Developing Diabetes Disease Classification Model using Sequential Forward Selection Algorithm

Emrana Kabir Hashi
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh

Md. Shahid Uz Zaman
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh

Md. Rokibul Hasan
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh

ABSTRACT

Data mining techniques are being used extensively in healthcare sector to discover hidden pattern and relationship between patients' record and their medical diagnosis dataset. In the concept of disease prediction, high classification accuracy can be obtained from accurately pre-processed and trained model. But existence of unimportant and irrelevant attributes in the training dataset may decrease the predictive accuracy and increase the time complexity in training phase. To increase the accuracy and efficiency, feature selection technique is frequently used in data mining. In this paper, a sequential forward selection based wrapper approach is proposed to select optimal and informative feature subset. It is known that diabetes mellitus is the most serious health problem and the complications lead to cause of death. So the aim of this research is to identify the significant attributes and classify diabetes dataset. The proposed approach is used to build the classifier models like Decision tree, K-Nearest Neighbor and Support Vector Machine produces the accuracies of 81.17%, 86.36% and 87.01% respectively. Finally, from results it is clear that the proposed model is performing better with high accuracy comparing the similar existing models. In the research, the Pima Indian diabetes dataset is used.

General Terms

Data mining, Knowledge Discovery, Feature selection, Sequential Forward Selection (SFS), Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM)

Keywords

Classification, Feature Selection, Wrapper Approach, Feature selection, SFS, Pima Indian Diabetes Dataset, C4.5, KNN, SVM

1. INTRODUCTION

Nowadays, medical data mining is playing a vital role in the healthcare sector to extract and analyze the hidden pattern from the clinical dataset that can be used for decision making in the biomedical research [1, 2]. Data mining and Knowledge Discovery process involves statistics, machine learning, database, information science and visualization [3]. Data mining application in healthcare is usefulness to predict disease which can provide better decision to physicians, doctors and a cost effective treatment to patients [4]. In the knowledge discovery process, after performing data cleaning, data integration, data selection and data transformation, different data mining task such as classification, regression, clustering, association rule and summarization are used to uncover the hidden relationship of data and evaluate the valuable knowledge [5-7].

Diabetes has become a serious disease and important to diagnosis diabetes at early stage due to its mortality rate. Diabetes is a disease in which the blood glucose levels get increase which is due to the defects in secretion of insulin, or its action, or both [8-10]. In diabetes, the cells of a person produce insufficient amount of insulin or unable to use insulin properly and efficiently that further leads to hyperglycemia and type-2 diabetes [11, 12]. A lot of rehearses has been done on this diabetes disease to predict this in early stage to overcome its complications such as strokes, blindness, kidney failure, damage and failure of several organs.

Every year a huge amount of clinical data are gathered from healthcare center for different diseases. In decision making process, it is important to discover the pattern and valuable information from those huge amount of data. This will assist doctors and reduce healthcare cost and waiting time for patient [6]. There are different classification algorithms such as Decision Tree, K Nearest Neighbor (KNN), Naïve Bayes, and Support Vector Machine (SVM) etc. help to diagnosis or predict the disease. But less significant features are responsible to decrease the performance of these classifiers. Better analysis can be obtained using feature selection which is the process of identifying and removing irrelevant and redundant data from dataset in order to improve the performance of the machine learning algorithms [13]. There are many advantages of feature selection such as increase the classification accuracy and decrease the computational time.

This paper aimed to predict diabetes using classification and feature selection approach. First of all developed an expert system to predict diabetes disease using Decision tree, KNN and SVM classifier. Then introduced and implemented a hybrid prediction model using sequential forward selection (SFS) based wrapper feature selection and classification. Finally compared the performance of different classifiers in terms of accuracy, sensitivity and specificity.

2. RELATED WORKS

Various feature selection based hybrid prediction models have been proposed to achieve better accuracy for different disease prediction. The present stage of some related research works and study papers is summarized in this section.

In paper [14], a Genetic Algorithm based Wrapper feature selection Hybrid Prediction Model (GWHPM) has proposed with different classifier. Here, Pima Indian Diabetes Dataset has used and different number of attributes are selected to build a classifier models using this GWHPM. After applying

K-means clustering 625 instances have selected to remove the outliers. Finally, different classifiers have used to evaluate the performance such as Decision Tree (accuracy- 94.75%, specificity- 88.489%, and sensitivity- 97.07%), Naive Bayes (accuracy- 97.86%, specificity- 98.2%, and sensitivity- 97.07%), k nearest neighbor (accuracy- 97.47%, specificity- 94.244%, and sensitivity- 98.67%) and Support Vector Machine (accuracy- 97.86%, specificity- 94.244%, and sensitivity- 98.2%).

