IJBPAS, February, 2026, 15(2): 506-527

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

APPLICATION OF MACHINE LEARNING MODEL FOR CLASSIFICATION OF HEALTHCARE ISSUES IN HUMANS

PRERANA M*[1], STAVELIN AK[2], RAMACHANDRAN S[3], MADHU B[4], PATIL V[5], BALASUBRAMANIAN S[6]

1: Research Scholar, Division of Medical Statistics, School of Life Sciences, JSS AHER, Mysuru, Karnataka – 570015, India

2: Assistant Professor, Division of Medical Statistics, School of Life Sciences, JSS AHER, Mysuru, Karnataka – 570015, India

3: Head, MCA, Ramakrishna Mission Vidyalaya, Coimbatore, Tamil Nadu - 641020, India, India

4: Deputy Dean (Research), Department of Community Medicine, JSS Medical College, JSS AHER, Mysuru, Karnataka – 570015, India

5: Dean (Clinical), Department of Radiology/Fetal Imaging, JSS Hospital, Mysuru, Karnataka – 570004, India

6: Former Research Director, School of Life Sciences, JSS AHER, Mysuru, Karnataka – 570004, India

***Corresponding Author: Dr. Stavelin Abhinandithe K: E Mail: stavelin.ak@jssuni.edu.in**

Received 13th Sept. 2024; Revised 25th Nov. 2024; Accepted 10th Jan. 2025; Available online 1st Feb. 2026

https://doi.org/10.31032/IJBPAS/2026/15.2.9513

ABSTRACT

Human disease prediction is the process of estimating the likelihood that a patient would develop a disease after analysing the constellations of their symptoms. Keeping track of a patient's health information and status during the initial assessment can help doctors treat a patient's condition successfully. Patients would receive a more streamlined and rapid course of therapy as a result of this analysis in the medical field. This study aims for the detection severity of multiple diseases in patients using machine learning (ML) models. Google Colab – python 3.0 with libraries from Sklearn, Numpy, Pandas, Matplotlib was the platform that was used to classify the data. There are 222 records of patients with varying health problems. Data from 2018 – 2023, among which 104 were female and 118 were male around the age group of 7-81 years were collected from a private hospital in Mysuru, Karnataka. The dataset included information on the patient's age, gender, blood pressure, respiratory rate, pulse,

haemoglobin, total leucocyte count, red blood cell count, platelet count, etc. In this study, the ability to categorise the severity of multiple diseases using six ML models - Logistic Regression (LR), K-Nearest Neighbour (K-NN), Support Vector Machine-Radial basis function Kernel (SVM-rbf), Gaussian, Decision Tree (DT) and Random Forest (RF) - was studied. DT model attained the best accuracy for detecting severity, with 99% accuracy, while RF model obtained 98%, Gaussian model obtained 80%, SVM model obtained 64% and LR model and the KNN obtained 62% accuracy.

Keywords: Human Health, Machine Learning, Classification, Decision Tree, Accuracy

INTRODUCTION

Diabetes develops when metabolic issues cause blood sugar levels to rise. This kind may harm a number of body organs and systems, including the heart, blood vessels, and eyes. It is significant to remember that hyperglycemia, or high blood sugar levels, is the direct cause of these negative effects. This is due to issues with the body. Either has difficulty controlling blood sugar levels or uses the insulin it generates improperly **[1]**. Insulin is a hormone that aids in glucose transport and cell accessibility. Keep in mind that diabetes classified into type 1 and type 2 as the two primary groups. To comprehend type 1 diabetes completely, it's crucial to understand that the condition is an autoimmune one, which means that the body's Insulin-producing cells are relentlessly attacked and destroyed by the immune system.

Type 2 diabetes is characterised by issues with the body's ability to utilise its own insulin properly as a result of lifestyle-related variables **[2]**. Regardless of socioeconomic class, a noticeable rise in the prevalence of type 2 diabetes has been seen during the past ten years in every country in the world. Whether they are developed or developing nations makes no difference **[3]**. Diabetes can result in myocardial infarction, stroke, and lower limb amputations in addition to blindness and renal failure. Diabetes patients with poor glycemic control also have a much higher chance of developing TB and cardiovascular disease. In 2022, 6.7 million people are expected to die from diabetes, according to the WHO/PAHO. Eighty one percent (81%) of diabetics come from middle-income countries.

Adults are particularly vulnerable to developing diabetes because 40% of them are undiagnosed and 90% of them live in middle-income nations. According to data, the amount spent globally on diabetes-related medical care would exceed USD 966 billion in 2021, a 316% rise from the previous ten years. According to the International Diabetes Federation (IDF), there are more than 541 million persons who experience glucose intolerance globally. Approximately 10% of Americans have a

high chance of having type 2 diabetes at some time in their lives, according to this data. According to estimates, 68% of adults with diabetes reside in nations with the highest incidence of the disease, including the United States of America **[2, 4]**. There were previously 27.9 million diabetics residing in these nations. However, one in ten persons worldwide were estimated to have diabetes in 2021, with 537 million adults globally being diagnosed with the disease. According to a study by the IDF, there would be 643 million diabetics worldwide by 2030 and 784 million by 2045.

Medical experts struggle to accurately diagnose symptoms and spot diseases at an early stage because of the enormous number of data. There are several applications being created right now that use data mining and machine learning algorithms to effectively anticipate health issues. The current machine learning algorithms for healthcare analysis are primarily focused on predicting one disease in a specific application. There isn't a single application that employs machine learning to incorporate different illness prediction models, such as one for liver analysis, one for cancer analysis, one for lung diseases, etc. Machine learning is the foundation of the suggested system. The field of machine learning makes predictions based on past data. The goal of putting this model into practise is to provide a user-friendly interface for properly predicting numerous diseases using a single application.

