Hindawi Publishing Corporation
Advances in Artificial Neural Systems
Volume 2015, Article ID 265637, 10 pages
http://dx.doi.org/10.1155/2015/265637

Research Article

Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes

Divya Tomar and Sonali Agarwal

Indian Institute of Information Technology, Allahabad 211012, India

Correspondence should be addressed to Divya Tomar; divyatomar26@gmail.com

Received 30 September 2014; Accepted 23 December 2014

Academic Editor: Chao-Ton Su

There is a necessity for analysis of a large amount of data in many fields such as healthcare, business, industries, and agriculture. Therefore, the need of the feature selection (FS) technique for the researchers is quite evident in many fields of science, especially in computer science. Furthermore, an effective FS technique that is best suited to a particular learning algorithm is of great help for the researchers. Hence, this paper proposes a hybrid feature selection (HFS) based efficient disease diagnostic model for Breast Cancer, Hepatitis, and Diabetes. A HFS is an efficient method that combines the positive aspects of both Filter and Wrapper FS approaches. The proposed model adopts weighted least squares twin support vector machine (WLSTSVM) as a classification approach, sequential forward selection (SFS) as a search strategy, and correlation feature selection (CFS) to evaluate the importance of each feature. This model not only selects relevant feature subset but also efficiently deals with the data imbalance problem. The effectiveness of the HFS based WLSTSVM approach is examined on three well-known disease datasets taken from UCI repository with the help of predictive accuracy, sensitivity, specificity, and geometric mean. The experiment confirms that our proposed HFS based WLSTSVM disease diagnostic model can result in positive outcomes.

1. Introduction

The advancement in the computer hardware technology to store data and the computerization of all fields has resulted in the generation and collection of huge amount of data. Along with it, the use of internet provides us with endless information. The need to tune this enormous data into useful patterns and knowledge has increased the demand of Data Mining among information system researchers. The problem with the large amount of data is its poor quality and redundant information. Moreover, the process of knowledge discovery is also affected. Hence, feature selection (FS) method has been used to tune the large data into useful and reliable data [1, 2]. Feature selection approach is used to select a minimal and relevant feature subset for a given dataset. A feature is said to be relevant, if the class is conditionally dependent on it, that is, if the feature is helpful in predicting class attribute. Another important criterion to check the usefulness of feature is tested on the basis of its redundancy where a feature is highly associated with other features. A good feature subset includes those features which are highly correlated with the class attribute or decision function but are uncorrelated with each other [1–3]. So the task of FS is to search for optimal feature subset depending on the problem to be solved. FS performs better with appropriate learning algorithm.

Nowadays, in medical diagnosis, the use of classification techniques is increasing gradually. The purpose of this research is to develop an effective disease diagnostic model for Breast Cancer, Hepatitis, and Diabetes patients. The maximum deaths of the women between 40 and 55 years of age are caused due to Breast Cancer [4]. As per World Health Organization (WHO) report, about 1.2 million women suffer from Breast Cancer every year [5]. Diabetes is a disease which finds its place among all age groups. The cause of Diabetes is high glucose level in the blood and insufficient secretion of insulin in the body of a person. Hence

the body cells do not react properly to insulin. Insulin is a hormone which helps in breaking down sugar, starch, and other food items into energy. Therefore, in the shortage of the insulin, the glucose is not properly converted to energy and is drained out by urination [6, 7]. Hepatitis is caused by high alcohol consumption, chemicals, and different viruses or drugs. Hepatitis causes liver problems [8]. Better diagnostic techniques and more effective treatments have resulted in decreased mortality rate caused due to these diseases [9].

The class imbalance is also a major problem with real healthcare data. In class imbalance, the data points of one class outnumber the data points of the other class. In medical domain, different classes contain different number of data points. For example, if there are two classes, healthy and sick, then there may be a possibility that both classes contain different number of records. Due to this problem a classifier may be biased towards the majority class (a class with large number of data points) and may produce wrong interpretation for the data points of minority class (a class with less number of data points). So a disease diagnosis system must be able to handle the class imbalance problem. Over the years, various approaches are proposed to deal with this issue such as undersampling, oversampling, and algorithm adjustment approaches [10, 11]. In undersampling, the data points of majority class are removed till both classes have the same number of data points. Sometimes, the useful information may be lost due to undersampling method because it deletes the data points of majority class [12–16], while in oversampling, the data points in minority class are added till both classes have equal number of data points. The problem with oversampling method is that it generates unnatural bias towards minority class. It is found from experimental study that undersampling performs better as compared to the oversampling method [15]. Assigning different cost to training data points is also another approach of handling class imbalance problem in which different costs are assigned to different data points [17–21]. Ensemble methods such as Boosting Support Vector Machine and Random Forest are also given by the researchers [10, 11]. Weighted SVM and weighted least squares SVM (LSSVM) are proposed by Yang et al. and Suykens et al., respectively [22, 23]. All these approaches have concluded that weighted classification approach improves the predictive performance of a classification system.

WLSTSVM is the weighted version of recently proposed least squares twin support vector machine (LSTSVM) in which distinct weights are assigned to the data points of each class to handle class imbalance problem. Here, we select LSTSVM because it has shown better generalization ability and faster computational speed. Since the healthcare data is imbalance in nature and contains many irrelevant features, this research work utilizes the advantages of correlation feature selection and sequential forward selection search strategy and proposes a HFS based WLSTSVM diagnostic model for Breast Cancer, Hepatitis, and Diabetes disease. Thus, in this way, this paper not only deals with redundant and irrelevant features but also solves the class imbalance problem.

The paper is organized into six sections as follows. Section 2 summarizes the methods and results of previous works on diagnosis of Breast Cancer , Diabetes, and Hepatitis diseases. Brief introduction of WLSTSVM is highlighted in Section 3. Proposed disease diagnostic model and experimental results are discussed in Sections 4 and 5, respectively. Finally conclusion is drawn in Section 6.

