Biomedical & Pharmacology Journal, June 2026. Vol. **19**(2), p. 1682-1698

A Two-Stage Ensemble Approach for Diabetes Prediction: Early Diagnosis with CatBoost and Advanced Diagnosis with LightGBM

Smita Kulkarni[1]*, Dnyanda Hire[2], Priya Charles[2] and Shweta Suryawanshi[2]

[1]Department of E and TC Engineering, MIT Academy of Engineering, Pune, India.
[2]Department of Semiconductor Engineering, School of Engineering, Management and Research, D. Y. Patil International University, Pune, India.
*Corresponding Author E-mail: sskulkarni@mitaoe.ac.in

https://dx.doi.org/10.13005/bpj/3449

(Received: 13 July 2025; accepted: 11 February 2026)

Diabetes is a common long-term sickness defined by elevated blood sugar levels as a result of impaired insulin physiological effects, faulty insulin secretion, or both. This condition can cause long-term damage and dysfunction to various tissues, including the kidneys, heart, blood vessels, eyes, and nerves. As living standards grow, diabetes is becoming increasingly common, making early and accurate detection crucial. This research focuses on predicting diabetes in two stages using machine learning. In this research, a two-stage approach was implemented for diabetes prediction using ensemble models. The first stage focused on the early diagnosis of diabetes to provide prior intimation about individuals' health status. This involved using the Sylhet dataset, which contains comprehensive information for detecting prediabetes. In the second stage, the Frankfurt dataset was utilized, which includes numerical parameters for further diabetes diagnosis, to predict diabetes based on pathological parameters and to provide appropriate treatment to prevent further health issues.The article falls under the domain of Biomedical Signals and Medical Sciences, with its scope focusing on the early and advanced diagnosis of diabetes using machine learning ensemble models, involving medical data analysis, preprocessing, and the use of biomedical signals or patient health parameters. Various ensemble models were employed in both stages. In stage one, the Categorical Boosting(CatBoost) algorithm demonstrated superior performance for early diagnosis using the Sylhet dataset, while in stage two, the Light Gradient Boosting Machine(LightGBM) algorithm proved to be most effective for diabetes prediction using the Frankfurt dataset. Selecting the appropriate classifier and correct featuresa critical challenge for machine learning techniques in this domain. The findings suggest that ML models are beneficial for diabetes prediction and can significantly contribute to improving human health.Future research could compare these ensemble models against a deep learning technique, such as CNNs or RNNs, on both datasets to enhance accuracy for early and advanced diabetes prediction.

Keywords: Categorical Boosting (CatBoost); Diabetes Prediction; Ensemble Learning; Light Gradient Boosting Machine (LightGBM); Machine Learning.

Diabetes Mellitus, which is occasionally abbreviated as "Diabetes", is a chronic metabolic disorder that is characterised by elevated blood sugar levels that are the result of inadequate or inefficient insulin administration. Diabetes, which is anticipated to continue increasing,

is estimated to affect 463 million individuals worldwide in 2019 and poses a substantial threat to global health. Machine learning integration is being investigated as a means of improving early diagnosis and intervention through the use of prediction tools. There are two types of diabetes: type 1 and type 2. Insulin resistance, or the insufficient responsiveness of cells to insulin, is the primary cause of type 2 diabetes. Type-1 diabetes is characterised by frequent urination, excessive thirst, loss of consciousness, irritation, fatigue, and intense hunger. Insulin replacement therapy, consistent exercise, healthful eating, and blood sugar monitoring are the primary prevention objectives. Nerve injury, kidney failure, hypoglycaemia, diabetic ketoacidosis, and heart issues are among the complications. Common clinical indications include elevated blood glucose levels, as well as increased thirst and urination. Insulin therapy is necessary for individuals with this type of diabetes, as oral medications are ineffective in managing it.

Diabetes mellitus, which is sometimes just called "diabetes,", is a long-term metabolic problem that causes blood sugar levels to be too high because the body doesn't use insulin properly. Diabetes is likely to get worse over time, and by 2019, it will affect about 463 million people around the world. It is a big danger to the health of the whole world. People are looking into how to use machine learning to improve early diagnosis and treatment with prediction tools. There are two types: Type 1 and Type 2. Insulin resistance is the main cause of type 2 diabetes. This condition means that cells don't respond to insulin the way they should. People with type 1 diabetes have to go to the toilet a lot, are always thirsty, lose their minds, and are always grumpy, worn out, and hungry. Insulin replacement therapy, regular exercise, eating well, and checking blood sugar levels are the most important things to do to avoid diabetes. Some of the problems that can occur are nerve damage, kidney failure, low blood sugar, diabetic ketoacidosis, and heart problems. High blood sugar and needing to pee, and drinking more often are common signs. People with this kind of diabetes need insulin therapy because oral medicines don't work to control it.

