Diabetics Prediction using ML

INT247

A Project report

Submitted in partial fulfillment of the requirements for the award of degree of

B.tech (Computer Science)

Submited To

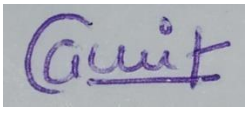
Sagar Pande

LOVELY PROFESSIONAL UNIVERSITY PHAGWARA,PUNJAB

From 15/02/22 to 30/03/22

Amit Kumar Ashutosh

11906959



Student Declaration

To whom so ever it may concern

I, Amit Kumar Ashutosh, 11906959, hereby declare that the work done by me on "Diabetics Prediction" from February 2022 to March, 2022, is a record of original work for the partial fulfillment of the requirements for the award of the degree, B.tech CSE.

Name-Amit Kumar Ashutosh

Registration Number-11906959

Dated: 30 March 2022

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people.All that I have done due to such supervision and assistance and I would not forget to thank them.

I respect and thank to my my university for providing me an opportunity to do the Community Development Project and giving me all support and guidance which made me complete the project.

I am thankful to and fortunate enough to get constant encouragement, support and guidance all Teaching staffs of Mittal School Of Lovely Professional University which helped me in successful completing our project work.

Secondly i would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

Index

Abstract

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

List of Figures

List of Tables

Introduction

Public health is a fundamental concern for protecting and preventing the community from health hazard diseases. Governments are spending a considerable amount of their gross domestic product (GDP) for the welfare of the public, and initiatives such as vaccination have prolonged the life expectancy of people. However, for the last many years, there has been a considerable emergence of chronic and genetic diseases affecting public health. Diabetes mellitus is one of the extremely life-threatening diseases because it contributes to other lethal diseases, i.e., heart, kidney, and nerve damage.

Diabetes is a metabolic disorder that impairs an individual's body to process blood glucose, known as blood sugar. This disease is characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. An absolute deficiency of insulin secretion causes type 1 diabetes (T1D). Diabetes drastically spreads due to the patient's inability to use the produced insulin. It is called type 2 diabetes (T2D) [4]. Both types are increasing rapidly, but the ratio of increase in T2D is higher than T1D. 90 to 95% of cases of diabetes are of T2D.

Inadequate supervision of diabetes causes stroke, hypertension, and cardiovascular diseases. To avoid and reduce the complications due to diabetes, a monitoring method of BG level plays a prominent role .A combination of biosensors and advanced information and communication technology (ICT) provides an efficient real-time monitoring management system for the health condition of diabetic patients by using SMBG (self-monitoring of blood glucose) portable device. A patient can check the changes in glucose level in his blood by himself . Users can better understand BG changes by using CGM (continuous glucose monitoring) sensors .

By exploiting the advantages of the advancement in modern sensor technology, IoT, and machine learning techniques, we have proposed an approach for the classification, early-stage identification, and prediction of diabetes in this paper. The primary objective of this study is twofold. First, to classify diabetes into predefined categories, we have employed three widely used classifiers, i.e., random forest, multilayer perceptron, and logistic regression. Second, for the predictive analysis of diabetes, long short-term memory (LSTM), moving averages (MA), and linear regression (LR) are used. To demonstrate the

effectiveness of the proposed approach, DROPBOX Indian Diabetes is used for experimental evaluation. We concluded that, in experimental evaluation, MLP achieved an accuracy of 86.083% in diabetes classification as compared to the other classifiers and LSTM achieved a prediction accuracy of 87.26% for the prediction of diabetes. Moreover, we have also performed a comparative analysis of the proposed approach with existing state-of-the-art approaches. The accuracy results of our proposed approach demonstrate its adaptability in many healthcare applications.

Besides, we have also presented the IoT-based hypothetical diabetes self-monitoring system that uses BLE (Bluetooth Low Energy) devices and data processing in real-time. The latter technique used two applications: Apache Kafka (for streaming messages and data) and MongoDB (to store data). By utilizing BLE-based sensors, one can collect essential sign data about weight and blood glucose. These data will be handled by data processing techniques in a real-time environment. A BLE device will receive all the data produced by sensors and other necessary information about the patient that resides in the user application, installed on the cell phone. The raw data produced by sensors will be processed using the proposed approach to produce results, suggestions, and treatment from the patient's server-side.

The rest of the paper is organized as follows. In Section 2, the paper presents the motivations for the proposed system by reviewing state-of-the-art techniques and their shortcomings. It covers the literature review about classification, prediction, and IoT-based techniques for healthcare. Section 3 highlights the role of physical activity in diabetes prevention and control. In Section 4, we proposed the design and architecture of the diabetes classification and prediction systems. Section 5 discusses the results and performance of the proposed approach with state-of-the-art techniques. In Section 6, an IoT-based hypothetical system is presented for real-time monitoring of diabetes. Finally, the paper is concluded in Section 7, outlining the future research directions.

The disease “Diabetes Mellitus” is one of the most common critical diseases in the world. According to the World Health Organization (WHO), diabetes affects 8.5% of adults over the age of 18 and is responsible for 1.6 million deaths worldwide. Although the rate of diabetes-related premature death in many developing countries fell from 2000 to 2010, the statistics again increased between 2010 and 2016. The four primary diseases, namely cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes, kill over 18% of people worldwide and have become a serious public health concern. For example, in 2000, deaths from diabetes climbed by 70%, and in 2020, mortality among males are expected to grow to 80%. Diabetes mellitus can result from obesity, age, lack of exercise, lifestyle, hereditary diabetes, high blood pressure, poor diet, etc. Over time, people with diabetes have a high risk of diseases such as heart disease, stroke, kidney failure, nerve damage, eye issues, etc.

The present clinical practice consists in collecting the data necessary to detect diabetes through a number of tests and then providing an appropriate diagnostic drug. In the healthcare sector (, supervised and non-supervised machine learning (ML) approaches are utilized to diagnose various kinds of diseases. ML methods enable researchers to investigate hidden patterns of medical datasets for predicting expected outcomes and reducing costs of identifying complex diseases. ML algorithms are trained with real medical datasets having different features and external variables.

If diabetes can be discovered at the primary stage, harmful effects can be avoided with adequate medical care. ML approaches can aid in early detection of this disease. Machine learning algorithms are adopted to create a prediction model since ML methods allow computers to learn and gain intelligence from previous experience or a pre-defined dataset. The predictive model can identify and understand the incoming data, allowing it to make more precise decisions.

Researchers such as , designed and proposed different kinds of machine learning based models to diagnose diabetes disease. However, only a few studies have concentrated on integrating the trained model into a user's app and designing a user interface so that consumers may monitor their health status on their smartphones. Furthermore, those models were trained using only one or two datasets, which does not guarantee that the

model would perform as expected in real-world scenarios. We have focused on addressing these gaps in our proposed machine learning model.

This work aims to apply machine learning algorithms to present the analytical results regarding physical components and circumstances that contribute to the development of diabetes in human body. The pre-processing on the datasets provides better accuracy of the model than the existing research works. The correlation based feature selection method discovers necessary attributes in the dataset. The classifier showing better performance for all the datasets is integrated in a web application. Considering the current research gap, we set the following research questions. *How does the accuracy of machine learning algorithms vary in predicting diabetes for different datasets including DROPBOX Indian dataset? What factors influence the most in detecting diabetics? How can an effective machine learning algorithm be discovered to integrate that in a web application?*

To predict diabetes of individuals, we have adopted several machine learning algorithms and evaluated their performances. We have examined the performance of seven different models, namely Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Gradient Boosting (GB) and k-nearest neighbor (k-NN) algorithm to develop a predictive model. Two datasets with various attributes including glucose level, insulin level, blood pressure, BMI and age are used for training the ML algorithms. Results of the algorithms are evaluated based on several performance metrics. Finally, based on the accuracy level, a web application is developed to predict the diabetes of any individual. Any users can get the prediction of diabetes by using this application on their smartphones or computers.

This Figure shows each phase of the proposed ML based diabetes prediction model. In the first phase, every dataset is pre-processed. In the second stage, the pre-processed datasets are feed into the different machine learning algorithms. In the third phase, the output of the models is then analyzed using various metrics. In the later phase, the model that provides the highest accuracy is adopted to detect diabetes of any individual and integrated with a web-based application. This web-based application is developed using Flask of python programming language.

