Classification of Diabetes Types using Machine Learning

Oyeranmi Adigun[1], Folasade Okikiola[2]
Department of Computer Science
Yaba College of Technology
Lagos, Nigeria

Nureni Yekini[3]
Department of Computer Engineering, Yaba College of Technology, Lagos, Nigeria

Ronke Babatunde[4]
Department of Computer Science
Kwara State University Malete
Kwara State, Nigeria

Abstract*—Machine learning algorithms have aided health workers (including doctors) in the processing, analysis, and diagnosis of medical problems, as well as the detection of disease patterns and other patient data. Diabetes mellitus (DM), commonly referred to as diabetes, is a gathering of a syndrome issue that is portrayed by high glucose levels in the blood over a drawn-out period. It is a long-term illness that is a great threat to humanity and causes death. Most of the existing machine learning algorithms used for the classification and prediction of diabetes suffer from embodying redundant or inessential medical procedures that cause complications and wastage of time and resources. The absence of a correct diagnosis scheme, deficiency of economic means, and a general lack of awareness represent the main reasons for these negative effects. Hence, preventing the sickness altogether through early detection may doubtless cut back a considerable burden on the economy and aid the patient in diabetes management. This study developed diabetes classification using machine learning techniques that will minimize the aforementioned drawbacks in the prediction of diabetes systems. Decision tree classifiers, logistic regression, random forest, and support vector machines are all examples of predictive algorithms that were tested in this paper. 1009 records of data set were obtained from the Diabetes dataset of Abelvikas, Data World. We used a confusion matrix to visualize the performance evaluation of the classifiers. The experimental result shows that the four machine learning algorithms perform well. However, Random Forest outperforms the other three, with a prediction accuracy of 100% and has a better prediction level when compared with others and existing work.

Keywords*—Machine learning; diabetes mellitus; predictive algorithm; correlation map; confusion matrix

I. Introduction

Diabetes is one of the most common and speedily increasing diseases within the world [1] and a serious pathological state in the world. This polygenic disease is a condition in which the body is unable to produce the required amount of internal secretion to keep blood sugar levels in check (National Center for Biotechnology Information, NCBI). In general, a higher risk of diabetes infection is associated with female gender, age over 35, and individuals who are overweight.

The day demands to identify and diagnose this diabetes condition at an early stage cannot be over-emphasized. The diagnosis and analysis of diabetes disease is an important issue in classification that is required and must be cost-effective, suitable, and valid to be built.

Diabetes mellitus, also known as diabetes, is a metabolic disorder that can result in elevated blood sugar levels (MSD Manual). It is a long-lasting disease that occurs when the pancreas fails to produce enough insulin or when the body fails to properly utilize the insulin that is produced. The insulin in the body system regulates the movement of sugar from the blood into the cells for energy use. Untreated high diabetes blood sugar can cause damage to the critical major organs such as the eyes, and kidneys, heart disease, sudden death which can lead to chronic damage to other organs, etc. [2, 3].

Therefore insulin is a catalyst in the regulation of blood sugar hormones. Hyperglycaemia (high blood sugar) is a common complication of uncontrolled diabetes that resulted in severe damage to nerves and blood vessels of the body's systems [4]. Diabetes is one of the most lethal diseases in the world, but with the introduction of machine learning, there is the potential to find a solution to this pandemic.

The crux of using a machine learning classifier and data mining is to derive knowledge from information stored in the dataset and produce a simple pattern description. A diabetes diagnostic tool using machine learning needs to be developed to predict patients with diabetes to detect the illness early before it is pathetic. Machine-learning algorithms (MLA) identify patterns from statistical quantities of data and feed them into the system to be digitally processed. Much has been achieved in the areas of using machine learning algorithms to solve many challenges in the health sector with the development of technology. Some of these are for the prognosis and/or diagnosis of diabetes for active and accurate decision-making [5]. Therefore, this paper focuses on the application of machine learning techniques to an online dataset to uncover hidden patterns in medical diagnosis and predict diabetes based on the data collected. To ensure that the information obtained from a system built using these techniques is reliable, a Support Vector Machine (SVM) and Random Forest (RF) are proposed for use in the prediction of diabetes in a patient.

It was discovered that there are three major kinds of Diabetes classified into three types: type 1, type 2, and gestational diabetes. Type 1 diabetes is distinguished by a lack of insulin production and necessitates daily insulin administration. Despite the fact that the exact cause of type 1

This work was financially supported by Yaba College of Technology in Lagos, Nigeria.

diabetes is unknown, it is unavoidable. The symptoms may appear unexpectedly and are caused by excessive urination (polyuria), fatigue (polydipsia), persistent hunger, loss of weight loss, and vision. Type 2 diabetes (non-insulin-dependent,) may be caused as a result of insufficient insulin in the body and is primarily caused by excess body weight and physical inactivity. The third type is gestational (hyperglycemia), which is defined as having blood glucose levels that are higher than normal but are lower than the conditions of diabetes that occur during pregnancy. This increases the likelihood of complications during pregnancy and childbirth and faces a greater chance of type 2 diabetes in the future too [6].

Patients with diabetes must undergo a series of tests and exams in order to properly diagnose the disease. These tests may include unnecessary or redundant medical procedures that result in complications and a waste of time and resources. Diabetes lowers the standard of living and reduces labor productivity, so the economic cost of the disease far outweighs the direct medical costs within the care sector. The main causes of these negative effects are a lack of a proper diagnosis scheme, a lack of financial resources, and a general lack of awareness. As a result, preventing the illness entirely through early detection will almost certainly reduce the economic burden and aid the patient in diabetes management. The following are the objectives of the study:

- Develop the Diabetes prediction system using a decision tree classifier, logistic regression, random forest, and support vector machine.
- Evaluate and compare the developed system and the performance of each algorithm in the ensemble of algorithms based on sensitivity, specificity, and accuracy.

