Comparative Study of Machine Learning Models for Diabetes Prediction

Amit Kumar Das
Dept. of ECE
Institute of Engineering &Management,
University of Engineering & Management
Kolkata, India

Saptarshee Basak
Dept. of ECE
Institute of Engineering &Management,
University of Engineering & Management
Kolkata, India

Shounak Choudhury
Dept. of ECE
Institute of Engineering &Management
University of Engineering & Management
Kolkata, India

2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech) | 979-8-3315-2076-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/IEMENTECH65115.2025.10959680

***Abstract* - Diabetes is one of the most widespread and critical health issues globally, attributed to factors such as obesity, sedentary lifestyles, and elevated blood sugar levels. These factors disrupt insulin function, resulting in metabolic irregularities and increased glucose concentrations in the bloodstream. This research aims to address the rising prevalence of diabetes by leveraging machine learning techniques to predict its onset. By providing early warnings, this approach seeks to motivate individuals to adopt healthier dietary and lifestyle choices, potentially reducing the risk of developing diabetes. The performance of several machine learning algorithms, such as classification and ensemble learning techniques, is investigated in this paper. K-Nearest Neighbours (KNN), Random Forest, Decision Tree, Support Vector Machine (SVM), Logistic Regression, XGBoost algorithm, Neural Network, and Ensemble learning voting classifier are some of the important methods that were assessed. Complementary preprocessing techniques, such as Label Encoding, Feature Engineering, replacing missing values, train-test split methods, are used to prepare and analyze the data. These methodologies are applied to large healthcare datasets, enabling the extraction of significant features, validation of results, and the generation of accurate predictions. The primary purpose of the research is to examine the effectiveness of different models to discover the most robust and reliable technique for predicting diabetes.The focus lies on developing models that can effectively differentiate between diabetic and non-diabetic individuals while achieving high validation scores. By uncovering hidden patterns and trends in the data, the study provides valuable insights for medical professionals, enhancing their ability to make early diagnoses and informed decisions.**

Keywords— Random Forest, Ensemble learning algorithms, machine learning, data visualization, data interpretation, logistic regression, Support Vector Machine, K-Nearest Neighbors, Neural Networks

I INTRODUCTION

The medical field has rapidly adopted machine learning as a transformative technology for improving decision-making and prediction. By analyzing medical datasets and generating insightful forecasts, machine learning enables the prevention of diseases and the identification of effective treatments and preventive measures. Among the many applications of machine learning in healthcare, this research focuses specifically on its use in predicting diabetes.

Diabetes is one of the fastest-growing global health challenges, requiring constant management to prevent severe complications. Machine learning algorithms offer promising solutions for early disease prediction, assisting in better diagnosis and timely intervention. This study explores various machine learning algorithms and their classifications, emphasizing their ability to support decision-making and enhance predictive accuracy in medical contexts. Using advanced analytics, medical professionals can leverage machine learning to help individuals take proactive measures, ultimately reducing the risk of illness. This article the application of machine learning in medicine, specifically in the prediction of diabetes. Diabetes is a condition caused by an inability of the body to produce sufficient insulin, causing blood glucose levels to rise. Serious consequences include organ damage, cardiovascular disease, and disturbances to several physiological systems may arise if it is not controlled. Diabetes is a major worldwide health issue. WHO reports that diabetes and cardiovascular-related illnesses contribute to millions of premature deaths annually, underscoring the need for improved management and early intervention strategies. Factors contributing to diabetes include aging, obesity, unintentional weight loss, and overeating, irritability, and muscle stiffness. Symptoms such as delayed wound healing, partial paralysis, irritation, itching, and blurred vision are often early indicators of the condition. To address these challenges, this research proposes a predictive model that combines diabetes-specific external factors with common symptoms to enhance diabetes classification.

Early prediction models powered by machine learning can can be extremely important in mitigating this growing burden, enabling better self-care and improved healthcare outcomes.

