Harnessing XGBoost and Other Techniques for enhanced diabetes diagnosis

Hrishit Madhavi
Department Of Computer Science and Engineering
Dr Vishwanath Karad MIT World Peace University, Pune, India
1032220164@mitwpu.edu.in

Jacob Cherian
Department Of Computer Science and Engineering
Dr. Vishwanath Karad MIT World Peace University, Pune, India
1032220333@mitwpu.edu.in

***Abstract*—In recent years, there has been a growing incidence of diabetes; hence, the necessity for the development of accurate and reliable risk prediction techniques is becoming increasingly important. Machine learning (ML) has been effective for the analysis of health-related data and prediction of patients at risk based on various factors, including age, lifestyle, and comorbid medical conditions. In the present study, we investigated the use of several machine learning techniques, including logistic regression, decision trees, random forest, and support vector machines (SVM), to predict diabetes risk in patients. Models were built from actual patient data in an attempt to uncover significant trends and risk factors for diabetes. We compared the performance of all algorithms to identify improvements in predictive accuracy and reduce the incidence of false positives. Our approach will yield a realistic and easy-to-interpret model for early detection of diabetes, allowing timely intervention and improved patient outcomes.**

Keywords—Diabetes prediction, Machine Learning, logistic regression, decision trees, random forest, Categorical Boost, Extreme Gradient Boost

I. Introduction

The hallmark of diabetes mellitus, one of the chronic illnesses that is developing the fastest in the globe is persistently elevated blood glucose levels. The disease's prevalence has emerged as a major public health concern, with millions of patients in all age groups and geographical locations. According to the World Health Organization (WHO), the global burden of disease is growing more rapidly each year, which increases healthcare costs and premature mortality. Severe side effects of diabetes, like heart failure, kidney disease, eye blindness, and neuropathy, can develop if the treatment is delayed.

Early detection of diabetes is important for avoiding long-term consequences. However, the traditional methods of diagnosis—blood test-dependent, detectable clinical symptoms, and patient history—typically diagnose the disorder when it is established. Though still the present gold standard, the traditional clinical methods are not likely to detect early symptoms, especially in prediabetes or asymptomatic individuals. In recent years, the failure to detect it earlier has prompted scientists and practitioners to look for technology-based methods.

The swift development of digital health infrastructure has enabled electronic recording and storage of patient data in large numbers. From wearable fitness trackers to electronic medical records, the information generated is a rich reservoir of data that, when subjected to rigorous analysis, can make disease detection possible earlier and with higher accuracy. To this end, machine learning (ML) has come into play. Since it can process quantities of advanced data that are massive, ML can find recurring patterns and their associations that are not necessarily immediately evident to human analysts.

Machine learning is about creating algorithms that can learn from past examples to predict new, unseen cases. In the case of diabetes, of utmost significance is that algorithms can be trained to recognize risk factors such as age group, mass, pressure of blood, genealogy, a person's way of living life and then apply these to make computations that a person will be having diabetes. Data-driven systems enable better comprehension of the patterns of health and the creation of tailored care interventions.

Researchers have compared different models that were developed to tackle the problem of diabetes. Models in supervised learning, such as logistic regression, decision trees, random forests, and support vector machines (SVM), have been effective when applied with labelled patient data. The target of these models is to predict whether an individual is diabetic or not diabetic from different input features. Also, unsupervised learning methods are being explored to uncover hidden patterns in unlabeled data with possible uses towards early detection and risk categorization.

Machine learning possesses a critical advantage in medicine: it improves and becomes more intelligent with experience over time with more data, becoming more accurate and less error-prone. Interpretability methods now allow clinicians to understand model predictions, becoming more confident in AI-powered devices. This allows an integration into the clinical routine. In this study we analyze different ML methods in predicting likelihood of diabetes to determine what type of algorithms are typically used, how these models are trained, and how they perform. The goal is to increase the diagnostic precision, reduce false predictions and enable timely interventions. With the continued development of ML, the potential returns for early detection and management of diseases are expected to grow substantially.

II. Literature Survey

Ali et al. [1] created dataset using the criteria of the American Diabetes Association; used 4900 samples were used for the training and 100 samples for testing. The results showed that the KNN achieved higher accuracy

Krati Saxena *et al.* [2] diagnosed diabetes mellitus using the K Nearest Neighbour Algorithm. For K = 3 and K = 5, accuracy and error rates were computed, and 70% and 69% accuracy were attained, respectively.

