Annals of Emerging Technologies in Computing (AETiC)
Vol. 3, No. 3, 2019

Research Article

A Diabetic Disease Prediction Model Based on Classification Algorithms

Ravinder Ahuja[1,*], Subhash C. Sharma[1] and Maaruf Ali[2]

[1]Electronics & Computer Discipline, Indian Institute of Technology, Roorkee Campus, India
ahujaravinder022@gmail.com; scs60fpt@iitr.ac.in
[2]International Association for Educators and Researchers (IAER), London, UK
maaruf@ieee.org
*Correspondence: ahujaravinder022@gmail.com

Received: 25th April 2019; Accepted: 15th May 2019; Published: 1st July 2019

Abstract: Diabetes is one of the chronic diseases in the world, 246 million people are inflicted by this disease and according to a World Health Organisation (WHO) report, this figure will increase to 380 million sufferers by 2025. Many other debilitating and critical health issues may further develop if this disease is not diagnosed or remain unidentified. Machine Learning (ML) techniques are now being used in various fields like education, healthcare, business, recommendation system, etc. Healthcare data is complex and high in dimensionality and contains irrelevant information - due to this, the prediction accuracy is low. The *Pima Indians Diabetes Dataset* was used in this research, it consisted of 768 records. Firstly, the missing values are replaced by the median followed by Linear Discriminant Analysis. Using the Python programming language, feature selection techniques is applied in combination with five classification algorithms: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Logistic Regression, Random Forest and Decision Tree. The aim of this paper is to compare the different classification algorithms in order to predict diabetes in patients more accurately. K-fold cross-validation is applied, considering k to be 2, 4, 5 and 10. The performance parameters taken are the: accuracy, precision, recall, F Score and area under the curve. Our study found that the MLP classifier gave the highest accuracy of 78.7% with a recall of 61.26%, precision of 72.45% and F_1 Score of 65.97% for k = 4.

Keywords: ***Classification Algorithms; Diabetes Prediction; Prediction; Feature Selection; Machine Learning; Neural Networks; Multi-layer Perceptron; MLP***

1. Introduction

Diabetes is one of the most chronic diseases in the world in which the sugar level of blood becomes too high [1]. It has become a fifth-ranked disease for disease related deaths [2]. Due to diabetes, other problems may arise like the increased risk of heart attack and stroke [3]. Unfortunately, these diseases cannot be cured, the only way is to manage the glucose level in the blood. Around 8.8% of adults were diabetic in 2017 around the world and the projected value is 9.9% by 2045 [4]. The diabetic disease is classified into three categories (i) type I (ii) type II and (iii) type III. Most of the people having diabetes are of type II [5]. The Pima is one of the most studied

Ravinder Ahuja, Subhash C. Sharma and Maaruf Ali, "A Diabetic Disease Prediction Model Based on Classification Algorithms", Annals of Emerging Technologies in Computing (AETiC), Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 44-52, Vol. 3, No. 3, 1st July 2019, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2019.03.005, Available: http://aetic.theiaer.org/archive/v3/v3n3/p5.html.

populations for diabetic analysis around the world [6]. Most of the Pima population is type-2 diabetic. Diabetic classification is a challenging issue due to nonlinear and complex data. Some other reasons are that the Pima dataset has null entries and outliers, which causes low performance of machine learning algorithms. Machine learning algorithms are presently being used in almost all fields like finances, marketing, medical, business, etc. Machine learning algorithms are of three types (i) supervised learning (ii) unsupervised and (iii) semi-supervised. In supervised learning, labeled data is available and the machine learns from some part of it and applies the learning to the unseen data. In the unsupervised machine learning algorithm, data is not labeled, the algorithms find an interesting relationship between the data points. Semi-supervised is a combination of supervised and unsupervised algorithms. With respect to diabetic prediction, we have applied various supervised machine learning algorithms as the dataset of Pima is labeled. The classifiers cannot correctly classify, due to the presence of missing values and outliers present [7-11]. In mathematics, outlier and missing value handling is an important issue which cannot be ignored. The dataset considered goes through two phases before applying the feature selection techniques. In the first phase, the missing values of the dataset are replaced by the median and the outliers are detected by using the Inter Quartile Range (IQR) and replaced by the group median. Further feature selection techniques have been applied and fifteen classification algorithms have been applied to classify the diabetic dataset. A major contribution is the application of the K-Fold cross-validation technique by considering the different value of k to be 2, 4, 5 and 10 and the application of the five classification algorithms.