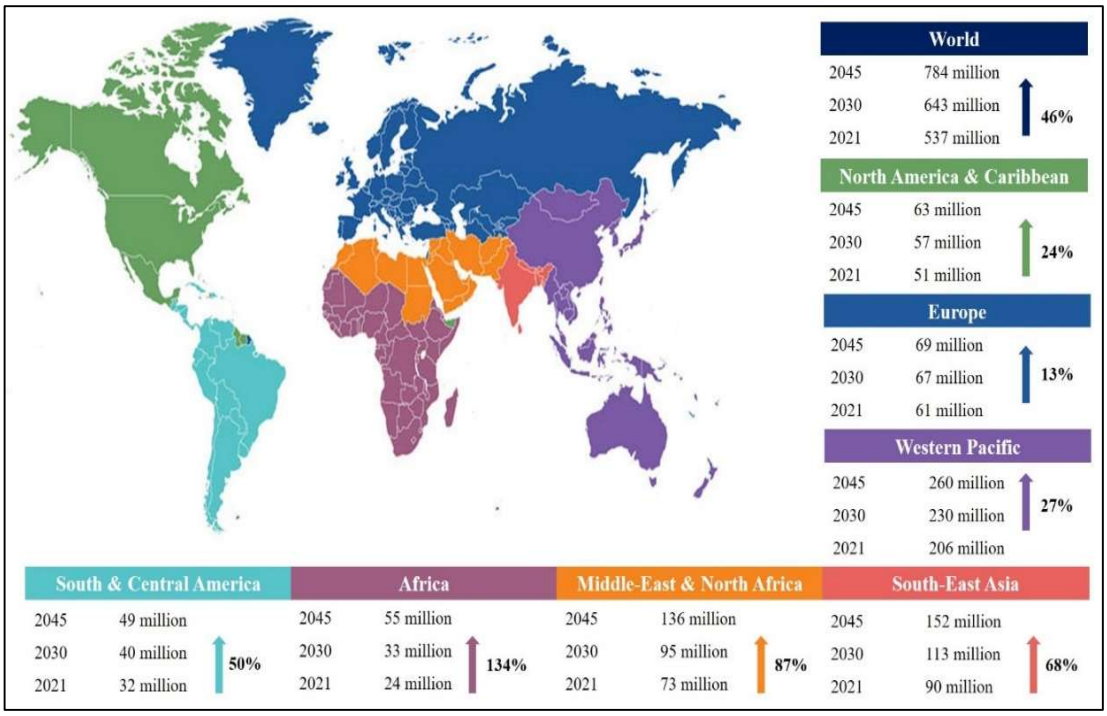


Figure 1: Region wise diabetic patients in the world

The majority of diabetic patients are currently found in the Western Pacific area **[5]**, as shown in **Figure 1**.

In particular, ML supervised learning models from artificial intelligence (AI) are useful tools for diagnosing and treating the chronic condition of diabetes. ML models have proven to be highly accurate diabetes development predictions. The models are based on details regarding a person's genetic make-up **[7]**, medical history **[6]**, and other risk factors. In order to find early indications of diabetes and other related disorders, ML models are being used to analyse medical pictures like CT scans and retinal scans **[8]**. With the use of computer science and statistics, the field of machine learning (ML) has developed breakthroughs that can identify and categorise gaps in patient care **[10, 11]**. These ML models are designed to help save medical expenses and improve the standard of patient care **[12, 13]**. This research is helpful in predicting pre-diabetes and identifying the risk factors linked to the development of diabetes from clinical data. A thorough evaluation of the patient's socio-demographics and health can help prevent diabetes, as can the implementation of an individualised treatment plan that takes into account each patient's unique risk factors and medical history **[14]**. This research aims to identify and categorise type 2 diabetes in patients using machine learning (ML) models, and to choose the optimal classification model to estimate the probability of developing diabetes. Retrospective data from 222 diabetic patients were obtained from JSS Hospital, Mysuru, Karnataka, in order to construct this paper.

Related work

By enhancing disease diagnosis, treatment, and patient outcomes, multimodal disease prediction utilising machine learning and deep learning algorithms has the potential to revolutionise the healthcare sector. In this context, with the aid of machine learning and deep learning algorithms, Akil Arsath J *et al* **[15]** developed a model that accurately predicted ailments like malaria, diabetes, and kidney, heart, and heart disease. They concentrated on predicting a variety of illnesses, paying close attention to the diabetes prediction assignment in particular. The effectiveness of different machine learning methods was demonstrated. Based on the results, it is clear that the random forest and stacked model algorithms perform better at predicting diabetes. For forecasting DM illness, Zou *et al* **[16]** applied ML algorithms and methodologies. DT, RF, and neural networks are the classifiers that are used. PIMA and hospital physical examination data from Luzhou, China are included in the dataset. The exami country dataset has 14 attributes, compared to 9 in the Te Pima dataset. The tool is called WEKA. The Random Forest Classifer has

the greatest reported accuracy of 80.8% for hospital data and 77% for the Pima dataset among the classifiers used. With various Classifers and Techniques, the accuracy can be increased even more. Self-organizing maps (SOMs) and hybrid wavelet neural networks (HWNNs) have been used by Zarkogianni *et al* **[17]** in their research. Te dataset is compiled from 560 people who have been diagnosed with both diabetes and cardiovascular disease (CVD). The maximum AUC curve results in 71.48%. By utilising methods to generate accurate CVD risk scores, the suggested strategy outperforms Binomial Linear Regression (BLR). 41 of the 560 individuals had non-fatal CVD in addition to having DM. Out of the 41, 4 had a stroke, while the remaining 35 had coronary heart disease. The paper's weaknesses include the necessity to increase accuracy percentage and the ability to concentrate on a single dataset rather than a hybrid model. Diabetes and cardiovascular disease (CVD) have been categorised by Alic *et al* **[18]** using artificial neural networks (ANN) and Bayesian networks (BN). A multilayer neural network with the Levenberg-Marquardt learning algorithm is the ANN that is being employed. Te BN is Naive Bayes, which has the highest DM and CVD accuracy at 99.51% and 97.92%, respectively. For diabetes disease, the accuracy of utilising ANN is 72.7%, 99%, and for cardiovascular disease, it is 80%, 9591%. The accuracy of BN for diabetes disease is 71%, and for cardiovascular disease it is 78% and 979.20%. The sigmoid transfer function is used by ANN while probability theory is used by BN. Sneha and Gangil's **[19]** research focuses on the early identification of diabetes utilising the best feature selection. With a specificity of 98.20% and 98%, the algorithms DT and RF are applied. The accuracy according to Naive Bayes is 82.30%. In order to improve classification accuracy, the authors' research additionally generalises the features. Five algorithms in total—SVM, RF, NB, DT, and KNN—are contrasted. It employs the data mining software rapid-miner. It is done to analyse the dataset's features. As previously noted, DecisionTree and RandomForest provide the maximum accuracy. The accuracy of SVM in the suggested technique by Applied Computational Intelligence and Soft Computing is 77% for SVM and 82.30% for NB, compared to 77.73% and 73.48% in the existing method. The goal of the research going forward is to raise the metrics' worth. Tafa *et al* **[20]** developed an integrated model for the prediction of diabetes using the algorithms SVM and Naive Bayes. On the model, several different datasets have been employed. The information was gathered in Kosovo. There are eight attributes in the dataset. Data from 402 patients were collected, of whom 80 had type 2 diabetes. The distinctiveness of some