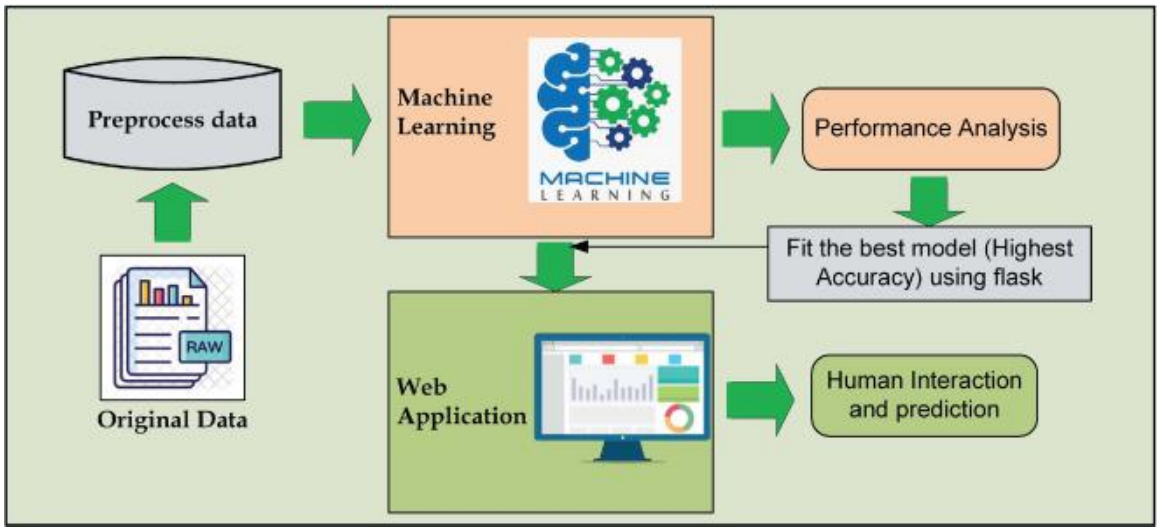


Fig. 1. Overview of the proposal

In a nutshell, the contributions of this research are listed as follows:

- Our first contribution is to train several machine learning algorithms using four different clinical datasets for detecting diabetes. All the datasets are pre-processed by applying different pre-processing techniques.

- Secondly, the performances of each ML algorithms with four datasets are analyzed with respect to several parameters like precision, Recall, f1-score, ROC curve and accuracy. Further, we have identified several important features or attributes using different feature selection methods such as correlation, chi-square, etc. The feature selection methods find out the mostly correlated features to diabetes disease. The performances of the ML algorithms were also analyzed on the reduced set of attributes.

•Thirdly, based on the performance results, web-based application is developed to predict the diabetes of individuals.

Literature Review

In this section, we discussed the classification and prediction algorithms for diabetes prediction in healthcare. Particularly, the significance of BLE-based sensors and machine learning algorithms is highlighted for self-monitoring of diabetes mellitus in healthcare. Machine learning plays an essential part in the healthcare industry by providing ease to healthcare professionals to analyze and diagnose medical data. Moreover, intelligent healthcare systems are providing real-time clinical care to needy patients The features covered in this study are compared with the state-of-the-art studies.

In most of the reseach works, Dropbox Indians Diabetes Dataset (PIDD) have been used by many researchers for diabetes prediction. Various supervised machine learning algorithms were used to predict diabetes. Radial basis function (RBF) kernel SVM, artificial neural network (ANN), multifactor dimensionality reduction (MDR), linear SVM and k-NN are some of them to mention. Based on p value and odds ratio (OR), Logistic Regression (LR) has been used to recognize the risk factors for diabetes. Four classifiers have been adopted to predict diabetic patients, such as NB, DT, Adaboost, and RF. Partition protocols like- K2, K5, and K10 were also adopted, repeating these protocols into 20 trails. For the performance measurement of the classifiers, accuracy (ACC) and area under the curve (AUC) were analyzed.
showed a comparison of widely utilized regression models such as Glmnet, RF, XGBoost, LightGBM for predicting type 2 diabetes mellitus. The goal of this work was to examine if innovative machine learning methodologies gave any advantages in early prediction of impaired fast glucose and fasting plasma glucose (FPGL) levels compared to classic regression techniques.
For the prediction of diabetic patients, have chosen four classifications such as naive bays (NB), decision tree (DT), adaboost and random forest. These methods were also implemented by three types of partition protocols (K2, K5, and K10). These classifers' performances are measured with precision (ACC) and curve surface (AUC).
A hybrid model to detect type 2 diabetes was suggested by. In order to extract unknown, hidden property from the dataset and to obtain more exact results, we use K-mean

clustering, which is followed by the execution of a Random Forest and XGBoost classifier.

suggested a Machine Learning Techniques (ML) DSS for anticipating diabetes. They compared traditional machine learning with approaches to the deep learning. The authors applied the classifiers most typically used for a standard machine learning method: SVM and the Random Forest (RF). In contrast, they used a full-scale neural network (CNN) for Deep Learning (DL) to forecast and identify patients who suffer from diabetes.

predicted diabetes using the decision tree, random forests, and neural network. The dataset is collected from the Luzhou physical exams in China. The PCA was applied to reduce the dimension of the dataset. They selected several ML approaches to execute independent test to verify the universal applicability of method.

Supervised machine learning models which explore data-driven approaches were used to identify patients with diabetes diseases . A complete research was conducted based on the National Health and Nutrition Examination Survey (NHANES) dataset. To develop models for cardiovascular, prediabetes, and diabetes detection, they have used all available feature variables within the data. Using various time frames and set of features within the data, different machine learning models, namely Support Vector Machines, logistic regression, gradient boosting and random forest were evaluated for the classification.

In the authors used NBs for the classification on all the attributes. Afterwards GA was used as an attribute selector and NBs used the selected attributes for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis. Three specific supervised machine learning methods are used by , namely SVM, Logistic regression and ANN. His goal for research was to predict diabetes patients and he has also proposed an effective model for the prior detection of diabetes disease. focused on machine learning classification algorithms for predicting diabetes disease with more accuracy. Their study in SVM classification algorithm achieved highest accuracy. Various measures have been used to calculate the performance of classification algorithms.

An intelligent model using machine learning practices is developed to identify diabetes disease. This model is constructed using approaches like clustering, removal of noise and classification, each of which made use of SOM, PCA and NN, respectively.
The adaboost and bagging ensemble techniques are used to detect diabetes. Along with standalone data mining technique, a base learner is used to identify patients with diabetes mellitus, namely J48 (c4.5) decision tree that makes use of multiple diabetes risk factors. In the Canadian Primary Care Sentinel Surveillance Network, three different ordinal adult groups are selected for classification. Experimental result shows that, the adaboost ensemble method shows better performance than both bagging and standalone J48 decision tree. For diagnosing T2DM, has taken in consideration four different classification models, namely SVM, K-NN, ANN and LR. A comparison is done among these algorithms to measure the diagnostic power of this algorithms. The algorithms are performed on six LncRNA variables and demographic data.

Overview

Real-time diabetes prediction is a complicated task. The emerging use of sensors in healthcare paved the path to handle fatal diseases . Several techniques have been presented in the literature to classify and predict diabetes. Acciaroli et al. exposed two accurate meters to measure diabetes in blood with less error rate. Furthermore, these commercial versions of glucometers are Accu-Chek with 6.5% error and CareSens with 4.0% error. Described the accuracy link of CGM with the calibration sensor. Alfian et al. [27] uncovered that the FDA had accepted CGM sensors for monitoring glucose in different trends and patterns. Moreover, at one particular time, one glucose reading should not be used to analyze the amount of insulin as not accepted in a glucometer. Rodríguez et al. [28] proposed a structural design containing a local gateway as a smartphone, cloud system, and sensors for advanced management of diabetes.

Filippoupolitis et al. [29] planned action to acknowledge a system using Bluetooth Low Energy (BLE) beacons and smartwatches. Mokhtari et al. considered technologies working with BLE for activity labeling and resident localization [30]. Gentili et al. [31] have used BLE with another application called Blue Voice, which can reveal the probability of multimedia communication of sensor devices and speech streaming service. Suárez et al. [32] projected a monitoring system based on the BLE device for air quality exposure with the environmental application. It aims at defining potential policy responses and studies the variables that are interrelated between societal level factors and diabetes prevalence [33, 34].

Wang et al. [39] have given a general idea of the up-to-date BLE technology for healthcare systems based on a wearable sensor. They suggested that low-powered communication sensor technologies such as a BLE device can make it feasible for wearable systems of healthcare because it can be used without location constraints and is light in weight. Moreover, BLE is the first wireless technology in communication for healthcare devices in the form of a wearable device that meets expected operating requirements with low power, communication with cellular directly, secure data transmission, interoperability, electronic compatibility, and Internet communications. Rachim and Chung [40] have suggested one transmission system that used deficient

power to observe the heart's activity through electrocardiograph signals using a BLE device for data transmission collecting by armband sensors and smartphones.

Mora et al. projected a dispersed structure using the IoT model to check human biomedically generated signals in reports using a BLE sensor device [41]. Cappon et al. [42] explored the study of CGM wearable sensors' prototypes and features of the commercial version currently used. Årsand et al. [43] offered the easiest method for monitoring blood glucose, physical activity, insulin injections, and nutritional information using smartphones and smartwatches. Morón et al. [44] observed the performance of the smartphone used in the medical field. Lee and Yoo [45] anticipated a structure using PDA (personal digital assistant) to manage diabetic patient's conditions better. It can also be used to send information about blood pressure, BG level, food consumption, and exercise plan of a patient with diabetes and give the direction of treatment by monitoring physical activity, food consumption, and insulin prescribed amount.

