ gene loci ($O$ dimensional features), which represents the potential solution of the problem. As shown in Figure. 1. In this stage, each chromosome is evaluated by SVM, DT, KNN, and random forest (RF). According to the maximum fuzzy strategy, chromosome is retained, if the predictive accuracy of the chromosome with one base classifier is larger than the threshold $\delta$. When the iteration finish, the frequency of gene loci could be counted, and then will be used to select the Top-$K$ genes based on the sort of the gene loci’s frequency as the new chromosome. The description process of the second stage feature selection is shown in Algorithm 1.

In this stage, 4 base models are used to evaluate each chromosome to avoid the overfitting problem in the application of the single classifier model. In addition, the maximum fuzzy strategy prevents the local optimal fast-fall trap and maintains the population diversity.

3) THIRD STAGE FEATURE SELECTION

The third stage feature selection needs to select the optimal chromosome with good generalization ability. In this stage, the population algorithm randomly generates $R$ chromosomes and each chromosome owns $O_1$ gene loci ($O_1$...
dimensional features), which represents the potential solution of the problem. Each chromosome meets the fitness function could be remained, and then Z random repeatability tests are carried out on the selected chromosome to obtain the final score ($P_{SUM}$). After a population iteration, it could obtain the optimal chromosome based on the final scores. The description process of the third stage feature selection is shown in Algorithm 2.

To ensure that the optimal chromosome can be obtained, we set some rules: (1) If the predictive accuracy of any base model is larger than the threshold $\delta$, it will be retained. (2) If the sum of predictive accuracy of every chromosome is less than threshold2, the top 20 chromosomes are remained.

**Algorithm 2 Third Stage Feature Selection**

**Input:** feature dataset, size of the population $T$.

**Output:** Select the optimal chromosome.

1. function TS(Data2, R, Data3).
2.     final = 0, num = 0;
3.     array arr1, arr2;
4.     while t0 to R do
5.         population algorithm to generate chrom($A_t$)
6.         sum = 0
7.         for i from 1 to k do
8.             sum + = $P(M_i(A_t))$;
9.             if $M_i(A_t)$ predictive accuracy larger than the threshold $\delta$ then
10.                 num+ = 1;
11.                 arr1[num] = t;
12.         end if
13.         sum1[num] = sum
14.         if num > M then
15.             break;
16.         end if
17.     end for
18.     end while
19.     if num == 0 then
20.         while t1:0 to 20 do
21.             randomly generated 4 training sets to test $A_t$;
22.             calculate sum of the 4 random repeatability test result; $P_{SUM}$
23.         end while
24.     else
25.         while t1:0 to num do
26.             randomly generated 4 training set to test $A_t$;
27.             calculate sum of the 4 random repeatability test result $P_{SUM}$;
28.         end while
29.     end if
30.     select the optimized of $A_t$
31. end function

**B. TWICE COMPETITIVE ENSEMBLE METHOD**

Twice competitive ensemble method is based on the bagging method (Figure 2). At the construction of training set, generate N training sets of the same proportion by resampling. Randomly select half of the samples from the minority class, and randomly select samples from the majority class as the training set at a ratio of 1:4 by undersampling. This method could avoid over-fitting problems of prediction model sets by a few samples of a single training set.

The four classifiers namely KNN, SVM, DT and RF are used to evaluate the data. Votes of even classifiers may be meaninglessness and discarded. The predictive model is demonstrated as follows:

Base model: KNN, SVM, DT, RF Combined model: $SVM\_DT\_KNN$, $SVM\_DT\_RF$, $SVM\_KNN\_RF$, $DT\_KNN\_RF$

Based on this, the predictive model set consists of base model and combined model. The twice competition is divided into the competition of predictive model sets and the competition of optimal model.

The competition of predictive model sets is aimed to select the top 5 prediction model set from the N predictive model sets. These model sets are trained by N training sets respectively. Calculating and summing the prediction accuracy of each prediction model set, and then, sorting N prediction model sets to select the top 5 prediction model sets.

The competition of optimal model is aimed to select the final model from each of the selected predictive model sets. There are 8 models in the prediction model set, and the classifier accuracy may exist in one or more cases with the highest accuracy. Therefore, different strategies can be used to select the optimal model. (1) There is a classifier model with the highest accuracy, which is regarded as the optimal solution of the predictive model set. (2) There are two or more classifier models with the highest predictive accuracy and only one is base model. It is adopted base model as the optimal model. (3) The other situations, which can select the optimal model by summing the predictive accuracy of each model in the N predictive model sets.