Huma Naz and Sachin Ahuja [3] used the PIMA Indian dataset. Deep learning was the presented model, resulting in an accuracy of 98.07%.

Kamrul Hasan et al. [4] used Multilayer Perceptron (MLP) for prediction and used various classifier algorithms like KNN, AdaBoost and XGBoost.The approach involved filling missing values, outlier removal and K-fold cross-validation. The research also proposed combining different models to improve prediction

Isfafuzzaman Tasin et al. [5] used PIMA dataset and Bangladeshi female patient dataset for prediction. The study used Decision Tree,KNN and other ensemble method. Among all of them XGBoost combined with ADASYN achieved highest accuracy of 81%.

Deepti Sisodia and Dilip Singh Sisodia [6]used three classification algorithm to detect diabetes which were decision tree, SVM and Naive Bayes. The study used PIMA Dataset and achieved highest accuracy of 76% on Navie Bayes

Mitushi Soni and Dr. Sunita Varma [7] The study used KNN,SVM,random forest and other algorithms for prediction of them Random forest achieved highest accuracy of 77%.

Sajida Perveen *et al.* [8] Using the standalone data mining algorithm J48 in conjunction with AdaBoost and bagging ensemble approaches that used the decision tree as a base learner to categorize patients with diabetes mellitus based on diabetes risk variables. In the Canadian Primary Care Sentinel Surveillance Network, this classification is carried across three distinct ordinal adult categories. The results of the evaluation showed that AdaBoost ensemble method performs better than alternative approaches.

Aishwarya Mujumdar and Dr. Vaidehi V [9] performed a comparative analysis of different models on PIMA Diabetes Dataset and a private dataset. The evaluation showed that the AdaBoost with application has accuracy of 98.8% on private dataset, while it has 77% accuracy on the PIMA dataset.

Rashmi Rane et al. [10] proposed a hybrid diabetes prediction model combining LightGBM and KNN in a voting classifier approach. They employed the GridSearchCV for the best parameter selection and validated the model on the Pima Indian Diabetes dataset. With suitable preprocessing and cross validation, their ensemble model yielded an accuracy of 90.1%, surpassing the individual models, i.e., KNN (86%) and LightGBM (88.7%). Their study proves the efficacy of using gradient boosting together with distance-based techniques for improved predictive diabetes diagnosis

III. Research Methodology

A. Model Architecture

We chose the Diabetes Dataset of Pima Indians for the analysis. This data is commonly used in the medical research community for predicting diabetes and has health information of Pima Indian women with an age of 21 years and more. This is best suited for this purpose because it has a broad range of medical features. The data set contains 768 instances, each consisting of eight input features and a binary output label representing diabetic (1) or non-diabetic (0) status. The features are the pregancies, Glucose level, pressure of blood, skinthickness, insulin, body mass index (BMI), diabetes pedigree function, and age.

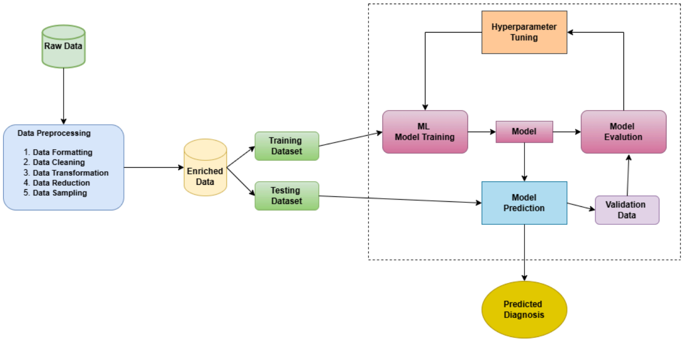

Fig. 1. Model Architecture

1] Raw Data: The pipeline begins with collecting raw data which is an essential step in model building.In this study the data used is Indian PIMA Diabetes Dataset that is widely used.

2] Data Preprocessing: The data which is collected is uncleaned meaning it has many null values,duplicate value in order to remove it we use data preprocessing techniques like data cleaning where missing values are filled,smoothing the noisy data, after cleaning, is then normalized and is reduced to eliminate redundant variables.