In [3 4] paper presents algorithms like 2GDNN-FL and CatBoost model and LightGBM (LGB) feature selection method proposed in [5] that improve the prediction of myocardial infarction (MI) risk using unbalanced datasets.

Machine learning algorithms can identify patterns and connections in extensive clinical and demographic data that may indicate the onset of diabetes. [6] In recent years, hybrid methods, which use more than one machine learning model, have become more popular as a way to make predictions more accurate and reliable. Some of these methods are neural networks, support vector machines, random forests, and gradient boosting. The ensemble method can make predictions more accurate and lower the chance of overfitting by using a number of models. using methods that work together to guess diabetes. The main goal of collaborative learning is to make predictions more accurate and dependable. This will make it possible to analyse and treat early. The primary aim of this study is to determine the most efficient method for diagnosing diabetes through the comparison of various bagging and boosting techniques. The study elucidates the application of predictive models to enhance diabetes management and proposes strategies to implement more effective interventions for diabetes prevention. When used this way, big data can help you quickly and correctly guess if you have diabetes. People who are thought to be at high risk for diabetes can join programmes that help them stay healthy and lower their risk. To make prediction models more accurate and reliable, we need to do more research in this area. This would help people with diabetes control their condition better and get better results.

This review used many group calculations. These are the best parts of different machine learning models that work together to improve performance. These methods make things even more accurate by using the best parts of each model and fixing the undesirable parts. By working together on these big-picture plans, we can make it much easier and more reliable to find diabetes. This will help doctors give their patients better care and outcomes in the long run.

Literature Survey

Through ensemble learning in deep learning frameworks, significant progress has been achieved in improving model resilience and prediction accuracy across a variety of domains, particularly in healthcare. Foundational studies

highlight the value of variety in base classifiers, highlighting techniques like boosting, stacking, and bagging, and highlighting how these methods can improve diagnostic accuracy, especially in the early diagnosis of disease. These studies do, however, also draw attention to certain issues, such as the requirement for bigger datasets and more robust validation procedures. Although individual classifiers are often outperformed by ensemble models, problems such as feature redundancy and class imbalance still exist. These results are consolidated in this review, which highlights the promising potential of ensemble learning to tackle challenging real-world issues and recommends areas for future study.

An extensive description of ensemble learning inside deep learning models can be found.[7] The significance of variety among base classifiers is examined, and the advantages and disadvantages of several ensemble approaches are discussed. The paper assesses the use of ensemble learning in a variety of fields, as well, while it points out that recent developments and computational challenges are not fully covered.

The use of ensemble techniques for diabetic retinopathy classification is examined in with a focus on the use of voting classifiers.[8] Al-Although the study shows that ensemble techniques are effective, a more thorough explanation of performance indicators and a comparison with other models would be beneficial.

Ensemble classification techniques are investigated to diagnose pre-diabetes early.[9] The study achieves excellent accuracy by combining machine learning algorithms with conventional diagnostic procedures, but it also recognizes that larger and more diverse datasets are necessary for improved model generalization.

Introduction to Diabetes Mellitus provides a comprehensive overview of diabetes,[10] covering its various forms, complications, and new advancements in treatment. The study emphasizes the worrying trend of untreated potential diabetics and draws attention to the challenges in precisely measuring diabetes prevalence because of different data collection methods.

The goal of the research paper is to enhance the prediction of diabetes onset by using heterogeneous ensemble learning.[11] While the study reveals notable advances in accuracy, it also points up certain problems, such as feature redundancy and class imbalance, which call for more investigation to resolve.

Data mining and algorithms are used to predict complications in individuals with diabetes.[12] The study highlights the value of ma-machine learning in healthcare; however, it does not provide specific implementation information for actual healthcare environments. The article explores the advantages of using ensemble models for early diabetes detection.[13] Its dependence on a single dataset, however, can restrict how broadly applicable its conclusions can be, emphasizing the necessity of testing on a variety of datasets.

In order to prevent end-organ damage in diabetic patients, the study emphasizes the significance of early detection and consistent diagnostic standards.[14] It draws attention to the rising incidence of diabetes worldwide, but it doesn't discuss contemporary treatment options or the psychological effects of the disease. The paper exhibits the higher performance of ensemble methods, such as Random Forest, in network traffic classification when compared to traditional algorithms.[15] Notwithstanding its merits, the study recommends further investigation into cutting-edge domains such as blockchain and cybersecurity.

