THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Prediction of Diabetes Based on Data Mining Techniques

Madhusmita Rout
Department of computer science and engineering
Lovely Professional University, Punjab
msrout77@gmail.com

Amandeep Kaur
Department of computer science and engineering
Lovely Professional University, Punjab
er.amandeep.kaur@gmail.com

***Abstract*— There have been many lethal diseases around the world causing various health issues that are getting very difficult for both prognosis and diagnosis of it. Nowadays, machine learning has become very popular in the health industry. As the amount of data is very high in volume and required a proper extraction to build a good predictive model, data mining is used extensively for this reason. The parameters can be chosen based on common symptoms. The frames need to be set well before a model is built. In this paper, we have applied various data mining techniques to a proposed predictive model to compare the accuracy of different algorithms. The experiment shows that logistic algorithms have an accuracy of 82.35% that is higher than other classifiers i.e. Support Vector Machine (SVM), Naive Bayes, Decision tree and K-nearest neighbour.**

Keywords*— *diabetes mellitus, machine learning, data mining techniques, predictive analysis, the Accuracy rate

I. INTRODUCTION

Diabetes mellitus (DM) is a deadly disease in the world, causing the death of millions per year. The WHO statistics show that more than 422 million people are suffering from diabetes [1]. The report also estimates that the death rate in diabetes is higher than other diseases in recent years. Earlier medical records show that there is still no cure to diabetes but with proper health maintenance and regular medical check-up the effect can be balanced well. Previously it was seen that mostly at old age the people were suffering in diabetes but now the youngsters are the most affected by it i.e. not only a specific age group is vulnerable to the disease but all. In general, diabetes occurs when a body can't produce insulin or can't correctly use the insulin thus leading to a rise in blood sugar level. People are sometimes unaware of the presence of disease in their body which can be for a long time. This increases the chance of the occurrence of diabetes. T2DM is such a condition that can be gone unnoticed for many years [3]. Diabetes can be classified into the following types:

- Type-1 diabetes- The beta cells are destroyed and no insulin is produced in the body [1]. It accounts for nearly 10% of all cases. It can occur in both kids and youngsters.
- Type-2 diabetes- It accounts for about 90% of all cases with a risk of developing many medical risks like diabetic retinopathy (damages blood vessels in eyes), diabetic neuropathy, kidneys and cardiovascular diseases [2].
- Gestational diabetes- It affects during the pregnancy which is identified as a risk of T2DM in the future for both mother and child [1]. It is usually cured after the childbirth.
- Impaired glucose tolerance (IGT) and impaired fasting glycaemia (IFG) are a condition where people have blood sugar level higher than the normal range but cannot be considered as diabetes but also a greater chance of progressing to type-2 diabetes can occur in future [1].

Many cases are undiagnosed because of no early check-up or awareness about the blood glucose level among different age groups which increases the risk of other diseases. There is no cure for diabetes but can be controlled with a healthy lifestyle and by following up a regular treatment.

A. Data mining in Diabetes

The main goal for using machine learning applications in healthcare is to improve the accuracy of a predictive model and to enhance a better quality of check up for the patient in a cost-friendly aspect within a very little time [4]. The vast amount of data is generalized using data mining methods. Machine learning algorithms are used for building a prognosis or diagnosis model with higher accuracy and efficiency [4]. There are many proposed systems already but the accuracy rate needs an improvement to ensure a better quality model.

An analysis is required at an early stage to determine the stage or severity of a disease. Data mining tools are used extensively with machine learning algorithms for examining the proposed system if it is valid or not for real-time decision making. The algorithms are used solely or in a hybrid manner to increase the accuracy of a model with various methods.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

The machine learning algorithms are classified into two categories as follows-

1. ***Supervised learning***- with a collection of a large number of training and test datasets which contains both input and desired output to make the algorithm learn for any new cases that are fed into the predictive model to produce a correct output [5]. Classifications, regression techniques are used here.
2. ***Unsupervised learning***- used with unlabelled data for inferring the hidden patterns or layers using methods like clustering, feature learning, anomaly detection and dimensionality reduction [5].