In paper [15], a feature selection based diabetes prediction has proposed with filter method and SVM classifier. F-score method and K-means clustering was used for feature selection from the Pima Indian diabetes dataset. Then SVM classifier has used to evaluate the performance of this model (accuracy- 98%, specificity- 97.77%, and sensitivity- 97.79%).

Another similar study [16], presented a comparison of different types of feature selection methods using SVM classifier on diabetes disease prediction. The study showed that the feature selection by using SVM-RFE (accuracy- 97%) gives better performance than F-score (accuracy- 86%), Genetic algorithm (accuracy- 94%), K-means (accuracy- 96%) and RelifF (accuracy- 87%) methods.

The research paper [17] developed a method using combined dataset of Diabetes disease. Here Fselect (accuracy- 63.54%, specificity- 43.00%, and sensitivity- 99.80%), wrapper (accuracy- 70.69%, specificity- 38.36% and Sensitivity- 89.95) and Ranker (accuracy- 72.61%, specificity- 41.04%, and sensitivity- 90.76%) methods are used for feature selection and LIBSVM for classification feature.

Another paper [18], developed a system where linear SVM classifiers combined with wrapper or embedded feature selection methods selected 39 features and gave 0.969 of the area under the curve.

In paper [19], proposed a feature selection model using Symmetrical Uncertainty Attribute set Evaluator and Fast Correlation-Based Filter (FCBF) and showed that LIBSVM classifier (selected 4 feature and accuracy- 77.99%) provided better performance than existing system (all feature accuracy- 77.47%).

The research [20] experimented GA-SVM model for predicting several real world datasets and the results showed that this model provided significant improvement in the performance of classification in comparison with Grid search.

The study [21], proposed GA-based approach with RBF kernel and the Grid algorithm on 11 real-world datasets from UCI database and result compared with the Grid algorithm.

This paper [22] proposed a feature selection model using artificial bee colony algorithm and SVM for diagnosis of hepatitis (accuracy- 94.92%), liver disorders (accuracy- 74.81%) and diabetes (accuracy- 79.29%) disease.

In paper [23], a feature selection via supervised model construction (FSSMC), an optimization of ReliefF and three complementary classification techniques (Naive Bayes, IB1 and C4.5) were applied to the data to predict diabetes and finally achieved 95% accuracy and 98% sensitivity.

This study [24] presented different feature selection models using F-score feature selection method and k-means clustering select the optimal feature subsets of the diabetes datasets that enhances the performance of the Support Vector Machine classifier.

In this paper [25], proposed a Backward Search feature selection approach for finding an optimum feature subset that enhances the classification accuracy of Naive Bayes and SVM classifier.

3. METHODOLOGY

3.1 Overview of Proposed System

In this paper, a Sequential Forward Selection (SFS) based Wrapper Feature Selection and classification model is proposed to predict diabetes disease. Fig.1 represents the block diagram of propose system. The main steps and detailed explanations are described in following section.

