IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 10, 2019, accepted January 25, 2019, date of publication February 12, 2019, date of current version March 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2898723

A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data

JING XU[1,2], PENG WU[1,2], YUEHUI CHEN[1,2], (Member, IEEE), QINGFANG MENG[1,2], HUSSAIN DAWOOD[3], AND MUHAMMAD MURTAZA KHAN[3]

[1]School of Information Science and Engineering, University of Jinan, Jinan, China
[2]Shandong Provincial Key Laboratory of Network Based Intelligent Computing, China
[3]College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

Corresponding authors: Peng Wu (ise_wup@ujn.edu.cn) and Yuehui Chen (yhchen@ujn.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671220, Grant 61640218, and Grant 61201428, in part by the National Key Research and Development Program of China under Grant 2016YFC106000, in part by the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation, China, under Grant ZR2016FB14, in part by the Project of the Shandong Province Higher Educational Science and Technology Program, China, under Grant J16LN07, in part by the Shandong Province Key Research and Development Program, China, under Grant 2016GGX101022, in part by the Doctoral Foundation of the University of Jinan under Grant XBS1748 and Grant XBS1806, and in part by the Shandong Province Natural Science Foundation under Grant ZR2018LF005.

ABSTRACT Classification of cancer subtypes is of paramount importance for diagnosis and prognosis of cancer. In recent years, deep learning methods have gained considerable popularity for cancer subtype classification, however, the structure of the neural network is difficult to determine and the performance of the deep network depends largely on its structure. To address this problem, a flexible neural tree (FNT) may be used. FNT is a special neural network with the advantage of automatic optimization of structure and parameters which cannot be used for multi-class classification. In this paper, a deep flexible neural forest (DFNForest) model is proposed, a novel ensemble of FNT model to aid with the classification of cancer subtypes. The proposed DFNForest model differs from the conventional FNT model because it transforms a multi-classification problem into many binary classification problems for each forest. We explore the cascade structure of DFNForest to deepen the flexible neural tree model so that the depth of the model is increased without introducing additional parameters. In addition to the DFNForest model, this paper proposes a combination of fisher ratio and neighborhood rough set for dimensionality reduction of gene expression data to obtain higher classification performance. The experiments on RNA-seq gene expression data show that our gene selection method has higher accuracy with fewer genes and the proposed DFNForest model has better performance for classification of cancer subtypes as compared to the conventional methods.

INDEX TERMS Cancer subtypes, cascade forest, classification, gene selection, machine learning.

I. INTRODUCTION

It is well known that cancer is a heterogeneous disease with different pathogenesis [1], [2]. Therefore, cancer has multiple molecular subtypes and its genesis and targeted treatments are different [3]–[5]. Recently, with the advent of high-throughput profiling technology, a large amount of genomic and transcriptomic data has been generated. This has provided an unprecedented opportunity to study cancer subtypes at the molecular level. Gene expression data usually consists of thousands of genes, however, the number of available samples is small. Among the thousands of features in gene expression data, only a small fraction of genes are associated with cancer subtypes, and the rest are redundant or noisy features. This converts the use of gene expression data for cancer subtype classification into a dimensionality reduction problem. The goal is to make the classification process efficient by using fewer features

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

while attempting not to compromise on efficiency. Learning knowledge from gene expression data is a hot research topic and has inspired many applications in medicine [6], [7].

Gene selection methods may be separated into three categories: filter, wrapper and embedded methods [8]. Filter methods focus on using a certain indicator to measure the close relationship between each attribute and sample category and sorting them according to the degree of association from large to small. Finally according to a preset threshold or the top k attributes to form a feature subset. Golub [9] proposed the use of signal-to-noise ratio (SNR) for gene selection and selected 50 genes for the diagnosis of acute lymphoblastic leukemia based on gene expression profiling. Zhu *et al.* [10] proposed to use t-test for overcoming the sparsity problem for feature selection for gene expression-based classification. Goh *et al.* [11] combined the Pearson correlation coefficient and signal-to-noise ratio with an evolving classification function to select minimum number of genes for gene expression-based classification. Muharram and Smith [12] proposed an effective Gini Index based approach for gene filtering. Liao *et al.* [13] proposed the use of Wilcoxon rank sum test along with support vector machine for assessing the importance of each gene. Kononenko [14] proposed the Relief method for gene selection, which is widely used for its excellent performance. The wrapper methods generally use a classifier to measure the performance of a feature subset, and then adjust the feature subset according to the result. The process is repeated until the optimal feature subset is obtained. In this regard, Li [15] combined genetic algorithm with K-nearest neighbor method to select feature genes. Khan *et al.* [16] proposed a hybrid selection method that comprised of a neural network and random feature extraction. The embedding method refers to embedding the selection of feature subsets into the training process of the classifier, which can be regarded as a combination of classifier training and gene selection process. Guyon *et al.* [17] proposed a recursive feature elimination (SVM-RFE) algorithm for gene selection, which is successfully applied to gene selection [18]–[20].

The rough set theory pioneered by Pawlak [21] has been successfully applied in bioinformatics to select informative genes [22], [23]. However, it has an obvious disadvantage that makes it unsuitable for processing continuous gene expression data. To address this shortcoming, Hu *et al.* [24], [25] proposed a neighborhood rough set model to process both discrete and continuous data sets. Directly using the neighborhood rough set model to eliminate redundant genes results in high calculation cost. To remedy this problem, we propose to use fisher ratio along with the neighborhood rough set model. This helps in removing a large number of unrelated genes, reduce the space-time consumption of the neighborhood rough set reduction process and reduce the training time of the classifier.

Neural networks have recently gained prominence in classification problems including cancer subtype classification [26], [27]. Neural networks generally result in high classification accuracy, which is purely dependent on the network structure. Unfortunately, there is no laid down procedure for selection of network structure. It depends upon the experience of the researcher and requires multiple tests for finding appropriate settings for their parameters. Chen [28], [29] proposed the flexible neural tree (FNT) which is a special neural network with automatic optimization of structure and parameters. FNT also has a few shortcomings related to multi-class classification. Firstly, there is a single root as output node making it unsuitable for multi-class classification. Secondly, to improve classification accuracy performance, it is necessary to deepen the model. Also, the cost of the parameter optimization algorithm increases significantly with increasing depth of the model.

