Diabetes Detection Using Machine Learning Algorithms

Nicole D’Souza   
Information Technology  
MPSTMEIndia

nicolemichelledsouza@gmail.com

Kunjal Shah

Information Technology  
MPSTMEIndia

kunjal2001@hotmail.com

Pranav Singh   
Information Technology  
MPSTMEIndia

singh.pranav161@gmail.com

*Abstract*— Diabetes is a critical disease. The timely prediction of this disease is essential in order to avoid drastic side effects. The current practice in the medical industry is to gather the required information for diabetes diagnosis through numerous tests and treatment is later provided on the basis of diagnosis. However, in a lot of cases, the early stages go undetected and it is also quite challenging for medical practitioners to diagnose it due to complex interdependence on various factors.  A single parameter is not very effective in accurately diagnosing diabetes and may be misleading in the decision-making process. To properly forecast diabetes at a preliminary phase, different criteria must be combined. The development of an early diabetes detection model is proposed in this research. The model will not only outperform humans in terms of accuracy, but it will also minimise the strain on medical practitioners.

Keywords— machine learning, diabetes, ensemble,

# Introduction

Diabetes Mellitus has become a significant public health problem. It caused by several factors, including heredity, obesity, insulin resistance, physical inactivity and many more.  As of 2014, that about 422 million people (World Health Organization [WHO], 2020) had diabetes worldwide,

and an estimated 700 million adults worldwide will have diabetes by 2045. (International Diabetes Federation [IDF], 2020). Due to this reason, an increasing number of researchers are focusing on finding a way to effectively predict diabetes at an early stage to avoid over complications.  Diabetes can be detected early, which reduces the risk of developing conditions such as renal failure, stroke, blindness, heart attacks, and lower limb amputation.

Because of the significant social impact of this particular condition, a massive volume of data is unavoidably generated. Thus, data mining and machine learning are effective methods for extracting usable information from vast databases. Various methods have been developed in the framework of this study, and as a result, an ensemble strategy employing machine learning and data mining for diabetes classification has been proposed.

Data mining is the study of strategies for gaining knowledge from databases. It is the process of identifying and analysing intriguing patterns in massive datasets. Data mining entails clustering, association, and categorization of data. There are various categorization approaches available, such nave bayes, k-nearest neighbours, decision trees, regression, and artificial neural networks. These categorization methods have been used in medical, commercial, and industrial applications.

For diabetes diagnosis, many researchers used the Pima Indian Diabetes Dataset. The Pima Indian Diabetes Dataset includes eight parameters. These characteristics include the number of pregnancies, BMI, plasma glucose, diastolic and systolic blood pressure, skinfold thickness, diabetic pedigree function, and Class 0 or 1 diabetes (0 means non-diabetic while 1 means diabetic patient).

# Related Work/Literature Survey

The study's goal was to create an artificial neural network model to figure out which variables were successful and how they affected diabetes. In terms of the evaluation system, these characteristics were grouped as input variables and output variables that reflect various probable levels of disease status. The data was entered into the JNN tool environment, then trained, validated, and tested to achieve an accuracy of 87.3 percent.

Compared the performance of different machine learning approaches and rated the prediction findings based on important risk variables The C4.5 decision tree was found to be superior for data classification in their study.

The prediction model was built and cross-validated using the Pima Indian dataset from UCI. Based on parameters discovered during the early stages of pregnancy, the classification model can predict the occurrence of gestational diabetes. The method was tested on a dataset of 768 samples and was found to be accurate to the tune of 77.8%.

Used a deep learning method to create a massive diabetes forecasting information processing system. The aim was to decrease the number of false positives and negatives as much as possible in order to improve recall and precision. For prediction of diabetes ELM classifier was applied due to their quick capability of learning. With a 98.07 percent accuracy rate, DL was the most efficient and promising of the four suggested classifiers for analyzing diabetes.

This work developed a random forest-based type II diabetes prediction algorithm for assessing several publicly available indicators. Random forest is an ensemble classifier that is made up of many decision trees and has great accuracy and robustness. The data was taken from the University of Virginia. The authors hope to build on other markers for predicting diabetes risk in the future, as well as improve the data mining perspective.

