IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 12, 2021, accepted December 21, 2021, date of publication December 30, 2021,
date of current version January 11, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3139595

A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm

ZEXIAN HUANG AND DAQI CHEN
School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006, China

Corresponding author: Daqi Chen (daqichen@gzhu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62003105; and in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China, under Grant ICT2021B09.

ABSTRACT Breast cancer is a neoplastic disease which seriously threatens women's health. It is regard as the most common cause of cancer death in women. Accurate detection and effective treatment are of vital significance to lower the death rate of breast cancer. In recent years, machine learning technique has been considered to be an effective method for accurate diagnosis of various diseases, among which Random Forest (RF) has been widely applied. However, decision trees with poor classification performance and high similarity may be generated during the training process, which affects the overall classification performance of the model. In this paper, a Hierarchical Clustering Random Forest (HCRF) model is developed. By measuring the similarity among all the decision trees, the hierarchical clustering technique is used to carry out clustering analysis on decision trees. The representative trees are selected from divided clusters to construct the hierarchical clustering random forest with low similarity and high accuracy. In addition, we use Variable Importance Measure (VIM) method to optimize the selected feature number for the breast cancer prediction. Wisconsin Diagnosis Breast Cancer (WDBC) database and Wisconsin Breast Cancer (WBC) database from the UCI (University of California Irvine) Machine Learning repository are employed in this study. The performance of the proposed method is evaluated by utilizing accuracy, precision, sensitivity, specificity and AUC (Area Under ROC Curve). Experimental results indicate that the classification based on HCRF algorithm with VIM as a feature selection method reaches the best accuracy of 97.05% and 97.76% compared to Decision Tree, Adaboost and Random Forest on both the WDBC and WBC datasets. The method proposed in this study is an effective tool for diagnosing breast cancer.

INDEX TERMS Breast cancer, hierarchical clustering random forest algorithm, feature selection.

I. INTRODUCTION

Breast cancer is one of the most important problems in women's health and has become the highest incidence of malignant tumor in women globally [1], [2]. According to the latest global cancer data in 2020, breast cancer has overtaken lung cancer as the world's leading cancer.

Accurate and early diagnosis can increase the probability for patients to gain timely and effective treatment and thus reduce the mortality of breast cancer [3]. The diagnosis for breast cancer mainly includes pathological diagnosis and imaging diagnosis. Compared with pathological diagnosis, imaging diagnosis is a non-invasive diagnostic means widely concerned in recent years [4]–[6]. However, imaging diagnosis often needs to be confirmed after the visualization of the tumor and may miss the early detection. Fine Needle Aspiration biopsy (FNA) is a minimally invasive pathological diagnosis method based on cell morphology [7], which have great potential to provide high accuracy and low false positive diagnosis. First, a fine needle is used to extract the cells from the breast tumor. Then the cell size, thickness, uniformity, smoothness and other data are statistically analyzed. Finally, these data are used to predict the new cases.

Machine learning is a process of utilizing data to discover latent information that may not been easily identified [8], which is suitable for prediction with FNA data.

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

Random forest is one of the ensemble learning methods widely used in disease detection. The randomness of it is reflecting in two aspects, the samples and features used in the decision trees. Therefore, compared with a decision tree, random forest reduces the possibility of over-fitting. Moreover, it is less susceptible to unbalanced samples, noise and outliers, and usually achieves high prediction accuracy [9]. Scholars have already applied random forest to the diagnosis of different kinds of diseases [10]–[14]. The exiting approaches improving the random forest mainly focus on the improvement of the decision tree algorithm, the modification of voting method, the preprocessing of the data set and optimization of feature selection. However, a random forest classifier is more effective only if the decision trees in the random forest classifier are diverse [15].

Clustering analysis is a significant process in the field of data mining [16], [17], which is a process of classifying objects into groups according to their similarities [18]. Various clustering methods have been proposed, mainly including k-means [19], hierarchical techniques [20], [21], density-based techniques [22], [23], grid-based algorithms [24]. In Chavent's work [25], he develops a technique which combines feature selection and variables clustering. Another study on the application of random clustering forest base on extended belief rule-based (EBRB) system has been published by Murugan *et al.* [26].

Feature selection is also a vital procedure before a classification task since biomedical data sets are often characterized by high dimensions which may include some irrelevant and redundant features [27]–[29]. Hou proposed a Sparse matrix regression (SMR) feature selection algorithm based on matrix data and sparse constraints [30], which was effective in the application of scene classification. Luo proposed a semi-supervised feature selection method based on insensitive sparse regression (ISR) and applied it to video semantic recognition [31]. In order to efficiently process high-dimensional non-Gaussian data in face recognition, an adaptive discriminant analysis (ADA) method was proposed in [32], which can distinguish the importance of each data point. Variable Importance Measure (VIM) is also a method of ranking the importance of features according to the Gini index [10], [33], which could help to pick up those most important ones to improve the classification performance.

In this paper, a breast cancer diagnosis methodlogy that uses VIM for feature selection and Hierarchical Clustering Random Forest (HCRF) for classification is proposed. We first generate a traditional random forest with several decision trees. Then, we group the decision trees into several clusters according to the similarity between them. Eventually, the decision trees in each cluster with the best performance are retained to construct the HCRF model. Then, VIM is adopted to extract the most significant tumor features of breast cancer for model construction. The optimal feature subset is obtained by deleting the less important features so as to improve the performance of the HCRF classifier, including training time, generalization ability and simplicity. Finally, grid search algorithm is used to optimize the parameters of our model. Experimental results show that the classification based on HCRF algorithm with VIM as a feature selection method is a practical way for in the early diagnosis of breast cancer.

The main contributions are summarized as follows:

1. An intelligent diagnose algorithm HCRF is proposed for the detection of breast cancer.

2. A feature selection method called VIM which uses the Gini index to measure the importance of each feature is used before classification to help us select the optimal feature subset.