Sometimes it's a tough task to derive only useful knowledge data from a large database. In such a case, data mining is very helpful. It is based on supervised and unsupervised learning. It is used for both predictive and descriptive analytics. Future trends and probabilities are obtained using predictive analysis. Descriptive analytics is limited to the past to know what has happened.

II. RELATED WORKS

Ding et al. [6] aimed to solve a multi label classification problem to improve diabetic complication prediction accuracy using clinical datasets. Used classifiers are Support vector machines (SVM)-seLDA. The outcome shows an improvement of 22.49% than the previous works.

Alkargole et al. [7] analysed and compared different data mining techniques as a hybrid framework using the Pima Indian dataset. For evaluation apache server was used. They applied classification algorithms like Decision tree, Naive Bayes and SVM. The proposed method achieved an accuracy of 94%.

Sneha et al. [8] designed a prediction algorithm to find an optimal classifier for prediction using a rapid miner tool. The dataset was collected from the UCI ML repository. Classification algorithms i.e. SVM, KNN, Decision tree, Random forest, Naive Bayes was applied for evaluation. The highest accuracy was achieved is 82.30% by naive Bayes

Faruque et al. [9] explored various risk factors to build a predictive model. They used a clinical dataset i.e. from the Diagnostic of medical center Chittagong. Various tools were used to perform this experiment. Classification algorithms like SVM, C4.5, Decision tree, and KNN was applied to the model. Better performance was shown by C4.5 i.e. 73.5% which is higher than other algorithms.

Liu et al. [10] designed a multi-task learning model for type-2 diabetes. They used a hierarchical Bayesian framework and evaluated the model using the R tool. The result shows an improvement of 9.1%.

Sonar et al. [11] developed a system to predict the diabetes risk level using the Pima dataset from the UCI machine learning repository. They used various data mining techniques to evaluate classifiers like SVM, ANN, decision tree and Naive Bayes. The result showed that ANN performed better with 82% that is higher than other algorithms.

Bai et al. [12] examined the efficiencies of various algorithms for the diagnosis of diabetes using the Pima dataset. Methods were used like Gaussian Naïve Bayes, BIRCH and OPTICS algorithms. The outcome shows that optics were more suitable for prediction.

Priya et al. [13] investigated the relative performance of various classifiers using the Weka tool. Here Pima dataset was used. Classifiers like SMO-SVM, Naive Bayes, decision tree and neural network was used. The outcome shows that 83.5% by Naïve Bayes was highest which outperforms others.

Srivastava et al. [14] analysed that artificial neural network to build a diabetes prediction model using the Pima dataset. Experimental results show that 92% accuracy was obtained by the ANN. A large dataset can be used further for obtaining more accuracy.

Abdallah et al. [27] applied a predictive data mining model based on past records retrieved from the database and conducted a questionnaire for both primary data and secondary data. The patient's records were obtained from the Bahrain Defence Force Hospital to examine the proposed software application. VS.net tool was used for writing the application and VB.net for coding purposes. Additionally, the oracle database was also used for storing and manipulating the data. This proposed software is already examined by professionals. The output was labelled with no risk, low risk, medium risk, and high risk. The results were satisfying as per the doctor's expectation. In various clinics, this software can be used by modifying its code to achieve accuracy.

Das et al. [28] aimed at analysing the prediction of diabetes using different classification algorithms like JV8, naive Bayesian to diagnose the disease quickly. The proposed model saves the time of both patients and doctors by generating reports quickly from the data repository. It was evaluated using both WEKA and MATLAB software. Further studies can be done with hybrid data mining techniques for an accurate and precise result.

Kavakiotis et al. [29] presented a review paper about the machine learning applications and data mining methods and tools that are used extensively for diabetes research to analyse diabetes prediction, diagnosis, complications, genetic background and environment, drugs and therapies. The review is summarized from several scientific journals published recently. It concluded that cardiovascular diseases and retinopathy were most common in diabetes patient

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Vijayan et al. [30] proposed a predictive system using the AdaBoost method using the Pima dataset. Classification algorithms like Support Vector Machine, Naive Bayes and Decision Tree were used for accuracy verification. The accuracy obtained is 80.72% which outperforms other classifiers. Further improvement can be done with the proposed decision support system using other classifiers like an artificial neural network, KNeighbor, etc. with dataset collected from other regions.