II METHODOLOGY

1) Data Acquisition: Data acquisition is a fundamental process in modern computational research, The concept of "data acquisition" integrates two distinct components: "data," referring to structured or unstructured facts and statistics that serve as the foundation for research, and "acquisition," which denotes the deliberate act of collecting

979-8-3315-2076-2/25/$31.00 ©□2025 IEEE

information for a defined purpose. Within the scope of this study, data acquisition encompasses the identification, gathering, cleansing, and preparation of datasets to ensure their readiness for analysis. In this research, the Pima Indian Diabetes dataset has been utilized as a representative source of critical medical information. This dataset provides valuable insights into the factors contributing to diabetes, emphasizing the importance of structured and reliable data in healthcare analytics. The majority of a data scientist's workflow is dedicated to the acquisition, cleaning, and exploration of datasets, as the efficacy of machine learning models hinges on the input data's accuracy, completeness, and quality.

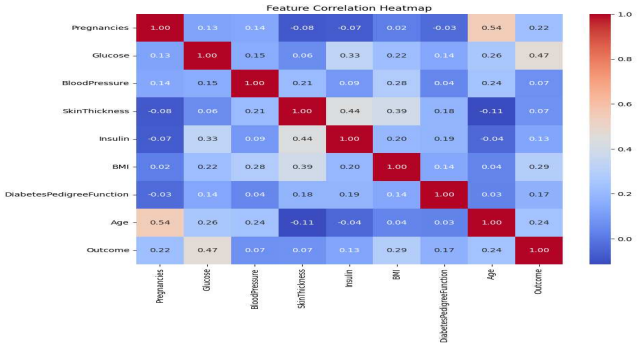

Fig 1: Feature Correlation Heatmap

2) Pre-Processing: Effective data preparation is a cornerstone of any successful machine learning project, particularly when working with a dataset as critical as the Pima Indian Diabetes dataset. This dataset, which contains vital medical information aimed at understanding the onset of diabetes, requires rigorous preprocessing to ensure accurate and reliable results. In the case of the Pima Indian Diabetes dataset, attributes such as glucose levels, blood pressure, and BMI occasionally contain zeros, which are physiologically improbable and could lead to model inaccuracies. To address this, we replaced these zeros with the median values of the respective columns, preserving the dataset's statistical integrity. Next, normalization was used to scale the information between 0 and 1. This step guarantees that all features contribute equally to the model's learning process, preventing bias towards attributes with greater numerical ranges. We also used Feature Engineering techniques on the dataset during pre-processing.Then, in order to properly assess model performance, the dataset was separated into training and testing subsets. An 80:20 split ratio was chosen, ensuring that the model has sufficient data for learning while reserving a portion for unbiased validation. Additionally, feature scaling was implemented to standardize all variables, through this meticulous preprocessing pipeline, the Pima Indian Diabetes dataset was transformed into a high-quality, model-ready form, enabling the development of robust predictive models.

3) Classification: The train-test split is a commonly used technique for assessing the performance of a machine learning model. This technique is applicable to various supervised learning tasks, including both regression and classification. The process involves dividing the dataset into two serate subsets: a test set and a training set. The training set provides the data required to identify patterns and relationships, and it forms the basis for the model's construction. However, the test set is utilized only for validation, enabling the model to generate predictions on data that has not yet been seen and compare these predictions to the actual target values. In this project, the train-test split was applied to the Pima Indian Diabetes dataset to ensure the reliability of the machine learning model. Additionally, the Min-Max Scaler was employed to normalize the dataset, transforming feature values into a standardized range between 0 and 1. Normalization is particularly important in datasets like this, where features such as glucose levels and BMI can have varying numerical ranges. By scaling all features to the same range, any biases towards features with bigger magnitudes are avoided and the model's training process is made more effective.

3) Data Analysis: In this study, a detailed comparison of machine learning algorithms—K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression, Decision Trees, and Support Vector Machines (SVM), Ensemble Learning (Voting classifier), XGBoost and Neural Network—was performed to predict the likelihood of diabetes. The dataset utilized contained medical parameters such as glucose concentration, blood pressure, BMI, age, and insulin levels. Data preprocessing comprised encoding category variables, standardizing numerical features, and managing missing information to guarantee uniformity across models. The models were evaluated using a wide range of parameters, such as the area under the ROC curve, F1-score, recall, accuracy, and precision (AUC-ROC), to capture both predictive performance and class imbalance handling. Through this analysis, the study provided insights about the strengths and weaknesses of each algorithm in terms of handling feature interactions, computational efficiency, and suitability for diabetes prediction in clinical datasets.

