Available online at **www.ijmrhs.com**

ISSN No: 2319-5886

International Journal of Medical Research & Health Sciences, 2022, 11(1): 41-48

Data Mining for Diagnosis of Diabetes

Neha Kamble* **and Rashmi Thakur**

Department of Computer Engineering, Thakur College of Engineering & Technology, Mumbai, India

**Corresponding e-mail: nehakamble19@gmail.com*

Received: 18-Dec-2021, Manuscript No. ijmrhs-21-48497 (M); **Editor assigned:** 20-Dec-2021, PreQC No. ijmrhs-21-48497 (P); **Reviewed:** 06-Jan-2022, QC No. ijmrhs-21-48497 (Q); **Revised:** 20-Jan-2022, Manuscript No. ijmrhs-21-48497 (R); **Published:** 29-Jan-2022, **Invoice:** J-48497

ABSTRACT

Diabetes is a sickness that is influencing numerous individuals nowadays. The majority of exploration is occurring around here. In this paper, we proposed a model to tackle the issues in the existing framework in applying information mining methods specifically bunching and arrangements which are applied to dissect the kind of diabetes and its reality level for each persevering from the data assembled. The continuous report of WHO exhibits an uncommon move in the number of diabetic patients and this will be in a comparative model in the coming many years as well. Early distinctive verification of diabetes is a fundamental test. Data mining has expected a fundamental occupation in diabetes. Data mining would be a significant asset for diabetes experts since it can uncover hidden gaining from a huge proportion of diabetes-related data. Diverse data mining frameworks assist diabetes with exploring and in the long run improving the idea of social protection for diabetes patients. This paper gives an audit of data mining strategies that have been typically associated with diabetes data examination and estimate of the infirmity. This paper aims to predict diabetes via. different machine learning methods including AdaBoost, Decision Tree classifier, XGBoost, Naive Bayes, voting classifier. We also calculate the highest accuracy out of all methods mentioned above. This task additionally plans to propose a powerful strategy for the prior location of the diabetes sickness.

Keywords: Data mining, AdaBoost, Decision Tree classifier, XGBoost, Naive Bayes, Voting classifier, KNN

INTRODUCTION

Diabetes mellitus is a metabolic condition defined by an abnormal rise in blood sugar concentrations caused by insulin insufficiency, inadequate insulin sensitivity of tissues, or both. Diabetes can cause significant complications and even death. To diagnose diabetes, however, multiple time-consuming tests and analyses of crucial components are required [1].

In hospitals and medical-related institutions, there is a massive amount of data. In the field of health care, information technology is critical. Diabetes is a chronic condition that has the potential to wreak havoc on the global healthcare system. The early expectation of diabetes is a very difficult errand for clinical experts because of the complex reliance on different components. Diabetes influences human organs like the kidney, eye, heart, nerves, foot, etc. Data mining is a process to extract useful information from a large database. It is a multidisciplinary field of software engineering which includes computational interaction, Artificial Intelligence (AI), measurable strategies, grouping, bunching, and finding designs. Data mining techniques have proved for early prediction of disease with higher accuracy to save human life and reduce the treatment cost (Figure 1) [2,3].

Data Mining insinuates isolating gaining from a great deal of data. It enables us to examine the huge models and explore the comparable strategies for Factual and Artificial Intelligence in far-reaching datasets. The data mining framework is used to predict possible future examples or to discover hid models in the direction of the data. Techniques, for instance, Fake Neural Networks, Decision Trees, Arrangement, Clustering, Association rule estimations, etc. are by and large utilized by trained professionals.

Data mining procedures are broadly being connected by analysts in Bioinformatics. Bioinformatics is the investigation of putting away, extricating, sorting out, deciphering, what's more, using data from organic arrangements what's more, and atoms. As of late, knowledge disclosure and data mining methods are generally utilized for removing the examples from the huge natural databases. The measure of organic information is developing quickly [1].