Table 1: Review of techniques used in the analysis of diabetes

Author	Description	Dataset	Tool	Methodology	Accuracy
Varma, K. M., & Panda, B. S [15]	Explores mining techniques and compare their performance analysis	PIMA	R tool	Naive Bayes, SVM, Logistic regression, Decision tree	74.67% by Decision tree
Zhu, C., Idemudia, C. U., & Feng, W. [16]	Improve accuracy for early diagnosis and prediction by hybrid techniques	PIMA	Anaconda	K-means, PCA, logistic regression	97.40%
Aada, M. T. S. A., & Tiwari, S. [17]	Upgrade the accuracy	PIMA	R tool	Naïve Bayes, Decision tree, KNN	94.44% by Ad boost and SVM
Saru, S., & Subashree, S. [18]	Compare different classifiers and accuracy for better prediction	PIMA	WEKA	Naive Bayes, Decision tree, KNN	94.44%
Sengamuthu, M. R., Abirami, M. R., & Karthik, M. D. [19]	Explore various techniques for better accuracy	PIMA	WEKA and MATLAB	KNN, ANN, C4.5, Naive Bayes, Genetic algorithms	99.87%
Mir, A., & Dhage, S. N. [20]	To recommend the best algorithm based on efficient performance	PIMA	WEKA	Naïve Bayes, SVM, CART, Random forest	79.13% by SVM
Yasen, M., Al-Madi, N., & Obeid, N. [21]	Compare and enhance optimization accuracy and speed	PIMA	WEKA	Dragonfly algorithm, ANN	79.77 by ANN-DA
Woldemichael, F. G., & Menaria, S. [22]	To classify diabetic and non-diabetic persons	PIMA	R studio	SVM, Naïve Bayes, Backpropagation algorithm	83.11% by the Backpropagation algorithm
Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. [23]	Improve accuracy and make the proposed technique to be more adaptive to multiple datasets	PIMA	WEKA	K-means, Logistic regression	applicable

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Kaur, H., & Kumari, V. [24]	Detect patterns with risk factors to classify as diabetic and non-diabetic	PIMA	R tool	SVM- linear, radial basis function (RBF) SVM, KNN	89% by SVM-linear
Sisodia, D., & Sisodia, D. S. [25]	Evaluate performances of various algorithms	PIMA	WEKA	Decision tree, Naive Bayes, SVM	76.30% by Naive Bayes
Singh, P. P., Prasad, S., Das, B., Poddar, U., & Choudhury, D. R. [26]	Classify the data correctly	PIMA	WEKA	ANN	88.6%

III. PROPOSED METHODOLOGY

According to the problem identification mentioned in the introduction section, we developed a classification model that will predict diabetes. In this model, different classifiers will be used like Logistic Regression, Gaussian naive Bayes, KNN, Decision tree and Support vector machine (SVM). Each individual algorithm will be applied to the model to obtain accuracy.