Rodríguez et al. [28] suggested an application for the smartphone, which can be used to receive the data from the sensor using a glucometer automatically. Rodríguez-Rodríguez et al. [46] suggested that checking the patient's glucose level and heart rate using sensors will produce colossal data, and analysis on big data can be used to solve this problem.

Generally, physical activity is the first prevention and control strategy suggested by healthcare professionals to diabetic or prediabetic patients [47]. Among diet and medicine, exercise is a fundamental component in diabetes, cardiovascular disease, obesity, and lifestyle rescue programs. Nonetheless, dealing with all the fatal diseases has a significant economic burden. However, diabetes mellitus emerged as a devastating problem for the health sector and economy of a country of this century.

Recently, the international diabetes prevention and control federation predicts that diabetes can affect more than 366 million people worldwide [49]. The disease control and prevention center in the US alarmed the government that diabetes can affect more than 29 million people [50]. While these alarming numbers are continuously increasing, they will burden the economy around the globe. Therefore, researchers and healthcare professionals worldwide are researching and proposing guidelines to prevent and control

this life-threatening disease. Sato [51] presented a thorough survey on the importance of exercise prescription for diabetes patients in Japan. He suggested that prolonged sitting should be avoided and physical activity should be performed every 30 minutes. Kirwan et al. [47] emphasized regular exercise to control and prevent type 2 diabetes. Particularly, they studied the metabolic effect on tissues of diabetic patients and found very significant improvements in individuals performing regular exercise. Moser et al. [48] have also highlighted the significance of regular exercise in improving the functionality of various organs of the body, as shown in Figure

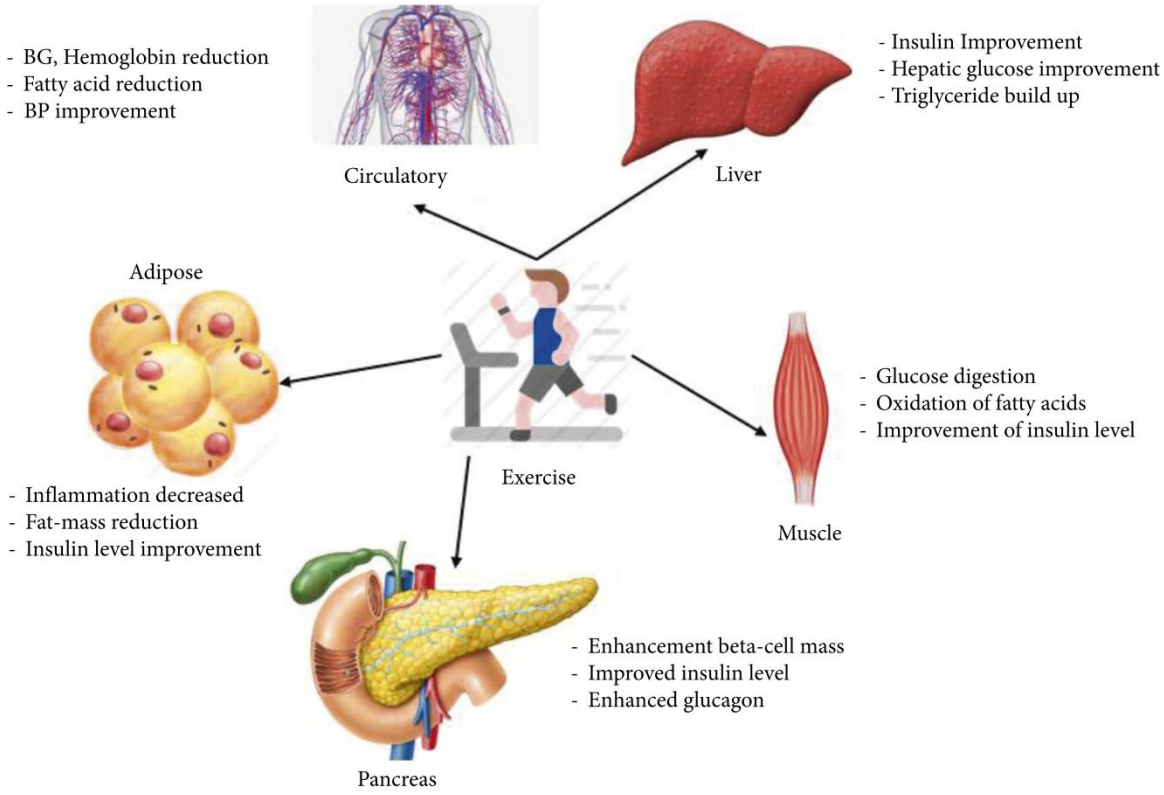


Fig 2. Impact of regular exercise on metabolism of diabetic patients

Circuit Description

Fig 3 depicts the proposed framework for diabetes prediction. Firstly, we pre-process two separate datasets. In the pre-processing stage, correlation between attributes of the datasets is analyzed for finding useful features in detecting diabetes. After that, the data is divided into two sets: training and testing. The training set is utilized to develop predictive ML models using a variety of machine learning algorithms. Next, we assess the proposal's performance with respect to different metrics. Finally, the best ML model is deployed in a web application using flask. Following this, we describe the workflow of each part briefly:

1.Data Collection: We collected two alternative datasets, each with a different number of factors or features, to ensure the model's robustness. The datasets were compiled from a wide variety of sources, including diabetes statistics and health characteristics obtained from people around the world and from various health institutes.

2.Data Analysis and Data Preprocessing: Several pre-processing techniques are applied on the datasets before feeding these datasets into the machine learning model so that the performance of the model is improved. The pre-processing tasks include removing outliers and dealing with missing values, data standardization, encoding, and so on.

- Outliers Removal - Attributes' values that are beyond acceptable boundaries and have high variation from the rest of the respective attribute's value might be present in the dataset. Such attributes' value might degrade the machine learning algorithm's performance. To eliminate such outliers, we applied the IQR (Inter-quartile Range) approach.

- Missing value Handling - To improve model performance, the mean value of each attribute was employed for handling the missing values.

•Label Encoding - Label encoding is the process of converting the labels of text/categorical values into a numerical format that ML algorithms can interpret. For example, the categorical values of Junkfood consumption status yes to ‘1′ and No to ‘0′ have been converted.

3.Model Construction and Prediction: To construct the predictive model, 80% of the pre-processed data has been used for training while the remaining 20% data is used for the testing purpose.

4.Performance Analysis: We have analyzed the results of the proposed model in terms of several performance metrics. The algorithm that provides highest prediction accuracy is selected as the best algorithm for the web application development.

5.Performance Comparison: In this step, the accuracy of the proposal has been compared with some recent works related to diabetes prediction. The performance results indicate that the proposal can improve the performance compared to the recent related research.

6.Web Application development: To develop a smart web application, we have used the Flask micro-framework and integrated the best model. To predict diabetes, a user is required to submit a form with necessary numbers of diabetes related parameters. The application uploaded in a server predicts the results using the adopted machine learning model. We describe the adopted machine learning algorithms in the following sections.

Software Tool

The proposed diabetes classification and prediction system has exploited different machine learning algorithms. First, to classify diabetes, we utilized logistic regression, random forest, and MLP. Notably, we fine-tuned MLP for classification due to its promising performance in healthcare, specifically in diabetes prediction. The proposed MLP architecture and algorithm are shown in Figure and Algorithm, respectively.

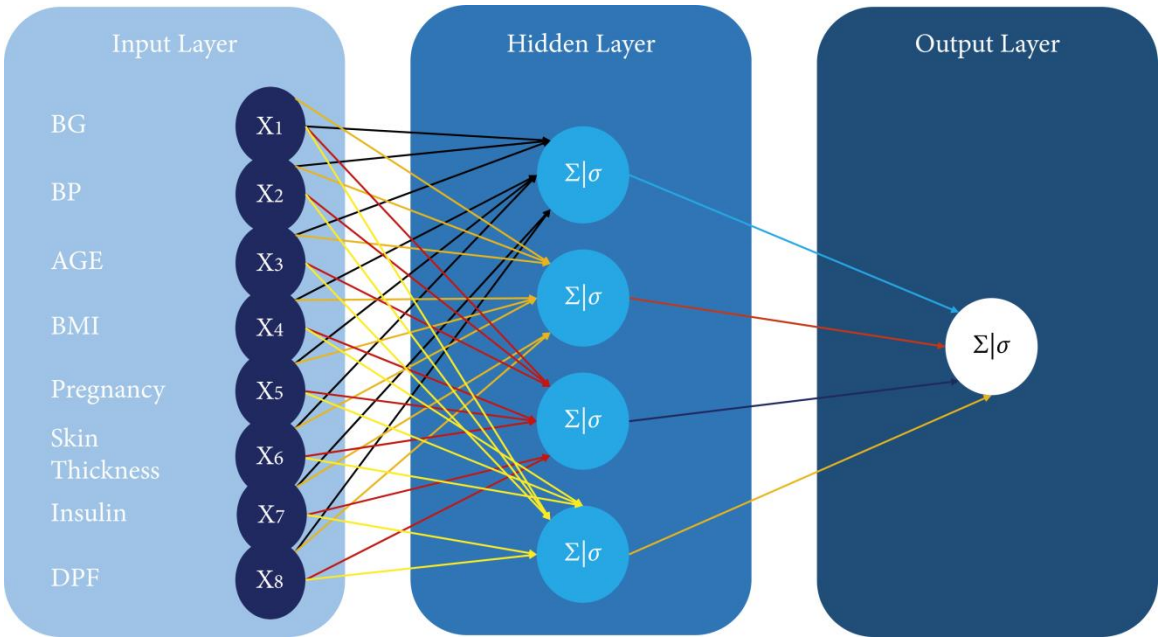


Fig. 3.Proposed MLP architecture with eight variables as input for diabetes classification