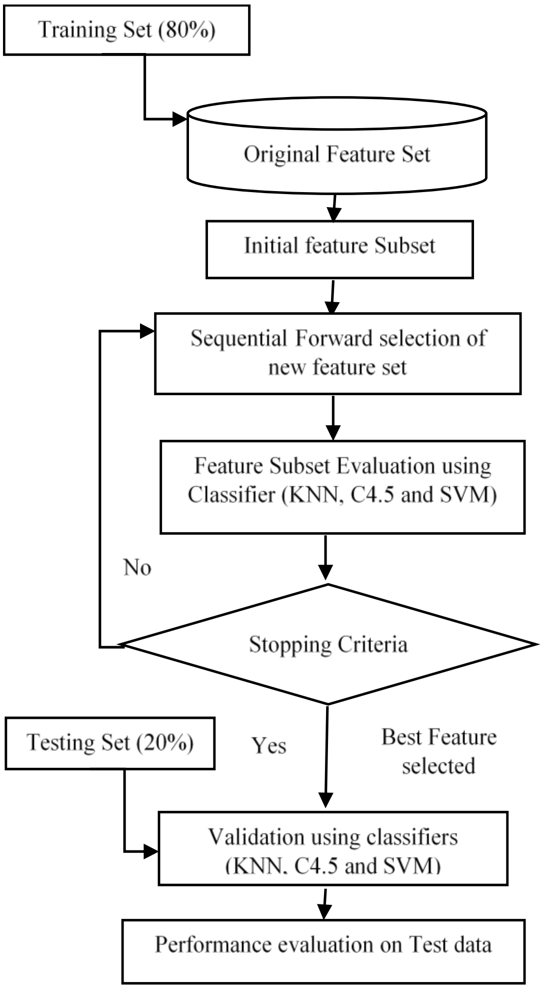
Training Set (80%)
Original Feature Set
Initial feature Subset
Sequential Forward selection of new feature set
Feature Subset Evaluation using Classifier (KNN, C4.5 and SVM)
No
Stopping Criteria
Testing Set (20%)
Yes
Best Feature selected
Validation using classifiers (KNN, C4.5 and SVM)
Performance evaluation on Test data

Fig.1: Overview of proposed system

3.2 Data Preprocessing

Data preprocessing is the first step in proposed system to identify and processing the noisy, incomplete, irreverent and inconsistent value of some attribute. For Pima Indian Dataset, data cleaning is performed as there any missing value or value zero are replaced with mean value of that attribute. Then scaling total dataset like 80% are selected as training set and rest 20% are selected as testing set.

3.3 Feature Selection

Sequential Forward Selection (SFS) based wrapper selection approach has used to select the best feature. Wrapper model approach uses the classifier as induction algorithm to measure the good subset [27]. Wrapper methods generally achieve better accuracy rate and use cross validation to avoid over-fitting and sometimes they are too expensive for large dimensional database in terms of computational complexity and time since each feature set considered must be evaluated with the classifier algorithm used [28, 29]. Sequential forward selection is a simplest greedy search algorithm. It starts with empty set and sequentially add features until all feature subsets are evaluated. Here, 10 fold cross validation is used on the training set for classification to evaluate the selected feature. The search space is 2N where N is the number of attribute. So stopping criteria is necessary based on evaluation function. Some of the commonly used criteria such as search is complete or next iteration fails to produce a better subset [30-32].

3.4 Classification

The classifiers KNN, C4.5 and SVM are used in this SFS based wrapper feature selection approach. A short description of these classifiers are presented in following sections.

3.4.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) approach has been used in different data analysis applications such as pattern recognition, data mining, databases and machine learning due to its simplicity and high accuracy [10][33]. This is very simple to understand but works very well in practical problems. In this algorithm, the classifier first computes the distance between the current instances to all the instances in the data set and then the k-nearest neighbors are identified. The test tuple is assigned to the most frequent class among these neighbors [12][14]. Shortcoming of KNN is that it is lazy algorithm more exactly all training set are required in testing phase. KNN is a supervised learning algorithm which classifies new data based on minimum distance from the new data to the K nearest neighbor [26]. The proposed work has used Euclidean Distance which formulated by equation (1) to define the closeness [26].

$$d(X,Y) = \sqrt{\sum_{i=0}^{n}(Xi - Yi)^2} \qquad (1)$$

Where, X=(x_1, x_2……,x_n) and Y=(y_1,y_2…..y_n)

3.4.2 Decision Tree

Decision tree is a tree structure model and widely used as a supervised classification algorithm. ID3, C4.5, C5.0, J48, CART and CHAID are powerful accepted decision tree algorithm. In this paper C4.5 classifier used as the dataset consists continuous values of attributes. Based on the attribute values recursively splitting a dataset into decision node and leaf node. Mathematically, entropy is degree of elements which calculated with the help of probability of the attribute item. In every stage of iteration, decision tree select an attribute with best information gain. The decision nodes are found by calculating the highest information gain from all attributes. The entropy, info (p, T) and information gain is calculated by equation (2) (3) and (4) respectively which is used to classify unknown records [1][14][26].

$$\mathrm{Gain(p) = F(Info\ (T) - \ Info\ (p, T)} \qquad (2)$$

$$\mathrm{Info(T) = Entropie(p) = -\sum_{i=1}^{n} pi \times \log(pi)} \qquad (3)$$

$$\mathrm{Info(p, T) = \sum_{j=1}^{n}\left(pj \times Entropie(pi)\right)} \qquad (4)$$

Here, F = number of known sample/total number of sample in the dataset for a given attribute, pi = the set of probability distribution, T= Test, pj= the set of all possible values for attribute T.