In the paper, ensemble approaches used to predict diabetes and diseases connected to cholesterol.[16] The work emphasizes the value of feature selection and data preprocessing, but it is limited by the use of a single dataset, suggesting that further research is needed to confirm the conclusions and enhance model resilience.

An ensemble of LightGBM and AdaBoost for Type-2 diabetes prediction, achieving over 90% accuracy on the PIMA dataset.[17] Their study emphasizes the strength of boosting methods in handling medical data with minimal preprocessing. A comparative analysis of boosting algorithms, including CatBoost, LightGBM, and XGBoost, as performed.[18] Their findings indicate that LightGBM performed best overall, while CatBoost showed superior performance in cases with categorical features and smaller datasets, supporting its role in early-stage diagnosis.

Multiple ensemble strategies were explored using CatBoost, LDA, and Random Forests, reporting that combining classifiers improved prediction accuracy and robustness.[19]

They demonstrated the advantage of ensemble diversity in clinical prediction tasks. The class imbalance problem in diabetes datasets was addressed using SMOTE,[20] coupled with CatBoost and LightGBM. Their study highlighted how proper preprocessing can significantly enhance model performance, particularly in the early detection phase.

The paper developed a CatBoost-based ensemble model was developed for early-stage diabetes risk prediction,[21] emphasizing its ability to perform well with small clinical datasets and imbalanced classes—aligning with your approach's first stage. A two-stage model, using CatBoost for early-stage screening and LightGBM for advanced diagnosis.[22] Their method showed improved performance over single-model approaches, reinforcing the value of progressive ensemble diagnosis strategies. An enhanced bagging method for time-series diabetes trend prediction, demonstrating LightGBM's adaptability in both classification and regression settings.[23]

A systematic review of AI methods for diabetes prediction was conducted,[24] noting that boosting algorithms like CatBoost and LightGBM are among the most effective due to their high accuracy, interpretability, and minimal need for feature scaling. The paper used CatBoost to predict insulin resistance, a precursor to diabetes.[25] Their model achieved high AUC scores, showcasing CatBoost's strength in early diagnosis using complex clinical features.

This literature review comprehensively captures the notable advancements in the application of ensemble learning in the healthcare domain, particularly in the context of diabetes-related classifications and predictions. Ensemble approaches are more accurate and reliable than individual classifiers; however, they still face challenges with validation procedures, feature redundancy, and dataset size. To get over these restrictions and investigate cutting-edge developments in ensemble learning for more widespread and useful applications, further study is required.

MATERIALS AND METHODS

As shown in Figure 1, the workflow of the proposed two-stage ensemble approach for diabetes prediction. The process begins by collecting data, focusing on categorical information in the case of Sylhet and numerical information in Frankfurt useful for early diabetes detection and further confirmation of the disease. The data is then pre-processed to handle missing values, normalize numerical features, identify important features, and remove outliers using- Interquartile Range (IQR) method to improve data quality. Multiple ensemble models are trained on the prepared data. After evaluating the models for accuracy, the best-performing model is selected and deployed for two-stage diabetes detection.

Machine learning provides a diverse range of methods for successful classification strategies, from fundamental classifiers to complex ensemble approaches. Different machine learning models are used in our research, each of which brings special talents and qualities to the classification tasks on different datasets. These models provide a thorough method for addressing classification problems across several domains, ranging from basic classifiers to sophisticated approaches.

Machine Learning Models Used for Diabetes Prediction

Bagging

Random Forest (RF): An extension of bagging that constructs numerous decision trees during training. Each tree is built using a random subset of the training data and a random subset of features. The ensemble of trees helps in enhancing predictive accuracy and controlling overfitting.

Extra Trees (ET): Similar to Random Forest, this model also builds an ensemble of decision trees. However, Extra Trees differ in the way they split nodes; they select cut points randomly, often leading to better performance through increased model diversity.

Boosting

AdaBoost: Short for Adaptive Boosting, this algorithm builds models sequentially, with each new model focusing on the errors made by the previous ones. It adjusts the weights of misclassified instances, directing the subsequent models to give more attention to those instances.

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x)$$

where $F_m(x)$ is the boosted model after m iterations, α_m is the weight applied to the m-th model, and $h_m(x)$ is the weak learner.

Equation 1.1 Adaboost

Gradient Boosting Machine (GBM): This method creates an ensemble of trees sequentially. Each tree corrects the errors made by the previous ones by minimizing a loss function, resulting in a powerful aggregate model.

$$F_m(x) = F_{m-1}(x) - \gamma_m \nabla_\theta L(y, F_{m-1}(x))$$

where $F_m(x)$ is the model after m iterations, γ_m is the learning rate, and $\nabla_\theta L(y, F_{m-1}(x))$ is the gradient of the loss function.

