A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes

Ratna Patil[1], Sharavari Tamane[2]
[1]Department of Computer Engineering, Babasaheb Ambedkar Marathwada University, India
[2]Department of Computer Engineering, JNEC, Aurangabad, India

Article Info

Article history:

Received Jan 9, 2018
Revised Mar 15, 2018
Accepted Jul 4, 2018

Keyword:

Data mining
Diabetes mellitus
Classification
Machine learning
ROC

ABSTRACT

Data mining techniques are applied in many applications as a standard procedure for analyzing the large volume of available data, extracting useful information and knowledge to support the major decision-making processes. Diabetes mellitus is a continuing, general, deadly syndrome occurring all around the world. It is characterized by hyperglycemia occurring due to abnormalities in insulin secretion which would in turn result in irregular rise of glucose level. In recent years, the impact of Diabetes mellitus has increased to a great extent especially in developing countries like India. This is mainly due to the irregularities in the food habits and life style. Thus, early diagnosis and classification of this deadly disease has become an active area of research in the last decade. Numerous clustering and classifications techniques are available in the literature to visualize temporal data to identify trends for controlling diabetes mellitus. This work presents an experimental study of several algorithms which classifies Diabetes Mellitus data effectively. The existing algorithms are analyzed thoroughly to identify their advantages and limitations. The performance assessment of the existing algorithms is carried out to determine the best approach.

Corresponding Author:

Ratna Patil,
Department of Computer Engineering,
Babasaheb Ambedkar Marathwada University, India.
Email: ratna.nitin.patil@gmail.com

1. INTRODUCTION

Diabetes mellitus is a group of metabolic diseases in which a person experiences high blood glucose levels either because the body produces inadequate insulin or the body cells do not respond properly to the insulin produced by the body. Patients with diabetes often experience frequent urination (polyuria), increased thirst (polydipsia) and increased hunger (polyphagia) [1], [2]. The 3 Types of Diabetes:

a. Type 1 Diabetes
   In this type of diabetes, the body does not produce enough insulin. This type pf diabetes is also referred to as insulin-dependent diabetes, juvenile diabetes or early-onset diabetes. Type 1 diabetes usually develops before a person is 40-years-old i.e., in early adulthood or teenage. Patients with type 1 diabetes will need to take insulin injections for the rest of their life. They must also ensure proper blood-glucose levels by carrying out regular blood tests and following a special diet.
b. Type 2 Diabetes
   In Type 2 Diabetes, the body does not produce enough insulin or the cells in the body display insulin resistance. Some people may be able to control their type 2 diabetes symptoms by losing weight, following a healthy diet, doing plenty of exercise, and monitoring their blood glucose levels. However, type 2 diabetes is typically a progressive disease – it gradually gets worse – and the patient will probably end up having to take insulin, usually in tablet form. Being overweight, physically inactive and eating the wrong

foods all contribute to our risk of developing type 2 diabetes. The risk of developing Type 2 diabetes also increases with age [3], [4].

c. Gestational Diabetes

This type affects females during pregnancy. Some women have very high levels of glucose in their blood, and their bodies are unable to produce enough insulin to transport all of the glucose into their cells, resulting in progressively rising levels of glucose. The majority of gestational diabetes patients can control their diabetes with exercise and diet. Between 10% to 20% of them will need to take some kind of blood-glucose-controlling medications. Undiagnosed or uncontrolled gestational diabetes can raise the risk of complications during child birth.

2. PROCESS WORK FLOW

Figure 1 shows the process of conceptual framework.

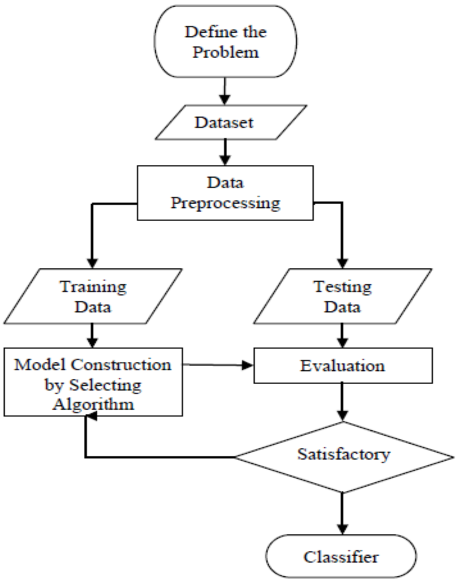

Figure 1. The process of conceptual framework

3. MODEL CONSTRUCTION

Model Construction will take place using Logistic Regression, K Nearest Neighbors (KNN), SVM, Gradient Boost, Decision tree, MLP, Random Forest and Gaussian Naïve Bayes and their performance will be evaluated [5], [6].

