An Effective Diabetes Prediction System Using Machine Learning Algorithms

Prof. Aditi Malkar[1], Devansh Trivedi[2], Altmash Siddique[3], Rishank Shah[4]
[1]Assistant Professor, *Department of Computer Engineering,*
[2]*BE Student, Department of Computer Engineering,*
[3]*BE Student, Department of Computer Engineering,*
[4]*BE Student, Department of Computer Engineering,*
[1,2,3,4]*MCT's Rajiv Gandhi Institute of Technology, Mumbai*

***Abstract*— Diabetes is a prevalent metabolic disorder that affects a significant number of people globally. Timely detection and treatment of diabetes can prevent complications and improve health outcomes. The healthcare industry is facing an increasing demand for better patient care and disease prediction systems. This study proposes a Disease Prediction System that integrates various features, including an AI ChatBot, Diabetes Prediction System, Chat and Appointment Booking System, to improve disease prediction accuracy. The Random Forest algorithm is utilized in the Diabetes Prediction System, which enhances the overall accuracy of the system. With multiple inputs, the system becomes proficient in accurately classifying diseases and predicting outputs. The system's accuracy was evaluated using a patient information dataset, resulting in an overall accuracy of 90.4%. These results demonstrate the Disease Prediction System's potential to improve healthcare outcomes by providing timely and accurate disease prediction. In conclusion, this study's proposed system has the potential to significantly benefit healthcare providers and the medical field. With its high disease prediction accuracy, efficient disease classification, and user-friendly features, this system can assist healthcare professionals in making precise diagnoses, providing effective treatments, and enhancing patient outcomes.**

Index Terms*— *Diabetes Prediction, Machine Learning, Manual Information, Random Forest, Decision Trees, Adaptive Boosting.

I. INTRODUCTION

The condition or disease which is permanent to the human body for more than three months is known as a chronic condition, There are different types of chronic diseases such as cancer. Alzheimer's. Arthritis. Asthma. Heart disease. and Diabetes. These types of diseases force people to change the daily lifestyle of the affected people. Diabetes occurs when the level of blood sugar is too high. Diabetes occurs when enough insulin can't be produced by the pancreas, or the produced insulin can't be utilized by our body [1]. Insulin regulates blood sugar in the human body. Uncontrolled diabetes occurs due to the abnormal rise of sugar in the blood. In the long run, it increases the chance of severe damage to organs such as nerves and blood vessels. As a result, this disease increases the peril of several fatal diseases [1]. According to WHO (World Health Organization), diabetes is of three types: (i) Juvenile diabetes (Type 1) that occurs when enough insulin can't be produced by the body [2]. Generally, children and young adults are affected by it [2]. (ii) Type II diabetes happens when the body can't utilize insulin effectively. Middle-aged people who are more than 45 years old are mostly affected by Type If diabetes [3]. But nowadays, it is also developed in children and young adults. At present, 95% of all diabetes is of Type II [3]. (iii) Gestational diabetes (Type III) occurs due to a high glucose level in the blood. Usually, it is diagnosed in women during pregnancy and had no experience of diabetes before. According to the latest report released by WHO, total diabetic patients have increased from 108 million to 422 million from 1980 to 2014 [1]. According to WHO, diabetes mellitus will be the seventh leading cause of mortality by 2030 [1]. A recent study has stated that 642 million young adults will be affected by diabetes by 2040 [3]. In 2016, diabetes directly affected the deaths of approximately 1.6 million people [1]. Unfortunately, this disease cannot be eradicated.

But we can control it by restraining the glucose level in blood. When diabetes is detected, its effect can be minimized. But this is not an easy task. To identify the disease, data are taken from patients like insulin, age, body mass index, family history of diabetes, etc. and then consulted to a doctor [4]. Then the doctor decides using his/her knowledge and experience. But this identification process is very time

consuming and sometimes most costly. Sometimes, it also misleads the diagnosis process due to the lack of experience of the doctors. Computer automated diagnosis can play an important character in the detection of diabetic patients. A lot of research has been done to identify diabetic patients at an early stage. Several medical datasets have been developed to accelerate the research in this field. Due to the non-linear, non-normal, and complex nature of the medical data, classification of the diabetic patients accurately is a challenging task for the researchers. That's why Machine Learning (ML) and deep learning methods are widely utilized to extract valuable knowledge from the dataset and predict diabetes disease. Various machine learning methods have been applied to classify diabetic patients. Their classification accuracy was not significant enough because most of them ignored the importance of handling missing values, removal of irrelevant features, and handling outliers. While interpreting the medical datasets, classification accuracy mostly depends on the pre-processing of the data. After considering all of these factors, the classification of diabetic patients remains an arduous task for the early diagnosis of patients [5].

