DIABETES PREDICTION ANALYSIS USING FEATURE SELECTION

[1]GURWINDER KAUR BAJWA, [2]PROF. ANIL SAGAR, [3]PROF. BALJINDER SINGH

[1]M.Tech. CSE, [2,3]Assistant Professor
Beant College of Engineering & Technology, Gurdaspur

***ABSTRACT*: Diabetes has influenced more than 246 million individuals worldwide with a dominant part of them being ladies. Recognition of diabetes in its beginning periods is the key for treatment. In this exploration work, number of choice trees is joined for the investigation procedure. This proposed research work also analyse the current computational insight strategies for anticipating diabetes. The dataset utilized in this examination work is gathered from National Institute of Diabetes and Digestive and Kidney Diseases and depends on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases. Proposed technique is compared with ANN, SVM, KNN, naïve bayes and logistic regression algorithm. Proposed technique gives more accuracy, precision, recall, and f-measure and less errors as compared to existing algorithms and hence performs better.**

Introduction

The real test looked by the different medicinal services associations like high innovation healing facilities and numerous restorative focuses is the conveyance of standard administrations at less expensive costs which can be managed by ever person. The choices made in logical condition that supply optimistic outcomes are fully is dependent healthcare professional's notion and proficient know-how reward in databases based on clinic. Some different types of choices are always taken that may now not a just right indifferent resolution and can result in disastrous outcomes that are not tolerable.

Any specialist can utilize shrouded data to care for his patient. Verdict of confirmation from learning extricate from clinical databases is incredible employment in front medicinal people. EBM utilizes an information mining innovation that creates it conceivable to consequently investigate colossal clinical Databases and to find designs behind them [1]. The combination of confirmation based solution rules into clinical choice emotionally supportive networks would both enhance quality and lessen expenses of care, by suggesting rules for just the most proficient medicines and drugs. Inner clinical involvement in coordination with outer clinical aptitude must be open to human services authorities at fitting time & in proper way.

Data warehousing & Data Mining offers an extensive help for social affair, dissecting and exhibiting therapeutic information. Clinical choices are frequently made utilizing the specialist's remedy and involvement in field instead of learning base which is rich in information covered up in database.

The objective of this paper is to study and analyze the existing diabetic prediction analysis techniques. It is also used to propose a hybrid feature selection with classification model comprising of Kmeans and Logistic Regression classifier (Cluster-Bagged Logistic Regression). It would also compare and analyse the performance of the proposed computational intelligent technique with the base techniques based on accuracy, precision, recall, root means squared error.

Rest of the paper is organised as follows. Overview of image forgery detection is presented in the first section, in second section image forgery types are discussed, in third section image forgery detection mechanisms are discussed and in forth section literature and comparison of various image forgery detection mechanisms is presented. Last section gives the conclusion of this paper.

1. PROBLEM DEFINATION

The inexhaustible measure of data concealed in clinical databases that can be capably use in finding of patient's diseases. In the existing paper, diabetic's patients have been analyzed using various machine learning algorithms including naïve Bayes, neural network, support vector machine, k nearest neighbor KNN, decision tree and Logistic Regression. Among all these, Logistic Regression performs better than all other algorithms.

2. Proposed Methodology

Logistic regression attempts to predict outcomes based on a set of independent variables, but logit models are vulnerable to overconfidence. Also, it doesn't perform well when feature space is too large. To overcome these disadvantages, A Hybrid Clustering with bagged classification model is proposed technique comprising of Kmeans (Clustering) technique with Bagged Logistic Regression (classification) technique which is used to mine the data to extract the useful patterns and to improve the accuracy of the classifier than the existing hybrid technique The benefits of Feature-boosted regression to the fact that the derived features of patient types are clustered the multivariate feature space into subspaces effectively, capturing important statistical relationships and differences that boosted the regression results.

2.1 Dataset collection: The dataset used in this research work is collected from National Institute of Diabetes and Digestive and Kidney Diseases and is based on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases. The Pima Indian diabetes database is a collection of medical diagnostic reports of 768 patients.

2.2 Pre-processing: The collected raw data is then pre-processed and formatted. If some missing values are there, it will handle all the missing values by either replacing those values by the mean/average of all the values or by removing.

