Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Diabetes Diagnosis using Machine Learning

Sadhasivam N[1], Harish J[2], Bharanidharan M[2]

[1]Associate professor, Computer Technology, Bannari Amman Institute of Technology, Erode, India

[2]Information Technology, Bannari Amman Institute of Technology, Erode, India

Email: [1]sadhasivamn@bitsathy.ac.in, [2]harish.it20@bitsathy.ac.in, [2]bharanidharan.it20@bitsathy.ac.in

Abstract

This abstract presents a study on utilizing the Gradient Boosting algorithm for diabetes diagnosis. The objective is to develop a reliable and effective model that uses patient data, to detect the presence of diabetes. For training and testing, a dataset made up of clinical parameters like age, body mass index, blood pressure, and glucose levels are used. The Gradient Boosting algorithm is implemented and optimized to achieve optimal predictive performance. The model's accuracy, precision, recall, and F1 score are evaluated to assess its effectiveness. The results of this study indicate that the Gradient Boosting algorithm's effectiveness in correctly identifying diabetes and highlight its potential as a trustworthy tool for clinical diagnosis. In order to improve the model's performance and expand its application in real-world healthcare settings, future study can concentrate on adjusting its parameters and investigating new characteristics.

Keywords: Gradient Boosting Algorithm, Diabetes, Diagnosis, Patient Data, Dataset, Accuracy, Precision, Performance.

1. Introduction

High blood sugar levels are a defining feature of diabetes, a chronic metabolic condition brought on by either inadequate insulin synthesis or inefficient insulin use. For efficient diabetes treatment and the avoidance of complications, an early and correct diagnosis of diabetes is essential [4]. Machine learning algorithms have advanced, and there is significant interest in using these methods to enhance diabetes diagnosis. Gradient Boosting is one such method that has produced encouraging results in a number of fields. An ensemble technique called gradient boosting combines a number of weak learners to produce a powerful predictive

DOI: https://doi.org/10.36548/jtcsst.2023.4.002
Received: 02.10.2023, received in revised form: 28.10.2023, accepted: 11.11.2023, published: 25.11.2023

model. It creates a sequence of decision trees in a sequential fashion, each one addressing the flaws in the one before it [5]. The purpose of this study is to investigate how the Gradient Boosting algorithm can be used to diagnose diabetes. The system can identify trends and forecast whether diabetes will be present by using a dataset with pertinent clinical parameters including age, body mass index (BMI), blood pressure, and glucose levels. There are various possible benefits in using gradient boosting to diagnose diabetes [6]. It is useful for combining many clinical variables since it can handle both numerical and categorical data well. Second, it may record intricate non-linear correlations between features, which are frequently found in datasets related to medical datasets. Finally, by modifying the weights throughout the training phase, it can handle unbalanced datasets, which are also frequent in medical data. The major goal of this research is to use the Gradient Boosting method to build a reliable and precise model for diabetes diagnosis [7]. We may judge a model's success in identifying diabetic cases by looking at parameters like accuracy, precision, recall, and F1 score. The results of this study have important ramifications for healthcare professionals since they can help with early intervention, individualized treatment strategies, and better patient outcomes [8]. By expediting the diagnostic process, it can also help lessen the strain on healthcare systems. The approach, including the dataset used, data pre-treatment methods, model implementation, and performance assessment, will be covered in the following upcoming sections. Following a discussion of the ramifications, restrictions, and prospective areas for further research. The machine learning model is also evaluated using the dataset [9,10].

2. Related Work

The use of machine learning algorithms, such as Gradient Boosting, for diabetes detection has been examined in a number of publications. These publications serve as a foundation for the current investigation and offer insightful information.

- Deberneh et al. Found Type 2 diabetes incidence in the Korean population was found to be reasonably well-predicted by the prediction model, which can also give patients and physicians useful information about the risk of acquiring T2D [2].
- Wang et al. In his study suggests an ensemble learning-based type 2 diabetes risk prediction algorithm. Extreme gradient boosting (XGBoost) and the weighted feature

selection technique based on random forest (RF-WFS) are utilised in the suggested model to identify the best features. [3].

