Received July 14, 2020, accepted August 2, 2020, date of publication August 5, 2020, date of current version September 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014362

A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data

NA LIU[1,2], XIAOMEI LI[1], ERSHI QI[1], MAN XU[3], LING LI[4], AND BO GAO[5]

[1]College of Management and Economics, Tianjin University, Tianjin 300072, China
[2]School of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China
[3]Business School, Nankai University, Tianjin 300071, China
[4]School of Political and Law, Shihezi University, Shihezi 832000, China
[5]School of Computer Science and Technology, Anhui University, Hefei 230601, China

Corresponding author: Xiaomei Li (lxm@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71571105 and Grant 71971123, in part by the National Social Science Foundation of China under Grant 17BMZ077, and in part by the High Level Talent Startup Project of Shihezi University under Grant RCSK2018C21.

ABSTRACT With the help of machine learning (ML) techniques, the possible errors made by the pathologists and physicians, such as those caused by inexperience, fatigue, stress and so on can be avoided, and the medical data can be examined in a shorter time and in a more detailed manner. However, while the conventional ML techniques, such as classification, achieved excellent performance in classification accuracy when applied in medical diagnoses, they have a fatal shortcoming of poor performance since the imbalanced dataset, especially for the detection of the minority category. To tackle the shortcomings of conventional classification approaches, this study proposes a novel ensemble learning paradigm for medical diagnosis with imbalanced data, which consists of three phases: data pre-processing, training base classifier and final ensemble. In the first data pre-processing phase, we introduce the extension of Synthetic Minority Oversampling Technique (SMOTE) by integrating it with cross-validated committees filter (CVCF) technique, which can not only synthesize the minority sample and thereby balance the input instances, but also filter the noisy examples so as to perform well in the process of classification. In the classification phase, we introduce ensemble support vector machine (ESVM) classification technique, which were constructed by multiple diversity structures of SVM classifiers and thus has the advantages of strong generalization performance and classification precision. Additionally, in the last phase of the final ensemble strategy, we introduce the weighted majority voting strategy and introduce simulated annealing genetic algorithm (SAGA) to optimize the weight vector and thereby enhance the overall classification performance. The efficiency of our proposed ensemble learning method was tested on nine imbalanced medical datasets and the experimental results clearly indicate that the proposed ensemble learning paradigm outperforms other state-of-the-art classification models. Promisingly, our proposed ensemble learning paradigm can effectively facilitate medical decision making for physicians.

INDEX TERMS Support vector machine, imbalanced data, ensemble learning, medical diagnosis.

I. INTRODUCTION

The World Health Organization (WHO) reports that, cancer has been listed the second leading cause of death and there estimated that about 9.6 million people die from cancer worldwide in 2018, mostly in developing countries [1]. Additionally, there are 422 million adults who have diabetes, and among them, diabetes is the cause of death in 1.5 million. Even worse, 17.5 million people die each year from cardiovascular diseases (CVDs). Furthermore, in China, it was estimated that 214,360 women died from breast cancer by 2008 and the number of deaths will reach up to 2.5 million by 2021 [2]. Due to such a serious situation, not only the patients but also their families suffer [3]. Thus, it is essential to identify the real reasons that cause ''such a large amount of deaths''. According to the WHO reports that many of the

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

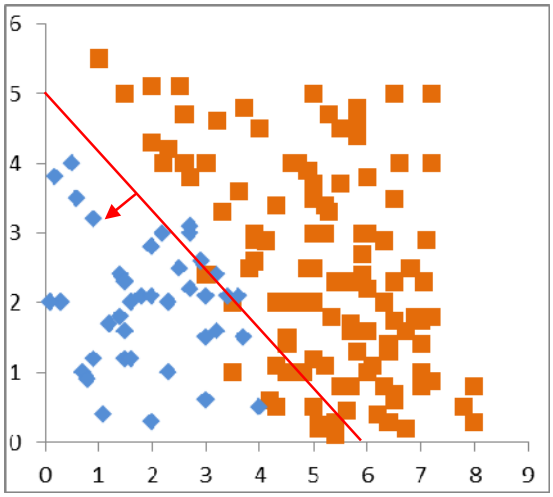

FIGURE 1. An example of unbalanced classification.

cancer cases are diagnosed too late, however, if they accepted accurate and early detection, more than 30% of these patients will be assured of long-term survival [4]. Consequently, it is of great significance for us to design an effective approach for early detection of diseases so as to improve the healthcare of our society.

Generally, due to machine learning (ML) techniques can effectively extract useful knowledge from large, complex, heterogeneous and hierarchical time series clinical data, these techniques have been widely utilized for medical diagnosis [5]–[8]. Moreover, with the help of ML techniques, the possible medical errors of the pathologists and physicians that caused by inexperience, fatigue, stress and so on can be avoided and the medical data can be examined in a short time and in a more detailed manner [9]–[11]. To the best of our knowledge, the problem of medical diagnosis has been attributed to classification issues, as previous studies demonstrated that various classification methods have been utilized for medical diagnosis, such as neural networks, Naïve Bayes, KNN, SVM, and most of these classification models achieved excellent performance. However, these state-of-the-art classification models only focused on classification accuracy, while neglecting the imbalanced characteristic of the input original data. More specifically, when the input data is imbalanced, the classifier will biased toward the majority class and the decision boundary (line) will biased toward minority class samples [12], [21], [56], [57], as can be observed from Figure 1. Therefore, the classification performance will dramatically deteriorate, and the poor performance of the classifier will become greatly troublesome, especially when applied in medical diagnosis. Motivated by the above deficiency, in this work, we only concentrate on research the binary classification problem and proposed a novel ensemble learning paradigm for medical diagnosis with imbalanced data, which consists of three phases. In the first phase, we introduce Synthetic Minority Oversampling Technique (SMOTE) integrated with cross-validated committees filter (CVCF) technique for resampling the examples. To the best of our knowledge, SMOTE has been proven to be superiority to under-sampling and can increase the number of instances in the minority class by creating new synthetic instances rather than relying on replication [13], [14]. However, SMOTE only focused on synthetic the minority instances, while neglecting the presence of class noise. Motivated by this drawback, we introduce CVCF noise filter technique to remove the noisy examples and thereby construct the integrated SMOTE-CVCF technique for data pre-processing. CVCF is a committee-based filter technique, which can obtain the excellent performance in noise filter [15], [16]. Then in the second phase, we introduced the ensemble learning technique for classification. It is worth noting that ensemble learning, as one of the state-of-the-art technologies in machine learning, can generate more accurate classification results than a single classifier because it reflects the benefits from both the performance of the different classifiers and the diversity of the errors [17]–[20]. Nevertheless, the most important concern is that when we apply ensemble learning techniques for classification, we should consider that there are typically two main challenges: the one is how to select diversity classifier members to form an ensemble, and the other is how to fuse the individual decisions of the base classifiers into a single decision result [21]. It is worth pointing out that SVM as the widely utilized classifier and has been proven to be one of the most effective approaches for addressing binary classification problems, and thus show superiority regarding low algorithmic complexity and high robustness [9]. Due to such advantages of SVMs, they have been widely utilized for classification. However, previous studies only focused on either tuning SVM classifier's parameters or performing feature selection [22], which may lead to overfitting and cannot produce the optimum results. Motivated by this deficiency, in this work, we apply multiple diversity structures of SVM classification models to construct the ensemble members. In the final phase, we introduce the weighted fusion strategy, which can not only overcome the shortcomings of majority voting but also can measure the importance of each ensemble member in the final classification. Herein, to find the optimal weight vector, we apply a hybrid algorithm of simulated annealing genetic algorithm (SAGA) for optimization and thereby find the optimum weight vector for the fusion strategy. To the best of our knowledge, no study has been performed to diagnose clinical diseases utilizing multiple diversity structures of SVM ensemble classifiers based on imbalanced datasets. To fill in this gap, we proposed a novel ensemble learning paradigm for medical diagnosis based on imbalanced data and envision that our proposed ensemble learning paradigm can be regarded as a useful clinical intelligent diagnosis tool for medical decision maker. The main contributions of this work can be summarized as follows:

- A novel multi-stage ensemble learning paradigm is proposed for medical diagnosis based on imbalanced data. To the best of our knowledge, this is the first comprehensive

ensemble learning technique employing multiple diversity structures of SVMs perform for classification. In addition, this is the first study, where SAGA algorithm has been employed to explore the optimal weight vector for weighted fusion strategy.

- We propose a novel data preprocessing strategy, which introduce SMOTE-CVCF integrated technique for data resampling. It can not only overcome the deficiency caused by SMOTE and thereby remove the noise examples effectively, but also can synthetic the minority instances and rebalance the input data set and thus perform well in the process of classification.

The rest of our paper is organized as follows. Section 2 presents related works on medical diagnosis. Section 3 introduces preliminaries of our study. Section 4 proposes the framework of our proposed method. Section 5 presents experimental analysis to validate the effectiveness of our proposed method. Finally, the conclusions of this research are summarized in Section 6.

II. RELATED WORK

In this section, we briefly review the related work about the imbalance learning and the classification approaches in medical diagnosis.

A. RELATED WORK ON THE IMBALANCED LEARNING

Imbalance learning techniques have been drawn a lot of attention from both the pattern recognition and the machine learning communities [68], [69]. Reference [70] proposed a novel ensemble method, which converts the imbalanced dataset into multiple balanced datasets firstly and then construct a number of classifiers on these multiple data with a specific classification algorithm. Reference [71] proposed a two-stage algorithm to deal with imbalanced data classification problems, in the first stage, the algorithm generates a set of IGs utilizing meta-heuristics approaches, which is a dynamic clustering using particle swarm optimization, genetic algorithm K-means and artificial bee colony K-means together, then in the second stage the classifier has been applied to classify the data. Reference [59] introduced a novel over-sampling technique. This approach which utilize the real-value negative selection method to generate artificial minority data, and then the generated minority data with actual minority data are combined with the original majority data as the input data which can be performed for classification. Reference [58] introduced a new self-adaptive cost weights-based SVM cost-sensitive ensemble for imbalanced data classification. The proposed method not only apply cost-sensitive SVMs as basic weak learner but also modify the standard boosting scheme to cost-sensitive ones, finally the extensive experimental results verify the efficacy of our proposed approach. In 2019, Reference [72] introduced a novel SMOTE based class-specific extreme learning machine approach, in their proposed approach they exploits the benefit of both the minority oversampling and the class-specific regularization. Reference [73] presented an effective oversampling method which combine k-means clustering and SMOTE together and the combination of the proposed method can avoid the generation of noise and effectively overcomes imbalanced between and within classes. In another study, Reference [74] use of heterogeneous ensembles for imbalance learning and the experimental results have shown that the heterogeneous ensembles provide significantly higher AUC and F 1 scores when compared to the ensembles utilizing a single classification method. In their proposed approach, they deal with the imbalanced problem from two aspects. The one is from the data level, they applied the related methods such as under-sampling or over-sampling, and the other is from the algorithm level, they adopted some improved algorithm such as adjusted the weight. However, they may not consider the noise problem then they perform for classification. Consequently, in order to deal with this issue, we applied the CVCF technique after rebalancing the data, which can remove the noise and thereby improve the performance of the final results.