2. Related Work

Several research techniques have been proposed on disease diagnosis and most of them obtained high predictive accuracies [24–42]. Quinlan used C4.5 decision tree classification model for Breast Cancer disease diagnosis and achieved 94.74% accuracy with 10-fold cross validation [24]. Hamilton et al. achieved 96% accuracy using RIAC approach while Ster and Dobnikar reached 96.8% using Linear Discreet Analysis [26, 27]. In another research work, Peña-Reyes and Sipper obtained 97.36% accuracy with Fuzzy Genetic Algorithm approach [28]. Several researches have been done in the Breast Cancer diagnosis using feature selection based classifier model. Akay and Huang et al. used F-score feature selection based SVM model while Chen et al. used SVM with rough-set based FS for the diagnosis of Breast Cancer disease [29–31]. Rathore and Agarwal have predicted the survivability of Breast Cancer patients using ensemble approach [32]. Polat and Güneş have developed a Breast Cancer disease diagnosis system by using least squares SVM [33]. Karabatak and Ince combined association rules and neural network approach to develop an expert system for the diagnosis of Breast Cancer disease. They applied 3-fold cross validation method to measure the performance of their proposed expert system [34]. Übeyli also used neural network for the same purpose [35]. Temurtas et al. performed a comparative study on the diagnosis of diabetes by using neural network [36]. They did a comparative analysis among multilayer neural network (MLNN) which was trained by Levenberg-Marquardt (LM) approach, probabilistic neural network (PNN), and other existing methods and found that MLNN with LM achieved highest accuracy for the diagnosis of diabetes disease among all methods. Liu and Fu proposed a PSO based SVM with Cuckoo search technique for the diagnoses of heart disease and Breast Cancer disease. Cuckoo search was used for selecting better initial parameters of kernel function and PSO approach searched for the best parameters of SVM [37]. Ganji and Abadeh developed a fuzzy classification system based on ant colony optimization approach, referred to as FCS-ANTMINER, for the diagnosis of Diabetes disease. FCS-ANTMINER system generated fuzzy rules for the diagnosis of Diabetes disease and achieved 84.24% accuracy [38]. Ashraf et al. developed a diseases diagnosis system using Correlation based feature selection and Naïve Bayes approach to diagnose Thyroid, Hepatitis, and Breast Cancer diseases [39]. They analyzed that the feature selection based disease diagnosis model produced more promising results. Polat and Güneş proposed a Hepatitis disease diagnosis system based on feature selection and artificial immune recognition system with fuzzy resource allocation mechanism [40]. Yang et al.

proposed an ensemble based wrapper methods for the feature selection from imbalanced data. They generated multiple balanced datasets from the original imbalance dataset by using sampling approach and evaluated each feature subset using an ensemble of base classifiers [41]. In another research work, Al-Shahib et al. focused on two issues, to identify discriminatory features and to overcome the challenge of data imbalance problem [42]. They applied feature subset selection approach followed by undersampling of majority class and generated a SVM classifier based system for the prediction of protein function from amino acid sequence. They analyzed that this combined approach outperformed other competitive learning algorithms. It is analyzed from all of the above research that feature selection plays an important role for the development of disease diagnosis system.

3. Weighted Least Squares Twin Support Vector Machine

It is evident from the literature results that support vector machine performs classification tasks more efficiently as compared to the decision tree, artificial neural network, and Naïve Bayes [9]. SVM solves a complex quadratic programming problem (QPP) for constructing a maximum margin hyperplane in order to handle the classification tasks. Recently, Jayadeva et al. proposed a twin support vector machine (TWSVM) that handles the classification tasks by generating two nonparallel planes, one plane for each class [43]. Different from SVM, TWSVM solves a pair of simple QPP and obtains two nonparallel planes in such a way that each hyperplane is nearer to the data points of one class while it is as far as possible from the data points of the other class. TWSVM is four times faster than that of traditional SVM and produces comparable results with existing methods. But again the problem with TWSVM classifier is to solve two QPPs. Further, Kumar and Gopal proposed least squares twin support vector machine (LSTSVM) which solves a pair of linear equations instead of two QPPs as in TWSVM [44]. LSTSVM takes lesser computational time and has better generalization ability as compared to the traditional TWSVM. But the above classification techniques are unable to handle class imbalance problem. So, in this research work we propose a weighted LSTSVM approach in which distinct weight is applied to the training data points. Number of data points in each class is used for the selection of appropriate weight parameters. The lower weight is assigned to the class with large number of data points while higher weight is assigned to the class with less number of data points [45]. Let n_1 and n_2 represent the size of positive and negative class, respectively. The weight is assigned to each class on the basis of the following formula:

$$\begin{aligned} &\text{For positive class: if } (n_1 \geq n_2) && \text{Weight} = 1 \\ &\qquad\qquad\quad \text{else} && \text{Weight} = \frac{n_2}{n_1}, \\ &\text{For negative class: if } (n_1 \geq n_2) && \text{Weight} = \frac{n_1}{n_2} \\ &\qquad\qquad\quad \text{else} && \text{Weight} = 1. \end{aligned} \tag{1}$$