3. Hierarchical clustering is introduced to improve the diversity and classification ability of decision trees in the random forest. This proposed method has great reference value for designing structural diversity using other types of basic learners or other ensemble learning algorithms.

4. The developed model is superior to other classifiers such as decision tree, Adaboost, random forest and performs better than other state of the art machine learning models for breast cancer detection.

The remaining parts of the paper are arranged as follows: In section II, we describe the detailed information of the database as well as the proposed method and give the evaluation metrics. Section III evaluates our proposed method and demonstrates the experimental results and discussion. Section IV is about the summary of this paper and future work.

II. MATERIAL AND METHODOLOGY

A. DATASET DESCRIPTION

The Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Breast Cancer (WBC) datasets used in this research are obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [34], [35]. The WDBC database contains 569 instances. Each instance consists of 30 attributes and a class label. Features are obtained from a digitized image of a FNA of a breast mass, which describe the traits of the cell nuclei [36], [37]. The WBC dataset includes 699 samples. Each sample has 9 features and a class label. 16 instances that include missing value are removed from the WBC dataset. TABLE 1 and TABLE 2 give the detailed information of the database.

B. RANDOM FOREST CLASSIFIER

Random Forest (RF) is an ensemble learning method proposed by Breiman *et al.* in 2001 [38]. It is implemented based on the idea of Bagging. The Bootstrap sampling technique is used to extract several different training subsets from the original training set. From each training subset, we train a decision tree. Finally, these decision trees form a random forest. For a binary task, the ultimate prediction is determined by the votes of all the trees.

Suppose that there is a training set X with M samples, and each sample consists of N input features and a classification

TABLE 1. Details of WDBC database.

Attribute number	Attributes	Attributes range		
		Mean	Standard Error	Maximum
1、11、21	Radius	6.98-28.11	0.112-2.873	7.93-36.04
2、12、22	Texture	9.71-39.28	0.36-4.89	12.02-49.54
3、13、23	Perimeter	43.79-188.50	0.76-21.98	50.41-251.20
4、14、24	Area	143.50-2501.00	6.80-542.20	185.20-4254.00
5、15、25	Smoothness	0.053-0.163	0.002-0.031	0.071-0.223
6、16、26	Compactness	0.019-0.345	0.002-0.135	0.027-1.058
7、17、27	Concavity	0.000-0.427	0.000-0.396	0.000-1.252
8、18、28	Concave points	0.000-0.201	0.000-0.053	0.000-0.291
9、19、29	Symmetry	0.106-0.304	0.008-0.079	0.157-0.664
10、20、30	Fractal dimension	0.050-0.097	0.001-0.030	0.055-0.208
Class distribution		Benign: 357(62.7%); Malignant: 212(37.3%)		
Number of instances		569		
PS: Ten real-valued features are computed for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features in the WDBC database.				

TABLE 2. Details of WBC database.

Attribute number	Attributes	Attributes range
1	Clump Thickness	1-10
2	Uniformity of Cell Size	1-10
3	Uniformity of Cell Shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial Cell Size	1-10
6	Bare Nuclei	1-10
7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10
Class distribution		Benign: 458(65.5%); Malignant: 241(34.5%)
Number of instances		699
PS: 16 instances that contain a single missing attribute value have been removed from the WBC database.		

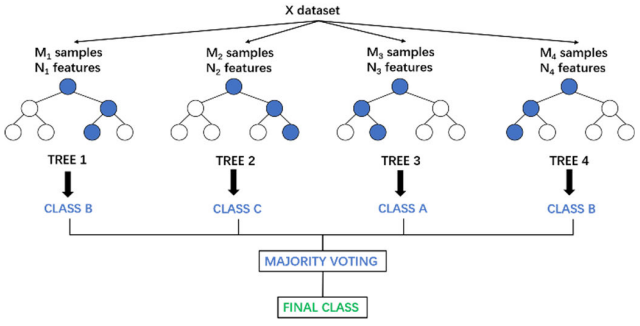


FIGURE 1. Random forest model.

label Y. The specific construction process of RF is as follows (FIGURE 1):

Step1: Select M samples from the training set X by applying Bootstrap technique.

Step2: Select n features ($n < N$) randomly and the feature with minimum Gini value is selected to split the node of the decision tree.

Step3: Repeat step1 and step 2 K times and obtain K decision trees.

Step4: Combine the decision trees into a random forest and determine the classification result by voting.

C. VARIABLE IMPORTANCE MEASURE (VIM) METHOD

Selecting the most discriminating features plays a crucial role in early cancer detection because it provides clinical information about potential biomarkers. Therefore, we need to find an optimal feature subset to improve the accuracy of classification [39], [40].

VIM method is used to calculate the importance of all the features and rank them according to their importance [10].

By deleting the less important features, the significant tumor features of breast cancer are extracted to form the optimal feature subset, which enhances the performance of the classifier.

Assume that the training set has N features $\{N_1, N_2, \ldots, N_N\}$, we use the ''Gini Index'' to select the optimal partitioning feature at each node when constructing decision trees in the random forest. Gini Index reflects the probability of category inconsistency of two samples randomly selected from the subset after node division. The smaller Gini Index is, the higher the purity of subset is. That means the partitioning features we choose are more conducive to classification. The calculation formula of ''Gini Index'' is

$$G_m = \sum_{c=1}^{C} p_{mc}(1 - p_{mc}) \tag{1}$$

where C is the number of categories on the training set and p_{mc} is the probability of a classification c at node m. In a binary task, $C = 2$, the ''Gini Index'' of node m is

$$G_m = 2p_m(1 - p_m) \tag{2}$$

where p_m is the probability of a classification 0 or 1.

The feature importance of N_i at node m is calculated by

$$I_{jm} = G_m - w_L G_L - w_R G_R \tag{3}$$

where G_L and G_R represent the ''Gini Index'' of the left and right nodes after node m is split, respectively. w_L and w_R are the number of weighted samples reaching the left and right nodes after node m is split, respectively.