The study is organized into five sections. Section I introduces the study by discussing the keywords briefly as well as the study's objectives Section II explains various related works in the field of diabetes type prediction Section III describes the study's methodology in detail. Section IV discusses the results of the algorithms. Section V concludes the study with recommendations for additional research.

II. Related Works

Several researchers have made contributions to fields where diabetes was predicted. Diabetes has a significant economic impact on society, and it is the most expensive chronic disease. The author [7] addressed the fact that majority of diabetes patients are asymptomatic, which leads to delayed standard clinical laboratory examinations that create large health datasets over a lifetime. They looked at machine learning algorithms to help with diabetes screening via routine laboratory tests, using data from 62,496 patients' lab tests. The following classifications were used; artificial neural networks, Bayes naïve, K-nearest neighbor, random forest, regression models, and support vector machines. In detecting diabetes, the artificial neural network model outperformed the others. Based on clinical data processing, computer processing has been used to identify diseases [8]. Knowledge extraction from data to aid decision-making by experts is a movement in the next generation of intelligent health systems [9].

The author [10] sought to develop effective models for predicting early gestational diabetes mellitus (GDM). The seven variables and 73 variables datasets were used to create models that predicted early GDM in different situations. In early pregnancy, ML models predicted GDM with high accuracy and were developed and tested in the Chinese population. The study [11] also employed ML and classification algorithms, with Logistic Regression providing the highest accuracy of 96 percent. Also, [12] carried out the use of random forest, KNN, Nave Bayes (NB), and J48 to develop diabetes analysis and prediction. The researchers used two datasets: PIDD (Pima Indian Diabetes Dataset) and 130 US hospital diabetes data sets. The developed system achieved 93.62 percent accuracy in the case of PIDD and 88.56 percent accuracy for a large dataset of 130-US hospitals. For large dataset analysis, the NB and J48 prediction algorithms were found to be superior. The author [13] reviewed diabetes types and treatments, as well as some emerging issues that may arise, and listed physical activities that will lead to healthy lifestyles.

Furthermore, [14] presented diabetes prediction based on big data from healthcare communities using various machine learning algorithms. Using SVM for classification and K-means for clustering, the developed system used an effective strategy for detecting diabetes disease earlier. The study [15] implemented a decision tree algorithm to predict diabetes. The experiments were carried out on the Pima Indians diabetes database, and the results achieved an accuracy of 87 percent. However, low sample sizes result in poor accuracy. The system that was developed can be used to predict or diagnose other diseases in the same family. In a similar vein, [16] used the most recent records of 13,309 Canadian patients aged 18 to 90 years, as well as their laboratory data. They developed predictive models using Logistic Regression and Gradient Boosting Machine (GBM) techniques and compared them to others such as Decision Tree and Random Forest. The GBM and LR models outperform the other two models. In this experiment, [17] proposed two machine learning classification algorithms, Fine Decision Tree and Support Vector Machine, which are used to detect diabetes at an early stage. When compared to the Fine Decision Tree algorithm, the SVM classification algorithm achieved a high percentage of accuracy.

The author [18] applied random forest, decision tree, and neural network in their study to predict diabetes mellitus with an accuracy of about 81 percent. The Pima Indians diabetes dataset from the UCI machine learning repository was used. The study [19] proposed using classification algorithms to predict diabetes. On a number of criteria, three machine learning classification algorithms were researched and assessed. According to the experimental findings, the Naive Bayes classification algorithm has an accuracy rate of 76.30 percent.

In [20] the author proposed the use of the Pima Indians diabetes dataset, using Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Random Forest to predict diabetes at various stages and compare the performance of

different classification techniques. While [21] presented a Unified Framework for Diabetes Prediction Based on Machine Learning. Six machine learning classifications for predicting diabetes and various evaluation criteria were used to investigate the performance of these classification techniques. The analysis results show that Naïve Bayes achieved the highest performance than the other classifiers, obtaining the F1 measure of 0.74. According to [22] in the prediction of diabetes using the classification algorithms. Naive Bayes, Multilayer Perceptron, and IBK algorithms were used. The Naive Bayes algorithm shows 100% accuracy compared to IBK 88% and Multilayer perceptron 88%.

Research work made by [23] developed a machine learning-based framework for detecting type 2 diabetes in electronic health records. The system created a semi-automated framework based on machine learning. A data-informed framework for identifying subjects with and without T2DM from EHR was proposed using machine learning and feature engineering. The author in [24] created an Ontology-based Diabetes Management system, a computer-based system that assists physicians in correctly diagnosing diabetes mellitus disease in patients. They used the Bayesian Optimization technique to boost prediction accuracy. Similarly, [25] developed a medical expert system for diabetes diagnosis, a diabetes ontology with 9 sub-classes, and a web-based application with web service architecture. With test data from 65 patients, an overall consistency rate of 90.7 percent was achieved. The author [26] demonstrated diabetes detection at an early stage using a computational intelligence fuzzy hierarchical model capable of performing early detection and identifying someone's susceptibility to DM. The model's accuracy is 87.46 percent. A number of techniques have been proposed over the years for the prediction of diabetes types. The comparison of diabetes techniques in Table I shows their performance and limitation. Four different classifiers will be used, and because Random Forest excels at working with non-linear data, the prediction will be more accurate and stable, with improved performance.

TABLE I. COMPARISON ANALYSIS OF EXISTING TECHNIQUES

S/N	Author(s)	Strategy	Performance %	Limitations
1	Aishwarya & Vaidehi (2019)	LR, RRF, RF.	96.	more time spent on their synthesis.
2	Minyechil et al (2019)	Random Forest, KNN, Naïve Bayes, and J48	93.62	time-consuming processes.
3	Quan Zou et al (2018)	D, RF, NN	80.8	could not predict the type of diabetes.
4	Zheng et al (2017)	KNN, Naïve Bayes, DT, RF, SVM, & LR	95	The model distinguishes patients with and without type 2 Diabetes Mellitus
5	Aiswarya et al (2015)[27]	DT & Naïve Bayes	J48 76.9, NB 79.5	not precise and a general conclusion for diabetes

III. METHODOLOGY

The importance of early diagnosis of diabetes mellitus to the life expectancy of the patient suffering from it cannot be over-emphasized. Early diagnosis will mean that, based on certain biological features found in the medical history of the patient, there is a predictive test. This section focuses on how predictive analysis of machine learning is used to predict the diabetes status of a patient accurately. Therefore, to develop and implement a diabetes recommendation prediction system, the proposed model employs machine learning techniques.