3.1. Linear WLSTSVM. Let the data points of positive and negative classes be indicated by two matrices A and B in real space R of d-dimension. For linearly separable data points, the WLSTSVM is formulated as

$$\begin{aligned} \min\,(w_1, b_1, \xi) \quad & \frac{1}{2}\left\|Aw_1 + e_1 b_1\right\|^2 + \frac{c_1}{2}\xi^T \sum_{i=1}^{n_2} W_i \xi \\ \text{s.t.} \quad & -\left(Bw_1 + e_2 b_1\right) + \xi = e_2, \\ \min\,(w_2, b_2, \eta) \quad & \frac{1}{2}\left\|Bw_2 + e_2 b_2\right\|^2 + \frac{c_2}{2}\eta^T \sum_{j=1}^{n_1} W_j \eta \\ \text{s.t.} \quad & \left(Aw_2 + e_1 b_2\right) + \eta = e_1, \end{aligned} \tag{2}$$

where "W" is a diagonal matrix containing weight for each of data points as per (1). c_1 and c_2 are penalty parameters and $e_1 \in R^{n_1}$ and $e_2 \in R^{n_2}$ are two vectors of 1's. w_1, $w_2 \in R^d$ are normal vectors to the hyperplanes and $b_1, b_2 \in R$ represent bias term. The Lagrangian of the above equations are obtained as

$$\begin{aligned} L\,(w_1, b_1, \xi, \alpha) &= \frac{1}{2}\left\|Aw_1 + e_1 b_1\right\|^2 + \frac{c_1}{2}\xi^T W_i \xi \\ &\quad - \alpha^T\left(-\left(Bw_1 + e_2 b_1\right) + \xi - e_2\right), \\ L\,(w_2, b_2, \eta, \beta) &= \frac{1}{2}\left\|Bw_2 + e_2 b_2\right\|^2 + \frac{c_2}{2}\eta^T W_j \eta \\ &\quad - \beta^T\left(\left(Bw_2 + e_1 b_2\right) + \eta - e_1\right), \end{aligned} \tag{3}$$

where $\alpha \in R^{n_1}$ and $\beta \in R^{n_2}$ are nonnegative Lagrangian multipliers. Optimization of the above equations returns the value of normal vectors and biases as follows:

$$\begin{aligned} \begin{bmatrix} w_1 \\ b_1 \end{bmatrix} &= -\left(P^T P\right)^{-1} Q^T \alpha, \\ \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} &= \left(Q^T Q\right)^{-1} P^T \beta, \end{aligned} \tag{4}$$

where $P = [A\ e_1]$ and $Q = [B\ e_2]$. The Lagrangian multipliers are obtained as

$$\begin{aligned} \alpha &= \left[\frac{D_i^{-1}}{c_1} + Q\left(P^T P\right)^{-1} Q^T\right]^{-1} e_2, \\ \beta &= \left[\frac{D_j^{-1}}{c_2} + P\left(Q^T Q\right)^{-1} P^T\right]^{-1} e_1, \end{aligned} \tag{5}$$

where D_i and D_j are two diagonal matrices having equivalent weight corresponding to W_i and W_j. Normal vector and bias are used to construct two nonparallel planes as

$$\begin{aligned} x^T w_1 + b_1 &= 0, \\ x^T w_2 + b_2 &= 0. \end{aligned} \tag{6}$$

The class is assigned to new data point as

$$\text{Class } i = \min\left|x^T w_i + b_i\right| \quad \text{for } i = 1, 2. \tag{7}$$

Table 1: Algorithm for WLSTSVM classifier.

Linear WLSTSVM	Nonlinear WLSTSVM
Define P and Q matrices as $P = [A \;\; e_1]$ and $Q = [B \;\; e_2]$. Calculate weight W for each class using (1). Choose penalty parameters c_1 and c_2 on the basis of validation. Weight and bias required for the construction of nonparallel planes are calculated by using (4). Generate two nonparallel planes by using (6). For new data point, calculate its perpendicular distances from both the planes and a class is assigned to it by using (7).	Define matrix D as $D = [A \;\; B]^T$. Define kernel function and P and Q matrices. $P = [K\ (X_1, D^T) \;\; e_1\ ]$ and $Q = [K\ (X_2, D^T) \;\; e_2\ ]$. Calculate weight W for each class using (1). Choose penalty parameters c_1 and c_2 on the basis of validation. Weight and bias required for the construction of nonparallel planes are calculated by using (10) and (12). Generate two nonparallel kernel generated surfaces by using (8). For a new data point, calculate its perpendicular distances from both kernel generated surfaces and a class is assigned to it by using (14).

3.2. Nonlinear WLSTSVM. WLSTSVM also works well for nonlinearly separable data points with the help of kernel function. Kernel function transforms the data points into higher dimensional space to make easier separation. Then, WLSTSVM constructs nonparallel planes in that space [44]. If "K" is any kernel function then equations of kernel generated surfaces are obtained as

$$\begin{aligned} K\left(x^T, D^T\right)\mu_1 + \gamma_1 = 0, \\ K\left(x^T, D^T\right)\mu_2 + \gamma_2 = 0, \end{aligned} \tag{8}$$

where $D = [A \;\; B]^T$. WLSTSVM for nonlinearly separable data points is formulated as

$$\begin{aligned} \min\left(\mu_1, \gamma_1, \xi\right) \quad & \frac{1}{2}\left\|K\left(A, D^T\right)\mu_1 + e_1\gamma_1\right\|^2 + \frac{c_1}{2}\xi^T\sum_{i=1}^{n_2} W_i\xi \\ \text{s.t.} \quad & -\left(K\left(B, D^T\right)\mu_1 + e_2\gamma_1\right) = e_2 - \xi, \\ \min\left(\mu_2, \gamma_2, \xi\right) \quad & \frac{1}{2}\left\|K\left(B, D^T\right)\mu_2 + e_2\gamma_2\right\|^2 + \frac{c_3}{2}\eta^T\sum_{j=1}^{n_1} W_j\eta \\ \text{s.t.} \quad & \left(K\left(A, D^T\right)\mu_2 + e_1\gamma_2\right) = e_1 - \eta. \end{aligned} \tag{9}$$