3.1. Logistic regression

Logistic regression is basically a linear model for classification rather than regression. It is also known as the logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, we use logistic regression to model probabilistically described outcomes of a single trial. It is a basic model which describes dichotomous output variables and can be extended for disease classification prediction [7], [8]. Suppose there are N input variables where their values are indicated by m_1, m_2, m_3,…,$m_{N.}$ Let us assume that the P probability of that an event will occur and 1- P be a probability that event will not occur. Logistic regression model is given by

$$\log\left(\frac{p}{1-p}\right) = \log it\ (P) = \beta_0 + \beta_1 m_1 + \cdots + \beta_N m_N \tag{1}$$

3.2. KNN

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). Case is classified by a majority vote of its neighbors, with the case being assigned to theclass most common amongst its K nearest neighbors measured by a distance function [9].

$$d(x,x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \tag{2}$$

If K = 1, then the case is simply assigned to the class of its nearest neighbor. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance. More formally, given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

a. It runs through the whole dataset computing d between x and each training observation. We'll call the K points in the training data that are closest to x the set A. Note that K is usually odd to prevent tie situations.
b. It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label. (Note I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise)

$$P(y = j|X = x) = \frac{1}{K}\sum_{i \in A} I\left(y^{(i)} = j\right) \tag{3}$$

Finally, our input x gets assigned to the class with the largest probability.

3.3. SVM

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane [10]. In the linear classifier model, we assumed that training examples plotted in space. These data points are expected to be separated by an apparent gap. It predicts a straight hyperplane dividing 2 classes. The primary focus while drawing the hyperplane is on maximizing the distance from hyperplane to the nearest data point of either class. The drawn hyperplane called as a maximum-margin hyperplane [11]. The classification process of SVM classifier. Figure 2 shows the SVM hyper planes.

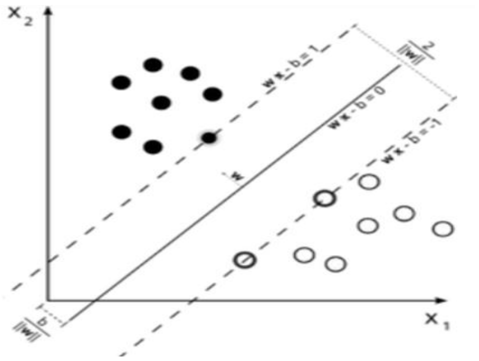

Figure 2. SVM hyper planes

$$\vec{w}x_i - b \geq 1 \text{ if } \theta_i = 1$$
$$\vec{w}x_i - b \leq 1 \text{ if } \theta_i = -1 \tag{4}$$

Where $||\vec{w}||$ is normal vector to the hyperplane, θ_i denotes classes and x_i denotes features. The Distance between two hyperplanes is $\frac{2}{||\vec{w}||}$, to maximize this distance denominator value should be minimized i.e, $||\vec{w}||$ shouldbe minimized. For proper classification, we can build a combined equation:

$$||\vec{w}||_{min} \text{ for } \theta_i(\vec{w}x_i - b) \geq 1 \; \forall i = 1, 2, \cdots, n \tag{5}$$

3.4. Gradient boost

Boosting refers to a family of algorithms that are able to convert weak learners to strong learners. The main principle of boosting is to fit a sequence of weak learners−models that are only slightly better than random guessing, such as small decision trees−to weighted versions of the data. More weight is given to examples that were misclassified by earlier rounds. The predictions are then combined through a weighted majority vote (classification) or a weighted sum (regression) to produce the final prediction. Gradient Tree Boosting s a generalization of boosting to arbitrary differentiable loss functions. It can be used for both regression and classification problems. Gradient Boosting builds the model in a sequential way.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{6}$$

At each stage the decision tree $h_m(x)$ is chosen to minimize a loss function L given the current model $F_{m-1}(x)$:

$$F_m(x) = F_{m-1}(x) + argmin_h \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + h(x_i)\big) \tag{7}$$

The algorithms for regression and classification differ in the type of loss function used.