- Liu et al. Demonstrated that GOSS, also known as LightGBM, can produce relatively accurate information gain estimation with a considerably less data quantity because the data instances with greater gradients are more significant in the computation of information gain. (XGBoost) [14].

- Rufo D.D et al. Suggested Light Gradient Boosting Machine (LightGBM). Because of its low computing complexity, it is appropriate for use in areas with limited resources, like Ethiopia. Therefore, the LightGBM concept was used in this work to create an accurate diabetes diagnosis model. The outcomes of the experiment demonstrate how useful the produced diabetes dataset is for predicting the presence of diabetes mellitus. With AUC, sensitivity, specificity, and accuracy of 98.1%, 98.1%, 99.9%, and 96.3%, respectively, the LightGBM model performed better on the ZMHDD dataset than KNN, SVM, NB, Bagging, RF, and XGBoost.[12].

- Hu et al. Developed a model employing the XG Boost ML and demonstrated that the model had superior predictive ability than the conventional LR model, and both models' calibration performance was strong [11].

- Birjais, et al, in his research has attempted to concentrate more on the diagnosis of diabetes, which the World Health Organisation identified in 2014 as one of the chronic diseases with the fastest rate of growth worldwide, and the various methods that can be used to achieve a correct diagnosis, such as Gradient Boosting, Logistic Regression, and Naive Bayes[1].

In combination, these related studies demonstrate the value of gradient boosting in the diagnosis of diabetes. For optimum performance, they stress the significance of integrating various data sources, feature selection, and interpretability. However, more study is needed to improve the Gradient Boosting algorithm's precision and generalizability in actual clinical contexts. In this study, we expand on these earlier works by concentrating explicitly on the use of gradient boosting for the diagnosis of diabetes. The proposed study aims to contribute to the existing body of knowledge by providing additional insights into the performance of the

algorithm and its implications for clinical practice. We seek to present a robust and reliable model for diabetes diagnosis using the Gradient Boosting technique by making use of a large dataset and stringent evaluation metrics.

3. Proposed Work

By employing the suggested methodology, the field of medical diagnostics is advanced to offer insightful information to academics and practitioners in the healthcare industry by creating a precise and dependable model for diabetes diagnosis utilising the Gradient Boosting algorithm. Figure 1 provides a better understanding of the following steps for the methodology used.

- Data Collection: An extensive Pima Indian Diabetes dataset with pertinent clinical characteristics, including age, gender, BMI, blood pressure, glucose levels, and other potential risk factors for diabetes is gathered from Kaggle to identify the diabetic and non-diabetic. The dataset included the pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age and outcome of 768 female patients both diabetic and non-diabetic [13].

- Data Pre-processing: To deal with missing values, outliers, and data normalisation, perform data pre-processing steps. To guarantee that the Gradient Boosting algorithm performs at its best, missing values should be imputed using the right methods, outliers should be identified and handled, and numerical characteristics should be scaled to a common range. The python code is developed to checks for the missing values, identifying the outliers and as well as removing the outliers and normalize the dataset

- Feature Selection: Use feature selection strategies to determine which features are most useful for diagnosing diabetes. Select a subset of pertinent features that greatly improve prediction accuracy using techniques like correlation analysis, feature importance ranking, or recursive feature elimination. The proposed study utilises recursive feature elimination, a backward feature elimination technique, to extract the features. The RFECV from scikit-learn is used to perform the feature extraction; it identifies the features with less importance and removes them from the feature set. The process is

repeated until the optimal number of features is reached. Figure 4 below shows the most suitable features selected for training.

- Dataset Split: Split the training and testing sets from the pre-processed dataset. Typically, a ratio of 70:30 or 80:20 is employed to ensure that there is enough data for the model's training while still allowing for thorough evaluation.

- Gradient Boosting Model Training: The selected features are used in training the model using the Gradient Boosting technique. To improve the model's performance and the training accuracy hyperparameters like learning rate, the number of trees, and maximum depth are tuned employing the randomized Search. The figure.1 below shows the steps involved in the proposed methodology.