characteristics is that they have not been frequently employed in previous studies, such as diet and physical activity. The data has been split equally between training and testing. The suggested model uses a combination of techniques to achieve an accuracy of 97.6%. When used alone, SVM offers an accuracy of 95.52% whereas Naive Bayes offers an accuracy of 94.52%. The model may be used in the future to analyse and evaluate matrices using various ML methods. Six ML classifiers, including Multilayer Perceptron, J48, JRip, Hoefding Tree, Random Forest, and Bayes Net, were utilised by Mercaldo *et al* **[21]**. PIMA is the dataset in use. The two most common algorithms are Greedy Stepwise and Best First. In order to improve the performance classification, they are utilised to express the qualities. The four factors taken are age, body mass index, diabetes pedigree function, and plasma glucose concentration. The dataset is validated using a 10 fold cross. The Hoefding Tree algorithm produced the following results: precision value 0.757, recall value 0.762, and F-measure value 0.759. For future development and accuracy improvisation, the parameters and the algorithm employed on the model can both be changed. J48, SVM, RF, and K-Nearest Neighbours (KNNs) are only a few of the classifiers utilised by Kandhasamy and Balamurali **[22]**. The UCI repository is where the dataset was taken. The matrices being compared are sensitivity, accuracy, and specificity. Using 5-fold cross validation, the classification was done on the dataset both with and without preprocessing. According to the results, KNN and Random Forest classifiers had the best accuracy rate of 100% after preprocessing, while the decision tree J48 classifier had the highest accuracy of 73.82% without it. OWDANN has been utilised by Annamalai and Nedunchelian **[23]** to predict diabetes mellitus. The suggested approach comprises two phases: predicting the presence of a disease and estimating its severity. The PIMA dataset is used for the preprocessing. Features are taken out of the preprocessing dataset, and OWDANN is used for classification. The preprocessed diabetes positive dataset is used in the severity level estimation phase, and GDHC predictions are made. The results show a 98.97% accuracy, a 94.98% sensitivity, and a 95.62% specificity. According to Davitt *et al* **[24]**, an increased blood glucose level is a sign of a metabolic problem that results in either insulin resistance, decreased insulin secretion, or both. The three types of etiologic diabetes are T1DM, T2DM, and GDM. The classification of diabetes includes genetic effects on beta-cell function, genetic insulin action deficiencies, drug or chemical-induced exocrine

pancreatic illness, endocrinopathies, post-transplant, and genetic syndrome.

METHODOLOGY

Data description

The study's methodology is covered in this section. The order is as follows: a description of the variables; a description of the data source; The ML models (LR, K-NN, SVM, Gaussian, DT, and RF) that were employed in this work are described in (3); (4) Data processing: an explanation of the methods for cleaning and preparing the data; (5) exploratory data analysis, considering methods for examining and visualising the data; (6) Data modelling: ML model training; and (7) Model validation: an explanation of methods for verifying and assessing the models.

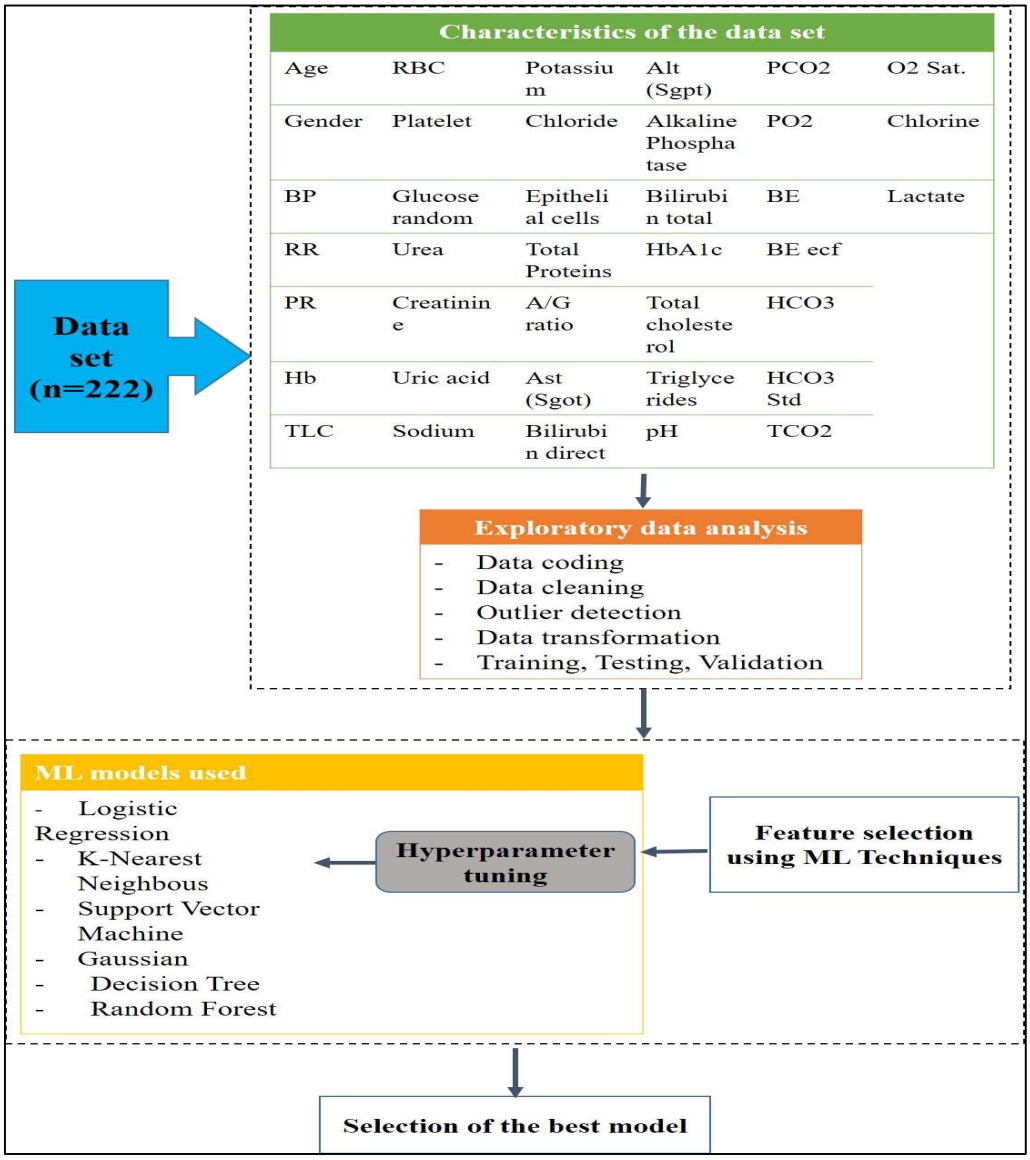


Figure 2: Model development process

Figure 2 shows the model development process. Database extraction and analysis, variable selection, training is shown in the process and on the basis of its performance, best model was selected.

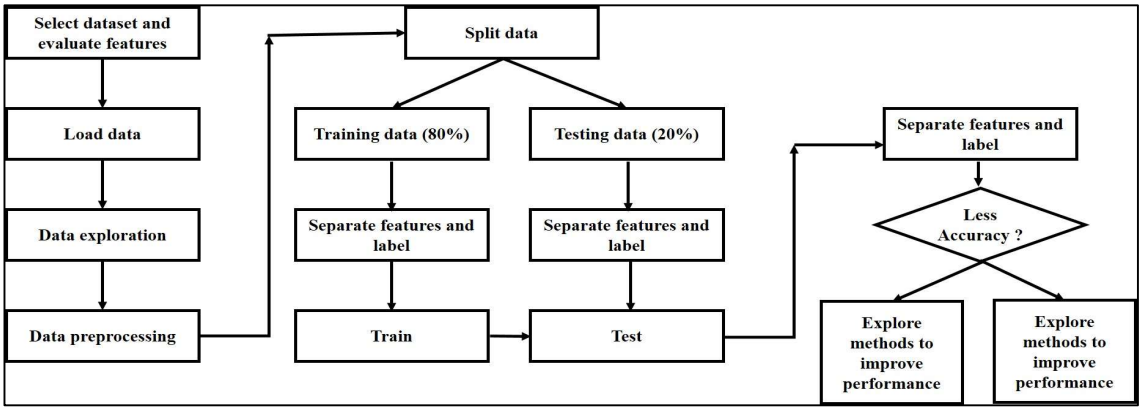


Figure 3: Flowchart of the prediction model

Description of classifiers

Logistic regression

It is a statistical technique used to examine a set of data that includes one or more independent factors that forecast a result. The result is measured using a dichotomous variable (takes just two potential values). By using a collection of independent variables and the dichotomous characteristic of interest (the dependent variable), such as the response or outcome variable, logistic regression determines the model that best describes the connection.