The paper is divided into the following sections: section two contains the related work, section three describes the methods and materials which include a description of the dataset, feature selection techniques used and classification algorithms applied, section four contains the overall methodology and performance parameters used, section five describes the experimental setup, section six the results followed finally by section seven which contains the conclusion and future scope.

2. Related Work

Lots of research work is present in the literature related to diabetes classification and diagnosis. In paper [12] authored by Karthikeyani, V. *et al*., a support vector machine with a Radial Basis Function (RBF) kernel was used whereby the zero values were replaced by entries with the mean achieving an accuracy of 74.80%. The same authors in [13] applied the Partial Least Square (PLS) technique for feature extraction and extracted three features out of eight features. They then applied Linear Discriminant Analysis (LDA) for classification and achieved an accuracy of 74.40%. A different approach taken by Kumari [14] was to delete the zero value entries, reducing the remaining entries to 460 from 768. Out of these 460 entries, 200 were used for training and the rest for testing, achieving an accuracy of 75.5%. Parashar *et al*. [15] have applied Linear Discriminant Analysis (LDA) for feature selection and selected two features out of eight and applied a Support Vector Machine (SVM) with a Feed Forward Neural Network to classify the diabetic dataset and achieved an accuracy of 75.65%. Bozkurt, M. R. *et al*. [16] applied Automatic Identification System (AIS) and Artificial Neural Network (ANN) methods of classification for the diabetic dataset and achieved an accuracy of 76% using their ANN algorithm. Iyer, A. *et al*. [17] applied the correlation-based feature selection technique to select two features out of eight and applied the Naïve Bayes and Decision Tree algorithm to classify diabetics. They also inserted the missing values with the average and reported an accuracy of 74.79%. Kumar Dewangan, A. *et al*. [18] applied a Multi-Layer Perceptron (MLP) and Bayes Net classifier and reported an output of 81.19% accuracy. [19] applied the J48 classification algorithm and achieved an accuracy of 76.58%. [20] applied four classification algorithms, these being the J48, Naïve Bayes, Logistic Regression and Random Forest, obtaining an accuracy of 80.43%. Maniruzzaman *et al*. [21] applied LDA, Quadratic Discriminant Analysus (QDA), Gaussian Process Classifier (GPC), and Naïve Bayes classification algorithms for diabetic patient classification and found that GPC gave the highest accuracy of approximately 82% using the radial basis kernel. Bashir, S. [22] brought in a Hierarchical Multi-level classifier with multi-objective voting technique (HM-Bag) for classification and compared their technique with the Naïve Bayes, Support Vector Machine, Logistic Regression, Quadratic Discriminant Analysis, K Nearest Neighbour, Random Forest and Artificial Neural

Network classifiers. They found that HM-Bag gave the highest accuracy of 77.21%. Deepti Sisodia *et al*. [23] applied SVM, Naïve Bayes and Decision Tree algorithms and found that Naïve Bayes gave the highest accuracy of 76.30%. Aishwarya, R. *et al*. [24] used Principal Component Analysis (PCA) for preprocessing and SVM for classification and reported an accuracy of 95%. Maniruzzaman, M. *et al*. [25] firstly replaced the zero entries with the median values and the outliers were detected using the Inter Quartile Range (IQR) method. If the outliers were detected they were then replaced with the median values. Six feature selection techniques, consisting of the Principal Component Analysis (PCA), Logistic Regression, mutual information, analysis of variance and the Fisher Discriminant Ratio (FDR) were applied in combination with ten classification algorithms (Random Forest, Linear Discriminant Analysis, Gaussian Process, Naïve Bayes,, Quadratic Discriminant Analysis Classifier, Artificial Neural Network, Support Vector Machine, Decision Tree, Logistic Regression and AdaBoost). They found that the Random Forest classification with Random Forest feature selection gave the highest accuracy of 92.26%.