This study emphasizes on how to improve the classification accuracy for the early diagnosis of diabetic patients. To classify diabetic patients, a Tree-Based prediction model has been proposed based on machine learning techniques. At first, missing values are handled by their group mean value. After handling the missing values. outliers are detected and handled using the Robust scaling normalization technique. An oversampling method is adopted to avoid the imbalanced class distribution problem. Then an effective feature selection technique is applied to build a Tree-Based prediction model for early diagnosis of diabetes patients.

II. LITERATURE REVIEW

Over the last few years, various researchers have worked on predicting diabetes using different machine learning approaches. Jeevan, Rajesh and Vijay. [1] employed three classifiers, namely Decision Tree, Support Vector Machine, and Naive Bayes, to detect diabetes at an early stage using the Pima Indians Diabetes Dataset (PIDD). Their results indicated that Naive Bayes outperformed the other two classifiers with an accuracy of 81.21%.

Another study by Vignana Jyothi. [2] used normalization and random forest to classify diabetes patients. They pre-processed the data using normalization and achieved an accuracy of 80.6%.

In a different study, Elif Nur, Belgium Erkal and Tulin [3] utilized K-Nearest Neighbor and Naive Bayes classifiers to predict diabetes. Their model achieved an accuracy of 84% on the PIDD dataset.

In a similar work, Vinay, Vasu, Shreya, Jay and Bhushan [4] used K-fold cross-validation to predict diabetes. They compared the performance of four classifiers, namely K-Nearest Neighbor, Decision Tree, Naive Bayes, and Support Vector Machine. Their results showed that K-Nearest Neighbor performed the best with an accuracy of 82.68%.

In conclusion, these studies show that machine learning algorithms can accurately predict diabetes, and different algorithms and techniques can be used to obtain high accuracy rates.

III. PROPOSED METHODOLOGY

A. Dataset Description

The Pima Indians Diabetes Dataset (PIDD) was utilized in this research, obtained from the University of California, Irvine (UCI) machine learning repository. The dataset comprises 768 instances, with a majority of female subjects at 21 years of age. There are two classes in the dataset based on the binary class attribute. A value of '0' indicates no diabetes (-ve), while '1' denotes diabetes (+ve). The dataset contains eight independent attributes, with 500 instances (65.1%) indicating no diabetes (-ve) and 268 instances (34.1%) indicating diabetes (+ve).

The details of the PIDD dataset is given in Table I.

TABLE I - DESCRIPTION OF PIDD DATASET

S.N.	Name of Attribute	Attribute Description	Percentage of Missing Values
1	Pregnancies	No. of times pregnant	0
2	Glucose	Glucose concentration for 120 mins	0.65%
3	Blood Pressure	Diastolic blood pressure	4.55%
4	Skin Thickness	Skin fold thickness	29.55%

5	Insulin	Serum-insulin (2 hours)	48.69%
6	BMI	Body mass index	1.43%
7	Pedigree Function	Diabetes pedigree function	0
8	Age	Age in years	0
9	Class	1 for diabetes positive and 0	0

B. Data Pre-processing

Real-world medical data is often complex in nature and may contain missing values, inconsistent data, and outliers, making it non-linear and non-normal. Data pre-processing plays a crucial role in the performance of any classification algorithm. In this research, our objective is to increase the accuracy of diabetic patient prediction, and thus we have taken steps to handle the data efficiently. Data pre-processing techniques utilized in this research include missing value imputation, removal of irrelevant features, outlier detection, oversampling, and feature scaling. To facilitate these steps, we have proposed a flowgraph model, illustrated in Fig.1.

Fig. 1 Workflow of Proposed Model

I) Missing Values Imputation

The Pima Indians Diabetes Dataset contains missing values for glucose (0.65%), blood pressure (4.55%), skin thickness (29.55%), insulin (48.69%), and BMI (1.43%). To handle these missing values, mean imputation was performed, whereby the missing values for each attribute were filled in with the mean value of that attribute.

2)Eliminating Outliers

Outliers in a dataset can significantly affect the calculated parameters; therefore, it is important to identify and remove them from the data. In this research, the Inter Quartile Range (IQR) method was used to detect and eliminate the outliers in the dataset.