Inspired by Deep Forest [30], this paper proposes to use deep flexible neural forest (DFNForest) for classification of cancer subtypes. DFNForest can directly handle multi-classification problems and deepen the model. The main idea of DFNForest is to transform a multi-classification into many binary classification problems in each forest. Meanwhile, through the cascade structure, the depth of the model is increased without introducing additional parameters. The main contributions can be summarized as follows. 1) proposing the use of fisher ratio in combination with neighborhood rough set to select most informative genes among a given gene expression. The fisher ratio is used to eliminate invalid genes and then neighborhood rough set is applied to reduce redundant genes. 2) proposing a model called DFNForest, which successfully solved the problem of FNT handling multi-classification problems and increased the depth of model. DFNForest transforms a multi-classification into many binary classification problems in each forest. Meanwhile, through the cascade structure, the depth of the model is increased without introducing additional parameters. 3) considering a small amount of gene expression data, so the number of cascading levels of DFNForest can be adaptively determined. When the accuracy of the next level no longer increases, it will stop increasing the number of levels.

The paper is organized as follows. The proposed gene selection method is introduced in Section 2. The classification with deep flexible neural forest is presented in Section 3. Section 4 presents the results of the proposed algorithm in comparison to state of the art methods and Section 5 provides a detailed discussion. Finally, the paper is concluded in section 6.

II. GENE SELECTION METHODS

Gene expression data generally comprises of thousands of genes, however, the number of available samples is usually small. Among thousands of features in gene expression data, only a few genes are actually associated with cancer subtypes whereas the rest may be considered as redundant or noisy features. Therefore, gene selection can be considered as dimensionality reduction problem that attempts to select important genes while maintaining the classification accuracy of original genes [31]. In the subsections below, we outline the proposed methodology, comprising of a combination of fisher ratio and neighborhood rough set for the purpose of gene selection.

A. FISHER RATIO

Fisher ratio [32] is a ratio between-class distances to within class distances. If there are two classes in a dataset, each sample can be labeled as $Y \in \{+1, -1\}$ and gene express vector i can be defined as $x_i = \{x_1^i, \ldots, x_n^i\}$. For each gene i, the standard deviation σ_i^+ (resp.,σ_i^-) and the mean μ_i^+(resp.,μ_i^-) are calculated and the fisher ratio F_i is calculated as:

$$F_i = \frac{(\mu_i^+ - \mu_i^-)^2}{(\sigma_i^+)^2 + (\sigma_i^-)^2} \tag{1}$$

Gene with highest F_i value is the most informative and the expression levels differ most on average in the two classes while also favoring those with small deviation in the respective classes [33]. Then the genes with high F_i values are selected as the top features.

B. NEIGHBORHOOD ROUGH SET

A neighborhood rough set (NRS) model [24], [25] can be used for processing of both discrete and continuous data sets while retaining information necessary to classify data accurately. Given a set of samples $U = \{x_1, x_2, \ldots x_n\}$, with A being a set of real-type features describing U, and D a decision attribute. If A generates a family of neighborhoods on the domain, it is called $NDT = \{U, A, D\}$ a neighborhood decision system. If D divides U into N equivalence classes: $X_1, X_2, \ldots, X_N$, $\forall B \subseteq A$, then the lower and upper approximations of decision D with respect to B can be represented as:

$$\underline{N}_B D = \bigcup_{i=1}^{N} \underline{N}_B X_i \tag{2}$$

$$\overline{N}_B D = \bigcup_{i=1}^{N} \overline{N}_B X_i \tag{3}$$

where $\underline{N_B} X = \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}$, $\overline{N_B} X = \{x_i | \delta_B(x_i) \cap X \neq \phi, x_i \in U\}$, $\delta_B(x_i)$ is a neighborhood information particle generated by attribute B and measure Δ.

The lower approximation of decision D is also called the decision positive domain, which is denoted as $POS_B(D)$ [34]. The size of the positive domain reflects the degree of separability of the classification problem in a given attribute space. The larger the positive domain, the fewer the overlapping of class boundaries. This ensures a better description of a classification problem using the selected set of attributes. Therefore, the dependency of the decision attribute D on the condition attribute B is defined as

$$\gamma_B(D) = Card(\underline{N}_B D)/Card(U) \tag{4}$$

Given a neighborhood decision system $NDT = \{U, A, D\}$, $B \subseteq A$, $\forall a \in A - B$, the importance of a relative to B can be defined as

$$SIG(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \tag{5}$$

Based upon the attribute importance index, a greedy attribute reduction algorithm is utilized. The algorithm takes the empty set as the starting point, calculates the attribute importance of all the remaining attributes each time, and selects the attribute with the largest attribute importance value to join the reduction set. This process is repeated until all remaining attributes have an importance of 0. This means that if new attributes are added, the system's dependency function values no longer change. The forward search algorithm ensures that important attributes are added to the reduction set first, so that important features are not omitted.

C. THE PROPOSED GENE SELECTION METHOD COMBINING FISHER RATIO AND NEIGHBORHOOD ROUGH SET

The fisher ratio method can effectively deal with noise in the gene expression data. It filters the noisy genes according to its contribution to classification, and thus effectively helps to identify the cancer subtype genes. The neighborhood rough set has the characteristics of not requiring discretization of continuous data and avoids information loss caused by data discretization, which can eliminate redundant genes.

If only fisher ratio is used as gene selection method, then the top k features are selected. Fisher ratio does not consider the relationship between genes and may select high correlated redundant genes, which not only increase the amount of calculation but also leads to incorrect classification results. When the neighborhood rough set is directly used to eliminate redundant genes, as the number of genes increases, the computational cost of the algorithm can be higher. In this paper, a feature gene selection algorithm based on fisher ratio and neighborhood rough set is proposed. The proposed algorithm effectively removes a large number of unrelated genes, reduce the space-time consumption of neighborhood rough set and training time of classifiers. The specific description can be seen in Algorithm 1.

Algorithm 1 Gene selection Using Fisher Ratio and Neighborhood Rough Set

```
Input: $NDT = \langle U, A, D \rangle$
Output: Reduction of $red$.
  Step 1: Using fisher ratio to select the top k features as A;
  Step 2: $\forall a \in A$: calculate neighborhood relationships $N_a$;
  Step 3: $\emptyset \rightarrow red$;
  Step 4: For any $a_i \in A - red$
          Calculate $SIG(a_i, red, D) = \gamma_{red \cup a}(D) - \gamma_{red}(D)$
  Step 5: Select $a_k$ to satisfy:
          $SIG(a_i, red, D) = \max(SIG(a_i, red, D))$
  Step 6: If $SIG(a_i, red, D) > 0$,
            $red \cup a_k \rightarrow red$
            go to Step 3
          else
            return $red$, end
```

III. CLASSIFICATION WITH THE PROPOSED DEEP FLEXIBLE NEURAL FOREST MODEL

In this section, a brief description of flexible neural tree followed by proposed flexible neural tree architecture into a deep flexible neural forest is presented.