K-Nearest Neighbours

KNN is a supervised learning algorithm that learns how to label additional points in the problem space using a collection of labelled points as input. When a new point needs to be marked, the closest previously labelled points are considered. These points are able to cast ballots for their neighbours, who can be thought of as the new point's neighbours. The new point is given the title that the majority of its neighbours hold.

Support Vector Machine – Radial Basis Function

Each data point is represented as a point in n-dimensional space (n is the number of features) and the value of each feature corresponds to the value of a certain coordinate in the discriminative classifier known as SVM. In the feature space, the algorithm creates one or more hyperplanes (decision boundaries) that categorise the data points into separate groups.

Gaussian

A continuous probability distribution that is frequently used in statistical modelling and machine learning is the Gaussian distribution, also referred to as the normal distribution. It is characterised by its mean and standard deviation, and has a bell-

shaped curve that is symmetrical around its mean.

Decision Tree

A decision tree uses a tree structure to develop classification or regression models. An accompanying decision tree is created sequentially when the data set is divided into smaller and smaller subsets. In the end, this produces a tree with leaf nodes and decision nodes. The decision node at the top of a tree is called the root node. A decision node is represented by a leaf node, which has two or more branches and denotes a categorization or judgement.

Random forest

A massive number of separate decision trees are built using the Random Forest technique, which then produces a class that is either the mean prediction (regression) or mode prediction (classification) of all the individual trees. They are employed to reduce variance while maintaining bias. They thereby defeat the decision tree's propensity for over-fitting.

Data Processing

The project took use of the retrospective data from a private hospital in Mysuru, Karnataka. The patients in the data set are men and women between the age group of 7 and 81 years. Each record in the data set has attributes and labels indicating whether or not the patient has been diagnosed with diabetes. The obtained data was converted into CSV format and loaded into a variable. Dataset has 222 rows and 53 columns, including 118 male and 104 female data. All results were analysed using Google Colab – python 3.0 with libraries from Sklearn, Numpy, Pandas, Matplotlib platform. Logistic regression, K-nearest neighbour, support vector machine, Gaussian, decision tree and random forest algorithms were used.

Table 1: Dataset and its characteristics

Sl. No.	Attributes	Normal Range	Health Issue
1	Age (Years)	7 - 81 Years	
2	Gender	Male/Female	
3	Type of Diabetes	IDDM or NIDDM	
4	Blood Pressure (mm/Hg)	120/80	
5	Respiratory Rate (bpm)	12 - 18	Heart
6	Pulse (bpm)	60 - 100	Heart
7	Haemoglobin (gm/dl)	Men – 13.8 to 17.2 Women – 12.1 to 15.1	Heart
8	Total Leucocyte Count (Cells/cumm)	4000 - 11,000	Risk of infections
9	Red Blood Cell Count (million/cumm)	Men – 4.35 to 5.65 Women – 3.92 to 5.13	Anemia
10	Platelet Count (Lakh/cumm)	150,000 – 450,000	Leukemia and other infections
11	Glucose Random (mg/dl)	< 125	Diabetes
12	Urea (mg/dl)	5 - 20	Diabetes
13	Creatinine (mg/dl)	Men – 0.7 to 1.3 Women – 0.6 to 1.1	Diabetes
14	Uric acid (mg/dl)	3.5 to 7.2	Diabetes
15	Sodium (mEq/L)	135 – 145	Dehydration
16	Potassium (mEq/L)	3.5 – 5.2	Irregular heart beat
17	Chloride (mEq/L)	96 – 106	Kidney Disease

18	Pus cells	0 – 12	UTI
19	Epithelial Cells (cells//HPF)	15 – 20	Kidney Disease
20	Total Proteins (gm/dl)	6 – 8.3	Inflammation/infection of kidney/liver
21	A/G Ratio	1 - 2	Liver function
22	Ast (Sgot) (U/L)	8 - 33	Liver function
23	Bilirubin Direct (mg/dl)	< 0.3	Liver function
24	Alt (Sgpt) (U//L)	7 – 56	Liver function
25	Alkaline Phosphatase (IU/L)	44 – 147	Liver function
26	Bilirubin Total (mg/dl)	0.1 to 1.2	Liver function
27	Glycated Haemoglobin – HbA1C (%)	4 – 5.6	Diabetes
28	Total Cholesterol (mg/dl)	< 200	Heart
29	Triglycerides (mg/dl)	< 150	Heart
30	pH	7.35 to 7.45	Heart
31	pCO2 (mmHg)	35 – 45	Heart
32	pO2 (mmHg)	75 – 100	Heart
33	Base Excess	-2 - +2	Heart
34	BE ecf	0 ± 4	Heart
35	HCO3	22 – 26	Heart
36	HCO3 Std	22 – 26	Heart
37	TCO2 (mEq/L)	22 – 30	Heart
38	O2.Sat. (%)	> 94	Heart
39	Chlorine (mEq/L)	96 - 106	Heart
40	Lactate (mEq/L)	< 2	Heart

There are 222 records of patients with different health problems. The health issues mentioned are recorded as 40 attributes shown in the **Table 1**.

With the help of the acquired data, 11 major health issues were classified into outcomes as shown in the **Table 2**.

Further, the 11 major Health issues were classified as low, medium and high as shown in the **Table 3**.