A. Predictive Analysis

This section illustrates the analysis of the proposed system and how the system that was designed works and is a feasible alternative to the existing one. The data used in this paper was collected from the dataset of Abelvikas, Data world. The data collected were subjected to different types of pre-processing, as will be addressed in subsequent sections to improve the system's performance. The proposed model implements the classification model with the highest accuracy level. These algorithms include Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine Classifiers. Fig. 1 shows the block diagram for the proposed model.

1) Data acquisition: This research was carried out using the dataset of Abelvikas, Data world. The dataset has multi-class problems of diabetes which separate it into individuals who have tested either negatively or positively (type1, type2, and normal) to diabetes. The dataset consists of 1009 total instances with eight attributes to provide adequate data from training after pre-processing requiring the removal of certain entries. Entries included Age (years), BS Fast (mmol/L), BS pp (mmol/L), Plasma R mmol/L), Plasma F (mmol/L), and HbA1c (mmol/L). The data collected from the Abelvikas, Data World Database was shown in the Table II.

2) The pre-processing stage: This handles inconsistencies in data to improve accuracy and precise outcomes. This dataset has missing values for a few selected attributes like Glucose level, Blood Sugar, and HBA1C because these attributes cannot have values of zero. The dataset is then scaled to normalize all values. Correlation is an amount of context between characteristics. It is a real number value that denotes the degree of significance between 0 and 1 and a negative value indicates an inverse relationship, while a direct relationship is indicated by a positive value. Fig. 2 shows the correlation map of the proposed model.

3) Training & classification: ML algorithms require training data to achieve the objective. This training dataset will be analyzed by the algorithm, which will then classify the inputs and outputs before analyzing it again. A sufficiently trained algorithm will effectively memorize all of the inputs and outputs in a training dataset. The prediction model consists of the best machine learning model after implementing different models, and the best was taken and deployed for application. The output of each model is taken to the next stage for testing. In training the classification algorithms and constructing the model, the steps taken were to

import the modules and dataset as a data frame and get insights from the dataset. From the entire data set, a feature set containing the first seven attributes is extracted, and the output set is extracted, which is the product of the prediction and the whole set is split into a 7:3 ratio train set and test set.

4) Testing: After the model was built, testing data validate to make accurate predictions. This is to confirm that the ML algorithms were trained effectively to evaluate the prediction models created.

5) Evaluation: Assessing the performance of the model using different metrics is integral to this research work. Based on the result from the test stage, the model was evaluated based on classification accuracy and specificity. A classification metric was employed to evaluate the developed model. There are four types of outcomes that could occur when performing classification predictions.

a) True positives happen when you predict that an observation belongs to a certain class and it turns out to be correct.

b) True negatives occur when you predict that an observation will not belong to a class and it actually does not belong to that class.

c) False positives happen when you assume an observation belongs to a class when it doesn't.

d) False negatives occur when you incorrectly predict that observation does not belong to a class when it does.

The results are frequently plotted on a confusion matrix. After making predictions based on the test data and then classifying each prediction as one of the four possible outcomes described above, the matrix was generated.

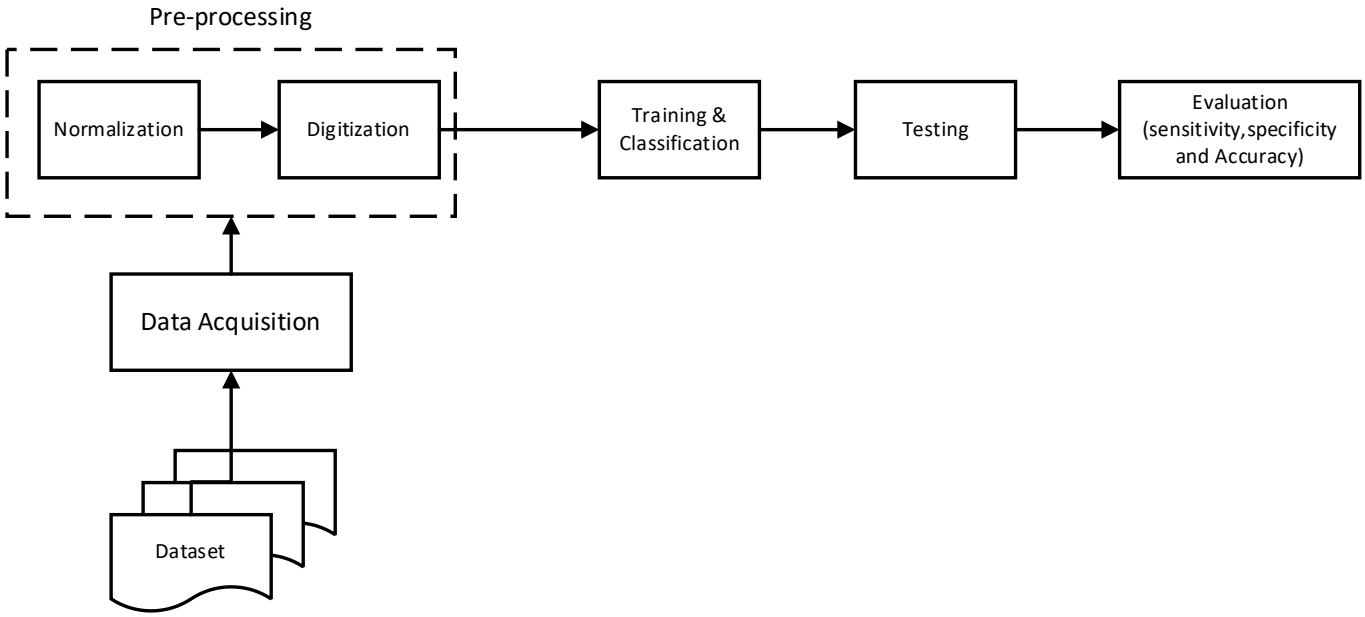

Fig. 1. Proposed Model for the Research.