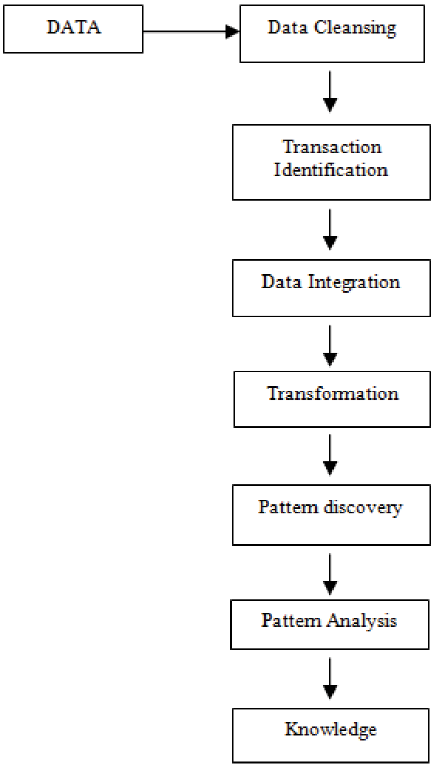

Figure 1 Data mining process

Diabetes is a state or on the other hand, a condition in which the body can't create or on the other hand use the insulin appropriately which brings about creating maladies influencing kidney, eyesight, nerve framework, veins, and heart-related issues. There are three sorts of diabetes. Type-1diabetes used to be called adolescent beginning diabetes. It is generally brought about by an auto-resistant response where the body's guard framework assaults the cells that produce insulin. Type-1 is likewise called insulin subordinate. Practically 90% of the diabetes experiences Type II diabetes which is on the other hand called non-insulin subordinate diabetes or Adult-beginning diabetes. This is portrayed by protection from insulin and insufficiency of relative insulin which may either or both together be present at the season of which diabetes is analyzed. In the instance of Gestational Diabetes Mellitus (GDM) which is frequently seen amid pregnancy coming about because of high glucose levels. This is regularly cultivated by renal intricacies, cardiovascular infections, and fringe vascular illnesses. Early recognizable proof of patients with undiscovered sort 2 diabetes or those at an expanded danger of creating type-2 diabetes is an imperative challenge in the field of medicine. The accessibility of immense measures of medicinal information prompts the requirement for incredible information investigation devices to remove helpful learning. Scientists have for quite some time been worried about applying factual what's more, information mining apparatuses to enhance information examination on vast informational indexes. Sickness conclusion, for example, diabetics is one of the applications where information mining devices are demonstrating victories in the ongoing years.

Literature Survey

A portion of the various techniques that have been applied on the PIMA Indian diabetes dataset is portrayed beneath with its outcomes. Bansal R., et al., utilized KNN classifier for the diagnosis of diabetes; the qualities are chosen using Particle Swarm Optimization (PSO) techniques. This method is proved to provide a prediction accuracy of 77% [4].

According to YA Christobel and C Sivaprakasam, a Class-wise KNN (CKNN) methodology for order of diabetes dataset was proposed where the preprocessing of the dataset is finished utilizing standardization and an ad-libbed model of KNN calculation, i.e., class shrewd KNN calculation is applied on the dataset for arrangement. This strategy accomplishes an exactness of 78.16% [5].

Harleen, et al. proposed a system based on a technique in data mining for diabetes disease prediction [6]. The proposed framework has three fundamental advances which are: pre-processing, including extraction and boundary assessment. In pre-processing step, the void and oddities sets are eliminated from the utilized dataset. Other than that, the obliging mystery models and connections of the dataset are researched in the component extraction step to additionally foster the unique result. Moreover, the proposed framework assessed dependent on utilizing J48, Naive Bayes, and the accomplished rates are 73.8%, 76.3% separately. Also, Ravi, et al., soft c means grouping and backing vector machine for making diabetes mellitus assumption [7]. The makers used a dataset that contains 768 cases and the got result was 59.5%.

Krishnaveni, et al. proposed six different procedures to foresee diabetic sickness [8]. The pre-owned strategies are Discriminant investigation, KNN Algorithm, Naive Bayes, SVM with Linear Kernel capacity, and SVM with RBF Kernel work. The obtained results of the proposed system for the used techniques are 76.3% using discriminant analysis, 71.1% using KNN Algorithm, 76.1% using Naive Bayes, 74.1% using SVM with Linear Kernel function, 74.1% using SVM with RBF kernel function. However, several authors in are used different methods to get the best prediction rate [9-11].