Through the twice competition, we obtained 5 models from N predictive model sets. And then, competitive ensemble is used to voting method to classify and diagnose.

**IV. RESULTS AND DISCUSSION**

**A. CONSTRUCTION OF DATASET**

Transcriptome profiles and functional proteomic profiles belong to the Cancer Genome Atlas (TCGA). Transcriptome profiles of breast cancer contain a total of 1109 patient samples and 113 normal samples, and each sample contains 62125-dimensional mRNA and ncRNA data. The transcriptome profiles contain a total of 901 samples, including 22 normal samples and 879 patients’ samples. Each sample contains 233 dimensions of functional proteomic profiles. The multi-omics datasets own 840 samples, 18 normal samples and 822 patients samples, of which each contains 62358 dimensions data. The specific information of data is shown as Table 1.
B. CONFUSION MATRIX

The confusion matrix is the case of training a predictive data model algorithm and the real situation, the use of accuracy (Accu), precision (Prec), sensitivity (Se), Specificity (Specificity, Sp) four indicators for performance evaluation, Accu, Prec, Sn, Sp are defined as follows:

\[
\text{Accu} = \frac{TP + TN}{TP + FP + FN + TN} \quad (6)
\]

\[
\text{Prec} = \frac{TP}{TP + FP} \quad (7)
\]

\[
\text{Se} = \frac{TP + FN}{TN} \quad (8)
\]

\[
\text{Accu} = \frac{TN}{FP + TN} \quad (9)
\]

TP, FP, TN, and FN is shown in Table 2.

C. PARAMETER OF EXPERIMENT

The FC threshold was 2, α was 0.05, and L was 60. In the second stage feature selection, the total number of chromosome gene loci \( O_1 \) was set to 40, the Top-K genes were 20, and the threshold1 was 0.99. In the third stage feature selection, the total number of chromosome gene loci \( O_2 \) was set to 10. The population number R was 600. Z repeatability tests was 4, and M was 50. In the TCEM, N was 50.

D. THE PERFORMANCE OF THE TSFS-TCEM

To evaluate the performance of the TSFS-TCEM method, we adopted multi-omics datasets to diagnose breast cancer through 5-fold cross-validation [32]–[35]. As was Table 3 shown, its accuracy, sensitivity, specificity, and F-Measure were 99.64%, 99.64%, 100%, and 99.81%, respectively. With multiple repeatability tests by 5-fold cross-validation, the average accuracy was 99.4% and diagnosis error was 0.24%.
In the TCGA database, only BRCA has transcriptome profiles and functional proteomic datasets with normal and cancer samples. Therefore, we evaluated TSFS-TCEM performance on the BRCA, LUAD and KIRC of transcriptome profiles (Table 4), and the diagnostic accuracy was 99.52%, 99.01%, and 99.32%, respectively. The result showed that the performance of the TSFS-TCEM fluctuated with the variation of imbalance rate, but the diagnostic accuracy of the model remained above 99%.

In view of the optimization mechanism of the TSFS, it aimed to find multiple chromosomes that meet the fitness function, so the TSFS need to traverse as many chromosomes as possible. To control the computational cost, we set the variable num to control the iteration. In the first stage feature selection, it cost 132.47s. In the second stage and the third stage feature selection, each single chromosome cost 0.56s. In the diagnose model, each evaluation cost 0.73s.

E. COMPARISON WITH OTHER METHODS

In this study, the TSFS-TCEM was compared with the different competition ensemble strategies through 5-fold cross-validation. The first competitive ensemble of the base models (Figure 3: M1-M4), the first competitive ensemble of the combined models (Figure 3: M5-M8), and the TCEM proposed in this article (Figure 3: M9). The first competition ensemble was the competition of the optimal competition by the base model or the combined model.

As illustrated in Figure 3, in the competitive ensemble of the base model, M1 (SVM) had the poorer diagnostic performance than the other three methods, and M4 (RF) had the optimal performance (99.28%). It indicated that different algorithms owned different performances on the diagnosis imbalanced datasets. Besides, the competitive ensemble of the combined model (M5-M8) was 0.33% higher than the competitive ensemble of the base model (M1-M4) on average, and the diagnosis accuracy reached 99.4% (M7, M8). It proved that the combined model achieved better performance than the base model in most cases. The TSFS-TCEM was superior to the other competitive ensemble methods. Therefore, we attest that the TSFS-TCEM is suitable for dealing with the classification of imbalanced datasets.