Equation 1.2 GBM

Extreme Gradient Boosting (XGBM): Known for its performance and speed, XGBM is an optimized imple- mentation of gradient boosting. It includes several advanced features for handling missing data, regularization, and parallel processing, making it a popular choice for many competitive machine learning tasks.

LightGBM (LGBM): A highly efficient gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with large datasets. LightGBM grows trees leaf-wise (best-first), which can lead to better accuracy compared to the level-wise growth used in traditional algorithms.

CatBoost: A gradient boosting algorithm that handles categorical features natively without extensive preprocess- ing. CatBoost is known for its high performance and robustness, especially when dealing with datasets that have categorical variables.

Stacking

This ensemble method combines multiple base models to form a meta-model. The base models are trained on the entire dataset, and then their predictions are used as input features for the meta-model, which is trained to produce the final prediction. This approach leverages the strengths of various models to improve overall performance.

Datasets Used

Dataset 1: The Sylhet Diabetes dataset is a symptomatic dataset containing data on 521 patients with 16 attributes. It is a pre-stage dataset that indicates, based on symptoms, whether a patient may have diabetes or not. The dataset includes features such as age, gender, and symptoms like polyuria, polydipsia, sudden weakness, general weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. These features capture various patient characteristics and symptoms that may aid in medical diagnosis or analysis.

Dataset 2: The Frankfurt dataset is a diagnostic stage dataset containing 2001 patients' data with 8 attributes, and 683 of them have developed diabetes. The dataset includes

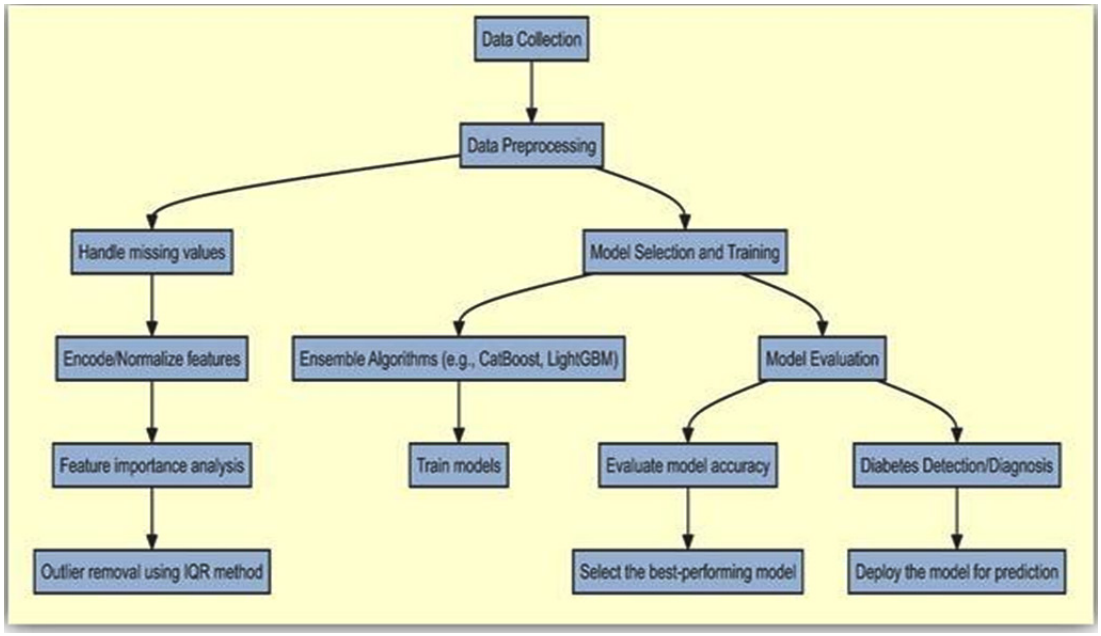

Fig. 1. Proposed Methodology for Diabetes Prediction

features such as pregnancies, skin thickness, diabetes pedigree function, glucose, insulin, age, blood pressure, and BMI. These features provide quantitative measurements used to assess and diagnose diabetes. Dataset Pre-processing: In the research, preprocessing was performed on the Sylhet and Frankfurt datasets by addressing missing values, normalizing numerical features, and encoding categorical variables. Then, numerical values were normalized in the Frankfurt dataset, and categories were encoded in the Sylhet dataset. Integration of key features from both datasets was done to ensure consistency and compatibility. Also, data visualization techniques were used, including box plots to identify outliers, heatmaps to understand correlations, and pair plots to uncover patterns. These steps improved data quality and reliability for subsequent analysis and model training.

Frankfurt dataset

Histogram

The histograms from the Frankfurt diabetes study highlight key insights, as shown in Fig. 2. The x-axis represents the range of values for each parameter, and the y-axis shows frequency. Most women reported no pregnancies, with fewer pregnancies as the number increased.