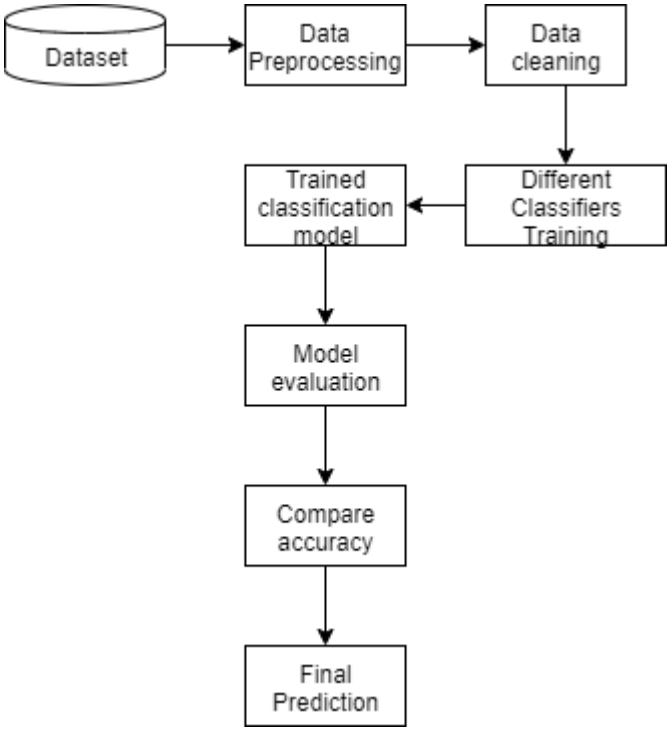

Figure 3.1 Proposed Framework

3.1 Dataset Collection

For this study, the PIMA Indian dataset is collected from the UCI Machine Learning Repository. It was originally collected from the Pima people of America. Due to a major change in their lifestyle, diabetes occurred in the dataset contains a record of 769 patients with nine attributes. This dataset is already used by many researchers for their experimental work to predict the onset of diabetes mellitus. But some missing values or noisy values are present in this dataset which can make the

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

prediction analysis slow. Hence normalization of data is necessary. We can first fill the missing values or null values then normalize them for further use. So, data pre-processing is required to obtain structured data.

Table 2- PIMA Attributes Description.

Attributes	Types
Pregnancies	Numeric
Glucose level	Numeric
Blood Pressure (mmHg)	Numeric
BMI (Body Mass Index)	Numeric
Skinfold thickness (mm)	Numeric
Insulin value in 2 hrs. (mu U/ml)	Numeric
Diabetes Pedigree function	Numeric
Age (years)	Numeric
Outcome	Nominal

3.2 Training and Test Dataset

The training dataset is a set of learning sets (a set of parameters) that are required to train the algorithms that can learn and predict. A test set is used only to assess the performance of the selected classifiers. It is included only for the testing of the classifiers. The size of the training and test dataset can differ for the evaluation of the model. Mostly heuristic methods are applied for it. If the model performs well in both datasets, that means the accuracy is better.

3.3 Data Pre-processing

A dataset may contain a noisy or missing value which can degrade the performance of a dataset. That is why data cleaning is required to fill up or smooth those values to remove any inconsistency in the dataset. Data pre-processing involves data cleaning, normalization, data transformation, feature extraction for dimensional reduction of the dataset. It helps to build structured data out of the unstructured data.

Feature Selection- It is also called as the variable selection. It helps a model to be simplified easily that allows for an easy interpretation, allows an algorithm to be trained faster, reduce the complexity and over fitting. Only useful features are selected using these methods.
Feature Extraction- It is generally used for the high dimensional dataset for accuracy improvements. Sometimes the number of features becomes similar so to avoid it dimensional reduction methods are applied.

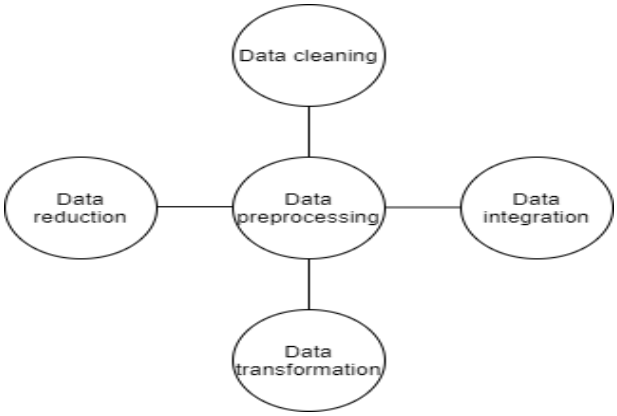

Figure 3.2 Data pre-processing method

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

3.4 Anaconda Tool

Anaconda distribution is an open source for many programming languages and provides more than 1500 packages. It includes a GUI, Anaconda Navigator and command-line interface and available for different OS platforms. In this experiment, we will be using python for evaluating the proposed method.