Due to the special of multi-omics, we do not find similar research to diagnose breast cancer through transcriptome profiles and functional proteomic datasets. Thus, we adopted the TSFS-TCEM with transcriptome profiles to compare the other methods. Literature [36] combined GA-SVM and SVM to analyze breast cancer of miRNA expression profiling. The predictive accuracy, sensitivity, specificity, and F-Measure were 95.29%, 95.8%, 96.97%, and 96.24%, respectively. Literature [37] adopted FDR and recursive feature selection to select the joint data, which combined mRNA, miRNA and lncRNA. And then, they used SVM to diagnose lung cancer. The predictive accuracy on the LUAD dataset was 95.3%. Literature [38] adopted DESeq as the feature method to select informative genes. They used KNN, SVM, DT, RF, and GBDT to the base models to train the informative genes, and then used deep neural network to master the predictive results of these base models. The predictive accuracies on the BRCA and LUAD datasets were 98.41% and 98.8%, respectively.

As illustrated in Table 5, compared to other methods on BRCA and LUAD datasets, the TSFS-TCEM improved diagnostic accuracy by 0.2-4.4%. It was demonstrated that the TSFS-TCEM kept more knowledge from cancer biological data. Therefore, we attest that the TSFS-TCEM algorithm is helpful for the cancer diagnosis with imbalanced datasets.

F. THE PERFORMANCE OF TSFS-TCEM PARAMETER

In this study, we assumed that the sampling ratio and the population number could affect the performance of the proposed model. We systematically analyzed the impact of the different parameters on the performance of TSFS-TCEM by single variable control (Table 6). When the population number was 100-400, fluctuations in prediction accuracy would shrink with the sampling ratio increased. The population number was 500 or more, the experimental results tended to be stable (99.4%). When the sampling ratio was 1: 4 and the population number was 600, the prediction accuracy reached the highest 99.64%. As shown in the Table.6, we also used transcription profiles to evaluate the performance of the TSFS-TCEM, and the results were similar to the multi-omics data experiment. Therefore, we believed that when the population number reached 500, the proposed method could ignore the parameters change.

| Cancer | Acc   | Unbalanced rate |
|--------|-------|-----------------|
| BRCA   | 99.52%| 47.5            |
| LUAD   | 99.01%| 7.48            |
| KIRC   | 99.32%| 8.39            |
TABLE 6. The precisions of HFS-TCEM at the different population number and sampling ratio. (%).

| BRCA  | Sampling ratio | Number of population |
|-------|----------------|----------------------|
|       | 100            | 200                  | 300                  | 400                  | 500                  | 600                  | 1000                |
| multi-omics | 1.1  | 98.92 | 99.04 | 99.04 | 99.17 | 99.4  | 99.4  | 99.4  |
|       | 1.2  | 98.8  | 98.92 | 98.28 | 99.28 | 99.4  | 99.4  | 99.4  |
|       | 1.3  | 99.04 | 99.17 | 98.92 | 99.04 | 99.4  | 99.4  | 99.4  |
|       | 1.4  | 99.28 | 98.92 | 99.4  | 99.28 | 99.4  | 99.4  | 99.4  |
|       | 1.5  | 98.92 | 98.92 | 98.8  | 99.28 | 99.4  | 99.4  | 99.4  |
|       | 1.6  | 99.4  | 99.17 | 99.28 | 99.28 | 99.4  | 99.4  | 99.4  |
|       | 1.10 | 99.4  | 99.28 | 99.4  | 99.4  | 99.4  | 99.4  | 99.4  |
| Transcription | 1.4  | 99.10 | 99.28 | 99.4  | 99.4  | 99.52 | 99.52 | 99.4  |

Sampling ratio represents the sampling ratio of positive and negative training set samples.

FIGURE 4. The most frequent predicted potential features.

### G. DISCUSSION ON ONCOGENES AND THERAPEUTIC TARGETS

At feature optimization, the selected features could have the randomness of characteristics since the learning set and the test set were randomly selected in each fold experiment. Therefore, the top-10 features would be varied with the test set. As is shown in Figure 4, the proportion of the functional protein data screened by the TSFS was 30%-40%, which also showed that functional proteins played an important role in the occurrence and development of breast cancer. We found that the expression values of functional proteins and their corresponding genes did not appear to be significantly abnormal, which means that the p-value of the corresponding genes was lower than the threshold of the FC-FDR. The result indicated that there was no absolute relevance between functional proteins and their corresponding genes.