III IMPLIMENTATION

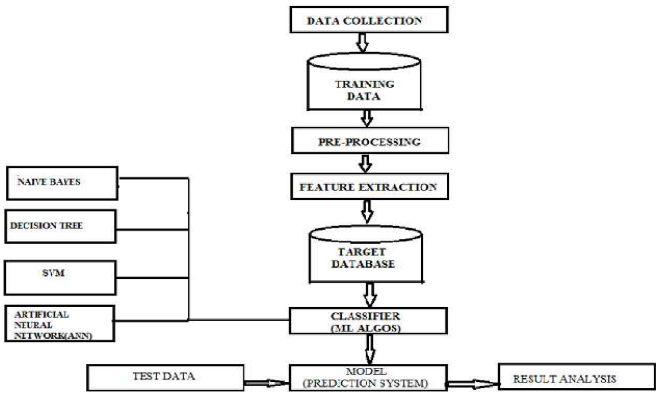

Fig 2: Data flow Diagram

MODELS USED

1. KNN (K-Nearest Neighbors) Classifier:

The K-Nearest Neighbors (KNN) algorithm is one of the simplest and most intuitive supervised learning methods utilized for jobs involving both regression and classification. The fundamental concept behind KNN is that similar data points are likely to be close to one another in feature space. uses the majority vote of its "K"- nearest neighbours in the feature space to determine a class label for a data item. The K value (the number of neighbours taken into account) is a crucial hyperparameter, as it directly influences the decision boundary and model performance.KNN evaluates the proximity of data points using a distance metric; the most widely used distance metrics are the Manhattan, Minkowski, and Euclidean distances. Once the distances are calculated, KNN simply selects the most frequent class from the nearest neighbors and allocates it to the newly created data point. When performing regression,KNN predicts the value of the target variable by taking the mean or median of the target values of the K nearest neighbors.

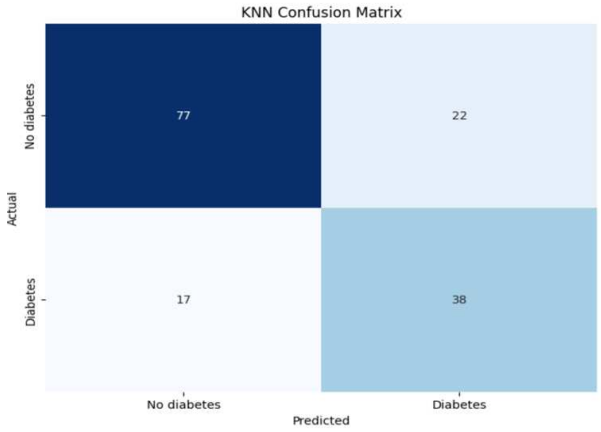

Fig 3: Classification Report and confusion matrix(KNN)

2. Random Forest:

Random Forest is a technique for group learning that combines the output of several decision trees to lower the chance of overfitting and enhance predictive performance. It is a type of bagging (bootstrap aggregating) technique, where a random portion of the data is used to train each tree, and at each split, a random feature is chosen. This randomness introduces diversity into the model, ensuring that there is little correlation between the different trees, thus improving the generalization of the model. During the training phase, multiple decision trees are constructed, and at prediction time, the results from all trees are aggregated. In classification tasks, the mode (most frequent class) is taken from the individual trees, while in regression tasks, the mean of the predicted values is used. Random Forest's strength resides in its ability to model complex relationships without requiring fine-tuning or feature engineering. One of Random Forest's main benefits is its ability to handle both numerical and categorical data, as well as its inherent robustness to noisy data and missing values. Furthermore, Random Forest provides an estimate of feature importance, helping identify the most significant variables.

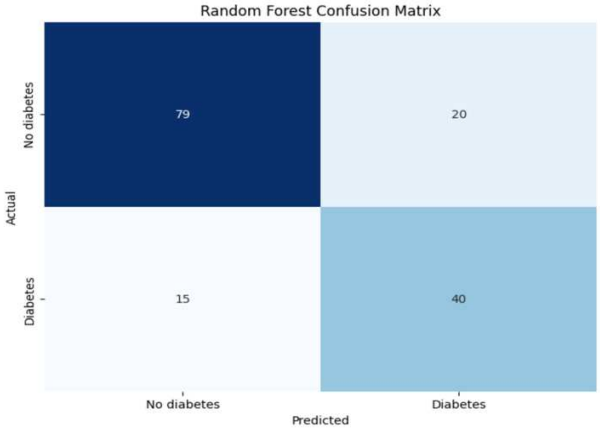

Fig 4: Classification Report and Confusion Matrix (RF)