The IQR method identifies outliers by first calculating the first quartile (Q1) and the third quartile (Q3) points. The IQR is further calculated as Q3-Q1.

The normal data range is defined with a lower limit of Q1-1.5IQR and an upper limit of Q3+1.5IQR.

To ensure accurate analysis, any data point that falls outside of the specified range is deemed an outlier and should be excluded from further analysis.

The concept of quartiles and IQR can be best visualized using a boxplot, where the minimum and maximum points are defined as

Q1-1.5IQR and Q3+1.5IQR, respectively,

and any point outside this range is considered an outlier. It is important to note that extreme data points do not always necessarily mean they are outliers.

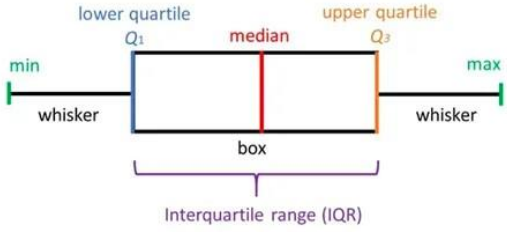

Fig. 2 Box plot showing Quartile distribution

3) Oversampling

Oversampling is a technique used to address class imbalance in a dataset. Class imbalance occurs when one class has significantly more samples than the other(s), which can result in biased classification models that favour the majority class. Oversampling involves randomly duplicating samples from the minority class until the dataset is balanced. This technique ensures that the classification model is not biased towards the majority class.

4) Feature Scaling

Feature scaling is the process of standardizing the range of independent feature values in a dataset. It is important to scale features before creating a

machine learning model because some algorithms are sensitive to the scale of the input features. Feature scaling ensures that the features are on a similar scale, which helps the model converge faster and reduces the impact of outliers. The Robust Scaler is a technique used to scale the data by using the interquartile range, which makes it robust to the presence of outliers in the dataset.

C. Classification

1) Decision Tree (DT)

The Pima Indians Diabetes Dataset has been classified using Decision Tree (DT), a supervised ML technique that constructs a model for classification and regression in the structure of a tree. DT has been widely used for the classification of disease diagnosis as it is easily explainable [16]. Moreover, it requires less data cleaning as outliers and missing values have less significance on the model's data. In DT, the class is predicted based on decision rules taken from input data. It splits the data into subsets of data and represents those subsets in a tree structure wherein each node a decision is made. The final classification is extracted from leaf nodes.

2) Random Forest (RF)

Random Forest is an ensemble ML algorithm that generates numerous classification models (decision trees) where each model is constructed using a feature selector such as Gini Index, Information Gain, and Gain Ratio. These models learn and make contributions to the prediction in a discrete manner. The final result is made from those obtained predictions.

3) Adaptive Boosting (AB)

Adaptive Boosting or AdaBoost integrates many weak classifiers to generate a strong classifier. AdaBoost sets weights to each weak classifier and ensures correct classification by training the sample data in each iteration while predicting outliers or unusual observations. The intuition behind this classification technique is that a single classifier can accurately predict a portion of the dataset giving incorrect results for other portions, but incorrect portions can be correctly predicted by other weak classifiers. The combination of weak classifiers is represented by

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where $h_t(x)$ is a weak classifier. The result of the final classifier is $H(x) = \text{sign}(f(x))$.

D. Evaluation Measures

To evaluate the effectiveness of our proposed model, we utilized K-fold cross-validation techniques to randomly split the dataset into k subsets to construct the training and test sets. During each iteration, k-1 subsets were used to train the model, while the remaining subset was used for testing. By repeating this process k times, we obtained the model's performance by averaging the test results of the independent k subsets. In this study, we used 10-fold cross-validation to reduce bias and variance. Various statistical measures, such as F1-score, precision, recall, and Receiver Operating Characteristic Curve (ROC AUC), were considered to evaluate the model's performance. Typically, the confusion matrix summarizes the overall performance of any prediction model, and accuracy, F1-score, recall, and precision can be derived from it. From the confusion matrix accuracy, F 1 score, recall, and precision are intended as follows.