According to Zahed Soltani, and Ahmad Jafarian, using artificial neural networks, they can design and implement complex medical processes as software [10]. These product frameworks thus are compelling for various fields of medication sciences like finding, treatment, and helping surgeons, doctors, and people in general. These frameworks can be carried out in various scales in an equal and disseminated way. As a rule, ANNs are equal to preparing frameworks that are utilized for recognizing complex examples among information. Thusly, in their paper PNNs are applied to recognizing diabetes infection type 2. They executed the PNN model in MATLAB. The Pima Indians Diabetes data set was used for diagnosing diabetes type 2, which consists of 768 data samples with 9 features. 90% of these 768 samples are used as training set and 10% are used as testing set. The method achieved 89.56% of diagnostic accuracy in the training phase, and 81.49% in the test phase. Both training and testing measures could identify the diabetes disease type 2 with good accuracy.

Thangaraju, et al. data mining is the act of looking at huge previous databases to produce new information. There are different sorts of data mining techniques are available [12]. Order, Clustering, Affiliation Rule, and Neural Network are the absolute most noteworthy methods in information mining. In healthcare businesses, data mining plays a noteworthy job. Most oftentimes information mining is utilized in human services enterprises as a way toward anticipating ailments. Diabetes is a ceaseless condition. This implies it goes on for quite a while, frequently for somebody's entire life. This paper contemplates the examination of diabetes estimating approaches utilizing grouping procedures. Here we are utilizing three various types of bunching strategies named Hierarchical grouping; Thickness-based bunching, and Simple K-Means grouping. Weka is utilized as a device.

Priyadarshini, et al., have used the concept of modified extreme learning machines to predict whether the patient is having diabetes or not based on the available diabetes dataset. The authors have drawn comparative inferences using neural networks and extreme learning classifiers [13]. A short survey of information digging procedures used for the conclusion of diabetes is talked about. In the survey, the creators have referenced that many information-digging calculations were utilized for the finding in that 85% were portrayed by managed learning draws near and 15% by unsupervised ones, and all the more explicitly, affiliation rules.

Proposed System

A. Data set: Pima Indians Diabetes (PID) dataset of National Institute of Diabetes and Digestive and Kidney Diseases. PID is composed of 768 instances as shown in Table 1. Eight numerical attributes are representing each patient in the data set.

Table 1 Dataset attributes

No.	Attributes
1	Number of times pregnant
2	Plasma glucose concentration in an oral glucose tolerance test
3	Diastolic blood pressure (mm/Hg)
4	Triceps skinfold thickness (mm)
5	2-hour serum insulin (mu U/ml)
6	Body mass index (kg/m^2)
7	Diabetes Pedigree function
8	Age (year)

B. Ada-boost: Ada-boost or Adaptive Boosting is one of the gathering boosting classifiers proposed by Yoav Freund and Robert Schapire in 1996. It joins different classifiers to build the exactness of classifiers. AdaBoost is an iterative gathering technique. AdaBoost classifier assembles a solid classifier by consolidating various inadequately performing classifiers so you will get high exactness solid classifier. The essential idea driving Adaboost is to set loads of classifiers and prepare the information test in every emphasis with the end goal that it guarantees the precise forecasts of uncommon perceptions. Any machine learning calculation can be utilized as a base classifier if it acknowledges loads on the preparation set. Adaboost should meet two conditions:

- The classifier should be trained interactively on various weighted training examples.
- In each iteration, it tries to provide an excellent fit for these examples by minimizing training errors.

C. Decision tree: A decision tree is a tree structure, which is a flowchart. It is utilized as a technique for arrangement and forecast with portrayal utilizing hubs and internodes. The root and interior hubs are the experiments that are utilized to isolate the occasions with various highlights. Interior hubs themselves are the consequence of characteristic experiments. Leaf hubs mean the class variable.

Decision tree gives an amazing method to order and expectation in diabetes finding issues. Different Decision tree calculations are accessible to order the information, including ID3, C4.5, C5, J48, CART, and CHAID Every hub for the Decision tree is found by ascertaining the most noteworthy data acquire for all credits and if a particular trait gives an unambiguous finished result (unequivocal arrangement of class quality), the part of this characteristic is ended and target esteem is allotted to it.