3.5. Decision tree

Decision tree is a simple, deterministic data structure for modelling decision rules for a specific classification problem. At each node, one feature is selected to make separating decision. We can stop splitting once the leaf node has optimally less data points. Such leaf node then gives us insight into the final result (Probabilities for different classes in case of classification). The most decisive factor for the efficiency of a decision tree is the efficiency of its splitting process as shown in Figure 3. We split at each node in such a way that the resulting purity is maximum. Well, purity just refers to how well we can segregate the classes and increase our knowledge by the split performed [12].

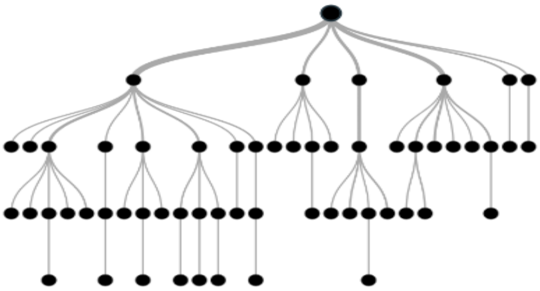

Figure 3. Decision tree

3.6. MLP

The Multilayer Perception (MLP) is perhaps the most popular network architecture in use today both for classification and regression. MLPs are feed forward neural networks which are typically composed of several layers of nodes with unidirectional connections, often trained by back propagation [13], [14]. The learning process of MLP network is based on the data samples composed of the N-dimensional input vector *x* and the M-dimensional desired output vector *d*, called destination. By processing the input vector *x*, the MLP produces the output signal vector *y*(*x*, *w*) where *w* is the vector of adapted weights. The error signal produced actuates a control mechanism of the learning algorithm. The corrective adjustments are designed to make the output signal $y_k(k = 1, 2,\ldots, M)$ to the desired responsed_k in a step by step manner. If a multilayer perceptron has a linear activation function in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron, then linear algebra shows that any number of layers can be reduced to a two-layer input-output model. In MLPs some neurons use a nonlinear activation function that was developed to model the frequency of action potentials, or firing, of biological neurons [15]. The two common activation functions are both sigmoids, and are described by

$$y(v_i) = \tanh(v_i) \text{ and } \mathrm{y}(v_i) = (1 + \mathrm{e}^{-v_i})^{-1} \tag{8}$$

The first is a hyperbolic tangent that ranges from -1 to 1, while the other is the logistic function, which is similar in shape but ranges from 0 to 1. Here y_i is the output of the ith node (neuron) and v_i is the weighted sum of the input connections. The learning algorithm of MLP is based on the minimization of the error function defined on the learning set (x_i, d_i) for i =1, 2,…, N using the Euclidean norm:

$$E(w) = \frac{1}{2}\sum_{i=1}^{N} \| y(x_i, w) - d_i \|^2 \quad (9)$$

The minimization of this error leads to the optimal values of weights. The most effective methods of minimization are the gradient algorithms, from which the most effective is the Levenberg–Marquard algorithm for medium size networks and conjugate gradient for large size networks. Figure 4 shows the MLP structure.

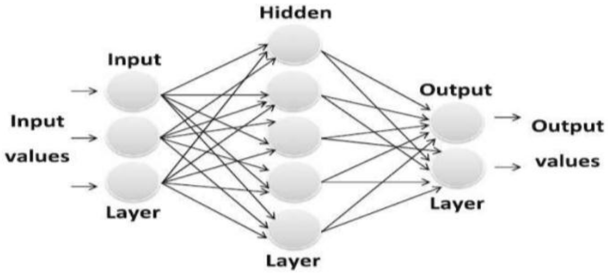

Figure 4. MLP structure

3.7. Random forest

Random forest is just an improvement over the top of the decision tree algorithm. The core idea behind Random Forest is to generate multiple small decision trees from random subsets of the data (hence the name “Random Forest”). Each of the decision tree gives a biased classifier (as it only considers a subset of the data). They each capture different trends in the data as shown in Figure 5.