A. FLEXIBLE NEURAL TREE

The function set F and terminal instruction set T are used to generate FNT model, which is defined as follows:

$$S = F \cup T = \{+_2, +_3, \ldots, +_N\} \cup \{x_1, \ldots, x_n\} \tag{6}$$

where $+_i(i = 2, 3, 4, \ldots, N)$ represents instructions of non-leaf nodes with i parameters. $x_1, x_2 \ldots, x_n$ are instructions of leaf nodes without parameters. To generate a flexible neural tree consider a nonterminal instruction $+_i(i = 2, 3, 4, \ldots, N)$, in which i values are randomly generated for non-leaf node and connecting weights between children [28], [29]. The following flexible activation function may be considered for the flexible neural tree:

$$f(x) = (1 + e^{-x})^{-1} \tag{7}$$

The output of a flexible neuron $+_n$ can be produced as follows.

$$\sum_n = \sum_{j=1}^{n} w_j * x_j \tag{8}$$

where $x_j(j = 1, 2, \ldots, n)$ are the inputs. The output of the node $+_n$ is computed as

$$out_n = f(sum_n) = (1 + e^{-sum_n})^{-1} \tag{9}$$

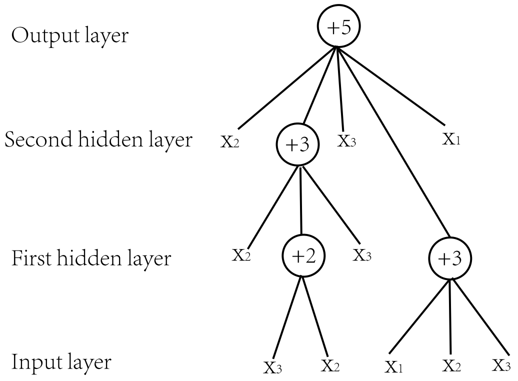

FIGURE 1. A typical representation of the FNT with function instruction set F ={$+_2, +_3, +_4, +_5$}, and terminal instruction set T ={x_1, x_2, x_3}.

A typical representation of the FNT is shown in Figure 1. The total output of the flexible neural tree can be recursively calculated from left to right by the depth-first method [35]. This flexible neural tree model allows over-layer connections and automatically select the structure, which is a sparse model and can get good generalization performance. The process of FNT optimization is mainly divided into two major steps: the optimization of the tree structure, followed by the parameter optimization. In this paper, the tree structure optimization algorithm is based on grammar guided genetic programming, and parameter optimization algorithm using particle swarm optimization.

B. TREE STRUCTURE OPTIMIZATION BY GRAMMAR GUIDED GENETIC PROGRAMMING

This paper uses grammar guided genetic programming (GGGP) to evolve the structure of FNT. An advantage of using GGGP is that it is not hindered by the disadvantage of having the function set and terminal instruction of the same type [36]. This helps avoid the problem of generating an invalid tree during crossover or mutation. A context-free grammar G is defined by the 4-tuple: $G = \{N, T, P, \Sigma\}$, where N is called nonterminal characters, and T is a set of terminals characters. The members of P are called rules of the grammar. Σ is the start symbol and an element of N. The rules of the grammar are expressed as x→y, where x belongs to N and y belongs to N∪T. There are four basic steps to generate grammar guided genetic programming:

(1) Initial population generation. In this step individual trees are generated randomly based on the grammar model.

(2) Evaluate each tree in the current generation. The tree has fitness values for each individual and hence they may be evaluated.

(3) Use either reproduction, mutation or crossover to produce the next generation. Follow this up be evaluating all trees in the next generation.

(4) Repeat until the best tree is found or termination criteria is met.

C. PARAMETER OPTIMIZATION WITH PSO

To optimize the parameters values we have used particle swarm optimization (PSO) algorithm [37]. Initially, the particles are generated randomly, and each particle represents a potential solution. Each particle has a position vector b_i associated with it. The complete population of particles moves through the problem space with a velocity a_i. At each step, a function f_i representing the fitness value of the solution is calculated to assess the suitability of the solution. Each particle keeps track of its own best position, and the best fitness of particle is stored in a vector p_i. Moreover, the best position among all the particles is stored as p_g. At each time step t, a new velocity for particle i is calculated by

$$a_i(t+1) = a_i(t) + c_1\varphi_1(p_i(t) - b_i(t)) + c_2\varphi_2(p_g(t) - b_i(t)) \tag{10}$$

where φ_1 and φ_2 are random numbers in [0,1], c_1 and c_2 are limited factors of position. Based on the changed velocities, each particle changes its position according to the following equation:

$$b_i(t+1) = b_i(t) + a_i(t+1) \tag{11}$$

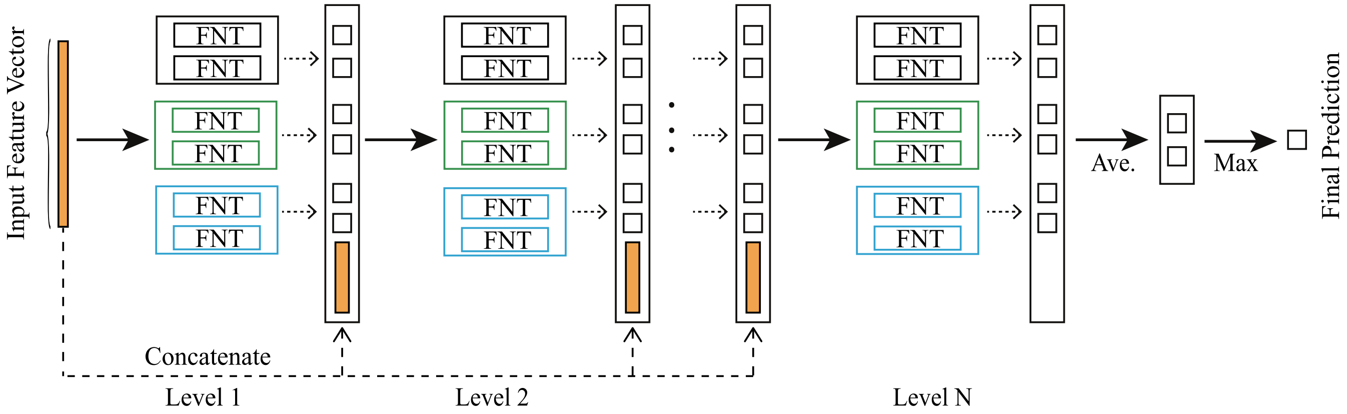

FIGURE 2. Illustration of the cascade forest structure. Three forests are generated by different grammar, the first forest (black) use function set F of $\{+_2, +_3, +_4\}$, the second forest (green) use $\{+_2, +_4, +_5\}$, and the last forest (blue) use function set F of $\{+_3, +_4, +_5\}$.