Glucose levels were mainly 100-150, and blood pressure readings were typically 60-80. Skin thickness generally ranged from 20-40 mm, and insulin levels were mostly 0-200. Diabetes pedigree function values were usually 0.0- 0.5, reflecting genetic risk. Most participants were aged 20-40, and there were more negative diabetes outcomes than positive ones.

Correlation Matrix

The correlation matrix reveals several notable observations as shown in Fig3. Positive correlations include blood pressure with age and BMI, skin thickness with age and BMI, and insulin with glucose, BMI, and age. Conversely, a negative correlation is observed between diabetes outcomes and pregnancies, indicating that individuals with diabetes tend to have fewer pregnancies.

Density plot

The density plots from the Frankfurt

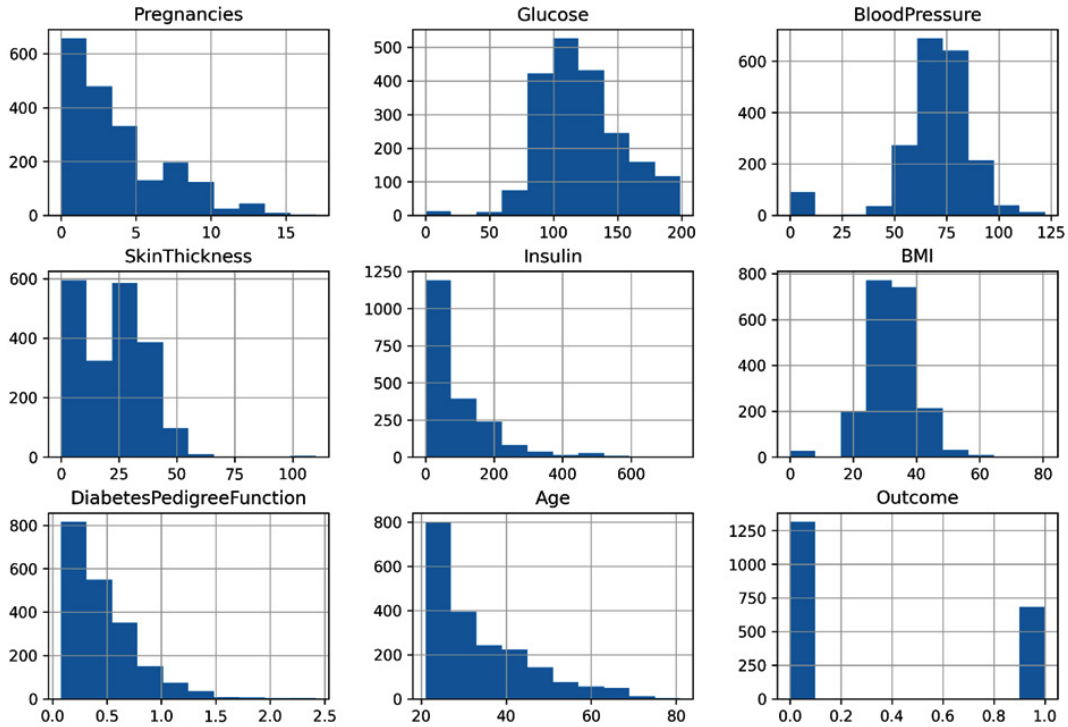

Fig. 2. Histogram

diabetes study reveal that most women had zero pregnancies, with a decreasing likelihood for higher numbers as shown in Fig. 4. The x-axis represents the range of values for each parameter, while the y-axis shows the density or probability of these values. Blood sugar levels were mainly between 100 and 150, and blood pressure commonly ranged from 60 to 80.

Skinfold thickness was typically 20-40 mm, and insulin levels were usually 0-200. Diabetes pedigree function scores were mostly 0.0-0.5, indicating a low genetic risk. Participants were primarily aged 25-50, with more women testing negative for diabetes than positive.

Box plot

The box plots from the Frankfurt diabetes study reveal key distribution patterns and potential outliers for various parameters, as shown in Fig. 5. Pregnancies show a right skew with higher-end outliers, while glucose levels suggest a normal distribution with outliers on both ends.

Blood pressure readings are slightly right-skewed with extreme values on both sides. Skin thickness appears normally distributed with high-end outliers, and insulin levels exhibit a right skew with higher outliers.

BMI is mostly centered but may lean towards higher values, with outliers on both ends. The diabetes pedigree function scores suggest a normal distribution with some higher-end outliers. Age shows a slight right skew with outliers at both extremes. The diabetes status box plot has limited data, making it hard to determine the distribution or outliers.