3.4.3 Support Vector Machine

SVM is a supervised learning algorithm which categorizes data by finding a model of optimal heperplane separating different dimensional data with a maximum interclass margin [1][5][33]. It was first introduced by Corinna Cortes and Vladimir Vapnik for classification and regression. It is one of the most popular machine learning and statistics algorithm. It is a classification techniques used for both linear (linear kernel) and non-linear dataset (RBF, sigmoid and polynomial kernel). In this proposed system, linear kernel has used as kernel function because of diabetes dataset contains two classes. The example of linearly separable problem is showed in Fig.2, assume some training data D contains a set of n point expressed as D= {(X_1,Y_2), (X_2,Y_2)….(X_n,Y_n)}. Here, Y_i= -1/1 that denotes the class to which data point X_n belongs and any hyper plane can be written as the set of points satisfying W^T.X + b=0 where W is normal to hyper plane, b is a constant. Now W^T. X_i + b<=-1 constraint added for xi of the first class and W^T. X_i + b>=1 for xi of second class then try to minimize ||w|| Subject to Y_i (W^T. X_i + b)>=1, for all i [15].

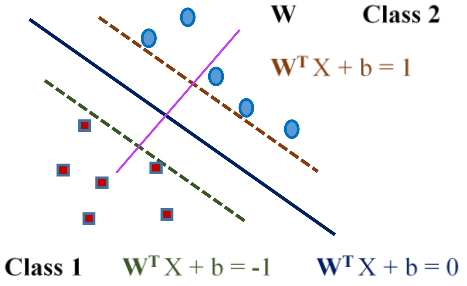

Fig.2: SVM training with two classes.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Dataset Description

The Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases [1-3] [14-16] [26] has used as diabetes dataset. It contains 768 instances with 9 numerical valued attribute. There are two classes, where 0 represent as tested negative and 1 represent as tested positive. In this dataset, 500 instances are tested negative and 268 instances are tested positive. The description of all attributes are listed in Table 1.

Table 1. Attribute description

Attribute name	Attribute description(all numeric valued)	Mean	Standard deviation
Pregnancy	Number of times pregnant	3.8	3.4
Plasma	Plasma glucose concentration	120.9	32.0
Pres	Diastolic blood pressure (mm Hg)	69.1	19.4
Skin	Triceps skin fold thickness (mm)	20.5	16.0
Insulin	2-Hour serum insulin (mu U/ml)	79.8	115.2
Mass	Body mass index (wgt in kg/(height in m)^2)	32.0	7.9
Pedi	Diabetes pedigree function	0.5	0.3
Age	Age (years)	33.2	11.8
Class	Class variable (0 or 1)	-	-

4.2 Performance Evaluation

In this proposed system, 80% of instances are randomly selected as training set and rest 20% as testing set. Attribute selection process is applied in the training set to select the best feature subset. Wrapper approach features selected on the classifier model which used as an induction algorithm to check the stopping criteria. 10 fold cross validation is performed to evaluate the performance of this classifier model on training data. In 10 fold cross validation criteria, first divide the whole dataset into 10 equal size subsets and trained nine subsets and testes the rest one, finally percentage these ten correctly classified accuracy [5]. When best model generated with optimal feature subset then classifiers are applied to evaluate the performance of test set.

A confusion matrix is a representation of classification results which is used to calculate accuracy, sensitivity and specificity [1]. Table 2 shows a general confusion matrix where TP, FN, FP and TN represent as True Positive, False Negative, false Positive and True Negative respectively. True Positive (TP) implies that diabetic patients who are classified as diabetic patients, whereas False Negative (FN) implies that diabetic patients who are classified as non-diabetic patients. On the other hand, False Positive (FP) implies that non-diabetic patients who are classified as diabetic patients. Commonly, the best learning algorithm is selected based upon the performance of the classifiers in terms of high accuracy [34].

Table 2. Confusion Matrix

Confusion Matrix	Classified as Healthy	Classified as not Healthy
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

The below formula (5), (6) and (7) are used to calculate sensitivity, specificity and accuracy [1] [3] [34]:

Sensitivity is calculated by dividing the true positive (TP) samples to the sum of true positive (TP) and false negative (FN) samples.

