2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-0009-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESC57686.2023.10193277

An Efficient Diabetes Prediction Model using Machine Learning

Esther Daniel
Department of Computer Science & Engineering
Karunya Institute of Technology & Sciences
Coimbatore, India
estherdaniell@gmail.com

Ujjal Amitya Victor
Department of Computer Science & Engineering
Karunya Institute of Technology & Sciences
Coimbatore, India
ujjalamitya@karunya.edu.in

Steve Abraham Sibby
Department of Computer Science & Engineering
Karunya Institute of Technology & Sciences
Coimbatore, India
steveabraham@karunya.edu.in

Jobin Johnson
Department of Computer Science & Engineering
Karunya Institute of Technology & Sciences
Coimbatore, India
jobinjohnson19@karunya.edu.in

G.V. Aditya
Department of Computer Science & Engineering
Karunya Institute of Technology & Sciences
Coimbatore, India
gvaditya@karunya.edu.in

Abstract— **Diseases are the ones that affect the continual life of human beings. One such illness which affects millions of people worldwide is diabetes. By examining huge datasets of patients, machine learning models offer a viable answer to the problems associated with diabetes detection. The proper management of diabetes depends on early identification This article aims to assess the effectiveness of diverse machine learning methodologies and strategies utilized for forecasting diabetes, utilizing the PIMA Indian Diabetes dataset as a reference. The dataset contains information about crucial factors such as the patient's age, BMI, blood pressure, and blood sugar levels, which are utilized for the analysis. The models which were taken into account in this research were compared to one another and then LightGBM was chosen as the primary model on the basis of its high accuracy. The Hyper parameters were then modified to produce the best performance possible. The results show that machine learning models can accurately detect diabetes and provide insights into the factors that contribute to the disease. This work provides a foundation for future research on diabetes prediction using machine learning models.**

Keywords: Healthcare, Machine Learning classifiers, Diabetes Prediction, Light Gradient Boosting Machine.

I. INTRODUCTION

The issue of diabetes is a significant health concern that affects both developed and developing nations. It is among the most widespread health problems worldwide. [1-3]. From 2000 to 2019, there was a rise of 3% in mortality rates associated with diabetes across different age groups. In the year 2019, diabetes, along with kidney diseases caused by diabetes, was responsible for approximately 2 million deaths worldwide [18]. Diabetes management and the reduction of consequences including heart disease, stroke, and kidney damage depend heavily on early detection and prevention of the condition. A multitude of machine learning models exist for forecasting the likelihood of an individual developing diabetes by analyzing their health information.

These models draw on many different types of data, such as patient-generated health data, electronic health records, and medical imaging. They use sophisticated algorithms to analyze this data, find patterns, and forecast who will be more prone to diabetes.Age, gender, BMI, blood pressure, a family history of diabetes, and glucose levels are some of the important variables that machine learning models consider when predicting diabetes. These models also account for lifestyle elements including dietary and exercise preferences.

A significant advantage of machine learning models designed for diabetes prediction is their ability to recognize individuals who are at a heightened risk of developing the disease. Preventive interventions can reduce the risk of getting the disease, such as dietary changes or medication. Another advantage of utilizing machine learning models to forecast diabetes is its capacity to enhance patient outcomes by facilitating early identification. Identifying and addressing diabetes in its early stages can lower the chances of experiencing adverse effects and improve the well-being of individuals.

To evaluate their effectiveness, the study utilizes commonly used machine learning algorithms based on classification, including Logistic Regression, K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Random Forest, Support Vector Machines (SVM), and Light Gradient Boosting Machine (LightGBM), while employing the training dataset. To attain optimal accuracy, the top-performing model will be selected for hyper-parameter tuning. Its performance will then be evaluated using a validation dataset and compared against the models presented in this study, as well as models utilized in other studies that utilized the same dataset.

II. RELATED WORKS

This section presents the state-of-the-art findings from different studies and analyses conducted on healthcare datasets using various algorithms and techniques.

In [19], authors analyzed the serious condition Diabetes Mellitus and identifies several contributing factors that may result in health problems. Even though existing hospital procedures involve performing diagnostic tests to gather data, the classification and prediction accuracy is subpar. Big Data Analytics provides a remedy by analyzing massive datasets to find hidden patterns and forecast outcomes. The proposed diabetes prediction model has shown enhanced classification accuracy in comparison to previous datasets and combines both irregular and exogenous parameters, such as glucose, BMI, age, and insulin. The article also suggests a pipeline strategy to boost diabetes prediction precision even more.