Second, we implement three widely used machine learning algorithms for diabetes prediction, i.e., moving averages, linear regression, and LSTM. Mainly, we optimized LSTM for crime prediction due to its outstanding performance in real-world applications, particularly in healthcare. The implementation details of the proposed algorithms are as follows.

Diabetes Classification Techniques

For diabetic classification, we fine-tuned three widely used state-of-the-art techniques. Mainly, a comparative analysis is performed among the proposed techniques for classifying an individual in either of the diabetes categories. The details of the proposed diabetes techniques are as follows.

- **Logistic Regression-**

  It is appropriate to use logistic regression when the dependent variable is binary, as we have to classify an individual in either type 1 or type 2 diabetes. Besides, it is used for predictive analysis and explains the relationship between a dependent variable and one or many independent variables, as shown in equation. Therefore, we used the sigmoid cost function as a hypothesis function ((x)). The aim is to minimize cost function . It always results in classifying an example either in class 1 or class 2.

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- **Random Forest (RF)-**

  As its name implies, it is a collection of models that operate as an ensemble. The critical idea behind RF is the wisdom of the crowd, each model predicts a result, and in the end, the majority wins. It has been used in the literature for diabetic prediction and was found to be effective [55]. Given a set of training examples $X = x_1, x_2, \ldots, x_m$ and their respective targets $Y = y_1, y_2, \ldots, y_m$, RF classifier iterates B times by choosing samples with replacement by fitting a tree to the training examples. The training algorithm consists of the following steps depicted in equation (2).

  (i)For $b = 1...B$, sample with replacement n training examples from X and Y.

  (ii)Train a classification tree f_b on X_b and Y_b.

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} f_i(x')$$

- **Multilayer Perceptron-**

  For diabetes classification, we have fine-tuned multilayer perceptron in our experimental setup. It is a network where multiple layers are joined together to make a classification method, as shown in Figure 2. The building block of this model is perceptron, which is a linear combination of input and weights. We used a sigmoid unit as an activation function shown in Algorithm 1. The proposed algorithm consists of three main steps. First, weights are initialized and output is computed at the output layer () using the sigmoid activation function. Second, the error is computed at hidden layers () for all hidden units. Finally, in a backward manner, all network weights () are updated to reduce the network error. The detailed procedure is outlined in Algorithm 1 for diabetes classification.
  Figure 2 shows the multilayer perceptron classification model architecture where eight neurons are used in the input layer because we have eight different variables. The middle layer is the hidden layer where weights and input will be computed using a sigmoid unit. In the end, results will be computed at the output layer. Backpropagation is used for updating weights so that errors can be minimized for predicting class labels. For simplicity, only one hidden layer is shown in the architecture, which in reality is much denser.
  Input data from the input layer are computed on the hidden layers with the input values and weights initialized. Every unit in the middle layer called the hidden layer takes the net input, applies activation function “sigmoid” on it, and transforms the massive data into a smaller range between 0 and 1. The calculation is functional for every middle layer. The same procedure is applied on the output layer, which leads to the results towards the prediction for diabetes.

Diabetes Prediction-

It is more beneficial to identify the early symptoms of diabetes than to cure it after being diagnosed. Therefore, in this study, a diabetes prediction system is proposed where three state-of-the-art machine learning algorithms are exploited, and a comparative analysis is performed. The details of the proposed approaches are as follows.

- **Moving Averages-**
  To predict diabetes, we used moving averages with the experimental setup due to its effectiveness in diabetes prediction for children. It is based on a calculation that analyzes data points by creating a series of averages of the subset of the data randomly. The moving average algorithm is based on the "forward shifting" mechanism. It excludes the first number from the series and includes the next value in the dataset, as shown in equation. The input values are calculated by averaging (P_{SM}) the train data at certain time stamps $P_M + P_M + \ldots P_{M-(n-1)}$. The algorithm used past observations as input and predicted future events.

$$P_{SM} = \frac{P_M + P_{M-1} + \ldots + P_{M-(n-1)}}{n}$$

$$= \frac{1}{n}\sum_{i=0}^{n-1} P^{M-i}.$$

Linear Search-

Second, a linear regression model is applied to the DROPBOX Indian dataset with the same experimental setup. We used this approach to model a relationship between a dependent variable, that is, outcome in our case, and one or more independent variables. The autonomous variable response affects a lot on the target/dependent variable, as shown in equation. We use a simplified hypothesis and cost function for multivariate linear regression, as there are eight different variables in our dataset We choose a very simplified hypothesis function (). The aim is to minimize cost function by choosing the suitable weight () parameters and minimizing sum of squared error (SSE).

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)} - y^{i}\right)\right)^2$$

$$h_\theta(x) = \theta^T x$$
$$= \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n.$$

Long Short-Term Memory

For diabetic forecasting, we have calibrated the long short-term memory algorithm with our experimental setup. The proposed approach outperformed as compared to other state-of-the-art techniques implemented, as shown in Table. LSTM is based on recurrent neural network (RNN) architecture, and it has feedback connections that make it suitable for diabetes forecasting . LSTM mainly consists of a cell, keep gate, write gate, and an output gate, as shown in Figure. The key behind using LSTM for this problem is that the cell remembers the patterns over a long period, and three portals help regulate the information flow in and out of the system. The details are presented in Algorithm.

Algorithm	Accuracy(%)	Recall(%)	Precision(%)
Logistic regression	73.05	72.7	73
Random forest	77.4	75.7	76.9
Proposed fine-tuned MLP	86.083	85.1	86.6

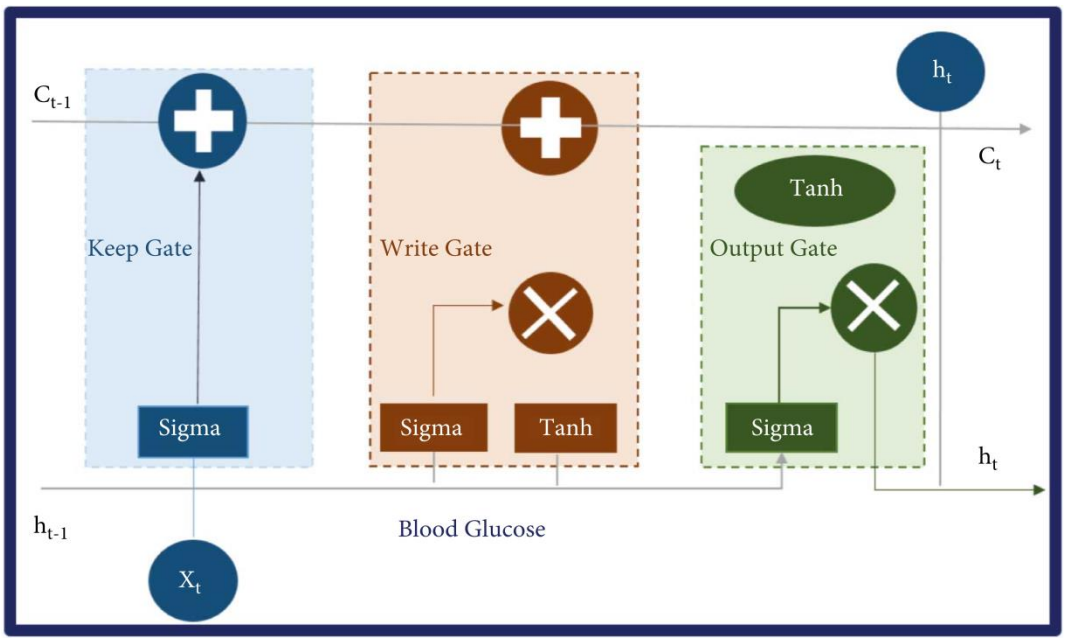


Fig 4. BG prediction using long short-term memory (LSTM) algorithm

Input to the algorithm is eight attributes enlisted in Table 3, measured from healthy and diabetic patients. The proposed LSTM-based diabetes prediction algorithm is trained with 80% of the data, and the remaining 20% is used for testing. We fine-tuned the prediction model by using a different number of LSTM units in the cell state. This fine-tuning helps to identify more prominent features in the dataset. These features will be kept in the cell state of the keep gate of the LSTM and will be given more weightage because they provide more insights to predict BG level. After that, we updated the network's weights by pointwise addition of the cell state and passed only those essential attributes for BG prediction. At this stage, we captured the dependencies between diabetes parameters and the output variable. Finally, the output gate updates the cell state and outputs/forwards only those variables that can be mapped efficiently on the outcome variable.