3.5 Classification Algorithms

These types of algorithms fall under a supervised learning approach that can be performed on any type of data. Classification learns from the input data and then based upon it classify the new data. This technique helps in identifying the class labels where new data can be fit.
Different classification algorithms are follows-

3.5.1 Logistic regression

It helps in understanding the relationship between multiple independent variables and a single dependent variable. The outcome is specific to certain events. In many of the medical cases, it is used to determine the severity of a patient using logistic regression models. Logistic regression can be binomial, ordinal or multi-nominal. It works only for the binary variables i.e. the outcome is '0' or '1'. Software used for it is R, SAS, Python, Java, and Matlab.

3.5.2 Decision Tree

A decision tree looks like a tree model with many leaf nodes and root nodes. It can be applied to both numerical and categorical data. It includes pros like easier visualization and understands. But sometimes due to little variation in data can generate the decision trees which could be very complex and unstable and leads to over-fitting. It can be overcome with bagging and boosting algorithms.

3.5.3 K-Nearest Neighbours

KNN is a widely used algorithm in classification problems. It's a non-parametric and lazy learning algorithm because it does not learn during the training phase. It is very simple to implement and can be applied to a large dataset. The value of k depends upon the data i.e. if the k value is large, the effect of noise is reduced. But the computation time can be high because each query instance will be calculated from all the training samples present in a dataset. Sometimes a confusion matrix can be used to validate the accuracy of the KNN algorithm. It can be used in a regression problem also.

3.5.4 Support vector machine Algorithm

SVM is used in both classification and regression problems. It works better for high dimensional spaces but the estimation about the probability is not direct. So, further fold-cross validation may be calculated. The output is understood by the hyper-plane that differentiates the given classes very clearly. It may take more training time for a larger dataset.

3.5.5 Naive Bayes

It is either known as simple Bayes. It is a very simple algorithm. This classifier is based on Bayesian theory. It is a highly scalable algorithm. It works best for binary classification. But, it may not work well for larger datasets.

$$P(z|o) = (P(o|z) * P(z)) / P(o)$$

where,

- P(z/o)-posterior probability of class z
- P(o/z)-likelihood probability of o
- P(z)-prior probability of class z
- P(o)-probability of o

IV. RESULTS AND ANALYSIS

The dataset was normalized using the panda library. First of all, the dataset is visualized to find outliers using the box plot. After that filling the missing values or null with the mean value. Then after we have applied classifiers like logistic regression, Naive Bayes, Decision tree, K-nearest neighbour and Support vector machine to the model.
The experiment is evaluated using a python tool which shows the performance of an individual algorithm. Logistic Regression is having a more accuracy rate than other classifiers i.e. 82.35%. Other classifiers obtained an accuracy of 76.62%, 75.97%,

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

66.23%, 64.28% by naive Bayes, decision tree, KNN, and SVM respectively. But this is only based on individual performance in aspect of accuracy.

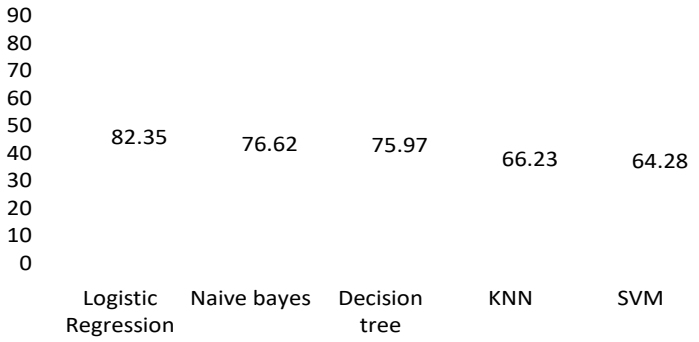

Figure 4.1- Comparison of different classification algorithms based on the accuracy

V. CONCLUSION

This paper focuses on a detailed study of various machine learning algorithms used for diabetes prognosis and diagnosis. Every individual algorithm has some drawbacks so to overcome that in the future we can prefer to make hybrid models in which will comprise multiple algorithms and also evaluating this on different datasets will show the accuracy of the model. The risk factors and complications involved in diabetes is the main concern. The parameters to be selected are examined in clinical datasets. Models can be also improved by hybrid methods as summarized in table-1. The accuracy differs in different proposed hybrid systems based on the chosen algorithms. The error rates are analysed and other factors also based on which an algorithm is to be selected. More datasets need to be evaluated so the decision making can be more accurate with very low error rates. Thus this can be further used for the prediction of diseases in the clinics in various regions.