Table 2: Classification into major health problems

Outcome	Disease
1	Heart Diseases
2	Risk of infections
3	Anemia
4	Leukemia and other infections
5	Diabetes
6	Dehydration
7	Irregular heart beat
8	Kidney Disease
9	UTI
10	Liver function test
11	Inflammation/infection of kidney/liver

Table 3: Classification of the major health problems into categories

Number of health issues	Category
4-6	Low-0
7-8	Medium-1
9-11	High-2

Table 4: Diabetes data

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 222 entries, 0 to 221
Data columns (total 53 columns):

#	Column	Non-Null Count	Dtype
0	Sl. No.	222 non-null	int64
1	Age (Years)	222 non-null	int64
2	Gender	222 non-null	object
3	Type	222 non-null	object
4	BP (mm/Hg)	222 non-null	object
5	Respiratory Rate (bpm)	222 non-null	int64
6	Pulse (bpm)	222 non-null	int64
7	Hemoglobin (gm/dl)	222 non-null	float64
8	Total Leucocyte Count (Cells/cumm)	222 non-null	int64
9	Red Blood Cell Count (million/cumm)	222 non-null	float64
10	Platelet Count (Lakh/cumm)	222 non-null	float64
11	Glucose Random (mg/dl)	222 non-null	int64
12	Urea (mg/dl)	222 non-null	int64
13	Creatinine (mg/dl)	222 non-null	float64
14	Uric Acid (mg/dL)	222 non-null	float64
15	Sodium (mEq/L)	221 non-null	float64
16	Potassium (mEq/L)	222 non-null	float64
17	Chloride (mEq/L)	222 non-null	float64
18	Pus Cells	222 non-null	int64
19	Epithelial Cells (cells/HPF)	222 non-null	int64
20	Total proteins (gm/dl)	222 non-null	float64
21	A/G Ratio	221 non-null	float64
22	Ast(Sgot) (U/L)	222 non-null	int64
23	Bilirubin Direct (mg/dl)	221 non-null	float64
24	Alt(Sgpt) (U/L)	221 non-null	float64
25	Alkaline Phosphatase (IU/L)	222 non-null	int64
26	Bilirubin Total (mg/dl)	222 non-null	float64
27	HbA1c	222 non-null	float64
28	Total cholesterol (mg/dl)	222 non-null	int64
29	Triglycerides (mg/dl)	222 non-null	int64
30	pH	222 non-null	float64
31	PCO2 (mmHg)	222 non-null	float64
32	PO2 (mmHg)	222 non-null	float64
33	Base Excess	222 non-null	float64
34	BEecf	222 non-null	float64
35	HCO3	222 non-null	float64
36	HCO3 Std	222 non-null	float64
37	TCO2 (mEq/L)	222 non-null	float64
38	O2.Sat (%)	222 non-null	float64
39	Potassium (mEq/L).1	222 non-null	float64
40	Chlorine (mEq/L)	222 non-null	int64
41	Lactate (mmol/L)	222 non-null	float64
42	outcome1	222 non-null	int64
43	outcome2	222 non-null	int64
44	outcome3	222 non-null	int64
45	outcome4	222 non-null	int64
46	outcome5	222 non-null	int64
47	outcome6	222 non-null	int64
48	outcome7	222 non-null	int64
49	outcome8	222 non-null	int64
50	outcome9	222 non-null	int64
51	outcome10	222 non-null	int64
52	outcome11	222 non-null	int64

dtypes: float64(25), int64(25), object(3)

Pandas library, read_csv(), was used, it is shown in **Table 4**. There were no null values in the dataset.

Histograms and libraries were used in the exploratory data analysis to discover characteristics with zero values and replace them with zero values during the cleaning phase. Following the application of the histogram, 222 diabetes patients—118 men and 104 women—make up the dataset. The statistical values of the dataset, including the number of observations, mean, standard deviation, minimum and maximum values, and the lower, median, and upper quartiles (Q1, Q2, and Q3), were then checked. These values provide information about the distribution of the data in percentile terms. **Table 5** presents the findings.

RESULTS AND DISCUSSIONS

Table 5: Statistical values of the given dataset

	Count	Mean	Std.	Min.	25%	50%	75%	Max.
Sl. No.	222	111.50	64.23	1	56.25	111.50	166.75	222
Age	222	57.10	15.15	7	50	58	66.75	92
Respiratory Rate	222	25.01	8.22	12	18.25	22	31.75	54
Pulse	222	88.54	14.20	60	80	86	92	184
Haemoglobin	222	11.84	2.32	6.30	10.10	12.20	13.40	17
Total Leucocyte Count	222	8567.92	6355.42	0	5187.50	8425	11542.50	33260
Red Blood Cell Count	222	3.27	1.99	0	2.01	4.07	4.72	6.31
Platelet Count	222	2.31	1.71	0	1.39	2.25	3.16	9.15
Glucose Random	222	261.90	104.15	67	184	249	321	654
Urea	222	40.42	29.32	9	23	32	47	239
Creatinine	222	1.05	1.39	0	0.60	0.77	1.17	12.55
Uric Acid	222	5.99	2.04	2.01	4.30	5.83	7.79	12.10
Sodium	222	105.07	54.92	0	124	132	136	149
Potassium	222	3.46	1.98	0	3.2	4.20	4.70	7.90
Chloride	222	75.69	40.72	0	86	95	100	113
Pus cells	222	135.02	375.03	0	0	12	24	2022
Epithelial cells	222	273.68	3030.46	0	0	12	23.75	45085
Total Proteins	222	3.68	3.33	0	0	5.69	6.60	8.40
A/G Ratio	222	0.70	0.70	0	0	0.80	1.40	2.20
Ast (Sgot)	222	17.63	49.61	0	0	13	23	690
Bilirubin Direct	222	0.12	0.39	0	0	0.07	0.16	5.62
Alt (Sgpt)	222	32.15	274.90	0	0	12	18	4088
Alkaline Phosphatase	222	87.95	115.83	0	0	72	122.50	948
Bilirubin Total	222	0.32	0.56	0	0	0.18	0.49	6.06
HbA1c	222	10.13	1.98	5.40	8.70	9.55	11.80	15
Total Cholesterol	222	162.35	48.12	85	121	152	205	342
Triglycerides	222	262.33	87.06	100	201	248	323.50	475
pH	222	1.19	2.71	0	0	0	0	7.49
pCO2	222	5.18	12.50	0	0	0	0	54.90
pO2	222	14.99	38.57	0	0	0	0	255
Base Excess	222	1.19	4.15	0	0	0	0	30.70
BE ecf	222	1.21	4.08	0	0	0	0	28.50
HCO3	222	3.03	7.54	0	0	0	0	31.10
HCO3 Std.	222	3.24	7.84	0	0	0	0	29
TCO2	222	4.01	10.79	0	0	0	0	54
O2. Sat	222	93.77	5.59	36.30	91.02	94.80	97	99.40
Chlorine	222	17.31	39.98	0	0	0	0	7.80
Lactate	222	0.28	0.76	0	0	0	0	134

Understanding the relationships between the variables in the provided information is crucial for utilising machine learning (ML) models to detect and classify diabetes. This is accomplished through the application of data analytic techniques and software tools. In order to do this, we imported a dataset using the Python programming language, computed a correlation matrix, and executed the correlation analysis technique (Figure 5). Additionally, we used a Python package to examine histograms related to various disease characterizations. Histograms representing a single variable and are used to visualize the shape of the distribution of the particular variable. This distribution conveys how often a value occurs. The histograms for all the features in the data set is as represented in the following **Figure 4**.

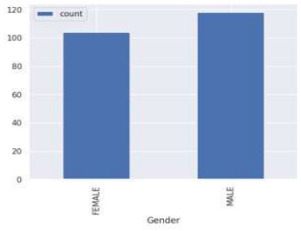
After completing the exploratory analysis of the dataset, the next step was to run the training, which divides the data into a ratio of 80% of training set and 20% of testing set. For this, we used the library SKlearn. Model_selection.train_test_split() since this library allows us to perform the analysis by specifying test sizes.