Removing outliers: Common outlier removal techniques include the Interquartile Range (IQR) method, which identifies outliers as data points outside Q1 - 1.5 IQR and Q3 + 1.5 IQR; the Standard Deviation method, defining outliers as those beyond a certain number of standard deviations from the mean, typically three; Grubbs' Test, a statistical test for outliers based on deviation

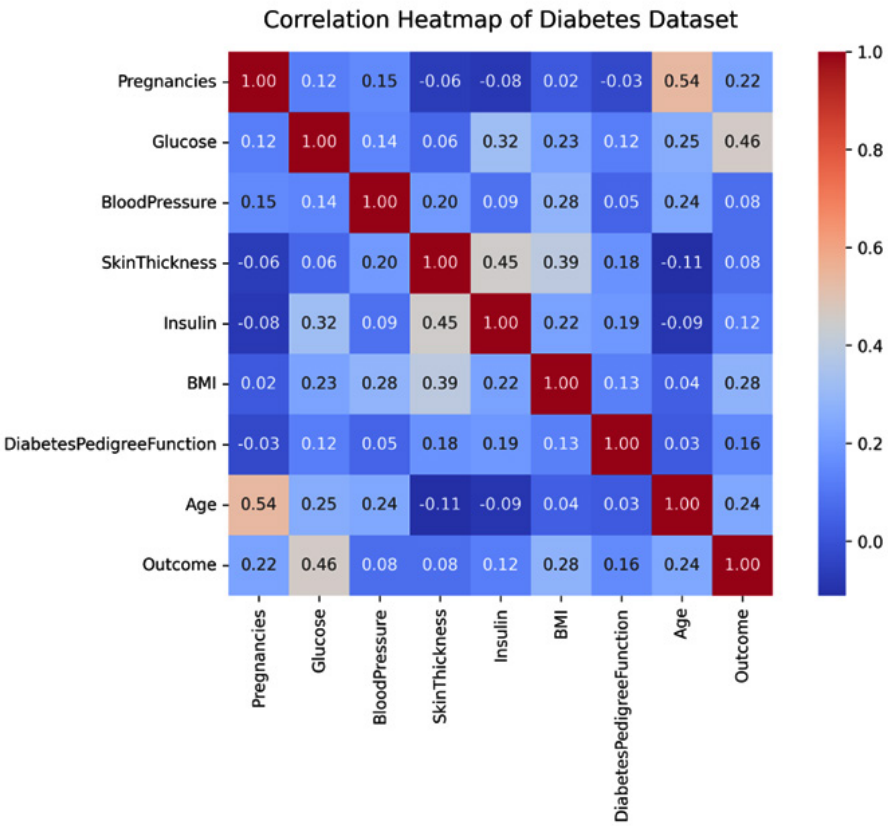

Fig. 3. Correlation Matrix

from the mean compared to the standard deviation; and Winsorization, which caps extreme values at the whisker ends instead of removing them, preserving potentially valuable information as shown in Fig.6.

Outcome

The dataset shows an imbalance as shown in Fig.7, with fewer data points for diabetes ("1") compared to non-diabetes ("0"). To address this, techniques like oversampling (e.g., SMOTE) create new minority class data, undersampling reduces majority class dominance, and cost-sensitive learning assigns higher weights to the minority class during training, enhancing diabetes classification.

Feature Importance

The bar chart in Fig.8 illustrates the importance of various factors in predicting diabetes.

Glucose emerges as the most critical factor, followed by BMI and age. Other variables, such as blood pressure and insulin levels, also contribute to the model but to a lesser extent.

Sylhet dataset

Correlation Matrix: A correlation matrix is a statistical tool that helps you understand the relationships between multiple variables at once about the diabetes dataset, as shown in Fig.9. The correlation matrix reveals strong positive correlations (dark red) between features like Polyuria and Weakness or Visual Blurring, indicating common underlying symptoms. Strong negative correlations (dark blue) are also noted, such as between Weakness and Obesity, suggesting conditions that lead to weight loss. Systematically examining these patterns can provide valuable insights into the relationships between various health measurements.

Strong negative correlations (dark blue) are also noted, such as between Weakness and Obesity, suggesting conditions that lead to weight loss. Systematically examining these patterns can provide valuable insights into the relationships between various health measurements.