The normal vector and bias in the higher dimensional space are measured as follows:

$$z_1 = \begin{bmatrix}\mu_1 \\ \gamma_1\end{bmatrix} = -\left(P^TP\right)^{-1}Q^T\alpha, \tag{10}$$

$$\alpha = \left[\frac{D_i^{-1}}{c_1} + Q\left(P^TP\right)^{-1}Q^T\right]^{-1}e_2, \tag{11}$$

$$z_2 = \begin{bmatrix}\mu_2 \\ \gamma_2\end{bmatrix} = \left(Q^TQ\right)^{-1}P^T\beta, \tag{12}$$

$$\beta = \left[\frac{D_j^{-1}}{c_2} + P\left(Q^TQ\right)^{-1}P^T\right]^{-1}e_2, \tag{13}$$

where $P = \left[K(A, D^T) \;\; e_1\right]$ and $Q = \left[K(B, D^T) \;\; e_2\right]$. A class is assigned to a new data point according to

$$\text{Class } i = \min\left|K\left(x^T, D^T\right)\mu_i + \gamma_i\right| \quad \text{for } i = 1, 2. \tag{14}$$

For a new data point, its perpendicular distance is calculated from each surface and a class is assigned to the data point depending upon the fact of which kernel surface lies nearest to the data point. If x_i and x_j represent vectors in input space, then Gaussian kernel function is formulated as

$$K_G = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right). \tag{15}$$

We use Gaussian kernel function to develop a nonlinear WLSTSVM classifier model. Hybrid feature selection and Grid Search parameter selection approach are also used with this classifier model which further enhances its performance for disease diagnosis. The algorithm of WLSTSVM classifier for both linearly and nonlinearly separable data points is given in Table 1.

4. Methodology and Experiments

4.1. Datasets. This research work proposes the disease diagnosis model for Breast Cancer, Hepatitis, and Diabetes disease. Datasets are taken from UCI Machine Learning Repository [46]. The details of datasets are shown in Table 2. It is clear from Table 2 that all the three datasets contain a different number of instances in each class and imbalance in nature. Imbalance ratio is calculated by taking the ratio of number of data points of majority class with minority class. For Breast Cancer, one class contains 241 and the other contains 458 instances. For Diabetes dataset, one class contains 500 and the other contains 268 instances and in Hepatitis one class contains 123 and the other class contains 32 instances. The numbers of features in Breast Cancer, Diabetes, and Hepatitis are 9, 8, and 19, respectively. Diabetes dataset is a very popular dataset and used for predicting diabetes in women during their pregnancy while Breast Cancer dataset is used to predict whether the tumor is benign or malignant. Hepatitis dataset is also useful for evaluating the performance of a classifier and is used to predict the patient survival rate.

Table 2: Dataset details.

Dataset	Number of features	Number of positive instances	Number of negative instances	Imbalance ratio
Breast Cancer	9	241	458	1.9
Pima Diabetes	8	500	268	1.8
Hepatitis	19	123	32	3.8

4.2. Hybrid Feature Selection Method. The main purpose of the feature selection approach is to select a minimal and relevant feature subset for a given dataset and maintaining its original representation. FS not only reduces the dimensionality of data but also enhances the performance of classifiers. There are two categories of feature selection—Filter and Wrapper methods. Filter method is useful to rank features according to their importance but the main problem is to decide the stopping criterion. Wrapper method is not suitable for large dataset due to exhaustive search. Since Filter is a less computationally expensive feature selection approach and Wrapper is more accurate, hybrid approach combines both, the advantages of Filter and Wrapper methods. In this paper, we used correlation feature selection as a Filter approach which considers the correlation of feature with target feature and selects only those features which show a strong correlation with the target feature and weak correlation with each other. Correlation coefficient between two features X_i and X_j is calculated as

$$\text{Correlation}\left(X_i, X_j\right) = \frac{E\left[\left(X_i - \mu_{Xi}\right)\left(X_j - \mu_{Xj}\right)\right]}{\sigma_{Xi}\sigma_{Xj}}, \quad (16)$$

where σ and μ represent standard deviation and expected values, respectively. Correlation coefficient indicates how strongly two features are related or associated with each other; that is, when the value of one feature is able to predict the value of another feature then they are said to be strongly correlated with each other. Correlation can be estimated from training data points as

$$r_{x(i),x(j)} = \frac{\sum_{k=1}^{m}\left(x_k^i - \overline{x^i}\right)\left(x_k^j - \overline{x^j}\right)}{(m-1)\, S_{x(i)} S_{x(j)}}, \quad (17)$$

where $S_{x(i)}$ and $S_{x(j)}$ indicate standard deviations of training data points, $\overline{x^i}$ and $\overline{x^j}$ are the mean value of sample, and x_k^i and x_k^j represent the value set of features X_i and X_j correspondingly. The above measurement value is used to rank the features according to their individual association with the target feature. A feature set is optimal only when it shows strong correlation with target feature and weak correlation with each other. So, high rank is given to the feature that satisfies strong correlation criteria. Using this condition, the merit of a subset of features is given as

$$M_F = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\,\overline{r_{ff}}}}, \quad (18)$$

where k denotes number of features, r is sample correlation coefficient, c is class, and f is predictive feature. $\overline{r_{cf}}$ and $\overline{r_{ff}}$ represent average value of feature-class correlation and feature-feature correlation, respectively.

The search strategy in Wrapper method either removes or adds features into candidate feature subset and finds an optimal feature subset that maximize the performance of learning algorithm. For example, in case of a classifier, the optimal feature subset maximizes its accuracy. Two common search strategies used in Wrapper methods are forward selection or backward elimination. In this research work we used sequential forward selection (SFS) approach because backward elimination does not work well when initial numbers of features are very large. In this situation, it takes more time and becomes infeasible whereas forward selection is faster when there is a need to select small number of features. At the beginning, SFS assumes an empty feature set and then it iteratively chooses and adds features one by one. In each iteration, the performance of the learning algorithm is evaluated by using generated feature subset. This search continues until adding new feature improves the performance of learning algorithm and stops when there is no improvement in the performance of learning algorithm with respect to the current feature subset.