3. Methods and Materials

3.1. Dataset Description

The original donor of the data set is the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset was from the website of the University of California, Irvine (UCI) [26]. The data set consists of 768 records (all women) out of which 500 are not diabetic and 268 are diabetic. There are eight attributes in the dataset as shown in Table 1. This dataset does contain zero values corresponding to the: (i) Glucose attribute in five records (ii) 35 records in the blood pressure (iii) 27 records in the BMI (Body Mass Index) attribute (iv) 227 records in the Skin Thickness attribute and (v) 374 records in the Insulin attribute. These zero values do not have significance so were replaced by the median of the corresponding attribute. The outliers are detected by using the IQR (Inter-Quartile Range) and the outliers that were found are replaced with the median value.

3.2. Feature Selection Techniques

Feature Selection Technique (FST) always boosts the classification accuracy and minimizes the computational cost. FST also eliminates the less important features and reduces the time complexity of the machine learning technique. The feature selection techniques that are used are as follows.

3.2.1. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a supervised technique which is used for extracting the important features from the dataset. It is used to avoid overfitting of the data as well as to minimize the computational costs. This is achieved by projecting a feature space onto a smaller lower dimensional space having optimal class separability. In LDA, more emphasis is given to the axes that are responsible for maximizing the partition amongst the multiple classes [27].

3.3. Classification Algorithms

3.3.1. Support Vector Classifier (SVC)

The Support Vector Machine (SVM) is a supervised machine learning technique that is useful in classification as well as in dealing with regression problems. It is a technique that best separates the two classes through a hyperplane or a line. It works on the assumption that the support vectors alone are important whereas other training samples can be ignored. This classifier is very effective in high dimensional spaces [28]. Furthermore, the radial basis function (RBF) kernel was utilised in the experiments.

3.3.2. MLP (Multi-layer Perceptron) Classifier

The concept of the multilayer perceptron is inspired by the human nervous system [29]. The advantages of MLP are that it is: (i) Highly fault tolerant, i.e. in case of failure of neurons and interconnections between them, they keep on working (ii) It is nonlinear in nature, by this it is suitable for all kinds of real-world problems. We have used 100 hidden layers in our experiments, the activation function is ReLU (Rectified Linear Unit) and the learning rate is 0.01.

3.3.3. Logistic Regression (LR)

It is a linear model for classification not for regression [30]. It applies the maximum likelihood estimation of the dependent variable. The main benefit of logistic regression is that it can handle nonlinear data and is robust. Let us consider that there are n number of features A_1, A_2, … A_n and let p be the probability of the event occurrence and $(1 - p)$ is the probability of the event not occurring. Then the model is given by:

$$Log\left(\frac{p}{1-p}\right) = logit(p) = \beta_0 + \beta_1 A_1 +, \ldots + \beta_n A_n \quad (1)$$

where β_i is the regression coefficients.

3.3.4. Decision Tree (DT)

The Decision Tree "is used for decision analysis. [In] Decision Trees, where target values can take continuous values are known as the regression trees. Considering the tree, [the] input values are represented [as] a path from the root to the leaves, [where] each leaf represents the target variable" [31]. The steps for the DT consists of: (i) "Form a tree with its nodes as features [(ii)] Select [one] feature to predict the output from the input, where the root node is that which contains the highest information gain [(iii)] Repeat the above steps to form subtrees based on features which are not used in the above nodes" [31].

3.3.5. Random Forest (RF)

This algorithm considers numerous decision trees, thus forming a forest. It is also called an ensemble of decision tree algorithms [32]. The building of the:

> "random tree begins at the top of the tree with [the] in-bag dataset. The first step involves selecting a feature at the root node and then splitting the training data into subsets for every possible value of the feature. This makes a branch for each possible value of the attribute. Tree design requires choosing a suitable attribute selection measure for splitting and the selection of the root node to maximize dissimilarity between [the] classes. [If] the information gain [is positive]; the node is split else the node will become a leaf node that would provide a decision of the most common target class in the training subset" [33].