Heart Disease

Heart Disease depends on 15 parameters:

	Gender	outcome1	count
0	FEMALE	1	104
1	MALE	1	118

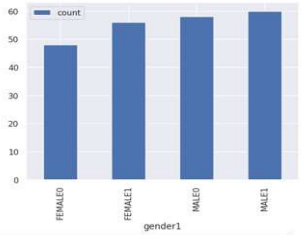
All patients are having heart issue.



Risk of infections

48 females and 58 males has no infections

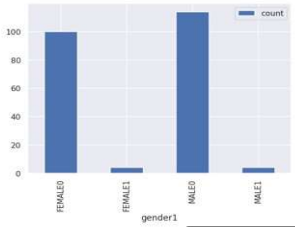
	Gender	outcome2	count
0	FEMALE	0	48
1	FEMALE	1	56
2	MALE	0	58
3	MALE	1	60



Iron deficiency – anemia

There is only 4 male and female have iron deficiency – anemia

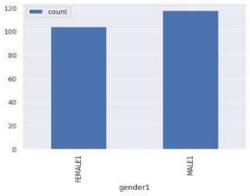
	Gender	outcome3	count
0	FEMALE	0	100
1	FEMALE	1	4
2	MALE	0	114
3	MALE	1	4



Leukemia and other infections

All of them found to have leukemia and other infections

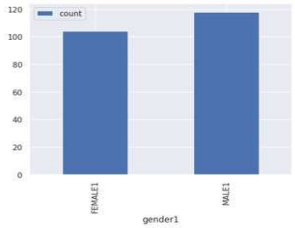
	Gender	outcome4	count
0	FEMALE	1	104
1	MALE	1	118



Diabetes

All of them are having Diabetes

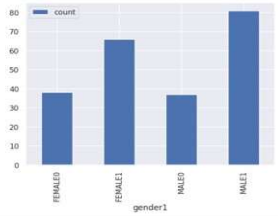
	Gender	outcome5	count
0	FEMALE	1	104
1	MALE	1	118



Dehydration

Dehydration is more among men than women

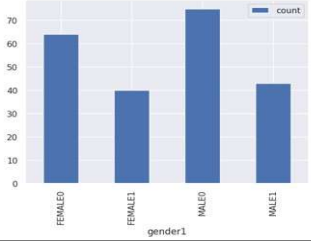
	Gender	outcome6	count
0	FEMALE	0	38
1	FEMALE	1	66
2	MALE	0	37
3	MALE	1	81



Irregular Heart Beat

Most of the females have irregular heart beat than males.

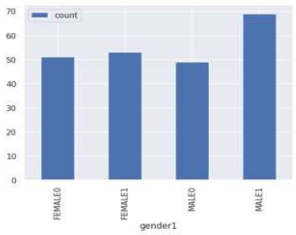
	Gender	outcome7	count
0	FEMALE	0	64
1	FEMALE	1	40
2	MALE	0	75
3	MALE	1	43



Kidney disease

More male are infected with kidney problems than female

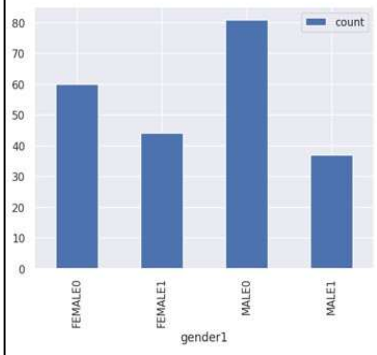
	Gender	outcome8	count
0	FEMALE	0	51
1	FEMALE	1	53
2	MALE	0	49
3	MALE	1	69



UTI

More females are having UTI problem than males

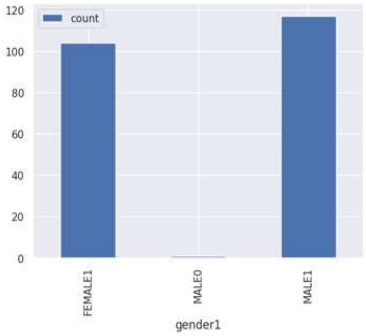
	Gender	outcome9	count
0	FEMALE	0	60
1	FEMALE	1	44
2	MALE	0	81
3	MALE	1	37



Liver Function Test

All of them are having liver function problem

	Gender	outcome10	count
0	FEMALE	1	104
1	MALE	0	1
2	MALE	1	117



Inflammation/infection of kidney/liver

More of males are having inflammation/ infection of kidney/liver problem.