TABLE II. CONFUSION MATRIX

Actual	Predicted	
	Positive	*Negative*
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{(TN + TP + FN + FP)}$$

$$\text{Precision} = \frac{(\text{TP})}{(FP + TP)}$$

$$\text{Recall} = \frac{(\text{TP})}{(FN + TP)}$$

$$\text{F I - Score} = \frac{2\,(\text{TP})}{(\text{FN} + \text{FP} + 2\text{TP})}$$

IV. RESULTS AND DISCUSSION

The experiment was conducted on Google Collab, a free cloud-based service. To evaluate the performance of the model, we used the 10-fold cross-validation technique. 80% of the dataset was used for training and the remaining 20% was used to test the accuracy of the model. The classification phase was divided into two steps, where all features were used for classification. The results, shown in Figure 2, indicate that the Random Forest Classifier, Decision Trees Classifier, and AdaBoost Classifier achieved accuracies of 90.4%, 85.1%, and 82.4%,

respectively. The highest accuracy was achieved using the Random Forest Classifier at 90.4%.

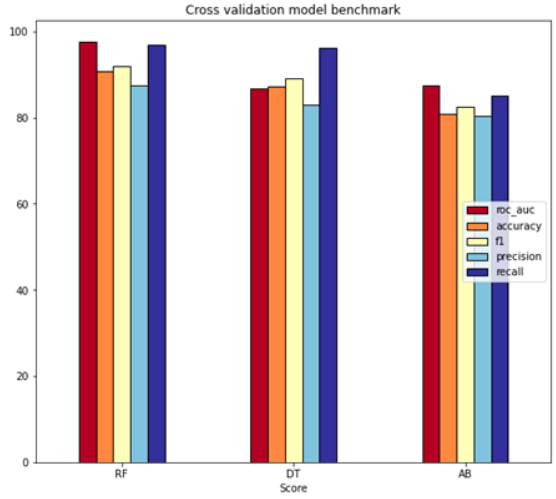

Fig. 3 Performance Analysis Using All Features

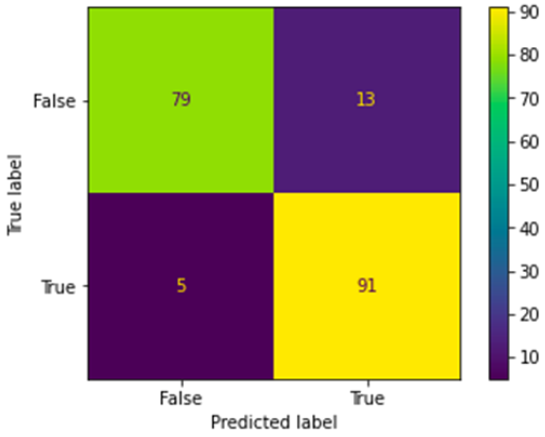

Fig. 4 Confusion Matrix

TABLE III. CLASSIFICATION PERFORMANCE USING ALL FEATURES

Evaluation Measure	Random Forest Classifier	Decision Tree Classifier	AdaBoost Classifier
Accuracy	90.4%	85.1%	82.4%
F1-Score	0.91	0.86	0.83
Precision	0.875	0.79	0.79
Recall	0.94	0.95	0.88
ROC AUC	0.90	0.84	0.82

V. CONCLUSION

Based on the findings of the experiment, it can be inferred that the Random Forest algorithm performs better than other algorithms in terms of accuracy. To prevent diabetes, individuals should strive to maintain consistent glucose levels, prioritize their mental and physical health, and consume a balanced diet to regulate their insulin levels. Those with a family history of diabetes should take extra precautions. Our Prediction model was able to achieve an accuracy rate of 90.5% using Random Forest Algorithm.

REFERENCE

[1] Kumar, Tiwari, Pande, “Diabetes prediction using machine learning tools”, “2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)”, 27 May 2022.

[2] Elif Nur Haner Kırğıl, Begüm Erkal, Tülin Erçelebi Ayyıldız, “Predicting Diabetes Using Machine Learning Techniques”, “2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE)”, 13 January 2023.

[3] Khilwani, Gondaliya, Patel, Hemnani Gandhi, Kumar Bhart, “Diabetes Prediction, using Stacking Classifier”, “2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)”, 10 January 2022.

[4] Kumari Shreya, Krishna Prathibha, “Machine Learning based Diabetes Detection”, “2021 6th International Conference on Communication and Electronics Systems (ICCES)”, 02 August 2021.

[5] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," in IEEE Access, vol. 7, pp. 102232-102238, 2019.

[6] Deepti Sisodia, Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Computer Science, vol. 132, pp. 1578-158, 2018.

[7] Abdullah Caliskan, Mehmet Emin Yuksel, Hasan Badem, Alper Rasturk, "Performance improvement of deep neural network classifiers by a simple training strategy," Engineering Applications of Artificial Intelligence, vol. 67, pp. 14-23, 2018.