D. THE PROPOSED DEEP FLEXIBLE NEURAL FOREST MODEL

Flexible neural tree is a special neural network with automatic optimization of structure and parameters. However, it also faces problems. First, there is only one root node as the output node, which is not suitable for directly dealing with multi-classification problems. Secondly, in order to obtain better performance, it is necessary to deepen the model. However, this will result in an increase in the number of parameters, thus resulting in the high cost of the parameter optimization algorithm. To solve the above problems of the flexible neural tree, we propose a deep flexible neural forest (DFNForest), a novel flexible neural tree ensemble method to classify cancer subtypes.

Representation learning and model complexity are the reasons because of which deep neural networks have achieved such great success in visual and speech recognition tasks [30]. Representation learning refers to the processing of features layer by layer. In order to make FNT deeper without adding additional parameters, we adopt the structure of cascade forest. As illustrated in Figure 2, by processing the features layer by layer, new features can be obtained, and new features along with the original features are passed as input to the next layer.

In proposed model, each level is an ensemble of FNT. Although decision tree is used in multi-grain cascade forest (gcForest), the obvious disadvantage of decision tree is that it cannot be directly applied to continuous data and needs to discretize the data first. This may result in loss of information. The gene expression data is continuous data, so FNT is more suitable as a base classifier. Proposed method maintains the following advantages of FNT: 1) FNT is a sparse model and allows cross-layer connections, which avoids overfitting and achieve better generalization performance. 2) FNT automatically optimizes the structure and parameters. In addition, the proposed ensemble learning improves overall performance through multiple FNTs. To improve the diversity of ensemble learning, we generate different structures of FNT through different grammars. For simplicity, suppose that we use three forests and two FNTs in each forest. As is illustrated in Figure 2, the first forest uses function set F of $\{+_2, +_3, +_4\}$, the second forest uses $\{+_2, +_4, +_5\}$, and the last one uses $\{+_3, +_4, +_5\}$. In order to solve the problem of FNT dealing with multi-classification problems, the M-ary method is used to transform the multi-class problem into several two-class problems in a forest. For example, if it is a four-class problem, each forest needs to contain $k = \log_2 4 = 2$ FNTs, so the number of trees in the forest is determined by the classification problem.

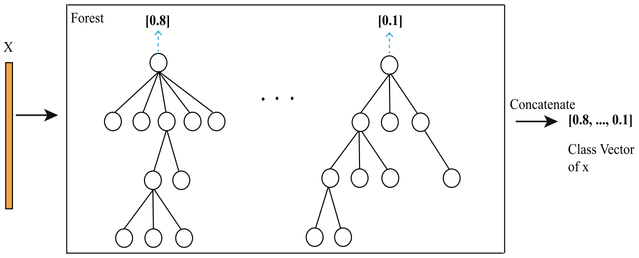

FIGURE 3. Illustration of class vector generation. Each FNT will generate an estimated value and then concatenate together.

As is demonstrated in Figure 3, given an example, each FNT will generate an estimated value. The estimated value forms a class vector concatenated with the original input feature vector as the input to the next level. For example, suppose there are four classes, then each forest will produce a two-dimensional class vector; thus, the next level of cascade will receive 6 (2 × 3)augmented features. The training set will be divided into two parts, one for training and one for validation. When a new level is added, the entire cascade will be verified by the validation set. When the accuracy no longer increases, the levels will stop increasing. In this way, the number of cascade levels is determined automatically. This allows it to be used on datasets of different sizes, suitable for small-scale gene expression data.

The DFNForest is a novel deep learning model that provides an alternative to deep neural networks. Compared with deep neural network, one of its significant advantages is the automatic design of the structure. The structure of FNT in each forest is automatically selected by the tree structure optimization algorithm, and the cascade levels are determined adaptively. DFNForest is an ensemble of FNT, however, it resolves the shortcoming of FNT to deal with multi-classification problems by converting it into many binary classification problems in each forest. Through the cascade structure, the depth of the model is increased without introducing additional parameters.

IV. EXPERIMENTAL RESULTS

A. DATASETS AND PARAMETERS

We conducted cancer subtype predictions on RNA-Seq gene expression datasets of three cancer types (BRCA (breast invasive carcinoma), GBM (glioblastoma multiforme), LUNG (lung cancer)) downloaded from The Cancer Genome Atlas (TCGA) [38]. Specifically, in BRCA data, there are four basic subtypes Basal-like, HER2-enriched, Luminal-A and Luminal-B in 514 samples. In GBM data, there are Classical, Mesenchymal, Neural and Proneural subtypes in 164 samples whereas for the LUNG data, there are Bronchioid, Magnoid and Squamoid subtypes in 275 samples. The details of three cancer types are shown in Table 1. The above-mentioned data were logarithm transformed and we filtered out genes with an average of less than 5.0 and a variance of less than 1.0 [39]. The remaining genes were used for later analysis. For each dataset, the corresponding clinical data is downloaded from TCGA and used the clinical subtype information to label each sample, as this provided the reference (ground truth) for assessing the performance of the proposed algorithm. Three different forests were developed for experiments. For these three forests, the function set F was set to $\{+_2, +_3, +_4\}$, $\{+_2, +_4, +_5\}$, $\{+_3, +_4, +_5\}$ respectively. The parameters used for FNTs are listed in Table 2.

TABLE 1. The detail of the three cancer types.

Dataset	Sample	Gene	Class
BRCA	514	4247	4
GBM	164	3398	4
LUNG	275	4596	3

TABLE 2. Parameter settings of FNT.

Parameter	value
Population size	50
Crossover probability	0.4
Mutation probability	0.01
C1	2.0
C2	2.0
Vmax	2.0

B. GENE SELECTION AND CANCER SUBTYPE CLASSIFICATION ON BRCA

The BRCA dataset was used to test both the proposed gene selection method and classification performance of proposed DFNForest model. To test the performance of proposed gene selection method, k-nearest neighbor (KNN), support vector machine (SVM), multi-layer perception (MLP), random forest (RF) and multi-grained cascade forest (gcForest) classifiers are used. We used raw data as the input as well as the data obtained from fisher ratio, neighborhood rough set (NRS), minimal redundancy maximal relevance (MRMR) and from fisher ratio combined with neighborhood rough set. Cross-validation results indicated that the best performance in BRCA subtype classification experiment was obtained when the parameter k in fisher ratio is set to 30 and the neighborhood parameter is set to 0.35. Next, the performance of DFNForest is compared with KNN, SVM, MLP, RF and gcForest classifiers. To be fair, 5-fold cross-validation to assess the overall accuracy of the different methods and randomly select 4/5 samples for training data and 1/5 samples for testing. The experimental results of classification accuracies are shown in Table 3.

TABLE 3. Classification accuracies for the selected BRCA genes.