The selected features from transcriptome profiles were ADAMTS-5, KIT, MAMDC2, PER1, TSN1, and VEGFD. The cleavage of ADAMTS-5 may affect the balance between the pro-tumor and anti-tumor effects induced by Fibulin-2 [39]. The expression of KIT varies according to the tumor diameter, tumor stage, and lymph node metastasis in breast cancer patients [40]. MAMDC2 is significantly correlated with the prognostic survival rate of the breast cancer patients [41]. The low expression of PER1 or PER2 in breast tissue may potentially inhibit apoptosis and eventually lead to cancer [42]. Studies have shown that the expression of TSN1 is negatively correlated with the invasion and metastasis of miR-548J [43]. VEGFD increases the chance of tumor invasion and metastasis by promoting the formation of tumor blood and lymphatic vessels. It can be therefore applied as a valuable prognostic evaluation index for breast cancer [44].

The functional proteins were AR, Akt, Bak, and 4E-BP1. AR is an intracellular protein that binds to androgen activated with high affinity and low energy; AR promotes the development of the breast through the growth factor MAPK signaling pathway [45]. Akt is cancer cells, it forms one of the main targets of the main effector downstream of this signaling pathway, which makes the pathway active after phosphorylation, and can block the pathway after dephosphorylation [46]. Besides, BH3-only protein, antibodies can directly activate Bak by binding the alpha1-alpha2 loop, thereby triggering cancer cell death [47]. 4E-BP1 is a highly conserved small molecule protein that is regulated by PI3K/AKT/mTOR, MAPK and other signaling pathways, thereby affecting the cell cycle Process [48].

We discovered that these functional proteins mainly participated in or regulate PI3K/AKT/mTOR and MAPK signaling pathways. PI3K/AKT/mTOR is an important intracellular signaling pathway that regulates the cell cycle. MAPK signaling pathway is transmitted from the cell surface to the interior of the nucleus. These signaling pathways play significant roles in the occurrence, development, and treatment of the tumor cells. As such, our findings indicated that these functional proteins could be utilized as the participants or regulators of these signaling pathways, and they could be used as potential drug targets for the treatment of breast cancer.

### V. CONCLUSION

To address the high dimension and imbalance of class challenge associated with multi-omics data, this paper proposed a novel framework with hybrid feature selection and ensemble method, named the TSFS-TCEM. The TSFS-TCEM was divided into three steps. Firstly, we combined transcriptome profiles and functional proteomic to construct the multi-omics data for diagnosing breast cancer, which is the first time to apply these biological datasets in this area. The TSFS-TCEM introduced multiple algorithms multiple models into the feature selection level and diagnostic model level. The TSFS was divided into three stages for ensuring diversity of multi-omics. In the first stage, FC-FDR and information gain wold quickly filter the features of high-dimensional transcriptome profiles, and functional proteomic was inappropriate for the FC-FDR method due to the data characteristics. The second stage reduced the dimensionality and eliminated redundancy of the multi-omics data. The
third stage feature selection used multiple repeatability tests to select the most stable and optimal subset. We proposed the TCEM to classify the imbalanced multi-omics of breast cancer datasets. The twice competition entailed the competition of the prediction model sets and the competition of the optimal model, which could improve the performance of the bagging method. The result showed that the diagnostic accuracy of the TSFS-TCEM was 99.64% through 5-fold cross-validation, and the four indicators of accuracy, sensitivity, specificity, and F-Measure were all above 99.63%. We discovered that there was no absolute correlation between the functional proteins and their corresponding genes, and functional proteins can be used as potential drug targets for the treatment of breast cancer.

The selected features of the TSFS-TCEM was based on the highest predictive probability of the multiple models, therefore, the performance would fluctuate by the population number and the sampling ratio. Besides, the computational cost of the TSFS-TCEM increased with the population number. Our future works is to improve the performance of the TSFS-TCEM under the low population number.

ACKNOWLEDGMENT

The authors would like to thank the members for their support and guidance.

REFERENCES

[1] L. J. van’t Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer.” Nature, vol. 415, no. 6871, pp. 530–536, 2002.