In our experiment, we have used 100 decision trees and the Gini Index for the impurity index. The steps for the Random Forest are as follows:

Step 1: From a total of n features, randomly m features are selected, $m << n$;

Step 2: A node d, which belongs to the set of m nodes, is calculated using the best split point;

"Step 3: Further, d is split into daughter nodes using the best split method;

[Step 4:] Repeat Steps 1-3 until a tree is formed with a root node and having the target as the leaf node;

[Step 5:] Steps 1-4 represent the creation of a tree. Repeat them the number of times to create a forest" [31].

4. Proposed Approach

Data pre-processing techniques are applied first, i.e. the missing values are identified and replaced by the group median. After this outlier is detected using the IQR (Inter-Quartile Range) method, it is then replaced by the group median. Further LDA feature selection techniques are applied to extract the important features from the processed data and five classification algorithms (RF, LR, DT, MLP and SVC) applied to predict the diabetic patients. K-fold cross-validation is also applied with the value of k to be 2, 4, 5 and 10. The overall methodology used is shown in Figure 1, below.

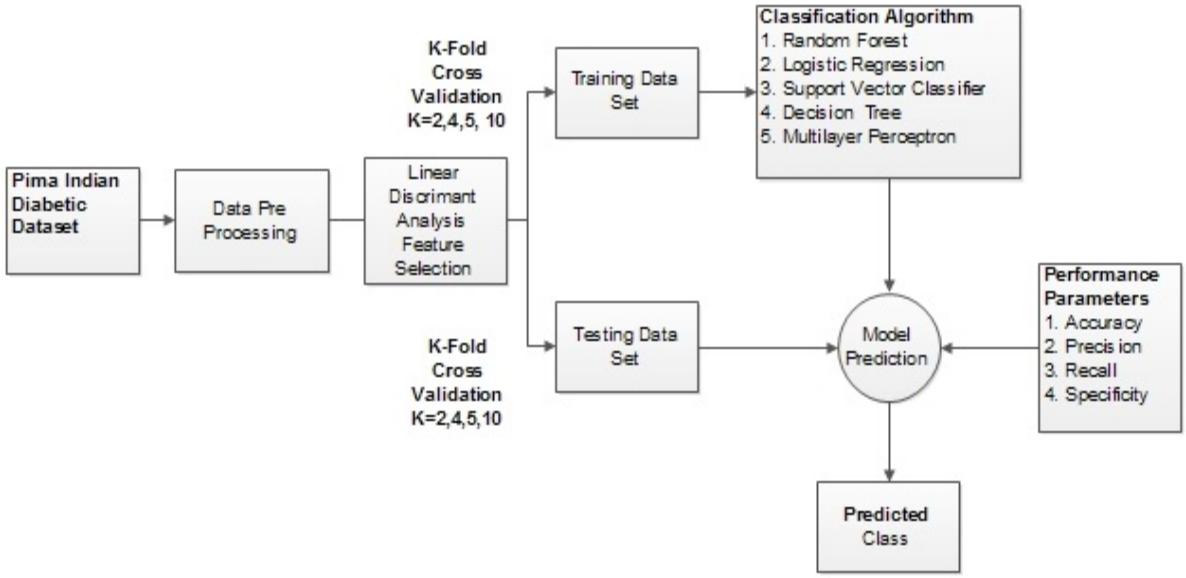

Figure 1. Overview of Methodology Used.

5. Experimental Setup

5.1. Performance Parameters

Four evaluation parameters are taken into consideration which are as follows:

5.1.1. Accuracy

It is the basis of measuring the quality of any predictive model. The accuracy measures the ratio of correct predictions to the total number of data points evaluated. This paper consists of the best accuracies that were obtained by various machine learning models after applying the feature selection and K-Fold techniques. Equation (2) gives the equation for the accuracy.

$$Accuracy = \frac{True\ Positive\ + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True\ Negative} \quad (2)$$