This research paper proposed a 2-stage feature selection algorithm that combines CFS and SFS feature selection techniques as shown in Figure 1. In the first stage, CFS technique is used to calculate the importance of each feature and then rank the features in descending order. In the second stage, SFS is used for feature selection. SFS constructs a candidate feature subset and selects one feature at a time. It then iteratively adds feature into this candidate feature subset. A temporary WLSTSVM classifier is constructed using current selected feature subset and then its performance is evaluated using 10-fold cross validation method. The feature subset, for which the WLSTSVM performs well, is selected as an optimal feature subset. Then this feature subset is used for final classifier construction.

4.3. Proposed Model. Figure 2 indicates the block diagram of the proposed hybrid feature selection based WLSTSVM model. This model uses hybrid feature selection (CFS + SFS) approach for selecting relevant feature subset. Initially, the preprocessing is performed on the original dataset. Here, we applied normalization to the original dataset. The next step is to divide the dataset into k-part using k-fold cross validation method. One part of this partitioned dataset is used for testing and the remaining $(k-1)$ part is used for training the classifier. CFS is used to rank the feature according to their relevance to the class and then features are arranged according to their rank. Next, SFS selects one best feature at a time and this feature is used to train and test the WLSTSVM classifier. After that, SFS iteratively adds one feature to the

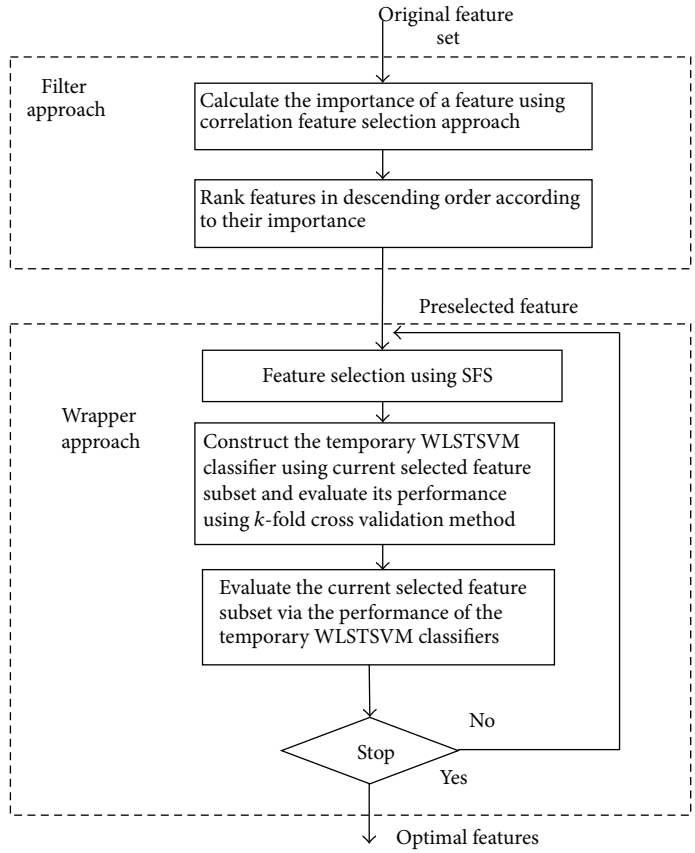

Figure 1: Hybrid feature selection method.

previous feature subset and again this feature subset is used to train and test the model. This process is repeated until all features appear in the subset. Finally, we obtain WLSTSVM disease diagnosis model which has the highest accuracy.

4.4. Performance Evaluation Parameters. In this paper, we used predictive accuracy, sensitivity, specificity, and geometric mean to evaluate the performance of the proposed disease diagnosis model. The formulations of the above mentioned parameters are given below:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\%, \\
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \\
\text{Specificity} &= \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\%, \\
\text{Geometric_Mean} &= \sqrt{\text{Sensitivity} \times \text{Specificity}},
\end{aligned} \tag{19}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Table 3: Accuracy comparison using 10-fold cross validation.

Methods	Pima Diabetes	Hepatitis	Breast Cancer
LSTSVM (without FS)	75.33%	84.28%	93.28%
LSTSVM + undersampling	74.31%	83.33%	92.83%
WLSTSVM	75.67%	86.67%	95.47%
Hybrid FS + LSTSVM	84.51%	85.01%	95.53%
Hybrid FS + LSTSVM + undersampling	82.02%	83.95%	93.31%
Hybrid FS + WLSTSVM	**89.71%**	**87.50%**	**98.55%**

5. Result and Discussion

WLSTSVM classifier model contains several parameters such as sigma and penalty parameters c_1 and c_2 which require some initial value. In this study, we used grid search approach for appropriate parameters selection. The penalty parameters are selected within $\{10^{-8}, 10^{-7}, \ldots, 10^{0}, \ldots, 10^{4}, 10^{5}\}$ range and sigma within $\{2^{-3}, 2^{-2}, \ldots, 2^{0}, \ldots, 2^{7}, 2^{8}\}$ range. The accuracy comparison of different methods for Diabetes, Hepatitis, and Breast cancer is shown in Table 3. We have evaluated the performance of the proposed model and compared

Table 4: Performance comparison of proposed WLSTSVM model.