3] Enriched Data: After applying preprocessing techniques, the data is splitted into training and testing dataset

4] Applying ML Model: Various models were used, like k-nearest neighbours (KNN), decision tree and ensemble techniques like Extreme Gradient Boosting (XGBoost) and Categorical Boost (CatBoost) for making prediction

5] Hyperparameter Tuning: Hyperparameter adjustment was done to improve the machine learning model's output after it had been evaluated.Tuning involves earning rates, maximum depth, or the number of estimators using techniques like GridSearch.

6] Model Evaluation: Once the model is trained, it is evaluated using evaluation metrics.The evaluation process involves finding Accuracy, Recall, Precision, F1-Score.

7] Predicted Diagnosis: After successful evaluation if the model it is applied on the testing dataset.The final outcome of the pipeline is the predicted diagnosis to determine if a patient is diabetic or non-diabetic, providing a basis for early intervention and medical decision-making.

B. Machine Learning Models

1] K-Nearest Neighbours: The K-Nearest Neighbors algorithm has been applied to the Diabetes Dataset of Pima Indians due to its ease of application and effectiveness in classification. It operates by determining the analogy between a new instance and current training instances in terms of distance measures, typically Euclidean distance. For predicting diabetes, KNN classifies a test instance into a class label to identify if a patient is diabetic or non-diabetic based on the majority class of its 'k' nearest neighbors. Despite its ease of use, KNN is computationally costly and sensitive to feature scaling and 'k' choice, particularly for a modestly big dataset like PIDD.

2] Decision Tree: Decision Trees present an understandable and interpretable model to predict outcome based on the PIDD. They utilize a recursive method of dividing data based on feature thresholds that either optimize information gain or reduce Gini impurity. Decision rules are frequently formulated using glucose levels, BMI, and age. A notable advantage of Decision Trees in medical contexts is their ability to highlight which features contribute most significantly to the final prediction.They tend to overfit training data, but this requires mechanisms such as pruning or imposing a maximum depth.

3] XGBoost: XGBoost is a powerful ensemble method that has performed very well on structured data like PIDD. XGBoost incorporates regularization, shrinkage, and column sampling on top of standard gradient boosting to make it very robust against overfitting and efficient trainin In predicting diabetes, XGBoost incorporates several decision trees in a framework of boosting where each one learns to correct residuals of the preceding ones. Its integrated handling of missing values and parallelization feature renders it a likely candidate for practical medical prediction tasks.

4] CatBoost: It is the most suited for categorical and numerical variables in datasets, whereas PIDD primarily has numerical features. Its implementation of ordered boosting and internal management of categorical encoding add more stability and generalizability. With minimal preprocessing and parameter adjustment in diabetes prediction on PIDD, CatBoost provides comparable accuracy and efficiency. According to the principles of gradient boosting and having more interpretability, it turns into an applied tool in clinical diagnosis and research in different medical prediction contexts.

C. Model Evaluation

1] Accuracy: This metric evaluates the overall performance of a model

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

2] Precision: It tells how many true prediction where actually correct.

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

3] Recall: It identify True Positive measure

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

4] F1: It represent balance between precision and recall and uses harmonic mean

$$F\ Measure \;=\; 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

D. Data Visualization

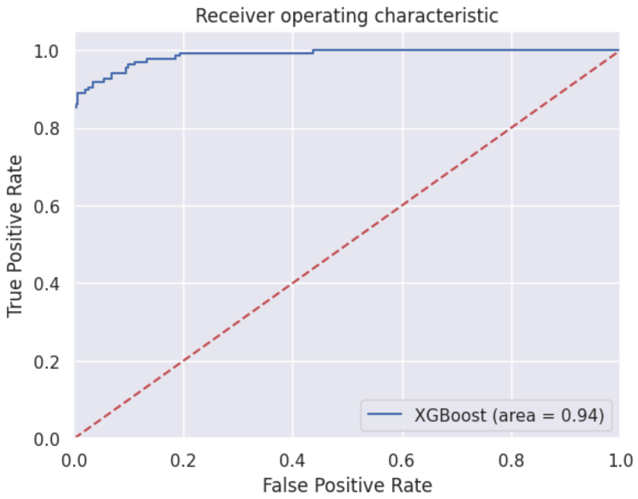

Fig. 2. ROC (Receiver Operating Characteristic) for XGBoost

In this research, the XGBoost classifier's performance was assessed for diabetes prediction on the Indian Pima dataset.The main metric used for evaluation was the Receiver Operating Characteristic (ROC) curve, which depicts how sensitivity and the false positive rate change with distinct threshold settings. The Area under the curve (AUC) serves as a numerical indicator of the classifier's ability to recognize patient with diabetes and not having diabetes. The XGBoost model reported an AUC of 0.94, which reflects good discriminative capability. This high value indicates that the model has good capability in identifying positive cases with few false positives and can be a potential candidate for use in medical diagnostic purposes.