If feature N_i is selected M times in the decision tree T_i, then the feature importance of N_i on decision tree T_i is calculated by

$$I_{ij} = \sum_{m=1}^{M} I_{jm} \tag{4}$$

Finally, the feature importance N_i in the random forest is defined as

$$I_j = \sum_{i=1}^{K} I_{ij} / \sum_{j=1}^{N} \sum_{i=1}^{K} I_{ij} \tag{5}$$

where K is the number of decision trees in the random forest and N is the number of input features on the training set.

D. HIERARCHICAL CLUSTERING RANDOM FOREST CLASSIFIER

The generalization ability of random forest is positively correlated with the classification ability of an individual decision tree and the diversity among decision trees [38]. Therefore, we try to improve the random forest model from two aspects: one is to improve the classification accuracy of a single decision tree; the other is to reduce the correlation between decision trees.

According to the random forest algorithm, we generate an initial random forest with K decision trees. We use the hierarchical clustering method to divide the decision trees into several clusters. We first regard each decision tree as an initial cluster. The two clusters with the highest similarity are found and they are combined into one cluster. Then the similarity between the new clusters is recalculated. The process of clustering is iterated multiple times and the decision trees in the initial random forest are clustered into several clusters. Finally, we select the decision tree with the highest AUC (Area Under ROC Curve) from each cluster as the representative of this cluster, and eliminate other decision trees from the clusters to obtain the HCRF model.

The method used in this paper to calculate the similarity between decision trees is disagreement measure (DIS), since it is a simple and effective method to measure diversity in decision forests [15]. DIS represents the ratio between the number of observations on which one classifier is correct and the other is incorrect to the total number of observations, which can be calculated with formula (6) using the variables defined in TABLE 3.

Suppose there are two decision trees T_i and T_j, their classification result of the samples on the training set is listed in TABLE 3.

TABLE 3. The relationship between decision trees Ti and Tj.

	T_j correct(1)	T_j incorrect(0)
T_i correct(1)	x_{11}	x_{10}
T_i incorrect(0)	x_{01}	x_{00}

x_{11} means quantity of training samples that is correctly classified by both T_i and T_j and x_{00} means quantity of training samples that is incorrectly classified by both T_i and T_j. x_{10} means quantity of training samples that is only rightly classified by T_i while x_{01} means quantity of training samples that is only rightly classified by T_j.

Then, the similarity calculation equation between decision trees T_i and T_j is as follows.

$$D_{i,j} = \frac{x_{01} + x_{10}}{x_{01} + x_{10} + x_{00} + x_{11}} \tag{6}$$

The smaller value of $D_{i,j}$ is, the greater the similarity is.

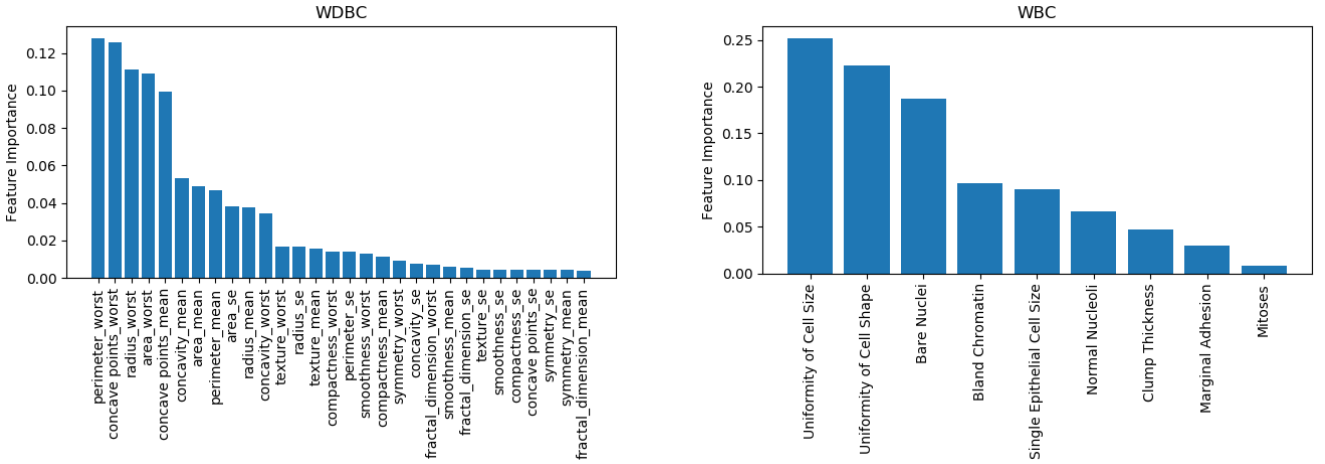
According to the above equation, we calculate the similarities between decision trees of the initial random forest and obtain a similarity matrix Sim which is defined as

$$Sim = \begin{bmatrix} D_{1,1} & \cdots & D_{1,K} \\ \vdots & \ddots & \vdots \\ D_{K,1} & \cdots & D_{K,K} \end{bmatrix} \tag{7}$$

The construction process of HCRF is summarized as follows.

E. EVALUATION METRICS

For a binary classification, the model divides the samples into two categories: Positive and Negative. If the prediction and the fact are both True, it is called True Positive (TP). If the prediction is False but the fact is True, it is called False



(a) feature importance of WDBC data set

(b) feature importance of WBC data set

FIGURE 2. The distribution of feature importance.