2.3 Proposed Technique: The proposed technique comprises of feature selection with clustering and classification i.e. clustered data is given to the classification for the evaluating the mining patterns. For the clustering of data, most commonly used algorithm is K-means.

2.4 Gini Index based Feature Selection : It is used to as a splitting method. In this algorithm we gather data sets for testing. Let A be the data sample, m be the divided no of subsets, P be the probability and Ci be the different classes then

GiniIndex(A)=1-∑_(i=1)^m▒〖P_i〗^2

At the point when the base of GiniIndex (A) is zero then it means that all the records have a place with a similar classification at this gathering; it demonstrates that the greatest helpful data can be acquired. Then at the point when every one of the examples of accumulation have a standard circulation to a specific classification, GiniIndex(A) achieves greatest, demonstrating the base valuable data got.for gini index the little contaminating influence is the higher is the quality. On the other hand,

GiniIndex(A)=∑_(i=1)^m▒〖P_i〗^2

measuring the contaminating influence of characteristic classify method, the greater is the contaminating influence the higher is the quality of attribute.

This optimized dataset is given as an input to Logistic Regression classifier to find the useful patterns. According to the existing techniques surveyed, when K-means is combined Logistic Regression it will give better.

K-means is applied on the input dataset by finding the Euclidean distance of each data point from the centroid and clusters are defined. If the distance of centroid of the present nearest cluster is less than or equal to the previous distance, then the data point remains in that cluster and there is no need to find its distance from other cluster centroids.

Then classify that optimized dataset using Bagged Logistic Regression classifier.

Bagging is a technique which decreases the variance of the prediction using dataset using combinations with repetitions to produce multi-sets of same size of the dataset. For each multi set the Logistic Regression learning algorithm is applied to classify the instances and a model is created and a vote related to that model is generated. The average of all the predicted votes is considered to be the result of the classifier. In this algorithm dataset is sampled with replacement into ten datasets with same number of tuples using bagging. And then for every bootstrap sample Logistic Regression classification is used as a base classifier and returns prediction results. At the end the final prediction is produced using average voting.

Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's).

3. Result Analysis:

3.1 Result Analysis

This area exhibits the reproduction consequences of the work implemented and the proposed approach.

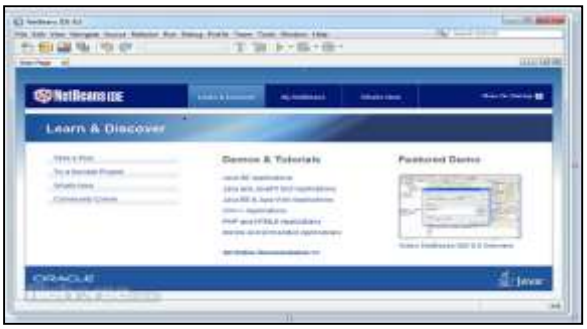


Fig 3.1 Netbeans IDE

3.2 Selection of Dataset

Fig 3.2 Choosing the dataset for diabetes patients

This is the first window we come across,here we select the files that our dataset have.that are preloaded when we create database at the backend.We have created diabetes. arff file .Here's it contains our emails dataset that we browse through,we upload them as such.here we upload from the main server thing. This interface helps us to choose the desired data set from any location and upload that data set. After the loading of the dataset the next step is to perform filtration.

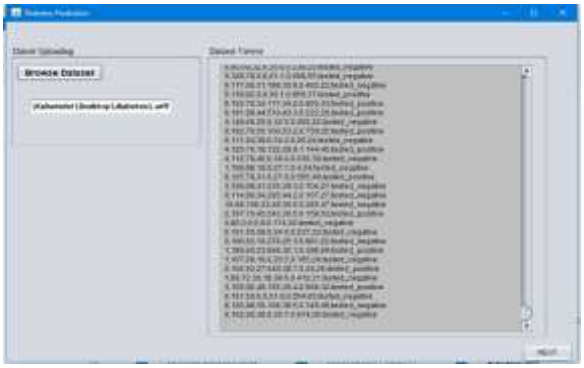


Fig 3.3 dataset selected

3.3 Randomizing the Dataset

The figure above shows the results of randomize filter. This filter randomly shuffles the order of instances passed through it.