B. RELATED WORK ON THE CLASSIFICATION APPROACHES IN MEDICAL DIAGNOSIS

1) MEDICAL DIAGNOSIS APPROACHES BASED ON A SINGLE CLASSIFIER

Due to the advantages of ML techniques, many basic data mining models, such as artificial neural networks (ANNs), decision tree analysis, support vector machines (SVMs), Naïve Bayes, KNN, have been utilized for medical diagnosis. To the best of our knowledge, neural networks have the advantage of capturing the correlations between attributes, so they have been widely utilized for medical diagnosis. Reference [23] developed a novel decision support algorithm to determine the most appreciate treatment method for rectal cancer survivals based on Analytic Hierarchy Process (AHP) and sequential decision trees. In the proposed method, the priorities of each sub-criteria were calculated by AHP method, and a sequential decision tree was constructed for the best treatment decision. In 2016, Reference [24] proposed a novel SVM parameter tuning scheme that uses the fruit fly optimization algorithm (FOA). In their proposed model, FOA algorithm can adjust the parameters of SVM effectively and thus output the optimization results. It can effectively reduce the computational time and improve the computational efficiency. In 2018, Reference [25] introduced a probabilistic neural network as a classification approach for the lung carcinomas. This method utilized a simple segmentation method and a probabilistic neural network to obtain the classification accuracy of lung carcinoma, which is also capable of detecting low-contrast nodules and lung cancers of minor or equal to 20 mm of diameter. In another study, Reference [26] introduced a novel intelligent classification model for breast cancer diagnosis. In their proposed model, they employed information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection and introduced the cost sensitive support

vector machine (CSSVM) learning algorithm perform for classification. Reference [27] presented a convolutional Neural Network Improvement for Breast Cancer Classification (CNNI-BCC). In their proposed algorithm it can classify the mammographic medical images into benign patient, malignant patient and healthy patient without prior information of the presence of a cancerous lesion. Reference [28] constructed a hybrid model which is based on artificial neural network and fuzzy logical for cardiac arrhythmia classification. The proposed hybrid model consists of two basic module units, each basic module unit includes three different classifiers based on the fuzzy KNN algorithm, multilayer perceptron with gradient descent and momentum (MLP-GDM), and multilayer perceptron with scaled conjugate gradient back propagation (MLP-SCG) model, and then the output of the classifiers can be combined utilizing a fuzzy system for integrated of the results. In another study, Reference [29] presented a novel heartbeat classification technique based on deep convolutional neural network and batch-weighted loss function for heartbeat classification. In their proposed model, it can not only well performed the classification task without noise removal or feature extraction, but also can effective quantify the loss bring by imbalance data with the help of introducing batch-weighted loss function.

2) MEDICAL DIAGNOSIS APPROACHES BASED ON CLASSIFIER ENSEMBLE

To overcome the weakness of a single classifier, in recent years, studies have increasingly focused on constructing ensemble models for medical diagnosis and the empirical results demonstrate that ensemble models have performed better than single model. In 2009, Reference [30] introduced neural network ensemble classifier for heart disease diagnosis. The proposed method creates new models by combining the posterior probabilities or the predicted values form multiple predecessor models. Reference [12] proposed an integrated of sampling technique, which incorporated both under-sampling and over-sampling technique together for resampling the input dataset and then employed the ensemble of SVMs perform for classification. Reference [31] presented boost SVM approach to predict the post-operative expectancy in the lung cancer patients. The weight of SVM learning criterion were determined by the ensemble learning approach, which can minimize the error of external sequential in boosting loop. Reference [32] proposed a novel dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) method for breast cancer diagnosis. The proposed DCE-MRI method based on a mixture ensemble of convolutional neural network which is a modular and image based ensemble, and can stochastically parathion the high-dimensional image space through simultaneous and competitive learning of its modules. Reference [33] proposed a novel nested ensemble learning technique for automated diagnosis of breast cancer, which utilized the combination of Stacking and Voting to detect the benign breast tumors from malignant cancers, as for each nested ensemble classifier which contain ''Classifiers'' and ''Meta-Classifiers'', and for each ''Meta-Classifiers'', which have two or more different classification algorithm. In another study, Wang *et al.* [11] designed a WAUCE (Weighted Area Under the Receiver Operating Characteristic Curve Ensemble) ensemble learning model for breast cancer diagnosis based on twelve different structures of SVM classifiers, and finally obtained the final results adopt the weighted area under the receiver operating characteristic. Reference [10] presented a novel technique for predicting types of kidney stone, and the proposed method was an ensemble learning method that included different individual classifiers, During the process of ensemble learning, each classifier was assigned a weight calculated by genetic algorithm (GA) and finally introduced the weight majority voting to fusion the final results of each classifier. In 2019, Reference [34] proposed a two-staged model based on tree ensemble to predict the survival of colorectal cancer. In their proposed method, they adopted ensemble classification model for the first stage to predict if the patient is survival or not and then introduced another ensemble learning regression model for the second stage to predict the remaining lifespan of the patients whose predicted output is death in the first stage. This two-stage prediction can effectively predict the survive time of the patient precisely, which overcome the deficiency of the traditional prediction methods. In another study, Reference [35] also proposed a novel stacking-based ensemble learning model for prostate cancer detection. In their proposed model, they simultaneously construct the diagnostic model and extract the interpretable diagnostic rule. Then in order to maximize the classification accuracy and minimize the ensemble complexity of the proposed model, they constructed a multi-objective optimization function, and then adopted the non-dominated sorting genetic algorithm-II (NSGA-II) algorithm to find the Pareto optimal solution. In summary, numerous previous studies have demonstrated that the results of ensemble learning have achieved superior performance than single classifiers, which can leverage the strength of individual classifiers and thereby output the optimal results [36]–[38]. Due to its advantages, in this work, we introduce ensemble learning technique for medical diagnosis.

III. PRELIMINARIES OF OUR STUDY

This section presents some preliminaries of our study, which include extend SMOTE with CVCF integrated filter technique, different types of kernel-based SVMs and the SAGA algorithm.

A. SMOTE-CVCF INTEGRATED FILTER METHOD

As a simple and effective oversampling method, SMOTE demonstrates performance that is superior to random oversampling [3], [6], [39]. By means of SMOTE, the number of instances for the minority category is increased by creating new synthetic instances [40]. The basic assumption of SMOTE is that the synthetic data points are generated on the line connecting the minority samples to their k nearest

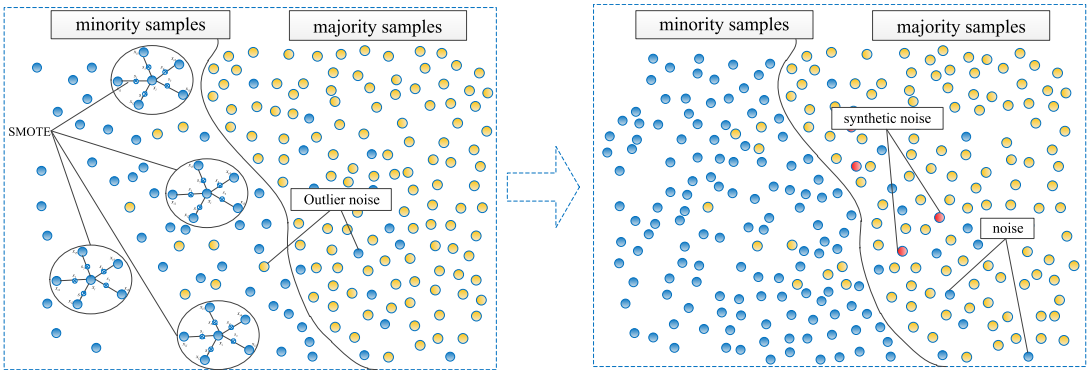

FIGURE 2. An example of SMOTE interpolation of random selected minority samples (k=5).

minority class neighbors, as shown in Figure 2, which presents an example of SMOTE interpolation of random selected minority samples when $k = 5$. As can be observed from Figure 2, SMOTE generates new minority samples around their original samples and thereby selects some of the new synthetic samples to balance the distribution of the dataset. The main steps of SMOTE are demonstrated below [39]:

Step 1: Calculate the distance between a feature vector in the minority category and one of its k nearest neighbors.

Step 2: Multiply the distance obtained in Step 1 by a random number between 0 and 1.

Step 3: Add the value obtained from Step 2 to the feature value of the original feature vector. Then, a novel feature vector is created by formula (1).

$$x_n = x_o + \delta * (x_{oi} - x_o) \tag{1}$$

where x_n denote the novel synthetic minority sample, and x_o denote the feature vector of each sample in the minority category. x_{oi} denote the i th selected nearest neighbor of x_o, and δ denotes a random number between 0 and 1. As shown in Figure 2, we notice that SMOTE can synthetic the minority instances and thereby rebalance the training dataset. However, when facing with noise examples, SMOTE can't perform well or even reinforce it. In order to overcome this deficiency, in this work, we introduce CVCF technique to filter the noisy examples first. As an effective noisy filter technique, CVCF was proposed by Verbaeten and Assche [15], which employed multiple classifiers using a single classification technique in a cross-validation strategy, if the misclassified examples by all cross-validation or most of the rounds, then it will be regarded as the noise and thereby be removed from the dataset [41]. The pseudo-code for the CVCF noisy filter technique is shown in the following.

Algorithm 1 CVCF Noisy Filter Technique

Input: Data set D; a parameter n
Output: Cleaned data set D'
1. Partitioning D into k subsets: $D_1, D_2, \ldots, D_k$
2. D' = D
3. for i = 1 to k do
4. $E_i = D \backslash D_i$
5. building a classifier with C4.5 algorithm H_i on E_i
6. for each instance X in D_i do
7. classify the instance X by the classifier Hi
8. if the instance X is misclassified by k classifiers (or most classifiers) then
9. D'←D'\X
10. end if
11. end for
12. end for
13. Return D'.

B. KERNEL-BASED SVMs

In this work, our ensemble learning technique was constructed by two different structures of SVM classifiers (i.e., $C - SVM$ and $v - SVM$) and five types of kernel functions (i.e., Linear kernel, Polynomial kernel, RBF kernel, Laplacian kernel, Sigmoid kernel). Additional details can be described as follows:

Given a training dataset $D = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \chi^n, y_i \in \gamma, i = 1, 2, 3, \ldots, \tau\}$, herein χ^n denote the n-dimensional feature space, and γ denote the category label, in this work we set $\gamma \in \{-1, +1\}$, then the dual form of the $C - SVM$ model is presented as below:

$$C - SVM : \begin{cases} \max\limits_{\alpha} -\dfrac{1}{2}\sum\limits_{i,j=1}^{\tau} \alpha_i \alpha_j y_i y_j \kappa(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum\limits_{i=1}^{\tau} \alpha_i \\ s.t.\ 0 \le \alpha_i \le C \quad \forall i \\ \sum\limits_{i=1}^{\tau} \alpha_i y_i = 0 \end{cases} \tag{2}$$

where in formula (2), α_i denotes the Lagrange multiplier, $\kappa(\mathbf{x}_i \cdot \mathbf{x}_j)$ denotes the kernel function, and C denotes the regularization term, which is used to balance the structural risk and empirical risk [11], [42], [43].