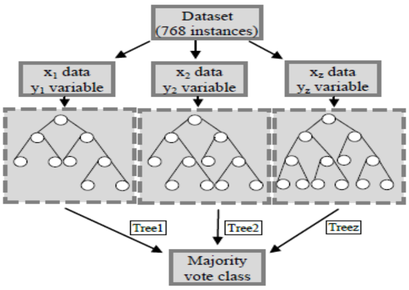

Figure 5. Random forest

This ensemble of trees is like a team of experts each with a little knowledge over the overall subject but thorough in their area of expertise. Now, in case of classification the majority vote is considered to classify a class. In analogy with experts, it is like asking the same multiple choice question to each expert and taking the answer as the one that most no. of experts vote as correct. In case of Regression, we can use the avg. of all trees as our prediction. In addition to this, we can also weight some more decisive trees high relative to others by testing on the validation data [16]. Majority vote is taken from the experts (trees) for classification.

3.8. Gaussian naïve bayes

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution [17]. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values as shown in Figure 6.

The likelihood of the features is assumed to be Gaussian, hence, conditional probability is given by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (10)$$

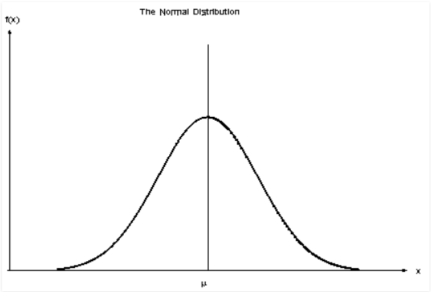

Figure 6. Gaussian curve

4. PERFORMANCE EVALUATION CRITERIA FOR MODEL

To analyze and compare the performance of the data mining methods presented in our study, we apply various statistics such as MAE, RMSE, NRMSE and Confusion Matrix computed as follows [18]-[20].

a. Mean absolute error (MAE)
MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{11}$$

b. Root mean square error (RMSE)
RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average of squared differences between prediction and actual observation.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{12}$$

c. Confusion matrix
The information about actual and predicted classification system is hold by the Confusion matrix. It demonstrates the accuracy of the solution to a classification problem. Table 1 shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study. Tp is the number of correct predictions that an instance is positive. Fn is the number of incorrect predictions that an instance is negative. Fp is the number of incorrect predictions that an instance is positive and $\mathrm{t_n}$ is the number of correct predictions that an instance is negative.

Table 1. The Confusion Matrix for a two class Classifier

		Predicted	
		Positive	Negative
Actual	Positive	$\mathrm{t_p}$	$\mathrm{f_n}$
	Negative	$\mathrm{f_p}$	$\mathrm{t_n}$

d. Precision
Precision looks at the ratio of correct positive observations. The formula is,

$$P = \frac{t_p}{t_p + f_p} \tag{13}$$

e. Recall/true positive rate/sensitivity
Recall is also known as sensitivity or true positive rate. It’s the ratio of correctly predicted positive events.

$$R = \frac{t_p}{t_p + f_n} \tag{14}$$

f. Accuracy
The proportion of the total number of predictions that were correct is known to be as Accuracy (AC). It shows overall effectiveness of classifier. It is determined using the equation:

$$AC = \frac{t_p + t_n}{t_p + f_n + f_p + t_n} \tag{15}$$

g. ROC
A receiver operating characteristics (ROC) graph is a method for conceptualize, organizing and selecting classifiers on the basis of their performance [21], [22]. ROC graphs are bi-dimensional graphs where on the Y axis t_p rate is plotted and on the X axis f_p rate is plotted. A ROC graph describe relative trade-offs between benefits (true positives) and costs (false positives) [23].