D. K-Nearest Neighbor (K-NN): It is the nearest neighbor computation. The K-nearest neighbor's computation is a methodology for gathering objects reliant upon the accompanying planning data in the component space. It is the most effortless among all instruments learning estimation. This estimation is presented by picking k concentrations in K_D as the basic k gathering representatives or "centroids". Methodology for selecting the fundamental seeds joins reviewing unpredictably from the dataset setting them as the course of action of clustering a little subset of the data or disturbing the overall mean of the data k-times. By then the estimation stresses between two phases till crossing point:

- Step 1: In data assignment, each datum point is designated to its associating centroid, with ties broken abstractly. This results in an allocating of the data.
- Step 2: Relocation of "connotes" each get-together specialist is moved to the point of convergence of all data coordinates delegated toward it. If the data centers go with a credibility measure, by then the movement is to the longings for the data sections.

E. XGBoost: XGBoost is perhaps the most famous machine learning algorithm nowadays. Notwithstanding the sort of forecast job needing to be done; relapse or grouping. XGBoost is notable to give preferred arrangements over other AI calculations. XGBoost (Extreme Gradient Boosting) has a place with a group of boosting calculations and utilizations the slope boosting (GBM) system at its centre. It is an advanced dispersed slope boosting library.

Boosting is a consecutive procedure that deals with the rule of a group. It consolidates a bunch of powerless students and conveys further developed expectation precision. At any moment t, the model results are weighed dependent on the results of past moment t-1. The results anticipated effectively are given a lower weight and the ones miss-arranged are weighted higher. Note that a powerless student is marginally better compared to arbitrary speculating. For instance, a choice tree whose forecasts are marginally better compared to half. We should comprehend boosting overall with a straightforward outline.

F. Naive Bayes: Naïve Bayes classifier depends on Bayes hypothesis. This classifier utilizes contingent freedom in which characteristic worth is autonomous of the upsides of different qualities. The Bayes hypothesis is as per the following:

Let X={x_1, x_2,… … ,x_n} be a bunch of n ascribes.

In Bayesian, X is considered as proof and H be some theory implies, the information of X has a place with explicit class C.

We need to decide P (H|X), the likelihood that the theory H holds given proof for example information test X. As indicated by Bayes hypothesis the P (H|X) is communicated as

P(H|X)=P(X|H) P(H)/P(X).

G. Voting Classifier: A Voting Classifier is a machine learning model that trains on a gathering of various models and predicts a yield (class) given their most noteworthy likelihood of picked class as the yield.

It just totals the discoveries of every classifier passed into Voting Classifier and predicts the yield class dependent on the most elevated greater part of casting a ballot. The thought is as opposed to making separate devoted models and discovering the precision for every them, we make a solitary model which trains by these models and predicts yield dependent on their consolidated larger part of deciding in favour of each yield class.

The Ensemble Vote Classifier carries out "hard" and "delicate" casting a ballot. In hard democratic, we anticipate the last class mark as the class name that has been anticipated most as often as possible by the older models. In delicate democratic, we foresee the class marks by averaging the class probabilities (possibly suggested if the classifiers are all around aligned).

H. Feature Selection Process: Feature Selection measures hold an extremely high significance in machine learning which gigantically impacts the exhibition of your model. The information includes that you use to prepare your machine learning models or calculations affects the presentation you can accomplish. Superfluous or missing highlights can contrarily affect the presentation of the framework.

Feature determination is one of the first and significant advances while playing out any machine learning task. An element in the event of a dataset essentially implies a segment. At the point when we get any dataset, not really every segment (Feature) will affect the yield variable. If we add this insignificant Feature in the model, it will simply make the model most noticeably awful and can diminish the precision of the models and cause your model to learn dependent on immaterial Features. This brings about the need of doing Feature determination.

METHODOLOGY

Ensemble Classifier is a mixture of several models that offers better quality than a single model, which in effect improves classification exactness or predictive efficiency.

In this proposed model, first, we are doing training and testing operations on the dataset (Pima Indians Diabetes (PID) dataset). After that, we are doing a feature selection process i.e. forward feature selection and backward feature selection for selecting the best features from dataset attributes. After selecting the best features, we are applying different machine learning algorithms like Ada-boost, XGBoost, Decision tree, KNN, Naive Bayes to calculate various parameters like accuracy, precision, recall, F1 score, etc. Now we had done a comparative analysis of all the mentioned algorithms and found the highest accuracy. After that clustering of the algorithm is done by selecting three algorithms from the above-mentioned algorithms and applying them to the voting classifier to get better accuracy from individual algorithms.