TABLE 1. Kernel function and default parameters.

No.	Kernel types	Kernel function	Default parameters
1	Linear kernel	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$	/
2	Polynomial kernel	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^d$	$d \geq 1$
3	RBF kernel	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$	$\sigma > 0$
4	Laplacian kernel	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma})$	$\sigma > 0$
5	Sigmoid kernel	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i^T \mathbf{x}_j + \theta)$	$\beta > 0, \theta < 0$

Additionally, in this work, we also introduce another *SVM* model, that is, the $v - SVM$ model.

$$v - SVM: \begin{cases} \max\limits_{\alpha} -\frac{1}{2} \sum\limits_{i,j=1}^{\tau} \alpha_i \alpha_j y_i y_j \kappa(\mathbf{x}_i \cdot \mathbf{x}_j) \\ s.t.\ 0 \leq \alpha_i \leq \frac{1}{\tau} \quad \forall i \\ \sum\limits_{i=1}^{\tau} \alpha_i y_i = 0 \\ \sum\limits_{i=1}^{\tau} \alpha_i \geq v \end{cases} \tag{3}$$

where in formula (3), α_i denotes the Lagrange multiplier, $\kappa(\mathbf{x}_i \cdot \mathbf{x}_j)$ denotes the kernel function, v represent the parameter to control the upper bound on the fraction of margin error. Herein $v \in [0, 1]$ is the parameter to control the upper bound on the fraction of the margin error and it also determines the lower bound of the fraction of the support vectors [44], [45].

From formula (2) and (3), we can clearly observe that the structures of these two SVM-based models vary greatly.

Additionally, in order to construct diversity classifiers to form an ensemble, we also introduce five types of different kernel functions, as demonstrated in Table 1.

C. SAGA HYBRID ALGORITHM

In this work, in order to find the optimal weight vector for the final ensemble, we introduce the SAGA hybrid algorithm. As an improved meta-heuristic approach, the hybrid SAGA algorithm can overcome the shortcomings of Simulated Annealing (SA) algorithms or Genetic Algorithm (GA), as well as the convergence to the global optimum solution. GA has the advantages of searching for the optimal solution quickly, but it has the fatal shortcoming of susceptibility to being trapped in the local optimum. Motivated by this deficiency, we introduce SA algorithm to improve GA, which can effectively change the annealing temperature during the iteration process and avoid becoming trapped in the local optimum [46]. Numerous empirical results have demonstrated that the hybrid algorithm of SAGA exhibits superior performance compared with those of the single Particle Swarm Optimization (PSO) or SA algorithms [13], [46]. Due to the advantages of the SAGA algorithm, in this work, we introduce this hybrid algorithm to optimize the weight vector, which can dynamically adjust the annealing temperature during the iteration process and avoid becoming trapped in the local optimum, thereby converging to global optimum solutions. The detailed steps of the SAGA algorithm can be described as follows:

Step 1: Set the initial parameters of SAGA: maxgen = 200; sizepop = 50; cross probability =0.7; mutation probability =0.05. Set the initial annealing temperature $T_0 = 100$; $T_{end} = 1$; $\xi = 0.8$.

Step 2: Set $T = T_0$; create the initial population, and calculate the fitness of each individual.

Step 3: Set the initial generation is 0.

Step 4: Select a chromosome with the largest fitness for replication, then perform cross and mutation.

Step 5: Then, generate a new population, and evaluate the fitness for each individual.

Step 6: Replace the least fitness population with the new best individual and then judge gen < maxgen? if "yes", then implement ***Step 4***; otherwise, execute the annealing operation.

Step 7: While $T_i < T_{end}$, if "yes", then output the optimal weight vector; otherwise, carry out the temperature operation of $T_{j+1} = \xi \times T_j$, then return to step 3. The overall flow chart of the SAGA algorithm is shown in Figure. 3.

IV. THE DESIGN OF THE PROPOSED METHOD

As noted before, an effective ensemble should consist not only of a set of models that are highly accurate but also the models that generate their errors on different parts of the input space as well. Thus, in our study, varying structures of SVMs have been utilized by each member of the ensemble to promote this necessary diversity. In general, our proposed novel ensemble learning paradigm can be structured in three consecutive stages, as it can be observed from Figure 5.

A. DATA PREPROCESSING

In the first stage, we introduce SMOTE-CVCF to integrate filter technique for resampling the input dataset. To the best of our knowledge, as an excellent oversampling technique, SMOTE can synthetic the minority instance so as to rebalance the dataset. However, SMOTE only focused on synthetic minority examples, while neglecting the noisy examples. Whereas when facing with noisy examples, SMOTE can't well handle it or even reinforce it, as it can be observed from Figure. 2. Motivated by this deficiency, we introduce CVCF noise filter technique to remove the noise examples. These two techniques must be applied in the correct order in order to obtain the ideal results. In this work, we employed SMOTE first and then CVCF noisy filter. This is due to that CVCF can filter the noisy example through the cross validation strategy, if we utilized CVCF before SMOTE, then it may carry the risk of removing all the minority examples which have been deemed as the noisy example in the cross validation. Through SMOTE-CVCF integrating filter technique, we can not only clean up the noisy examples both presented in the original data set and synthetic noise examples created by SMOTE, but

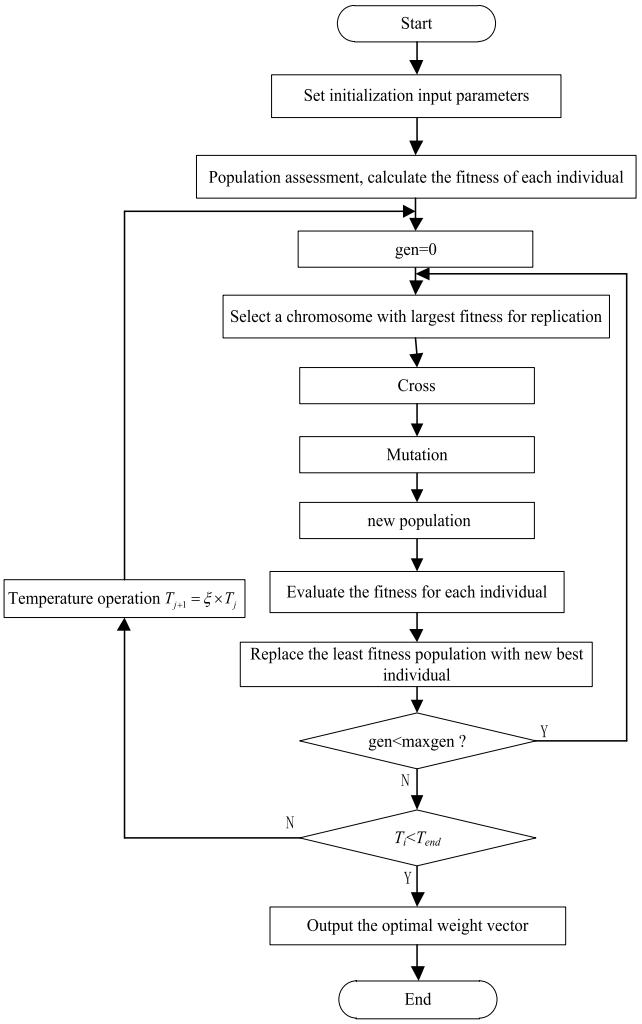

FIGURE 3. Overall flow chart of the SAGA algorithm.

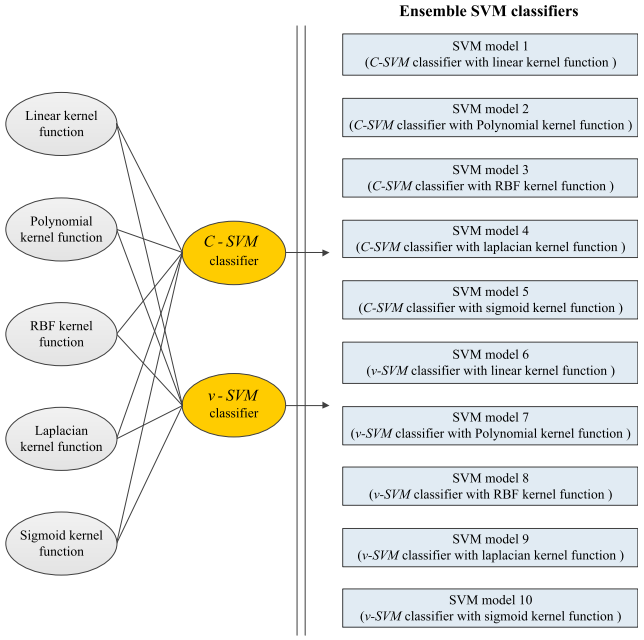

FIGURE 4. The details of our proposed SVM ensemble learning strategy.

also can rebalance the input dataset and thus perform well in the process of classification.

B. ENSEMBLE PRUNING FOR CLASSIFICATION

In this work, we design multiple diversity structures of SVM classifiers to form an ensemble. Due to the strong generalization performance and classification precision, SVMs have demonstrated their superior performances to other conventional classification methods [12], [47]. Based on the different structures of the *C-SVM* and *v-SVM* models, we fully consider five different types of kernel functions, and thereby form an ensemble classifier. As it can be obviously observed from Figure 4, our proposed SVM ensemble learning paradigm consists of two different structures of SVM classifiers and five different of kernel functions. In particular, the diversity of the ensemble member mostly relying on different options of kernel functions and the structure of the SVM classifiers. According to each diversity SVM classifier, we employed the grid search approach to obtain the penalty parameter C and kernel function parameter g of the SVM classifier. Furthermore, the 10-fold fold cross-validation approach with five replications are been utilized for the training and thereby obtain the final results of each classifier.

C. ENSEMBLE FUSION STRATEGY

It is well-known that conventional ensemble fusion strategies, such as majority voting, consider the decision from each classifier equally, and thereby neglect the influence of those classifiers with low accuracy [11]. Consequently, in order to overcome the deficiency of majority voting, in our work we introduced a weighted fusion strategy, which can not only overcome the deficiency by majority voting but also has the advantages of considering the contribution of each classifier. The formula of the weighted fusion strategy as follows [10], [11]:

$$\gamma(x) = F\left[\sum_{i=1}^{T} w_i \cdot h_i^j(\mathbf{x})\right] \tag{4}$$

$$H(x) = \begin{cases} -1 & \sum_{i=1}^{n} w_i h_i^j(\mathbf{x}) < 0 \\ +1 & otherwise \end{cases} \tag{5}$$

In formula (4), w_i represents the weight of each basic classifier, and $h_i^j(\mathbf{x})$ represents the decision results of i-th classifier corresponding to j-th pattern. $F[\cdot]$ denotes the ensemble fusion strategy. Formula (5) represents the weighted ensemble fusion strategy. $H(x)$ denotes the final result of the ensemble classifiers. Given the set of $(h_i^1(\mathbf{x}); h_i^2(\mathbf{x}); \ldots; h_i^N(\mathbf{x}))$, which denote the output results of each classifier, where N represents the numbers of classifiers, w_i is the corresponding weight of each classifier. Then according to formula (4) and formula (5), we can obtain the final results of ensemble learning.