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \qquad (5)$$

Specificity is calculated by dividing the true negative (TN) samples to the sum of true negative (TN) and false positive (FP) samples.

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \qquad (6)$$

Calculation of accuracy is performed by taking ratio of truly classified samples (true negative, true positive) to the total number of samples.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \qquad (7)$$

4.3 Baseline Performance with All Attribute

The first approach of this work is to apply KNN, C4.5 and SVM classifier without feature selection on diabetes dataset of Pima Indian dataset which obtained from UCI machine learning repository. Here, 10 fold cross validation has used to evaluate the performance. The performance of the KNN, C4.5 and SVM classifier is given in Table 3. Using all attribute KNN, C4.5 and SVM has obtained 74.47%, 76.82% and 77.17% accuracy respectively.

Table 3. Performance comparison of different classifiers on the original diabetes dataset

Sl. No	Classifier	Accuracy
1	KNN	74.48%
2	C4.5	76.82%
3	SVM	77.17%

4.4 Results of Proposed System

This research work experimented for feature selection and classification based model construction on diabetes data of Pima Indian dataset. As mentioned before, wrapper feature selection approach depends on classification model. As a part of feature selection approach SFS wrapped with different classifiers namely KNN, C4.5 and SVM. The diabetes dataset contains 8 attributes so that 28 or 256 number of search space is generated during the selection process. The termination criteria set as traverse and complete all search space or accuracy value does not improve during the iteration. In this phase, 80% of total instances or 614 number of instances are used of training purpose. To select the best feature subset 10 fold cross validation has used to obtain the accuracy of classifier.

In the process of selecting optimal feature subset and training the model, the validation of KNN classifier has evaluated with the value of K=11. The experimental results showed that five attributes Pregnancy, Plasma, Mass, Pedi and Age are selected. Then these attributes are used for classification of testing samples. The second algorithm decision tree C4.5 has wrapped with sequential forward selection to select the best feature. The experimental results show that four attributes Plasma, Age, Mass and Pedi are selected using this model that are used in for the test set classification purpose. The third one is SVM classification algorithm which uses linear kernel function and the C=10 is selected here. During the feature selection phase SVM trained with best feature using 10 fold cross validation. The experimental results show that five attributes Plasma, Age, Mass, Pedi and Pres are selected for classification.

Then, rest 20% of total instances or 154 number of instances are used of testing phase. Three classifiers classify the test set and obtained confusion matrix is showed in Table 4. Here presented the number of true positive, false positive, true negative and false negative values obtained from KNN, C4.5 and SVM classifier.

Table 4. Confusion matrix of different classifiers

Classifier	True Positive	False Positive	True Negative	False Negative
KNN	26	18	99	11
C4.5	29	15	104	6
SVM	28	16	106	4

The performance of classifier has calculated using formula (5) (6) and (7). Table 5 shows the comparison of performance of different classifiers in terms of accuracy, specificity and sensitivity.

Table 5. Performance comparison of different classifiers of proposed system

Classifier	No. of selected attribute	Accuracy	Specificity	Sensitivity
KNN	5	81.17%	84.62%	70.27%
C4.5	4	86.36%	87.39%	82.86%
SVM	5	87.01%	86.89%	87.50%

This experimental results shows that after wrapper based feature selection the accuracy of classifiers have improved. By implementing proposed system, an accuracy 81.17%, 86.36% and 87.01% are obtained by using KNN, C4.5 and SVM classifier. So it can be observed that SVM classifier performed better than KNN and C4.5 classifier for this dataset.

4.5 Result Analysis

A graphical view of a comparison between existing model and proposed model is shown in Fig.3.

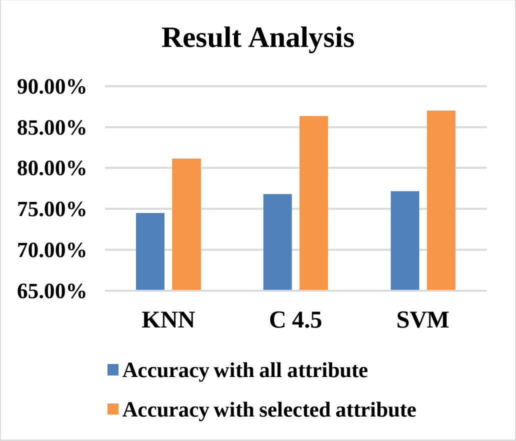

Fig.3: Graphical comparison of classifiers accuracy between existing model with all attributes and proposed model with selected attributes.