The authors used the Pima Indian Dataset and applied 5 different data mining algorithms to conduct early diabetes prediction. The 5 algorithms and their corresponding accuracies are Gaussian Mixture Model (815 accuracy), Artificial Neural Networks (89 percent accuracy), ELM (82 percent accuracy), Logistic Regression (64 percent accuracy), and Support Vector Machine (74 percent accuracy). The highest level of precision was discovered in ANN.

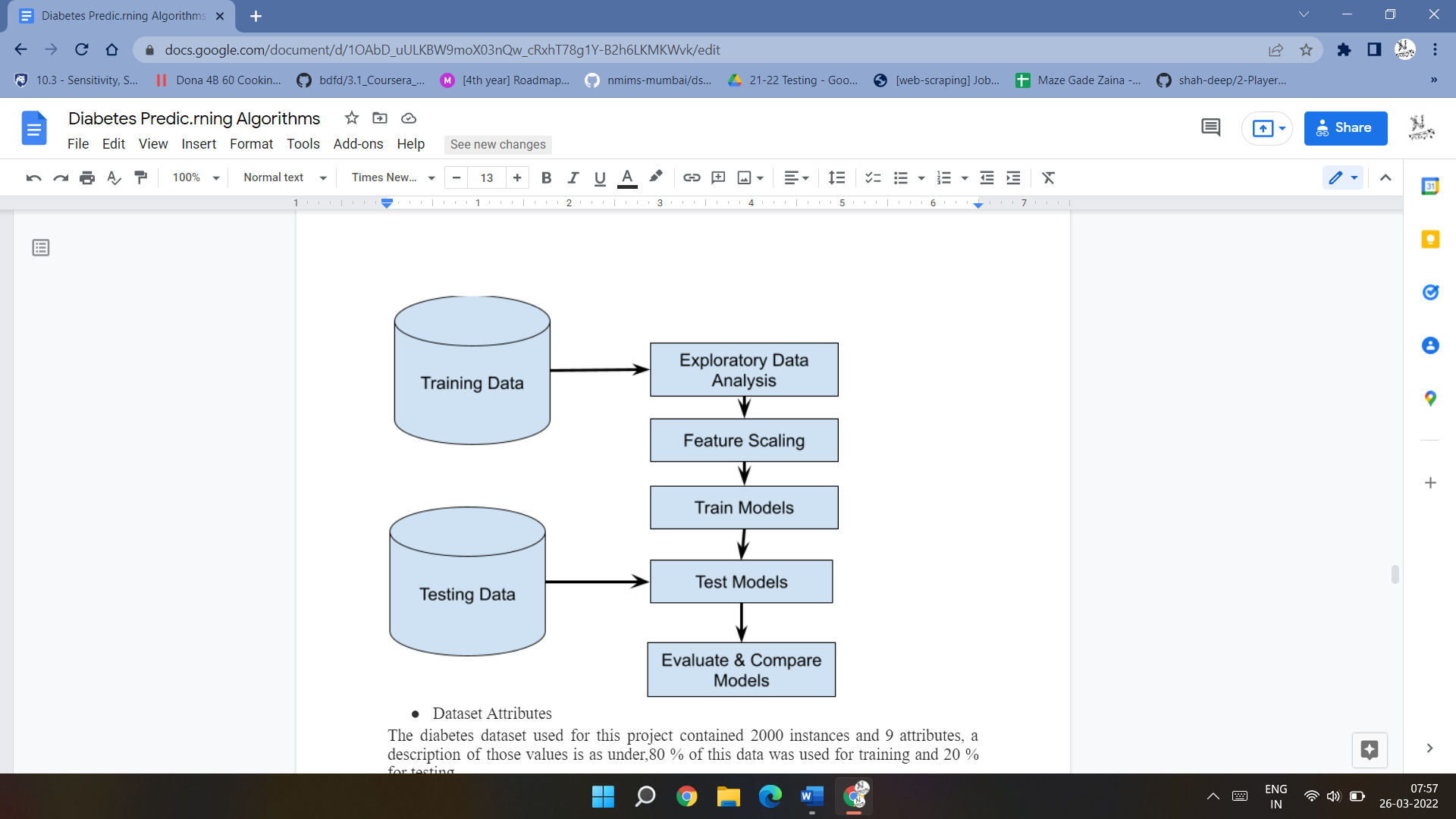
Enthought Canopy, an online tool, was used by the authors of this work. Enthought Canopy is a Python package distribution with core integrated capabilities for application development, iterative data analysis, and data visualisation. KNearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF) are the six algorithms tested, with SVM and KNN providing the greatest accuracy (77%) and SVM and KNN providing the highest accuracy (77%) respectively.