The diabetes prediction algorithm consists of three fundamental steps. First, weights are initialized and a sigmoid unit is used in the forget/keep gate to decide which information should be retained from previous and current inputs (). The input/write gate takes the necessary information from the keep gate and uses a sigmoid unit which outputs a value between 0 and 1. Besides, a Tan_h unit is used to update the cell state C_t and combine both outputs to update the old cell state to the new cell state.

Attributes	Description	Mean	Std. deviation	Range
Pregnancies	No. of pregnancies	3.85	3.37	0–17
Glucose	2 hours of oral glucose tolerance test for plasma glucose concentration	121	32	0–199
Blood pressure	Blood pressure in mm Hg	69.1	19.3	0–122
Skin thickness	Skinfold thickness of triceps (mm)	20.5	15.9	0–99
Insulin	Two hours of serum insulin (mu U/ml)	79.8	115	0–846
BMI	Body mass index (weight in kg/(height in m)^2)	32	7.88	0–67
Diabetes Pedigree Function	Attribute used in diabetes prognosis	0.47	0.33	0.078–2.4
Age	Age (years)	33.2	11.8	21–81
Outcome	Class variable (0 or 1)	0.35	0.48	Y/N

Finally, inputs are processed at the output gate and again a sigmoid unit is applied to decide which cell state should be output. Also, $\mathrm{Tan_h}$ is applied to the incoming cell state to push the output between 1 and −1. If the output of the gate is 1, then the memory cell is still relevant to the required production and should be kept for future results. If the output of the gate is 0, the memory cell is not appropriate, so it should be erased. For the write gate, the suitable pattern and type of information will be determined written into the memory cell. The proposed LSTM model predicts the BG level (h_t) as output based on the patient's existing BG level (X_t).

Code

In [8]:
```
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import accuracy_score
```

Data Collection and Analysis

PIMA Diabetes Dataset

In [7]:
```
# loading the diabetes dataset to a pandas DataFrame
diabetes_dataset = pd.read_csv('C:/Users/ashu/Desktop/INT247/diabetes.csv')
```

In [9]:
```
pd.read_csv?
```

In [10]:
```
# printing the first 5 rows of the dataset
diabetes_dataset.head()
```

Out[10]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

In [11]:
```
# number of rows and Columns in this dataset
diabetes_dataset.shape
```

Out[11]:
```
(768, 9)
```

In [12]:
```
# getting the statistical measures of the data
diabetes_dataset.describe()
```

Out[12]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

In [13]:
```
diabetes_dataset['Outcome'].value_counts()
```

Out[13]:
```
0    500
1    268
Name: Outcome, dtype: int64
```

0 --> Non-Diabetic

1 --> Diabetic

In [14]:
```
diabetes_dataset.groupby('Outcome').mean()
```

Out[14]:

Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164

In [15]:
```
# separating the data and labels
X = diabetes_dataset.drop(columns = 'Outcome', axis=1)
Y = diabetes_dataset['Outcome']
```

In [16]:
```
print(X)
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age
0                       0.627   50
1                       0.351   31
2                       0.672   32
3                       0.167   21
4                       2.288   33
..                        ...  ...
763                     0.171   63
```

Data Standardization

In [20]:
```
scaler = StandardScaler()
```

In [21]:
```
scaler.fit(X)
```

Out[21]:
```
StandardScaler()
```

In [22]:
```
standardized_data = scaler.transform(X)
```

In [23]:
```
print(standardized_data)
```

```
[[ 0.63994726  0.84832379  0.14964075 ...  0.20401277  0.46849198
   1.4259954 ]
 [-0.84488505 -1.12339636 -0.16054575 ... -0.68442195 -0.36506078
  -0.19067191]
 [ 1.23388019  1.94372388 -0.26394125 ... -1.10325546  0.60439732
  -0.10558415]
 ...
 [ 0.3429808   0.00330087  0.14964075 ... -0.73518964 -0.68519336
  -0.27575966]
 [-0.84488505  0.1597866  -0.47073225 ... -0.24020459 -0.37110101
   1.17073215]
 [-0.84488505 -0.8730192   0.04624525 ... -0.20212881 -0.47378505
  -0.87137393]]
```

In [24]:
```
X = standardized_data
Y = diabetes_dataset['Outcome']
```

In [25]:
```
print(X)
print(Y)
```

Train Test Split

In [26]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.2, stratify=Y, random_state=2)
```

In [27]:
```
print(X.shape, X_train.shape, X_test.shape)
```

```
(768, 8) (614, 8) (154, 8)
```

Training the Model

In [28]:
```
classifier = svm.SVC(kernel='linear')
```

In [39]:
```
#training the support vector Machine Classifier
classifier.fit(X_train, Y_train)
```

Out[39]:
```
SVC(kernel='linear')
```

Model Evaluation

Accuracy Score

In [32]:
```
# accuracy score on the training data
X_train_prediction = classifier.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

In [33]:
```
print('Accuracy score of the training data : ', training_data_accuracy)
```

```
Accuracy score of the training data :  0.7866449511400652
```

In [34]:
```
# accuracy score on the test data
X_test_prediction = classifier.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

In [35]:
```
print('Accuracy score of the test data : ', test_data_accuracy)
```

```
Accuracy score of the test data :  0.7727272727272727
```

Making a Predictive System

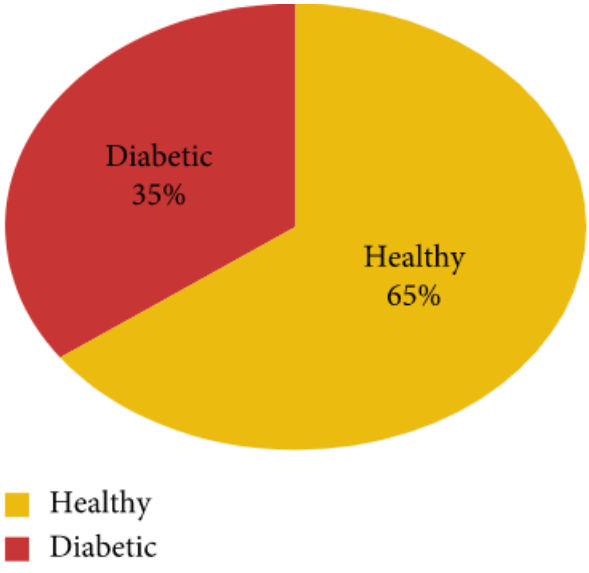
In [38]:
```
input_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input_data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print('The person is not diabetic')
else:
  print('The person is diabetic')
```