Hasan et al. [20] concentrates on the challenge of precisely predicting diabetes when faced with outliers or missing values in the dataset, as well as limited labelled data. The suggested framework makes use of several machine learning classifiers, multilayer perceptron, data standardization, feature selection, outlier rejection, filling in missing values, and K-fold cross-validation. Ensembling various classifiers with weights calculated from the AUC metric is advised to increase prediction accuracy. On the Pima Indian Diabetes Dataset, the effectiveness of the suggested framework is assessed, and it is discovered to outperform existing approaches, attaining an AUC of 0.950, a 2% improvement over the results of the most recent studies. The source code for predicting diabetes is accessible to anyone interested.

Sarwar et al. [4] investigate the implementation of machine learning algorithms for anticipatory analytics in healthcare. The study examines six distinct machine learning methods to anticipate diabetes using medical records of patients. The algorithms are contrasted based on performance and accuracy to determine which algorithm is best for predicting diabetes. The major goal is to use machine learning approaches to enable early diabetes prediction, ultimately helping practitioners and healthcare providers.

The effects of diabetes on a global scale as well as the difficulties in making an early prognosis because of the many relationships it has with several other factors. Despite the existence of conventional methods for diabetes diagnosis, data science techniques, particularly machine learning, may help with early prediction with improved accuracy. A system that combines the results of three different supervised machine learning approaches—namely, Support Vector Machine, logistic regression, and Artificial Neural Network—to predict diabetes and provide a useful strategy for early illness diagnosis is critical [5].

Amani Yahyaoui et al. [6] focuses on predicting and detecting diabetes by evaluating the effectiveness of medical Decision Support Systems (DSS) in supporting healthcare professionals in making clinical decisions. The suggested decision support system (DSS) employs Machine Learning (ML) techniques and contrasts conventional machine learning methodologies with deep learning approaches. To achieve this, the research employs Support Vector Machine (SVM) and Random Forest (RF), which are the two most commonly used classifiers in conventional machine learning, and a fully convolutional neural network (CNN) in deep learning. When tested on the Pima Indians Diabetes database, RF outperformed deep learning and ML methods in predicting diabetes. The study described in [7] investigates the escalating incidence of diabetes, which is associated with high blood sugar and obesity. The primary objective is to identify the critical factors that contribute to diabetes and underscore the attributes that are essential for predicting an individual's likelihood of developing the disease. The authors also note that in fields where enormous datasets are available, variable and feature selection represent vital areas of research. Priyanka Sonar et al. [8], focuses on how dangerous diabetes is and how it can cause serious illnesses like heart failure, renal problems, and blindness. In addition to developing machine learning techniques to forecast the risk of diabetes in patients, routine checks are essential. The goal is to construct a system that employs Naive Bayes, Decision Tree, Artificial Neural Network and SVM algorithms to properly forecast a patient's probability of getting diabetes. 85% accuracy is achieved by the Decision Tree model, according to the results, while 77% and 77.3% accuracy are attained by the Naive Bayes and SVM models, respectively. The results imply that machine learning techniques can accurately forecast a patient's probability of developing diabetes.

Quan Zou et al. [9] utilized physical examination data from a Chinese hospital located in Luzhou and established three distinct machine learning models: decision tree, random forest, and neural network. The researchers enhanced the models' performance by employing dimensionality reduction methods like principal component analysis and minimum redundancy maximum relevance. When all characteristics were taken into account, the random forest model generated the most precise outcome of 0.8084 accuracy.

The objective of the authors in [12] was to devise a precise model for predicting the probability of developing diabetes, utilizing four different machine learning classification algorithms on two databases. The study found that the random forest algorithm yielded the highest accuracy of 97.6% on the database obtained from Frankfurt Hospital, Germany, while the SVM algorithm yielded the highest accuracy of 83.1% on the Pima Indian database. Additionally, the authors in [13] explored the applicability of Long Short-Term Memory (LSTM) neural network to predict patient health status, specifically for diabetes, and integrated it into a patient management information system, demonstrating a 6.5% improvement in accuracy compared to traditional MLP and LSTM approaches, making it viable for homecare assistance and de-hospitalization processes. Tushar et al. [16] and Zou et al. [15] also presented a dependable diabetes prediction system that employed a dropout technique to address data overfitting. The proposed neural network surpassed other state-of-the-art approaches in providing better prediction scores for the Pima Indians Diabetes Data Set. The collective findings of these studies accentuate the potential of machine learning and neural network techniques in accurately predicting and managing diabetes.