5. EXPERIMENTAL RESULTS AND OBSERVATIONS

In Experimental studies the dataset have been partitioned between 70–30 % (538–230) for training and testing purpose. Table 2 shows Logistic Regression being the simplest classifier have performed well with an accuracy of 79.54%, while having relative absolute error 21.65%. Among the applied algorithms Logistic Regression has higher accuracy which is quite well and having the lowest RMSE value 46.52%. Table 2 shows comparative analysis of algorithm in terms of Mean Absolute Error, Root Mean Square Error and Accuracy score [4]. ROC is plotted for all the algorithms. More the area covered better is the classifier. These measurements are taken by using Spyder tool on Pima Indian Diabetes Data set taken from UCI repository. The results are shown in Table 2. The results may be improved by applying large size updated data sets of realistic context. However we need to apply other machine learning algorithms using real data set before generalizing the results.

Table 2. Summary of Prediction for different Algorithms

Algorithm	MAE	RMSE	Accuracy Score
Logistic Regression	0.2165	0.4652	0.7954
KNNeighbors	0.2511	0.5011	0.7489
Linear SVM	0.3203	0.5660	0.6797
Gradient Boosting	0.2078	0.4558	0.7922
Decision tree	0.2684	0.5181	0.7316
MLP	0.3593	0.5994	0.6407
Random Forest	0.2381	0.4880	0.7619
Gaussian Naïve Bayes	0.2381	0.4880	0.76

Figure 7 shows the comparative analysis in terms of accuracy.

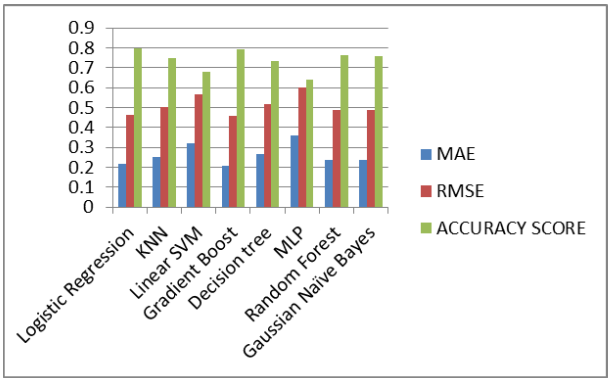

Figure 7. Comparative Analysis in terms of Accuracy

Table 3 shows Comparison of Algorithms for training time, Training and Score. Logistic Regression gives the best testing score of 77%. Neural Net Classifier takes the longest time to train the dataset. Recall, Precision, Accuracy calculated using confusion matrix and the comparison is done.

Table 3. Comparison of Algorithms for training time and Score

Classifier	Train_Score	Test_Score	Training_time
Naïve Bayes	0.7672	0.7619	0.0041
Logistic Regression	0.7672	0.7836	0.0190
Random Forest	0.9963	0.7706	0.1146
K Nearest Neighbors	0.7896	0.7489	0.0030
Gradient Boosting	0.9330	0.7836	0.3414
Decision tree	1.0000	0.7403	0.0079
Linear SVM	1.0000	0.6797	0.1777
Neural Net	0.7523	0.7143	0.9177

Figure 8 shows the comparative analysis in terms of score and training time.

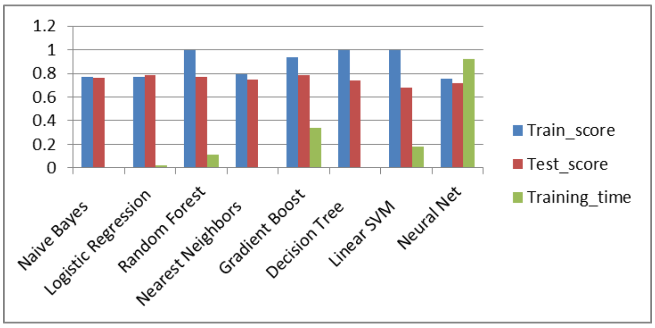

Figure 8. Comparative Analysis in terms of Score and training time

Table 4 shows the results for PIMA on algorithms.

Table 4. Results for PIMA on algorithms

Classifier	Precision	Recall	F1–Measure	ROC
Naïve Bayes	0.7299	0.6962	0.7070	0.70
Logistic Regression	0.7622	0.7157	0.7298	0.75
Random Forest	0.7288	0.6998	0.7096	0.70
K Nearest Neighbors	0.7110	0.6903	0.6978	0.69
Gradient Boosting Classifier	0.7736	0.7471	0.7540	0.75
Decision Tree	0.6960	0.7061	0.70	0.71
Linear SVM	0.3398	0.50	0.4046	0.50
Neural Net	0.6123	0.6249	0.6128	0.62

Figure 9 shows the comparative analysis of algorithms in terms of ROC.