Method	Gene	KNN	SVM	MLP	RF	gcForest	DFNForest
Original data	4247	0.816	0.888	0.870	0.880	0.880	0.928
Fisher ratio	30	0.880	0.888	0.880	0.896	0.920	0.928
NRS	8	0.816	0.816	0.816	0.870	0.880	0.896
MRMR	28	0.888	0.896	0.888	0.896	0.928	0.930
Fisher + NRS	7	0.880	0.904	0.896	0.904	0.928	0.936

As we can see, fisher ratio selects the top 30 genes from 4247 genes. After selecting the informative genes using the fisher ratio, classification accuracies increase in most classifiers as compared to when the original data was used. This indicates that fisher ratio can remove redundant genes. NRS selects 8 informative genes and therefore the number of features is greatly reduced however this results in low classification accuracy. This may be associated to elimination of non-redundant genes in the process of dimensionality reduction. The classification accuracy of MRMR method is higher, however, the number of genes is more than proposed gene selection methods. When the 7 selected genes obtained by the combination of fisher ratio and neighborhood rough set are used, it achieves the highest classification accuracy with the fewest number of genes. This may be attributed to removal of noise or redundant information from the sample dataset. A detailed description of the selected 7 significant genes of breast cancer subtypes is shown in Table 4. All of these selected genes have been identified in other studies [40], [41] to be biologically relevant to BRCA cancer.

Compared with the conventional methods KNN, SVM MLP and RF, gcForest and DFNForest have higher accuracy. DFNForest (93.6%) performs better than gcForest (92.8%), which may be associated to the fact that DFNForest is more suitable for processing continuous data. Furthermore, to evaluate the overall performance of KNN, SVM, MLP, RF, gcForest and DFNForest on the BRCA dataset, the average precision, recall and F-1 score of each method were considered. As shown in Figure 4, DFNForest model has achieved good performance in the BRCA subtype classification.

TABLE 4. Descriptions of the BRCA genes selected by fisher ratio and neighborhood rough set.

Gene name	Gene description
ERBB2	ErbB2 is a 185kDa cell membrane receptor encoded by the proto-oncogene erbB-2.
TMEM45B	Transmembrane protein 45B is a member of TMEMs.
AURKA	AURKA acts as an enzyme to directly or indirectly activate an oncogene or inactivate a tumor suppressor gene.
FOXC1	FOXC1 protein activates transcription by a promoter or other transcription factor acting on target cells.
CEP55	CEP55 is a mitotic phosphoprotein that plays a key role in cytokinesis, the final stage of cell division.
ESR1	ESR1 gene mutation is closely related to ER-positive breast cancer, and ER is a nuclear protein encoded by ERS1 gene.
SFRP1	The SFRP1 gene is one of the secreted glycoprotein families and is currently recognized as an epigenetic marker. Its epigenetics involves multiple tumors.

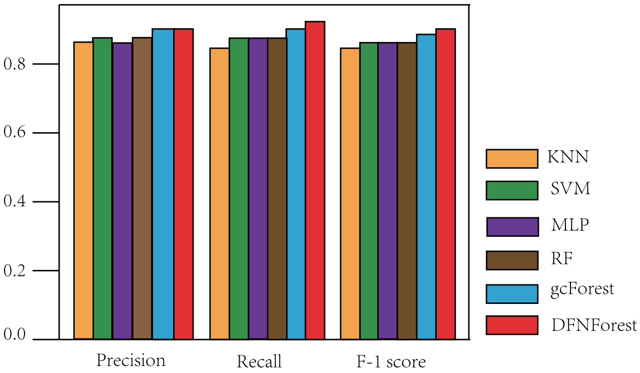

FIGURE 4. Comparison of overall performance of KNN, SVM, MLP, RF, gcForest and DFNForest on BRCA datasets. The average precision, recall and F-1 score were evaluated.

C. GENE SELECTION AND CANCER SUBTYPE CLASSIFICATION ON GBM

To test the performance of proposed gene selection method, we compared it with the raw data, fisher ratio, neighborhood rough set (NRS) and minimal redundancy maximal relevance (MRMR). Cross-validation has shown that the parameter k in fisher ratio is set to 28 and the neighborhood parameter is set to 0.35, which gives better performance in the GBM subtype classification experiment. The second part was classification experiments for comparing DFNForest with KNN, SVM, MLP, RF and gcForest classifiers. We used 5-fold cross-validation to assess the overall accuracy of the different methods and randomly selected 4/5 samples for training and 1/5 samples for testing.

The experimental results of classification accuracies on GBM datasets are shown in Table 5. It can be seen that fisher ratio selects the top 28 genes from 3398 genes. Only selecting the discriminant gene using the fisher ratio, the classification accuracies increase in most classifiers compared with the raw data. NRS selects 6 informative genes and the number of features is greatly reduced but the accuracy is the lowest, which may be that the genes associated with the classification are also eliminated. The classification accuracy of proposed gene selection method is just lower than MRMR, but the number of genes is much smaller than MRMR. A detailed description of the selected 6 significant genes of glioblastoma cancer subtypes is demonstrated in Table 6. The location where glioblastoma cancer occurs is the brain, and it has been proved in other studies [42], [43] that the expression of the six selected genes is closely related to the function of the brain.

Compared with the traditional methods KNN, SVM, MLP and RF, gcForest and DFNForest have higher accuracy. DFNForest (84.2%) performs better than gcForest (76.3%). Furthermore, to evaluate the overall performance of KNN, SVM, MLP, RF, gcForest and DFNForest on the GBM datasets, we examined the average precision, recall and F-1 score of each method. As presented in Figure 5, our proposed classification model has achieved better results as compared to KNN, SVM, MLP, RF and gcForest methods.

TABLE 5. Classification accuracies for the selected GBM genes.

Method	Gene	KNN	SVM	MLP	RF	gcForest	DFNForest
Original data	3398	0.684	0.632	0.658	0.684	0.737	0.816
Fisher ratio	28	0.710	0.658	0.684	0.710	0.763	0.816
NRS	6	0.658	0.658	0.632	0.684	0.710	0.737
MRMR	23	0.710	0.710	0.737	0.763	0.816	0.854
Fisher + NRS	6	0.710	0.684	0.710	0.737	0.763	0.842

TABLE 6. Descriptions of the GBM genes selected by fisher ratio and neighborhood rough set.