Performance	Pima Diabetes		Hepatitis		Breast Cancer	
Evaluation	Original features	Reduced features	Original features	Reduced features	Original features	Reduced features
Parameters	8	5	19	12	9	5
Accuracy	75.67%	**89.71%**	86.67%	**87.50%**	95.47%	**98.55%**
Sensitivity	87.09%	**96.94%**	91.67%	**95.14%**	98.00%	**100%**
Specificity	69.57%	**83.11%**	75.54%	**82.96%**	93.55%	**97.37%**
Geometric mean	77.84%	**89.76%**	83.22%	**88.84%**	95.75%	**98.68%**

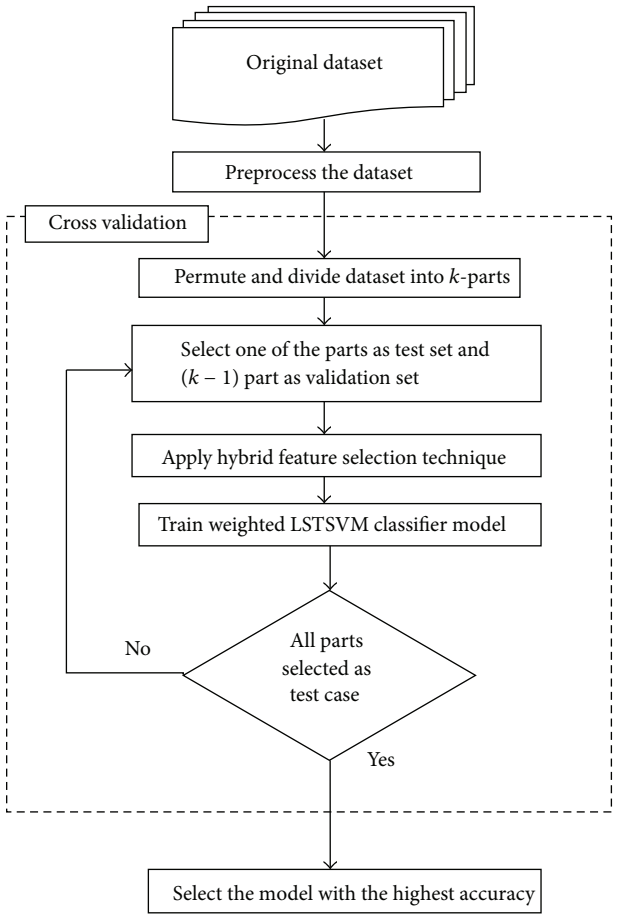

Figure 2: Flow diagram of proposed methodology.

it with the traditional LSTSVM classifier without feature selection, WLSTSVM classifier without feature selection, LSTSVM with undersampling and without feature selection, and HFS based LSTSVM and HFS based LSTSVM with undersampling approach. 10-fold cross validation is used for the performance comparison. Here, we used undersampling for balancing the imbalanced class as compared to the oversampling because undersampling has shown better performance as compared to the oversampling and it does not introduce unnatural bias towards minority class. First of all, LSTSVM is implemented using matlab R2012a. We have checked its accuracy for all given disease datasets without FS. The accuracies of LSTSVM without FS for Diabetes, Hepatitis, and Breast Cancer dataset are 75.33%, 84.28%, and 93.28%, respectively. Then, we applied undersampling

Table 5: Comparison of predictive accuracies of proposed and other classifiers for Pima Diabetes.

Reference	Approach	Predictive accuracy
This Study	**HFS + WLSTSVM**	**89.71%**
Polat et al. [48]	LSSVM	78.21%
Temurtas et al. [36]	MLNN with LM	82.37%
Kahramanli and Allahverdi [49]	Hybrid system	84.2%
Statlog [47]	Logdisc	77.7%
Jankowski [50]	IncNet	77.6%
Weka [47]	Logistic	77.08%
Statlog [47]	SMART	76.8%
Weka [47]	Naïve Bayes	76.04%
Weka [47]	SMO	76.43%
Statlog [47]	BP	75.2%
Ster and Dobnikar [27]	ASI	76.6%
Ster and Dobnikar [27]	MLP + BP	76.4%
Ster and Dobnikar [27]	Fisher disc. analysis	76.5%
Ster and Dobnikar [27]	QDA	59.5%
Ster and Dobnikar [27]	KNN	71.9%
Ster and Dobnikar [27]	CART	72.8%
Zarndt [47]	Bayes	72.2 ± 6.9
Zarndt [47]	C4.5 DT	72.7 ± 6.6
Zarndt [47]	ID3	71.7 ± 6.6
Zarndt [47]	IB3	71.7 ± 5.0
Zarndt [47]	OCN2	65.1 ± 1.1

Table 6: Comparison of predictive accuracies of proposed and other classifiers for Hepatitis.

Reference	Approach	Predictive accuracy
This study	**HFS + WLSTSVM**	**87.50%**
Ster and Dobnikar [27]	CART decision tree	82.7%
Ster and Dobnikar [27]	LVQ	83.2%
Ster and Dobnikar [27]	MLP with BP	82.1%
Ster and Dobnikar [27]	ASR	85%
Ster and Dobnikar [27]	QDA	85.8%
Rafal Adamczak [47]	RBF (Tooldiag)	79.0%
Rafal Adamczak [47]	MLP + BP (Tooldiag)	77.4%

Table 7: Comparison of predictive accuracies of proposed and other classifiers for Breast Cancer.

Reference	Approach	Predictive accuracy
This study	HFS + WLSTSVM (10 × CV)	98.55%
Karabatak and Ince [34]	AR + NN	97.4%
Chen et al. [31]	GA	96.99%
Sousa et al. [51]	Discrete particle swarm optimization	94%
Akay [29]	FS + SVM (train: 75%-test-25%)	99.51%
Quinlan [24]	C4.5 (10 × CV)	94.74%
Hamilton et al. [26]	RIAC (10 × CV)	95.00%
Ster and Dobnikar [27]	LDA (10 × CV)	96.80%
Polat and Güneş [33, 40]	LS-SVM	98.53%
Abonyi and Szeifert [52]	Supervised fuzzy custering	95.57%
Goodman et al. [53]	AIRS	97.20%
Bennett and Blue [54]	SVM (5 × CV)	97.20%
Şahan et al. [55]	**Fuzzy AIS - KNN (10 × CV)**	**99.14%**