This work provides a new methodology for overcoming the problem of correlation, which makes it difficult for the classification algorithm to detect correlations among the data, by utilising PCA to change the initial collection of features. The PCA application will aid in the filtering of irrelevant features, reducing training time and cost while also improving model performance. Because of k-means' capacity to resolve outliers, the output of PCA analysis is then forwarded for unsupervised clustering using K-means. After cleaning the K-means cluster result, Logistic Regression is used to create a supervised classification model for the dataset. The model's accuracy in terms of performance is 89 percent. According to the authors, accuracy improves even more with a real-time dataset.

In this study, the model is used to understand trends in the Pima Indian diabetes dataset using supervised machine learning techniques in R. They looked at 5 different models for detecting diabetes in female patients, including linear kernel support vector machine (SVM-linear), radial basis kernel support vector machine (SVM-RBF), k-nearest neighbour (k-NN), artificial neural network (ANN), and multifactor dimensionality reduction (MDR). A dataset of female Pima Indian patients with a minimum age of twenty one years was taken from the UCI machine learning repository. In both classification and regression, the support vector machine (SVM) is used. The data points in the SVM model are displayed on a space and categorised into groups and points.

The authors talk about a novel technique based on NDDM to accommodate irregular and sparsely sampled EMRs data with the objective to overcome the analytical problems. The dataset acquired is from CPCSSN. It contains unique data of about 170,000 patient, Spanning 13 years timeframe.  The final dataset comprised 775 (61.03%) female and 1143 (38.96%) male and among them 584 (23.49%) are diabetic patients. The experimental results showed that the INDDM performed better than the NDDM. As a secondary analysis, a logistic regression analysis is also conducted to assess the significant p-value of each risk factor with the incident of T2DM.In this study, a series of experiments are designed to analyse and compare INDDM with standard NDDM for handling irregular and sparsely sampled EMRs data as input for prognostic prediction task. All experiments are conducted using the previously described data obtained from CPCSSN.

# methodology

We used a diabetes dataset with 2000 instances and 9 attributes to train the model, then displayed the data and conducted feature scaling with a Standard Scaler. Exploratory data analysis was performed to better comprehend the data we obtained. Following that, we pre-processed our data, which means we had to clean it by removing duplicate, missing, or strange values and scaling it properly. The next stage was to choose the machine learning models that would be used to train the data. The model's accuracy was then tested using the test dataset. Finally, we compared the models based on several performance indicators such as accuracy, f1-score, recall, and so on.

## Dataset Attributes

The diabetes dataset used for this project contained 2000 instances and 9 attributes, a description of those values is as under, 80 % of this data was used for training and 20 % for testing.

|  |  |  |
| --- | --- | --- |
| Attribute | Range | Description |
| Pregnancies | 0-17 | Number of times pregnant |
| Glucose | 0-199 | Plasma Glucose Concentration |
| Blood Pressure | 0-122 | Diastolic Blood Pressure |
| Skin Thickness | 0-110 | Triceps skin fold thickness |
| Insulin | 0-744 | 2 hour serum Insulin |
| BMI | 0-80 | Body Mass Index = (Weight in Kg)/(height in m)^2 |
| Diabetes Pedigree Function | 0.078-2.42 | Likelihood of diabetes based on family history |
| Age | 21-81 | Age in year |
| Outcome | 0 &1 | 0= health, 1=diagnosed diabetes |

## Algorithms Used:

1. K Nearest Neighbours

The K Nearest Neighbour algorithm is a classification and regression Supervised Learning approach. It is a flexible approach that is frequently used to fill in missing values and resample datasets. As the name implies, it assigns a class to the unknown data point based on the K nearest known data points.

1. Logistic Regression

As the initial model, the Logistic Regression Classifier was utilized. It is similar to the linear regression model in that it computes a weighted sum of the input characteristics, but instead of producing the outcome, it provides the logistic of the result. It computes the probability of a certain outcome depending on individual factors.