3. SVM (Support Vector Machine):

Support Vector Machines (SVMs) are supervised learning algorithms used for tasks such as classification and regression. The core idea of this model is to identify a hyperplane that distinctly separates data points into different categories. The best hyperplane is defined as the one that maximizes the margin, which is the gap between the closest points of opposing classes, known as support vectors. This maximization helps the model achieve better generalization on unseen data. While SVM is particularly effective for datasets where the classes can be separated by a plane or a line. It can also handle more complex, non-linear data. This is achieved using kernel functions, which allow SVM to map data into higher dimensions efficiently without explicitly calculating those dimensions, thanks to the kernel trick.

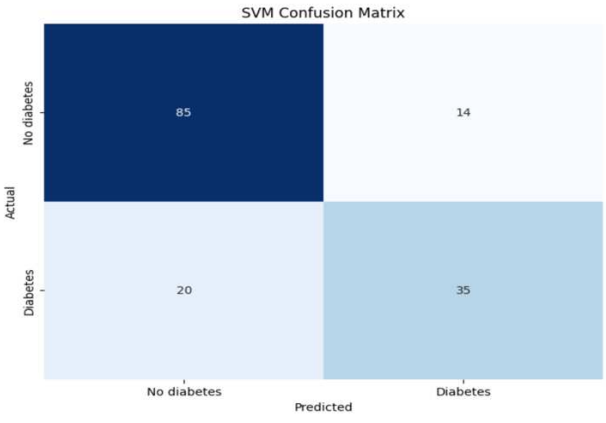

Fig 5: Classification Report and confusion matrix (SVM)

4. Decision Trees:

This is a type of non-parameterized, supervised learning algorithm. It is used for both regression and classification tasks. The model works by recursively splitting the data into subsets based on the most significant feature at each decision node, using criteria such as Gini impurity, information gain, or variance reduction. The process continues until the data split is sufficient or a stopping criterion (e.g., maximum depth or minimum samples per leaf) is met. Each decision node in the decision tree represents a feature, while the branches represent possible outcomes, and the leaves contain the final prediction. The structure of a decision tree allows it to be highly interpretable, as one can easily trace the decisions made by the model from input to output.

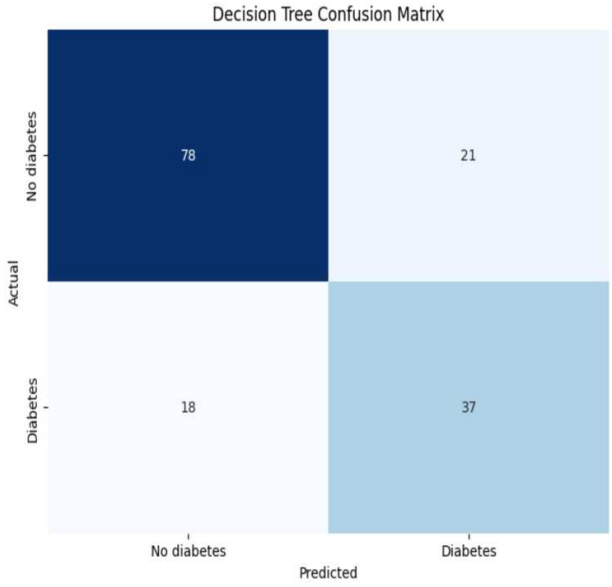

Fig 6: Classification Report and confusion matrix (DT)

5. Logistic Regression:

Logistic regression is a machine learning model or classifier used to plot the relationship between one dependent variable and one or more independent variables. Unlike linear regression, logistic regression is used for predicting categorical outcomes, particularly binary outcomes (e.g., success or failure, yes or no). It predicts the probability that a given input belongs to a particular class by the sigmoid function, which is also known as the logistic function, which transforms any number (real number) into a probability score between 1 and 0. This makes Logistic regression highly interpretable, as the output can be directly interpreted as the probability of class membership. The model operates by estimating the coefficients of the input variables through maximum likelihood estimation (MLE).