Here, using all attributes classifier KNN, C4.5 and SVM have provided 74.47%, 76.82% and 77.17% accuracies respectively. Using feature selection model with optimal attributes classifier KNN, C4.5 and SVM have provided 81.17%, 86.36% and 87.01% accuracies respectively. So, it is clear that proposed model performing better with high accuracy comparing with existing model.

5. CONCLUSION

This research work proposes a sequential forward selection based wrapper feature selection approach to select the optimal feature subset to improve the classification accuracy. This is experimented on Pima Indian diabetes dataset and the proposed system shows the better performance than existing system. This model provides 37.5% and 50% feature reduction with about 10% increase of classification accuracy. The performance of this system is evaluated by comparing KNN (81.17%), C4.5 (86.36%) and SVM (87.01%) classifier. It can conclude that for both existing system and proposed system SVM provides better accuracy than other classifiers for diabetes dataset. This proposed approach helps the doctors, physicians and medical practitioners to predict and diagnosis disease with important and relevant feature. It will provide an informative way to plan a strategy of efficient and accurate treatment. Finally this system can be tested on other disease dataset with using different classifiers. The future work focuses on increasing classification accuracy using different feature section approach wrapped with different classifier.

6. REFERENCES

[1] Karthikeyani, V., I. Parvin Begum, K. Tajudin, and I. Shahina Begam. "Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction." *International Journal of Computer Applications* 60, no. 12 (2012): 26-31.

[2] Raghavendra, S., and M. Indiramma. "Classification and Prediction Model using Hybrid Technique for Medical Datasets." *analysis* 127, no. 5 (2015): 20-25.

[3] Vijayan, Veena, and Aswathy Ravikumar. "Study of data mining algorithms for prediction and diagnosis of diabetes mellitus." *International journal of computer applications* 95, no. 17 (2014): 12-16.

[4] Almarabeh, Hilal, and Ehab F. Amer. "A Study of Data Mining Techniques Accuracy for Healthcare." *International Journal of Computers and Applications* 168, no. 3 (2017): 12-16.

[5] Parthiban, G., A. Rajesh, and S. K. Srivatsa. "Diagnosing Vulnerability of Diabetic Patients to Heart Diseases using Support Vector Machines." *International Journal of Computer Applications*48, no. 2 (2012): 45-49.

[6] Vispute, Nilesh Jagdish, Dinesh Kumar Sahu, and Anil Rajput. "An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set." *International Journal of Computer Applications* 131, no. 2 (2015): 6-11.

[7] Parthiban, G., A. Rajesh, and S. K. Srivatsa. "Diagnosis of heart disease for diabetic patients using naive bayes method." *International Journal of Computer Applications* 24, no. 3 (2011): 7-11.

[8] Tambade, Shital, Madan Somvanshi, Pranjali Chavan, and Swati Shinde. "SVM based Diabetic Classification and Hospital Recommendation." *International Journal of Computer Applications*167, no. 1 (2017): 40-43.

[9] Karthikeyan, T., and K. Vembandadsamy. "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus." *International Journal of Computers and Applications* 5, no. 5 (2015): 96-104.

[10] Sethi, Harsha. "Diabetes Diagnoser: Expert System for Diagnosis of Diabetes Type-II." *International Journal of Computer Applications* 148, no. 11 (2016): 19-25.

[11] Sumathy, Mythili, Mythili Thirugnanam, Praveen Kumar, T. M. Jishnujit, and K. Ranjith Kumar. "Diagnosis of Diabetes Mellitus based on Risk Factors." *International Journal of Computers and Applications* 10, no. 4 (2010): 1-4.

[12] Karthikeyan, T., and K. Vembandadsamy. "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus." *International Journal of Computers and Applications* 5, no. 5 (2015): 96-104.

[13] Asir, D., S. Appavu, and E. Jebamalar. "Literature Review on Feature Selection Methods for High-

Dimensional Data." *International Journal of Computer Applications* 136, no. 1 (2016): 9-17.

