Predictive Modelling for Diabetes: Assessing the Efficacy of Ensemble and Standard Machine Learning Algorithms

Prof. Ashish Talekar
Dept. of Artificial Intelligence
G.H Raisoni College of Engineering
Nagpur, India

Ishita Savale
Dept. of Artificial Intelligence
G.H Raisoni College of Engineering
Nagpur, India

Shrishti Singh
Dept. of Artificial Intelligence
G.H Raisoni College of Engineering
Nagpur, India

Siddhi Jain Khilosiya
Dept. of Artificial Intelligence
G.H Raisoni College of Engineering
Nagpur, India

Ashik Shambharkar
Dept. of Artificial Intelligence
G.H Raisoni College of Engineering
Nagpur, India

Aasirbad Biswal
Dept. of Artificial Intelligence
G.H Raisoni College of Engineering
Nagpur, India

***Abstract*— Diabetes is a chronic condition with increasing global prevalence, necessitating early and precise prediction for effective management. Various machine learning (ML) models, including Support Vector Machines, Logistic Regression, Decision Trees and ensemble techniques like Gradient Boosting, Random Forest, XGBoost and LightGBM, were assessed in this study for their ability to predict diabetes. The dataset contains important clinical parameters such as diabetes pedigree function, blood pressure, insulin level, BMI, age, number of pregnancies, glucose level, and insulin level. Preprocessing techniques such as feature scaling and missing-value imputation are utilized to enhance the model's predictive strength. The experimental results show that ensemble models outperform standard methods, with Logistic Regression at 75% accuracy, Decision Trees at 86%, Random Forest at 90%, Gradient Boosting at 88%, and XGBoost and LightGBM at 90.5% and 93%, respectively. These results highlight how ML-driven predictive models can aid in early intervention and clinical decision-making. To improve generalizability and practical usefulness, future studies should look into bigger, more varied datasets and hybrid modelling approaches**.

***Keywords:* Supervised tree model, Random Forest, Gradient Boosting, Diabetes Prediction, Logistic Regression, Machine Learning, Ensemble Techniques, Support Vector Machine, XGBoost.**

I. INTRODUCTION

Diabetes, a long-term health issue when the body doesn't properly process blood sugar, resulting in it staying too high. It is brought on by either insufficient pancreatic synthesis of insulin or a cellular responsiveness to insulin. After consuming food, glucose, the primary energy source for cells, enters the bloodstream, and insulin is the hormone that aids its absorption into cells. When insulin levels are insufficient or cells become resistant, glucose accumulates in the bloodstream, resulting in diabetes. The global adult population with diabetes has grown significantly in recent thirty years, from over 200M in 1990 to over 830M in 2022, it is expected to surpass 1.3 billion by the middle of the century. This increase is particularly noticeable in countries with low and moderate incomes, where access to treatment is still limited. According to projections, there may be 1.3 billion people with diabetes worldwide by 2050. In 2021, diabetes was directly responsible for 1.6 million deaths, while renal disease brought on by diabetes was responsible for an additional 530,000 deaths. Furthermore, complications from diabetes, like conditions affecting the heart also apoplexy, account for around 11% of all cardiovascular deaths worldwide. About 59% among individuals with diabetes in less affluent nations do not receive enough medication, despite this concerning increase. Three distinct types of diabetes are recognized: IDDM (type 1), NIDDM (type 2), and diabetes during pregnancy(gestational) diabetes. Insulin-dependent diabetes mellitus (IDDM) usually affects children and teenagers which is initiated due to the body's defence system targeting insulin-secreting biological units. Non-insulin-dependent diabetes mellitus (NIDDM), the most common type, is caused by the body's cells becoming the organism's biological units, eventually decreasing pancreatic function. Gestational diabetes occurs during pregnancy as a result of hormonal changes that interfere with insulin function, raising the risk of problems for both mother and child.