	Gender	outcome11	count
0	FEMALE	0	49
1	FEMALE	1	55
2	MALE	0	50
3	MALE	1	68

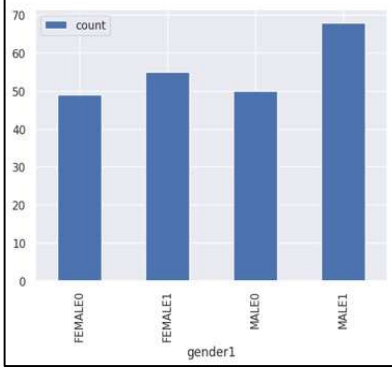


Figure 4: Histogram of features for all the records in the data set

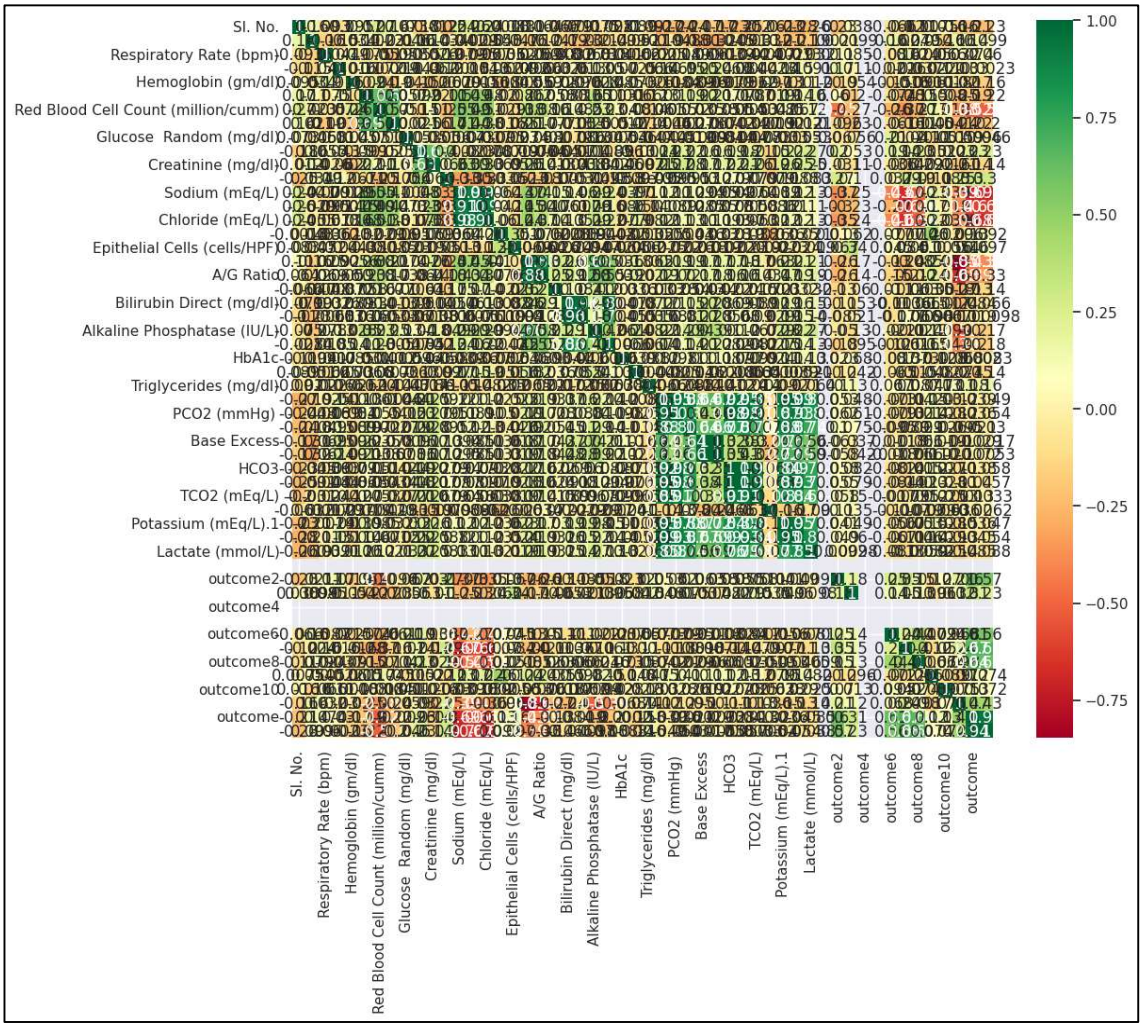


Figure 5: Correlation values and heat map, showing the correlation between the features for the dataset

MALE 118
FEMALE 104
Name: Gender, dtype: int64

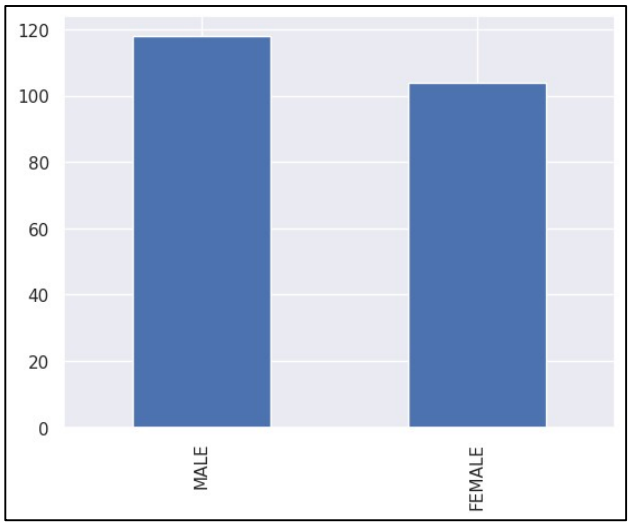


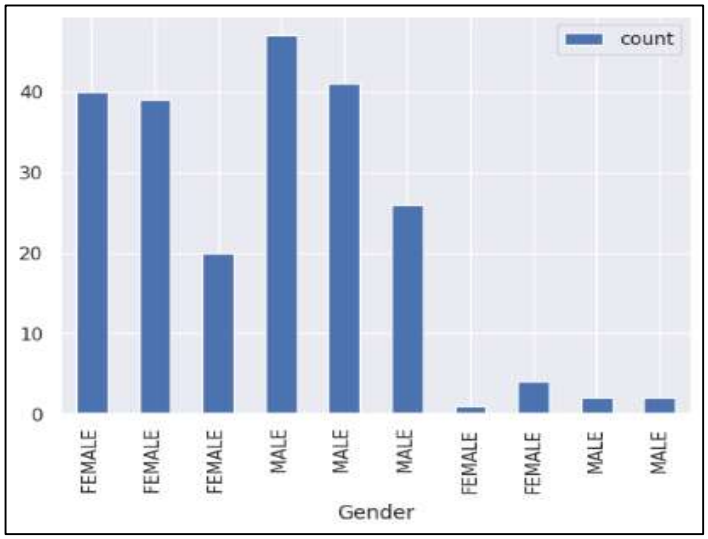
Figure 6: Total number of patients

Figure 6 Illustrates the total number of diabetic patients including male and female patients in the data set.

Following the 222 patient original data set's training for the LR, KNN, RF, DT, and SVM models. Distributing the data so that, for training and validation, respectively, 80% and 20% were used. We continued with the training with the goal of utilising machine learning (ML) models to identify and categorise individuals with various illnesses and then choose the most accurate model to predict the risk of diabetes. We had six classification models, among which LR achieved 62% accuracy, KNN reached 62% accuracy, SVM reached 64% accuracy, Gaussian reached 80%, DT reached 99% accuracy and RF reached 98% accuracy. According to the results, DT came out to be the best model to identify and classify different diseases with respect to NIDDM using ML model.

```
diabetes_df_group = (
    diabetes_df.groupby(["Gender", "final_output"])
        .size()
        .reset_index(name='count')
)
print(diabetes_df_group)
# plotting graph
#diabetes_df_group.plot(x="Gender", y=["final_output","count"], kind="bar")
diabetes_df_group.plot(x="Gender", y=["count"], kind="bar")
```

```
Gender  final_output  count
0   FEMALE         0.0     40
1   FEMALE         1.0     39
2   FEMALE         2.0     20
3     MALE         0.0     47
4     MALE         1.0     41
5     MALE         2.0     26
6   FEMALE         0.0      1
7   FEMALE         2.0      4
8     MALE         1.0      2
9     MALE         2.0      2
<Axes: xlabel='Gender'>
```



```
#classifying Outcome

#diabetes_df1=diabetes_df.drop(["Gender","Type","Sl. No.","Age (Years)","BP
(mm/Hg)"],axis=1)
y=diabetes_df1['final_output']
i=len(diabetes_df1.columns)
X=diabetes_df1.iloc[:,0:i-1].values
print(X)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y)


#Using Logistic Regression Algorithm to the Training Set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)
print('Logistic regression',classifier.score(X_test,y_test))

#Using KNeighborsClassifier Method of neighbors class to use Nearest Neighbor algorithm
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
classifier.fit(X_train, y_train)
print('Kneighbors',classifier.score(X_test,y_test))

#Using SVC method of svm class to use Kernel SVM Algorithm
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)
classifier.fit(X_train, y_train)
print('SVM-rbf',classifier.score(X_test,y_test))

#Using GaussianNB method of naïve_bayes class to use Naïve Bayes Algorithm
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)

print('Gaussian',classifier.score(X_test,y_test))

#Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm

from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
classifier.score(X_test,y_test)
print('Decision Tree',classifier.score(X_test,y_test))

#Using RandomForestClassifier method of ensemble class to use Random Forest
Classification algorithm
```

```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state =
0)
classifier.fit(X_train, y_train)
classifier.score(X_test,y_test)
print("Random forest",classifier.score(X_test,y_test))
```

Table 6 summarizes accuracies of all the six algorithms.