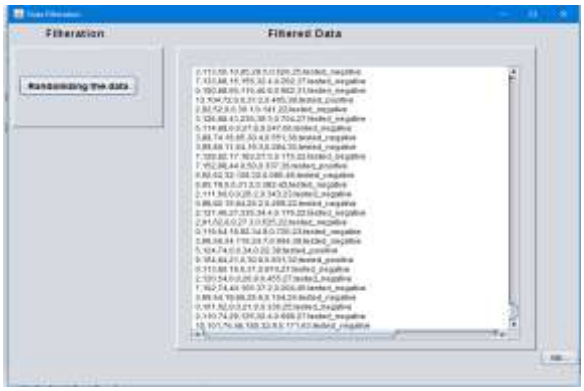


Fig 3.4 Filtering the Dataset using Randomizing dataset.

3.4 Results of ANN algorithm

The figure above shows the classification results of ANN algorithm. The results show the accuracy of 75.2604% i.e. 578 instances are correctly classified out of 768 instances. The kappa statistics for ANN algorithm is 0.438 which denotes the enhancement in the algorithm its value should lie between 0 and 1 closer to 1 means algorithms performs better. Class details parameters are also shown like precision which is 0.747, recall 0.7573, F Measure 0.748, TP Rate 0.753 and FP rate 0.328.

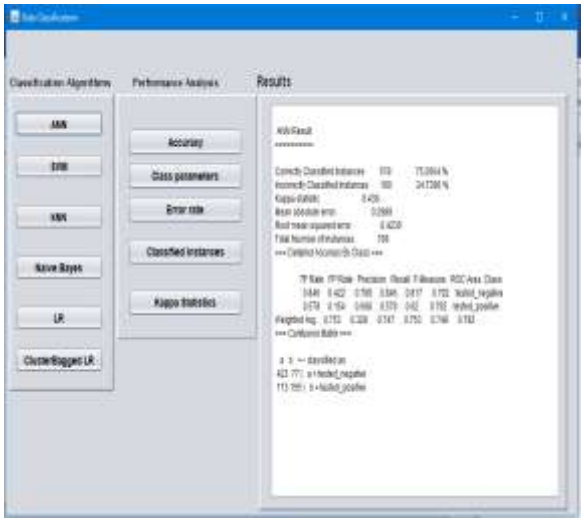


Fig 3.4 Showing the results of ANN classification algorithm

3.5 Results of SVM Algorithm

The figure above shows the classification results of decision tree algorithm. The results show the accuracy of 65.1042% i.e. 500 instances are correctly classified out of 768 instances. Class details parameters are also shown like precision which is 0.424, recall 0.651, F Measure 0.513, TP Rate 0.651 and FP rate 0.651. The mean absolute error and root mean squared error in this case is 0.349 and 0.5907.

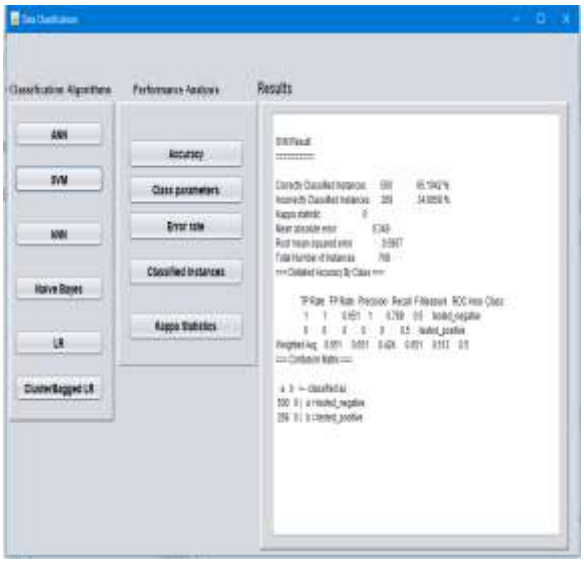


Fig 3.5 Showing the results of SVM classification algorithm

3.6 Results of KNN Algorithm

The figure above shows the classification results of KNN algorithm. The results show the accuracy of 70.3125% i.e. 540 instances are correctly classified out of 768 instances. Class details parameters are also shown like precision which is 0.697, recall 0.703, F Measure 0.699, TP Rate 0.703 and FP rate 0.379.