According to the study, machine learning can be a useful tool for correctly forecasting diabetes, and this strategy may help to combat the disease's rising incidence.

III. PROPOSED METHODOLOGY

The proposed method, as depicted in Figure 1, involves several steps. Initially, we obtain the Pima Dataset as our data source. Subsequently, data is pre-processed to ensure its compatibility and quality for use with machine learning algorithms. Next, we use the K-Fold algorithm to partition the data into separate sets for training and testing. Afterward, we employ grid search to tune the model with the assistance of hyperparameters. Finally, we select the best-performing model for testing and evaluate its performance based on various metrics.

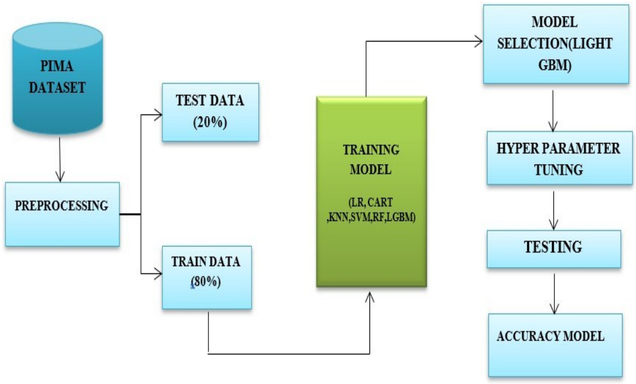

Fig 1: Proposed Architecture for Diabetes Prediction

Data Collection: Collect the diabetes patient data from PIMA Diabetes dataset. The dataset encompasses various parameters, including pregnancies, glucose levels, blood pressure readings, skin thickness, insulin measurements, BMI values, diabetes pedigree function scores, age, and outcome.

Table 1. Dataset Information

Column Name	Data Type
Pregnancies	Int64
Glucose	Int64
Blood Pressure	Int64
Skin Thickness	Int64
Insulin	Int64
BMI	float64
Diabetes Pedigree Function	float64
Age	Int64
Outcome	Int64

Data Preprocessing: Missing values were filled. Outliers were identified by comparing observations to the 25% and 75% quartiles. Insulin variable underwent standalone review, suppressing contradictory values. Local Outlier Factor (LOF) method was used to detect outliers across all variables, and those exceeding the threshold were deleted. To determine the most crucial factors that influence the prediction of diabetes, do feature selection. To make sure that every variable has the same scale, normalize the data.

Figure 2 represents the frequency of each unique output value in a dataset. It is useful for analyzing the distribution of outputs and identifying any imbalances or biases in the data and this information can inform model selection and training strategies.

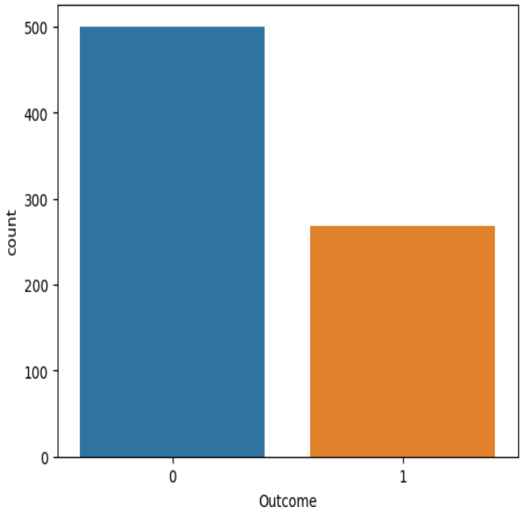

Fig 2: Outcome count plot

The degree of correlation between an attribute and the target variable determines the extent to which that attribute influences the outcome. To visualize this correlation, a correlation matrix is used, which indicates the correlation value of each attribute with the target variable. Figure 3 presents a visualization of this correlation matrix.