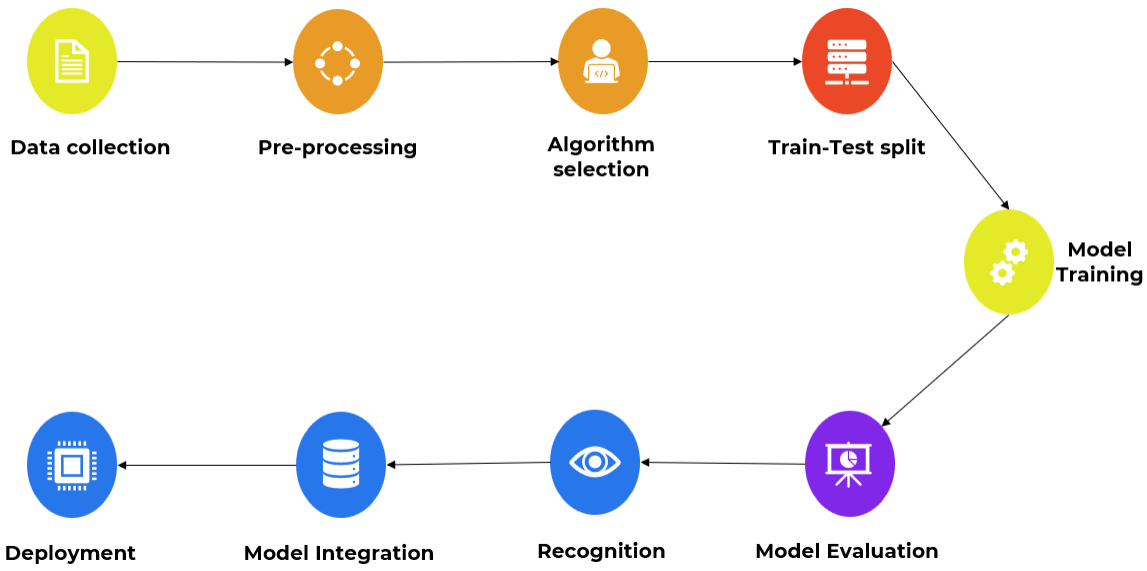


Figure 1. Proposed Methodology

Gradient boosting is an ensemble machine learning strategy that combines a number of weak learners (usually decision trees) to produce a powerful prediction model by repeatedly focusing on the errors produced by earlier models, thereby progressively increasing accuracy. Under the gradient boosting method, the two most popular models are XGBoost and LightGBM. Since it was first put forth in 2017, LightGBM has drawn a lot of interest. It is an effective implementation of gradient boosting trees known as an adaptive gradient boosting model. It seeks to increase computing effectiveness in order to address big data prediction challenges more successfully. LightGBM primarily uses the histogram technique to combine

mutually incompatible information in order to increase the processing power and prediction accuracy of the algorithm. To speed up training and use less memory, LightGBM uses a histogram-based method and a tree leaf-wise growth plan with a maximum depth limit. The fundamental concept underlying the histogram-based strategy is to discretize the continuous feature values into X integers first, after which a histogram of some variable's width W is built. The data is traversed to find the decision tree based on the discretized values of the histogram. Since the decision tree is a weak model, the histogram approach considerably reduces the time complexity, and this fuzzy partitioning method frequently produces better results. On the other hand, the gradient boosting technique has two different growth strategies. They are level-wise and leaf-wise growth strategies respectively (Figure 2). The leaves on the same layer are divided at the same time in the level-wise growth strategy.

However, despite the fact that they have differing information gains, leaves on the same layer are regarded indifferently. Information gain denotes the anticipated decrease in entropy brought on by dividing the nodes according to different properties. The XGBoost algorithm employs a method like this. By just splitting the leaf with the greatest information gain on the same layer, the leaf-wise development method is more effective. LightGBM algorithm uses a leaf-wise split [12].