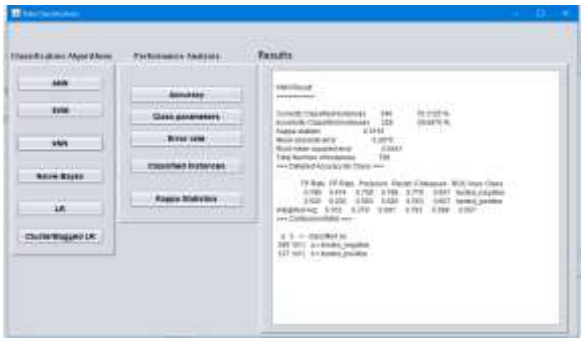


Figure 3.7: Showing the results of KNN classification algorithm

3.7 Results of naïve bayes

The figure above shows the classification results of naïve bayes classification algorithm. The results show the accuracy of 75.9115% i.e. 583 instances are correctly classified out of 768 instances. Class details parameters are also shown like precision which is 0.755, recall 0.759, F Measure 0.756, TP Rate 0.759 and FP rate 0.311.

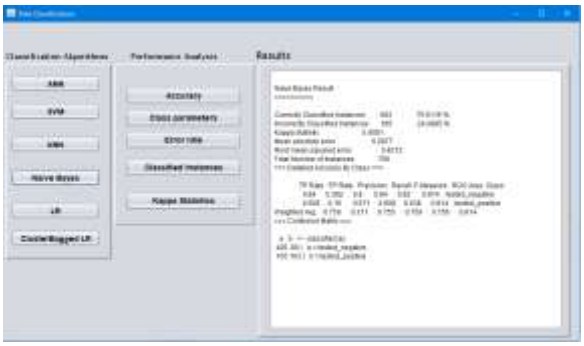


Fig 3.8: Showing the results of Naïve Bayes classification algorithm.

3.8 Results of Logistic Regression classification algorithm

The figure above shows the classification results of logistic regression classification algorithm. The results show the accuracy of 77.7944% i.e. 597 instances are correctly classified out of 768 instances. Class details parameters are also shown like precision which is 0.772, recall 0.777, F Measure 0.77, TP Rate 0.777 and FP rate 0.317.

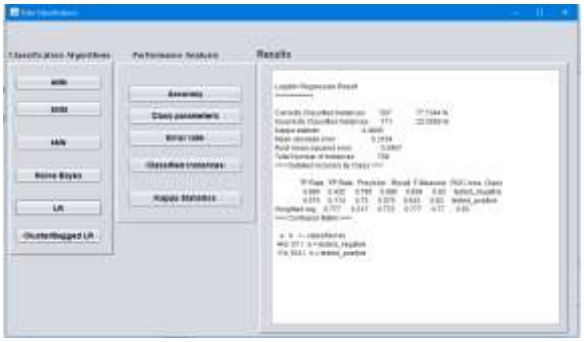


Fig 3.9 Showing the results of Logistic Regression classification algorithm.

3.9 The results of proposed algorithm

The figure above shows the classification results of proposed algorithm. The results show the accuracy of 99.2188% i.e. 762 instances are correctly classified out of 768 instances. Class details parameters are also shown like precision which is 0.992, recall 0.992, F Measure 0.992, TP Rate 0.992 and FP rate 0.01.

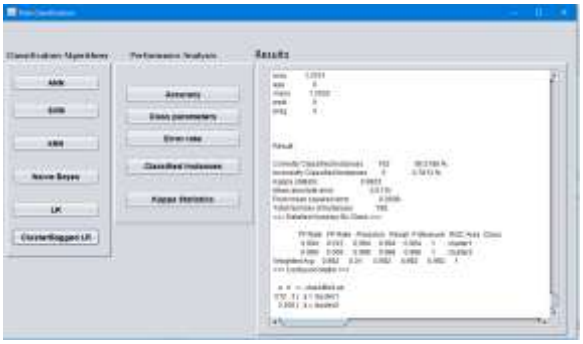
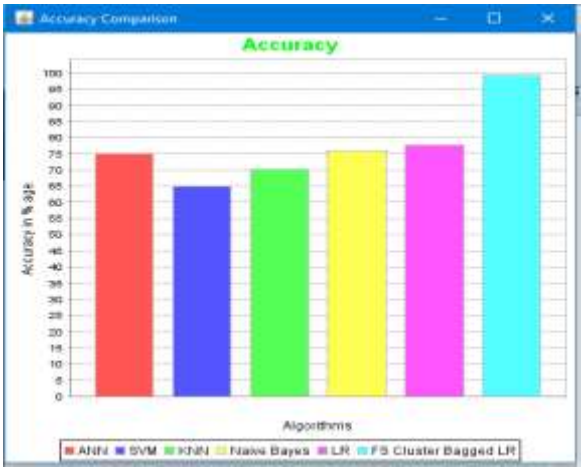