Early diabetes detection is critical for avoiding serious complications like cardiovascular disease, amputations, and kidney failure. Diabetes is frequently diagnosed with diagnostic procedures. With technological improvements, ML models have developed as a significant tool in early detection plus management. When applied to clinical data, ML algorithms can assist in predicting the likelihood of diabetes onset, allowing for prompt intervention and improved patient outcomes. In this study, we investigate the use of machine learning (ML) approaches to predict diabetes risk based on clinical indicators. To find the best accurate prediction model, we test a variety of ML models. The goal is to examine the difficulties and constraints of these models, as well as suggestions for further study and advancements, while highlighting the potential of ML-driven analytics in enhancing diabetes care.

II. LITERATURE REVIEW

Al Gharabawi et al. [1] used Random Forest for Identifying key attributes as well as triple-ANN for classification on a 13-feature Kaggle dataset. Their method's 98.73% accuracy highlights the value of integrating deep learning and ensemble techniques for early diabetes identification. Riveros Perez et al. [2] evaluated five + and an AUC of 0.8168, XGBoost outperformed SVM, which had a higher sensitivity. This brought to light the trade-offs in performance amongst classifiers for screening at the population level. Faraz et al. [3] evaluated linear and non-linear SVMs against Random Forest on the Pima dataset. The nonlinear SVM's accuracy of about 95% indicates that kernel-based models are capable of handling intricate, non-linear correlations in health data. Refat et al. [4] evaluated various ML and DL models. Using a 17-feature UCI dataset. With a test accuracy of about 99%, XGBoost performed better than any other, demonstrating the resilience of boosting algorithms. Sneha et al. [5] used to optimize attributes to boost prediction power for several classifiers on a diabetic data. Based on their findings, RF provided the most specificity and Naïve Bayes the highest total accuracy. The study emphasized how informed feature reduction enhances diagnostic reliability and precision across models. Soni et al. [6] Conducted a contrastive study of six predictive models, including Random Forest, SVM, KNN, and Logistic Regression, using the Pima Indian dataset. With an accuracy of almost 91%, Random Forest was the best performer among them. The outcomes highlighted the ensemble approaches' resilience and dependability in clinical prediction tasks. Febriana [7] contrasted the efficiency of Naïve Bayes and KNN on the Pima dataset using confusion matrix measures. When compared to KNN, Naïve Bayes showed a higher recall, suggesting a greater capacity to detect real cases of diabetes. Because of this, it was a better option for medical applications where reducing false negatives is essential.

III. METHODOLOGY

A. Dataset Description

This research utilizes the Pima Indians Diabetes Dataset, Collected from the UCI ML Archive, a widely recognized source for benchmarking in diabetes-related machine learning research. The dataset is frequently cited in the literature due to its inclusion of a comprehensive set of clinical and demographic attributes relevant to Type 2 diabetes. It comprises 768 records, each representing a female patient of Pima Indian descent, thereby offering a consistent and standardized foundation for evaluating predictive models in diabetes diagnosis and risk assessment.

Feature Descriptions:

The dataset consists of eight predictive variables and one target variable (Outcome). A detailed description of each feature is provided below:

- Pregnancies: How many times a woman has been pregnant.
- Glucose: Commonly referred to as blood sugar.
- Blood Pressure: The pressure of blood in the body's arteries when the heart is resting, measured in mmHg.
- Skin Thickness: How thick the skin is on the back of the arm — gives clues about body fat and health.
- Insulin: The level of insulin in the blood two hours after eating sugar, showing how well the body controls blood sugar.
- BMI: A value that indicates whether a person's weight is appropriate for their height.
- Diabetes Pedigree Function (DPF): A score that shows how much diabetes runs in a person's family.
- Age: Patient's age in years
- Outcome: An outcome of 0 signifies absence of diabetes, whereas 1 signifies the presence of diabetes.

B. Data Preparation

IT is vital for improving the accuracy and reliability of ML models, especially when used with healthcare datasets like the Pima Indians Diabetes Dataset. These datasets frequently exhibit issues such as missing values, outliers, and inconsistencies, all of which can adversely affect model performance. To mitigate these challenges and ensure high data quality, a series of preprocessing steps were undertaken. These included data cleaning procedures to address missing and anomalous entries, as well as feature scaling techniques to standardize variable distributions, thereby optimizing the dataset for effective model training.