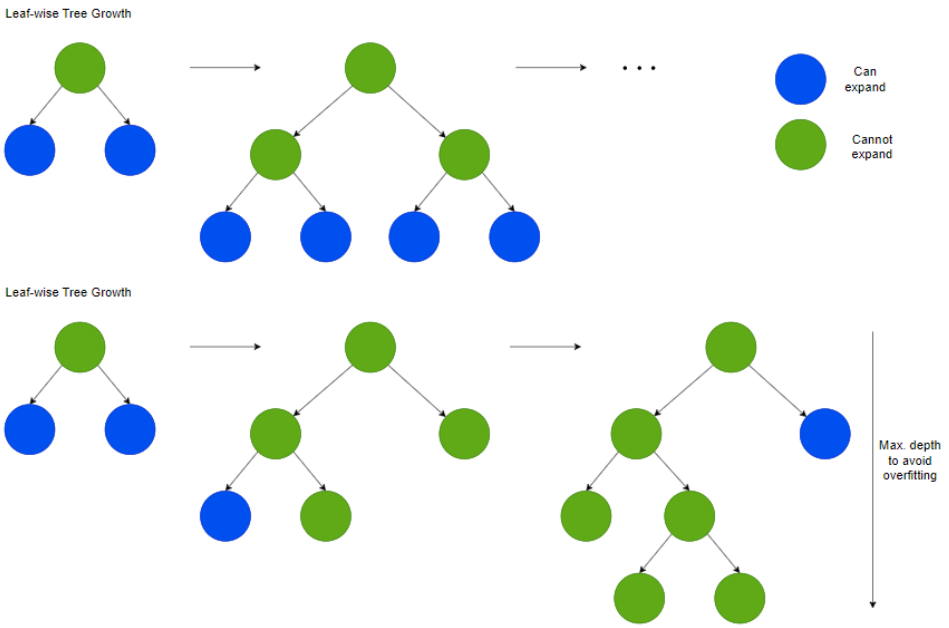


Figure 2. Comparing Level-wise and Leaf-wise Approach [14]

Furthermore, a maximum depth restriction is implemented during the formation of trees because this technique may result in trees with excessive depth, which could lead to overfitting. Therefore, LightGBM augments Leaf-wise with a maximum depth limit to maintain high efficiency while avoiding overfitting. In summary, LightGBM is utilized for machine learning jobs that require accuracy and efficiency. With quick training and prediction timeframes, it excels at handling massive datasets (Figure 3). It is favoured in competitions and practical applications since it is primarily utilized for classification and regression.

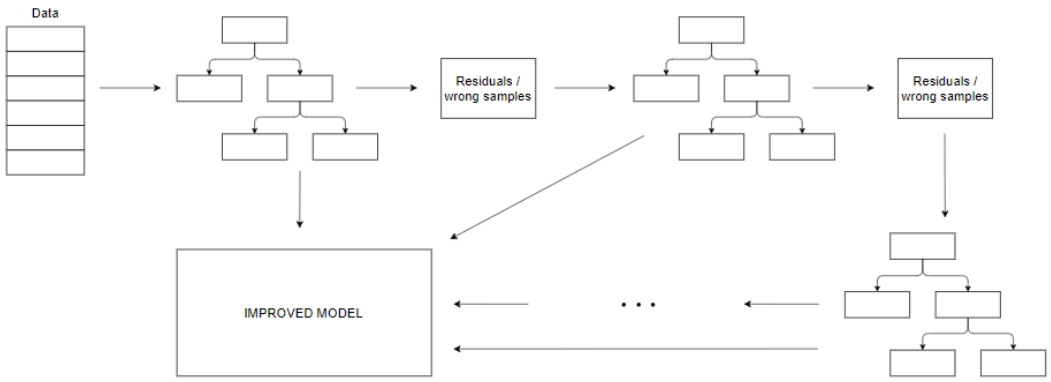


Figure 3. Methodology of LightGBM Algorithm

4. Results and Discussion

Utilizing the testing dataset, the trained model is evaluated. To evaluate the model's propensity to correctly forecast diabetic cases, the performance metrics such accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) are computed. The findings are validated by comparing the proposed method with the state of art approaches that are existing. The model is trained and tested using python.