Gene name	Gene description
PLAUR	This gene encodes the receptor for urokinase plasminogen activator, given its role in localizing and promoting plasmin formation, likely influences many normal and pathological processes.
B4GALT1	This gene is one of seven beta-1,4-galactosyltransferase genes. They encode type II membrane-bound glycoproteins that appear to have exclusive specificity for the donor substrate UDP-galactose.
BASP1	This gene encodes a membrane bound protein with several transient phosphorylation sites and PEST motifs. Alternative splicing results in multiple transcript variants that encode the same protein.
ZDHHC22	Restricted expression toward brain.
SEPP1	This gene encodes a selenoprotein that is predominantly expressed in the liver and secreted into the plasma. Mice lacking this gene exhibit neurological dysfunction, suggesting its importance in normal brain function.
CDH4	This gene is a classical cadherin from the cadherin superfamily. This cadherin is thought to play an important role during brain segmentation and neuronal outgrowth.

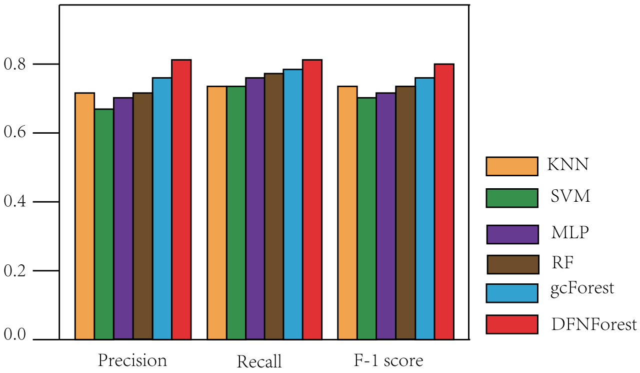

FIGURE 5. Comparison of overall performance of KNN, SVM, MLP, RF, gcForest and DFNForest on GBM datasets. The average precision, recall and F-1 score were evaluated.

TABLE 7. Classification accuracies for the selected LUNG genes.

Method	Gene	KNN	SVM	MLP	RF	gcForest	DFNForest
Original data	4596	0.710	0.776	0.746	0.791	0.830	0.865
Fisher ratio	29	0.716	0.791	0.780	0.791	0.830	0.865
NRS	6	0.710	0.716	0.746	0.746	0.776	0.791
MRMR	26	0.780	0.791	0.780	0.806	0.830	0.880
Fisher + NRS	8	0.746	0.806	0.780	0.806	0.865	0.880

D. GENE SELECTION AND CANCER SUBTYPE CLASSIFICATION ON LUNG

The proposed gene selection method is also tested on third dataset along with existing and DFNForest classification methods. It was observed that if parameter k in fisher ratio is set to 29 and the neighborhood parameter is set to 0.35, the best performance is obtained for LUNG subtype classification. For assessing the performance of the proposed DFNForest classifier, 5-fold cross-validation was performed. In each data set, 4/5 samples were randomly selected for training and 1/5 samples for testing. The experimental results of classification accuracy are shown in Table 7.

From Table 7, it is clear that classification accuracies increase for most classifiers after dimensionality reduction using fisher ratio. This is because of removal of noise in the raw data, which affects the classification results. NRS selects 6 informative genes and the number of features is greatly reduced but the accuracy is the lowest. Although classification accuracy of MRMR method is higher, the number of genes is more than proposed gene selection methods. In addition, the classification accuracy of proposed gene selection method is higher with fewer number of genes. A detailed description of the selected 8 significant genes of lung cancer subtypes is shown in Table 8. The selected 8 genes have been shown to be closely related to the occurrence and development of cancer in other studies [44].

Compared with the conventional methods KNN, SVM, MLP, RF and gcForest, the proposed DFNForest method has higher classification accuracy. DFNForest with a classification accuracy of 88.0% outperforms gcForest which has a classification accuracy of 86.5%. Moreover, to compare the overall performance of the proposed method, the results are compared with KNN, SVM, MLP, RF, gcForest and DFNForest on the LUNG datasets. Average precision, recall and F-1 score for each method were ananlyzed. As shown in Figure 6, proposed classification model has achieved good performance in the LUNG subtype classification as it outperforms the state-of-the-art methods.

TABLE 8. Descriptions of the LUNG genes selected by fisher ratio and neighborhood rough set.

Gene name	Gene description
HLA-DPA1	HLA-DPA1 belongs to the HLA class II alpha chain paralogues. It plays a central role in the immune system by presenting peptides derived from extracellular proteins.
RAC2	The encoded protein localizes to the plasma membrane, where it regulates diverse processes, such as secretion, phagocytosis, and cell polarization.
ARHGAP20	The protein encoded by this gene is an activator of RHO-type GTPases, transducing a signal from RAP1 to RHO and impacting neurite outgrowth.
FAM72B	Broad expression in lymph node.
NAMPT	The protein belongs to the nicotinic acid phosphoribosyltransferase family and is thought to be involved in many important biological processes, including metabolism, stress response and aging.
FGA	This gene encodes the alpha subunit of the coagulation factor fibrinogen, which is a component of the blood clot.ITGB2 The encoded protein plays an important role in immune response and defects in this gene cause leukocyte adhesion deficiency.
DIAPH3	This gene encodes a member of the diaphanous subfamily of the formin family. Members of this family are involved in actin remodeling and regulate cell movement and adhesion
ITGB2	The encoded protein plays an important role in immune response and defects in this gene cause leukocyte adhesion deficiency.

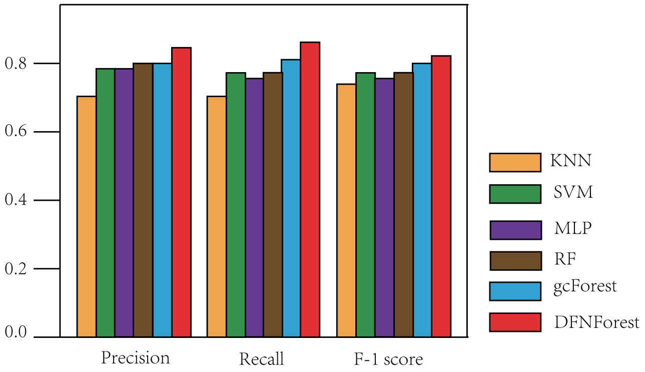

FIGURE 6. Comparison of overall performance of KNN, SVM, MLP, RF, gcForest and DFNForest on LUNG datasets. The average precision, recall and F-1 score were evaluated.

V. DISCUSSION

The DFNForest is a novel deep learning model that provides an alternative to deep neural networks. Compared with the deep neural network, one of its significant advantages is the automatic design of its structure. The structure of FNT in each forest is automatically selected by the tree structure optimization algorithm, and the cascade levels are determined adaptively. DFNForest is an ensemble of FNT, but it solves the shortcomings of FNT. It is capable of dealing with multi-classification problems by transforming a multi-classification problem into many binary classification problems in each forest. Through its cascade structure, the depth of the model can be increased without introducing additional parameters. We compared DFNForest with KNN, SVM, MLP, RF and gcForest classification methods on RNA-Seq gene expression cancer datasets. We found: 1) the deep forest models (DFNForest and gcForest) are superior to other traditional classification methods on most of cancer datasets. This may be through a cascading structure, which can extract more meaningful features layer by layer. 2) DFNForest outperformed the gcForest on most of cancer datasets. This illustrates that proposed model which utilizes FNT as the base classifier is more suitable for processing continuous gene expression data.