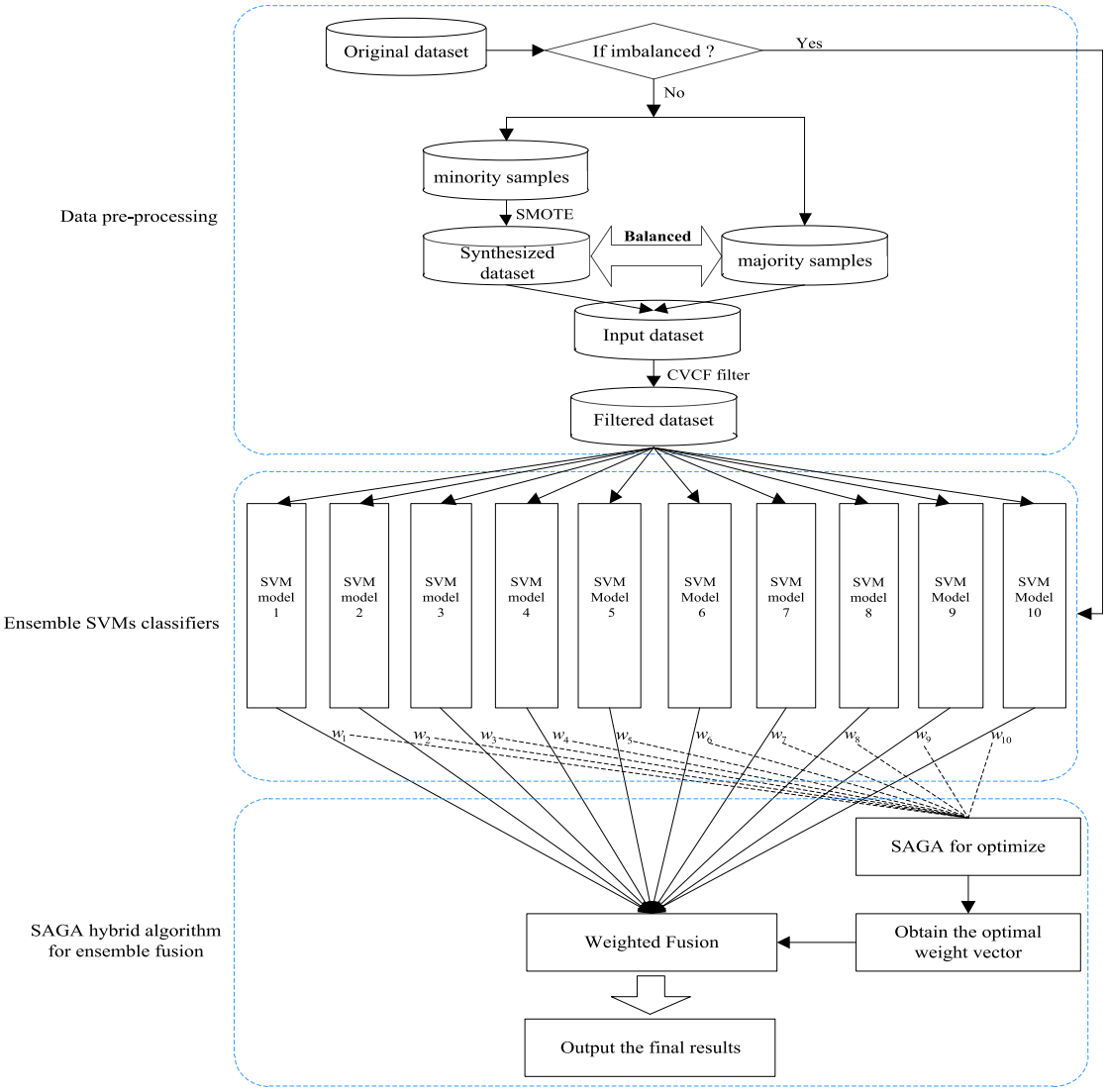

FIGURE 5. The framework of our proposed ensemble learning paradigm.

As noted before, one of the most critical issues in the weighted fusion strategy is determining the weight vector of each classifier [10]. However, how to determine the optimal weight vector of each classifier and thereby output the optimal results has become an imperative issue we should to solve. To the best of our knowledge, improve meta-heuristic algorithm of SAGA has the advantages of fast convergence to the global optimal solution [46]. In this regard, we introduce the SAGA algorithm to optimize the weight vector. In our work, the SAGA will find an optimal weight vector of $\mathbf{w}$, measuring the importance of each SVM classifier in the final classification. The final classification results can be obtained by a simple linear combination of the decision values of the SVMs with the weight vector. In this way, the representation of each individual of our population in the GA is defined as a vector containing the weights of each classifier members.

$$chromosome = [w_1, w_2, \ldots, w_n] \tag{6}$$

where n represents the number of SVMs, and in our study, we take full account of the different structures of the *C-SVM* and the *v-SVM* models and five different types of kernel functions. Then, we set $n = 10$. When we applied the SAGA algorithm to optimize the weight vector, the most important consideration is to determine the fitness function. For this task, the fitness function of our SAGA is the accuracy of combined classifiers using the given weights. We define the fitness function as following:

$$\begin{aligned} fitness &= \sum_{i=1}^{n} w_i \cdot h_i^j(x) \\ s.t. &\begin{cases} \sum_{i=1}^{n} & w_i = 1 \\ w_i \geq 0, & i = 1, 2, \ldots, n \end{cases} \end{aligned} \tag{7}$$

In formula (7), w_i is the weight vector, and $h_i^j(x)$ represent the output result of the i-th classifier corresponding to the j-th

TABLE 2. Details of the imbalanced medical datasets.

No	Dataset	Attributes	Instances (minority/majority)	Minority category	Majority category	IR
1	Wisconsin Breast Cancer(WBCD)	9	683(239/444)	"4"	"2"	1.86
2	Pima Indians Diabetes(Pima)	8	768(268/500)	"2"	"1"	1.87
3	Haberman	3	306(81/225)	"Positive"	"Negative"	2.77
4	Parkinson 2	754	756(192/564)	"0"	"1"	2.94
5	Parkinson 1	22	197(48/147)	"0"	"1"	3.06
6	Wisconsin Prognostic Breast Cancer (WPBC)	34	198(47/151)	"R"	"N"	3.21
7	Thyroid Disease	5	215(35/180)	"Positive"	"Negative"	5.14
8	SPECT heart data	22	267(55/212)	"0"	"1"	3.85
9	SPECTF heart data	44	267(55/212)	"0"	"1"	3.85

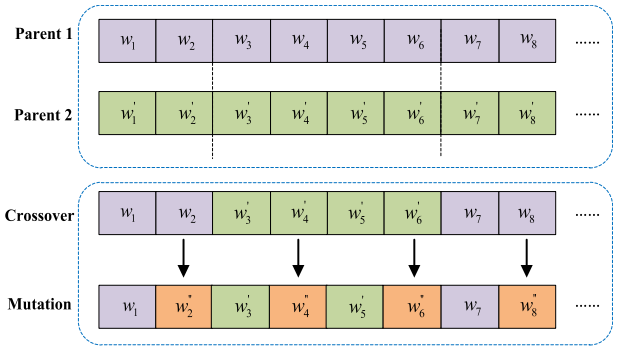

FIGURE 6. Crossover and mutation steps for the genetic algorithm.

pattern. In the GA, a population of 300 individuals is initially created using random weights for each individual. In each generation, the fitness function is performed on each instance of the population, and the population is sorted. From Figure. 6, we can clearly find that the crossover and mutation steps for the GA are shown, and each column represents a weight corresponding to each classifier.

V. EXPERIMENTAL ANALYSIS

In order to examine the efficacy and rationality of our proposed ensemble learning paradigm, we conduct an empirical analysis together with the experiment on the MATLAB 2016a platform. The performance parameters of the executing host are Windows 10 with an Intel (R) Core(TM) i5-8250U CPU at 1.80 GHz, X64, and 16 GB (RAM).

A. DATASETS

In order to evaluate the performance of the proposed ensemble learning approach in dealing with imbalanced medical datasets, we introduced nine imbalanced datasets with different imbalanced ratio (IR) in this work. The value of IR is computed as the ratio between the number of instances belonging to the minority category and the number of samples belonging to the majority category. Of these imbalanced online medical datasets, which come from UCI machine learning repository[1] and KEEL-dataset repository.[2] The detailed description of these datasets is presented in Table 2.

[1] https://archive.ics.uci.edu/ml/datasets.php [Last Accessed: Qct 23, 2019]

TABLE 3. Confusion matrix.

		Actual class	
		Class 0	Class 1
Predicted class	Class 0	True negative (TN)	False positive(FP)
	Class 1	False negative(FN)	True positive(TP)

B. EVALUATION MEASURES

To evaluate the performance of our proposed hybrid algorithm, the precision, recall, G-mean, F-measure and AUC have been utilized as the evaluation approaches for the imbalanced data. The AUC represents the area under the ROC curve, and it provides a single-number summary for the performance of a learning algorithm and is one of the best methods for comparing classifiers in two-class problems, especially in imbalanced learning [48]–[50]. The value of G-mean [51] represents the geometric mean of the true positive rate (TPR) and true negative rate (TNR), which has been widely utilized for evaluation measure for imbalanced data. The calculation formulas are presented as follows:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}} \tag{10}$$

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{11}$$

The evaluation measures listed above are based on the confusion matrix, which is shown in Table 3.

In Table 3, class 0 denotes the absence of disease and class 1 denotes the presence of disease. TP is the number of true positives, which represents cases that are correctly categorized in the 'positive' class; FN is the number of false

[2] http://keel.es/datasets.php [Last Accessed: Qct 23, 2019]

TABLE 4. The details information of the different classification models as for comparison.

No	Author/reference	Year	The details information for each comparison method.
1	V. N. Vapnik.[52]	1999	Single SVM without re-sampling the input dataset.
2	Akbani et al.[53]	2004	Single SVM classifier with SMOTE over-sampling.
4	Chen et al.[54]	2004	Random forest with balanced training data utilizing under-sampling pre-processing.
5	Liu et al.[12]	2011	Combine both under-sampling and over-sampling techniques together, and then introduce an ensemble SVMs perform for classification.
6	Bashir et al.[55]	2015	Ensemble learning approach which composed of five heterogeneous classifiers, and then adopt weighted voting technique for fusion.
7	Bashir et al.[60]	2015	The ensemble prediction approach utilizing an ensemble of five heterogeneous classifiers: Naïve Bayes, linear regression, quadratic discriminant analysis, instance based learner and support vector machines.
8	Bashir et al.[63]	2016	The proposed framework consists of three modules: the first is data acquisition and preprocessing, the second is ensemble classifier perform for training, which constructed by seven heterogeneous classifiers, and then the last module is prediction and evaluation.
9	Bashir et al. [62]	2016	The proposed "HM-BagMoov" model utilizing an ensemble of seven heterogeneous classifiers: Naïve Bayes, Linear Regression, Quadratic Discriminant Analysis, Instance Based Learner, Support Vector Machine, Artificial Neural Network and Random Forest, and then it adopted bagging with multi-objective optimized weighted vote technique for ensemble.
10	Bashir et al. [61]	2016	The proposed ensemble learning model consist of five heterogeneous classifiers: Naive Bayes, decision tree induction using information gain (DT-Info), SVM, decision tree induction using Gini index (DT-Gini), IBL, and finally introduced weighted voting for ensemble.
11	Kazemi et al.[10]	2018	Heterogeneous ensemble learning technique which composed of four data mining algorithms and then adopt weighted fusion strategy to obtain the final results
12	Wang et al.[35]	2019	Stacking-based three diversity decision tree members for ensemble.

negatives, which represents 'positive' class cases that are classified as 'negative'; TN is the number of true 'negatives', which represents cases that are correctly categorized in the 'negative' class; and FP is the number of false 'positives', which represents 'negative' class cases that are classified as 'positive'. In this work, we set positive cases as the majority category and set negative cases as the minority category. Additionally, in order to evaluate the performance of the ensemble classification models, we employed 10-fold cross-validation and the overall performances of the classifiers are calculated by averaging the performance of the 10 subsets.