Algorithm 1 HCRF Algorithm

Input: Initial random forest $\{T_i, i = 1, 2, \ldots, K\}$
Similarity measure function D
Number of decision trees in the HCRF Q
Process:
1: **for** $i = 1, 2, \ldots, K$ **do**
2: $C_i = T_i$ (Regard each decision tree as a cluster)
3: **end for**
4: **for** $i = 1, 2, \ldots, K$ **do**
5: **for** $j = 1, 2, \ldots, K$ **do**
6: $Sim(i, j) = D(C_i, C_j)$
7: **end for**
8: **end for**
9: **repeat**
10: Find the two clusters with the highest similarity C_i and C_j in the matrix Sim
11: Merge C_i and C_j: $C_i = C_i \cup C_j$
12: Update the matrix Sim
13: $K = K$-1
14: **until** $K = Q$
15: **for** $i = 1, 2, \ldots, Q$ **do**
14: Select the decision tree T_i with the highest AUC as the representative of cluster C_i and delete other decision trees
15: **end for**
Output: HCRF $\{T_i, i = 1, 2, \ldots, Q\}$

Negative (FN). If the prediction is True, but the fact is False, it is called False Positive (FP). If the prediction and the fact are both False, it is called True Negative (TN). According to the above four cases, a confusion matrix can be obtained (TABLE 4).

The evaluation metrics used in this paper include: accuracy, precision, sensitivity, specificity and AUC.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$precision = \frac{TP}{TP + FP} \tag{9}$$

TABLE 4. Confusion matrix.

Actual class	Predicted class	
	positive	negative
positive	TP	FN
negative	FP	TN

$$sensitivity = \frac{TP}{TP + FN} \tag{10}$$

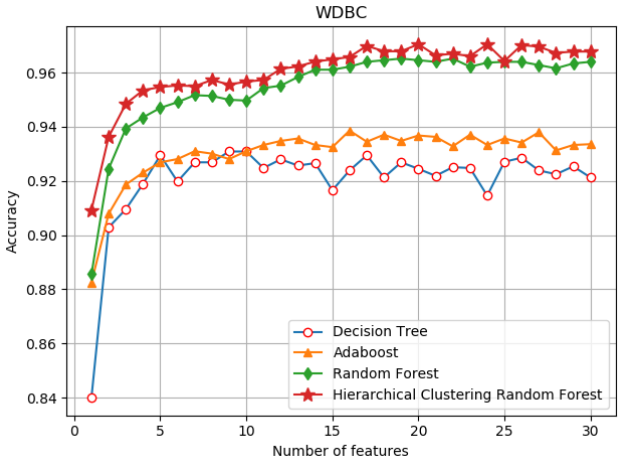
$$specificity = \frac{TN}{TN + FP} \tag{11}$$

The TP, FP, TN and FN measures can be collected to construct a plot, which is a Receiver Operating Characteristic (ROC) curve, to show the tradeoff of FN and FP rates to model classification errors. ROC curve is typically plotted using FP rate vs. TR rate. By calculating the area under the ROC curve, we can get the AUC value [41].

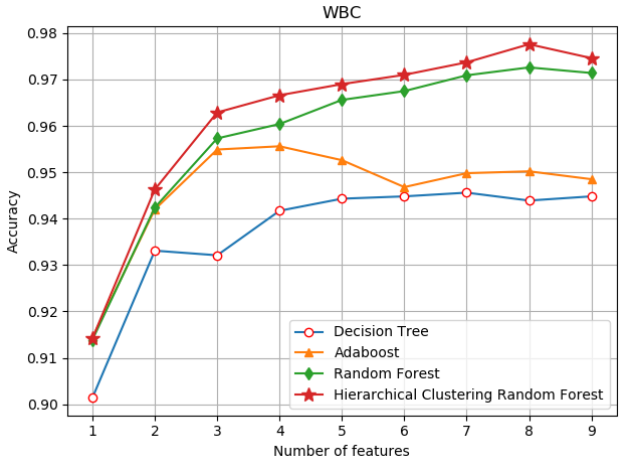
III. EXPERIMENTAL RESULTS AND DISCUSSION

In the experiment, we used 70% of the data as the training set and the remaining 30% of data for the test set. Feature selection was first carried out using VIM method. Different subsets of features of different sizes from 1 to N (a subset of features of size N means no feature selection) were produced according to the ranking of the feature importance, where N denotes the number of all features in the dataset. For each subset of features, grid search algorithm is used to optimize the parameters of the models. FIGURE 2 shows the distribution of all the features' importance of the WDBC and WBC datasets.

Then, DT, Adaboost, RF and our HCRF model was tested on the WDBC and WBC database and the accuracy when using all subsets of features size from 1 to N is shown in FIGURE 3. From the comparison of accuracy on the above four models, it is proved that HCRF has a higher accuracy

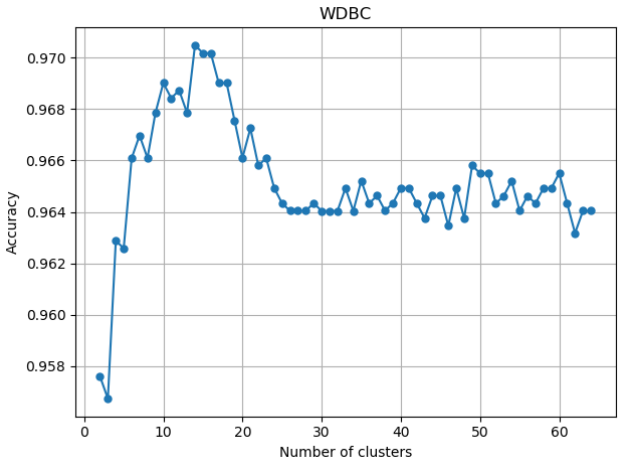


(a) accuracy of different model on WDBC data set

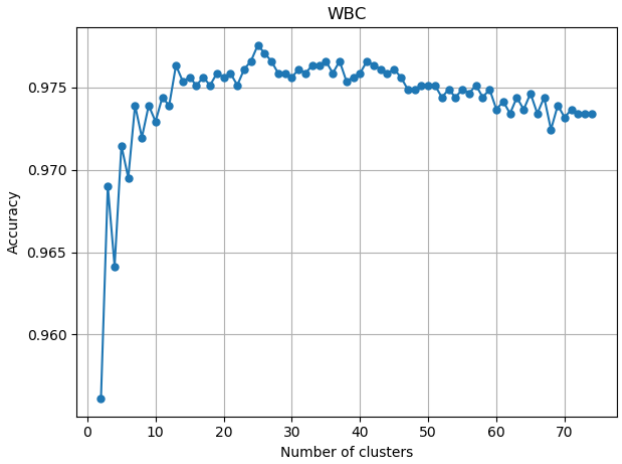


(b) accuracy of different model on WBC data set

FIGURE 3. Accuracy of DT, Adaboost, RF and HCRF based on different subsets of features.