1. Decision Tree

Very sophisticated tools can be supported by decision trees. A tree or graph-like structure is built using characteristics such as cost, categorization categories, and effort. The selection is made by moving from root to leaf until all of the requirements are met. Gini indices are used to determine node split. The incorporation of Gini indexes aids in node splitting. The concept of random forest classifier is likewise derived from a random collection of decision trees. These classifiers also aid in the identification of diabetes.

1. Naive Bayes

It is a supervised learning approach founded on the Bayes theorem. It is a type of algorithm in which the value of one property is assumed to be independent (naive) of the value of another feature. It considers the conditional probability, which determines the likelihood of an event occurring if some of the other occurrences have already occurred.

1. Random Forest

This method is one of the simplest and most versatile algorithms for both classification and regression tasks; it use numerous separate decision trees to function as a single one. Each tree classifies the class to which an instance belongs, and the projected class is the class with the most votes.

1. Gradient Boosting

To generate the final predictions, a Gradient Boosting Machine (GBM) integrates the predictions from several decision trees. Keep in mind that in a gradient boosting machine, all of the weak learners are decision trees. Each decision tree's nodes use a distinct group of features to determine the best split. This means that the individual trees aren't all the same, and as a result, they can extract distinct signals from the data.

Furthermore, each new tree considers the errors or faults committed by prior trees. As a result, each succeeding decision tree is constructed on the mistakes of the prior trees. The trees of a gradient boosting machine method are created successively in this manner.

1. XGB

XGBoost is nothing more than an improved version of the GBM algorithm! The operation of XGBoost is the same as that of GBM. XGBoost builds trees consecutively, attempting to repair the mistakes of earlier trees. One of the most fundamental differences is that XGBM uses parallel pre-processing (at the node level), which allows it to be quicker than GBM. In addition, XGBoost incorporates a number of regularization algorithms that prevent overfitting and increase overall performance.

1. LGBM

Because of its speed and efficiency, the LightGBM boosting algorithm is becoming increasingly popular. LightGBM can easily handle large volumes of data, but it does not function well with a limited number of data points. The trees in LightGBM grow leaf-by-leaf rather than level-by-level. Following the first split, the following split is performed exclusively on the leaf node with the highest delta loss.

# Result

 Metrics for Testing Data

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall |
| K Nearest Neighbours | 88.2% | 97% | 87% |
| Logistic Regression | 77.2% | 90% | 79% |
| Decision Tree | 98% | 97% | 100% |
| Naive Bayes | 75.8% | 85% | 80% |
| Random Forest | 98.1% | 97% | 98% |
| XGB | 97.6% | 96% | 99% |
| LGBM | 97.6% | 96% | 99% |
| Gradient Boosting | 83.44% | 81% | 86% |

On Analysis of the Model Accuracies we have found that Random Forest has the highest accuracy of 98.1% followed by Decision Tree at 98% and XGB and LGBM at 97.6%

# Research gaps

Upon an extensive literature survey, we have found that there are various methods to tackle this situation but each of them comes with certain disadvantages or a set of problems. The first challenge is related to data acquisition. One of the major problems while conducting a literature survey was finding articles and papers which were unrelated to the PIMA India dataset. This dataset is well established and provides fairly decent results but there are others that are either too small or inadequate or lack real-time data. Small datasets lead to overfitting on a model, which leads to higher accuracy, but these do not work that well on new testing data. Thus, it is not feasible for real-time implementation. It is important to bring in real and recent patient data for continuous training and optimization of models. The quantity of the data-set should be large enough to train appropriately and predict with higher efficiency. Another problem is feature selection. Some authors neglected a few of the features, while some grouped them for feasible training. It is necessary to select a model which uses features to optimize the performance. Debugging was also a problem because tools like Jupyter Notebooks which divide the code into cells cannot be used on automation batch processes. Authors in some cases required a time-series dataset. It's hard to replicate such work because it's not available on any online resource. For both training and testing, such specialized models demand extensive tuning and large datasets. Another challenge is the construction of an actual model. Many parameters must be adjusted to achieve perfect accuracy. While building a model, factors such as random states, kernel, number of trees, hyper parameter tuning, and others are taken into account. It is really essential to choose the correct algorithm with the appropriate hyper parameters. Some classification models only train on a single parameter, due to which the model’s real-time detection accuracy is reduced. The analysis of these schemes in all cases reveals that the majority of them have either a single data input parameter or a feature selection that is not optimal. Along with such parameter constraints, few classification-based schemes are solely dependent on specific types of hardware devices, making their availability and adaptability more difficult.