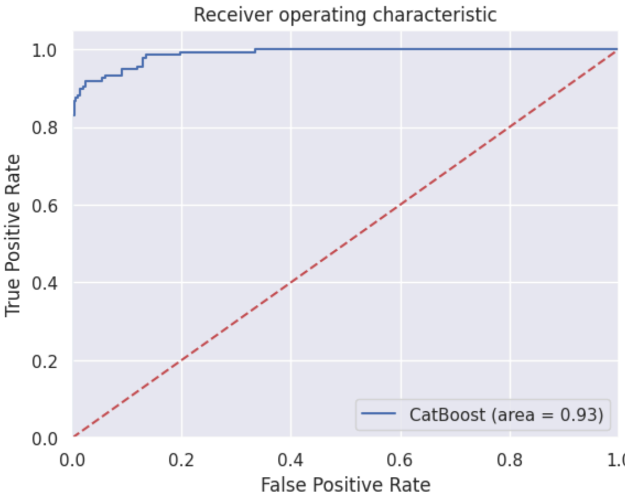

Fig. 3. ROC (Receiver Operating Characteristic) for CatBoost

The CatBoost classifier was also evaluated with the same dataset and evaluation process. The ROC curve for CatBoost provided an AUC of 0.93, which also indicates a high degree of classification performance. While slightly less than XGBoost, the CatBoost model still shows strong ability in separating the two classes. The findings verify that CatBoost is extremely effective for diabetes prediction, providing stable performance with little compromise. Both models outperform the baseline random classifier (AUC = 0.5) significantly, with XGBoost having a slight advantage in terms of overall predictive accuracy.

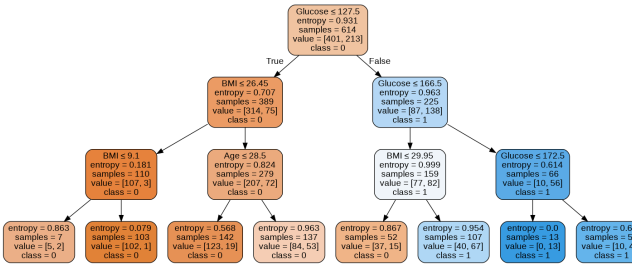

Fig. 4. Decision Tree Visualization

The graphical representation of the decision tree explains how various attributes of the Indian PIMA diabetes dataset are utilized to predict the outcome of diabetes. The root node of the tree initiates the split based on glucose level with a threshold value of 127.5.This means that glucose is a highly relevant feature in the prediction of diabetes, as revealed through clinical observation.Further splits are based on BMI and age, and entropy values at every node are provided to quantify the level of uncertainty.Low entropy for leaf nodes describes the purity of classification. The tree indicates patients with higher values of glucose and BMI have a higher probability of being in the diabetic class (class = 1) with smaller values usually in non-diabetic (class = 0). Hierarchical form depicts that decision trees employ feature thresholds such that they partition data and generate a prediction.