For the proposed combination of the fisher ratio and the neighborhood rough set gene selection method, we found: 1) our proposed gene selection method performed better than using fisher ratio and neighborhood rough set separately on cancer datasets, perhaps because these two methods complement each other very well. Their combination considers both the relationship between genes and class labels and the relationship within genes. 2) the proposed gene selection method outperformed than classic MRMR method on most cancer datasets with fewer number of genes. Although proposed gene selection method and DFNForest give better predictions for cancer subtype than other methods, only RNA-seq gene expression data is used in this paper. It has been demonstrated that integration of different types of genomic data contributes to cancer subtype classification [45]–[47]. Therefore, we can consider using the proposed DFNForest model to integrate different kinds of genomic data for cancer subtype classification.

VI. CONCLUSIONS

Classification of cancer subtypes is important for the diagnosis and treatment of cancer. However, the RNA-seq gene expression data used for cancer subtype classification has the nature of high dimensionality and small sample size. In order to avoid overfitting, we proposed the combination of the fisher ratio and the neighborhood rough set to select the informative genes which can remove both noise and redundancy from the dataset. Most importantly, a model called DFNForest was proposed, which successfully solved the problem of FNT handling multi-classification problems and increased the depth of FNT. The main idea of the proposed DFNForest is to transform a multi-classification into many binary classification problems in each forest. Meanwhile, through the cascade structure, the depth of the proposed model is increased without introducing additional parameters. DFNForest is well suited for processing small-scale biology data because its number of levels can be adaptively determined. Experiments show that proposed gene selection method has higher accuracy even after dimensionality reduction and the proposed DFNForest model has better performance in the classification of BRCA, GBM, LUNG cancer subtypes as compared to conventional methods. In conclusion, the proposed DFNForest model provides an option to classify cancer subtypes by using deep learning on small-scale biology datasets.

REFERENCES

[1] J. Stingl and C. Caldas, "Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis," *Nature Rev. Cancer*, vol. 7, no. 10, pp. 791–799, Oct. 2007.

[2] G. Bianchini *et al.*, "Prognostic and therapeutic implications of distinct kinase expression patterns in different subtypes of breast cancer," *Cancer Res.*, vol. 70, no. 21, pp. 8852–8862, 2010.

[3] L. M. Heiser *et al.*, "Subtype and pathway specific responses to anticancer compounds in breast cancer," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 109, no. 8, pp. 2724–2729, Feb. 2012.

[4] A. Prat, J. S. Parker, and O. Karginova, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer," *Breast Cancer Res.*, vol. 12, no. 5, p. R68, 2010.

[5] M. J. Jahid, T. H. Huang, and J. Ruan, "A personalized committee classification approach to improving prediction of breast cancer metastasis," *Bioinformatics*, vol. 30, no. 13, pp. 1858–1866, Mar. 2014.

[6] S. L. Wang, X. Li, S. Zhang, J. Gui, and D. S. Huang, "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction," *Comput. Biol. Med.*, vol. 40, no. 2, pp. 179–189, Feb. 2010.

[7] J. Dai and Q. Xu, "Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 211–221, 2013. doi: 10.1016/j.asoc.2012.07.029.

[8] W. H. Chan *et al.*, "Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme," *Comput. Biol. Med.*, vol. 77, pp. 102–115, Oct. 2016.

[9] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[10] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature selection for gene expression using model-based entropy," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 1, pp. 25–36, Jan. 2010.

[11] L. Goh, Q. Song, and N. Kasabov, "A novel feature selection method to improve classification of gene expression data," in *Proc. Asia-Pacific Bioinf. Conf.*, Jan. 2004, pp. 161–166.

[12] M. A. Muharram and G. D. Smith, "Evolutionary feature construction using information gain and gini index," in *Proc. Eur. Conf. Genetic Program.*, 2004, pp. 379–388.

[13] C. Liao, S. Li, and Z. Luo, "Gene selection using wilcoxon rank sum test and support vector machine for cancer classification," in *Proc. Int. Conf. Comput. Inf. Sci.*, 2007, pp. 57–66.

[14] I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Proc. Eur. Conf. Mach. Learn.*, 1994, pp. 171–182.

[15] L. Li, C. Weinberg, T. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, Dec. 2001.

[16] J. Khan *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Med.*, vol. 7, no. 6, pp. 673–679, 2001.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[18] Y. Liang, F. Zhang, J. Wang, T. Joshi, Y. Wang, and D. Xu, "Prediction of drought-resistant genes in arabidopsis thaliana using SVM-RFE," *PLoS ONE*, vol. 6, no. 7, Jul. 2011, Art. no. e21750.

[19] E. Tapia, P. Bulacio, and L. Angelone, "Sparse and stable gene selection with consensus SVM-RFE," *Pattern Recognit. Lett.*, vol. 33, no. 2, pp. 164–172, Jan. 2012.

[20] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, "Gene selection for tumor classification using neighborhood rough sets and entropy measures," *J. Biomed. Inform.*, vol. 67, pp. 59–68, Mar. 2017.

[21] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, Oct. 1982.

[22] P. Maji and S. Paul, "Rough set based maximum relevance-maximum significance criterion and Gene selection from microarray data," *Int. J. Approx. Reasoning*, vol. 52, no. 3, pp. 408–426, Mar. 2011.

[23] T. K. Liu, Y. P. Chen, Z. Y. Hou, C. C. Wang, and J. H. Chou, "Noninvasive evaluation of mental stress using by a refined rough set technique based on biomedical signals," *Artif. Intell. Med.*, vol. 61, no. 2, pp. 97–103, Jun. 2014.

[24] Q. Hu, J. Liu, and D. Yu, "Mixed feature selection based on granulation and approximation," *Knowl.-Based Syst.*, vol. 21, no. 4, pp. 294–304, May 2008.

[25] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Inf. Sci.*, vol. 178, no. 18, pp. 3577–3594, 2008.

[26] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics*, vol. 34, no. 21, pp. 3727–3737, Nov. 2018.

[27] A. K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data," *Neural Comput. Appl.*, vol. 29, no. 12, pp. 1545–1554, Jun. 2016.

[28] Y. Chen, B. Yang, J. Dong, and A. Abraham, "Time-series forecasting using flexible neural tree model," *Inf. Sci.*, vol. 174, nos. 3–4, pp. 219–235, 2005.