C. EXPERIMENTAL DESIGN

To demonstrate the superior performance of our proposed ensemble learning paradigm, we design the experiment from two aspects. In the first aspect, we explore the effect of our proposed ensemble learning approach with other state-of-the-art classification models. To the best of our knowledge, these classification models have been considered state-of-the-art classifiers and proved excellent in previous studies. The details of these classification models can be found in Table 4. Then in the second aspect, in order to verify the robustness of our proposed approach, we comprehensively compared its previous obtained results on nine imbalanced medical datasets with other state-of-the-art classification models. In our comparative experiments, for all the SVM classifiers we employ RBF kernel function, and the penalty parameters C and kernel function parameters g are determined utilizing grid search method, The variation range of parameter C is 2^(-5) to 2^(15); the variation range of parameter g is 2^(-9) to 2^ (3); the step of average classification accuracy is 0.1; the step of parameter C is 0.2. For each comparative experiment, we repeat our experiment ten times and computed the average values.

D. EXPERIMNETAL RESULTS AND ANALYSIS

The results of the confusion matrix of our proposed ensemble learning paradigm for different imbalanced medical datasets are demonstrated in Table 5. Additionally, in this work, we also add a working example and explain the detail steps of our proposed method, as can be obviously observed in APPENDIX.

1) COMPARISON WITH OTHER STATE-OF-THE-ART CLASSIFICATION METHODS

In this subsection, we compare our proposed novel ensemble learning paradigm with other state-of-the-art classification models. As for each imbalanced medical dataset, we set a comparison experiment and the results are demonstrated in Table 6.

The ensemble model runs on each test instance individually, and each instance in test set can be classified into absence of the disease and presence of the disease. We employed 10-fold cross validation and the average performances of these results are calculated and analyzed to verify the superiority performances of our proposed approach. As can be seen in Table 6, in order to verify the effectiveness of our proposed approach, we set precision, recall, F-measure,

TABLE 5. The results of the confusion matrix for each imbalanced medical datasets.

No	Datasets		Confusion matrices		Results				
			Class 0	Class 1	Precision	Recall	F-measure	G-mean	AUC
1	WBCD	Class 0	170	6	0.9412	0.9796	0.9600	0.9535	0.9727
		Class 1	2	96					
2	Pima	Class 0	157	40	0.6460	0.7277	0.6518	0.7550	0.7561
		Class 1	38	73					
3	Haberman	Class 0	11	17	0.8431	0.8958	0.8687	0.8041	0.8438
		Class 1	10	85					
4	Parkinson2	Class 0	54	22	0.9116	1.0000	0.9538	0.8429	0.8552
		Class 1	0	227					
5	Parkinson1	Class 0	12	5	0.9519	0.9672	0.9440	0.8563	0.8965
		Class 1	2	59					
6	WPBC	Class 0	50	5	0.7222	0.7701	0.6647	0.8876	0.7545
		Class 1	12	13					
7	Thyroid	Class 0	14	0	1.0000	1.0000	1.0000	1.0000	1.0000
		Class 1	0	72					
8	SPECT heart data	Class 0	15	8	0.9024	0.8810	0.8916	0.7580	0.7945
		Class 1	10	74					
9	SPECTF heart data	Class 0	26	3	0.9117	1.0	0.9538	0.9469	0.9598
		Class 1	0	31					

G-mean and AUC as the evaluation measures. Table 6 indicates the results comparison of the proposed ensemble approach with some other state-of-the-art methods for different imbalanced medical data sets. As it can be obviously observed that the proposed ensemble learning paradigm can achieve higher precision, recall, F-measure, G-mean and AUC in all the imbalanced medical datasets in comparison with the other state-of-the-art classification models.

2) ROBUSTNESS ANALYSIS OF DIFFERENT CLASSIFICATION APROACHES

In order to further verify the superiority of our proposed ensemble learning paradigm, we also compared the robustness of our proposed ensemble learning method with the other state-of-the-art models.

According to ref [75] the relative performance of the algorithm n (n represent one of the eleven algorithms) on a certain benchmark dataset can be measured by the following ratio:

$$a_n = \frac{MPrecision_n}{\max\limits_n MPrecision_n} \tag{12}$$

$$b_n = \frac{MGMean_n}{\max\limits_n MGMean_n} \tag{13}$$

$$c_n = \frac{MFmeasure_n}{\max\limits_n MFmeasure_n} \tag{14}$$

$$d_n = \frac{MRecall_n}{\max\limits_n MRecall_n} \tag{15}$$

$$e_n = \frac{MAUC_n}{\max\limits_n MAUC_n} \tag{16}$$

where $MPrecision_n$, $MGMean_n$, $MFmeasure_n$, $MRecall_n$, $MAUC_n$ denote the mean of Precision, recall, F-measure, G-mean and AUC obtained by the algorithm n with different imbalanced ratios and classifiers on a certain dataset, respectively.

According to formula (12)∼(16), we can infer that the performance of the best performing algorithm $n*$ on a certain medical dataset would be equal to 1, while the relative performance of other algorithms would be less than 1. The larger a_n ,b_n, c_n, d_n, e_n values indicate that the algorithm n has the better relative performance. Based on the above analysis, we can conclude that the sum of a_n ,b_n, c_n, d_n, e_n is, then the better the robustness of the algorithm is [59]. We calculated the a_n,b_n, c_n, d_n, e_n values of the eleven algorithms in all medical datasets, and the results are shown in Figure 7∼Figure 11. In Figure 7∼ Figure 11, the numerical value labeled at the top of each histogram represent the sum of the $a_n(b_n, c_n, d_n, e_n)$ values of its corresponding algorithm in imbalanced medical datasets. From these figures we can clearly found that the our proposed ensemble learning paradigm always has the maximum sum values which are 7.847,7.396,7.787,7.968,7.833 respectively.

TABLE 6. Comparison of our proposed ensmeble approach with state of the art rechnique for different imbalanced medical datasets.

Dataset	Author/reference	Results				
		Precision	Recall	F-measure	G-mean	AUC
WBCD	V. N. Vapnik.[52]	0.9393	0.9117	0.9253	0.9271	0.9017
	Akbani et al.[53]	0.9410	0.9237	0.9367	0.9301	0.9161
	Chen et al.[54]	0.9285	0.9285	0.9454	0.9516	0.9235
	Liu et al.[12]	0.8532	0.9490	0.8986	0.8807	0.8529
	Bashir et al.[55]	0.8654	0.9184	0.8911	0.8925	0.8547
	Kazemi et al.[10]	0.8519	0.9388	0.8932	0.8800	0.8747
	Wang et al.[35]	0.8911	0.9184	0.9045	0.9140	0.8963
	Bashir et al.[63]	0.9406	0.9462	0.9472	0.9529	0.9525
	Bashir et al. [62]	0.8947	0.9444	0.9189	0.9525	0.9679
	Bashir et al. [61]	0.9407	0.9250	0.9367	0.9280	0.9642
	Bashir et al.[60]	0.9411	0.9411	0.9411	0.9420	0.9720
	Our proposed	**0.9412**	**0.9796**	**0.9600**	**0.9535**	**0.9727**
Pima	V. N. Vapnik.[52]	0.6142	0.4166	0.5263	0.6206	0.6076
	Akbani et al.[53]	0.6431	0.4789	0.5576	0.6987	0.6852
	Chen et al.[54]	0.5714	0.5157	0.5333	0.6714	0.6563
	Liu et al.[12]	0.6372	0.6486	0.6429	0.7103	0.6138
	Bashir et al.[55]	0.6148	0.6757	0.6438	0.6842	0.6518
	Kazemi et al.[10]	0.5840	0.6577	0.6186	0.6556	0.6456
	Wang et al.[35]	0.5615	0.6577	0.6058	0.6317	0.6669
	Bashir et al.[63]	0.5963	0.6105	0.5957	0.7033	0.7272
	Bashir et al. [62]	0.5916	0.7144	0.6434	0.7505	0.7219
	Bashir et al. [61]	0.6248	0.6646	0.6403	0.7246	0.7499
	Bashir et al.[60]	0.5993	0.6880	0.6346	0.7495	0.7247
	Our proposed	**0.6460**	**0.7277**	**0.6518**	**0.7550**	**0.7561**
Haberman	V. N. Vapnik.[52]	0.8421	0.7272	0.7804	0.6001	0.6217
	Akbani et al.[53]	0.8089	0.8911	0.8645	0.6026	0.6098
	Chen et al.[54]	0.8421	0.8	0.8205	0.6027	0.7536
	Liu et al.[12]	0.8218	0.8737	0.8469	0.5418	0.6177
	Bashir et al.[55]	0.8235	0.8842	0.8528	0.5423	0.7068
	Kazemi et al.[10]	0.8081	0.8421	0.8247	0.5096	0.6018
	Wang et al.[35]	0.8113	0.8953	0.8557	0.5200	0.6538
	Bashir et al.[63]	0.7842	0.8171	0.7771	0.5609	0.5476
	Bashir et al. [62]	0.7771	0.8145	0.7893	0.5876	0.6315
	Bashir et al. [61]	0.6177	0.8631	0.6988	0.5408	0.5751
	Bashir et al.[60]	0.8401	0.7962	0.8198	0.6007	0.6092
	Our proposed	**0.8431**	**0.8958**	**0.8687**	**0.8041**	**0.8438**
Parkinson2	V. N. Vapnik.[52]	0.9024	0.9780	0.9187	0.8073	0.8043
	Akbani et al.[53]	0.9107	0.9798	0.9226	0.8231	0.8056
	Chen et al.[54]	0.8805	0.8939	0.8872	0.4228	0.8217
	Liu et al.[12]	0.8968	0.9956	0.9436	0.7681	0.8345
	Bashir et al.[55]	0.8980	0.9692	0.9322	0.7763	0.8147
	Kazemi et al.[10]	0.8908	0.9339	0.9118	0.7655	0.7861
	Wang et al.[35]	0.8947	0.9736	0.9325	0.7672	0.7965
	Bashir et al.[63]	0.8095	0.6938	0.7472	0.6727	0.7689
	Bashir et al. [62]	0.8333	0.7291	0.7770	0.7186	0.8431
	Bashir et al. [61]	0.9102	0.7017	0.80	0.7687	0.8023
	Bashir et al.[60]	0.9112	0.8711	0.8932	0.7257	0.7529
	Our proposed	**0.9116**	**1.0000**	**0.9538**	**0.8429**	**0.8552**
Parkinson1	V. N. Vapnik.[52]	0.9136	0.8689	0.9138	0.7927	0.8718
	Akbani et al.[53]	0.9213	0.9376	0.94	0.8231	0.8189
	Chen et al.[54]	0.9205	0.7222	0.8125	0.6009	0.6574
	Liu et al.[12]	0.8814	0.8525	0.8667	0.7200	0.7365
	Bashir et al.[55]	0.8548	0.8689	0.8618	0.6343	0.7742
	Kazemi et al.[10]	0.8500	0.8361	0.8430	0.6325	0.7266
	Wang et al.[35]	0.9322	0.9016	0.9167	0.8443	0.8175
	Bashir et al.[63]	0.9339	0.9329	0.9167	0.2096	0.8101
	Bashir et al. [62]	0.9504	0.8146	0.9308	0.3271	0.8681
	Bashir et al. [61]	0.9117	0.9668	0.8791	0.3020	0.8861
	Bashir et al.[60]	0.9035	0.9601	0.8645	0.2955	0.8812
	Our proposed	**0.9519**	**0.9672**	**0.9440**	**0.8563**	**0.8965**
WPBC	V. N. Vapnik.[52]	0.6667	0.6400	0.6531	0.7480	0.7473
	Akbani et al.[53]	0.71	0.7016	0.6601	0.7989	0.7018
	Chen et al.[54]	0.5	0.6666	0.5714	0.7601	0.7171