Python is chosen for this application. Python is an intelligent tool for development. Python is widely regarded as the language of choice of computer models of education and learning.

RESULTS AND DISCUSSION

The outcome was obtained with the proposed approach on various parameters, such as accuracy, error rate, confusion matrix, precision, recall, F1 score, and the algorithm tests performance.

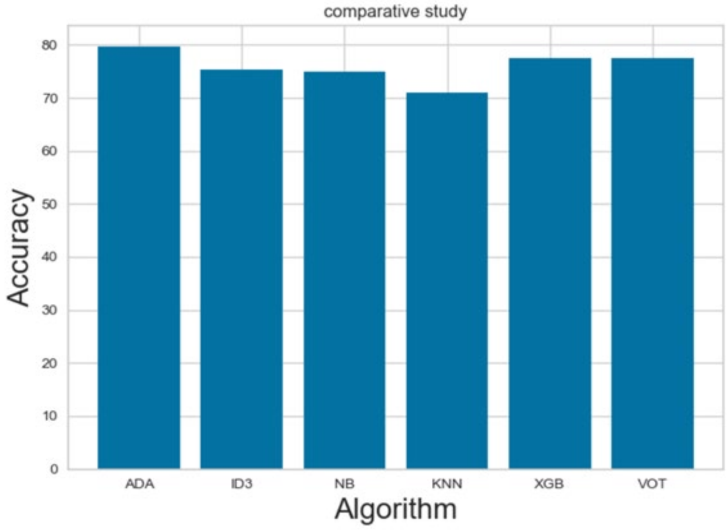

Figure 1 Comparative study of the accuracy of all algorithms used

So in this proposed approach, we had done a feature selection process on the data set and selected the best features. Then we calculate the accuracy of all algorithms used and find out the maximum accuracy. As in the above Figure 1, we come to know that the Ada-Boost algorithm is having maximum accuracy. We had also done a comparative study of AUC-Score as mentioned in Figure 2. We had also done a comparative study of all factors like accuracy precision, recall, F1-Score as well (Table 2).

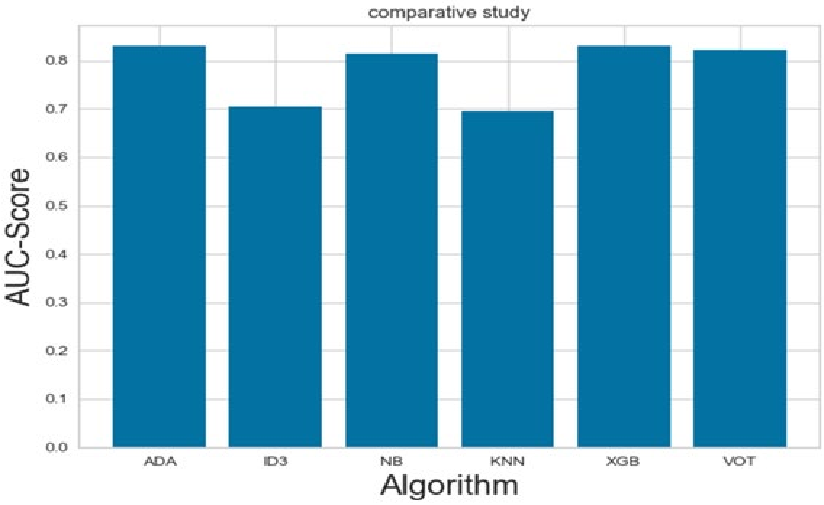

Figure 2 Comparative study of AUC-Score of all algorithms used

Table 2 Comparative study of all factors of all algorithms

Algorithm	Accuracy	AUC	log loss	Precision	Recall	Fl-score	Support
Ada Boost	80.0865	0.83	0.4734	Y=0.82	Y=0.90	Y=0.86	Y=157
				N=0.73	N=0.59	N=0.66	N=74
Decision tree	75.3246	0.70	8.5226	Y=0.80	Y=0.84	Y=0.82	Y=157
				N=0.63	N=0.57	N=0.60	N=74
Naïve Bayes	74.8917	0.81	0.5339	Y=0.80	Y=0.85	Y=0.82	Y=157
				N=0.62	N=0.54	N=0.58	N=74
Knearest Neighbour	70.9956	0.69	3.7622	Y=0.76	Y=0.84	Y=0.80	Y=157
				N=0.56	N=0.43	N=0.49	N=74
XGBoost	77.489	0.83	0.5223	Y=0.80	Y=0.90	Y=0.84	Y=157
				N=0.70	N=0.51	N=0.59	N=74
Voting classifier	75.7575	0.83	0.4829	Y=0.81	Y=0.83	Y=0.82	Y=157
				N=0.63	N=0.59	N=0.61	N=74