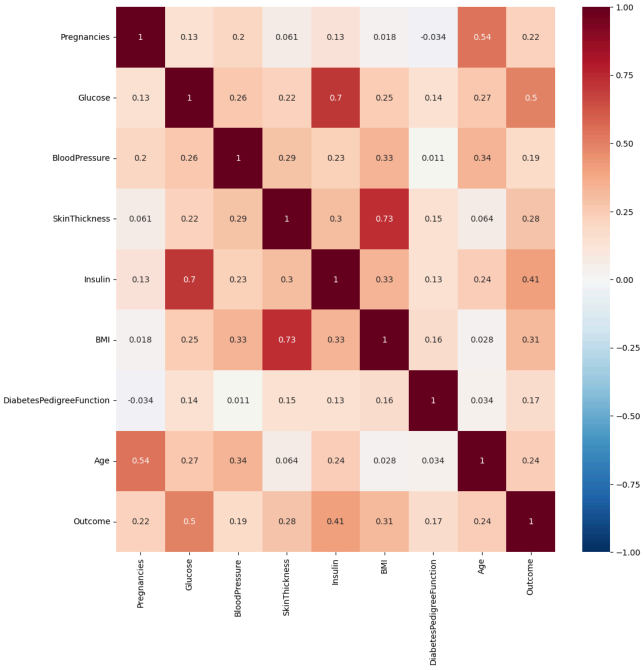

Fig. 5. Heatmap for Diabetes Prediction

The correlation heatmap provides an intuitive picture of the relationships between features in the dataset as identified by their Pearson correlation coefficients. The strongest correlations observed are between insulin and glucose levels (r = 0.70) and BMI and skin thickness (r = 0.73). The strong correlations show that these pairs of features vary together, indicating some association with patient health profiles. Furthermore, glucose was correlated with the outcome variable (0.50), which holds some importance for diabetes prediction. Other characteristics such as age and pregnancies showed association with the outcome but were weak. As evidenced by the strong correlations shown in the heatmap, they are useful for feature selection and will enhance the performance of machine-learning algorithm.

IV. Results

Table I. Comparison of models

Model	**Accuracy (%)**	**Limitations**
KNN	80	performance drops with high-dimensional data.
Decision Tree	79.87	Prone to overfitting without pruning or ensemble methods.
XGBoost	97	Uses L1 and L2 regularization,so reduces overfitting
CatBoost	96	Uses L1 and L2 regularization,so reduces overfitting

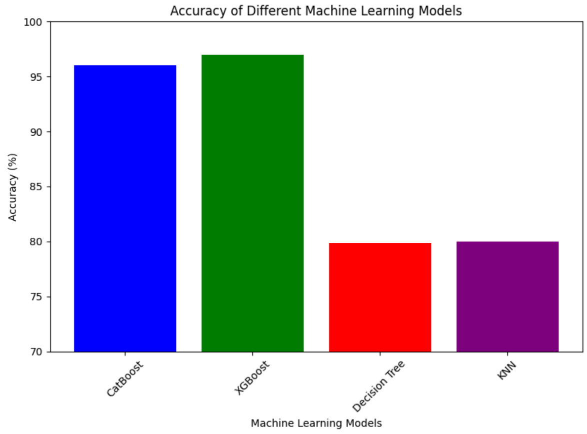

Fig 6. Bar Chart to compare different ML Models

V. Conclusion

In order to improve diabetes diagnostic reliability and accuracy, we experimented in this work with the use of XGBoost and other ML techniques. A comparative study shows how XGBoost systematically surpasses standard models due to the way it addresses imbalanced datasets better, exhibits stronger regularization, and captures sophisticated nonlinear dependencies better. Upon combining it with effective preprocessing and feature engineering methods, XGBoost brings dramatically increased diagnosis accuracy with much-needed tools for early diabetes detection and control. Upcoming research might involve deep learning methods or hybrid models to further improve predictive performance and overcome real-world clinical dataset limitations.

References

[1] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, and S. Abdulwahed, "Diabetes Classification Based on KNN," *IIUM Engineering Journal*, vol. 21, no. 1, 2020. [Online]. Available: https://doi.org/10.31436/iiumej.v21i1.1206

[2] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbour Algorithm," *Int. J. Comput. Sci. Trends Technol.*, vol. 2, no. 4, pp. –, 2014.

[3] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metab. Disord.*, 2020. [Online]. Available: https://doi.org/10.1007/s40200-020-00520-5

[4] K. Hasan *et al.*, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, 2020. DOI: 10.1109/ACCESS.2020.2989857

[5] I. Tasin *et al.*, "Diabetes Prediction Using Machine Learning and Explainable AI Techniques," *Healthcare Technology Letters*, 2022. DOI: 10.1049/htl2.12039

[6] D. Sisodia and D. S. Sisodia, "Diabetes Prediction using Machine Learning Algorithm," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

[7] M. Soni and S. Varma, "Diabetes Prediction using Machine Learning Techniques," *Int. J. Eng. Res. Technol.*, 2020.

[8] S. Perveen *et al.*, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Symposium on Data Mining*, 2016.

[9] A. Mujumdar and V. V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[10] R. Rane *et al.*, "A Novel Prediction Model for Diabetes Detection Using Gridsearch and A Voting Classifier between LightGBM and KNN," *2021 2nd Global Conference for Advancement in Technology (GCAT)*, 2021. DOI: 10.1109/GCAT52182.2021.9587551