TABLE 6. *(Continued.)* **Comparison of our proposed ensmeble approach with state of the art rechnique for different imbalanced medical datasets.**

	Liu et al.[12]	0.7143	0.4000	0.5128	0.8138	0.6770
	Bashir et al.[55]	0.7143	0.6000	0.6522	0.7977	0.6546
	Kazemi et al.[10]	0.6500	0.5200	0.5778	0.7532	0.6801
	Wang et al.[35]	0.6154	0.6400	0.6275	0.7096	0.6377
	Bashir et al.[63]	0.3258	0.5150	0.3834	0.7141	0.5752
	Bashir et al. [62]	0.4267	0.7611	0.4889	0.7986	0.6825
	Bashir et al. [61]	0.4858	0.5119	0.4036	0.6432	0.6781
	Bashir et al.[60]	0.4225	0.6833	0.5598	0.7716	0.6399
	Our proposed	**0.7222**	**0.7701**	**0.6647**	**0.8876**	**0.7545**
Thyroid	V. N. Vapnik.[52]	0.9863	1.0000	0.9931	0.9570	0.9884
	Akbani et al.[53]	0.9899	1.000	0.9934	0.9616	0.9764
	Chen et al.[54]	0.9333	1.0	0.9655	0.8165	0.8036
	Liu et al.[12]	0.9859	0.9722	0.9790	0.9568	0.9197
	Bashir et al.[55]	0.9726	0.9861	0.9793	0.9130	0.8839
	Kazemi et al.[10]	0.9722	0.9722	0.9722	0.9129	0.865
	Wang et al.[35]	1.0000	1.0000	1.0000	1.0000	1.0000
	Bashir et al.[63]	0.9444	1.00	0.9714	0.8944	0.9847
	Bashir et al. [62]	1.00	0.9941	0.9969	0.9970	0.9972
	Bashir et al. [61]	0.9473	1.00	0.9729	0.8660	0.9898
	Bashir et al.[60]	0.9440	1.00	0.9714	0.8944	0.9974
	Our proposed	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
SPECT heart data	V. N. Vapnik.[52]	0.8391	0.8902	0.8639	0.6076	0.6651
	Akbani et al.[53]	0.8901	0.90	0.8812	0.7013	0.7856
	Chen et al.[54]	0.8888	0.8	0.8421	0.7559	0.7876
	Liu et al.[12]	0.8481	0.8171	0.8323	0.6641	0.6438
	Bashir et al.[55]	0.8553	0.7927	0.8228	0.6921	0.6154
	Kazemi et al.[10]	0.8077	0.7683	0.7875	0.5684	0.5852
	Wang et al.[35]	0.8500	0.8293	0.8395	0.6648	0.7149
	Bashir et al.[63]	0.8164	0.8785	0.8382	0.5910	0.6711
	Bashir et al. [62]	0.9007	0.8751	0.8899	0.7015	0.7201
	Bashir et al. [61]	0.8340	0.8747	0.8484	0.6620	0.6945
	Bashir et al.[60]	0.9014	0.9030	0.8831	0.7358	0.7202
	Our proposed	**0.9024**	**0.9052**	**0.8916**	**0.7580**	**0.7945**
SPECTF heart data	V. N. Vapnik.[52]	0.7777	1.0	0.8750	0.8019	0.8216
	Akbani et al.[53]	0.7898	0.8817	0.9213	0.8824	0.8986
	Chen et al.[54]	0.7777	1.0	0.8750	0.8819	0.8991
	Liu et al.[12]	0.7857	0.7333	0.7586	0.8044	0.7383
	Bashir et al.[55]	0.9091	0.6667	0.7692	0.9250	0.7742
	Kazemi et al.[10]	0.7308	0.6333	0.6786	0.7618	0.6801
	Wang et al.[35]	0.7500	0.9000	0.8182	0.7426	0.7953
	Bashir et al.[63]	0.9050	0.9750	0.9514	0.9460	0.9588
	Bashir et al. [62]	0.8984	0.9019	0.9409	0.9432	0.9569
	Bashir et al. [61]	0.8340	0.8747	0.8484	0.6620	0.9563
	Bashir et al.[60]	0.9100	0.8916	0.8925	0.9346	0.9431
	Our proposed	**0.9117**	**1.0**	**0.9538**	**0.9469**	**0.9598**

Based on the above analysis, we can draw the conclusion that our proposed novel ensemble learning paradigm obtains the superior performances compared with other state-of-the-art classification models in terms of classification accuracy, G-mean and AUC (area under the receiver operating characteristic (ROC) curve). Moreover, it can also obtain the best robustness among the eleven algorithms as compared.

VI. DISCUSSION

The aim of this study is to propose a novel ensemble learning paradigm for medical diagnosis that performs at the same level or better than the other state-of-the-art comparison methods. In this section, we will provide in-depth discussion of the performance of our proposed ensemble learning paradigm.

In this regard, an extensive empirical analysis has been carried on nine imbalanced medical datasets. In order to evaluate the performances of our proposed ensemble learning paradigm, we compare five evaluation measures (i.e., precision, recall, G-mean, F-measure and AUC). In Table 6, we report the results comparison of the proposed ensemble approach with some other state-of-the-art methods for different imbalanced medical data sets. In general, we draw the conclusion that selection of diversity structures of SVM classifiers and thereby construct the ensemble learning paradigm outperforms other state-of-the-art classification models. Form the results we can infer that two types of SVM structures with five different kernel functions are adopted in our ensemble, which can not only increase the ensemble model structure diversities, but it can also increase the

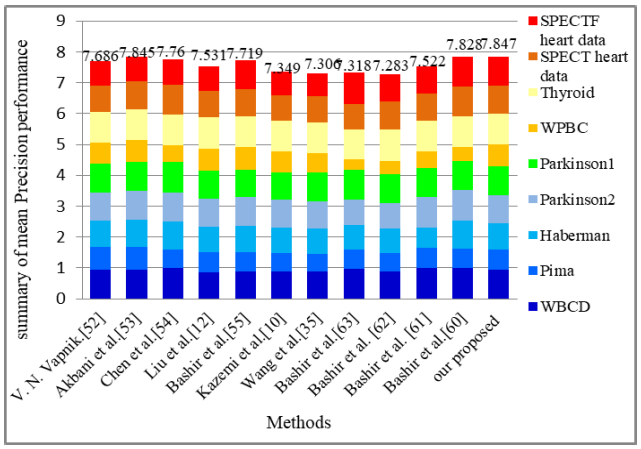

FIGURE 7. Comparison of robustness performance of different algorithms on nine imbalanced medical datasets in terms of Precision.

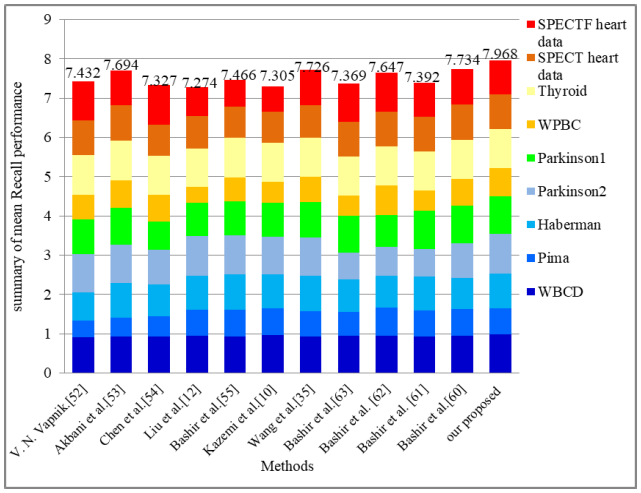

FIGURE 8. Comparison of robustness performance of different algorithms on 9 imbalanced medical datasets in terms of Recall.

parameter diversities. Moreover, we applied improved weighted ensemble mechanism which also considering the contribution of good basic classifier and thereby improve the final results. Finally the results of the experimental analysis indicate that our proposed method can achieve higher precision, recall, F-measure, G-mean and AUC in all the imbalanced medical datasets in comparison with the other state-of-the-art classification models. As it obviously indicated that the proposed ensemble classifier can improve the classification accuracy for medical diseases diagnosis, which also confirms our initial objective of this research.

Moreover, regarding the robustness of different algorithm, we calculated the a_n ,b_n, c_n, d_n, e_n values of the eleven algorithms in all medical datasets and the results are shown in Figure 7∼ Figure 11. As it can be observed from the results presented in Figure 7∼ Figure 11, our proposed ensemble learning paradigm can achieve the best robustness among the eleven algorithms as compared. From the results, we can infer that our proposed ensemble learning paradigm can effectively avoid the generation of noisy instances which often occurred

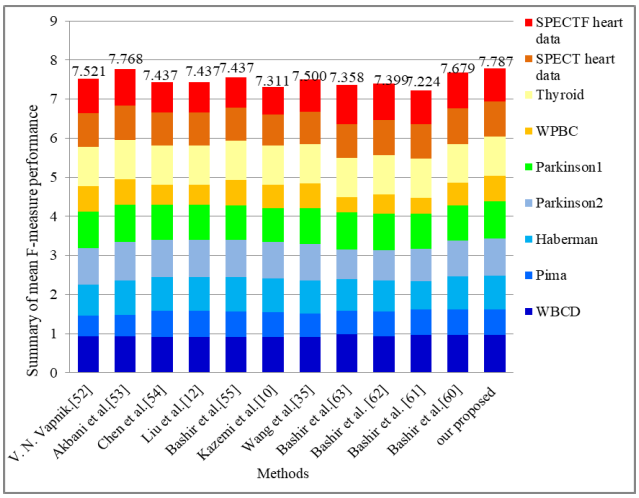

FIGURE 9. Comparison of robustness performance of different algorithms on nine imbalanced medical datasets in terms of F-measure.