(a) clustering iterative process of HCRF on WDBC data set



(b) clustering iterative process of HCRF on WBC data set

FIGURE 4. Accuracy of HCRF based on different number of clusters.

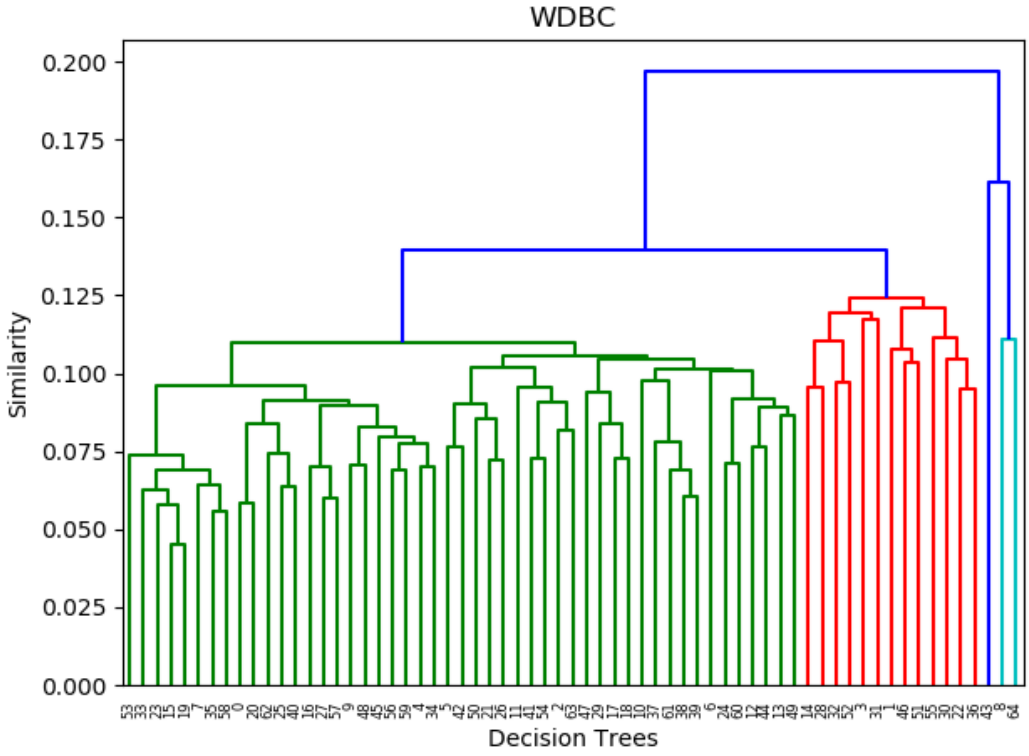
TABLE 5. Selected features after applying VIM method.

Datasets	The most important features
WDBC	1,2,3,4,5,6,7,8,11,12,13,14 17,20,21,22,23,24,25,26,27,28,29,30 (9,10,15,16,18,19 were removed)
WBC	1,2,3,4,5,6,7,8 (9 was removed)

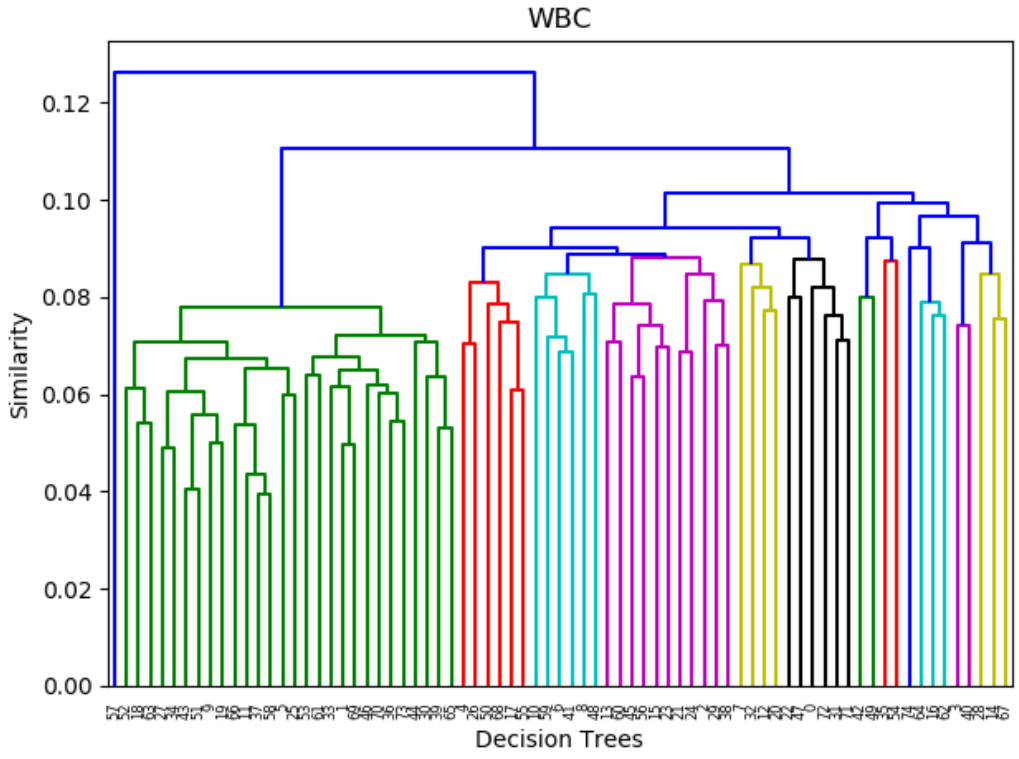
TABLE 6. The representative decision trees in each cluster.