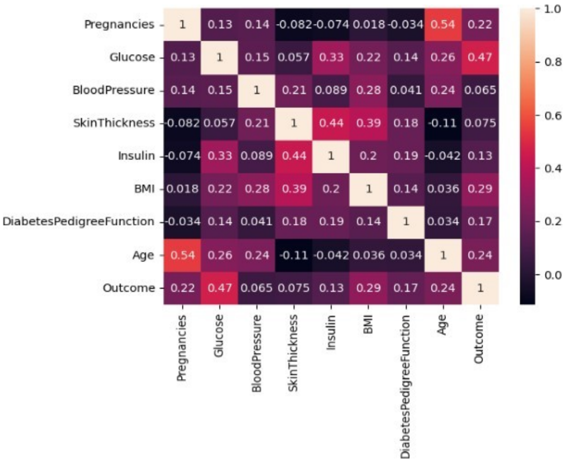

Fig3: Correlation matrix

A confusion matrix is depicted in Figure 4, which presents a structured display of the true positives, true negatives, false positives, and false negatives. This visual representation is an effective tool for evaluating the precision of a model's predictions and calculating key performance metrics like recall, precision, sensitivity, and F-1 score.

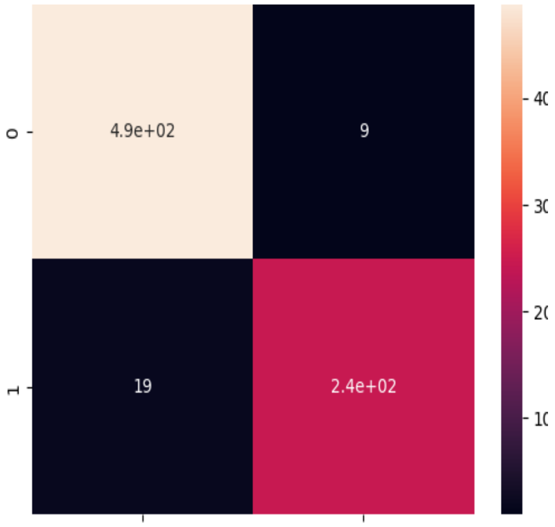

Fig 4: Confusion Matrix

K-fold cross-validation: A popular machine learning technique for model selection and evaluation is cross-validation. The method entails training the model K times on K equal-sized subsets, or "folds," of the dataset. During the K-fold cross-validation process, one of the folds is selected as the validation set, while the remaining folds are employed as the training set in each iteration. Throughout the K iterations, each fold is used as the validation set once. By exposing the model to diverse subsets of data, the use of cross-validation diminishes the possibility of overfitting. The value of K, which is typically set to 5 or 10, may be adjusted depending on factors such as the dataset's size and the model's complexity.

Data Splitting: Once the data has been pre-processed, it is segregated into three distinct sets, namely the training set, validation set, and testing set. The dataset is divided into 80% for training and 20% for validation and testing. The validation set plays a crucial role in fine-tuning the hyper-parameters, while the training set is employed to train the machine learning models. Finally, the testing set is utilized to gauge the model's efficacy.

Model Selection: For predicting diabetes, we are employing machine learning techniques. LR,KNN,CART,SVM,RF and LightGBM are some of the algorithms used.

Algorithm:

Input: A dataset with n samples and m features is given, where $(x1, y1)$, $(x2, y2)$, . . . , (xn, yn) are the feature vectors for the individual samples, and (yi) is the target variable.

Preprocessing the data: The data preprocessing can be represented as a set of operations applied to a dataset D to obtain a preprocessed dataset

D': D' = f(D) where f() represents a series of data transformation functions that can include:

Handling missing values: D' = g(D)
Handling outliers: D'' = h(D')
Feature selection: D''' = j(D'')

The steps for implementing LightGBM model:
1. Set the value of the prediction function (x) to the target variable's mean as its initial value.