5.1.2. Precision

The Precision of a model is the fraction of relevant occurrences among the retrieved occurrences. It is also referred to as a positive predictive value. It is calculated by taking the ratio of true positives by the total positives in a model. In simple words, a high precision means that the algorithm returns more relevant results than the irrelevant ones. Equation (3) gives the equation for the accuracy.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

5.1.3. Recall

The Recall is also known as the sensitivity of the model. It is the fraction of relevant occurrences that have been retrieved over the total amount of relevant occurrences. A high recall means that most of the occurrences returned were relevant. It is measured as the ratio of true positives to the summation of true positives and false negatives, given in (4):

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

5.1.4. F-Score

The F-Score is a measure that combines the precision and recall by taking its harmonic mean. It is approximately the average of the two when they are close, else their harmonic means. The harmonic mean is the ratio of the square of the geometric mean divided by the arithmetic mean. In F_1 measure, both the precision and recall are equally weighted, as defined in (5):

$$F_1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (5)$$

5.2. K-Fold Cross Validation

In K-fold cross-validation method, the original data set of 768 patients is partitioned into equal-sized sub-segments [31]. The number of segments depends upon the value of *k* taken, in our case we have taken *k* to be 2, 4, 5, or 10. Out of all the sub-segments, one partition was used as the "testing data and remaining nine [was used for the] training data. This cross-validation technique is repeated [*k*] times, where each sub-partition is taken as [the] testing data at least once. These results obtained from the above repetitions are averaged or otherwise combined to produce a single estimation. The advantage of using this validation [strategy] is that every single data is used for [the] training as well as [in] testing the model and each entry in the dataset is used for validation of the result at least once" [31]. This helps to increase the accuracy of the model.

6. Results

The diabetic dataset from the UCI repository was utilised. Further outliers and missing values were replaced with the median values. The linear discriminant analysis (LDA) feature selection technique was applied in order to extract the important features from the pre-processed dataset. Five classification algorithms were used, which are the: Support Vector Classifier (SVC), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and the Multilayer Perceptron (MLP) on the LDA processed data for *k* values of 2, 4, 5 and 10 for K-Fold cross-validation. The performance parameters considered are the precision, recall, F_1 Score and accuracy. K-fold cross-validation is applied if the dataset is small, as in our case since only 768 patient records exist. The results of the five classification algorithms based on accuracy are presented in Table 1.

Table 1. Comparison of classifiers using **accuracy** with different k-fold values.

k-fold	Support Vector Classifier	Decision Tree	Random Forest	Logistic Regression	Multi-Layer Perceptron
K = 2	77.6	69	69.9	77.8	77.5
K = 4	77.6	69.9	70	77.6	78.7
K = 5	77.5	71.5	72.9	77.6	78.2
K = 10	77.5	69.5	70	77.6	77.6

As shown in Table 1 the highest accuracy 78.7% is achieved in the case of multilayer perceptron. Table 2 shows the recall performance recorded with different values of K (2, 4, 5, and 10). The highest recall value of 61.26% is achieved in the case of multilayer perceptron.

Table 2. Comparison of classifiers using **Recall** with different *k*-fold values.

k-fold	Support Vector Classifier	Decision Tree	Random Forest	Logistic Regression	Multi-Layer Perceptron
K = 2	57.75	59.14	60.08	58.21	59.15
K = 4	57.6	59.15	56.81	57.7	61.26
K = 5	55.86	58.5	57.68	56.86	60.36
K = 10	57	56.42	56.37	57.25	60.92

Table 3 shows the precision performance of the five classification algorithms considering different values of k to be 2, 4, 5, and 10. The highest value of precision of 72.45% is achieved in the case of using the multilayer perceptron classifier for a value of $k = 4$.

Table 3. Comparison of classifiers using **Precision** with different *k*-fold values.

k-fold	Support Vector Classifier	Decision Tree	Random Forest	Logistic Regression	Multi-Layer Perceptron
K = 2	70.43	55.14	55.72	72.53	70.33
K = 4	69.11	56.05	59.02	72.12	72.45
K = 5	71.35	58.92	61.33	71.52	69.27
K = 10	70.48	56.39	57.51	71.02	71.1

Table 4, show the the F_1 Score performance of five classification algorithm considering different values of *k* to be 2, 4, 5, and 10. The highest value of F_1 Score 65.97% is achieved in the case of using the multilayer perceptron classifier for a value of $k = 4$.