TABLE II. DATABASE FILE REPRESENTATION

S/N	Field	Type	Range
1.	Age	Integer	21 – 81
2.	Blood Sugar in fasting	Real	0 – 54
3.	Blood Sugar after a meal	Real	4.2 - 8.1
4.	Plasma Glucose in fasting	Real	3.9 - 9.1
5.	Plasma Glucose	Real	7.9 - 13.1
6.	Glycated hemoglobin (HBA1C)	Real	28 – 69
7.	Type	String	0 – 255
8.	Class	Boolean	1 - Diabetic, 0 - Non-Diabetic

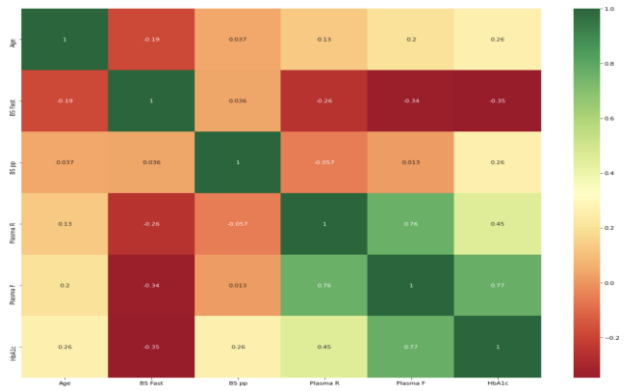

Fig. 2. Correlation Map.

B. Modelling Methods

In training the classification algorithms and constructing the model, the following steps were taken (see Fig. 3).