# Future scope

On further reading, we inferred that in order to improve the accuracy of the models and provide a better performance, we must work towards feature selection techniques like LDA, SVD, and PCA. This will enable us to better choose important features and will also consequently reduce the training time of the model. The implementation of Neural Networks like ANN, CNN, and RNN in combination along with other algorithms will help achieve higher accuracy in diabetes detection. This is because hybrid schemes play an important role in improving the performance of the models. Patients can be treated more effectively owing to early detection which will help avoid further risks in cases of diabetes. One limitation of this study is that a structured dataset has been utilized for training. However, in the future, unstructured data must also be considered, as this will paint a clearer picture of the accuracy of the model. Patients can be treated more effectively owing to the early detection of diseases which will help avoid further risks. The learnings from diabetes detection models can be applied to other medical domains for prediction, such as for different types of cancer, sleep apnoea, patterns of mental health illnesses, etc. Additionally, attributes like family history, lifestyle habits, etc can also be incorporated into the model.

# Conclusions

In this study various machine learning algorithms are applied on the dataset and the classification is done using various algorithms of which Random Forest gives the highest accuracy of 98.1%. We have seen the comparison of different machine learning algorithms on the data set proving that the model provides an improved and more accurate prediction of diabetes than the existing data set. Further this work can be extended to see how many of non-diabetic people will get diabetes in the next few years.

# References

[1] Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V et al (2020) Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. Diabetes Metab Res Rev 36(2):e3252. <https://doi.org/10.1002/dmrr.3252>

[2] Thenappan S, Rajkumar MV, Manoharan PS (2020) Predicting diabetes mellitus using modified support vector machine with cloud security. IETE J Res. <https://doi.org/10.1080/03772063.2020.1782781>.

[3] Bai BGM, Nalini BM, Majumdar J (2019) Analysis and detection of diabetes using data mining techniques-a big data application in health care. In: Shetty NR, Patnaik LM, Nagaraj HC, Hamsavath PN, Nalini N (eds) Emerging research in computing, information, communication and applications, vol 882. Springer, Singapore

[4]  Islam MT, Raihan M, Farzana F, Aktar N, Ghosh P, Kabiraj S (2020) Typical and non-typical diabetes disease prediction using random forest algorithm. In: Abstracts of the 11th international conference on computing, communication and networking technologies, IEEE, Kharagpur, 1-3 July 2020. <https://doi.org/10.1109/ICCCNT49239.2020.9225430>

[5] Islam MMF, Ferdousi R, Rahman S, Bushra HY (2020) Likelihood prediction of diabetes at early stage using data mining techniques. In: Gupta M, Konar D, Bhattacharyya S, Biswas S (eds) Computer vision and machine intelligence in medical image analysis. Advances in intelligent systems and computing, vol 992. Springer, Singapore, pp 113–125. <https://doi.org/10.1007/978-981-13-8798-2_12>

[6] Daanouni O, Cherradi B, Tmiri A (2019) Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In: Abstracts of the 4th international conference on smart city applications, ACM, Casablanca, 2-4 October 2019. https://doi.org/10.1145/3368756.3369072

[7] Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585.<https://doi.org/10.1016/j.procs.2018.05.122>

[8] El\_Jerjawi, N. S., & Abu-Naser, S. S. (2018). Diabetes prediction using artificial neural network. *International Journal of Advanced Science and Technology*, *121*.

[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] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, *6*(1), 1-19.

[11] Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, *19*(1), 391-403.

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

[13] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. *IEEE Access*, *8*, 76516-76531.

[14] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, *16*, 100204.

[15] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)* (pp. 1-6). IEEE.

[16] 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.

[17] 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*, *17*, 100179.

[18] Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.

[19] Fiarni, C., Sipayung, E. M., & Maemunah, S. (2019). Analysis and prediction of diabetes complication disease using data mining algorithm. *Procedia computer science*, *161*, 449-457.

[20] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017, February). Risk prediction of type II diabetes based on random forest model. In *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)* (pp. 382-386). IEEE.