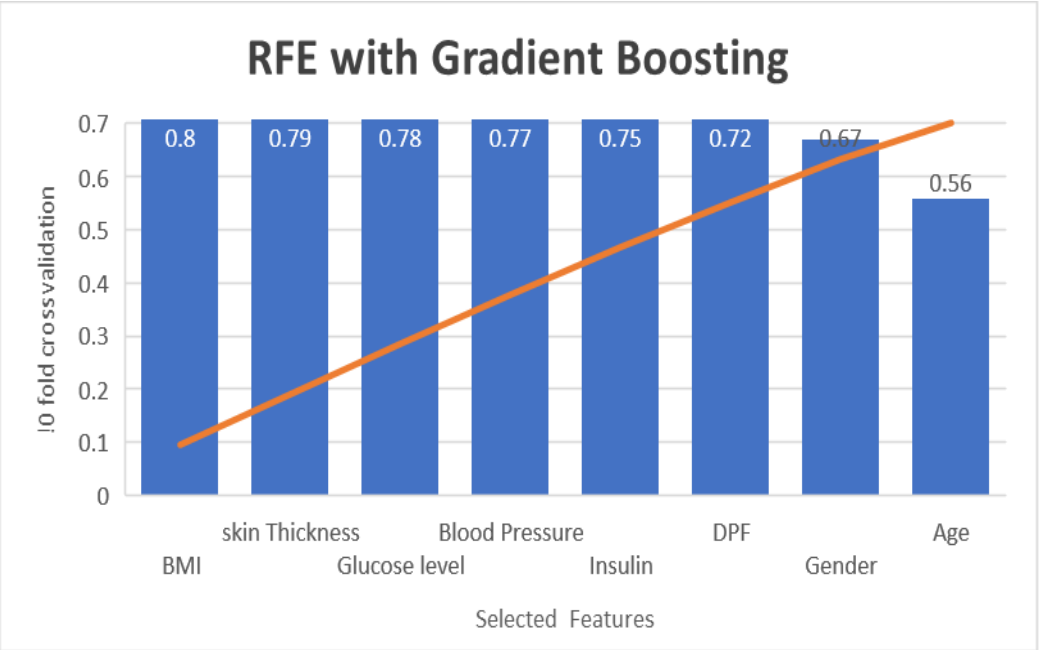
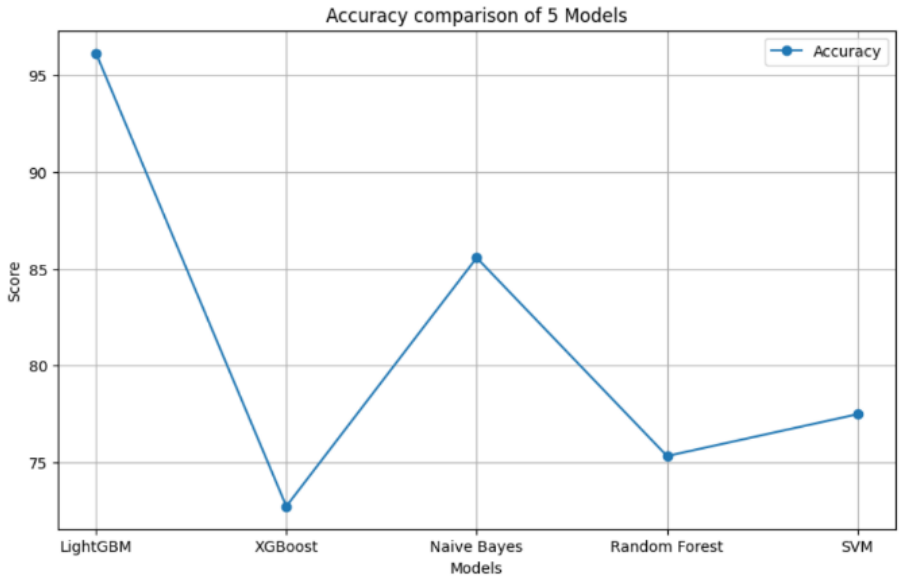
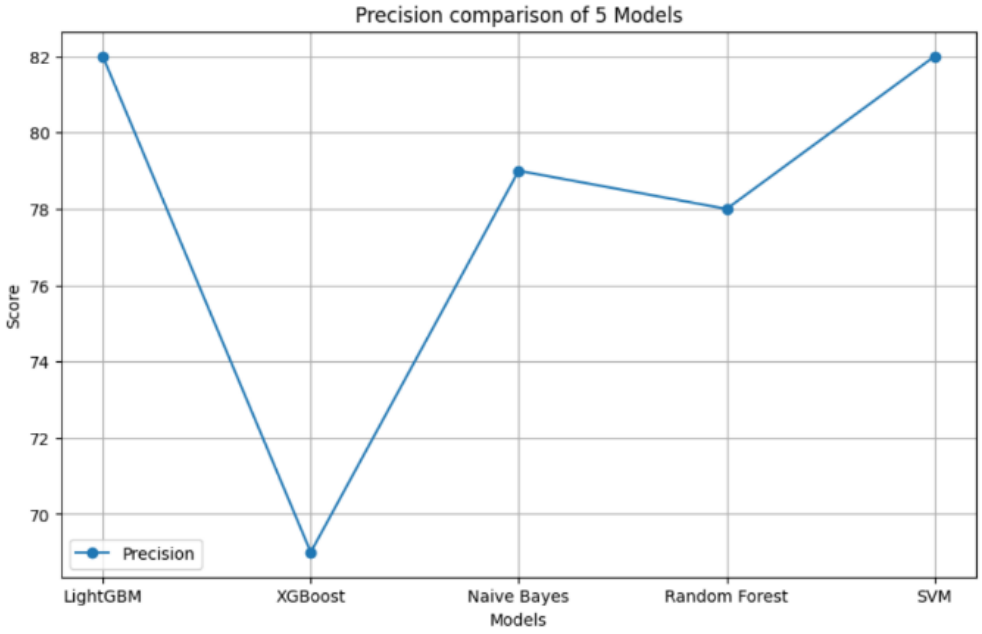


