https://doi.org/10.46335/IJIES.2023.8.3.5 **e-ISSN: 2456-3463**
Vol. 8 , No. 3, 2023, PP. 24-34

International Journal of Innovations in Engineering and Science, www.ijies.net

Effectiveness of Machine Learning & Deep Learning Models for Diabetes Prediction

Priyabrata Sahu[1], Jibendu Kumar Mantri [2]

[1] *Research Scholar , P G Department of Computer Application , MSCBD University , Odisha , India*
[2] *Associate Professor, P G Department of Computer Application , MSCBD University , Odisha , India*

priyabsahu@gmail.com

***Received on**: 29 March,2023* ***Revised on**: 09 April,2023,* ***Published on**: 11 April,2023*

Abstract: *Hyperglycemia alters blood sugar levels. Hyperglycemia, often known as high blood sugar, is the result of uncontrolled diabetes, which may cause nerve and blood vessel problems. Hyper-glycemia, or high blood sugar, is a typical result of insufficient glucose management and is associated with several significant health complications, most notably those affecting the nerves and blood vessels. Machine learning (ML) and deep learning (DL) predictive models have seen tremendous development throughout industries, including health care, making early diagnosis of diabetes a breeze. The treatment of chronic diabetes, one of the world's most prevalent illnesses, might benefit greatly from improved diagnostic efficiency. Here, we examine the relative merits among several ML and DL approaches to the problem of early diabetic illness prediction. The primary objective of this research study is to organize and conduct out diabetes diagnosis and prognosis with several machine learning approaches and then analyze the results of these methods to determine which one is the most accurate classifier. In this work, we take a multifaceted approach to diabetes and its prediction by investigating a wide range of disease-related characteristics. Many Machine Learning classification methods, including Random Forest (RF), Logistic regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Decision Tree (DT), Gradient Boosting, are applied to the canonical Pima Indian Diabetes Dataset (PIDD) (GB). There is a wide range of precision amongst the models used here. A technology that can accurately predict diabetes is shown in this research. The results of this research indicate that one of the Data mining models, random forest network models have superior accuracy in making diabetes forecasts.*

Keywords—Diabetes prediction; Machine learning; Deep learning; Classification; Artificial Neural Network (ANN).

I. INTRODUCTION

Diabetes (DM), the most common condition that impairs insulin production and response, may raise blood glucose levels. According to global diabetes statistics over 382 million individuals worldwide were diagnosed with diabetes in 2013 [1,2]. It was the fifth highest cause of mortality for women in 2012, and it was the eighth top cause of death overall that year. Diabetes may cause CAD, CKD, HTN, and hypothyroidism. Thus, the early identification and treatment of these diseases may help patients recover [2]. T1D and T2D are the two types of diabetes (T2D). Type 1 diabetics are usually under 30. High blood glucose, thirst, and peeing are typical symptoms [4]. Patients need insulin for this kind of diabetes. Those over the age of 50 have a greater risk of developing type 2 diabetes, which is associated with overweight or obese mellitus, cardiovascular, and other illnesses [5]. Diabetes kills many people worldwide. Diabetes early detection may save lives. This study predicts diabetes using diabetic symptoms. We predict diabetes using the "Pima Indian Diabetes Dataset" PIDD with ML classification and outfit methodologies. ML instructs computer systems or machines explicitly. Designing grouping and fitting algorithms from datasets with various ML approaches yields excellent knowledge

https://doi.org/10.46335/IJIES.2023.8.3.5 e-ISSN: 2456-3463
Vol. 8 , No. 3, 2023, PP. 24-34

gathering. Finding the correct machine learning algorithm for predicting is tricky. Thus, this research utilized "Random Forest (RF), Logistic regression (LR), Support Vector Machine (SVM), Multilayer perceptron (MLP), Decision Tree (DT), Gradient Boost (GB)" algorithms to predict diabetes and evaluates their effectiveness.

Diabetics cannot utilize insulin, a hormone produced by pancreatic islets (eyelets) [6]. It is mainly the primary reason of heart disease, amputation, renal failure, and early death [7]. Type-1, type-2, and gestational diabetes are most frequent. Due to significant advances in deep learning and data availability, diabetes diagnosis, glucose control, and complications assessment may be predicted. Contemporary deep learning frameworks help diagnose diabetes [8]. Performance parameters like "Precision, recall, f1-score, execution time, and ROC" value can be employed to discover the most accurate method, even though accuracy is the best approach to do it.

Diabetics may benefit from early recognition and treatment. That dataset was classified using a classification method. Thus, we analyzed classifiers like GBoost, DT, RF, SVM, MLP, and LR to choose the best method. We employed 520 data from a benched-mark UCI repository related dataset [9] with 16 characteristics, 416 for training along with 104 for testing. We also compare performance using accuracy, recall, f1-score, processing time, and ROC value. RF classifier is best for early diabetes prediction with 74 % accuracy.

The primary goals of this research were to:
a. Assess the availability of publicly accessible datasets in diabetes research.
b. Conducted an in-depth comparison of ML and DL methods.
c. Early diabetes detection effectiveness assessment using performance metrics.
d. The next steps in the field's research, which must be taken by the next generation of experts.

Following this introduction part and outline, the rest of the proposed research paper is basically composed of where in Section II, we review the relevant literature. In section III, we provide a high-level summary of our study procedures and methodologies. Section IV details the proposed approach and Performance Evaluation Metrics, whereas Section V presents experimental results. The results and suggestions for further research are highlighted in the last section, VI.

II. LITERATURE REVIEW

In this Research work , we emphasize the work of a select group of academics those who have primarily contributed substantial findings to the field of diabetes mellitus prognostication by mining public health related medical data for insights with the help of machine learning ML models and deep learning DL models. The next portion of the article discusses some of the studies conducted to identify or forecast diabetes via the use of machine learning. By combining the RF approach with other machine learning methods such as SVM, DT, KNN, RF, LR and GB the authors of [10] were able to reach an accuracy of 77%. As the authors of [11] describe State vector machines and K- nearest neighbour classification Approaches here provide the best accuracy of diabetes forecasting. This 768-record sample provides an accuracy of 77%. Paper [12] presents Prediction of diabetes with the use of machine learning ML techniques and proposes to foresee diabetes via three classification methodologies : SVM, LR and ANN. This research recommends an effective method for early detection of diabetes. According to the authors, the LDA approach was presented in article [13], and the authors subsequently merged the SVM classifier with Feed Forward Neural Networks to create a classification method. The SVM classifier has an accuracy of 75.65%. The KNN, Naive Bayes, and RF classification models created by the authors of study [14] achieved final accuracy rates of 66.19 percent, 72.66 percent, and 73.72 percent, respectively. The Weka system relied on them. Naive Bayes, SVM, and ANN classifiers are all compared in study [15], which uses a diabetic dataset for testing. They conducted a study adjusting for body mass index, from which they inferred that the likelihood of developing diabetes was high. Ultimately, they arrived to the conclusion that mixing models improves classification accuracy over using only one. Using the diabetes dataset, the study [16] offered a predictive analytic model based on J48 (C4.5), K-Nearest Neighbours (KNN) classifier, Random Forest (RF) classifier, and Support Vector Machines. They anticipated that the J48 method would outperform the others by an accuracy of 73.82% before pre-processing the data, but that the KNN and Random Forest algorithms would achieve superior accuracy after pre-processing.

Alternatively, Naiarun et al. [17] developed an application for the aim of predicting diabetes that is hosted on the web using accuracy of prediction as a key factor. They then examined numerous prediction

https://doi.org/10.46335/IJIES.2023.8.3.5 **e-ISSN: 2456-3463**
Vol. 8 , No. 3, 2023, PP. 24-34

methods from machine learning and deep learning, such as RF, DT, LR, CNN, NB, bagging, and boosting. In the end, they discovered that Random Forest was superior, with an accuracy of 85.55% and a ROC value of 0.912. Paper [18] describes an assessment of three machine learning methodologies , including Logistic regression (LR), Naive Bayes (NB), and State vector machines (SVM) , utilising a 10-fold cross-validation evaluation technique. In which SVM's 84% accuracy much surpassed that of competing methods. In addition, the author of research [19] employed many machine learning methods to make a pre-diagnosis of diabetes mellitus and to show where enhancements may be made to the filtering procedures. The experiment used a 10-fold cross-validation method using a diverse set of algorithms, including RF, KNN, SVM, NB, DT, and LR. Researchers then, as described in article [20], created a model based on the J48 Decision Tree (DT) classifier for the management of patients with type 2 diabetes. They looked at the different factors like age, gender, renal issue, smoking, hypertension, cardiac problem, and diabetes as well as seven individual patient characteristics. Both the ROC value (0.624) and the accuracy rate (70.80) have been exceeded by their findings. In order to guarantee the use of important variables and produce findings employing techniques for machine learning, they found results that were comparable to clinical ones., the authors of article [11] developing a novel classification algorithm using a dataset for predicting type 2 diabetes's early stages [9]. They developed three different machine learning algorithms for diabetes mellitus prediction, including classifiers like a "Random Forest (RF), a Multi-layer Perceptron (MLP), and a Radial Basis Function Network (RBF)". They demonstrated that the RBF algorithm is superior, with a 98.80 percent success rate. Many machine learning and deep learning categorization approaches were used to increase performance in the aforementioned research. Most research employed some of the performance measures including F-score, ROC-score accuracy, precision, along with execution time to evaluate different methods and choose the optimum one. Our study concludes with a comparison of state-of-the-art categorization strategies based on machine learning that are used for early stage diabetic mellitus prediction utilising a variety of risk variables applied to actual diagnostic healthcare records.

III. METHODOLOGY

The primary objective of this research work is to organize, implementation and further analyze the results of Diabetes Prediction using many Machine Learning approaches in order to identify the most effective classifier. Along this, we briefly discuss the steps. For an overview of the suggested approach for diabetes forecasting, see Figure 1.

3.1. Dataset Description

In this study, we make use of the Pima Indian Diabetes Dataset (PIDD). The UCI AI respiratory dataset is publicly available and may be accessed from their website (Dataset, PIDD). This dataset has 9 characteristics, which included an outcome attribute, and 768 entries. Two hundred and sixty-eight of the findings indicate that the patient does indeed have diabetes, whereas the remaining five hundred reports indicate that the patient does not have diabetes. The positive results account for 268 of the total 768 reports.

3.2. Data Pre-processing

The Indian Diabetes dataset has no missing value (NAN) values, although several useless characteristics have zero values. We calculate the required mean and median of all columns with zero values for diabetes and non-diabetic patients. Diabetic and non-diabetic patients substitute zero. We employed 70% of the standardized PIDD for validation and training and 30% for assessment. Python is used for programming the model.

3.3. Algorithms used for Classification.

Once our dataset is ready, we apply Machine Learning techniques to it in order to categorise it. In this research work, we incorporate the LR, RF, SVM,MLP, DT, and GB algorithms using features including pregnancy, glucose, blood pressure, skin thickness, insulin, body mass index, pedigree, and age, as well as two additional functionalities that it has been extracted from the dataset using Exploratory Data Analysis-EDA. A diabetic is defined as any individual with a blood pressure (BP) of more than 80 and a Anyone with a blood pressure reading above 80 is also considered diabetes. The ANN classification method achieves a maximum accuracy of 76% when used with the aforementioned strategies.

3.3.1. Random Forest (RF) : RF is a simple ML method that produces good results without meta parameters. Due of its flexibility and applicability, this technique is widely utilised classification and regression.

https://doi.org/10.46335/IJIES.2023.8.3.5 e-ISSN: 2456-3463
Vol. 8 , No. 3, 2023, PP. 24-34

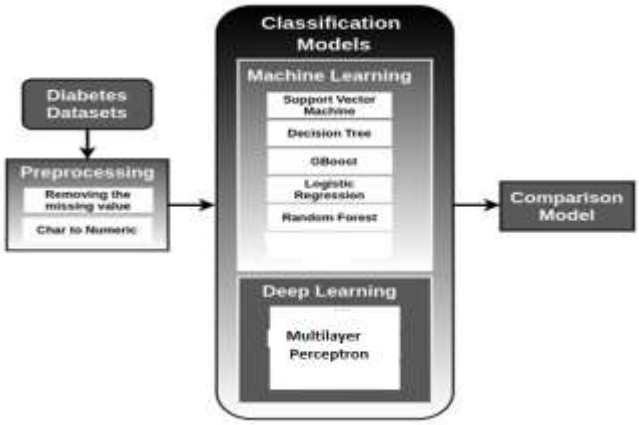

Fig 1. Proposed Research Model

Table 1. PIDD Dataset Parameters

Number	Attributes	Description
1	Pregnancies	Number of times pregnant
2	Insulin	2-Hour insulin serum (µU/ml)
3	BMI	The index of body mass
4	Age	The Age (years)
5	Glucose	Concentration of plasma glucose for 2 hours in an oral glucose tolerance check
6	Blood Pressure	Blood Pressure Diastolic (mm Hg)
7	Diabetes PedigreeFunction	Diabetes pedigree function
8	Skin Thickness	Skinfold triceps thickness (mm)
9	Outcome	Range of value: 0 and 1(0 means no 1 means yes)

This method generates forests randomly. The created "forest" is a "Decision Trees" band. This method handles massive datasets. Leo Bremen developed Random Forest. It randomly chooses dataset samples and creates decision trees for each. Decision trees forecast. After voting, use the model with the most votes [21]. A random forest (RF) classifier is basically a group of tree-based classifiers that, given a set of random, independent vectors that are uniformly distributed, each vote for the category that best fits the majority of the data in regression and classification tasks [22]. This classifier is a basic, adaptable machine learning technique that usually gives great results requiring extra parameters. This research tests the RF approach using PIDD characteristics.

3.3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is based on ML-machine learning. In 1963 Vapnik and Chervonenkis invented SVM. The SVM seeks a hyperplane that isolates any class's instances. This classifier basically defines the hyperplane that eliminates spots to place most of the significant no. of points of a related similarity class on a relevant similar side while expanding the duration of each class to such a related hyperplane. The nearest hyperplane points are support vectors. The smallest gap between a class and a hyperplane is its spots [23]. Grouping and regression are also possible using the hyperplane. SVM groups instances and classifies compounds without data. Hyperplane plays out the division to any group's closest training place to disconnect. This study tests the Support Vector Machine technique using Pima Indians Diabetes dataset characteristics.

3.3.3. Decision Tree (DT)

Basic classification and regression is DT. A tree-structured DT model may categorise instances by attributes (Quinlan et 1986) [23]. Decision trees are used for categorical results. Decision trees give nominal and numerical properties. It can handle fluctuating values. The decision tree categorises the qualifying dataset by splitting the nodes from the topmost to the class node. Each node indicates a possible response for the instance's test property. Decision trees may simply create significant patterns from the top node to the attack class node. Decision trees (DTs) are supervised learning methods used for categorization, discriminate, and predictive modelling [24]. Each Decision tree (DT) node, or Decision node as well as Leaf node, tests a characteristic to construct one of the tree's two branches. This research tests the Decision tree algorithm using PIDD characteristics.

3.3.4 Logistic Regression:

Logistic regression- LR, often referred as a basically Logit Model, models binary response variables statistically. Logistic regression (LR) and logistic regression employ linear regression to estimate the probability of one outcome class relative to another, such as good or bad treatment [25]. Logit is the regression co-efficient β in the basic logistic model [26].

$$\text{Logit } (Y)= \text{naturallog(odds)}=\ln((\pi/(\pi-1))=\alpha+\beta x \quad (1)$$

3.3.5 An Overview of Boosting Methods:

The purpose of ensemble learning approach is to train the model using as many different kinds of learning algorithms as possible. The Bagging technique is an example of ensemble learning in which several models are simultaneously applied to independent subsamples of the same dataset. In contrast to parallel construction, the boosting technique tries to train both the methodology

https://doi.org/10.46335/IJIES.2023.8.3.5 **e-ISSN: 2456-3463**
Vol. 8 , No. 3, 2023, PP. 24-34

and the model, and it is also widely employed in practise. The model is trained using a simple technique, then restructured in light of the training outcomes to facilitate learning. The updated model is passed on to the subsequent method, which benefits from the simpler learning. This article presents a variety of boosting techniques, each of which offers a unique take on the sequential approach.

Adaboost was getting better by adding the decision stump to its most recent update to its weighting system (1 node divided into 2 leaves). Gradient boost [27], which is another sequential approach, makes trees bigger because the loss is optimised by making 8 to 32 leaves. (Tax loss: Look at the residual in linear models. The residual error is equal to (y test-y prediction), and the loss is equal to the sum of the squares of all the data points. Why is the square used? Since the target value is the difference between what was predicted and what actually happened, forecast errors are very important. Even if a negative number is not zero, squaring it would lead to a small loss, so negative numbers are squared. In short, the next technique is given a set of residual values, which are then lowered so that they can be sent to the next algorithm.

3.3.6 Multi-layer Perceptron: This is Deep Learning model. The Multi-Layer Perceptron (MLP) classifier is an FFANN is made up of more than just two levels; the input layer is the first one, and the output layer is the last one. The HIDDEN layer is an additional, more advanced layer sandwiched between the INPUT and OUTPUT layers. Increases in the number of layers will result in a proportional rise in the temporal complexity. For illustration, each neuron takes in data in the form of X1, X2, X3,..Xn, and the bias and weight are respectively indicated by (b) and (w); the product of the input and weight is the output, which in turn be represented as y depending on the activation function() [28].

$$y = \phi(\sum_{i=1}^{n}((w_i \cdot x_i) + b)) \quad (2)$$

IV. PROPOSED METHODOLOGY AND PERFORMANCE METRICS

Figure 1 shows our methodology , which achieves our research study goal. We first pre-processed the diabetes dataset. Tenfold cross-validation separated the dataset into train and test sets after pre-processing. Then, the recommended methods are basically used on the training set as a diagnostic tool for diabetes mellitus in its early stages. Finally, assessment metrics are utilised to compare effectiveness on the test set. This section briefly discusses these periods.

*4.*1 Dataset and Attributes

Here, we examined impact of machine learning ML and deep learning DL methods in the early phases of diabetes detection using the UCI repository diabetes dataset [9]. From PIDD dataset to obtain this data from 768 individuals who were either newly diagnosed with diabetes or displaying symptoms that are associated with diabetes. There are 9 characteristics, some of which are positive, and other of which are negative; the positive and negative indicators are used to evaluate the patient's likelihood of acquiring diabetes.

4.2 Pre-Processing:

Addressing missing values in the pre - processed data was an essential part of the data pre-processing that allowed us to reach our study aim. For instance, Prediction of diabetes using machine learning ML and deep learning DL are not suitable for use with minimal values of the attributes. We quantify nominal attribute values, such as "male" and "female" in the sex category, "yes" and "no" in the other attributes category, and "positive" and "negative" in the class category, by assigning a 1 to "yes" and a 0 to "no."

4.3 Evaluation Metrics for Performance Measurement
After the process of cross validation of recommended procedures in this research work , we will require some means of assessing how well they functioned. In this study, we employed a variety of standard criteria for assessing the efficacy of classification systems to assess the results of our studies. Important performance Metrics like as precision, recall, f1-score, ROC-curve, and accuracy are used to determine a model's level of predictive performance [29].

Precision: Precision is determined by dividing the number of accurate diagnoses by the combined total of correct and incorrect diagnoses.
Recall: Recall is calculated by dividing the number of correct responses by the total number of responses.
F1-score: Geometric average of accuracy and recall.
Accuracy: Divide the number of accurate predictions by the total number of guesses shown below.

These performance metrics Accuracy, precision, recall, and F1-score value are used to assess machine learning algorithms [30]. Our confusion matrix assessed

https://doi.org/10.46335/IJIES.2023.8.3.5 e-ISSN: 2456-3463
Vol. 8 , No. 3, 2023, PP. 24-34

accuracy, F1-score, recall, and precision for each classification system. The machine learning confusion matrix shows algorithm performance. User input datasets affect performance [31].

V. RESULTS AND DISCUSSIONS

In this research, we used many categorization methods to make diabetes prognoses. The suggested method uses several classifiers including RF, SVM, GB, LR , DT, and MLP with features including pregnancy, glucose, blood pressure, skin thickness, insulin, body mass index, pedigree, and age, as well as two additional features extracted from the dataset using Exploratory Data Analysis. Anyone with a blood pressure reading above 80 is also considered diabetes. For making Analysis of the results , it is done by performing five tasks.

Task 1 - Importing libraries and dataset,
Task 2 - Exploratory Data Analysis (EDA) ,
Task 3 - Data Preparation for model evaluation ,
Task 4 - Model Evaluation ,
Task 5 - Final Results Summary

Task 1 - Importing libraries and dataset (Table 2)

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.63	50	1
1	1	85	66	29	0	26.6	0.35	31	0
2	8	183	64	0	0	23.3	0.67	32	1
3	1	89	66	23	94	28.1	0.17	21	0
4	0	137	40	35	168	43.1	2.29	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.17	63	0
764	2	122	70	27	0	36.8	0.34	27	0
765	5	121	72	23	112	26.2	0.24	30	0
766	1	126	60	0	0	30.1	0.35	47	1
767	1	93	70	31	0	30.4	0.32	23	0

Task 2 - Exploratory Data Analysis (EDA) and Bar Plot to check for unique number of values in each feature of the dataset .

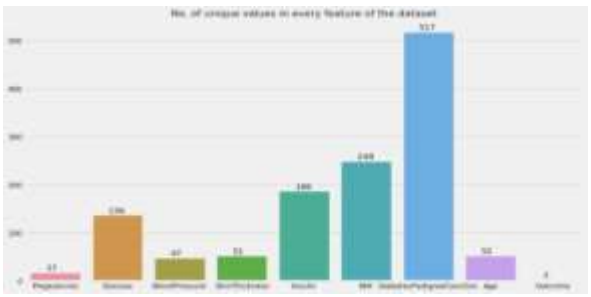

Fig 2: Density plot of all features

Density plot of all features : In Density plot of all features as shown in figure 2 , If we carefully see , then we would come to know that some of the features are having zero value which is unrealistic. The features having zero value are :- Glucose ,Blood Pressure ,Skin Thickness ,Insulin ,BMI .For these features, the value as zero can be considered as missing value and therefore we will be replacing them to Nan and will do some arrangements to fill these missing values. So, we have some NULL values. We will fill the missing values using some strategy. Most of the features are loosely gaussian distributed, which is good for us. Some points to note here :-

a. Features like Age and pregnancies are having good correlation which is quite obvious.
b. Features like Glucose and Insulin are having good correlation. ,c. Features Insulin and Skin Thickness are having highest correlation value. ,d. Our target Feature i.e., Outcome is having some correlation with glucose, Insulin and BMI.

1). Bar plot for Pregnancies

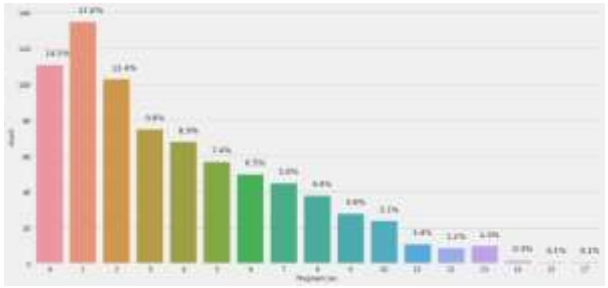

Fig 3: Bar plot -Pregnancies

From this plot (Figure 3) , we can conclude that, generally chances of having diabetes increases with increase in number of pregnancies.

2). Plot for Pregnancies and Outcome

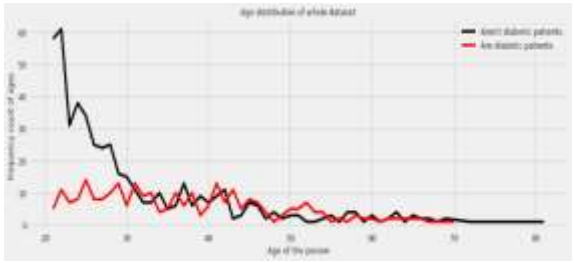

Fig 4: Plot for Pregnancies and Outcome

From this plot (Figure 4) , we get to know that chances of having diabetes increases with age (more nearly after 30 Years).

https://doi.org/10.46335/IJIES.2023.8.3.5 e-ISSN: 2456-3463
Vol. 8 , No. 3, 2023, PP. 24-34

3). Plot shows "Glucose level" and "Frequency count of persons" which indicates "Glucose distribution of whole dataset" and Plot shows "Glucose distribution of whole the dataset (Figure 5) .

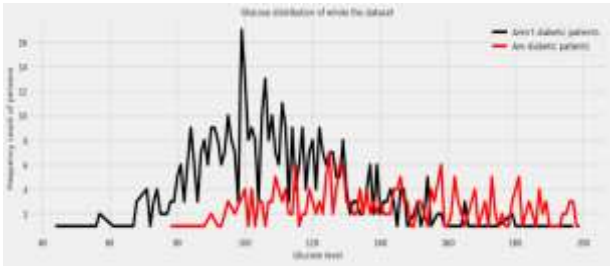

Fig 5: Plot for Glucose level and Outcome

4). Plot of Age , 'Blood Pressure' and 'Outcome'

Plot of "Blood Pressure level" ,"Frequency count of persons" for Blood Pressure distribution of whole dataset and Plot of "Blood Pressure level" ,"Frequency count of persons" for "Blood Pressure distribution of whole the dataset (Figure 6)

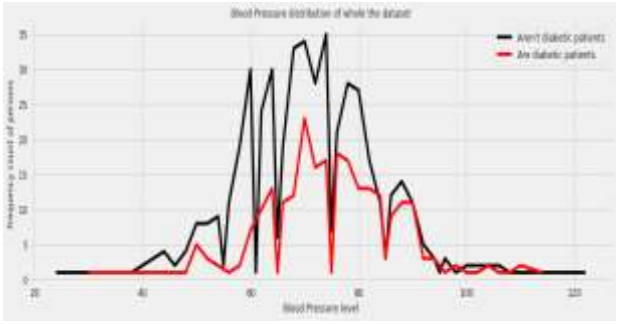

Fig 6: Plot of Age , 'Blood Pressure' and 'Outcome'

5). Plot of "Skin Thickness level", "Frequency count of persons" for "Skin Thickness distribution of whole dataset")

Plot of "Skin Thickness level" and "Frequency count of persons" for "Skin Thickness distribution of whole the dataset" and plot of Age' 'SkinThickness'and Outcome' (figure 7)

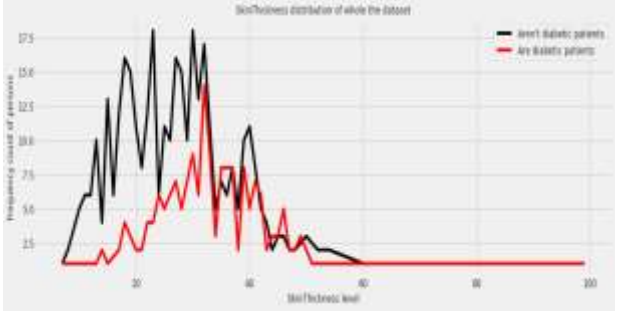

Fig 7: Plot of Age , Skin Thickness and 'Outcome'

6). Plot of Insulin Distribution : A. Insulin distribution of whole dataset ,B. Plot of "Aren't diabetic patients" and 'Are diabetic patients'. And Insulin distribution of whole the dataset (figure 8)

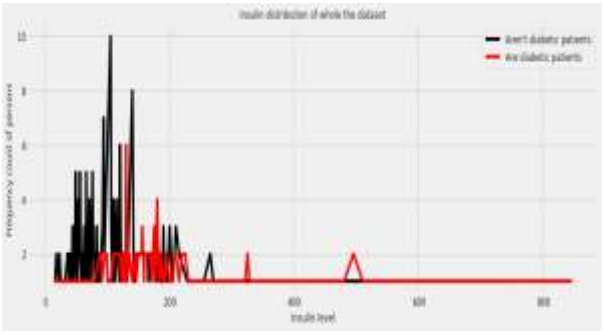

Fig 8: Plot of Age , Skin Thickness and 'Outcome'

7). Plot of 'Age', and 'BMI' : A.BMI distribution of whole dataset B. Plot of "Aren't diabetic patients and 'Are diabetic patients. Of "BMI distribution of whole the dataset (figure 9)

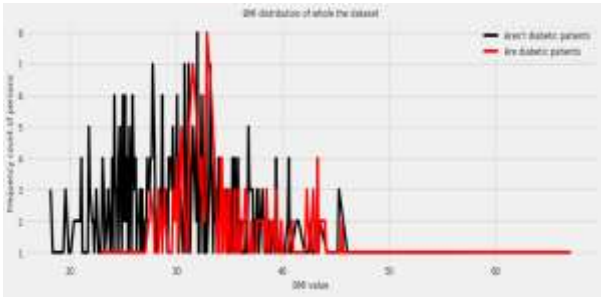

Fig 9: Plot of 'Age', and 'BMI'

Task - 3 - Data Preparation for model evaluation : In this Task , Checking the Outliers is done. Plot of 'Pregnancies' (Table 3a)

Pregnancies' > 13

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
88	15	136	70	32	110	37.1	0.15	43	1
159	17	163	72	41	114	40.9	0.82	47	1
298	14	100	78	25	184	36.6	0.41	46	1
455	14	175	62	30	NaN	33.6	0.21	38	1

a. Plot of 'Blood Pressure' (Table 3b)

'BloodPressure' < 40

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
18	1	103	30	38	83	43.3	0.18	33	0
125	1	88	30	42	99	55	0.5	26	1
597	1	89	24	19	25	27.8	0.56	21	0
599	1	109	38	18	120	23.1	0.41	26	0

https://doi.org/10.46335/IJIES.2023.8.3.5 **e-ISSN: 2456-3463**
Vol. 8 , No. 3, 2023, PP. 24-34

International Journal of Innovations in Engineering and Science, www.ijies.net

c. Plot of 'Skin Thickness' (Table 3c)

'SkinThickness' > 65

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
579	2	197	70	99	NaN	34.7	0.57	62	1

d. Plot of 'Insulin' (Table 3d)

'Insulin' > 500

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
8	2	197	70	45	543	30.5	0.16	53	1
13	1	189	60	23	846	30.1	0.4	59	1
228	4	197	70	39	744	36.7	2.33	31	0
247	0	165	90	33	680	52.3	0.43	23	0
286	5	155	84	44	545	38.7	0.62	34	0
409	1	172	68	49	579	42.4	0.7	28	1
584	8	124	76	24	600	28.7	0.69	52	1
655	2	155	52	27	540	38.7	0.24	25	1
753	0	181	88	44	510	43.3	0.22	26	1

e. Dealing with NULL values and splitting data into train and test data set. (Table 3e)

Dealing with NULL values	
Pregnancies	4
Glucose	5
BloodPressure	39
SkinThickness	228
Insulin	383
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0

Here , it is necessary to split dataset into train and test data set

The shape of training data : (576, 8) (576,)
The shape of testing data : (192, 8) (192,)
using smote before scaling
The shape of training data : (576, 8) (576,)
The shape of testing data : (754, 8) (754,)
using smote after scaling
The shape of training data : (754, 8) (754,)
The shape of testing data : (192, 8) (192,)

Task 4 - Model Evaluation

A. Performing cross validation for Logistic Regression Classifier model. (Table4a: Logistic regression)

Logistic Regression

	precision	recall	f1-score	support
0	0.8	0.69	0.74	123
1	0.56	0.7	0.62	69
accuracy			0.69	192
macro avg	0.68	0.69	0.68	192
weighted avg	0.71	0.69	0.7	192

F1 score : 0.681 for Logistic regression .

B. Performing cross validation for SVC Classifier model. (Table 4b : Linear SVC)

Linear SVC

	precision	recall	f1-score	support
0	0.8	0.69	0.74	123
1	0.56	0.7	0.62	69
accuracy			0.69	192
macro avg	0.68	0.69	0.68	192
weighted avg	0.71	0.69	0.7	192

F1 score : 0.681 for Linear SVC

C1. Performing cross validation for Decision Tree Classifier model. (Table 4c1) : Decision Tree

DecisionTree Classifier

	precision	recall	f1-score	support
0	0.82	0.64	0.72	123
1	0.54	0.75	0.63	69
accuracy			0.68	192
macro avg	0.68	0.7	0.68	192
weighted avg	0.72	0.68	0.69	192

F1 score : 0.676 for Decision Tree.

C2. Performing cross validation for hyperparameter tunning of Decision Tree Classifier model. F1 score : 0.663 .

Table 4c2: Decision Classifier

Hyper tunning of DecisionTreeClassifier model.

	precision	recall	f1-score	support
0	0.83	0.6	0.7	123
1	0.52	0.78	0.63	69
accuracy			0.67	192
macro avg	0.68	0.69	0.66	192
weighted avg	0.72	0.67	0.67	192

D1. Performing cross validation for MLP Classifier model. (Table 4d1: MLP Classifier)

MLP Classifier

	precision	recall	f1-score	support
0	0.82	0.67	0.74	123
1	0.55	0.74	0.63	69
accuracy			0.69	192
macro avg	0.69	0.7	0.68	192
weighted avg	0.72	0.69	0.7	192

F1 score : 0.684(Table 4d) : MLC

D2. Performing cross validation for hyperparameter tunning of MLP Classifier model. Table 4d2: MLP

Hyperparamter tunning of MLPClassifier model.

	precision	recall	f1-score	support
0	0.8	0.64	0.71	123
1	0.53	0.71	0.6	69
accuracy			0.67	192
macro avg	0.66	0.68	0.66	192
weighted avg	0.7	0.67	0.67	192

F1 score : 0.658

E1. Performing cross validation for Random Forest Classifier model.(Table 4e1: Random Forest Model)

https://doi.org/10.46335/IJIES.2023.8.3.5 **e-ISSN: 2456-3463**
Vol. 8 , No. 3, 2023, PP. 24-34

International Journal of Innovations in Engineering and Science, www.ijies.net

RandomForest Classifier

	precision	recall	f1-score	support
0	0.88	0.7	0.78	123
1	0.61	0.83	0.7	69
accuracy			0.74	192
macro avg	0.74	0.76	0.74	192
weighted avg	0.78	0.74	0.75	192

F1 score : 0.739 (Table 4e) MLC

E2. Performing cross validation for hyperparameter tunning of Random Forest Classifier model.

Hyperparamter tunning of Random Forest Classifier model.

	precision	recall	f1-score	support
0	0.88	0.72	0.79	123
1	0.63	0.83	0.71	69
accuracy			0.76	192
macro avg	0.75	0.77	0.75	192
weighted avg	0.79	0.76	0.77	192

F1 score : 0.754

F1. Performing cross validation for Gradient boosting Classifier model. Table 4f: GB : F1 score : 0.692

GradientBoosting Classifier

	precision	recall	f1-score	support
0	0.81	0.7	0.75	123
1	0.57	0.71	0.63	69
accuracy			0.7	192
macro avg	0.69	0.7	0.69	192
weighted avg	0.72	0.7	0.71	192

F2. Performing cross validation for hyperparameter tunning of GradientBoostingClassifier model.

Hyperparamter tunning of GradientBoostingClassifier model.

	precision	recall	f1-score	support
0	0.8	0.72	0.76	123
1	0.58	0.68	0.63	69
accuracy			0.71	192
macro avg	0.69	0.7	0.69	192
weighted avg	0.72	0.71	0.71	192

F1 score : 0.694

TASK 5: FINAL SCORES OF ALL MODELS

The F1 score of Logistic Regression : 0.681

The F1 score of Linear SVC : 0.681

The F1 score of Decision Tree Classifier : 0.663

The F1 score of MLP Classifier : 0.658

The F1 score of Random Forest Classifier : 0.754

The F1 score of Gradient Boosting Classifier : 0.694

Utilizing the dataset, a computer-assisted diabetes detection system is built using deep learning models and machine learning models . Six distinct machine learning models, including GB, Random Forest (RF), Decision Tree (RF) , Support Vector Machine (SVM), and Logistic Regression (LR) have been presented in this research. MLP classification methods used in Deep Learning, have been implemented. We have pre-processed individual data point in the dataset before using categorization methods. Table 5 shows Accuracy, Precision, Recall, F1-Score, for each and every Deep Learning (DL) and Machine Learning (ML) method, in order to facilitate a speedy selection of the optimal model while considering all possible ratings. Here, in our case we got the maximum F1 score from RF Classifier. The data was imbalanced; therefore, SMOTE has been used oversampling technique to balance the positive and negative instances. We have tried varieties of model here. More Careful tunning and feature engineering may bring even better results.

Table 5 :Comparison of DM-ML Models for accuracy and other parameters.

Classification	Accuracy	Precision	Recall	F1-Score
LR	0.69	0.68	0.69	0.68
SVC	0.69	0.68	0.69	0.68
DT	0.68	0.7	0.68	0.68
MLP	0.69	0.69	0.7	0.68
RF	**0.74**	**0.74**	**0.74**	**0.74**
GB	0.7	0.69	0.7	0.69

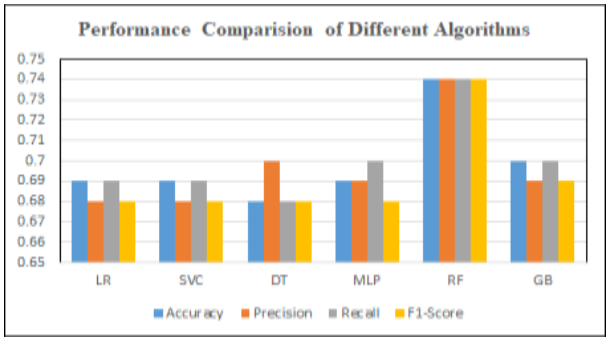

Fig 10: Performance analysis of different algorithms

VI. CONCLUSION

According to the research so far, numerous artificial intelligence (AI) strategies for diabetes prevention, diagnosis, and treatment are being developed, evaluated, and applied. Machine Learning (ML) , a subset of Artificial Intelligence (AI) , might revolutionise diabetes risk prediction and early detection. Successfully managing diabetes requires early detection. We prepared and evaluated and performed Prediction of Diabetes Using a Number of Different Machine Learning ML Methods and conducted output assessment to determine the optimum classifier with the greatest accuracy. We collected data set characteristics and evaluated ML classification methods in this article to attain high accuracy. RF algorithms outperform ML