2. Calculate the loss function's negative gradient $(-gi)$ with respect to the current prediction function $(f\_t - 1, x)$ for each sample i:

$$gi = \frac{\partial(yi,\ f_{\{t-1\}(xi)})}{\partial f\_\{t-1\}(xi)}$$

3. Create a regression tree $h(x)$ using the feature vectors xi to fit the negative gradients $-gi$. The tree is trained to reduce the mean squared gradient error that is negative:

$$ht(x) = argmin\ \ \{h\} \sum_{n}^{i=0}(gi - h(xi))^{2}$$

4. Calculate t using line search or another optimization technique to determine the ideal step size. To reduce the loss function for the current iteration, the step size is selected:

$$\lambda t = argmi\{\lambda\}L\ (y, f_{\{t-1\}(x)} + \lambda\, ht(x))$$

5. Update prediction function:

$$f\_\{t\}(x) = f\_\{t-1\}(x) + \lambda t\ ht(x)$$

6. Output the final prediction function $fT(x)$

Evaluate the model using test data:

$$Accuracy = \frac{Number\ of\ correct\ Prediction}{Total\ number\ of\ predictions\ made}$$

Output: The model with tuned parameters has been built and validated.
We have carried out the implementation of CART, Logistic Regression, SVM, KNN, and Random Forest in a similar manner.

Hyper-parameter Tuning: Fine-tune the selected models by adjusting their hyper-parameters using techniques like grid search. The aim is to find the best hyper-parameters that yield the highest performance on the validation set.

Grid search: It is a machine learning technique used to optimize hyper parameters by searching through a pre-defined grid of possible hyper parameter values. Hyper parameters are set before the training process which includes things like learning rate and regularization. Grid search trains and evaluates models for every combination of hyper parameters using cross-validation. Then it selects the optimal set of hyper parameters that produce the best performance on the validation set. While grid search can be computationally expensive, it effectively finds the best hyper parameters and improves model performance.

IV. EXPEREMENTAL ANALYSIS

To enhance the accuracy of the diabetes prediction model, a methodology was proposed that involved utilizing machine learning techniques and carrying out data preprocessing on the PIMA dataset, which is commonly used for diabetes prediction. The model was trained on a representative sample of the dataset, which contributed to achieving a high level of performance. The results were analyzed to assess the effectiveness of the approach. The data preprocessing and the use of a generalized sample of the dataset contributed to the model's- high level of accuracy, which is crucial for accurate diabetes prediction.

Table 2. Accuracy of ML models for Diabetes Prediction

Research Work	Method	Accuracy (%)
Edeh et al. [12]	SVM	83.1
Massaro et al. [13]	LSTM-AR	89
Zou et al. [15]	mRMR-RF	77.21
Tushar et al. [16]	Deep MLP	88.41
Proposed Method	LightGBM	90.2

The proposed model outperformed all other models compared in this study, achieving an impressive accuracy of 90.2%. This accuracy level was the highest among all other works used in the comparison. The performance of the model was analyzed against various other models used in different studies, and the comparison was visualized in Table 2.

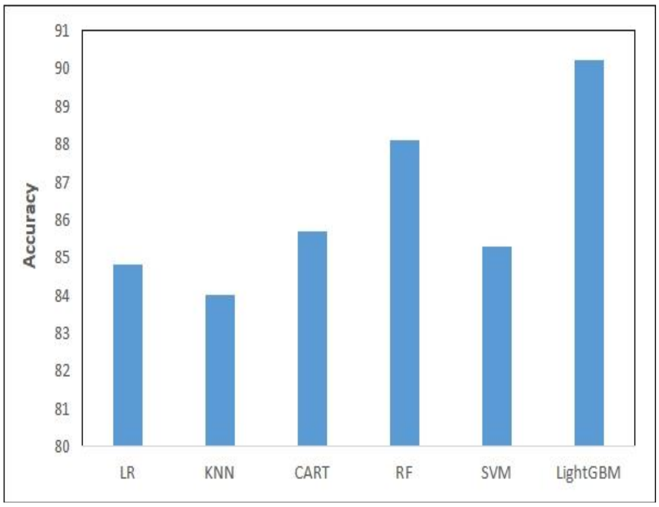

Fig 5: Accuracy of Algorithms after hyper parameter tuning

From various state-of-art methodologies the LightGBM model, which has not previously been investigated for diabetes prediction, in contrast to earlier studies that employed standard binary classification techniques. Fig 5 represents the proposed strategy outperformed previous models in terms of accuracy by optimizing the model's hyperparameters.