Figure 4. Selected Features

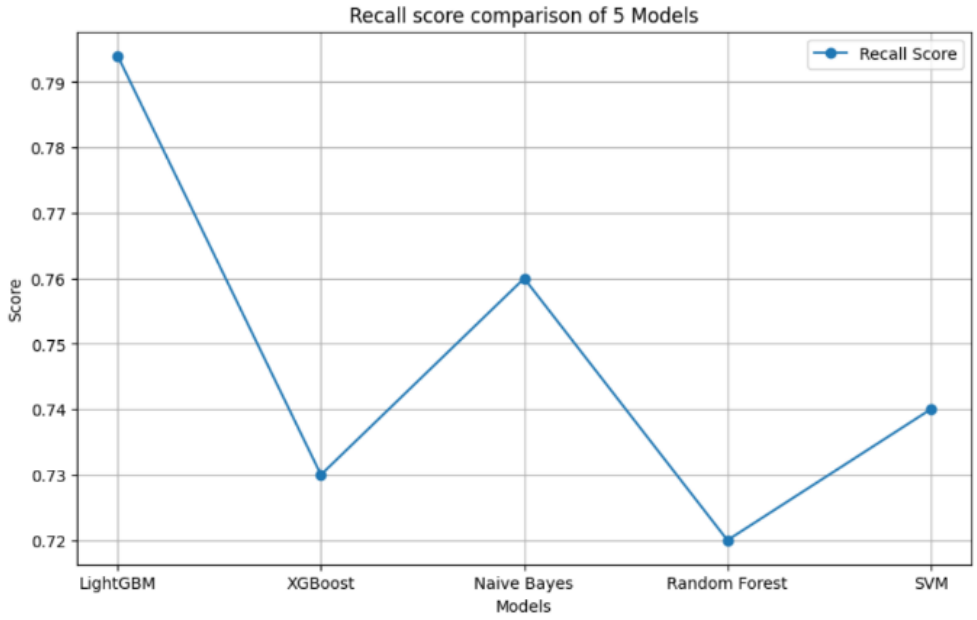
The figure .5 illustrates the comparison of different algorithms with the LightGBM based on the metrics accuracy, precision, recall, and F1 score. The results observed through each approach is satisfactory with an average an average accuracy of 80%