1. Dataset Balancing

The original Pima Indians Diabetes Dataset exhibited a class imbalance, characterized by a disproportionately higher number of non-diabetic (Class 0) instances compared to diabetic (Class 1) cases. Such imbalance poses a significant challenge in supervised learning, often resulting in biased models that favour the majority class, thereby compromising the detection of minority class instances. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was used. In order to improve class representation without merely reproducing instances, IT forms new minority class samples by creating interpolations between current examples. By increasing the dataset size from 768 to 1,000 observations, this method improved the model's generalization across both diabetic and non-diabetic categories and achieved a more balanced class distribution.

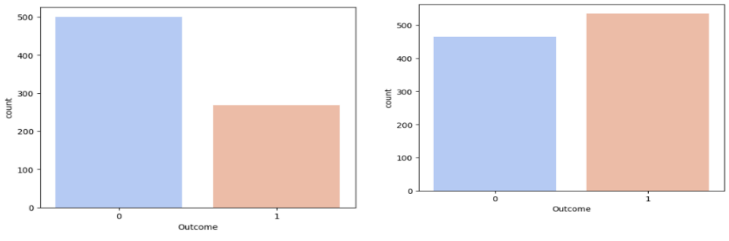

Figure 1: Class distribution before and after SMOTE.

2. Data Cleaning

Data cleaning constitutes a fundamental preprocessing step in developing reliable and accurate machine learning models for diabetes prediction. Raw healthcare datasets, such as those used in this study, frequently contain issues including missing values, duplicate records, inconsistencies, and outliers. If left unaddressed, these anomalies can significantly degrade model performance, leading to biased predictions and reduced generalizability. Prior to model training, comprehensive data cleaning was performed to remove noise and improve the dataset's quality and integrity.

a. Managing Absent Data:

Incomplete patient records or errors in data collection might result in missing values in diabetes datasets. Different imputation methods can be used according to the type of missing data:

- Mean/Median Imputation: Missing values in numerical attributes, such as glucose level or BMI, might be substituted by the mean or median of the feature.
- Mode Imputation: Missing values for categorical variables, including smoking status or gender, can be filled in using the most prevalent category.
- Predictive Imputation: ML methods like KNN and regression models can forecast missing values by analysing existing feature associations

b. Removing Duplicate and Inconsistent Entries:

Duplicate records can distort the dataset distribution and lead to biased predictions. Identifying and removing redundant patient records ensure data integrity. Additionally, inconsistencies such as incorrect age values (e.g., negative values or unrealistic ages) or contradictory entries (e.g., non-diabetic individuals with high fasting blood glucose levels) are corrected to improve dataset quality.

c. Outlier detection:

Outlier detection was essential to improve model accuracy and prevent extreme values from skewing predictions. Outliers were detected in key numerical attributes such as glucose, blood pressure, BMI, skin thickness, insulin, and age within the Pima Indians Diabetes Dataset.

3. Feature Scaling and Normalization

Feature scaling was used to make sure every feature contributed uniformly to the model and to speed up the optimization process.

- Min-Max Scaling: This method was applied to features such as Glucose, Insulin, BMI, and Age, transforming their values into a standardized range of [0,1].
- Z-score Standardization: Standardization was used to achieve a mean of 0 and a standard deviation of 1, features like blood pressure, skin thickness, and diabetes pedigree function were standardized.

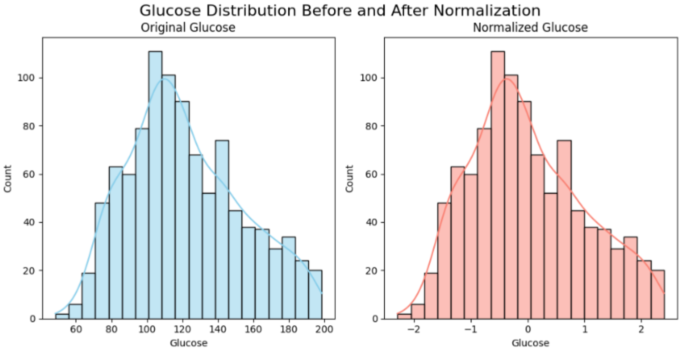

Figure No.2: Comparison of feature value before and after normalization.