Datasets	Representative decision trees in each cluster
WDBC	2, 3, 9, 15, 31, 32, 44, 45, 47, 49, 53, 55, 56, 65
WBC	1, 5, 8, 9, 11, 13, 23, 25, 29, 30, 33, 36 40, 41, 42, 49, 50, 55, 56, 57, 58, 63, 68, 73, 75

than other models whatever subsets of features we choose. The maximum accuracy was achieved by using the first 24 features on the WDBC dataset and the first 8 features on the WBC dataset. Thus, these features were regarded as the most discriminating features obtained using VIM technique, which are shown in TABLE 5. For WDBC database, Radius-Mean, Texture-Mean, Perimeter-Mean, Area-Mean, Smoothness-Mean, Compactness-Mean, Concavity-Mean, Concave points-Mean, Radius-SE, Texture-SE, Perimeter-SE, Area-SE, Concavity-SE, Fractal dimension-SE, Radius-Max, Texture-Max, Perimeter-Max, Area-Max, Smoothness-Max,



(a) tree-shaped clustering structure of random forest on WDBC data set



(b) tree-shaped clustering structure of random forest on WBC data set

FIGURE 5. **The tree-shaped clustering structure of the initial random forest.**

TABLE 7. The performance of different models on WDBC AND wbc datasets.

DATASET	WDBC						WBC					
Models	Acc (%)	Pre (%)	Sen (%)	Spe (%)	AUC (%)	Testing Time (s)	Acc (%)	Pre (%)	Sen (%)	Spe (%)	AUC (%)	Testing Time (s)
DT	91.46	87.67	90.00	92.33	91.16	78	94.39	93.32	90.55	96.46	93.51	46
Adaboost	93.33	91.15	91.10	94.67	92.88	214	95.02	92.82	93.13	96.05	94.59	179
RF	96.37	95.97	94.37	97.57	98.88	199	97.26	95.16	97.22	97.29	99.34	165
HCRF	97.05	97.32	94.77	98.41	98.96	148	97.76	95.65	98.12	97.56	99.39	134

Compactness-Max, Concavity-Max, Concave points-Max, Symmetry-Max, Fractal dimension-Max are selected and Radius, Perimeter, Area and Texture are considered as the most important features. On WBC database, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli are chosen as the most significant features. The shape and size (including radius, perimeter and area) of cells are closely related to the lesion of breast cancer. The obvious enlargement of epithelial cells may be indicative of malignancy. Rough texture is a sign of malignancy and texture uniformity is a sign of benign. The cells grouping in the breast is related to the Clump Thickness. Benign cells are usually monolayer while malignant cells are usually multilayered. Normal nucleoli is used to describe small structures present in the nucleus. Nucleoli are usually small, but begin to protrude in malignant cells. The bare nuclei mean a nuclei lacking cytoplasm. Cells that exhibit this phenomenon are likely to be malignant. We consider them as the optimal feature subsets which help doctors find the most essential features in each dataset [42].

We also optimize our HCRF model by selecting different number of clusters based on an initial random forest. As shown in FIGURE 4, the accuracy gradually rises when the number of clusters increases and reaches a peak when the number of clusters is 14(WDBC) and 25(WBC). Then, the accuracy drops with the number of clusters keep on increasing, indicating that either too much or too little clustering will affect the accuracy of the HCRF model. FIGURE 5 shows the process of clustering according to the similarity between the clusters. In each cluster, the decision tree with the highest AUC is picked out to construct the HCRF. TABLE 6 lists the representative decision trees in each cluster when HCRF achieves the highest accuracy.

The comparison results are given in TABLE 7. The proposed HCRF model outperforms other models in term of accuracy, precision, sensitivity, specificity and AUC. The highest accuracy achieved by HCRF using a subset of 24 features is 97.05% on the WDBC dataset. The number of decision trees is reduced from 65 in the initial random forest to 14 in the hierarchical clustering random forest. Compared with DT, Adaboost and RF, the accuracy increases 5.59%, 3.72% and 0.68% respectively. For the WBC dataset, the highest accuracy of 97.76% is also obtained by HCRF using a subset of 8 features. The number of decision trees is reduced from 75 in the initial random forest to 25 in the hierarchical clustering random forest. The accuracy increases 3.37%, 2.74% and 0.5% respectively compared with DT, Adaboost and RF.

TABLE 8. Comparison of accuracy with other models on WDBC.

Models(year)	Accuracy(%)
Weighted vote-based ensemble(2015)[43]	95.09
GA-classifier(2016)[44]	96.6
EM-PCA-CART-Fuzzy Rule-Based(2017)[45]	94.1
LMNN-SRA(2018)[46]	96.66
PCA+CNN(2019)[47]	96.4
Bayesian Network(2020)[48]	96.31
Fuzzy-ID3+FUZZTDBD(2021)[49]	94.53
HCRF (This work)	97.05

TABLE 9. Comparison of accuracy with other models on WBC.

Models(year)	Accuracy(%)
Weighted vote-based ensemble(2015)[43]	97.42
GA-classifier(2016)[44]	96.6
WAUCE(2017)[50]	97.10
DBN Re-RX(2018)[51]	96.83
SVM Linear Kernel(2019)[52]	96.72
ICA-Fuzzy-SR(2019)[53]	95.45
RBFNN(2020)[54]	97.0
APC(2020)[55]	97.31
Fuzzy-ID3+FUZZTDBD(2021)[49]	94.36
HCRF (This work)	97.76

We also make a comparison of the performance of the HCRF algorithm with others. The compared models also use feature selection methods to remove the redundant features. TABLE 8 and TABLE 9 shows the result of the our proposed HCRF model and some published studies on the WDBC and WBC datasets. From the table, we can obviously see that our proposed model which uses VIM as a feature selection method and HCRF as a classifier achieves the best performance.