```
[[ 0.3429808   1.41167241  0.14964075 -0.09637905  0.82661621 -0.78595734
   0.34768723  1.51108316]]
[1]
The person is diabetic
```

Result

This study used the DROPBOX Indian Diabetes dataset taken from the National Institute of Diabetes and Kidney Diseases center. The primary objective of using this dataset is to build an intelligent model that can predict whether a person has diabetes or not, using some measurements included in the dataset. There are eight medical predictor variables and one target variable in the dataset. Diabetes classification and prediction are a binary classification problem. The details of the variables are shown in Table.

The dataset consists of 768 records of different healthy and diabetic female patients of age greater than twenty-one, as shown in Figure 4. The feature value distribution is shown in Figure 5. The target variable outcome contains only two values, 0 and 1. The primary objective of using this dataset was to predict diabetes diagnostically. Whether a user has a chance of diabetes in the coming four years in women belongs to DROPBOX Indian. The dataset has a total of eight variables: glucose tolerance, no. of pregnancies, body mass index, blood pressure, age, insulin, and Diabetes Pedigree Function. All eight attributes shown in Table 3 are used for the training dataset in the classification model in this work.



Data distribution

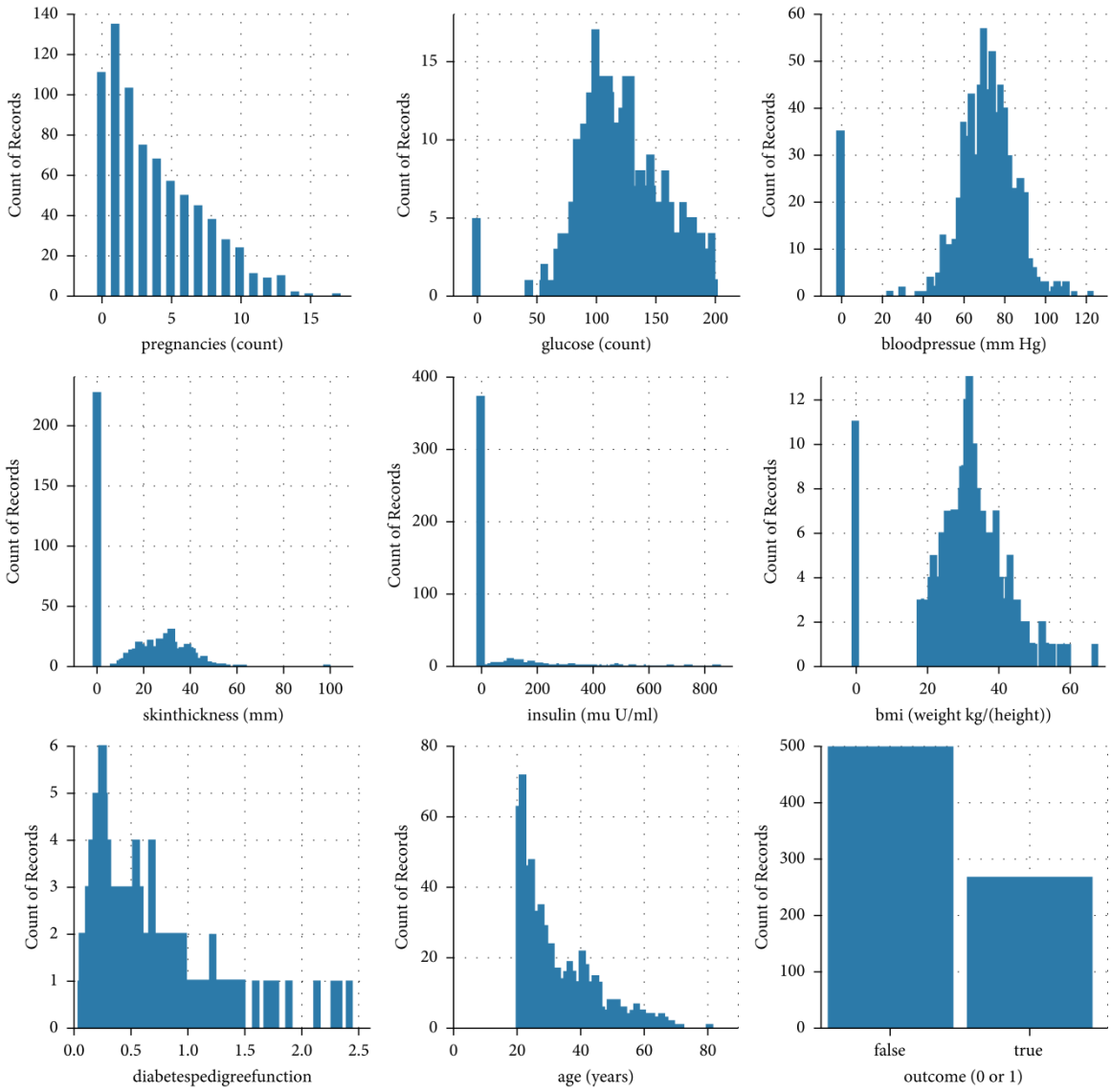


Fig. Dataset features' distribution visualization

This paper compares the proposed diabetes classification and prediction system with state-of-the-art techniques using the same experimental setup on the DROPBOX Indian dataset. The following sections highlighted the performance measure used and results attained for classification and prediction, and a comparative analysis with baseline studies is presented.

Performance Metrics-

Three widely used state-of-the-art performance measures (Recall, Precision, and Accuracy) are used to evaluate the performance of proposed techniques, as shown in Table 4. TP shows a person does not have diabetes and identified as a non diabetic patient, and TN shows a diabetic patient correctly identified as a diabetic patient. FN shows the patient has diabetes but is predicted as a healthy person. Moreover, FP shows the patient is a healthy person but predicted as a diabetic patient. The algorithm utilized 10-fold cross-validation for training and testing the classification and prediction model**.**

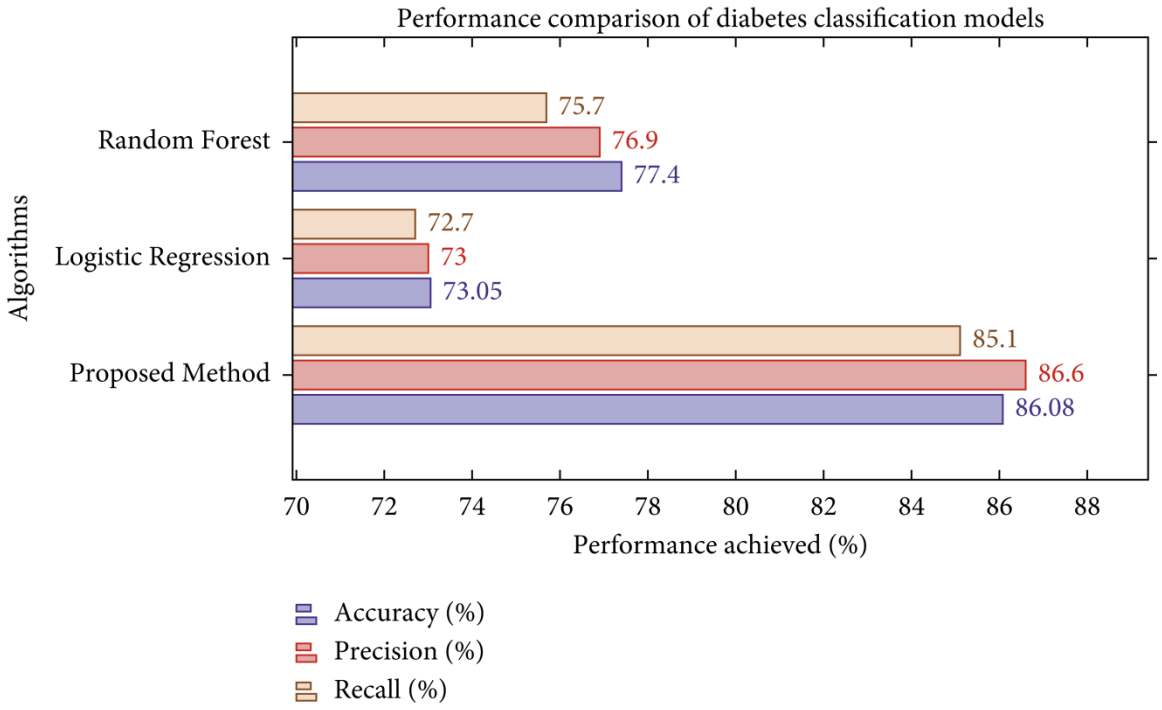
For diabetes prediction, the two most commonly used performance measures are the means correlation coefficient (*r*/Pearson *R*) and root mean square error (RMSE), as shown in Table 5. *R* is mainly used to measure the linear dependence strength among the two variables. One variable is for actual value, and another variable is for predicted values. RMSE generates a hint of the overall correctness of the estimate. There can be three values for correlation: 0 for no relation, 1 for positive correlation, and −1 for the negative correlation. RMSE shows the difference between actual values and predicted values.

Performance measure for diabetes prediction.

Performance metric	Formula
r	$(n(\sum XY) - (\sum X)(\sum Y))/\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}$
Root mean square error (RMSE)	$\sqrt{[\sum_{i=1}^{N}(y_{fi} - y_{f0})^2/N]}$
Accuracy	$(TP + TN)/(TP + TN + FP + FN)$

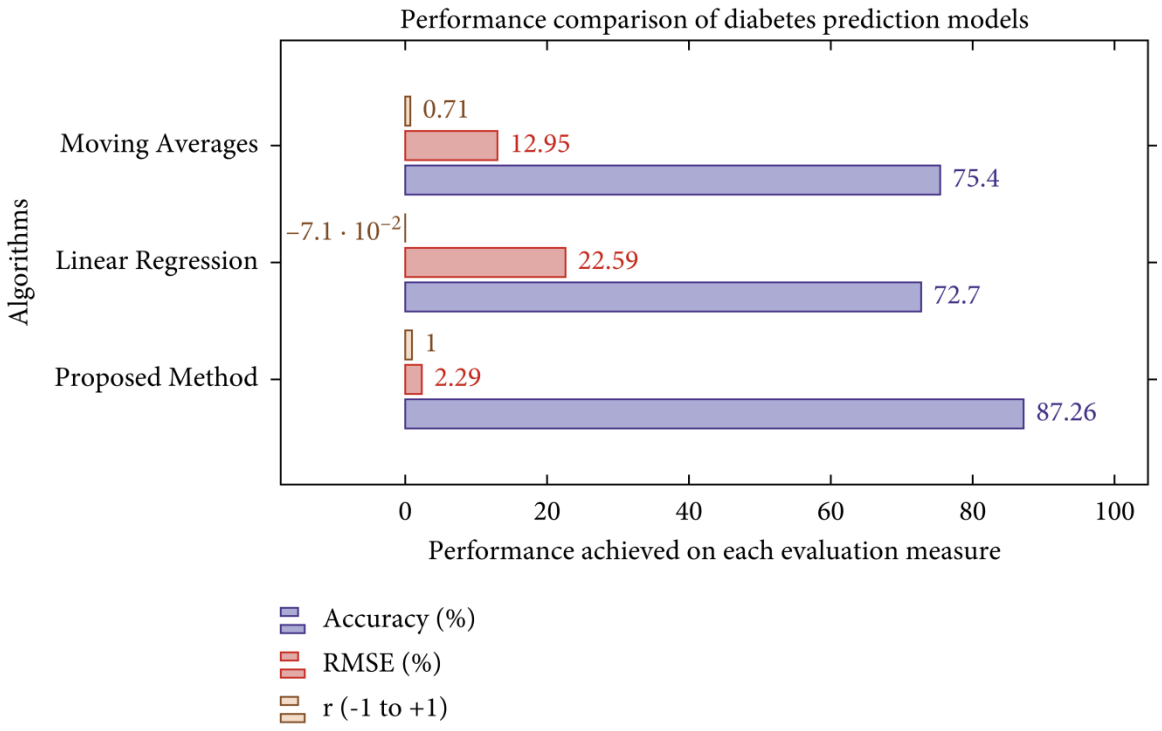
For diabetic classification, three state-of-the-art classifiers are evaluated on the DROPBOX dataset. The results illustrate that the fine-tuned MLP algorithm obtained the highest accuracy of 86.083% as compared to state-of-the-art systems, as shown in Table 2.

It is evident from the results that our proposed calibrated MLP model could be used for the effective classification of diabetes. The proposed classification approach can also be beneficial in the future with our proposed hypothetical system. Data of weight scales, blood pressure monitor, and blood glucometer will be collected through sensor devices such as BLE and input of user's demographic data (for example, date of birth, height, and age). The proposed MLP algorithm outperforms with 86.6% Precision, 85.1% Recall, and 86.083% Accuracy, as shown in Figure 6. These results are outstanding for decision-making with the proposed hypothetical system to determine patient diabetes, T1D or T2D.