REFERENCES

[1] *diabetes*. (2018). Retrieved from https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] The Editors of Encyclopaedia Britannica. (2019, September 6). Diabetes mellitus. Retrieved from https://www.britannica.com/science/diabetes-mellitus

[3] American Diabetes Association. (2019). 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. *Diabetes Care*, *42*(Supplement 1), S13-S28.

[4] Chen, P. H. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare Nature, *Materials*, *18*(5), 410.

[5] Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), 719.

[6] Ding, S., Li, Z., Liu, X., Huang, H., & Yang, S. (2019). Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model. *Information Sciences*, *499*, 12-24.

[7] Alkaragole, M. L. Z., & Kurnaz, A. P. S. (2019). COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS.

[8] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1), 13.

[9] Faruque, M. F., & Sarker, I. H. (2019, February). Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-4). IEEE.

[10] Liu, B., Li, Y., Ghosh, S., Sun, Z., Ng, K., & Hu, J. (2019). Complication Risk Profiling in Diabetes Care: A Bayesian Multi-Task and Feature Relationship Learning Approach. *IEEE Transactions on Knowledge and Data Engineering*.

[11] Sonar, P., & JayaMalini, K. (2019, March). Diabetes Prediction Using Different Machine Learning Approaches. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 367-371). IEEE.

[12] Bai, B. M., Nalini, B. M., & Majumdar, J. (2019). Analysis and Detection of Diabetes Using Data Mining Techniques—A Big Data Application in Health Care. In *Emerging Research in Computing, Information, Communication, and Applications* (pp. 443-455). Springer, Singapore.

[13] Priya, D. T., & Udayan, J. D. (2019). Comparative Study of Classification Algorithm for Diabetics Data. In *Advances in Big Data and Cloud Computing* (pp. 327-351). Springer, Singapore.

[14] Srivastava, S., Sharma, L., Sharma, V., Kumar, A., & Darbari, H. (2019). Prediction of Diabetes Using Artificial Neural Network Approach. In *Engineering Vibration, Communication and Information Processing* (pp. 679-687). Springer, Singapore.

[15] Varma, K. M., & Panda, B. S. (2019). Comparative analysis of Predicting Diabetes Using Machine Learning Techniques.

[16] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, 100179

[17] Aada, M. T. S. A., & Tiwari, S. (2019). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques.

[18] Saru, S., & Subashree, S. (2019). Analysis and Prediction of Diabetes Using Machine Learning. *International Journal of Emerging Technology and Innovative Engineering*, *5*(4).

[19] Sengamuthu, M. R., Abirami, M. R., & Karthik, M. D. (2018). Various Data Mining Techniques Analysis to Predict Diabetes Mellitus.

[20] Mir, A., & Dhage, S. N. (2018, August). Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-6). IEEE.

[21] Yasen, M., Al-Madi, N., & Obeid, N. (2018, July). Optimizing Neural Networks using Dragonfly Algorithm for Medical Prediction. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)* (pp. 71-76). IEEE.

[22] Woldemichael, F. G., & Menaria, S. (2018, May). Prediction of Diabetes Using Data Mining Techniques. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 414-418). IEEE

[23] Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, *10*, 100-107.

[24] Kaur, H., & Kumari, V. (2018). Predictive modeling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*.

[25] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, 1578-1585.

[26] Singh, P. P., Prasad, S., Das, B., Poddar, U., & Choudhury, D. R. (2018). Classification of Diabetic Patient Data Using Machine Learning Techniques. In *Ambient Communications and Computer Systems* (pp. 427-436). Springer, Singapore.

[27] Aldallal, A., & Al-Moosa, A. A. A. (2018, September). Using Data Mining Techniques to Predict Diabetes and Heart Diseases. In 2018 4th International Conference on Frontiers of Signal Processing (ICFSP) (pp. 150-154). IEEE

[28] Das, H., Naik, B., & Behera, H. S. (2018). Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In Progress in Computing, Analytics and Networking (pp. 539-549). Springer, Singapore

[29] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104

[30] [46] Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.