C. Machine Learning Models

1. Logistic Regression (LR)

It is a statistical approach for forecasting two valued outcomes by evaluating the likelihood of a given class. It uses the sigmoid function to transform linear combinations of input information into probability values ranging from 0 to 1. A fundamental model for diabetes prediction, LR is frequently employed in medical research due to its ease of use and interpretability. The logistic regression model can be mathematically represented as follows:

$$\text{logit}(Y) = \ln\left(\frac{p}{1-p}\right) = a + bX$$

where b represents the regression coefficient.

2. K-Nearest Neighbors (KNN)

It is a method that classifies a new sample based on the most common class among its k-nearest neighbors. It is particularly helpful when there is little prior knowledge about the data distribution, and it is frequently used as a baseline model for diabetes classification.

3. Decision Tree (DT)

It is a supervised learning technique mostly used for categorization. It creates a tree-like structure by dividing data according to feature values, with each node standing for a decision rule. Decision trees are useful for determining important diabetes factors and are simple to understand.

4. Random Forest (RF)

It is a combination method technique that combines numerous decision trees. Each tree is trained on a randomly selected subset of the data, and the final prediction is made by majority voting. RF is a reliable option for diabetes classification since it is very resistant to overfitting and offers good accuracy without requiring a lot of hyperparameter modification.

5. AdaBoost (ADB)

It is a combination method technique that iteratively modifies the weights of feeble classifiers to improve their performance. In each iteration, it gives misclassified examples more weights, which helps the model concentrate on more difficult cases. By lowering bias and variance, ADB improves diabetes prediction and is frequently used with decision trees as base learners.

6. Gradient Boosting Classifier (GBC)

It is a collective strategy algorithm that builds decision trees in sequence, with each new tree fixing the faults of the ones before it. It is a strong option for diabetes classification since it successfully reduces lost functions. Compared to XGBoost, GBC is slower, but it has better predictive capabilities.

7. Support Vector Machine (SVM)

It is a potent classification technique that determines the best hyperplane for separating distinct classes. It can handle nonlinear interactions with kernel functions and is especially efficient in high-dimensional areas. SVM is frequently used to analyse medical data, including the prediction of diabetes.

8. XGBoost (XGB)

It is an optimized version of the Gradient Boosted Trees technique that is intended for excellent performance and scalability. It is well known for managing massive datasets with millions of samples and effectively builds decision trees in parallel under distributed conditions. Because of its capacity to capture intricate feature interactions, XGB is frequently utilized in diabetes prediction and uses regularization strategies to avoid overfitting.

9. LightGBM (LGB)

It is a gradient boosting framework engineered for enhanced speed and efficiency. In contrast to XGBoost, it employs a histogram-based methodology, facilitating faster computations and reduced memory consumption. It constructs trees in a leaf-wise manner rather than level-wise, rendering it particularly effective for handling large datasets.LGB is renowned for its high accuracy and is frequently utilized in medical diagnostics, including the prediction of diabetes.

D. Performance Evaluation Metrics

Different models are compared using the following evaluation metrics:

1. Accuracy:

The correctness of a model’s predictions compared to actual outcomes, measured as the ratio of correct prediction to total cases.

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

2. Precision:

It is the percentage of anticipated positive cases that were actually positive. It is particularly crucial in situations when false positives have serious repercussions.

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

3. Recall:

This metric assesses how well the model can recognize positive examples among all real positive examples.

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

4. F1-score:

A metric that combines precision and recall, calculated as their harmonic mean. It is very handy when working with unbalanced datasets.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC-AUC:

A metric that evaluates a model's ability to distinguish between classes by measuring the area under the curve plotting true positive rate against false positive rate across various thresholds.

$$\text{FPR} = \frac{FP}{FP + TN}$$

Interpretation:
AUC = 1 indicates perfect class separation.
AUC = 0.5 indicates the model performs equivalently to random chance (no prediction power).
AUC < 0.5 indicates poor model performance (misclassification).