We also have explored the dataset used in Andy Choens' study. This dataset consists of records of only one patient. The information was recorded every five minutes. The collection of data was made by using a sensor device (a CGM device). This device allows the patient to store information about BG every five minutes. So, the recorded data by using this device are in massive amounts. Dataset was limited, and most data were noisy that can affect the accuracy of the proposed system, so we neglected it.

For diabetic prediction, we implemented three state-of-the-art algorithms, i.e., linear regression, moving averages, and LSTM. Notably, we fine-tuned LSTM and compared its performance with other algorithms. It is evident from Figure and Table that the LSTM outperformed as compared to other algorithms implemented in this study.



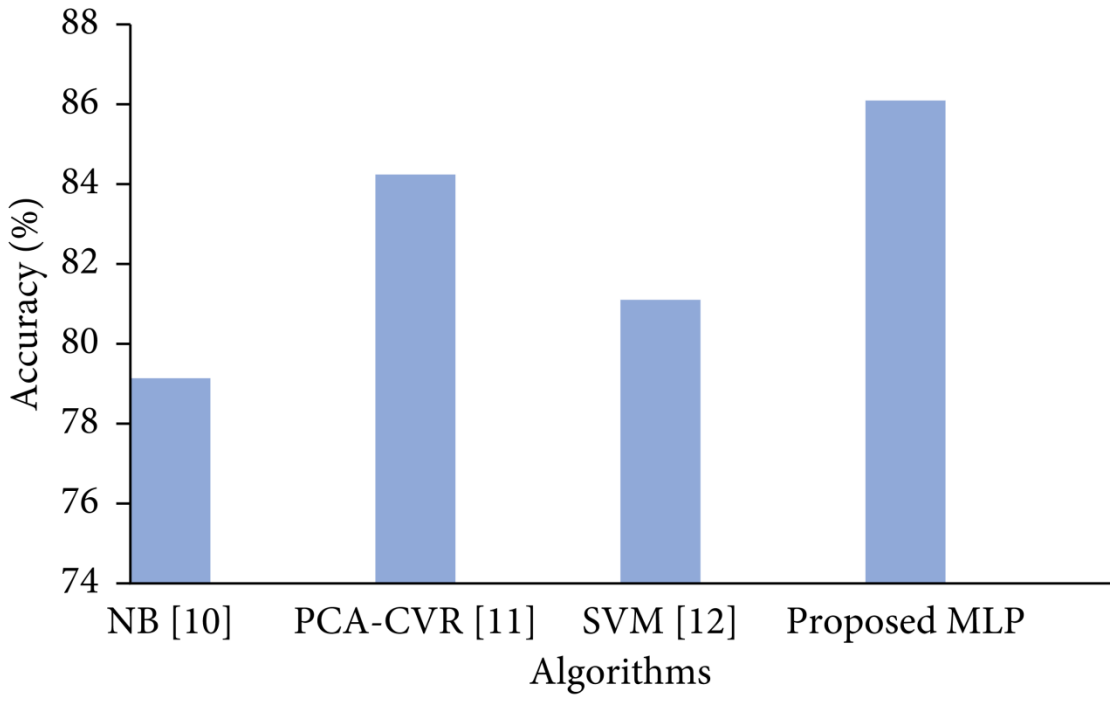
Algorithm	*r*	RMSE	Accuracy
Moving average	0.71	42.946	75.4
Linear regression	−0.071	82.592	72.7
Proposed fine-tuned LSTM	0.999	2.285	87.26

Forecasting model comparison for BG

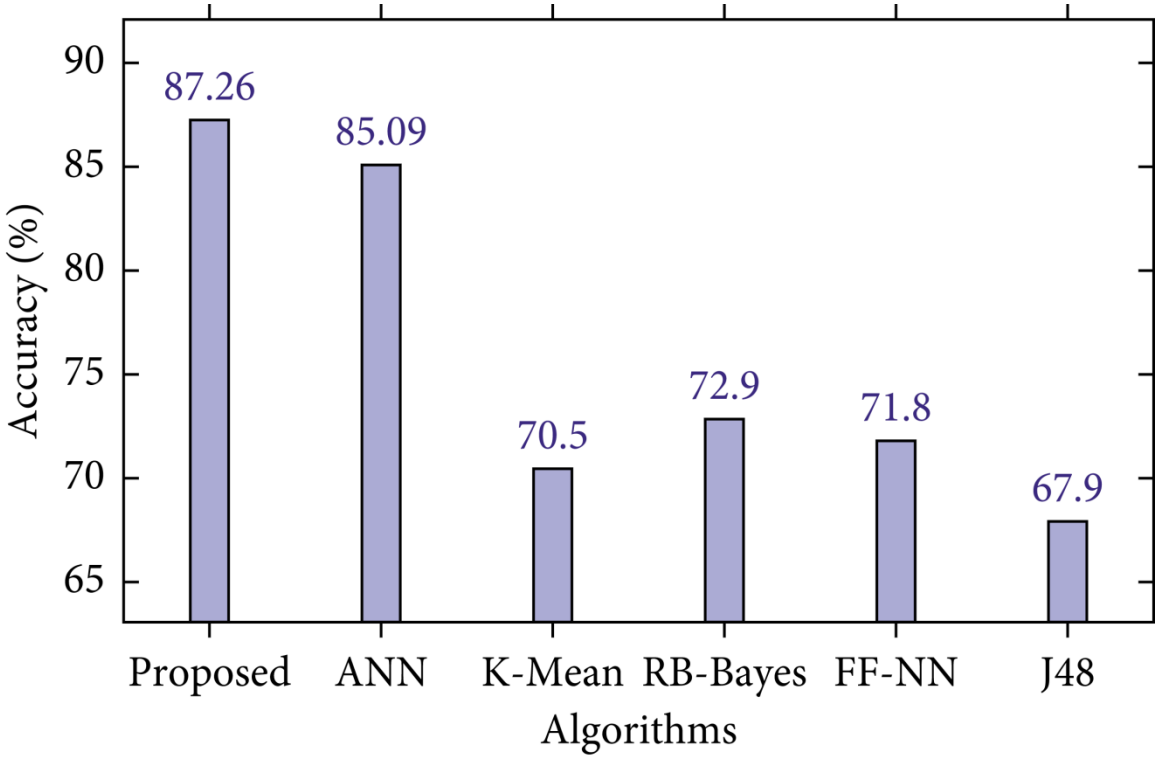
Table shows the performance values of prediction models with RMSE and *r* evaluation measures. The proposed fine-tuned LSTM produced the highest accuracy, 87.26%, compared to linear regression and moving average. We can see in Table 6 that the correlation coefficient value is 0.999 using LSTM, −0.071 for linear regression, and 0.710 for moving average.

Different baseline studies have been implemented and compared with the proposed system to verify the performance of the proposed diabetes classification and prediction system. Mainly, we focus on those studies that used the DROPBOX dataset.

First, we compare the state-of-the-art diabetes classification techniques with the proposed technique. All the baseline techniques used the DROPBOX dataset and the same evaluation measures used in this study. In particular, the authors compared naïve Bayes, PCA_CVR (classification via regression) , and SVM with different machine learning techniques for diabetes classification. However, the proposed fine-tuned MLP-based diabetes classification technique outperformed as compared to baseline studies, as shown in Figure.

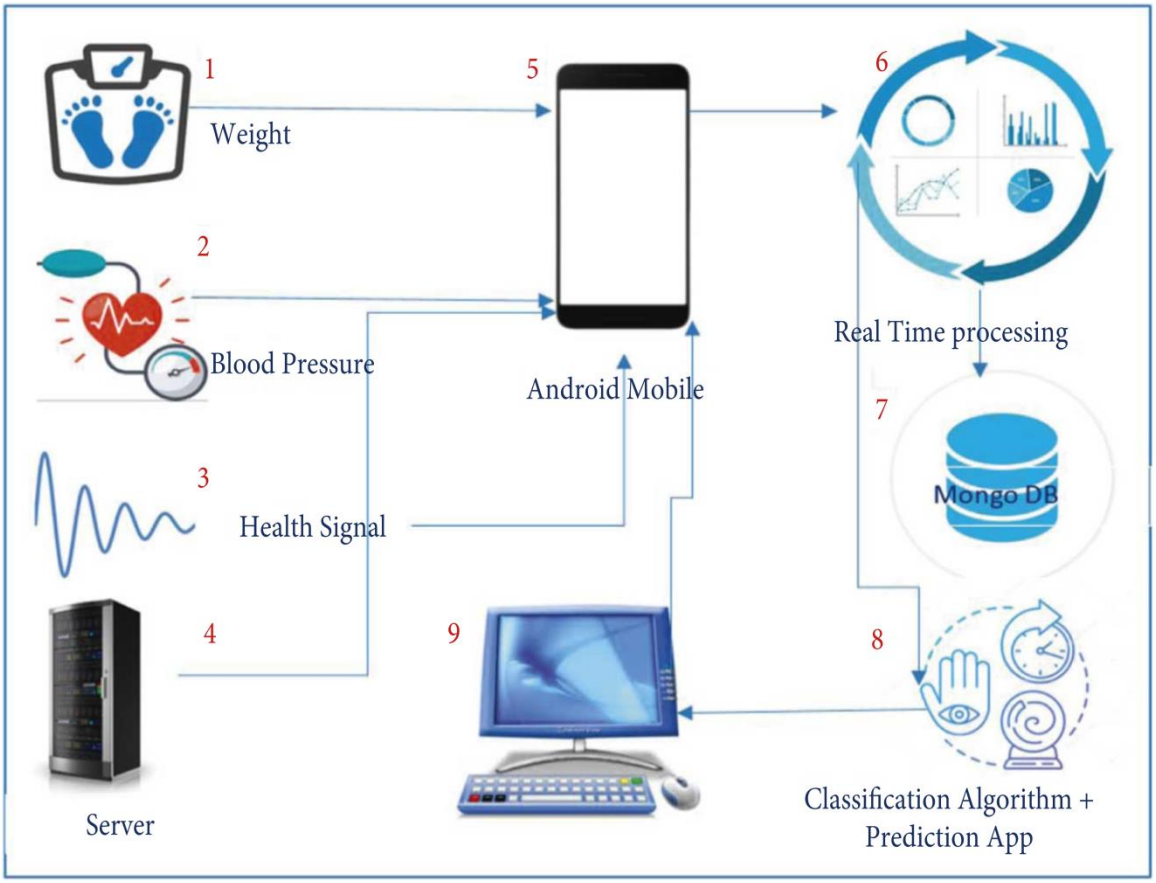


Several attempts have also been made in the literature for diabetic prediction due to its importance in real life. For this comparison, we have chosen the most recent and state-of-the-art techniques. We compare the proposed system performance with the recent state-of-the-art systems as shown in Figure s. The proposed method outperformed as compared to state-of-the-art systems with an accuracy of 87.26%, all the compared systems evaluated on the PID with the same experimental setup.



Future Scope

This study has also proposed the architecture of a hypothetical diabetic monitoring system for diabetic patients. The proposed hypothetical system will enable a patient to control, monitor, and manage their chronic conditions in a better way at their homes. The monitoring system will store the health activities and create interaction between patients, smartphones, sensor medical devices, web servers, and medical teams by providing a platform having wireless communication devices, as shown in Figure. The central theme of the proposed healthcare monitoring system is the collection of data from sensors using wireless devices and transmitting to a remote server for diagnosis and treatment of diabetes. Knowledge-based data are stored. Rule-based procedures will be applied for the suggestions and treatment of diabetes, informing the patient about his current health condition, prediction, and recommendation of future changes in BG.



First, essential data about patient health will be collected from sensors such as BLE wireless devices. Data comprised weight, blood pressure, blood glucose, and heartbeat, along with some demographic information such as age, sex, name, and CNIC (Social Security Number). Some information is required in the application installed on the user's mobile and sensor data. All completed data in the application will be transferred to the real-time data processing system. On the other side, aggregate data will be stored in MongoDB for future processing. Analysis and prepossessing techniques are performed to extract rules from the knowledge base for the treatment and suggestions about the user. Results and treatment procedures will be sent to the monitoring system, and finally, the user can get the output by interacting with their android mobile phone. In the end, the patient will know about the health condition and risk prediction of diabetes based on the data transferred by their application and stored data from history about the user.