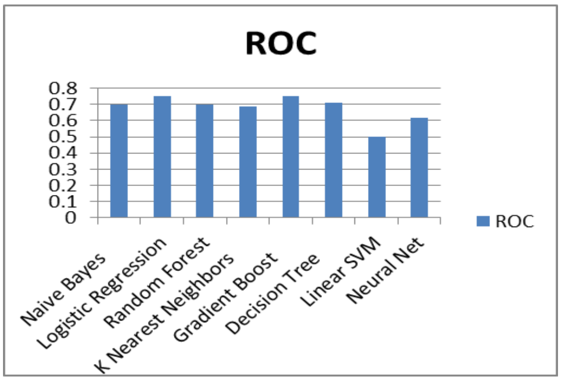

Figure 9. Comparative analysis of algorithms in terms of ROC

Figure 10 shows the comparative analysis of algorithms in terms of recall, precision, accuracy, ROC.

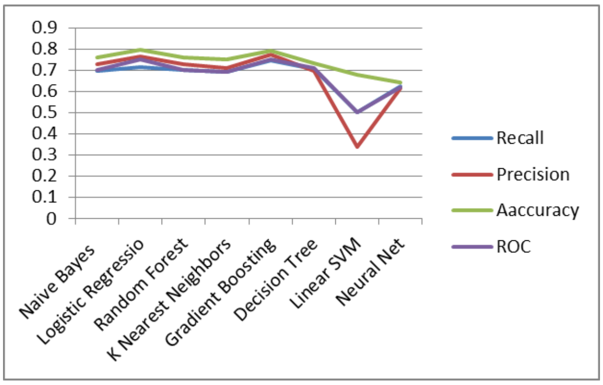

Figure 10. Comparative analysis of algorithms in terms of Recall, Precision, Accuracy, ROC

6. CONCLUSION

In this paper, we have inspected the execution of eight machine learning algorithms namely Logistic Regression, K Nearest Neighbors (KNN), SVM, Gradient Boost, Decision tree, MLP, Random Forest and Gaussian Naïve to predict the population who are most likely to develop diabetes on Pima Indian diabetes data. The performance measurement is compared in terms of MAE, RMSE, ROC, Test Accuracy, Precision and Recall obtained from the test set. Here the studies conclude that Logistic Regression and Gradient Boost classifiers achieve higher test accuracy of 79 % than other classifiers. Further, we plan to recreate our study of Classification models by introducing the intelligent machine learning algorithms applied to a large collection of real life data set. Using Gaussian Fuzzy decision tree algorithm for the diagnosis accuracy obtained was 75% [24]. Design of a Diabetic Diagnosis System Using Rough Sets accuracy obtained was 76% [25]. The results obtained by our experimental algorithms can be further improved by applying outlier detection before classification. This study can be used to select best classifier for predicting diabetes.

REFERENCES

[1] K. Selvakuberan, *et al.*, "An Efficient Feature Selection Method for Classification in Health Care Systems Using Machine Learning Techniques", *IEEE,* pp. 8610-8615, 2011.

[2] M. Seera, *et al.*, "A Hybrid Intelligent System for Medical Data Classification", *Expert Elsevier: Systems with Applications*, vol. 41, pp. 2239-2249, 2014.

[3] T. Karthikeyan, *et al.*, "An Intelligent Type-II Diabetes Mellitus Diagnosis Approach using Improved FP-growth with Hybrid Classifier Based Arm Research", *Journal of Applied Sciences, Engineering and Technology*, vol. 11, no. 5, pp. 549-558, 2015.

[4] D. K. Karumanchi, *et al.*, "Early diagnosis of Diabetes mellitus through the eye", *2nd International Conference on Endocrinology*, 2014.