IV. CONCLUSION

In conclusion, we have developed a model for breast cancer diagnosis which uses VIM for feature selection and HCRF

for classification. Both of these two processes not only enhance the performance and generalization ability of the classifier but also reduce the complexity and testing time of the model. In the end, our proposed method achieves 97.05% accuracy on the WDBC dataset and 97.76% accuracy on the WBC dataset. Compared to the traditional random forest, the proposed HCRF model increases accuracy by 0.68% and 0.5% on the WDBC and WBC datasets respectively. This is of vital importance in actual diagnosis scenario, which means more breast cancer can be detected in time and more lives could be saved.

Our proposed method has great reference value for designing structural diversity using other types of basic learners, such as neural networks and support vector machines or other ensemble learning algorithms. The proposed method could be also applied in the detecting cancers of other types and provide doctors with guidance for early diagnosis, which have many useful medical applications in clinical breast tumor diagnosis. For patients who are with breast cancer history, such a model can lead to a more rapid intervention with the most appropriate treatment.

As future work, we plan to visualize the decision trees and take structural diversity into consideration to further enhance the diversity among decision trees of random forest. Moreover, we will use the heuristic algorithms to optimize the relevant parameters to make our method more intelligent.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA Cancer J. Clinicians*, vol. 60, no. 1, pp. 277–300, 2015.
[2] L. Chang, L. S. Weiner, S. J. Hartman, S. Horvath, D. Jeste, P. S. Mischel, and D. M. Kado, "Breast cancer treatment and its effects on aging," *J. Geriatric Oncol.*, vol. 10, no. 2, pp. 346–355, Mar. 2019.
[3] H. Danish and S. Goyal, "Early diagnosis and treatment of cancer series: Breast cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 80, no. 3, pp. 956–957, 2011.
[4] D. L. Miglioretti, J. Lange, J. J. Van Den Broek, C. I. Lee, and R. A. Hubbard, "Radiation-induced breast cancer incidence and mortality from digital mammography screening: A modeling study," *Ann. Internal Med.*, vol. 164, no. 4, pp. 205–214, Jan. 2016.
[5] I. D. Naranjo, P. Gibbs, J. S. Reiner, R. Lo Gullo, C. Sooknanan, S. B. Thakur, M. S. Jochelson, V. Sevilimedu, E. A. Morris, P. A. T. Baltzer, T. H. Helbich, and K. Pinker, "Radiomics and machine learning with multiparametric breast MRI for improved diagnostic accuracy in breast cancer diagnosis," *Diagnostics*, vol. 11, no. 6, p. 919, May 2021.
[6] W. Tao, M. Lu, X. Zhou, S. Montemezzi, G. Bai, Y. Yue, X. Li, L. Zhao, C. Zhou, and G. Lu, "Machine learning based on multi-parametric MRI to predict risk of breast cancer," *Frontiers Oncol.*, vol. 11, p. 226, Feb. 2021.
[7] D. Maruti and G. Vandana, "Fine-needle aspiration cytology of colloid carcinoma breast in correlation with histopathology," *Apollo Med.*, vol. 12, no. 4, pp. 264–266, Dec. 2015.
[8] David and Edwards, "Data mining: Concepts, models, methods, and algorithms," *J. Proteome Res.*, vol. 2, no. 3, p. 334, 2003.
[9] M. Hosni, I. Abnane, A. Idri, J. M. C. de Gea, and J. L. F. Alemán, "Reviewing ensemble classification methods in breast cancer," *Comput. Methods Programs Biomed.*, vol. 177, pp. 89–112, Aug. 2019.
[10] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, Dec. 2020.
[11] V. R. E. Christo, H. K. Nehemiah, J. Brighty, and A. Kannan, "Feature selection and instance selection from clinical datasets using cooperative co-evolution and classification using random forest," *IETE J. Res.*, pp. 1–14, Jan. 2020.
[12] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre, and G. Isabelle, "Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier," *Comput. Med. Imag. Graph.*, vol. 60, pp. 42–49, Sep. 2017.
[13] J. J. van der Zande, M. Dauwan, E. Van Dellen, P. Scheltens, A. W. Lemstra, and C. J. Stam, "P2–131: Applying random forest machine learning to diagnose Alzheimer's disease and dementia with Lewy bodies: A combination of electroencephalography (EEG), clinical parameters and biomarkers," *Alzheimer's Dementia*, vol. 12, pp. P661–P662, Jul. 2016.
[14] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Appl. Soft Comput.*, vol. 86, Jan. 2020, Art. no. 105941.
[15] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003.
[16] A. Amaricai, "Design trade-offs in configurable FPGA architectures for K-means clustering," *Stud. Informat. Control*, vol. 26, no. 1, pp. 43–48, Mar. 2017.
[17] L. Xiangxiao, O. Honglin, and X. Lijuan, "Kernel-distance-based intuitionistic fuzzy c-means clustering algorithm and its application," *Pattern Recognit. Image Anal.*, vol. 29, no. 4, pp. 592–597, Oct. 2019.
[18] H. Jiawei and K. Micheline, "Data mining: Concepts and techniques," *Data Mining Concepts Models Methods Algorithms Second Ed.*, vol. 5, no. 4, pp. 1–18, 2006.
[19] M. Jasmine and G. Kesavaraj, "Implementation of K-means clustering algorithm in the crime data set," *Program. Device Circuits Syst.*, vol. 12, no. 1, pp. 13–18, 2020.
[20] L. Billard and J. Kim, "Hierarchical clustering for histogram data," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 9, no. 5, p. e1405, Sep. 2017.
[21] S. Lee, J. Jung, I. Park, K. Park, and D.-S. Kim, "A deep learning and similarity-based hierarchical clustering approach for pathological stage prediction of papillary renal cell carcinoma," *Comput. Struct. Biotechnol. J.*, vol. 18, no. 2, pp. 2639–2646, 2020.
[22] C. Malzer and M. Baum, "A hybrid approach to hierarchical density-based cluster selection," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2020, pp. 223–228.
[23] M. C. Thrun and A. Ultsch, "Using projection-based clustering to find distance- and density-based clusters in high-dimensional data," *J. Classification*, vol. 38, no. 2, pp. 280–312, Jul. 2021.
[24] Y.-H. Chiang, C.-M. Hsu, and A. Tsai, "Fast multi-resolution spatial clustering for 3D point cloud data," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 1678–1683.
[25] M. Chavent, R. Genuer, and J. Saracco, "Combining clustering of variables and feature selection using random forests," *Commun. Statist.-Simul. Comput.*, vol. 50, no. 2, pp. 426–445, 2019.
[26] N. N. Chen, X. T. Gong, Y. M. Wang, C. Y. Zhang, and Y. G. Fu, "Random clustering forest for extended belief rule-based system," *Soft Comput.*, vol. 25, pp. 4609–4619, Jan. 2021.
[27] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.
[28] A. H. Alsaeedi, A. L. Albukhnefis, D. Al-Shammary, and M. Al-Asfoor, "Extended particle swarm optimization (EPSO) for feature selection of high dimensional biomedical data," *arXiv:2008.03530*. Accessed: Aug. 1, 2020. [Online]. Available: https://arxiv.org/abs/2008.03530v1
[29] T. Luo, C. Hou, D. Yi, and J. Zhang, "Discriminative orthogonal elastic preserving projections for classification," *Neurocomputing*, vol. 179, pp. 54–68, Feb. 2016.
[30] C. Hou, Y. Jiao, F. Nie, T. Luo, and Z.-H. Zhou, "2D feature selection by sparse matrix regression," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4255–4268, Sep. 2017.
[31] T. Luo, C. Hou, F. Nie, H. Tao, and D. Yi, "Semi-supervised feature selection via insensitive sparse regression with application to video semantic recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1943–1956, Oct. 2018.
[32] T. Luo, C. Hou, F. Nie, and D. Yi, "Dimension reduction for non-Gaussian data by adaptive discriminative analysis," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 933–946, Mar. 2019.
[33] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 483–519, 2013.
[34] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Oper. Res.*, vol. 43, no. 4, pp. 570–577, Aug. 1995.