Table 4. F_1 Score performance of five classification algorithms with different *k*-fold values.

k-fold	Support Vector Classifier	Decision Tree	Random Forest	Logistic Regression	Multi-Layer Perceptron
K = 2	61.62	57.02	55.47	64.59	65.15
K = 4	60.85	57.51	59.76	64.09	65.97
K = 5	59.36	58.54	58.91	63.24	63.94
K = 10	63.94	55.91	56.18	63.52	65.58

From all the four tables, it was observed that the multilayer perceptron algorithm is performing better amongst all the other algorithms for $k = 4$.

7. Conclusion

The research was undertaken using the dataset from the UCI repository. All zero-valued entries and outliers were replaced with the group median values. Further, a Linear Discriminant feature selection technique was applied to select the best features. On the optimal features selected, five classification algorithms were applied (SVC, RF, DT, MLP, and LR) along with *k*-fold (K =2, 4, 5 and 10) cross-validation. This enabled extensive data analysis to be performed to conclusively determine the optimal result. This was found by the result with the highest accuracy of 78.7% that was achieved by using the Multilayer Perceptron classifier with $k = 4$ in *k*-fold cross-validation. From the results, it is also observed that the Multilayer Perceptron was performing better compared to all the other classification algorithms in term of all the other performance parameters like precision, recall and the F_1 Score. Future work will include a study with other classifiers and feature selection techniques for further analysis of the dataset.

References

[1] P. Muntner, L.D. Colantonio, M. Cushman, D.C. Goff, G. Howard, V.J. Howard and M.M. Safford, "Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations", *JAMA* 311(14):1406–1415, 2014.

[2] B.A. Hamburg and G.E. Inoff, "Relationships between behavioral factors and diabetic control in children and adolescents: A camp study", *Psychosomatic Medicine*, *44*(4), 321-339, 1982.

[3] American Diabetes Association, "Diagnosis, and classification of diabetes mellitus", Diabetes Care 37 (Supplement 1): S81–S90, 2014.

[4] C. Fitzmaurice, C. Allen, R.M. Barber, L. Barregard, Z.A. Bhutta, H. Brenner and T. Fleming, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study", *JAMA Oncol*. 3(4):524–548, 2017.

[5] Y. Shi and F.B. Hu, "The global implications of diabetes and cancer", *Lancet* 9933(383):1947–1948, 2014.

[6] W.C. Knowler, D.J. Pettitt, M.F. Saad and P.H. Bennett, "Diabetes mellitus in the Pima Indians: incidence, risk factors, and pathogenesis", *Diabetes/metabolism Reviews* 6, no. 1, 1-27, 1990.

[7] S. Bashir, U. Qamar and F.H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework", *J. Biomed. Inform*. 59:185–200, 2016.

[8] N.A. Zainuri, A.A. Jemai and N.A. Muda, "A Comparison of various imputation methods for missing values in air quality data", *Sains Malays*, 44(3):449–456, 2015.

[9] J. Kaiser, "Dealing with missing values in data", *J. Syst. Integr*. 5(1): 42–43, 2014.

[10] C. Leys, C. Ley, O. Klein, P. Bernard and L. Licata, "Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median", *J. Exp. Soc. Psychol*. 49(4):764– 766, 2013.

[11] M.R. Baneshi and A.R. Talei, "Does the missing data imputation method affect the composition and performance of prognostic models?", *Iran Red Crescent Med. J*. 14(1):30–31, 2012.

[12] V. Karthikeyani, I.P. Begum, K. Tajudin and I.S. Begam "Comparative of data mining classification algorithm in diabetes disease prediction", *Int. J. Comput. Appl*. 60(12):26–31, 2012.

[13] V. Karthikeyani and I.P. Begum, "Comparison performance of data mining algorithms in the prediction of diabetes disease", *Int. J. Comput. Sci. Eng*. 5(3):205–210, 2013.