Fig 3.10 Showing the results of proposed algorithm.

3.10 Evaluation Parameters

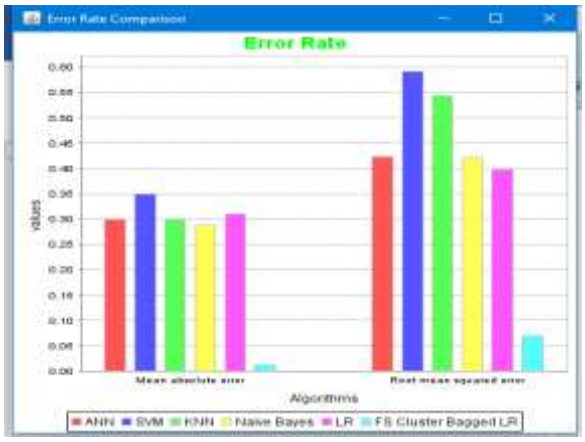
The proposed project was being evaluated on following parameters:

The following graph shows the accuracy comparison of existing with the proposed classification algorithm.



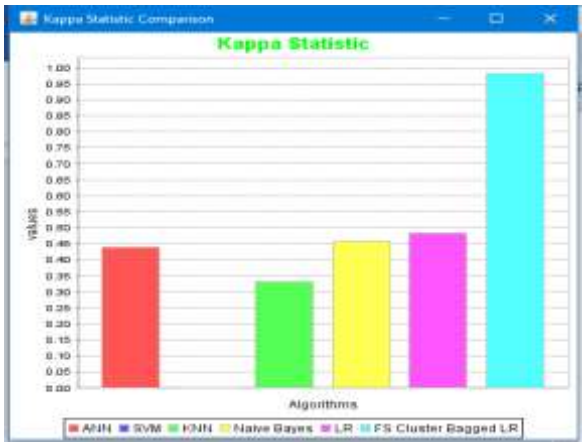
The following graph shows the class parameters comparison of existing with the proposed classification algorithm:

The following graph shows the error rate comparison of existing with the proposed classification algorithm:



The following graph shows the Classified Instances Comparison:

The following graph shows the kappa statistic Comparison:



Conclusion

This research work implements a hybrid feature selection clustering with classification model comprising of enhanced k-means and Logistic Regression classifier (cluster-bagged Logistic Regression). The proposed technique comprises selecting the features first then clustering with classification i.e. clustered data is given to the classification for the evaluating the mining patterns. This research work also analyse the current computational insight strategies for anticipating diabetes. The dataset utilized in this examination work is gathered from National Institute of Diabetes and Digestive and Kidney Diseases and depends on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases.The Pima Indian diabetes database is a collection of 768 diabetic patients analyzed from the medical diagnostic reports. Results are evaluated on the basis of accuracy, precision, recall, root means squared error and to compare the performance of the proposed technique with the existing techniques. Proposed technique gives accuracy of 99.2188. Proposed technique is compared with ANN, SVM, KNN, naïve bayes and logistic regression algorithm. Proposed technique gives more accuracy, precision, recall, and f-measure and less errors as compared to existing algorithms and hence performs better.

Future scope:

This research work focuses on the classification of diabetes using data mining technique. Similarly in future, we can use data mining technique on any other medical condition. Here, we perform clustering on classification. In future, feature selection may be combined with classification so that clustering time is reduced as feature selection optimizes the features and selects the best features.

REFERENCES

[1] Candice MacDougall, Jennifer Percival and Carolyn McGregor (2009), "Integrating Health Information Technology into Clinical Guidelines", IEEE Annual International Conference, 2009, pp. 4646-4649.