[5] M. B. Wankhade and A. A. Gurjar, "Analysis of Disease using Retinal Blood Vessels Detection", *International Journal of Engineering and Computer Science,* vol. 5, no. 12, pp. 19644-19647, 2016.

[6] S. B. Choi, *et al.*, "Screening for Prediabetes Using Machine Learning Models", *Hindawi Publishing CorporationComputational and Mathematical Methods in Medicine*, 2014.

[7] M. S. Klein and J. Shearer, "Metabolomics and Type 2 Diabetes: Translating Basic Research into Clinical, Application", *Hindawi Publishing Corporation Journal of Diabetes Research*, 2015.

[8] M. Kothainayaki and P. Thangaraj, "Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm". Article can be accessed online at http://www.publishingindia.

[9] M. N. Devi, *et al.*, "An Amalgam KNN to Predict Diabetes mellitus", *IEEE*, 2013.

[10] N. H. Barakat, *et al.*, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, 2010.

[11] T. Santhanam and M. S. Padmavathi, "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis", *Procedia Computer Science,* vol. 47, pp. 76-83, 2015.

[12] A. G. Karegowda, *et al.*, "Rule based classification for diabetic patients using cascaded K-means and decision tree C4.5", *International Journal of Computer Applications*, vol. 45, no. 12, pp. 0975-8887, 2012.

[13] H. Temurtas, *et al.*, "A Comparative Study on Diabetes disease Diagnosis using Neural Networks", *Elsevier: Expert Systems with Applications*, vol. 36, 2009.

[14] K. Srinivas, *et al.*, "Hybrid Approach for Prediction of Cardiovascular Disease Using Class Association Rules and MLP", *International Journal of Electrical and Computer Engineering*, vol. 6, no. 4, pp. 1800-1810, 2016.

[15] R. Kala, *et al.*, "Diagnosis of Breast cancer by Modular Evolutionary Neural Networks", *Inderscience: International Journal of Biomedical Engineering and Technology (IJBET)*, vol. 7, no. 2, pp. 194-211, 2011.

[16] C. Hsieh, *et al.*, "Novel Solutions for an old disease: Diagnosis of Acute Appendicitis with random forest, Support Vector Machines, and Artificial Neural Networks", *Surgery*, vol. 149, no. 1, pp. 87-93, 2011.

[17] Md. M. Mottalib, *et al.*, "Detection of the Onset of Diabetes Mellitus by Bayesian Classifier Based Medical Expert System", *Transaction on Machine Learning and Artificial Intelligence*, 2016.

[18] P. Yasodha and M. Kannan, "Analysis of a population of diabetic patient's databases in Weka tool", *Proceedings of the International Journal of Scientific & Engineering Research*, vol. 2, no. 5, 2011.

[19] H. Mahajan, *et al.*, "Health Intervention Impact Assessment on Glycemic Status of Diabetic Patients", *International Journal of Diabetes Research*, pp. 73-80, 2012.

[20] M. Abdar, *et al*., "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases", *International Journal of Electrical and Computer Engineering*, vol. 5, no. 6, pp. 1569-1576, 2015.

[21] R. M. Rahman and F. Afoz, "Comparison of Various Classification Techniques using different Data Mining Tools for Diabetes Diagnosis", *Journal of Software Engineering and Applications*, vol. 6, pp. 85-97, 2013.

[22] V. Pellakuri, *et al*., "Performance Analysis and Optimization of Supervised Learning Techniques for Medical Diagnosis Using Open Source Tools", *International Journal of Computer Science and Information Technologies*, vol. 6, no. 1, pp. 380-383, 2015.

[23] R. R. Rao and K. Makkithaya, "Learning from a Class Imbalanced Public Health Dataset: A Cost-based Comparison of Classifier Performance", *International Journal of Electrical and Computer Engineering*, vol. 7, no. 4, pp. 2215-2222, 2017.

[24] K. V. S. R. P. Varma, *et al.*, "A Computational Intelligence Approach for a better Diagnosis of Diabetic Patients", *Computers and Electrical Engineering*, vol. 40, pp. 1758-1765, 2014.

[25] M. Anouncia S., *et al.*, "Design of a Diabetic Diagnosis System using Rough Sets", *Cybernetics and Information Technologies*, vol. 13, no. 3, 2013.