[14] V.A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine", *Int. J. Eng. Res. Appl*. 3(2):1797–1801, 2013.

[15] A. Parashar, K. Burse and K. Rawat, "A Comparative approach for Pima Indians diabetes diagnosis using lda-support vector machine and feed-forward neural network", *Int. J. Adv. Res. Comput. Sci. Softw. Eng*. 4(4):378–383, 2014.

[16] M.R. Bozkurt, N. Yurtay, Z. Yilmaz and C. Sertkaya, "Comparison of different methods for determining diabetes", *Turk. J. Electr. Eng. Comput. Sci*. 22(4):1044–1055, 2014.

[17] A. Iyer, S. Jeyalatha and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques", *Int. J. Data Min. Knowl. Manag. Process.* 5(1):1–14, 2015.

[18] A.K. Dewangan and P. Agrawal, "Classification of diabetes mellitus using machine learning techniques", *Int. J. Eng. Appl. Sci*. 2(5):145–148, 2015.

[19] R. Sivanesan and K.D.R. Divya, "A Review on diabetes mellitus diagnoses using classification on Pima Indian diabetes data set", *Int. J. Adv. Res. Comput. Sci. Manag. Stud*. 5(1):12–17, 2017.

[20] M. Nabi, A. Wahid and P. Kumar, "Performance analysis of classification algorithms in predicting diabetes", *Int. J. Adv. Res. Comput. Sci*. 8(3):456–461, 2017.

[21] M. Maniruzzaman, N. Kumar, M.M. Abedin, M.S. Islam, H.S. Suri, A.S. El-Baz and J.S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm", *Comput. Methods Prog. Biomed*. 152:23–34, 2017.

[22] S. Bashir, U. Qamar and F.H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework", *J. Biomed. Inform*. 59:185–200, 2016.

[23] D. Sisodia and D.S. Sisodia, "Prediction of Diabetes using Classification Algorithms", *Procedia Computer Science*, 132: 1578-1585, 2018..

[24] R. Aishwarya, P. Gayathri and N. Jaisankar, "A Method for Classification Using Machine Learning Technique for Diabetes", Int. J. of Eng. And Tech., 5(3):2903-2908, June (2013).

[25] M. Maniruzzaman, M.J. Rahman, M. Al-MehediHasan, H.S. Suri, M.M. Abedin, A. El-Baz and J.S. Suri, J.S., "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers", *Journal of Medical Systems*, *42*(5), p.92, 2018.

[26] A. Frank and A. Asuncion, (2010). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[27] F. Song, D. Mei and H. Li, "Feature Selection Based on Linear Discriminant Analysis", *2010 Int. Conf. on Intelligent System Design and Engineering Application*, Changsha, pp. 746-749, 2010.

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

[29] T.H. Reinhardt, "Using neural networks for prediction of the sub cellular location of proteins", *Nucleic Acids Res.* 26(9):2230–2236, 1998.

[30] B. Tabaei and W.A. Herman, "Multivariate logistic regression equation to screen for diabetes", *Diabetes Care* 25:1999–2003, 2002.

[31] R. Ahuja, V. Vivek, M. Chandna, S. Virmani and A. Banga, "Comparative Study of Various Machine Learning Algorithms for Prediction of Insomnia", *Advanced Classification Techniques for Healthcare Analysis*, ed. Chinmay Chakraborty, 234-257 (2019), accessed July 01, 2019. DOI:10.4018/978-1-5225-7796-6.ch011

[32] L. Breiman, "Random Forests", *Mach. Learn*. 45(1):5–32, 2001.

[33] W. Almayyan, "Lymph Diseases Prediction Using Random Forest and Particle Swarm Optimization", *J. of Intelligent Learning Systems and Applications*, Vol. 8, No.3:51-62 (2016). DOI: 10.4236/jilsa.2016.83005

© 2019 by the author(s). Published by Annals of Emerging Technologies in Computing (AETiC), under the terms and conditions of the Creative Commons Attribution (CC BY) license which can be accessed at http://creativecommons.org/licenses/by/4.0/.