IV. IMPLEMENTATION & RESULTS

This research utilizes PIMA Indian Diabetes dataset with a comprehensive machine learning framework to predict diabetes. The dataset was split into 80% for training and 20% for testing. To ensure statistical reliability and reduce the risk of overfitting, a 5-fold cross-validation approach was employed to evaluate the model. Data pre-processing included the imputation of missing values, outlier detection, and feature normalization—crucial steps for enhancing data quality and ensuring consistent model performance. Furthermore, hyper parameter optimization was performed through both Grid Search and Random Search methodologies, particularly benefiting complex models like Random Forest, Gradient Boosting, and XGBoost. This tuning process significantly contributed to improving model accuracy and generalizability in predicting diabetes outcomes.

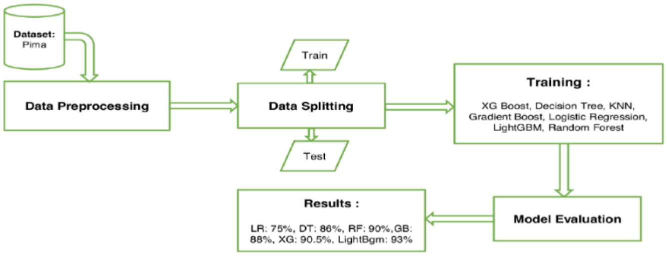
Dataset:
Pima
Data Preprocessing
Train
Data Splitting
Test
Training :
XG Boost, Decision Tree, KNN, Gradient Boost, Logistic Regression, LightGBM, Random Forest
Results :
LR: 75%, DT: 86%, RF: 90%,GB: 88%, XG: 90.5%, LightBgm: 93%
Model Evaluation

Figure No.3: Machine learning framework for diabetes Prediction

A. Model Performance Analysis

The comparative analysis of model accuracy following hyperparameter optimization indicates that LightGBM achieved the highest accuracy at 93%, with XGBoost (90.5%), Random Forest (90%), and KNN (90%) closely following. Traditional models such as SVM (77%) and Logistic Regression (75.5%) demonstrated lower performance, highlighting the superiority of ensemble methods. Decision Tree and AdaBoost exhibited moderate performance, achieving accuracies of 86% and 80.5%, respectively. These findings underscore the substantial enhancement in classification performance afforded by boosting algorithms compared to standalone classifiers.

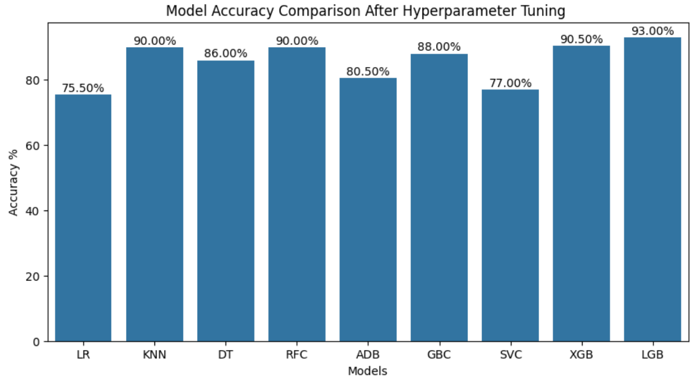

Figure No.4: Accuracy Result of Machine learning methods

B. ROC Curve and AUC Analysis

The ROC curve assesses model performance by comparing their AUC (Area Under the Curve) values. LBM has the best AUC (0.98), followed by XGB and RF (0.97), demonstrating its superior prediction skills. GBC and KNN performed well (0.96), but traditional models like LR, SVC, and Decision Tree had the lowest AUC (~0.86), demonstrating poor classification between diabetes and non-diabetic cases.