approach to the dataset in order to balance the data of both classes and checked the predictive accuracy of LSTSVM with the balanced data. LSTSVM with undersampling has obtained 74.31%, 83.33%, and 92.83% accuracies for Diabetes, Hepatitis, and Breast Cancer disease correspondingly. Again, we checked the performance of our proposed approach WLSTSVM to deal with class imbalance problem without any FS approach. As shown in Table 3, it is clear that WLSTSVM performs well as compared to the LSTSVM + undersampling approach. Next, we applied HFS approach to LSTSVM which enhances its accuracy for all given disease datasets. In the same way, HFS based LSTSVM with undersampling also performs well compared to the LSTSVM with undersampling. Then we applied HFS to the proposed WLSTSVM classifier model and achieved far better accuracy for Diabetes, Hepatitis, and Breast Cancer disease dataset. The results indicate that HFS based WLSTSVM is a better choice for the construction of the disease diagnosis model. The proposed disease diagnosis model not only selects the relevant and useful features but also handles the class imbalance problem very well.

The performance of our proposed WLSTSVM classifier model is checked with and without HFS approach. It is clear from Table 4 that the predictive accuracy and geometric mean of WLSTSVM with reduced features are much better compared to the WLSTSVM with original features. The proposed model selects 5 features for Diabetes, 12 features for Hepatitis, and 5 features for Breast Cancer datasets as shown in Table 4. For Breast Cancer patients, the proposed model selects Bare Nuclei, uniformity of cell shape and cell size, clump thickness, and marginal adhesion features. The proposed model predicts the Diabetes in pregnant women using their Plasma Glucose Level, Body Mass Index, number of times a woman is pregnant, Pedigree, and their age. For Hepatitis, it selects Bilirubin, age, Histology, Antivirals, Alk Phosphatase, SGOT, Liver Big, Liver Firm, Steroid, Spleen Palpable, Varices, and Ascites to predict a patient suffering from hepatitis will either die or survive. The value of sensitivity and specificity is also improved with the feature selection approach. The sensitivity of the proposed approach is increased from 0.8709 to 0.9694 for Diabetes, from 0.9167 to 0.9514 for Hepatitis, and from 0.9800 to 1 for Breast Cancer dataset. In the same manner, specificity is also varied from 0.6957 to 0.8311 for Diabetes, from 0.7554 to 0.8296 for Hepatitis, and from 0.9355 to 0.9737 for Breast Cancer dataset.

Tables 5, 6, and 7 indicate the comparison of predictive accuracy of our proposed HFS based WLSTSVM model with other existing classifiers for Diabetes, Hepatitis, and Breast Cancer diseases [47]. Our proposed approach based on WLSTSVM with HFS obtains 89.71% accuracy for Diabetes which is better compared to the other existing approaches. It again obtains 87.50% accuracy for Hepatitis and 98.55% accuracy for Breast Cancer diagnosis which is comparable to the other existing models.

6. Conclusion

This paper is concerned with two issues-class imbalance problem and the need of feature selection. We propose a WLTSVM-based disease diagnosis model with HFS for diagnosis of Breast Cancer, Diabetes, and Hepatitis diseases. HFS combines the positive aspects of Filter and Wrapper approaches. This model not only is useful for selecting significant features but it also handles the class imbalance problem very effectively. The predictive accuracy for Pima Diabetes dataset is 89.71% which is far better than other existing approaches. For Hepatitis and Breast Cancer diseases, the predictive accuracy is 87.50% and 98.55%, respectively, which is also comparable with other existing models. The accuracy improvement is showing more variation from 74.31% to 89.71% for Diabetes dataset, 92.83% to 98.55% for Breast Cancer dataset, and 83.33% to 87.50% for Hepatitis dataset. Experimental results show the effectiveness of the proposed disease diagnosis model. Thus, the above results indicate that feature selection with proper handling of class imbalance problem may enhance the accuracy of classifiers up to a significant level which is necessary while handling real time data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] C. Lemnaru, *Strategies for dealing with real world classification problems [Ph. D. thesis]*, Faculty of Computer Science and Automation, Universitatea Technica, Din Cluj-Napoca, Cluj-Napoca, Romania, 2012.

[2] D. Tomar and S. Agarwal, "A survey on pre-processing and post-processing techniques in data mining," *International Journal of Database Theory & Application*, vol. 7, no. 4, 2014.

[3] H.-H. Hsu, C.-W. Hsieh, and M.-D. Lu, "Hybrid feature selection by combining filters and wrappers," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8144–8150, 2011.

[4] 2014, http://www.imaginis.com/breast-cancer-resource-center.

[5] A. Jemal, M. M. Center, C. DeSantis, and E. M. Ward, "Global patterns of cancer incidence and mortality rates and trends," *Cancer Epidemiology Biomarkers and Prevention*, vol. 19, no. 8, pp. 1893–1907, 2010.

[6] E. L. Mohamed, R. Linder, G. Perriello, N. Di Daniele, S. J. Pöppl, and A. De Lorenzo, "Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis," *Diabetes, Nutrition and Metabolism*, vol. 15, no. 4, pp. 215–221, 2002.

[7] D. Tomar and S. Agarwal, "Predictive model for diabetic patients using hybrid twin support vector machine," in *Proceedings of the 5th International Conferences on Advances in Communication Network and Computing (CNC '14)*, pp. 1–9, 2014.

[8] http://www.medicalnewstoday.com/articles/145869.php.

[9] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science & Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[10] X. Wang Benjamin, *Boosting support vector machine [M.S. thesis]*, 2005.

[11] B. X. Wang and N. Japkowicz, "Boosting support vector machines for imbalanced data sets," *Knowledge and Information Systems*, vol. 25, no. 1, pp. 1–20, 2010.