From Figure 1, we come to know that after applying three algorithms i.e. Ada-boost, XGBoost, KNN to the voting classifier we get low accuracy than a single Ada-Boost algorithm.

CONCLUSION

In this paper, various techniques are discussed to predict the diagnosis of diabetes. Using the data mining technique the health care management predicts the disease and diagnosis of diabetes and then the health care management can alert the human being regarding diabetes based upon this prediction. From the results, we conclude that Ada-Boost Algorithm is giving maximum accuracy of 80.086% then various other techniques. After a comparative study, we come to know that Ada-Boost is giving maximum accuracy even though using the voting classifier.

Future Scope

As in this proposed system we had used boosting algorithm for the voting classifier. So instead of that we can also use other algorithms and can find the maximum accuracy in the future. We had done feature selection and select the best features from data set attributes. But we can also do and check the accuracy and all other factors without feature selection in the future as well.

DECLARATIONS

Acknowledgment

I express many thanks to Dr. Rashmi Thakur Asst. Prof. of Thakur College of Engineering and Technology (Computer Science Dept.) for her great effort in guiding me to accomplish this task. She was great support and encouragement. I would also like to thank the National Institute of Diabetes and Digestive and Kidney Diseases for maintaining Dataset.

Conflict of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] Han, Jiawei, Jian Pei, and Micheline Kamber. "Data mining: Concepts and techniques." Elsevier, 2011.

[2] Papageorgiou, Eirini, Ioanna Kotsioni, and Athena Linos. "Data mining: A new technique in medical research." *Hormones,* Vol. 4, No. 4, 2005, pp. 189-91.

[3] Zaki, Mohammed J., Jason TL Wang, and Hannu TT Toivonen. "BIOKDD01: Workshop on data mining in bioinformatics." *ACM SIGKDD Explorations Newsletter,* Vol. 3, No. 2, 2002, pp. 71-73.

[4] Mahajan, Aakanksha, Sushil Kumar, and Rohit Bansal. "Diagnosis of diabetes mellitus using PSO and KNN classifier." *2017 International Conference on Csomputing and Communication Technologies for Smart Nation (IC3TSN)*, 2017.

[5] Christobel, Y. Angeline, and P. Sivaprakasam. "A new Classwise K Nearest Neighbor (CKNN) method for the classification of diabetes dataset." *International Journal of Engineering and Advanced Technology,* Vol. 2, No. 3, 2013, pp. 396-400.

[6] Harleen, Bhambri. "A prediction technique in data mining for diabetes mellitus." *Journal of Management Sciences and Technology,* Vol. 4, No. 1, 2016, pp. 1-12.

[7] Sanakal, Ravi, and T. Jayakumari. "Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine." *International Journal of Computer Trends and Technology,* Vol. 11, No. 2, 2014, pp. 94-98.

[8] Krishnaveni, G., and T. Sudha. "A novel technique to predict diabetic disease using data mining-Classification techniques." *International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (IJASTEMS),* Vol. 3, No. 1, 2017, pp. 5-11.

[9] Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." *International Journal of Engineering Research and Applications,* Vol. 3, No. 2, 2013, pp. 1797-801.

[10] Soltani, Zahed, and Ahmad Jafarian. "A new artificial neural networks approach for diagnosing diabetes disease type II." *International Journal of Advanced Computer Science and Applications,* Vol. 7, No. 6, 2016, pp. 89-94.

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

[12] Thangaraju, P., B. Deepa, and T. Karthikeyan. "Comparison of data mining techniques for forecasting diabetes mellitus." *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 3, No. 8, 2014.

[13] Priyadarshini, Rojalina, Nilamadhab Dash, and Rachita Mishra. "A novel approach to predict diabetes mellitus using modified extreme learning machine." *2014 International Conference on Electronics and Communication Systems (ICECS)*, IEEE, 2014.