[29] Y. Chen, B. Yang, and A. Abraham, "Flexible neural trees ensemble for stock index modeling," *Neurocomputing*, vol. 70, nos. 4–6, pp. 697–703, Jan. 2007. doi: 10.1016/j.neucom.2006.10.005.

[30] Z. H. Zhou and J. Feng, "Deep forest: Towards an alternative to deep neural networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1–6.

[31] W. Lu, Z. Li, and J. Chu, "A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning," *Comput. Biol. Med.*, vol. 83, pp. 157–165, Apr. 2017.
[32] K. Z. Mao, "RBF neural network center selection based on Fisher ratio class separability measure," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1211–1217, Sep. 2002.
[33] Y. Chen and Y. Zhao, "A novel ensemble of classifiers for microarray data classification," *Appl. Soft Comput.*, vol. 8, no. 4, pp. 1664–1669, Sep. 2008.
[34] M. L. Hou *et al.*, "Neighborhood rough set reduction-based gene selection and prioritization for gene expression profile analysis and molecular cancer sclassification," *J. Biomed. Biotechnol.*, vol. 2010, pp. 1110–1117, Aug. 2015.
[35] Y. Chen, L. Peng, and A. Abraham, "Gene expression profiling using flexible neural trees," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, 2006, pp. 1121–1128.
[36] P. Wu and Y. Chen, "Grammar guided genetic programming for flexible neural trees optimization," in *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science), vol. 4426. Sep. 2007, pp. 964–971.
[37] J. Kennedy, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw.*, 2002.
[38] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, Sep. 2013.
[39] Y. Guo, S. Liu, and Z. Li, "BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data," *BMC Bioinf.*, vol. 19, p. 118, Jul. 2018.
[40] A. M. Sieuwerts *et al.*, "mRNA and microRNA expression profiles in circulating tumor cells and primary tumors of metastatic breast cancer patients," *Clin. Cancer Res.*, vol. 17, no. 11, pp. 3600–3618, Feb. 2011.
[41] M. P. Piechocki, G. H. Yoo, S. K. Dibbley, and F. Lonardo, "Breast cancer expressing the activated HER2/neu is sensitive to gefitinib *in vitro* and *in vivo* and acquires resistance through a novel point mutation in the HER2/neu," *Cancer Res.*, vol. 67, no. 14, p. 6825, Jul. 2007.
[42] W. B. Pope *et al.*, "Relationship between gene expression and enhancement in glioblastoma multiforme: Exploratory DNA microarray analysis," *Radiology*, vol. 249, no. 1, p. 268, Oct. 2008.
[43] W. Y. Cheng *et al.*, "A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma," *Plos One*, vol. 7, no. 4, 2012, Art. no. e34705.
[44] D. T. Ross *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genet.*, vol. 24, no. 3, pp. 227–235, Mar. 2000.
[45] M. Bhattacharyya, J. Nath, and S. Bandyopadhyay, "MicroRNA signatures highlight new breast cancer subtypes," *Gene*, vol. 556, no. 2, pp. 192–198, Feb. 2015.
[46] N. G. Bediaga *et al.*, "DNA methylation epigenotypes in breast cancer molecular subtypes," *Breast Cancer Res.*, vol. 12, no. 5, p. R77, 2010.
[47] L. Cantini *et al.*, "MicroRNA–mRNA interactions underlying colorectal cancer molecular subtypes," *Nature Commun.*, vol. 6, p. 8878, Nov. 2015.

JING XU received the B.S. degree in computer science from Jinan University, Jinan, China, in 2017, where she is currently pursuing the master's degree with the School of Information Science and Engineering. Her research interests include computational intelligence, bioinformatics, and machine learning.

PENG WU received the M.S. degree from the School of Information Science and Engineering, University of Jinan, Jinan, China, in 2007, and the Ph.D. degree from the College of Information Science and Technology, Beijing Normal University, Beijing, China, in 2016. He is currently a Researcher with the Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan. His main research interests include computational intelligence, bioinformatics, and machine learning.

YUEHUI CHEN received the B.Sc. degree from the Department of Mathematics (major in control theory) from Shandong University, China, in 1985, and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Kumamoto University, Japan, in 2001. From 2001 to 2003, he was a Senior Researcher with Memory-Tech Corporation, Tokyo. Since 2003, he has been a member with the faculty of the School of Information Science and Engineering, University of Jinan, where he is currently the Head of the Laboratory of Computational Intelligence. He has authored and co-authored over 100 papers. His research interests include evolutionary computation, neural networks, fuzzy systems, hybrid computational intelligence and their applications in time-series prediction, system identification, and intelligent control. He is a member of the IEEE Systems, Man and Cybernetics Society and the Computational Intelligence Society. He is also a member of the Editorial Boards of several technical journals and a member of the program committee of several international conferences.

QINGFANG MENG received the master's and Ph.D. degrees from the School of Information Science and Engineering, Shandong University, China, in 2005 and 2008, respectively. Since 2008, she has been an Associate Professor and a tutor of master with the School of Information Science and Engineering, University of Jinan, Jinan, China. She has published more than 40 research papers, where 11 are indexed by SCI and more than 30 are indexed by EI. Her research interests include nonlinear time series analysis, biomedical signal processing, and computational intelligence.

HUSSAIN DAWOOD received the M.S. and Ph.D. degrees in computer application technology from Beijing Normal University, Beijing, China, in 2012 and 2015, respectively. He is currently an Assistant Professor with the College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia. His current research interests include image processing, pattern recognition, and machine learning.

MUHAMMAD MURTAZA KHAN received the B.E. degree (Hons.) in electrical engineering from the University of Engineering and Technology at Taxila, Pakistan, in 1999, the M.S. degree in computer software engineering from the National University of Sciences and Technology (NUST), Rawalpindi, Pakistan, in 2005, and the M.S. and Ph.D. degree in image processing from the Grenoble Institute of Technology, Grenoble, France, in 2006 and 2009, respectively. He has been an Assistant Professor with the College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia, since 2015. Prior to his appointment at UJ, he has been an Assistant Professor of electrical engineering with the School of Electrical Engineering and Computer Science, NUST, Islamabad, Pakistan, since 2010. During his time at NUST, he was the CoPI of a collaborative project, related to real-time stitching and rendering of video from multiple cameras using multiple projectors, with the Electronics and Telecommunication Research Institute, South Korea, from 2011 to 2015. From 2000 to 2004, he was a Software Engineer and then a Senior Software Engineer with Streaming Networks Pvt. Ltd., Islamabad, where his duties revolved around the development of video drivers and codecs optimized for Philips TriMedia Processor.

• • •