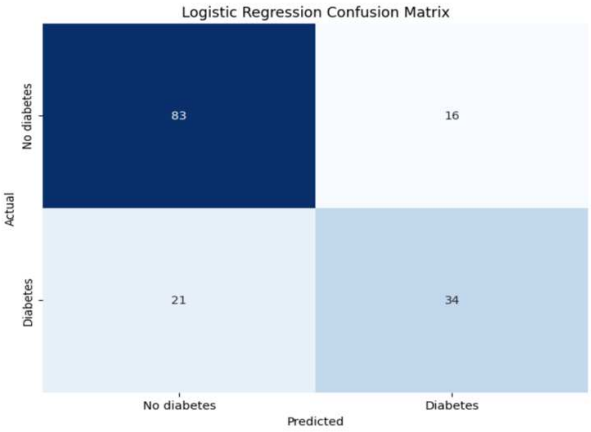

Fig 7: Classification Report and confusion matrix (LR)

6.Ensemble Learning Voting Classifier:

Ensemble Learning Algorithm is a powerful paradigm in machine learning. It uses a combination of ML models to improve its accuracy and performance. The Voting Classifier is a type of ensemble method that aggregates the predictions of several models to make a final decision. The main idea of ensemble methods is that combining the outputs of multiple diverse models leads to better generalization compared to relying on a single model, as it reduces overfitting and bias. The Voting Classifier uses two primary techniques for aggregation: hard voting and soft voting. This approach is simple but effective, especially when the individual models in the ensemble are diverse and complementary. In soft voting, the individual classifiers output probabilities for each class, and the final output prediction is made by taking an average of these probabilities and selecting the class with highest average probability. The strength of the Voting Classifier lies in its ability to leverage the diversity of base models, which may include different algorithms or varying hyperparameters. This helps reduce model overfitting

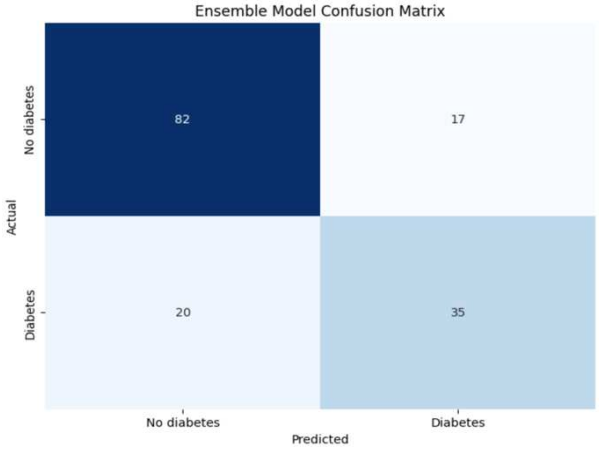

Fig 8:Classification Report and confusion matrix(Ensemble)

7. XGBoost:

XGBoost or Extreme gradient boosting is a highly efficient and scalable machine learning algorithm designed to optimize the performance of gradient boosting models. XGBoost builds an ensemble or collection of decision tress where the mistakes made by one decision tree is corrected by another. Gradient boosting itself is a technique where each new model is added minimizing the residual errors of the combined ensemble. XGBoost enhances this process by introducing several key innovations that improve its efficiency, speed, and generalization. Notably, XGBoost also uses regularization techniques, such as L1 (Lasso) and L2 (Ridge) regularization, to tackle and prevent overfitting by penalizing overly complex models. Additionally, tree pruning is employed to remove branches of the tree that have little contribution to the model, thus improving its generalization capability. XGBoost also utilizes column subsampling and row sampling, similar to techniques used in random forests, which further improve its robustness. Its parallelized execution and distributed processing capabilities make it suitable for handling large datasets and complex models.