Table 6: Algorithm accuracies

Sl. No.	Algorithms	Accuracy
1	Logistic Regression	62%
2	K-Neighbours	62%
3	Support Vector Model	64%
4	Gaussian	80%
5	Decision Tree	99%
6	Random Forest	98%

CONCLUSION

People today are prone to a wide range of illnesses brought on by environmental factors and lifestyle decisions. Therefore, it is now essential to predict diseases before they manifest. In this research, systematic efforts were made to design a system which results in predicting the type 2 diabetes mellitus. Six machine learning algorithms were applied, various models for the diagnosis of multiple diseases were made. Each algorithm was analysed and compared with each other using numerous valuation parameters. After pre-processing the null values and removing missing data, logistic regression was used to first create a predictive model. Accuracy and execution time were improved by using approaches of feature selection. A significant difference in the accuracies provided by various classifiers were noted. The result of the test shows that the decision tree performed well of the dataset of diabetic patients collected from a private hospital in Mysuru, Karnataka with an accuracy of 99%. In future, the predicted system with the used ML classification algorithms can be used to predict various diseases for a larger dataset and also to classification with respect to type 1 and type 2 diabetes mellitus, which will in turn help to predict early diseases.

REFERENCES

[1] Li, Z.; Han, D.; Qi, T.; Deng, J.; Li, L.; Gao, C.; Gao, W.; Chen, H.; Zhang, L.; Chen, W. Hemoglobin A1c in Type 2 Diabetes Mellitus Patients with Preserved Ejection Fraction Is an Independent Predictor of Left Ventricular Myocardial Deformation and Tissue Abnormalities. BMC Cardiovasc. Disord. 2023, 23, 49.

[2] OMS Diabetes—World Health Organization. Available online:

https://www.who.int/es/news-room/fact-sheets/detail/diabetes (accessed on 20 February 2023).

[3] OPS/OMS Diabetes—PAHO/WHO: Pan American Health Organization. Available online: https://www.paho.org/es/temas/ diabetes (accessed on 20 February 2023).

[4] PAHO PAHO/WHO|Pan American Health Organization. Available online: https://www.paho.org/en (accessed on 25 February 2023).

[5] International Diabetes Federation. IDF Diabetes Atlas|Tenth Edition. Available online: https://diabetesatlas.org/ (accessed on 25 February 2023).

[6] El-Attar, N.E.; Moustafa, B.M.; Awad, W.A. Deep Learning Model to Detect Diabetes Mellitus Based on DNA Sequence. Intell. Autom. Soft Comput. 2022, 31, 325–338.

[7] Mohamed, A.T.; Santhoshkumar, S. Deep Learning Based Process Analytics Model for Predicting Type 2 Diabetes Mellitus. Comput. Syst. Sci. Eng. 2022, 40, 191–205.

[8] Philip, N.Y.; Razaak, M.; Chang, J.; Suchetha, M.S.; Okane, M.; Pierscionek, B.K. A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes. IEEE Access 2022, 10, 13460–13471.

[9] Susana, E.; Ramli, K.; Murfi, H.; Apriantoro, N.H. Non-Invasive Classification of Blood Glucose Level for Early Detection Diabetes Based on Photoplethysmography Signal. Information 2022, 13, 59.

[10] Zhou, H.; Myrzashova, R.; Zheng, R. Diabetes Prediction Model Based on an Enhanced Deep Neural Network. EURASIP J. Wirel. Commun. Netw. 2020, 2020, 148.

[11] American Diabetes Association. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. Diabetes Care 2018, 41, S13–S27.

[12] Thotad, P.N.; Bharamagoudar, G.R.; Anami, B.S. Diabetes Disease Detection and Classification on Indian Demographic and Health Survey Data Using Machine Learning Methods. Diabetes Metab. Syndr. Clin. Res. Rev. 2023, 17, 102690.

[13] Azit, N.A.; Sahran, S.; Leow, V.M.; Subramaniam, M.; Mokhtar, S.; Nawi, A.M. Prediction of Hepatocellular Carcinoma Risk in Patients with Type-2 Diabetes Using Supervised Machine Learning Classification Model.

Heliyon 2022, 8, e10772. [CrossRef]

[14] Aggarwal, S.; Pandey, K. Early Identification of PCOS with Commonly Known Diseases: Obesity, Diabetes, High Blood Pressure and Heart Disease Using Machine Learning Techniques. Expert Syst. Appl. 2023, 217, 119532.

[15] Akil Arsath J, Suganthi S, Rakeshwaran S, Karthiga S. Multimodal Disease Prediction using Machine Learning and Deep Learning Techniques. International Research Journal on Advanced Science Hub. 2023, 5 (5S), 2582-4376.

[16] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Frontiers in Genetics, vol. 9, p. 515, 2018 Nov 6.

[17] K. Zarkogianni, M. Athanasiou, A. C. Tanopoulou, and K. S. Nikita, "Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication," IEEE J Biomed Health Inform, vol. 22, no. 5, pp. 1637–1647, 2017.

[18] B. Alic, L. Gurbeta, and A. Badnjevi ´ c, "Machine learning ´ techniques for classifcation of diabetes and cardiovascular diseases," in Proceedings of the 2017 6th Mediterranean Conference on Embedded Computing (MECO), pp. 1–4, Bar, Montenegro, June 2017.

[19] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," Journal of Big Data, vol. 6, no. 1, p. 13.

[20] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," in Proceedings of the 2015 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378–382, Budva, Montenegro, June 2015.

[21] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients classifcation and diagnosis through machine learning techniques," Procedia Computer Science, vol. 112, pp. 2519–2528, 2017.

[22] J. pradeep Kandhasamy and S. Balamurali, "Performance analysis of classifer models to predict diabetes mellitus," Procedia

Computer Science, vol. 47, pp. 45–51, 2015.

[23] R. Annamalai and R. Nedunchelian, "Diabetes mellitus prediction and severity level estimation using OWDANN algorithm," Computational Intelligence and Neuroscience, Article ID 5573179, 11 pages, 2021.

[24] C. Davitt, K. E. Flynn, R. K. Harrison, A. Pan, and A. Palatnik, "Current practices in gestational diabetes mellitus diagnosis and management in the United States: survey of maternal fetal medicine specialists," American Journal of Obstetrics and Gynecology, vol. 225, no. 2, pp. 203-204, 2021 Aug.

[25] B. S. Ahamed and M. S. Arya, "Prediction of type 2 diabetes using the LGBM classifier methods and techniques," Turkish Journal of Computer and Mathematics Education, vol. 12, No.12.