Step One: Import the modules and dataset as a data frame

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
warnings.filterwarnings('ignore', category=DeprecationWarning)
#data source: https://data.world/dabeleska/diabetes-type-dataset
df = pd.read_csv('/content/Diabetestype.csv')
```

Fig. 3. Code Snippet to Import Dataset.

Step Two: Get insights from data

The Insight derived from the dataset is shown in Table III

TABLE III. INSIGHTS GOTTEN FROM DATASET

	Age	BS Fast	BS pp	Plasma R	Plasma F	HbA1c	Type	Class
0	50	6.8	8.8	11.2	7.2	62	T1	1
1.	31	5.2	6.8	10.9	4.2	33	N	0
2.	32	6.8	8.8	11.2	7.2	62	T1	1
3.	21	5.7	5.8	10.7	4.8	49	N	0
4.	33	6.8	8.8	11.2	7.2	62	T1	1

Key: T1=Type1, N=Normal

Step Three: Specify features and test sets

A training set containing the first seven attributes of the data set is extracted and the test set which is the eighth attribute is also extracted, which is the product of the prediction and the entire dataset is split into a 7:3 ratio train set and test set.

Step Four: Train prediction model

The prediction model consists of the best machine learning model after implementing different models, and the best was taken and deployed for application. The output of each model is taken to the next stage for testing.

Step Five: Test model

The test set is used to assess the prediction models that have been created. This step is carried out four times to ascertain consistency.

Step Six: Evaluate

From the result of the test stage, the model is evaluated based on classification accuracy and specificity. The Table IV shows the accuracy of the four models based on these parameters.

1) Prediction methods for diabetes: The following machine learning strategies are used for comparative analysis of the diabetes predictive model. Classifiers include logistic regression, decision trees, random forests, and support vector machines.

a) Logistic Regression (LR): It is another supervised learning classification algorithm that models the relationship between a categorical response variable and its covariates. It computes probabilities using a logistic function, which is the accumulative logistic distribution, to assess the association between a categorical dependent variable and more than one independent variable. It is another probabilistic-based statistical model used in machine learning to solve classification problems. The logistic regression model uses the sigmoid function to predict the probability of outcomes of positive and negative class and can be derived from a sigmoid function obtained below,

$$\mathrm{P} = \frac{1}{1+e} - a + bx \quad (1)$$

where P = probability, a and b = parameter of Model.

b) Decision Tree Algorithm: A DT is one of the supervised machine learning algorithms that employ the classification regression trees algorithm, which can handle both classification and regression. It aids decision-making by generating a decision-tree-like model in which data is continuously split according to a specific parameter. There are two types of units in the tree: decision nodes and leaves. The data is split at the decision nodes, and the final decisions or outcomes are at the leaves. To solve classification and regression problems, the algorithm generates decision trees from training data. The classification error rate is defined as the proportion of the training set that does not belong to the most common class:

$$\text{Entropy (S)} = \sum_{i=1}^{n} - Pi\ Log(\ Pi) \quad (2)$$

where Pi is the percentage of the training set from the ith class in the region.

c) Random Forest (RF): It is one of the machine learning prediction algorithms. It lends itself better to the ensemble approach. It is capable of handling large datasets with ease. Random Forest is an ensemble classifier made up of many decision trees, with the ensemble implying that it employs multiple machine-learning algorithms to achieve predictive performance. It outperforms others in terms of diabetes mellitus prediction.

The following is the algorithm:

1 Create an N-Tree bootstrap sample using the input data.
2 Grow an unpruned regression for each bootstrap sample by splitting the node from all predictor nodes. Predictors select the best split from the input variables.
3 Predict new data by aggregating N-Tree predictions.

The random forest formulas are given below using Gini Index formulae for classification.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n}(Pi)^2 \quad (3)$$

It measures total variance across ith classes that true positives and true negatives. It should be noted that the Gini index is a measure of node purity that has a small value if all of the Pi are close to zero or one.

Support Vector Machine

A Support Vector Machine (SVM) is a type of supervised classification algorithm that has been widely and successfully applied to text classification tasks. This helps with regression and classification tasks and can work with multiple variables. This algorithm effectively performs nonlinear classification and also maps the inputs into a high-dimensional feature space that is used for classification, detection, and regression.

Step 1: Identify the appropriate hyperplane.