V. CONCLUSION

Various machine learning algorithms for classification were analysed. The accuracy of each algorithm was evaluated, with Logistic Regression, KNN, CART, Random Forest, SVM, and LightGBM achieving accuracy rates of 84.8%, 84%, 85.7%, 88.1%, 85.3%, and 88.2%, respectively. After conducting hyper-parameter tuning, the LightGBM model was chosen and further improved the accuracy to 90.2%. According to this study, the implementation of the LightGBM model resulted in a notable improvement in the accuracy and precision of predicting diabetes compared to previously available datasets. Future studies could explore the probability of non-diabetic individuals developing diabetes in the future.

VI. REFERENCES

[1] A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza-Albarrán, and K. L. Ramaiya, “Diabetes in developing countries," J. Diabetes, vol. 11, no. 7, pp. 522_539, Mar. 2019.

[2] Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou."Predicting the Onset of Diabetes with Machine Learning Methods." Journal of Personalized Medicine 13, no. 3 (2023): 406.

[3] Febrian, Muhammad Exell, Fransiskus Xaverius Ferdinan,Gustian Paul Sendani, Kristien Margi Suryanigrum, and Rezki Yunanda. "Diabetes prediction using supervised machine learning." Procedia Computer Science 216 (2023): 21-30.

[4] Sarwar, Muhammad Azeem, Nasir Kamal, Wajeeha Hamid,and Munam Ali Shah. "Prediction of diabetes using machine learning algorithms in healthcare." In 2018 24th international conference on automation and computing (ICAC), pp. 1-6. IEEE, 2018.

[5] Joshi, Tejas N., and P. P. M. Chawan. "Diabetes prediction using machine learning techniques." Ijera 8, no. 1 (2018): 9- 13.

[6] Yahyaoui, Amani, Akhtar Jamil, Jawad Rasheed, and MirsatYesiltepe. "A decision support system for diabetes prediction using machine learning and deep learning techniques." In 2019 1st International informatics and software engineering conference (UBMYK), pp. 1-4. IEEE, 2019.

[7] Dutta, Debadri, Debpriyo Paul, and ParthajeetGhosh."Analysing feature importances for diabetes prediction using machine learning." In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 924-928. IEEE, 2018.

[8] Sonar, Priyanka, and K. JayaMalini. "Diabetes prediction using different machine learning approaches." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 367-371. IEEE, 2019.

[9] Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, YingJu,and Hua Tang. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[10] Zolfaghari, Rahmat. "Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm." Int. J. Comput. Eng. Manag 15 (2012): 2230-7893.

[11] Sneha, N., and Tarun Gangil. "Analysis of diabetesmellitusfor early prediction using optimal features selection." Journal of Big data 6, no. 1 (2019): 1-19.

[12] Edeh, Michael Onyema, Osamah Ibrahim Khalaf, Carlos Andrés Tavera, Sofiane Tayeb, Samir Ghouali, Ghaida Muttashar Abdulsahib, Nneka Ernestina Richard-Nnabu, and AbdRahmane Louni. "A classification algorithm-based hybrid diabetes prediction model." Frontiers in Public Health 10 (2022).

[13] Massaro, Alessandro, Vincenzo Maritati, Daniele Giannone, Daniele Convertini, and Angelo Galiano. "LSTM DSS automatism and dataset optimization for diabetes prediction." Applied Sciences 9, no. 17 (2019): 3532.

[14] Dadgar, Seyyed Mohammad Hossein, and Mostafa Kaardaan. "A hybrid method of feature selection and neural network with genetic algorithm to predict diabetes." International Journal of Mechatronics, Electrical and Computer Technology (IJMEC) 7, no. 24 (2017): 3397-3404.

[15] Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[16] Ashiquzzaman, Akm, Abdul Kawsar Tushar, Md Rashedul Islam, Dongkoo Shon, Kichang Im, Jeong-Ho Park, Dong-Sun Lim, and Jongmyon Kim. "Reduction of overfitting in diabetes prediction using deep learning neural network." In

IT Convergence and Security 2017: Volume 1, pp. 35-43. Springer Singapore, 2018.

[17] Kalagotla, Satish Kumar, Suryakanth V. Gangashetty, and Kanuri Giridhar. "A novel stacking technique for prediction of diabetes." Computers in Biology and Medicine 135 (2021): 104554.

[18] https://www.who.int/news-room/fact-sheets/detail/diabetes

[19] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

[20] Hasan, Md Kamrul, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.