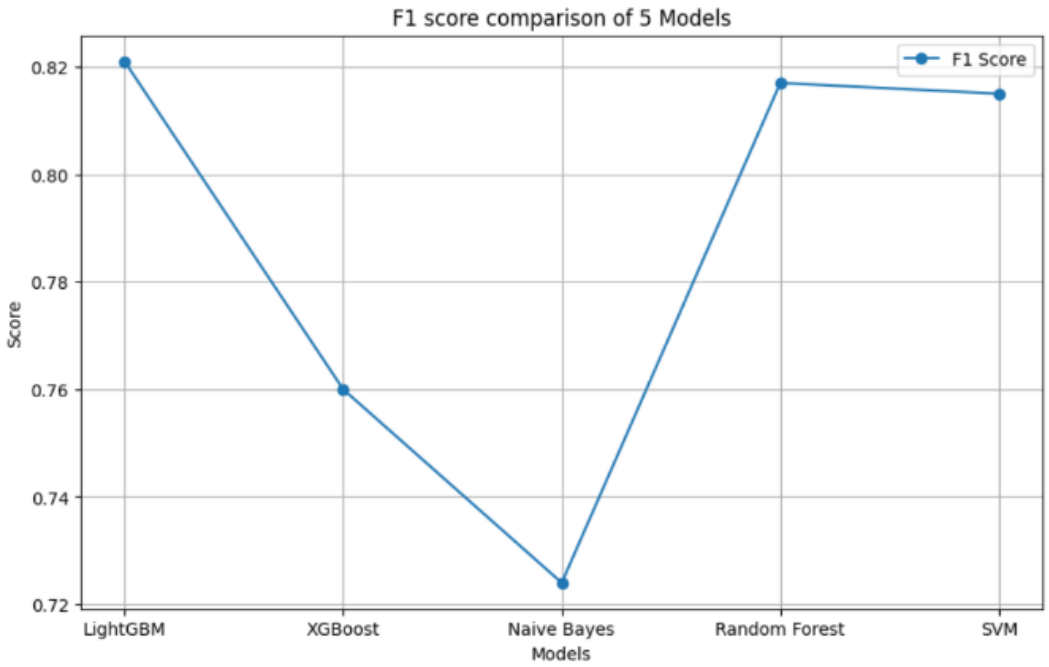
(a)


Precision comparison of 5 Models
Score
82
80
78
76
74
72
70
Precision
LightGBM
XGBoost
Naive Bayes
Random Forest
SVM
Models

(b)


Recall score comparison of 5 Models
Recall Score
Score
0.79
0.78
0.77
0.76
0.75
0.74
0.73
0.72
LightGBM
XGBoost
Naive Bayes
Random Forest
SVM
Models

(c)



(d)

Figure 5. (a,b,c,d). Comparison between Various Models

The findings of the study suggest that Light Gradient Boosting is an effective machine learning technique for diagnosing diabetes. It captures intricate non-linear correlations between variables and handles a variety of clinical aspects with ease. The model's interpretability enables one to comprehend the significance of various features in the diagnosis procedure. The established model has important consequences for patients and healthcare professionals. By making a timely diagnosis of diabetes possible, LightGBM is much better and can clearly help with early intervention, individualised treatment strategies, and improved patient outcomes. By expediting the diagnostic procedure, the model can also lessen the strain on healthcare systems. Although the Gradient Boosting method demonstrated encouraging results in our investigation, there is still room for growth. The performance and generalizability of the model might be improved by tweaking the hyperparameters, investigating ensemble methods, and incorporating other data sources. In conclusion, using the Gradient Boosting algorithm to diagnose diabetes gives a trustworthy and effective method to help medical professionals correctly detect and treat diabetic cases. The results of this study add to the expanding body of knowledge in the area of medical diagnostics and demonstrate how machine learning has the potential to enhance healthcare outcomes.

The robustness of the model, can be tested by performing the sensitivity analysis changing the dataset or a few key features. This is usually done to assure the model's dependability in real-world applications, by evaluating the model's stability and performance under various scenarios which is the future work of the research. The table.1 below shows the experimental results and the computational time observed for PIDD dataset.