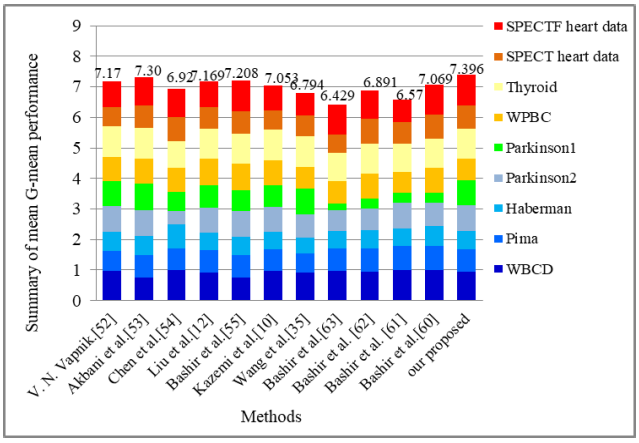

FIGURE 10. Comparison of robustness performance of different algorithms on nine imbalanced medical datasets in terms of G-mean.

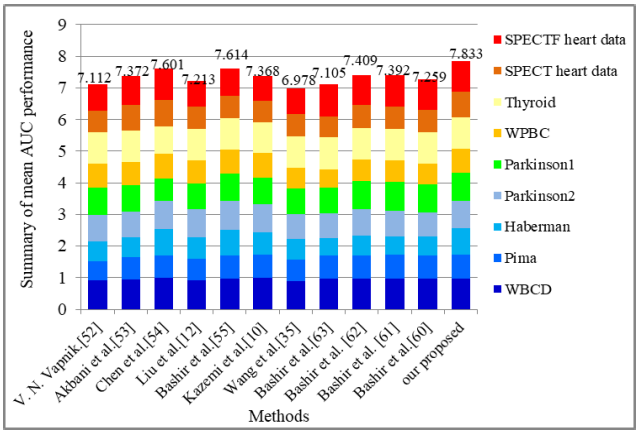

FIGURE 11. Comparison of robustness performance of different algorithms on nine imbalanced medical datasets in terms of AUC.

in other oversampling methods and can effective rebalancing the input data. From the empirical results, we can conclude that on one hand our proposed approach can effectively solve the problems encountered by other over-sampling methods

and can obtain the best robustness, on the other hand from the results presented in Table 6, our proposed approach can also achieve the best performances in terms of precision, recall, G-mean, F-measure and AUC. Furthermore, the analysis as described above has verified the effectiveness of our proposed ensemble learning paradigm.

VII. CONCLUSION

Medical diagnosis plays an important role in the healthcare system of our society, and the most important aspect is that the diagnostic results directly affect the patient's treatment and safety. To extract valuable knowledge for medical decision making can make our healthcare community better [64]–[67]. In this regard, we proposed a novel ensemble learning paradigm for medical diagnosis with imbalanced data, which consists of three phases: data pre-processing, training base classifiers and final ensemble. First, we introduce SMOTE-CVCF integrate filter technique for data pre-processing, which can not only filter the noisy examples, but also rebalance the input dataset and thus perform well in the process of classification. Then, in the next phase, the $C-SVM$ and the $v-SVM$ with five kernel functions are been utilized to increase the diversity of the ensemble model. In the last phase, we adopt the weighted fusion strategy. Then, in order to obtain the optimum weight vector, we introduce the SAGA algorithm to optimize the weight vector to improve the reasonableness of our ensemble fusion strategy. To evaluate the performance of our proposed method, nine benchmark medical imbalanced datasets are introduced, and the empirical results of these medical datasets demonstrate that our proposed ensemble learning paradigm can achieve the superior performances than other state-of-the-art classification models. To the best of our knowledge, this is the first study that employs multiple diversity structures of SVM classifiers to form an ensemble for medical diagnosis with imbalanced data. The main objective of this work was to apply our proposed ensemble learning paradigm in a clinical disease diagnostic system and thereby facilitate clinicians in making high-quality and effective decisions in the future.

APPENDIX

A WORKING EXAMPLE OF OUR PROPOSED ENSEMBLE CLASSIFIER

The working of the proposed ensemble learning paradigm can be clearly demonstrated by the following example:

(1) Suppose that the classifier training is performed for training data and the metrics of F-measure results are generated by each classifier:
SVM 1=94.40%; SVM 2=91.38%; SVM 3=94.40%; SVM 4=83.93%; SVM 5=87.50%; SVM 6=78.85%; SVM 7=89.47%; SVM 8=83.33%; SVM 9=78.85%; SVM 10=83.33%.

(2) According to the results of each classifier, we adopted SAGA algorithm to calculate the weight of each classifier and the resultant weights are as follows:
SVM 1=0.1008; SVM 2=0.1232; SVM 3=0.0967; SVM 4=0.0924; SVM 5=0.1058; SVM 6=0.0866; SVM 7=0.1162; SVM 8=0.0962; SVM 9=0.0855; SVM 10=0.0964.

(3) Suppose, the classifiers have predicted the following classes for a test instance:
SVM 1 = class 1; SVM 2 = class 1; SVM 3 = class 0; SVM 4 = class 0; SVM 5 = class 1; SVM 6 = class 0; SVM 7 = class 1; SVM 8 = class 0; SVM 9 = class 1; SVM 10 = class 0.

(4) The weighted vote-based ensemble classifier will generate the following prediction results. We simply add up the weight of each classifier that have voted for a particular class. For instance:
Class 0: SVM 3+SVM 4+SVM 6+SVM 8+SVM 10 ➔ *0.0967+0.0924+0.0866+0.0962+0.0964 0.4862* ➔ *Class 1: SVM 1+SVM 2+SVM 5+ SVM 7+SVM 9* ➔ *0.1008+0.1232+0.1058+0.1162+0.0855* ➔ *0.5315*

(5) Hence, according to weighted vote-based ensemble classifier, the class 1 has a higher value as compared with class 0. Therefore, the test instance will be classified as class 1.

CONFLICT OF INTEREST

The authors have no conflict of interest to mention.

REFERENCES

[1] World Health Organization (WHO). *Cancer*. Accessed: Oct. 23, 2019. [Online] Available: http://www.who.int/cancer/en/

[2] L. Fan, K. Strasser-Weippl, J.-J. Li, J. S. Louis, D. M. Finkelstein, K.-D. Yu, W.-Q. Chen, Z.-M. Shao, and P. E. Goss, "Breast cancer in China," The Lancet Oncology, vol. 15, no. 7, e279-e289, 2014.

[3] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," *Appl. Soft Comput.*, vol. 20, pp. 15–24, Jul. 2014.

[4] J. Jiang, X. Li, C. Zhao, Y. Guan, and Q. Yu, "Learning and inference in knowledge-based probabilistic model for medical diagnosis," *Knowl.-Based Syst.*, vol. 138, pp. 58–68, Dec. 2017.

[5] S. V. Kovalchuk, E. Krotov, P. A. Smirnov, D. A. Nasonov, and A. N. Yakovlev, "Distributed data-driven platform for urgent decision making in cardiological ambulance control," *Future Gener. Comput. Syst.*, vol. 79, pp. 144–154, Feb. 2018.

[6] S. Piri, D. Delen, and T. Liu, "A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets," *Decis. Support Syst.*, vol. 106, pp. 15–29, Feb. 2018.

[7] M. Eshtay, H. Faris, and N. Obeid, "Improving extreme learning machine by competitive swarm optimization and its application for medical diagnosis problems," *Expert Syst. Appl.*, vol. 104, pp. 134–152, Aug. 2018.

[8] R. Nagarajan and M. Upreti, "An ensemble predictive modeling framework for breast cancer classification," *Methods*, vol. 131, pp. 128–134, Dec. 2017.

[9] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 9014–9022, Jul. 2011.

[10] Y. Kazemi and S. A. Mirroshandel, "A novel method for predicting kidney stone type using ensemble learning," *Artif. Intell. Med.*, vol. 84, pp. 117–126, Jan. 2018.

[11] 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.

[12] Y. Liu, X. Yu, J. X. Huang, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Inf. Process. Manage.*, vol. 47, no. 4, pp. 617–631, Jul. 2011.