[2] S. Müller, S. Raulefs, P. Bruns, F. Afonso-Granz, A. Plötner, R. Thermann, C. Jäger, A. M. Schäffer, B. Kong, I. Regeli, W. K. Roth, B. Rotter, K. Hoffmeier, G. Kahil, I. Koch, F. J. Theis, J. Kleeff, F. Winter, and C. W. Michalski, “Next-generation sequencing reveals novel differentially regulated mRNAs, IncRNAs, miRNAs, sRNAs and a piRNA in pancreatic cancer.” Mol. Cancer, vol. 14, no. 1, p. 94, Dec. 2015.

[3] J. Vey, L. A. Kapsner, M. Fuchs, P. Unberath, G. Veronesi, and M. Kunz, “A toolbox for functional analysis and the systematic identification of diagnostic and prognostic gene expression signatures combining meta-analysis and machine learning.” Cancers, vol. 11, no. 10, p. 1606, Oct. 2019.

[4] J. Xu, L. Cai, B. Liao, W. Zhu, P. Wang, Y. Meng, J. Lang, and J. Yang, “Identifying potential mirna-disease associations with probability matrix factorization.” Frontiers Genet., vol. 10, p. 1234, 2019.

[5] A. Lopez-Rincon, L. Mendoza-Maldonado, M. Martinez-Archundia, A. Schönhuth, A. D. Kraneveld, J. Garsen, and A. Tonda, “Machine learning-based ensemble recursive feature selection of circulating miRNAs for cancer tumor classification.” Cancers, vol. 12, no. 7, p. 1785, Jul. 2020.

[6] H. U. Osmany, P. Théron, G. Peleux, J. F. Broomberg, and C. S. Leslie, “Linking signaling pathways to transcriptional programs in breast cancer.” Genome Res., vol. 24, no. 11, pp. 1869–1880, Nov. 2014.

[7] K. Njoku, D. Chiasserini, A. D. Whetten, and E. J. Crobish, “Proteomic biomarkers for the detection of endometrial cancer.” Cancers, vol. 11, no. 10, p. 1572, Oct. 2019.

[8] M.-J.-M. Chen, J. Li, Y. Wang, R. Akbani, Y. Lu, G. B. Mills, and H. Liang, “TCPA v3.0: An integrative platform to explore the pan-cancer analysis of functional proteomic data.” Mol. Cellular Proteomics, vol. 18, no. 8, pp. S15–S25, Aug. 2019.

[9] J.-Y. Xu, C. Zhang, X. Wang, L. Zhai, Y. Ma, Y. Mao, K. Qian, C. Sun, Z. Liu, S. Jiang, and M. Wang, “Integrative proteomic characterization of human lung adenocarcinoma.” Cell, vol. 182, no. 1, pp. 245–261, 2020.

[10] N. Alnugren and H. Alshamal, “A survey on hybrid feature selection methods in microarray gene expression data for cancer classification.” IEEE Access, vol. 7, pp. 78533–78548, 2019.
LIJUN CAI received the Ph.D. degree in computer application technology. He is currently a Professor of computer science and technology with Hunan University. His main research interests include cloud computing and bioinformatics.

YAJIE MENG received the M.S. degree in software engineering from Hunan University, Hunan, China, in 2018, where she is currently pursuing the Ph.D. degree in computer science and technology. Her current research interests include machine learning, cancer diagnosis, predicting ncRNA and disease association, and artificial intelligence.

CHANGLONG GU received the Ph.D. degree from Hunan University. He has published three articles related to miRNAs and diseases associations. His research interest includes biological big data analysis.

JIALIANG YANG received the Ph.D. degree from the Department of Mathematics, National University of Singapore, in 2009. From 2010 to 2011, he was an Assistant Professor with the CAS-MPG Partner Institute for Computational Biology, China. After that, he moved to USA and became a Postdoctoral Fellow with the Department of Basic Sciences, Mississippi State University. Since 2013, he has been a Postdoctoral Fellow and a Senior Scientist with the Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai. He is currently the Vice President of Genesis Beijing Company Ltd. He has published more than 70 peer-reviewed articles. His main research interests include bioinformatics, machine learning, oncology, aging, and evolutionary biology.

JIASHENG YANG received the Ph.D. degree from the Department of Automation, Northeastern University, China, and the Ph.D. degree from the Department of Civil and Environmental Engineering of Singapore University of Technology. He was a Research Fellow with the National University of Singapore. He is currently a Data Scientist Lead with DBS. His main research interests include machine learning, control theory, and bioinformatics.