[35] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *Proc. SPIE*, vol. 1905, pp. 861–870, Jul. 1993.
[36] W. H. Wolberg, W. N. Street, D. M. Heisey, and O. L. Mangasarian, "Computer-derived nuclear features distinguish malignant from benign breast cytology," *Hum. Pathol.*, vol. 26, no. 7, p. 792, 1995.
[37] W. H. Wolberg, "Computerized breast cancer diagnosis and prognosis from fine-needle aspirates," *Arch. Surg.*, vol. 130, no. 5, p. 511, May 1995.
[38] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[39] H.-D. Zhu and Y. Zhong, "New feature selection algorithm based on multiple heuristics," *J. Comput. Appl.*, vol. 29, no. 3, pp. 849–851, May 2009.
[40] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.
[41] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.
[42] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104089.
[43] S. Bashir, U. Qamar, and F. H. Khan, "Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble," *Qual. Quantity*, vol. 49, no. 5, pp. 2061–2076, Sep. 2015.
[44] S. Aalaei, H. Shahraki, A. Rowhanimanesh, and S. Eslami, "Feature selection using genetic algorithm for breast cancer diagnosis: Experiment on three different datasets," *Iranian J. Basic Med. Sci.*, vol. 19, no. 5, pp. 476–482, May 2016.
[45] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telematics Informat.*, vol. 34, no. 4, pp. 133–144, 2017.
[46] A. Sanchez, C. Soguero-Ruiz, I. Mora-Jiménez, F. J. Rivas-Flores, D. J. Lehmann, and M. Rubio-Sánchez, "Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions," *Expert Syst. Appl.*, vol. 100, pp. 182–196, Jun. 2018.
[47] M. M. Hasan, M. R. Haque, and M. M. J. Kabir, "Breast cancer diagnosis models using PCA and different neural network architectures," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME2)*, Jul. 2019, pp. 1–4.
[48] N. Krishnakumar and T. Abdou, "Detection and diagnosis of breast cancer using a Bayesian approach," *Adv. Artif. Intell.*, vol. 12109, pp. 335–341, May 2020.
[49] N. F. Idris and M. A. Ismail, "Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: Automatic fuzzy database definition," *PeerJ Comput. Sci.*, vol. 7, p. e427, May 2021.
[50] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, Jun. 2018.
[51] Y. Hayashi, "Use of a deep belief network for small high-level abstraction data sets using artificial intelligence with rule extraction," *Neural Comput.*, vol. 30, no. 12, pp. 3309–3326, 2018.
[52] M. I. H. Showrov, M. T. Islam, M. D. Hossain, and M. S. Ahmed, "Performance comparison of three classifiers for the classification of breast cancer dataset," in *Proc. 4th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Dec. 2019, pp. 1–5.
[53] I. Masoudiasl, S. Vahdat, S. Hessam, S. Shamshirband, and H. Alinejad-Rokny, "Proposing an integrated method based on fuzzy tuning and ICA techniques to identify the most influencing features in breast cancer," *Iranian Red Crescent Med. J.*, vol. 21, no. 9, pp. 1–11, Sep. 2019.
[54] A. H. Osman and H. M. A. Aljahdali, "An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model," *IEEE Access*, vol. 8, pp. 39165–39174, 2020.
[55] R. Santiago-Montero, H. Sossa, D. A. Gutiérrez-Hernández, V. Zamudio, I. Hernández-Bautista, and S. Valadez-Godínez, "Novel mathematical model of breast cancer diagnostics using an associative pattern classification," *Diagnostics*, vol. 10, no. 3, p. 136, Mar. 2020.

ZEXIAN HUANG is currently pursuing the degree in robotics engineering with Guangzhou University. His research interests include machine learning and machine vision.

DAQI CHEN received the Ph.D. degree in control science and engineering from Zhejiang University, in 2018. He is currently a Lecturer with Guangzhou University. His research interests include biosensors and deep learning.