[13] H. Choi, H. Son, and C. Kim, "Predicting financial distress of contractors in the construction industry using ensemble learning," *Expert Syst. Appl.*, vol. 110, no. 15, pp. 1–10, Nov. 2018.
[14] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci.*, vol. 477, pp. 47–54, Mar. 2019.
[15] S. Verbaeten and A. Van Assche, "Ensemble methods for noise elimination in classification problems," in *Multiple Classifier Systems*. Berlin, Germany: Springer, 2003.
[16] L. P. F. Garcia, A. C. P. L. F. D. Carvalho, and A. C. Lorena, "Noise detection in the meta-learning level," *Neurocomputing*, vol. 176, pp. 14–25, Feb. 2016.
[17] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 814–833, Jul. 2017.
[18] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, May 2015.
[19] M. Papouskova and P. Hajek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," *Decis. Support Syst.*, vol. 118, pp. 33–45, Mar. 2019.
[20] X. Zhang and S. Mahadevan, "Ensemble machine learning models for aviation incident risk prediction," *Decis. Support Syst.*, vol. 116, pp. 48–63, Jan. 2019.
[21] B. Krawczyk, M. Woźniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Appl. Soft Comput.*, vol. 14, pp. 554–562, Jan. 2014.
[22] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, Mar. 2014.
[23] A. Suner, C. C. Çelikoğlu, O. Dicle, and S. Sökmen, "Sequential decision tree using the analytic hierarchy process for decision support in rectal cancer," *Artif. Intell. Med.*, vol. 56, no. 1, pp. 59–68, Sep. 2012.
[24] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, "Evolving support vector machines using fruit fly optimization for medical data classification," *Knowl.-Based Syst.*, vol. 96, pp. 61–75, Mar. 2016.
[25] M. Woźniak, D. Połap, G. Capizzi, G. L. Sciuto, L. Kośmider, and K. Frankiewicz, "Small lung nodules detection based on local variance analysis and probabilistic neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 173–180, Jul. 2018.
[26] N. Liu, E.-S. Qi, M. Xu, B. Gao, and G.-Q. Liu, "A novel intelligent classification model for breast cancer diagnosis," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 609–623, May 2019.
[27] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," *Expert Syst. Appl.*, vol. 120, pp. 103–115, Apr. 2019.
[28] E. Ramirez, P. Melin, and G. Prado-Arechiga, "Hybrid model based on neural networks, type-1 and type-2 fuzzy systems for 2-lead cardiac arrhythmia classification," *Expert Syst. Appl.*, vol. 126, pp. 295–307, Jul. 2019.
[29] A. Sellami and H. Hwang, "A robust deep convolutional neural network with batch-weighted loss for heartbeat classification," *Expert Syst. Appl.*, vol. 122, pp. 75–84, May 2019.
[30] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675–7680, May 2009.
[31] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Appl. Soft Comput.*, vol. 14, pp. 99–108, Jan. 2014.
[32] R. Rasti, M. Teshnehlab, and S. L. Phung, "Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," *Pattern Recognit.*, vol. 72, pp. 381–390, Dec. 2017.
[33] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognit. Lett.*, vol. 132, pp. 123–131, Apr. 2020.
[34] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, "A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction," *Inf. Sci.*, vol. 474, pp. 106–124, Feb. 2019.
[35] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin, and Y. Jin, "Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection," *Appl. Soft Comput.*, vol. 77, pp. 188–204, Apr. 2019.
[36] N. Rooney, H. Wang, and P. S. Taylor, "An investigation into the application of ensemble learning for entailment classification," *Inf. Process. Manage.*, vol. 50, no. 1, pp. 87–103, Jan. 2014.
[37] L. Yu, "An evolutionary programming based asymmetric weighted least squares support vector machine ensemble learning methodology for software repository mining," *Inf. Sci.*, vol. 191, pp. 31–46, May 2012.
[38] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Inf. Sci.*, vol. 425, pp. 76–91, Jan. 2018.
[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[40] G. C. Cawley, "Model selection for support vector machines via adaptive step-size Tabu search," in *Artificial Neural Nets and Genetic Algorithms*. Vienna, Austria: Springer, 2001, pp. 434–437.
[41] L. P. F. Garcia, J. Lehmann, A. C. P. L. F. de Carvalho, and A. C. Lorena, "New label noise injection methods for the evaluation of noise filters," *Knowl.-Based Syst.*, vol. 163, pp. 693–704, Jan. 2019.
[42] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, 2000.
[43] P. Chen, C. Lin, and B. Schölkopf, "A tutorial on ν-support vector machines," *Appl. Stochastic Models Bus. Ind.*, vol. 21, no. 2, pp. 111–136, 2005.
[44] L. Sørensen and M. Nielsen, "Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination," *J. Neurosci. Methods*, vol. 302, pp. 66–74, May 2018.
[45] E. Ardjmand, O. S. Bajgiran, and E. Youssef, "Using list-based simulated annealing and genetic algorithm for order batching and picker routing in put wall based picking systems," *Appl. Soft Comput.*, vol. 75, pp. 106–119, Feb. 2019.
[46] M. Shokouhifar and A. Jalali, "An evolutionary-based methodology for symbolic simplification of analog circuits using genetic algorithm and simulated annealing," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1189–1201, Feb. 2015.
[47] B. Al-Salemi, M. Ayob, G. Kendall, and S. A. M. Noah, "Multi-label Arabic text categorization: A benchmark and baseline comparison of multi-label learning algorithms," *Inf. Process. Manage.*, vol. 56, no. 1, pp. 212–227, Jan. 2019.
[48] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci.*, vol. 291, pp. 184–203, Jan. 2015.
[49] H. Guo, H. Liu, C. Wu, W. Zhi, Y. Xiao, and W. She, "Logistic discrimination based on G-mean and F-measure for imbalanced problem," *J. Intell. Fuzzy Syst.*, vol. 31, no. 3, pp. 1155–1166, Aug. 2016.
[50] H. Qiu, H.-Y. Yu, L.-Y. Wang, Q. Yao, S.-N. Wu, C. Yin, B. Fu, X.-J. Zhu, Y.-L. Zhang, Y. Xing, J. Deng, H. Yang, and S.-D. Lei, "Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy," *Sci. Rep.*, vol. 7, no. 1, p. 16417, Dec. 2017.
[51] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. Int. Conf. Mach. Learn.*, 1997, pp. 179–186.
[52] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
[53] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. ECML*, 2004, pp. 39–50.
[54] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Dept. Statist., Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. 666, 2004.
[55] 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.
[56] N. Liu, J. Shen, M. Xu, D. Gan, E.-S. Qi, and B. Gao, "Improved cost-sensitive support vector machine classifier for breast cancer diagnosis," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Nov. 2018.
[57] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis," *Comput. Ind. Eng.*, vol. 140, pp. 1–9, Feb. 2020.
[58] X. Tao, Q. Li, W. Guo, C. Ren, C. Li, R. Liu, and J. Zou, "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification," *Inf. Sci.*, vol. 487, pp. 31–56, Jun. 2019.

[59] X. Tao, Q. Li, C. Ren, W. Guo, C. Li, Q. He, R. Liu, and J. Zou, "Real-value negative selection over-sampling for imbalanced data set learning," *Expert Syst. Appl.*, vol. 129, pp. 118–134, Sep. 2019.
[60] S. Bashir, U. Qamar, and F. H. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting," *Australas. Phys. Eng. Sci. Med.*, vol. 38, no. 2, pp. 305–323, Jun. 2015.
[61] S. Bashir, U. Qamar, and F. H. Khan, "A multicriteria weighted vote-based classifier ensemble for heart disease prediction," *Comput. Intell.*, vol. 32, no. 4, pp. 615–645, Nov. 2016.
[62] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Informat.*, vol. 59, pp. 185–200, Feb. 2016.
[63] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "HMV: A medical decision support framework using multi-layer classifiers for disease prediction," *J. Comput. Sci.*, vol. 13, pp. 10–25, Mar. 2016.
[64] J. Su, Q. Bai, S. Sindakis, X. Zhang, and T. Yang, "Vulnerability of multinational corporation knowledge network facing resource loss," *Manage. Decis.*, Mar. 2020, doi: 10.1108/MD-02-2019-0227.
[65] S. Jiafu, Y. Yu, and Y. Tao, "Measuring knowledge diffusion efficiency in R&D networks," *Knowl. Manage. Res. Pract.*, vol. 16, no. 2, pp. 208–219, 2018.
[66] J. Su, Y. Yang, and X. Zhang, "Knowledge transfer efficiency measurement with application for open innovation networks," *Int. J. Technol. Manage.*, vol. 81, nos. 1–2, pp. 118–142, Jun. 2019.
[67] D. Gan, J. Shen, and M. Xu, "Adaptive learning emotion identification method of short texts for online medical knowledge sharing community," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–10, Jun. 2019.
[68] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103465.
[69] V. H. A. Ribeiro and G. Reynoso-Meza, "Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets," *Expert Syst. Appl.*, vol. 147, Jun. 2020, Art. no. 113232.
[70] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognit.*, vol. 48, no. 5, pp. 1623–1637, May 2015.
[71] R. J. Kuo, "Integrating cluster analysis with granular computing for imbalanced data classification problem—A case study on prostate cancer prognosis," *Comput. Ind. Eng.*, vol. 125, pp. 319–332, Nov. 2018.
[72] B. S. Raghuwanshi and S. Shukla, "SMOTE based class-specific extreme learning machine for imbalanced learning," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104814.
[73] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci.*, vol. 465, pp. 1–20, Oct. 2018.
[74] H. G. Zefrehi and H. Altınçay, "Imbalance learning using heterogeneous ensembles," *Expert Syst. Appl.*, vol. 142, Mar. 2020, Art. no. 113005.
[75] Y. Zhang, X. Li, L. Gao, L. Wang, and L. Wen, "Imbalanced data fault diagnosis of rotating machinery using synthetic oversampling and feature learning," *J. Manuf. Syst.*, vol. 48, pp. 34–50, Jul. 2018.

NA LIU received the master's degree from the Department of Industrial Engineering, Chongqing University, Chongqing, China, in 2012. She is currently pursuing the Ph.D. degree with the Department of Management Science and Engineering, College of Management and Economics, Tianjin University, Tianjin, China. She is also an Associate Professor with the Department of Industrial Engineering, School of Mechanical and Electrical Engineering, Shihezi University, Xinjiang, China. She has published more than 20 articles in international journals, including *Information Processing & Management*, *Mathematical Problems in Engineering*, and *Computers & Industrial Engineering*. She has presented her research in several national and international conferences. Her current research interests include intelligent decision, machine learning, imbalanced data learning, ensemble models, and application of intelligent decision methods in healthcare domain.

XIAOMEI LI received the master's and Ph.D. degrees from the Department of Management Science and Engineering, College of Management and Economics, Tianjin University, Tianjin, China, in 2004 and 2008, respectively. She is currently an Associate Professor with the Department of Management Science and Engineering, College of Management and Economics, Tianjin University. She has published more than 30 articles in international journals, including the *Journal of Cleaner Production* and the *Journal of General Management*. She has presented her research in several national and international conferences, including the IEEE, POMS, and INFORMS. Her current research interests include innovation management, knowledge management, and intelligent decision. She is a member of AOM and IACMR.

ERSHI QI received the Ph.D. degree from the College of Management Science, Tianjin University, Tianjin, China, in 1992. He is currently a Professor with the Department of Management Science and Engineering, College of Management and Economics, Tianjin University. He is also the Dean of the College of Management Science. He has published several books/textbooks in the broader are of production management. He is often invited to national and international conferences for keynote addresses on topics related to intelligent decision, production system optimization, healthcare analytics, and decision support systems. He has published more than 100 articles in international journals, including *Information Processing & Management*, *Mathematical Problems in Engineering*, EJOR, *Omega*, *Industrial Management and Data Systems*, *Kybernetes*, the *Journal of Applied Sciences*, *Information Technology Journal*, and *Sustainability* (Switzerland). He has presented his research in several national and international conferences, including CIE, POMS, and INFORMS annual meeting. His current research interests include intelligent decision, machine learning, imbalanced data learning, production system optimization, ensemble models, and application of intelligent decision methods in healthcare domain. He regularly serves and chairs tracks for various production management and analytics conferences, and serves on several academic journals as the editor-in-chief, a senior editor, an associate editor, and an editorial board member.

MAN XU received the Ph.D. degree from the Department of Management Science and Engineering, College of Management and Economics, Tianjin University, Tianjin, China, in 2011. She is currently an Associate Professor with the Department of Information Resource Management, Business School, Nankai University, Tianjin, China. She has published more than 30 papers in international conferences and journals, including CIE, POMS, INFORMS, *Information Sciences*, *Information Processing & Management*, *Mathematical Problems in Engineering*, the *Journal of Systems Engineering and Electronics*, the *Journal of Computational Information Systems*, and the *Journal of Information & Computational Science*. Her current research interests include data-driven intelligent decision making, machine learning, data mining, and application of intelligent decision methods in healthcare domain. Moreover, she is also an Editorial Board Member and a Reviewer of the international journal of *Computers & Industrial Engineering*. She was awarded the Excellent Doctoral Thesis, in 2013. In 2010, she received the Academic Newcomer Award of Ministry of Education, China.

LING LI received the master's degree from the School of Mechanical and Electrical Engineering, Shihezi University, Xinjiang, China, in 2007, and the Ph.D. degree from the Department of Economics and Management, Shihezi University, in 2018. She is currently a Professor with the School of Political Science and Law, Shihezi University. She has presided over a number of national, provincial, and ministerial projects and has published more than 20 articles. Her current research interests include knowledge management and intelligent decision.

BO GAO is currently pursuing the master's degree in intelligent decision and machine learning with the School of Computer Science and Technology, Anhui University, Hefei, China. As a Research Assistant, he has participated in research projects in a variety of fields, including database modeling and analysis, data visualization, intelligent decision, and application of intelligent decision methods in healthcare domain. His current research interests include machine learning, imbalanced data learning, and ensemble models.