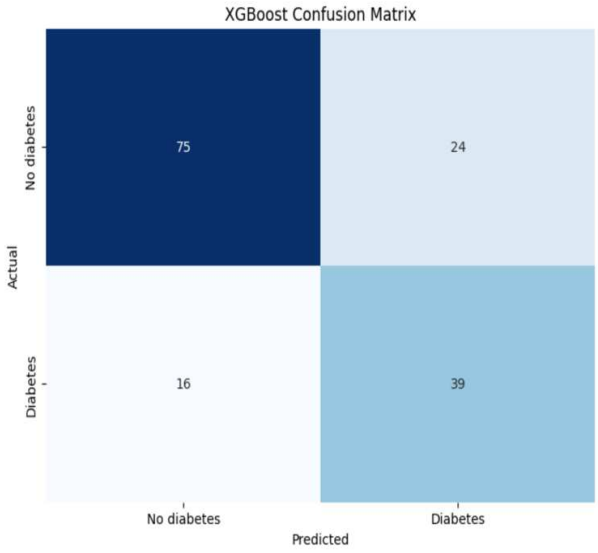

Fig 9: Classification Report and confusion matrix (XGBoost)

8. Neural Network:

Neural Networks are a class of ML- models that attempts to copy or mimic the information processing in biological neural systems. At its core, it consists of several layers of interconnected units known as neurons. Every neuron takes input data, applies a set of learned weights, and passes the result through a function called an activation function to get the desired output. These outputs are then passed to subsequent layers, eventually leading to the final prediction or classification. Raw data is received by the input layer, the hidden layers learn hierarchical representations of the data, the output layer generates the final prediction output. Neural Networks are especially powerful for tasks where the data has complex and non-linear relationships. The deep learning variant of neural networks, known as Deep Neural Networks (DNNs),. The flexibility of neural networks and their capacity to model highly complex relationships in large datasets have made them one of the most influential tools in modern machine learning, with applications across virtually every domain of AI research and real-world implementation.

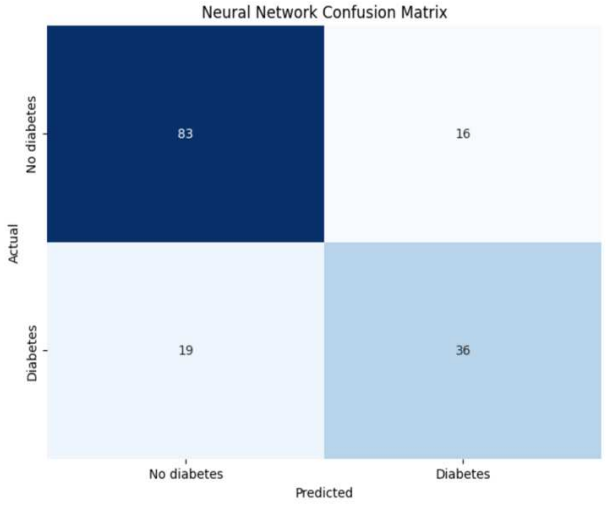

Fig 10: Classification Report and confusion matrix (NN)

IV RESULTS

Our work has carried out a thorough comparison of several machine learning models to predict diabetes, evaluating their efficacy with the use of important parameters AUC-ROC curves also known as receiver operating characteristics recall, F1-score, accuracy, and precision. The analysis revealed that the Random Forest and Ensemble models outperformed others, each attaining the highest of 0.84. Logistic Regression closely followed by an AUC of 0.83, demonstrating its capability as a lightweight yet effective predictive tool. Notably, Random Forest also had the highest accuracy (77.27%), indicating its superior ability to correctly classify diabetes cases. Logistic Regression maintained a solid accuracy of 75.97%, reflecting its reliability. Decision Tree, however, performed the least well in terms of prediction with an AUC of 0.73, highlighting its limited suitability for this problem compared to other methods. Thus, through this project we can conclude that Random Forest and Ensemble Learning models were clearly better for the given use case as compared to the other models. This study thus fulfills its goal to compare various types provide of such machine learning models and provide a clear standpoint as to what models should be used and developed further to make diabetes prediction using Machine Learning even more accurate and useful in a real-world scenario.