The proposed structural design for hypothetical real-time processing and monitoring of diabetes is shown in Figure. The data from the user's mobile will be transmitted in the JavaScript Object Notation (JSON) format to the Application Program Interface (API) in any language. The data produced at this stage will be in the form of messages, which are then transferred to the Kafka application . Kafka will store all the data and messages and deliver the required data and processed output to the endpoints that could be a web server, monitoring system, or a database for permanent storage. In Kafka, application data are stored in different brokers, which can cause latency issues. Therefore, within the system architecture, it is vital to consider processing the readings from the sensors closer to the place where data are acquired, e.g., on the smartphone. The latency problem could be solved by placing sensors close to the place, such as a smartphone where data are sent and received.

This inclusion will make the overall network architecture compliant to the emerging Edge and Fog computing paradigms, whose importance in critical infrastructures such as hospitals is gaining momentum. It is essential to consider the Edge and Fog computation paradigm while sending and receiving data from smartphones to increase the performance of the hypothetical system. Edge computing utilizes sensors and mobile devices to process, compute, and store data locally rather than cloud computing. Besides, Fog

computing places resources near data sources such as gateways to improve latency problems.

Apache Kafka will be used in real time as a delivery agent for messages in a platform that allows fault-tolerant, tall throughput, and low-latency publication. The vital signs' data collected by the patients are placed using the JSON format and then transmitted using wireless devices with the help of an android application having HTTP along with REST API for the confined remote server for the design [28]. Moreover, Node.js for web design will be used as a REST API to collect sensor data. Kafka application will receive it in the form of streams of records.

The sensor data that comes from the Kafka application is continuously generated and stored on the server. In the proposed system, the MongoDB NoSQL database will be used for data storage due to its efficiency in handling and processing real-world data [29]. The stored diabetes patient data can be input into our proposed diabetes classification and prediction techniques to get useful insights.

Conclusion

Diabetes mellitus is a disease, which can cause many complications. How to exactly predict and diagnose this disease by using machine learning is worthy studying. According to the all above experiments, we found the accuracy of using PCA is not good, and the results of using the all features and using mRMR have better results. The result, which only used fasting glucose, has a better performance especially in Luzhou dataset. It means that the fasting glucose is the most important index for predict, but only using fasting glucose cannot achieve the best result, so if want to predict accurately, we need more indexes. In addition, by comparing the results of three classifications, we can find there is not much difference among random forest, decision tree and neural network, but random forests are obviously better than the another classifiers in some methods. The best result for Luzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721, which can indicate machine learning can be used for prediction diabetes, but finding suitable attributes, classifier and data mining method are very important. Due to the data, we cannot predict the type of diabetes, so in future we aim to predicting type of diabetes and exploring the proportion of each indicator, which may improve the accuracy of predicting diabetes.

References

[1] https://www.who.int/health-topics/diabetes.

[2] https://www.medicalnewstoday.com/articles/325018#how-is-the-pancreas-linked-with-diabetes

.

[3] https://www.webmd.com/diabetes/diabetes-causes.

[4] https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284.

[5] https://www.niddk.nih.gov/healthinformation/diabetes/overview/symptoms-causes.

[6] https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html.

[7] https://www.healthgrades.com/right-care/diabetes/is-there-a-cure-for-diabetes.

[8] https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects.