

# Effective Comparative Algorithms for Diabetes Prediction

Doaa Trabay<sup>1</sup>, Mai. A. Elnady<sup>2</sup>

Doaatrabay@oi.edu.eg, doctormaielnady@gmail.com

<sup>1</sup>Lecture in Obour Institute for Management & Information Systems & Computer Science

<sup>2</sup>lecture in Computers & Information System-Sadat academy for management science

## Abstract

One of the deadliest chronic conditions that elevate blood sugar is Diabetes. Diabetes increases the risk of kidney disease, stroke, visual problems, nerve damage, and other diseases. Currently, various tests are used in hospitals to obtain the necessary data for the diagnosis of Diabetes, and depending on this diagnosis, and appropriate treatment is provided. The healthcare sector benefits significantly from big data analytics as databases are vast. One can examine a large set of data using big data analytics. If Diabetes is not diagnosed and not treated, several consequences may result in death. However, developing Machine Learning (ML) techniques solves this critical problem. The motive of this study is to make a comparative analysis using different algorithms for Diabetes prediction. The goal of this study is to conduct a close examination of several Diabetes prediction systems using nine machine learning algorithms for diabetes prediction as detect it according to classification methodologies like SVM, NB, DT, LR, KNN, NN, random forest, ada-boost, gradient boost. The "National Institute of Diabetes and Digestive and Kidney Diseases" is where the dataset first came from. After analyzing the prediction result of different algorithms,

we observed that Neural Network (NN), Logistic Regression (LR), and Gradient-Boosting (GB) are more accurate than others.

**Keywords:** Diabetes diseases, Machine learning algorithms (ML), Classification models, prediction model

## 1. Introduction

The chronic disease Diabetes is brought on by inadequate insulin production or irregular insulin release. Hormonal imbalances lead to abnormal insulin secretion, and an unstable insulin level prevents glucose from being removed from the bloodstream. There are serious health consequences when blood glucose levels rise. Diabetes is the direct cause of high blood sugar, and some of its typical symptoms include weight loss, excessive urination, hunger, and thirst [1]. Diabetes frequently affects people of all ages, from young children to the elderly. It is possible to control Diabetes if it is caught early enough.

A crucial step in Diabetes prevention and management in healthcare is accurate Diabetes classification. Therefore, detecting Diabetes early is more beneficial for controlling it. Because the patient needs to see the doctor frequently, the procedure of identifying Diabetes at an early stage

may appear laborious. Through disease prediction, improvements in machine learning techniques have resolved this crucial and fundamental issue in healthcare. Therefore, more methods for predicting Diabetes have been suggested in the literature. PreDiabetes develops when blood glucose levels rise above the usual range, but a doctor may have trouble diagnosing Diabetes if the patient's symptoms are mild. Exercise and weight loss thereby lower the risk of developing pre-Diabetes .

Three forms of Diabetes must be identified to assess the severity of the condition: [2] Diabetes of type I typically develops throughout infancy. Type I Diabetes is a condition in which the body produces no or very little insulin. Patients with type 1 Diabetes have access to insulin injections to manage their condition. Unusual weight loss, unusual appetite and thirst, abnormal urine, and situations involving the kidneys and eyes are symptoms of this kind of DM. Type 1 Diabetes symptoms will raise the already elevated risk of heart disease and stroke.

When the body doesn't respond to insulin, Type II Diabetes (T2D) develops. This condition typically affects adults. Type 2 Diabetes is characterized by weight gain and a sharp increase in blood pressure. T2D raises the risk of developing heart conditions, including stroke. Type 1 Diabetes Mellitus (DM), also known as Insulin-Dependent Diabetes Mellitus, is the most common type (IDDM). Because of this type of DM, a patient needs to receive insulin injections because their body cannot produce enough insulin on its own. Non-Insulin-Dependent Diabetes Mellitus is another name for type 2 (NIDDM). This form of Diabetes manifests as improper insulin uses by bodily cells [2].

Type-III: Gestational Diabetes develops when a pregnant woman's blood sugar level rises, and the Diabetes is not diagnosed sooner. Long-term consequences are linked to DM. A diabetic person also faces significant risks of developing several medical conditions [2].

Rapid advances in ML have enhanced the efficiency of decision-making processes in a wide range of applications, including medical diagnosis. ML-based diagnostics has gained attention in recent years due to its faster inference results and ability to perform complex tasks requiring specialized expertise and experience. ML is used to improve the accuracy of diagnosing different diseases by detecting patterns. To predict and assess the correctness of the input dataset, ML employs a range of classifiers, including supervised, unsupervised, and ensemble learning [3]. To classify disease severity, diseases are classified by multiple algorithms such as KNN, DT, Genetic Algorithm (GA), NB, etc. [4-6].

This paper focuses mainly on nine ML techniques: KNN, NB, ANN, DT, SVM, Random-Forest, LR, GB, and Ada-boost . The accuracy of the three best-achieved techniques is that the LR model is the best algorithm compared to other algorithms. NN technology came in second, and then Gradient Boosting had the third-best resolution technology. Then, the three models were predicted to predict Diabetes disease.

Our paper is organized in the following manner: The related work of several classifications approaches for Diabetes prediction is summarized in Section 2. The dataset is presented in Section 3. The methodology is suggested in Section 4. Classification models are presented in Section 5. The findings and analyses are presented in Section 6. Section 7 Prediction of best three models. Finally, section 8, Conclusion and future work.

## 2. Related Work

The paper discussed the use of multiple supervised ML techniques for Diabetes prediction. Radial Function Kernel (RBF) SVM, Artificial Neural Network (ANN), Multifactorial Dimensionality Reduction (MDR), linear SVM, and KNN were used by Kaur & Kumari [7]. Logistic regression (LR) was used to identify risk factors for Diabetes based on the p-value and odds ratio (OR). Maniruzzaman et al. [8] used the NB, DT, Ada-Boost, and

Random Forest (RF) classifiers to predict Diabetes patients. The K2, K5, and K10 fragmentation techniques have also been adopted. Accuracy (ACC) and area under the curve (AUC) of classifiers were tested with these replicated techniques in 20 paths.

In the study [9], Kopitar et al. compared many commonly used regression models for type 2 Diabetes mellitus prediction, including Glnnet, RF, XGBoost, and LightGBM. Yahyaoui et al. [10] proposed ML techniques and contrasting methods of Deep Learning (DL) with those of classical ML. They used the SVM and RF classifiers frequently employed in conventional ML techniques. They utilized a full-Scale Neural Network (CNN) for DL to detect people with Diabetes. With the aid of clinical data, Nazin, Ahmed et al. [11] proposed efficient ML-based classifier models for diagnosing Diabetes. In this study, the algorithms DT, NB, KNN, RF, Gradient Boosting (GB), LR, and SVM were trained on various datasets. The findings of this paper imply that Diabetes can be accurately and successfully predicted using a clinical data preparation pipeline and ML-based categorization. The suggested model by Aeshah Saad et al. [12] combines the SVM and RF ML methods to predict Diabetes. They used actual data sets gathered from primary healthcare for a security force. The outcome demonstrated that the RF method outperforms the SVM in terms of accuracy.

A stacking-based ensemble technique for predicting type 2 Diabetes mellitus was put forth by Singh and Namrata [13]. They trained the SVM, DT, RBF, and poly SVM stacking ensemble four base learners using the bootstrap approach using cross-validation. However, the state-of-the-art comparison is absent, and variable selection is not specified. A deep neural network-based medical decision system for predicting Diabetes was created by Tawfik Begrich et al. [14]. These algorithms represent some of the most recent developments in image analysis, language processing, and computer vision. The algorithm has

demonstrated accuracy. Additionally, it can be used with medical knowledge to increase decision-making efficiency, adaptability, and transparency. By merging the outcomes of many machine learning algorithms, Jyoti Rani [15] created a system that can predict a patient's early onset of Diabetes with a higher degree of accuracy. KNN, LR, RF, SVM, and DT are the techniques employed.

### 3. Data Source

Diabetes develops when the pancreas is unable to produce enough insulin or when the body cannot use the insulin produced. The hormone insulin controls the amount of glucose in the blood. According to the World Health Organization (WHO), most diabetic patients are women, especially during pregnancy. We collected data from the National Institute of Diabetes and Digestive and Kidney Diseases for 768 pregnant patients aged 20 to 50 years with family history in mind [16], and all patients are females. The patient's medical parameters include the number of pregnancies, glucose, blood pressure, insulin, age, etc. The paper aims to estimate the patient's likelihood of developing Diabetes.

### 4. Methodology

The research focused on nine ML techniques: KNN, Tree, SVM, Random-Forest, NN, NB, LR, GB, and Ada-boost. These phases apply to the mentioned techniques as follows:

Phase 1: The data set is entered and ready for preprocessing.

Phase 2: From the preprocessed Diabetes data collection, we select the nine features: (Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree Function, Age, and Target).

Phase 3: With the help of the confusion matrix, all models are evaluated.

Phase 4: Predict the three best models.

So, the attributes and its description will be shown in Table 1 as follows.

Table 1. The attributes and descriptions of our paper

Attributes	Abbreviation	Description	Value	Data Type
Pregnancies	Pr	number of pregnancies	0:15	Numeric Value
Glucose	Glu	a -f hour oral glucose tolerance test's plasma glucose concentration	50:200	
Blood Pressure	BP	BP in diastole (mm Hg)	50:150	
Skin Thickness	ST	Resting blood pressure	80:180	
Insulin	Ins	(-f hour serum insulin, mu U/ml)	0:900	
Body Mass Index	BMI	BMI (weight in kilograms divided by height in meters) ^f	0:70	
Diabetes Pedigree Function	DBI	Skin fold thickness of the triceps (mm)	0:3	
Age	Ag	All patients are females	20:50	
Outcome	Ou	Exercise-induced angina	0 = false; 1 = true	

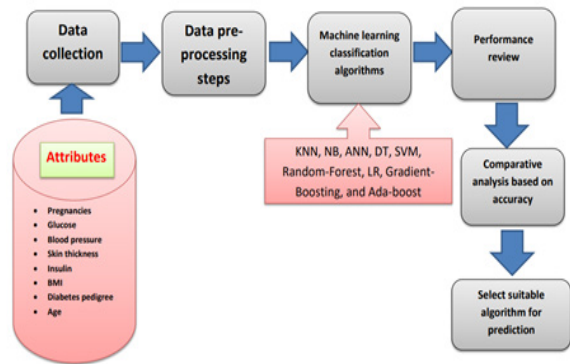


Fig1: Proposed methodology

## 5. Classification models

We divide classical ML into two categories: directed learning (supervised learning) and Undirected learning (unsupervised learning) [3].

### • Supervised Learning

The machine has a 'supervisor' or 'teacher' who provides the machine with all correct and accurate answers.

### • Unsupervised Learning

Undirected learning would leave the machine alone with a large pile of data, and its job would be to classify that data. The data is not classified. A machine will learn much faster with a teacher, so directed learning is frequently used in real-world tasks. There are two main types of directed learning methods: classification and regression.

In our research paper, we present some algorithms used to detect Diabetes according to the classification method, which is shown as follows:

### 5.1 Support Vector Machine (SVM) Technique:

The SVM is one of the often-used supervised ML model types for classification. Given a two-class training sample, an SVM's objective is to determine the optimal, highest-margin separation hyperplane between the two classes. For greater generalization, the hyperplane shouldn't be positioned closer to data points from the opposite class. It is best to select the hyperplane that is the furthest away from the data points in each category. The points positioned closest to the classifier's boundary are those that make up the SVM. The SVM chooses the ideal separating hyperplane. The distance between the hyperplane defined by  $wT x + b = 1$  and the hyperplane defined by  $wT x + b = -1$  will be maximized mathematically. This separation is equivalent to  $2w$ . So, our goal is to solve no more than two  $w$ . In other words, we want  $\min w | 2$ . All  $x(i)$  should be accurately classified by the SVM, which implies that  $y_i (wT x_i + b) \geq 1, \forall i \in \{1, \dots, N\}$  [17].

### 5.2 Naive Bayes (NB) Classifier

NB is a categorization strategy based on the idea that all features are distinct and independent. It states that the status of a particular feature within a class has no bearing on the status of any additional features. It is regarded as a strong method used for classification since it is based on conditional probability. For data with imbalance issues and missing values, it performs well. A machine

learning classifier called Naive Bayes uses the Bayes Theorem. Using posterior probability and the Bayes theorem,  $P(C)$ ,  $P(X)$ , and  $P(X|C)$  can be used to derive  $P(C|X)$  [18].  $P(C|X) = (P(X|C) P(C))/P$  as a result ( $X$ ), Where  $P(C|X)$  is the posterior probability for the target class. The possibility of predicting class is  $P(X|C)$ .  $P(C)$  is the likelihood that class  $C$  is true:  $p(X) =$  Prior Probability of Predictor.

### 5.3 Decision Tree (DT) Classifier

Using a decision rule drawn from past data, the DT's primary objective in this study is to forecast the target class. It uses nodes and internodes for categorization and prediction. Root nodes classify the instances according to many traits. The root nodes may have two or more branches, but the leaf nodes signify categorization. The DT considers the highest information gain among all the attributes at each stage when choosing a node [19].

### 5.4 Logistic Regression (LR) Technique

It is one of the most widely used classification algorithms because it uses straightforward procedures to divide data into different classes [20]. Pass/Fail will be the outcome. We refer to these values as discrete values.

### 5.5 k-Nearest Neighbors (KNN) Technique

One of the supervised learning group's ML algorithms, the KNN method, is "one of the simplest algorithms due to its ease of use and time consumption." Because the KNN technique separates pre-sorted data, it uses the complete data set as a training set rather than dividing it into a training and test set [21]. When the desired result is for a new data element, the KNN algorithm goes through the entire data set to find the  $k$  nearest instances of the new one.

### 5.6 Random Forest Technique (RF)

It is an improvement over DT, consisting of many individual DTs working as a group to get more

accurate and stable predictions. From the results of this group's prediction, the highest vote-vote result is obtained, which is better than using the best model alone [22].

### 5.7 Ada-Boost Technique

It is used with many types of learning algorithms to improve performance. On each iteration, Ada-Boost will identify a misclassified data point, thereby increasing its weight (in other words decreasing the weight of the integer point) so that the next classifier will be more attentive to correcting it [23].

### 5.8 Gradient Boost (GB) Technique

It's one of the standard boosting algorithms, this algorithm works by adding prediction models sequentially, and each model corrects the previous model [24]. Still, instead of adjusting the weights of the examples at each step as is done in Ada-Boost, this method tries to fit the new prediction model with the remaining errors from the model that preceded it and focuses on the difference between prediction and fundamental truth.

### 5.9 Neural Network (NN) Technique

Inputs  $x_i$ , hidden layers, and output  $u_i$  are elements that make up a neuron. The result is produced using an activation function, such as a sigmoid and a constant bias  $b$ . (See equation) [25].

$$f(b + \sum_{i=1}^n x_i u_i)$$

## 6. Evaluation Results for Used classification Models and Discussions

The research focused on nine ML techniques: KNN, Tree, SVM, Random Forest, NN, NB, LR, Gradient Boosting, and Adaboost. we explains the results of Confusion Metrix (CM), Classification Accuracy (CA), F1, Recall, Precision, Area Under the Curve (AUC), Log-loss, and Specificity in the following Table 2.



Table 2. Evaluation results for the techniques used

Used Techniques	Confusion Matrix (CM)	Comprative Performance																								
LR technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>442</td> <td>58</td> <td>500</td> </tr> <tr> <td>1</td> <td>113</td> <td>155</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>555</td> <td>213</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	442	58	500	1	113	155	268		Σ	555	213	768	
		Predicted																								
		0	1	Σ																						
Actual	0	442	58	500																						
	1	113	155	268																						
	Σ	555	213	768																						
NN Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>440</td> <td>60</td> <td>500</td> </tr> <tr> <td>1</td> <td>113</td> <td>155</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>553</td> <td>215</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	440	60	500	1	113	155	268		Σ	553	215	768	
		Predicted																								
		0	1	Σ																						
Actual	0	440	60	500																						
	1	113	155	268																						
	Σ	553	215	768																						
GB Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>420</td> <td>80</td> <td>500</td> </tr> <tr> <td>1</td> <td>107</td> <td>161</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>527</td> <td>241</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	420	80	500	1	107	161	268		Σ	527	241	768	
		Predicted																								
		0	1	Σ																						
Actual	0	420	80	500																						
	1	107	161	268																						
	Σ	527	241	768																						
NB Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>383</td> <td>117</td> <td>500</td> </tr> <tr> <td>1</td> <td>85</td> <td>183</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>468</td> <td>300</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	383	117	500	1	85	183	268		Σ	468	300	768	
		Predicted																								
		0	1	Σ																						
Actual	0	383	117	500																						
	1	85	183	268																						
	Σ	468	300	768																						
RF Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>418</td> <td>82</td> <td>500</td> </tr> <tr> <td>1</td> <td>115</td> <td>153</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>533</td> <td>235</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	418	82	500	1	115	153	268		Σ	533	235	768	
		Predicted																								
		0	1	Σ																						
Actual	0	418	82	500																						
	1	115	153	268																						
	Σ	533	235	768																						
Ada-Boost Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>379</td> <td>121</td> <td>500</td> </tr> <tr> <td>1</td> <td>113</td> <td>155</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>492</td> <td>276</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	379	121	500	1	113	155	268		Σ	492	276	768	
		Predicted																								
		0	1	Σ																						
Actual	0	379	121	500																						
	1	113	155	268																						
	Σ	492	276	768																						
KNN Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>409</td> <td>91</td> <td>500</td> </tr> <tr> <td>1</td> <td>126</td> <td>142</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>535</td> <td>233</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	409	91	500	1	126	142	268		Σ	535	233	768	
		Predicted																								
		0	1	Σ																						
Actual	0	409	91	500																						
	1	126	142	268																						
	Σ	535	233	768																						
DT Technique	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>383</td> <td>117</td> <td>500</td> </tr> <tr> <td>1</td> <td>85</td> <td>183</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>468</td> <td>300</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	383	117	500	1	85	183	268		Σ	468	300	768	
		Predicted																								
		0	1	Σ																						
Actual	0	383	117	500																						
	1	85	183	268																						
	Σ	468	300	768																						
SVM	<table border="1"> <tr> <td colspan="2"></td> <td colspan="2">Predicted</td> <td></td> </tr> <tr> <td colspan="2"></td> <td>0</td> <td>1</td> <td>Σ</td> </tr> <tr> <td rowspan="2">Actual</td> <td>0</td> <td>343</td> <td>157</td> <td>500</td> </tr> <tr> <td>1</td> <td>105</td> <td>163</td> <td>268</td> </tr> <tr> <td></td> <td>Σ</td> <td>448</td> <td>320</td> <td>768</td> </tr> </table>			Predicted					0	1	Σ	Actual	0	343	157	500	1	105	163	268		Σ	448	320	768	
		Predicted																								
		0	1	Σ																						
Actual	0	343	157	500																						
	1	105	163	268																						
	Σ	448	320	768																						

From these results, it is clear that even if the majority of studies used some of these algorithms, such as KNN, Adaboost, or DT, to identify patients with Diabetes disease, we found that LR, which reached 95% accuracy, was the best algorithm compared to the other algorithms. The NN technique came in second with 94%; then, GB achieved the third-best technique with 93% accuracy. By distributing the data, it was noted that Diabetes occurred to them 34.90% of the time in the data set, while 64.10% had no Diabetes disease, as shown in Fig.(2).

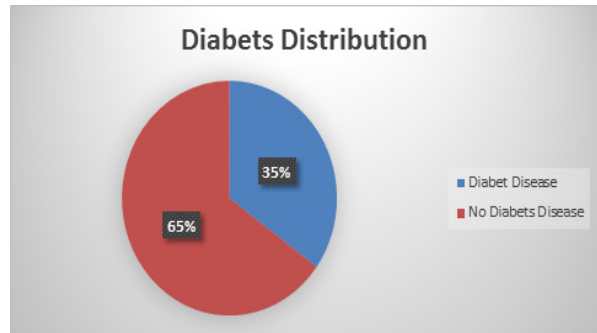


Fig 2 : Result of Diabets distribution

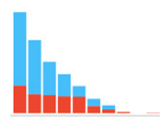
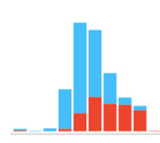

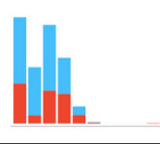
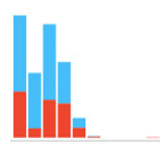
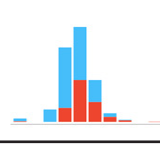
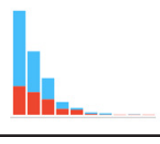
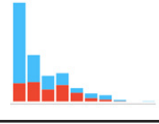
We will also show which features are essential and non-significant for Diabetes disease, where important factors show different variances, meaning they are crucial. Table (3) shows an information gain, information gain ratio, and Gini decrease for each attribute vital for diabetic disease.

Table 3: The rank of Diabetes diseases attributes

NO	Attributes	Info. gain	Gain ratio	Gini
1	Glu	0.170	0.085	0.102
2	Ag	0.081	0.041	0.048
3	BMI	0.079	0.039	0.044
4	Ins	0.055	0.030	0.031
5	Pr	0.043	0.021	0.028
6	ST	0.036	0.018	0.022
7	DPF	0.022	0.011	0.015
8	BP	0.0148	0.008	0.009

From Table (3), We discovered that the features of Glu, ag, then BMI are the three most important attributes affecting our accuracy results.

The attributes with distribution statistics, mean, median, dispersion, min., and max. Values are shown in Table (4).

Name	Distribution	Mean	Median	Dispersion	Min.	Max.
Pre		3.85	3	0.88	0.88	17
Glu		120.89	117	0.26	0.26	199
BP		69.11	72	0.28	0.28	122
ST		20.54	23	0.78	0.78	99
Ins		79.80	30.50	1.44	1.44	846
BMI		31.993	32.0	0.246	0.246	67.1
DBF		0.47188	0.37250	0.70169	0.078	2.420
Ag		33.24	29	0.35	0.35	81

show a sample of dataset in the following Table (5). The prediction is for 239 instances of Diabetes , nine attributes, and all of which are numeric values.

Table 5. Sample predictable Diabetes data

	LR	NN	GB	Pr	Glu	BP	ST	Ins	BMI	DPF	Ag
0	0	0	0	6	148	72	35	0	33.6	0.627	50
0	0	0	1	1	85	66	29	0	26.6	0.351	31
1	0	1	0	8	183	64	0	0	23.3	0.672	32
1	1	1	0	1	89	66	23	94	28.1	0.167	21
1	1	1	0	0	137	40	35	168	43.1	2.288	33
0	0	0	1	5	116	74	0	0	25.6	0.201	30
0	0	0	0	3	78	50	32	88	31.0	0.248	26
0	0	0	0	10	115	0	0	0	35.3	0.134	29
0	0	0	1	2	197	70	45	543	30.5	0.158	53
0	1	0	0	8	125	96	0	0	0.0	0.232	54
0	0	0	0	4	110	92	0	0	37.6	0.191	30
0	0	0	0	10	168	74	0	0	38.0	0.537	34
1	0	1	0	10	139	80	0	0	27.1	1.441	57
0	0	0	0	1	189	60	23	846	30.1	0.398	59
0	0	0	0	5	166	72	19	175	25.8	0.587	51
0	0	0	0	7	100	0	0	0	30.0	0.484	32
1	1	1	0	0	118	84	0	230	45.8	0.551	31
0	0	0	0	7	107	74	0	0	29.6	0.254	31
1	1	1	1	1	103	30	38	83	43.3	0.183	33
115	103	107	107	103	103	30	38	83	43.3	0.183	33
70	30	74	84	0	72	0	19	175	25.8	0.587	51
30	38	0	47	0	118	84	0	230	45.8	0.551	31
96	83	0	230	0	107	74	0	0	29.6	0.254	31
34.6	43.3	29.6	45.8	0.254	0.183	0.254	0.183	0.254	0.183	0.254	0.183
0.529	0.183	0.254	0.183	0.254	0.183	0.254	0.183	0.254	0.183	0.254	0.183
32	33	31	31	31	31	31	31	31	31	31	31

### 7 . Prediction for the best three Models:

We will predict for the higher three models and

After analyzing the prediction result of three models, we observed that LR and NN are 97% more accurate than GB.

## 8. Conclusion

One of the most common diseases today is Diabetes, and early detection is crucial to saving lives. We provided nine strategies in this paper that underwent comparative analysis and produced good results. We concluded that ML approaches are more effective for this analysis. Numerous researchers have proposed employing ML when the data set is not very large, and this study supports that approach. We concluded that the ML technique better supports this analysis. Numerous researchers have proposed employing ML when the data set is not very large, and this study supports that approach. The comparison approaches used are CM, CA, AUC, F1 score, precision, and specificity. We obtained a data set from the National Institute of Diabetes and Digestive and Kidney Diseases that includes 768 patients' medical histories, including details about their ages, pregnancies, blood pressure, insulin, and other factors. According to the evaluation's findings, LR is 95% more accurate than the other algorithms. According to the distribution data, we discovered that 34.90% of them had Diabetes diseases. Based on different variances, we chose the significant factors crucial to the doctor staff, conducted the distribution statistics, and extracted the mean, median, and both the min and max values. The three most reliable ways, which assist clinicians in diagnosing diseases before they manifest, were anticipated based on our utilized features. GB was 97% less accurate than the LR and NN.

The amount of the data set may grow as this study progresses. The DL can then be applied alongside

numerous additional advancements to produce more encouraging outcomes. The data can be normalized in various ways, and the results can be compared. There are more ways to incorporate ML and DL models that have been trained into Diabetes for quick disease identification.

## References

- [2] American Diabetes Association. (2021). 2. Classification and diagnosis of diabetes: standards of medical care in diabetes-2021. *Diabetes care*, 44(Supplement 1), S15-S33.
- [3] Ieracitano, C., Paviglianiti, A., Campolo, M., Hussain, A., Pasero, E., & Morabito, F. C. (2020). A novel automatic classification system based on hybrid unsupervised and supervised machine learning for electrospun nanofibers. *IEEE/CAA Journal of Automatica Sinica*, 8(1), 64-76.
- [4] Firdaus, H., Hassan, S. I., & Kaur, H. (2018). A comparative survey of machine learning and meta-heuristic optimization algorithms for sustainable and smart healthcare. *African Journal of Comput. ICT Ref. Format*, 11(4), 1-17.
- [5] Alharan, A. F., Algelal, Z. M., Ali, N. S., & Al-Garaawi, N. (2021, September). Improving Classification Performance for Diabetes with Linear Discriminant Analysis and Genetic Algorithm. In *2021 Palestinian International Conference on Information and Communication Technology (PICICT)* (pp. 38-44). IEEE.
- [6] Nabeel, M., Majeed, S., Awan, M. J., Muslih-ud-Din, H., Wasique, M., & Nasir, R. (2021). Review on Effective Disease Prediction through Data Mining Techniques. *International Journal on Electrical Engineering & Informatics*, 13(3).
- [7] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*.



- [8] Maniruzzaman, M., Rahman, M., Ahammed, B., & Abedin, M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, 8(1), 1-14.
- [9] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1), 1-12.
- [10] Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)* (pp. 1-4). IEEE.
- [11] Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*, 2, 229-241.
- [12] Bharat, A., Pooja, N., & Reddy, R. A. (2018, October). Using machine learning algorithms for breast cancer risk prediction and diagnosis. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)* (pp. 1-4). IEEE.
- [13] Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, 40(1), 1-22.
- [14] Beghriche, T., Djerioui, M., Brik, Y., Attallah, B., & Belhaouari, S. B. (2021). An efficient prediction system for diabetes disease based on deep neural network. *Complexity*, 2021.
- [15] Swathi, P., Jyothi, S., & Revathi, A. (2021). Machine Learning Techniques for Identifying Diabetes and Its Complications Based on Long Non-coding RNAs. In *Proceedings of the 2nd International Conference on Computational and Bio Engineering* (pp. 93-105). Springer, Singapore.
- [16] Abouzid, M. R., Ali, K., Elkhawas, I., & Elshafei, S. M. (2022). An overview of diabetes mellitus in Egypt and the significance of integrating preventive cardiology in diabetes management. *Cureus*, 14(7).
- [17] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- [18] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).
- [19] Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, 27(3), 221-234.
- [20] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1), 3-14.
- [21] Bailey, T., & AK, J. (1978). A NOTE ON DISTANCE-WEIGHTED K-NEAREST NEIGHBOR RULES.
- [22] Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
- [23] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.
- [24] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- [25] Hand, D. J., & Yu, K. (2001). Idiot's Bayes-not so stupid after all?. *International statistical review*, 69(3), 385-398.