[2] K.Srinivas, B.Kavihta Rani and Dr. A.Govrdhan (2010), "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, pp. 250-255.

[3] Mai Shouman, Tim Turner, Rob Stocker (2012), "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment", IEEE Japan-Egypt Conference on Electronics, Communications and Computers, 2012, pp. 173-177.

[4] Bata Sundar V, T Devi and N Saravanan (2012), "Development of data Clustering Algorithm for Predicting Heart" International Journal of Computer Applications, Volume 48, Issue 7, 2012.

[5] Nirmala Devi M., Appavu alias Balamarugan. S, Swathi U.V (2013), "An Amalgam KNN to predict Diabetes Mellitus" IEEE International Conference on Emerging Trends in Computing, 2013.

[6] GunasekarThangarasu, Assoc. Prof. Dr. P.D.D. Dominic (2014), "Prediction of Hidden Knowledge from Clinical Database using Data mining Techniques", IEEE, 2014.

[7] ShravanKumarUppin and M A Anusuya (2014), "Expert System Design to Predict Heart and Diabetes Diseases", International Journal of Scientific Engineering and Technology, Volume No.3, Issue No.8, 2014, pp : 1054-1059.

[8] AiswaryaIyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of Diabetes UsingClassification Mining Techniques", International Journal of Data Mining & Knowledge Management Process, Vol.5, No.1, 2015, pp. 1-14.

[9] Veena Vijayan V., Anjali C., "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach", IEEE Recent Advances in Intelligent Computational Systems, 2015, pp. 122-127.

[10] Amit kumarDewangan, Pragati Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques", International Journal of Engineering and Applied Sciences, Volume-2, Issue-5, 2015, pp. 145-148.

[11] Thirumal P. C. and Nagarajan N, "Utilization of Data Mining Techniques for Diagnosis ofDiabetes Mellitus - A Case Study", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 1, 2015, pp. 8-13.

[12] Srideivanai Nagarajan and R. M. Chandrasekaran, "Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques", Indian Journal of Science and Technology, Vol 8(8), 2015, pp. 771–776.

[13] C. Kalaiselvi and G. M. Nasira, "Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques", Indian Journal of Science and Technology, Vol 8(14),, 2015, pp. 1-7.

[14] Tahani Daghistani, Riyad Alshammari, "Diagnosis of Diabetes by Applying Data Mining Classification Techniques", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016, pp. 329-332.

[15] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques toPredict Diabetes", Science Direct Symposium on Data Mining Applications, 2016, pp. 115-121.

[16] B. Senthil Kumar, Dr. R. Gunavathi, "A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 12, 2016, pp. 463-467.

[17] Panigrahi Srikanth, DharmaiahDeverapal, "A Critical Study of Classification AlgorithmsUsing Diabetes Diagnosis", IEEE 6th International Advanced Computing Conference, 2016, pp. 245-249.

[18] Ekta, Sanjeev Dhawan, "Classification of Data Mining and Analysis for Predicting Diabetes Subtypes using WEKA", International Journal of Scientific & Engineering Research, Volume 7, Issue 12, 2016, pp. 100-103.

[19] Dr. M. Renuka Devi, J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research, Volume 11, Number 1, 2016, pp 727-730.

[20] P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, 2017, pp. 705-709.

[21] IoannisKavakiotis, Olga Tsave, Athanasios Salifoglou, NicosMaglaveras,IoannisVlahavas, IoannaChouvarda, "Machine Learning and Data Mining Methods in Diabetes Research", Elsevier Computational and Structural Biotechnology Journal 15, 2017, pp. 104–116.

[22] S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar, "Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques", International Journal of Statistics and Systems, Volume 12, Number 2, 2017, pp. 183-188.

[23] Messan Komi, J un Li, Y ongxinZhai, Xianguo Zhang, "Application of Data Mining Methods in Diabetes Prediction", IEEE 2nd International Conference on Image, Vision and Computing, 2017, pp. 1006-1010.

[24] Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar, "Predictive Analysis of Diabetic Patient DataUsing Machine Learning and Hadoop", IEEE International conference on I-SMAC, pp. 619-624.

[25] Ashok Kumar Dwivedi, "Analysis of computational intelligence techniques for diabetesmellitus prediction", Springer, 2017.