[14] Anirudha, R. C., Remya Kannan, and Nagamma Patil. "Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data." In *Industrial and Information Systems (ICIIS), 2014 9th International Conference on*, pp. 1-6. IEEE, 2014.

[15] Gandhi, Khyati K., and Nilesh B. Prajapati. "Diabetes prediction using feature selection and classification." *International Journal of Advance Engineering and Research Development* (2014).

[16] Kaur, Sandeep, and Sheetal Kalra. "Feature Extraction Techniques Using Support Vector Machines In Disease Prediction." In Proceedings of the4th International Conference on Science, Technology and Management (ICSTM-16), India International Centre, New Delhi. 2016.

[17] Negi, Anjli, and Varun Jaiswal. "A first attempt to develop a diabetes prediction method based on different global datasets." In *Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on*, pp. 237-241. IEEE, 2016.

[18] Cho, Baek Hwan, Hwanjo Yu, Kwang-Won Kim, Tae Hyun Kim, In Young Kim, and Sun I. Kim. "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods." *Artificial intelligence in medicine* 42, no. 1 (2008): 37-53.

[19] Balakrishnan, Sarojini, and Ramaraj Narayanaswamy. "Feature selection using FCBI in type II diabetes databases." *International Journal of the Computer, the Internet and the Management* 17, no. 1 (2009): 50-8.

[20] Phan, Anh Viet, Minh Le Nguyen, and Lam Thu Bui. "Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems." *Applied Intelligence* 46, no. 2 (2017): 455-469.

[21] Huang, Cheng-Lung, and Chieh-Jen Wang. "A GA-based feature selection and parameters optimizationfor support vector machines." *Expert Systems with applications* 31, no. 2 (2006): 231-240.

[22] Uzer, Mustafa Serter, Nihat Yilmaz, and Onur Inan. "Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification." *The Scientific World Journal* 2013 (2013).

[23] Huang, Yue, Paul McCullagh, Norman Black, and Roy Harper. "Feature selection and classification model construction on type 2 diabetic patients' data." *Artificial intelligence in medicine* 41, no. 3 (2007): 251-262.

[24] Gandhi, Khyati K., and Nilesh B. Prajapati. "Diabetes prediction using feature selection and classification." *International Journal of Advance Engineering and Research Development* (2014).

[25] Balakrishnan, Sarojini, Ramaraj Narayanaswamy, Nickolas Savarimuthu, and Rita Samikannu. "SVM ranking with backward search for feature selection in type II diabetes databases." In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pp. 2628-2633. IEEE, 2008.

[26] Hashi, Emrana Kabir, Md Shahid Uz Zaman, and Md Rokibul Hasan. "An expert clinical decision support system to predict disease using classification techniques." In Electrical, Computer and Communication Engineering (ECCE), International Conference on, pp. 396-400. IEEE, 2017.

[27] Kohavi, Ron, and George H. John. "Wrappers for feature subset selection." Artificial intelligence 97, no. 1-2 (1997): 273-324.

[28] Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Feature subset selection problem using wrapper approach in supervised learning." International journal of Computer applications 1, no. 7 (2010): 13-17.

[29] Sun, Ming-an, Qing Zhang, Yejun Wang, Wei Ge, and Dianjing Guo. "Prediction of redox-sensitive cysteines using sequential distance and other sequence-based features." BMC bioinformatics17, no. 1 (2016): 316.

[30] Laimighofer, Michael, Jan Krumsiek, Florian Buettner, and Fabian J. Theis. "Unbiased prediction and feature selection in high-dimensional survival regression." Journal of Computational Biology 23, no. 4 (2016): 279-290.

[31] Mao, Kezhi Z. "Orthogonal forward selection and backward elimination algorithms for feature subset selection." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34, no. 1 (2004): 629-634.

[32] Bagherzadeh-Khiabani, Farideh, Azra Ramezankhani, Fereidoun Azizi, Farzad Hadaegh, Ewout W. Steyerberg, and Davood Khalili. "A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results." Journal of clinical epidemiology 71 (2016): 76-85.

[33] Thirumal, P. C., and N. Nagarajan. "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study." ARPN Journal of Engineering and Applied Science 10, no. 1 (2015).

[34] Daghistani, Tahani, and Riyad Alshammari. "Diagnosis of Diabetes by Applying Data Mining Classification Techniques." *International Journal of Advanced Computer Science and Applications (IJACSA)* 7, no. 7 (2016): 329-332.