Step 2: Following the first step, the second step is to maximize the distances between neighboring data points.

Step 3: Insert a feature z $= x^2 + y^2$ It implies that SVM can solve such a problem.

Step 4:-Use an SVM classifier to classify the binary class.

SVM formulae are derived from the equation of hyperplane function to obtain the below,

$$W^* = \arg_w \text{Max} \frac{1}{||w||2} [Min\ Yn \mid W^T\ (\emptyset(x) + b] \quad (4)$$

Where $\arg_w$Max is an acronym for arguments of the maxima, which are simply the locations of a dynamic array domains where a function's particular value is maximized. The inner phrase [Min Yn | WT ((x)+b] essentially indicates the shortest distance between two points and the closest point to the decision boundary.

IV. Results and Discussion

This section aims to get acquainted with results obtained after performing various activities on the dataset obtained from the dataset of Abelvikas, Data World. Fig. 4 shows the registration page for the patient.

A. Results

The implementation tools used in this research are Python programming language, google collaboratory, and libraries containing algorithms used for artificial intelligence development, and Anaconda houses a large amount of these libraries. Fig. 5 depicts the recent patient, daily added patient and diabetes rate charts. Fig. 6 shows the disease diagnosis and report.

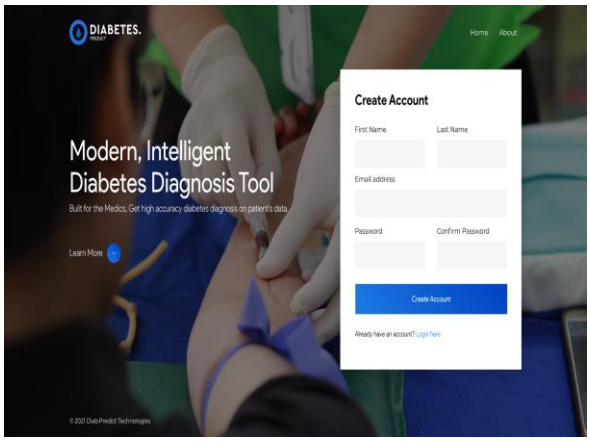

Fig. 4. Landing Page and Registration Page.

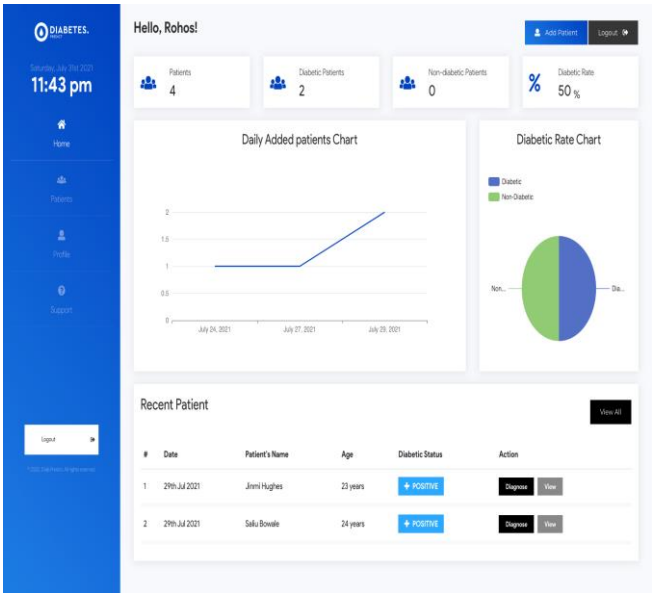

Fig. 5. Dashboard Page.

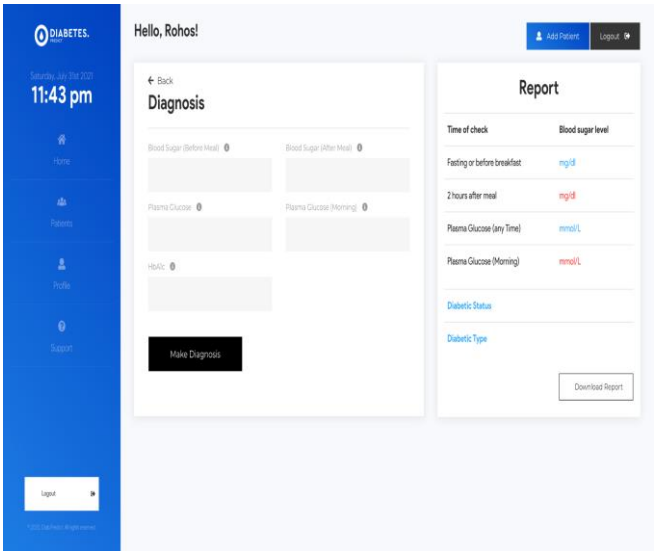

Fig. 6. Diagnose Patient Page.

1) Performance metrics: The classifiers we used were then applied to the dataset individually and ran five iterations to ensure that the results obtained from the average of each implementation of a particular algorithm are accurate. Also, these tests were done on randomly selected samples of the dataset to avoid the problem of overfitting. Various parameters were used to evaluate the system, but for this research, three performance indexes were used: Sensitivity (SE), Specificity (SP), and accuracy, as shown in equations (5)–(7). True positives (TP) and true negatives (TN), as well as the false positives (FP) and false negatives (FN).

$$(SE) = \frac{no_{of\_predicted\_true_{positive}}}{true\_positive + false\_negative} \quad (5)$$

$$(SP) = \frac{no_{of\_predicted\_true_{negative}}}{true\_negative + false\_positive} \quad (6)$$

$$Accuracy = \frac{number\_of\_correct\_predictions\_}{total\_number\ of\_prediction} \quad (7)$$

The prepared model was integrated into a Python Web Framework, and Flask Framework and hosted on a server for testing. To test the solution, random records from the dataset were used, and an average of the following was calculated for each algorithm. From the above list, it is shown amongst our Ensemble of algorithms why the Random forest algorithm was chosen as the eventual algorithm used for the implementation of this work, as it has the highest average accuracy among the four algorithms.

2) Confusion matrix evaluation: A confusion matrix is also referred to as a contingency table or error matrix, used to visualize the performance of a classifier, it's a good way of evaluating a good effective classification model. This means that the high performance of any classification model can be visualized in its confusion matrix having a strong main diagonal shown in Fig. 7.

3) Implementation of confusion matrix: The confusion matrix was implemented for each algorithm in the ensemble of algorithms leading to the results are shown in Table IV, while Table V shows the confusion matrix for Classification Models using Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT).

The above shows the result of the confusion matrix for classification algorithms with 70% training data and 30% testing data of 1009 records. The result yielded the Table V below.

$$y = \frac{1}{N}\sum_{i=1}^{5} Xi \quad (8)$$

where y is the mean, Xi is the result of the confusion matrix and the i-th attribute value of the no of iterations.

$$y = \frac{1}{N}\sum_{j=1}^{3} yi \quad (9)$$

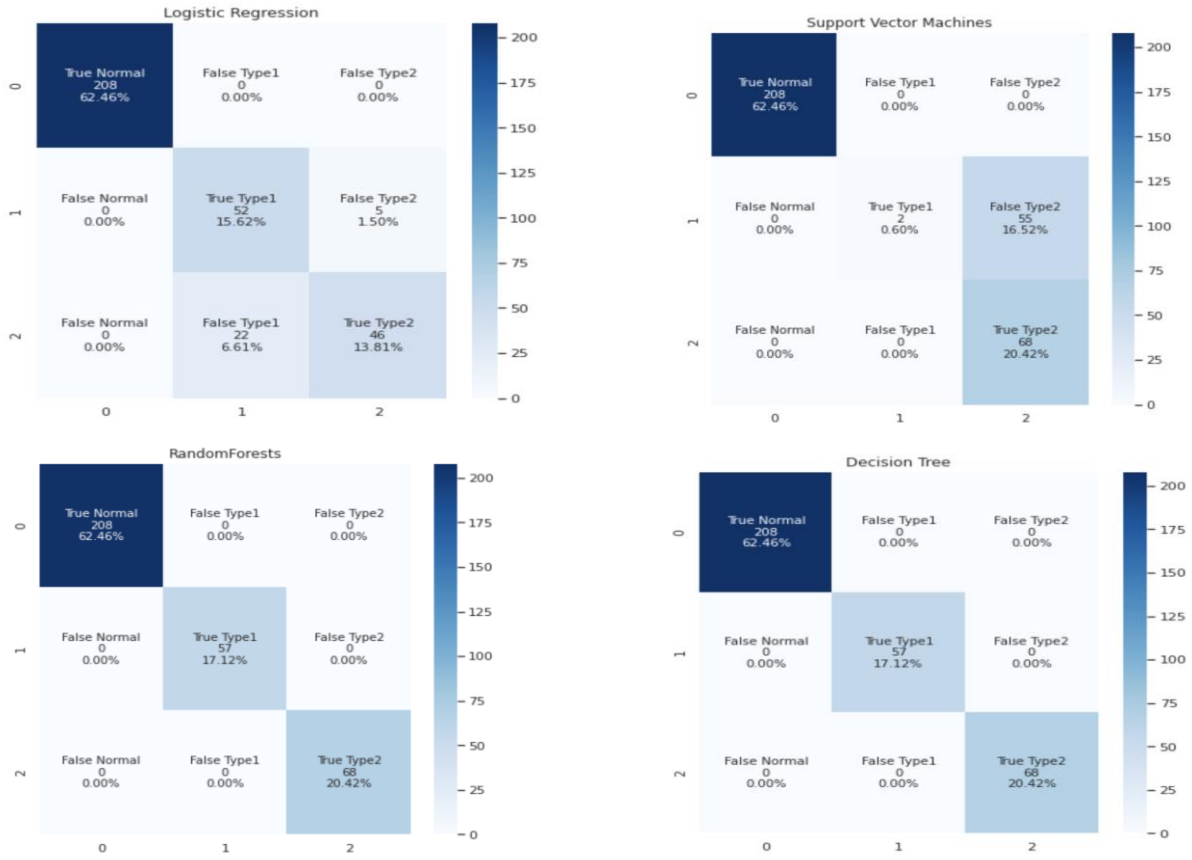

Fig. 7. Confusion Matrix for Classification Models using (i) LR (ii) SVM (iii) RF (iv) DT.

TABLE IV. PERFORMANCE OF THE STUDIES CLASSIFICATION MODEL USING NORMAL, TYPE 1 AND TYPE 2 DIABETES

Algorithm	Class	1st Iteration	2nd Iteration	3rd Iteration	4th Iteration	5th Iteration	Mean Accuracy
LR	Normal	0.990099	0.955446	0.955446	1.000000	0.940594	0.997030
	Type 1	0.940594	1.000000	0.940594	0.940594	0.995050	0.955426
	Type 2	0.960396	0.955446	1.000000	0.980100	0.980100	0.954436
RF	Normal	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
DT	Normal	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
SVM	Normal	0.995050	0.831683	0.826733	1.000000	0.856436	0.998020
	Type 1	0.856436	1.000000	0.851485	0.851485	0.995050	0842417
	Type 2	0.831683	0.826733	1.000000	0.840796	0,840796	0.840436

Key: LR=Logistic Regression, RF= Random Forest, DT=Decision Tree, SVM = Support Vector Machine

where y is the mean average, y*i* is the mean of the result of the confusion matrix, and the *i*-th attribute value of the number of classes.