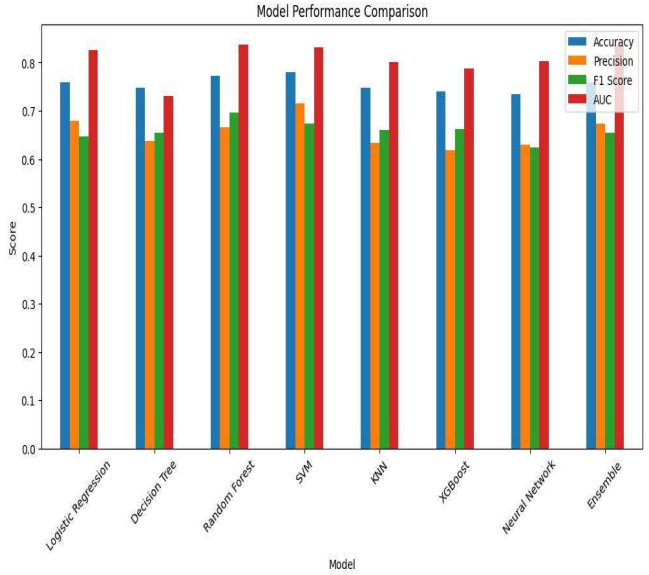

Fig 11: COMPARISON BAR GRAPH

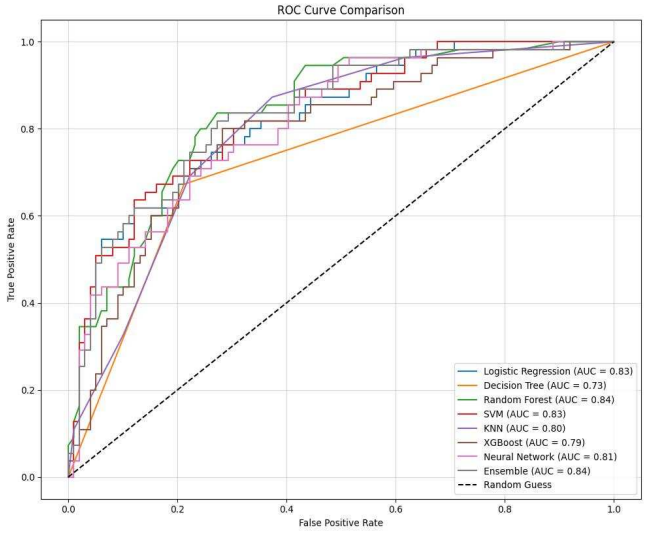

Fig 12: Comparison ROC Curve

CONCLUSIONS

This study highlights the importance of selecting appropriate machine learning models for clinical prediction tasks such as diabetes diagnosis. Random Forest and the Ensemble model emerged as top performers, combining strong predictive accuracy with robustness in distinguishing between diabetic and non-diabetic cases. Their high AUC scores and accuracy suggest they are well-suited for use in real-world clinical applications where reliability and precision are critical. While Logistic Regression demonstrated competitive results, making it a simpler and interpretable alternative, models like Decision Tree fell short in comparison, indicating their limitations in handling the complexity of diabetes prediction. These results emphasize the necessity of leveraging advanced algorithms, such as Random Forest or ensemble techniques, to ensure reliable and accurate predictions in healthcare scenarios. Future work could explore optimizing these models further or combining them with domain-specific features to enhance predictive capabilities even more.

Model	Accuracy	Precision	F1 Score	AUC
Logistic Regression	0.7597	0.68	0.6476	0.8251
Decision Tree	0.7467	0.6379	0.6548	0.7303
Random Forest	0.7727	0.6666	0.6956	0.8374
SVM	0.7792	0.7142	0.6730	0.8312
KNN	0.7467	0.6333	0.6608	0.8
XGBoost	0.7402	0.6190	0.6610	0.7882
Neural Network	0.7337	0.6296	0.6238	0.8018

TABLE 1: MODEL COMPARISON TABLE

REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[4] Wu,S.Yang,Z.Huang,J.He,andX.Wang,"Type2diabetesmellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018.

[5] S. Islam Ayon and M. Milon Islam, "Diabetes prediction: A deep learning approach," International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 21–27, 2019.

[6] H. Naz and S. Ahuja, "Deep Learning Approach for diabetes prediction using Pima Indian Dataset," Journal of Diabetes; Metabolic Disorders, vol. 19, no. 1, pp. 391–403, 2020.

[7] D.SisodiaandD.S.Sisodia,"Predictionofdiabetesusingclassification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[8] M. Komi, Jun Li, Yongxin Zhai, and Xianguo Zhang, "Application of data mining methods in diabetes prediction," 2017 2nd International .

[9] M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis," *Expert systems with applications*, vol. 38, no. 12, pp. 14 650–14 659 .

[10] World Health Organization, "Diabetes," 16 September 2022. [Online].

[11] M. Komi, Jun Li, Yongxin Zhai, and Xianguo Zhang, "Application of data mining methods in diabetes prediction," 2017 2nd International .