[12] J. Laurikkala, "Instance-based data reduction for improved identification of difficult small classes," *Intelligent Data Analysis*, vol. 6, no. 4, pp. 311–322, 2002.

[13] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[14] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, vol. 97, pp. 179–186, 1997.

[15] C. Ling and C. Li, "Data mining for direct marketing—specific problems and solutions," in *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, pp. 73–79, 1998.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[17] P. Domingos, "MetaCost: a general method for making classifiers cost-sensitive," in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, San Diego, Calif, USA, August 1999.

[18] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI '01)*, pp. 973–978, Seattle, Wash, USA, August 2001.

[19] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659–665, 2002.

[20] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM '03)*, pp. 435–442, Melbourne, Fla, USA, November 2003.

[21] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.

[22] X. Yang, Q. Song, and Y. Wang, "A weighted support vector machine for data classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 5, pp. 961–976, 2007.

[23] J. A. K. Suykens, J. de Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness and sparce approximation," *Neurocomputing*, vol. 48, pp. 85–105, 2002.

[24] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996.

[25] D. Tomar and S. Agarwal, "Feature selection based least square twin support vector machine for diagnosis of heart disease," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 2, pp. 69–82, 2014.

[26] H. J. Hamilton, N. Shan, and N. Cercone, *RIAC: A Rule Induction Algorithm Based on Approximate Classification*, Computer Science Department, University of Regina, 1996.

[27] B. Ster and A. Dobnikar, "Neural networks in medical diagnosis: comparison with other methods," in *Proceedings of the International Conference on Engineering Applications of Neural Networks (EANN '96)*, pp. 427–430, 1996.

[28] C. A. Peña-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artificial Intelligence in Medicine*, vol. 17, no. 2, pp. 131–155, 1999.

[29] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[30] C.-L. Huang, H.-C. Liao, and M.-C. Chen, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis," *Expert Systems with Applications*, vol. 34, no. 1, pp. 578–587, 2008.

[31] 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 Systems with Applications*, vol. 38, no. 7, pp. 9014–9022, 2011.

[32] N. Rathore and S. Agarwal, "Predicting the survivability of breast cancer patients using ensemble approach," in *Proceedings of the International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT '14)*, pp. 459–464, IEEE, February 2014.

[33] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing*, vol. 17, no. 4, pp. 694–701, 2007.

[34] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3465–3469, 2009.

[35] E. D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," *Expert Systems with Applications*, vol. 33, no. 4, pp. 1054–1062, 2007.

[36] H. Temurtas, N. Yumusak, and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8610–8615, 2009.

[37] X. Liu and H. Fu, "PSO-based support vector machine with Cuckoo search technique for clinical disease diagnoses," *The Scientific World Journal*, vol. 2014, Article ID 548483, 7 pages, 2014.

[38] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14650–14659, 2011.

[39] M. Ashraf, G. Chetty, D. Tran, and D. Sharma, "Hybrid approach for diagnosing thyroid, hepatitis, and breast cancer based on correlation based feature selection and Naïve bayes," in *Neural Information Processing*, Lecture Notes in Computer Science, pp. 272–280, Springer, Berlin, Germany, 2012.

[40] K. Polat and S. Güneş, "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation," *Digital Signal Processing*, vol. 16, no. 6, pp. 889–901, 2006.

[41] P. Yang, W. Liu, B. B. Zhou, S. Chawla, and A. Y. Zomaya, "Ensemble-based wrapper methods for feature selection and class imbalance learning," in *Advances in Knowledge Discovery and Data Mining*, pp. 544–555, Springer, Berlin, Germany, 2013.

[42] A. Al-Shahib, R. Breitling, and D. Gilbert, "Feature selection and the class imbalance problem in predicting protein function from sequence," *Applied Bioinformatics*, vol. 4, no. 3, pp. 195–203, 2005.

[43] Jayadeva, R. Khemchandani, and S. Chandra, "Twin support vector machines for pattern classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[44] M. A. Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7535–7543, 2009.

[45] D. Tomar, S. Singhal, and S. Agarwal, "Weighted least square twin support vector machine for imbalanced dataset," *International Journal of Database Theory and Application*, vol. 7, no. 2, pp. 25–36, 2014.

[46] Dataset, 2014, http://archive.ics.uci.edu/ml/datasets.html.

[47] http://www.is.umk.pl/projects/datasets.html.

[48] K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine," *Expert Systems with Applications*, vol. 34, no. 1, pp. 482–487, 2008.

[49] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1, pp. 82–89, 2008.

[50] N. Jankowski, "Controlling the structure of neural networks that grow and shrink," in *Proceedings of the 2nd International Conference on Cognitive and Neural Systems*, 1998.

[51] T. Sousa, A. Silva, and A. Neves, "A particle swarm data miner," in *Progress in Artificial Intelligence*, vol. 2902 of *Lecture Notes in Computer Science*, pp. 43–53, Springer, Berlin, Germany, 2003.

[52] J. Abonyi and F. Szeifert, "Supervised fuzzy clustering for the identification of fuzzy classifiers," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2195–2207, 2003.

[53] D. E. Goodman Jr., L. C. Boggess, and A. B. Watkins, "Artificial immune system classification of multiple-class problems," in *Proceedings of the Artificial Neutral Networks in Engineering Conference (ANNIE '02)*, pp. 179–184, 2002.

[54] K. P. Bennett and J. A. Blue, "Support vector machine approach to decision trees," in *Proceedings of the IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence*, vol. 3, pp. 2396–2401, Anchorage, Alaska, USA, May 1998.

[55] S. Şahan, K. Polat, H. Kodaz, and S. Güneş, "A new hybrid method based on fuzzy-artificial immune system and *k*-nn algorithm for breast cancer diagnosis," *Computers in Biology and Medicine*, vol. 37, no. 3, pp. 415–423, 2007.