TABLE V. CONFUSION MATRIX DATA FOR CLASSIFICATION MODELS USING (I) LR (II) SVM (III) RF (IV) DT (%)

Classifier	Normal	Type1	Type2
LR	62.46	22.23	15.31
SVM	62.46	0.60	36.94
RF	62.46	17.12	20.42
DT	62.46	17.12	20.42

B. Discussion

Diabetes has recently become one of the leading causes of death in humans. Diabetes is becoming more common every year for a variety of reasons, including poor eating habits, and the prevalence of unhealthy foods. Diabetes detection early on can help with clinical management decision-making. We have employed numerous measures of evaluation throughout this research to determine and quantify the performance of each algorithm in our ensemble of algorithms, which comprises the Logistic Regression algorithm, Decision Tree, Random Forest, and Support Vector Machine Classifier algorithms, all these algorithms were tested on the diabetes dataset of Abelvikas in five iterations, and the result of the test gave a model that we eventually used for the implementation. However, with all these algorithms it was important to realize which was the most effective of all them, and this was achieved by getting an accurate reading of each and, including algorithm over five iterations. An average of the accuracy reading from each algorithm was used as a measure to determine the eventual algorithm that was used to form our model (Fig. 9), which turned out to be the Random forest and Decision tree Algorithms. From Table VI, the outcomes of the average of the accuracy tests on each algorithm are displayed, this also includes the specificity accuracy and sensitivity accuracy as well as the classification accuracy. When we compare the values in Tables V and VI, we see that the classification results after the confusion matrix are similar to the classification model results. Examining the confusion matrix revealed the same similarity. Table VII shows the mean average score of the algorithms.

In Fig. 8 the use of an ensemble of algorithms aids data mining in determining the most effective algorithm that can be used to generate an effective model. The accuracy report obtained from multiple tests shows that the random forest and decision tree algorithms on our dataset proved to be better prediction algorithms than the other algorithms. The results were compared to the results of works of literature. The Table VIII demonstrated that the developed system's accuracy was higher than [28] accuracy of 91.32 percent because RF excels at working with non-linear data, constructing multiple decision trees, and merging them to produce a more accurate and stable prediction with improved performance.

TABLE VI. COMPARISON OF RESULTS OF DIFFERENT CLASSIFIERS

Metrics Average	Logistic Regression	Decision tree Classifier	Random Forest	Support Vector Machine
Accuracy (%)	92	100	100	83

TABLE VII. THE MEAN AVERAGE SCORE OF THE ALGORITHMS

Metrics Average	Logistic Regression	Decision tree Classifier	Random Forest	Support Vector Machine
Sensitivity	0.921705	1.0	1.0	0.692391
Specificity	0.980548	1.0	1.0	0.933251
Accuracy	0.968964	1.0	1.0	0.893624

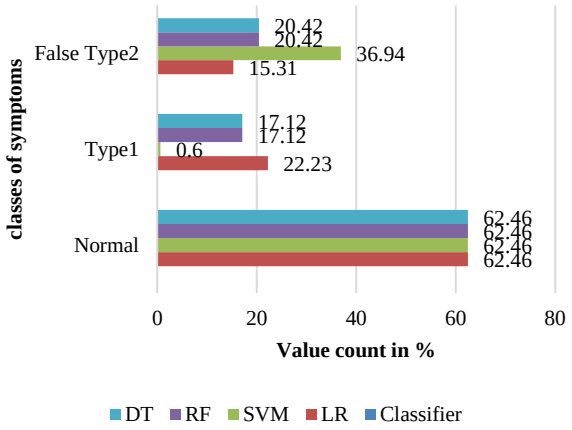

Fig. 8. Confusion Matrix Data for Classification Models.

TABLE VIII. RESULTS COMPARISON TABLE

Author	Model / Method	Dataset Used	% Accuracy
1. Deepti & Dillip. 2018	Naive Bayes & SVM	PIMA Indian Diabetes dataset	76.3%
2. Radha, et al. (2014)	C4.5	A hospital repository	86%
3. Song et al.. (2017)	ANN	Small undefined number of data	74.8%
4. Rashid, & Abdullah, 2016	Decision Tree	A hospital repository	75.5%
5. Afrand, (2012)	Combination of Classifier algorithms	A hospital repository	91.3%
6. Adidela (2012)	Fuzzy 1D3 and Estimation maximization algorithm	A private hospital Repository	91.3%
7. Developed System	LR RF DT SVM	Abelvikas, Data world	92% 100% 100% 83%

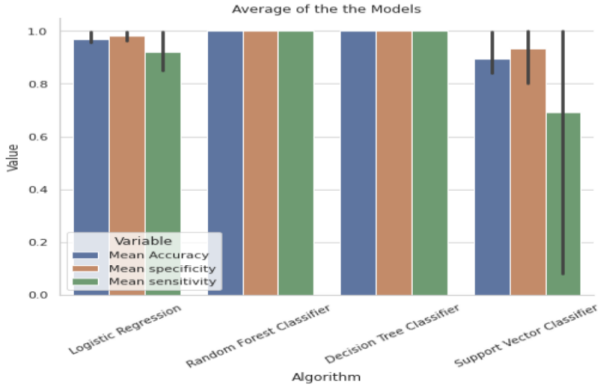

Fig. 9. Average Accuracy of the Models.

V. CONCLUSION

This study compares and evaluates the performance of four machine learning algorithms in the classification of diabetes. The Abelvikas datasets from the Data World repository are used to train and test the system. For diabetes classification, a host of machine learning models have been applied with 1009 instances and eight critical variable features were extracted and identified: age, blood sugar in fasting, blood sugar after a meal, plasma glucose in fasting, plasma glucose, glycated hemoglobin, type, and class. The results of the analysis revealed that the Random forest and Decision tree models were the most accurate in predicting diabetes. The system developed ensures a stable prediction. As a result, the models can be more effectively applied to other diseases. A combination of algorithms, rather than just the most performant algorithm in the ensemble, may be more beneficial in the future.

REFERENCES

[1] W.H.O. "About diabetes". World Health Organization, 2014

[2] A. Krasteva,., V. Panov., A. Krasteva., A. Kisselova, and Z. Krastev, "Oral cavity and systemic diseases-Diabetes Mellitus." Biotechnol. Biotechnol. Equip. 25, 2183-2186. Doi: 10.5504/BBEQ.2011.