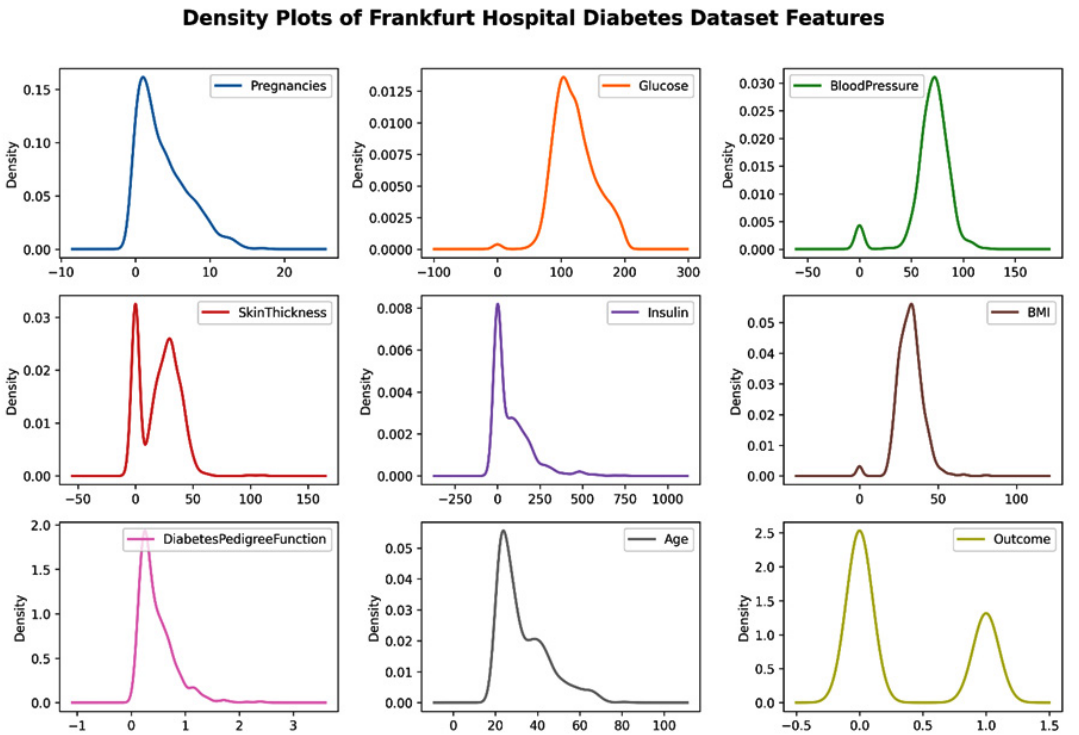

Fig. 4. Density Plot

Density Plot

In the density plot as shown in Fig. 10, the x-axis represents symptom intensity or levels, while the y-axis shows the frequency of occurrence. Patient ages range from 0 to 150, with a majority under 50 years old.

For symptoms such as polyuria, polyphagia, and alopecia, densities peak around values of 0.5, 1.0, and 2.0, indicating medium levels are most common. Sudden weight loss shows a density skewed towards the lower end, suggesting high prevalence. Weakness is widely distributed but peaks at higher levels. Genital itch and visual blurring are prevalent, with densities shifting towards positive values, while throat irritation has a broad distribution. Itching and irritability vary in intensity, with itching leaning towards higher values. Delayed healing is common, partial paresis is rare, and muscle stiffness has moderate density across the spectrum. Obesity patterns vary regionally. The "Class" feature reflects different diagnoses, with density patterns indicating the prevalence of various conditions.

Feature Importance: The analysis identifies BMI (Body Mass Index) as the most important feature, followed by Blood Glucose and Diabetes Pedigree Function, indicating their significant role in identifying individuals at risk for diabetes, as shown in Fig. 11.

Moderately important features include Age, Skin Thickness, Insulin, and Number of Pregnancies, which contribute to the model's predictive ability in combination with the more critical features. Although it is challenging to definitively determine the least important features, Plasma Insulin and Blood Pressure (Diastolic) appear to have relatively lower importance, suggesting a weaker contribution to the model's predictions compared to other features.