classification methods. Random forest classification accuracy was 74%. Age and diabetes are unrelated, despite scientific evidence to the contrary. The risk assessment process for diabetes in its early stages has been revolutionized by machine learning and deep learning. We used machine learning and deep learning classification algorithms and diabetes risk variables to predict diabetes early in our research work. For evaluating effectiveness of machine learning and deep learning models for diabetes prediction , six classification algorithms—LR, GB, RF, DT, SVM, MLP were tested on the diabetes dataset. RF beat other machine learning ML and deep learning DL methods for early stage diabetes detection by over 74%. The research may help doctors recognise diabetes early and make better diabetes treatment choices, saving lives. Our study can properly predict diabetes but has limits. The study's tiny sample size made statistical significance difficult to establish. Our research is highly recommended because it is made up of research articles from various sources that can help other academics working on alternative diabetic prediction models.

REFERENCES

[1] *Tao Z, Shi A, Zhao J. "Epidemiological perspectives of diabetes". Cell Biochem Biophys 2015;73:181–5.*

[2] *Lonappan, A., Bindu, G., Thomas, V., Jacob, J., Rajasekaran, C., and Mathew, K. T. (2007). "Diagnosis of diabetes mellitus using microwaves". J. Electromagnet. Wave. 21, 1393–1401. doi: 10.1163/156939307783239429*

[3] *Kang, Hyun. (2013). "The prevention and handling of the missing data". Korean journal of anesthesiology.*

[4] *Iancu, I., Mota, M., and Iancu, E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883.*

[5] *Robertson, G., Lehmann, E. D., Sandham, W., and Hamilton, D. (2011). "Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proofof-concept pilot study". J. Electr. Comput. Eng.2011:681786. doi: 10.1155/2011/681786.*

[6] *"Learning about diabetes and its type.".Available:https://www.diabetesresearch.org/what-is-diabetes*

[7] *D. Gahlan, R. Rajput, and V. Singh, "Metabolic syndrome in north indian type 2 diabetes mellitus patients: A comparison of four different diagnostic criteria of metabolic syndrome." Diabetes & metabolic syndrome, vol. 13, no. 1, pp. 356–362, 2018.*

[8] *T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: a systematic review," IEEE Journal of Biomedical and Health Informatics, 2020.*

[9] *Dataset, P. I. D. UCI Machine Learning Repository, diambil dari http://archive. ics. uci. edu/ml/datasets. Pima+ Indians+ Diabetes. Accessed (October 2020)*

[10] *Soni. M and Varma. S (2020)," Diabetes Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology (IJERT)*

[11] *Sarwar. M, Kamal. N, Hamid. W and Shah. A (2018), International Conference on Automation and Computing (ICAC).*

[12] *Tejas N.Joshi, Prof.Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques", January 2018, Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part-II), pp.-09-13*

[13] *Parashar, A., Burse, K., & Rawat, K. (2014)."A Comparative approach for Pima Indians diabetes diagnosis using ldasupport vector machine and feed forward neural network". International Journal of Advanced Research in Computer Science and Software Engineering, 4(11), 378-383.*

[14] *Al Helal, M., Chowdhury, A. I., Islam, A., Ahmed, E., Mahmud, M. S., & Hossain, S. (2019, February). "An optimization approach to improve classification performance in cancer and diabetes prediction'. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.*

[15] *L. Li, "Diagnosis of diabetes using a weight-adjusted voting approach," in 2014 IEEE International Conference on Bioinformatics and Bioengineering, 2014, pp. 320–324.*

[16] *J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," Procedia Computer Science, vol. 47, pp. 45–51, 2015.*

[17] *N. Naiarun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," Procedia Computer Science, vol. 69, pp. 132– 142, 2015, the 7th International Conference on Advances in Information Technology.*

[18] *I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.*

[19] *T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," International Journal of Medical Informatics, vol. 97, 2016.*

[20] *T. Ahmed, "Developing a predicted model for diabetes type 2 treatment plans by using data mining," Journal of Theoretical and Applied Information Technology, vol. 90, pp. 181–187, 2016.*

[21] *Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). "Random forests: from early developments to recent advancements". Systems Science & Control Engineering: An Open Access Journal, 2(1), 602-609.*

[22] *E. Goel, E. Abhilasha, E. Goel, and E. Abhilasha, "Random Forest: A review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, 2017.*

[23] *Quinlan, J. R. (1986)." Induction on decision tree'. Mach. Learn.1, 81–106. doi: 10.1007/BF00116251*

[24] *A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 18, no. 6, pp. 275–285, 2004.*

[25] *J. M. Hilbe, "Logistic regression models". Chapman and hall/CRC, 2009.*

[26] *C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," The journal of educational research, vol. 96, no. 1, pp. 3–14, 2002.*

[27] *Friedman, J. (2001). "Greedy boosting approximation: a gradient boosting machine". Ann. Stat. 29, 1189–1232. doi: 10.1214/aos/1013203451*

[28] *A. Odeh, I. Keshta, and E. Abdelfattah, "Efficient detection of phishing websites using multilayer perceptron," 2020.*

[29] *D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.*

[30] *Sokolova M., Japkowicz N., Szpakowicz S., (2006), "Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation", American Association for Artificial Intelligence (www.aaai.org).*

[31] *Yağanoğlu, M., & Köse, C., (2018)," "Real-time detection of important sounds with a wearable vibration-based device for hearing-impaired people". Electronics, 7(4), 50.*