[3] Mahmud, S M Hasan, Hossin , Md Altab, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkarm. "Association for Computing Machinery. ACM ISBN 978-1-4503-6582-6/18/08 DOI: https://doi.org/10.1145/3297730.3297737n, 2018.

[4] Angela Betsaida B Laguipo. "COVID-19 could trigger diabetes in Healthy people". News Medical Life Science, 2020.

[5] Lee, Yong-ho, Ban,g Heejung and Kim Dae Jung, "How to establish Clinical Prediction Model", Journal List Endocrinol Metab, 2016.

[6] P. Samant., and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images. Computer Methods and Programs in Biomedicine". 157, 121–128. DOI: https://doi.org/10.1016/J.CMPB.2018.

[7] Glauco Cardozo , Guilherme Brasil Pintarelli , Guilherme Rettore Andreis ,Annelise Correa Wengerkievicz Lopes, and Jefferson Luiz Brum Marques, "Use of Machine Learning and Routine Laboratory Tests for Diabetes Mellitus Screening, Hindawi BioMed Research International Volume, pp1-14, 2022.

[8] M. E. Hossain,. A. Khan, M. A. Moni, and S. Uddin, "Use of electronic health data for disease prediction: a comprehensive literature review," Transactions On Computational Biology And Bioinformatics, vol. 18, no. 2, pp. 745–758. 2021.

[9] De Silva K., N. Mathews, H. Teede et al. "Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: a retrospective cohort analysis using machine learning and unstructured big data," Computers in Biology and Medicine, vol. 132, article 104305. 2021.

[10] Wu et al. (2021) Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning, The Journal of Clinical Endocrinology & Metabolism, Vol. 106, No. 3, e1191–e1205 doi:10.1210/clinem/dgaa899 Clinical Research Article.

[11] Aishwarya, Mujumdar, V. Vaidehi "Diabetes prediction using machine learning algorithms International Conference on Recent Trends in Advanced Computing", ICRTAC, 2019.

[12] Minyechil, Alehegn, Rahul, J. & Dr. Preeti, M. "Diabetes Analysis And Prediction Using Random Forest, KNN, Naive Bayes, And J48: An Ensemble Approach".*International Journal of Pure and Applied Mathematics*, Volume 118 No. 9,871-878m, 2019.

[13] Nail, Rachel, and Suzane Falck, "An overview of diabetes types and treatments".https://www.medicalnewstoday.com/articles/323627, 2020.

[14] Vijayakumar, Kavin Prasad Arjunan, Manivel Sivasakthi, Karthikeyan Lakshmanan "Diabetes Prediction By Machine Learning Over Big Data From Healthcare Communities", International Research Journal of Engineering And Technology(Irjet)E-Issn: 2395-0056volume: 06 Issue: 04| Apr2019.

[15] Thomas, Jesia , Anumol Joseph, Irene Johnson, Jeena Thomas, "Machine Learning Approach For Diabetes Prediction" International Journal of Information Systems and Computer Sciences: Volume 8, No.2, 2019.

[16] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao "Predictive models for diabetes mellitus using machine learning techniques" Lai et al. BMC Endocrine Disorders https://doi.org/10.1186/s12902-019-0436-6, October 2019.

[17] H. R. Divakar, D Ramesh, B R Prakash "An Ontology-Driven System to Predict Diabetes with Machine Learning Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, pp 4005-4011, Vol-9 Issue-2, 2019.

[18] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang. "Predicting Diabetes Mellitus with Machine Learning techniques". https://dx.doi.org/10.3389%2Ffgene .2018.

[19] Deepti S. & Dilip S. S. "Prediction of Diabetes using Classification Algorithms". *International Conference on Computational Intelligence and Data Science*, pp 1578-1585 *(ICCIDS 2018),*

[20] Farooqui, Ritika, and Tyagi, "Prediction Model for Diabetes Mellitus Using Machine Learning Techniques. International Journal of Computer Sciences and Engineering Open Access Research Paper Volume-6, Issue-3 E-ISSN: 2347-2693, 2018.

[21] Mahmud, S M Hasan, Hossin , Md Altab, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkarm. "Association for Computing Machinery. ACM ISBN 978-1-4503-6582-6/18/08 DOI: https://doi.org/10.1145/3297730.3297737n, 2018.

[22] K. Nandhini.M, "Prediction of diabetes using Classification algorithms," International Journal of Science, Engineering and Management, vol. 2, no. 12, pp. 287-291, 2017.

[23] Zheng, Tao, Wei Xie , Liling Xu , Xiaoying He, Ya Zhang, Mingrong You, Gong Yang , You Chen "A machine learning-based framework to identify type 2 diabetes through electronic health ", 2017.

[24] F. M. Okikiola, O.S. Adewale, A.M. Mustapha, A.M. Ikotun, O.L. Lawal "A framework for Ontology-based diabetes diagnosing system using bayesian optimization technique. https://doi.org/10.51406/jnset.v17i1.1906. 2018.

[25] Sakorn Mekruksavanich."Medical expert system based ontology for diabetes disease diagnosis"IEEE International Conference on Software Engineering and Service Science (ICSESS) pp 383-389, 2016.

[26] Rian Budi, Lukmanto E.Irwansyah The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model," Procedia Computer Science Volume 59, 2015, Pages 312-319.

[27] I. Aiswarya., S. Jeyalatha., S. Ronak.," Diagnosis of Diabetes using classification mining techniques". *International Journal of Data Mining & Knowledge Management Process (IJDKP),* Vol. 5, No. 1, pp. 1-14, 2015.

[28] D. R. Adidela, "Application of fuzzy ID3 to predict diabetes." Int J Advanced Computer Math Sci 3.4.541-5., 2012.