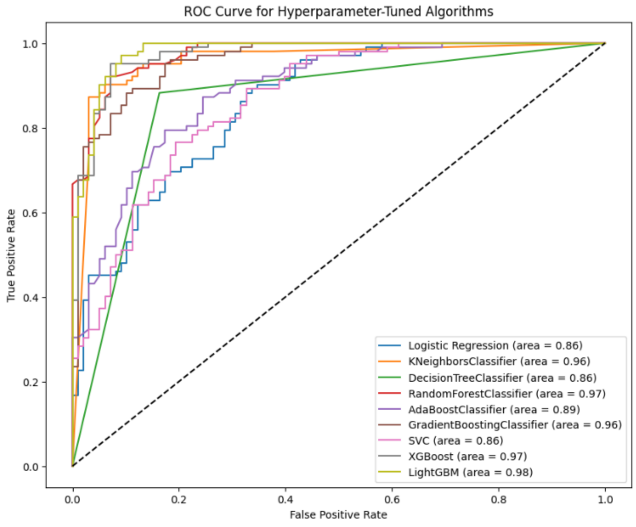

Figure No.5: ROC-AUC Curve For all algorithms

V. CONCLUSION

The primary goal of this research was to build and implement a diabetes prediction system utilizing machine learning techniques, as well as to evaluate its effectiveness. The proposed technique used a variety of classification algorithms, including SVM, KNN, Decision Tree, Logistic Regression, and Gradient Boosting, as well as advanced ensemble approaches like XGBoost and LightGBM, which were improved through intensive hyperparameter tuning. Experimental results showed that ensemble approaches outperformed standard classifiers, with XGBoost and LightGBM yielding strong performance. Notably, LightGBM achieved the greatest classification accuracy (93%). These findings can help healthcare providers make early predictions and educated decisions, ultimately improving diabetes treatment and patient outcomes. Future research could focus on improving the model by combining deep learning architectures with real-time health monitoring data. Furthermore, adding electronic health records (EHR) and distributing the system as a web or mobile application may improve its accessibility and practical value. Exploring explainable AI (XAI) strategies may also improve prediction interpretability, leading to increased trust and adoption among medical professionals.

VI. REFERENCES

[1] Al-Gharabawi, F. W., & Abu-Naser, S. S. (2023). Machine learning-based diabetes prediction: Feature analysis and model assessment. International Journal of Academic Engineering Research,7(9),55–63.

[2] Riveros Perez, E., & Avella-Molano, B. (2025). Learning from the machine: Is diabetes in adults predicted by lifestyle variables? BMJ Open, 15(1), e067894.

[3] Faraz, S., & Singh, P. (2022). Diabetes prediction using machine learning. Journal of Applied Science and Education, 2(2), 1–12.

[4] Refat, M. A. R., Al Amin, M., Kaushal, C., Yeasmin, M. N., & Islam, M. K. (2021). A comparative analysis of early-stage diabetes prediction using machine learning and deep learning approach. IEEE TechRxiv.

[5] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology, 9(9), 921–925. Retrieved from

[6] Febriana, M. E., Ferdinana, F. X., Sendania, G. P., Suryanigruma, K. M., & Yunandaa, R. (2020). Diabetes prediction using supervised machine learning. Procedia Computer Science, 179, 443–450.

[7] Fatima, M., & Pasha, M. (2024). Predictive modelling for type 2 diabetes using hybrid machine learning techniques. Journal of Medical Systems, 48(1), 12

[8] Islam, S. M. S., et al. (2023). Explainable AI for diabetes prediction: A comparative study of ensemble methods. Computers in Biology and Medicine, 165, 107416.

[9] Ali, L., et al. (2023). A smart healthcare framework for diabetes prediction using machine learning techniques. IEEE Access, 11, 11023–11

[10] Yang, Z. (2025). Application of machine learning in diabetes prediction based on electronic health record data analysis. ITM Web of Conferences, 70, 04015.

[11] Sonia, J. J., Jayachandran, P., Md, A. Q., Mohan, S., Sivaraman, A.K., & Tee, K. F. (2023). Machine-learning-based diabetes mellitus risk prediction using multi-layer neural network no-prop algorithm. Diagnostics, 13(4), 723.

[12] Thakur, D., Gera, T., Bhardwaj, V., AlZubi, A. A., Ali, F., & Singh, J. (2023). An enhanced diabetes prediction amidst COVID-19 using ensemble models. Frontiers in Public Health, 11, 1331517.

[13] Patro, K. K., Allam, J. P., Sanapala, U., & others. (2023). An effective correlation-based data modelling framework for automatic diabetes prediction using machine and deep learning techniques. BMC Bioinformatics, 24, 372.

[14] Luo, J., Kumbara, A., Shomali, M., Han, R., Iyer, A., Agarwal, R., & Gao, G. (2024). Let curves speak: A continuous glucose monitors based large sensor foundation model for diabetes management. arXiv preprint arXiv:2412.09727.