Table 1. Experimental Results Observed

Machine Learning Models	**Accuracy %**	**Precision %**	**Recall %**	**F1 %**	**Training Time (s)**	**Testing Time (s)**
LightGBM	96.5	82.5	80	82	.621	.001455
XGBoost	72.8	61.5	73	76	1.89	.00389
Random Forest	85.4	79.3	76	72 .4	2.34	.01099
Navie Bayes	76.2	78.1	72	81.9	.0582	.00599
SVM	78.4	81.9	74	81.8	.0798	.00655

5. Conclusion

In this study, the use of the Light Gradient Boosting algorithm for the diagnosis of diabetes is investigated. The goal was to create a reliable and effective algorithm that uses patient data to predict the presence of diabetes. LightGBM, XGBoost, Naïve Bayes, Random Forest Regression and Support Vector Machine algorithms were used for performing a comparison study. The “PIMA Indian Diabetes Dataset (PIDD)” that consists of 8 factors namely Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree function and age was chosen. All these factors are considered for evaluating the Outcome factor. In Figure 5, visualization for Accuracy, Precision, Recall Score, and F1 score for all 5 models are provided for a better understanding. The Gradient Boosting model was trained using a large dataset with pertinent clinical features. High accuracy, precision, recall, and F1 score all indicated that the model performed well at correctly detecting diabetic cases.

References

[1] Birjais, Roshan, Ashish Kumar Mourya, Ritu Chauhan, and Harleen Kaur. "Prediction and diagnosis of future diabetes risk: a machine learning approach." SN Applied Sciences 1 (2019): 1-8.

[2] Deberneh, Henock M., and Intaek Kim. "Prediction of type 2 diabetes based on machine learning algorithm." International journal of environmental research and public health 18, no. 6 (2021): 3317.

[3] Xu, Zhongxian, and Zhiliang Wang. "A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier." In 2019 eleventh international conference on advanced computational intelligence (ICACI), pp. 278-283. IEEE, 2019.

[4] Chaki, Jyotismita, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan. "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review." Journal of King Saud University-Computer and Information Sciences 34, no. 6 (2022): 3204-3225.

[5] Alassaf, Reem A., Khawla A. Alsulaim, Noura Y. Alroomi, Nouf S. Alsharif, Mishael F. Aljubeir, Sunday O. Olatunji, Alaa Y. Alahmadi, Mohammed Imran, Rahma A. Alzahrani, and Nora S. Alturayeif. "Preemptive diagnosis of diabetes mellitus using machine learning." In 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1-5. IEEE, 2018.

[6] Faruque, M. F., & Sarker, I. H. (2019, February). Performance analysis of machine learning techniques to predict diabetes mellitus. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-4). IEEE.

[7] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." Ict Express 7, no. 4 (2021): 432-439.

[8] Choudhury, Ambika, and Deepak Gupta. "A survey on medical diagnosis of diabetes using machine learning techniques." In Recent Developments in Machine Learning and Data Analytics: IC3 2018, pp. 67-78. Springer Singapore, 2019.

[9] Palimkar, Prajyot, Rabindra Nath Shaw, and Ankush Ghosh. "Machine learning technique to prognosis diabetes disease: Random forest classifier approach." In Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021, pp. 219-244. Springer Singapore, 2022.

[10] Insani, M. Ilham, Alamsyah Alamsyah, and Anggyi Trisnawan Putra. "Implementation of expert system for diabetes diseases using naïve Bayes and certainty factor methods." Sci. J. Informatics 5, no. 2 (2018): 185-193.

[11] Hu, Xiaoqi, Xiaolin Hu, Ya Yu, and Jia Wang. "Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm." Frontiers in Endocrinology 14 (2023): 1105062.

[12] Rufo, Derara Duba, Taye Girma Debelee, Achim Ibenthal, and Worku Gachena Negera. "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)." Diagnostics 11, no. 9 (2021): 1714.

[13] https://www.kaggle.com/code/gifarihoque/pidd-missing-data-ml-iterimputer-tut-86

[14] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).