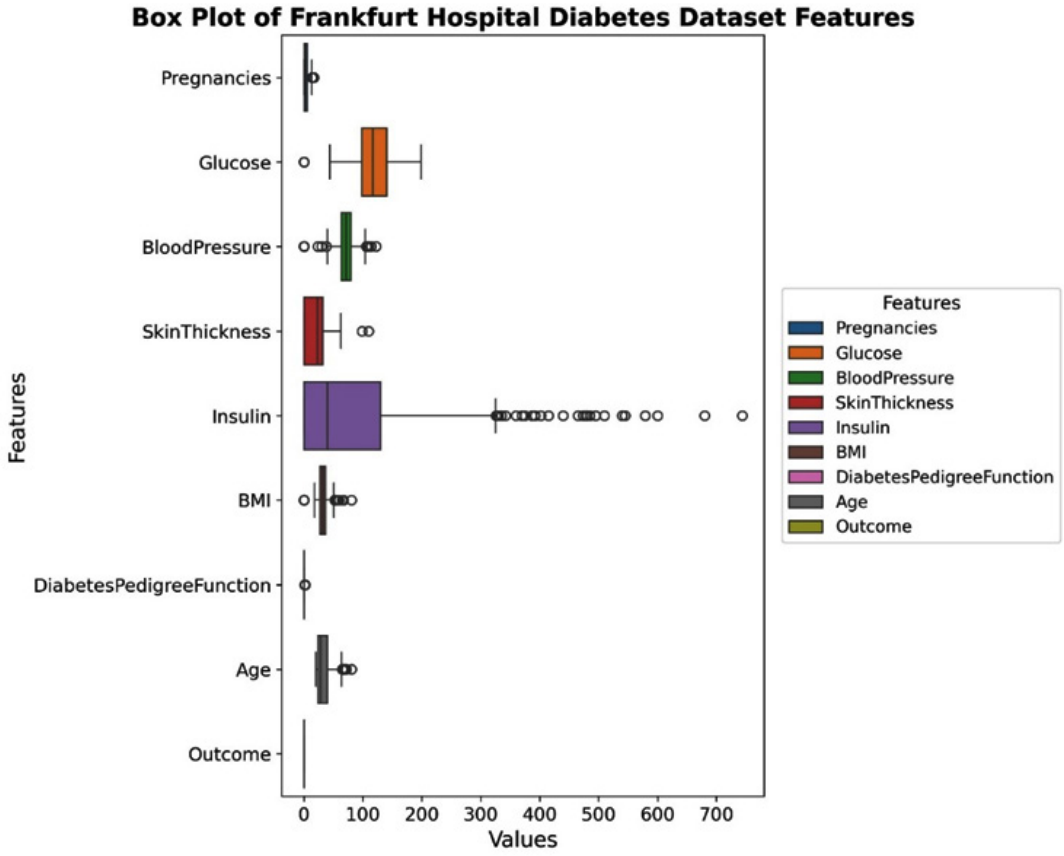

Fig. 5. Box Plot

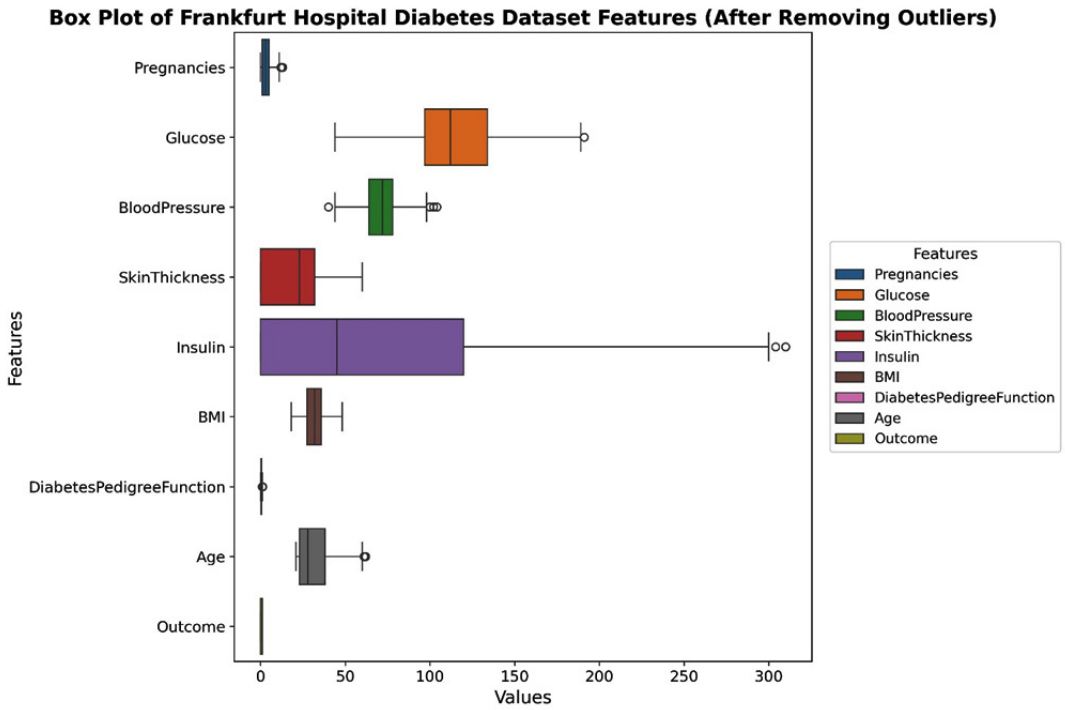

Fig. 6. After removing outliers using the interquartile range

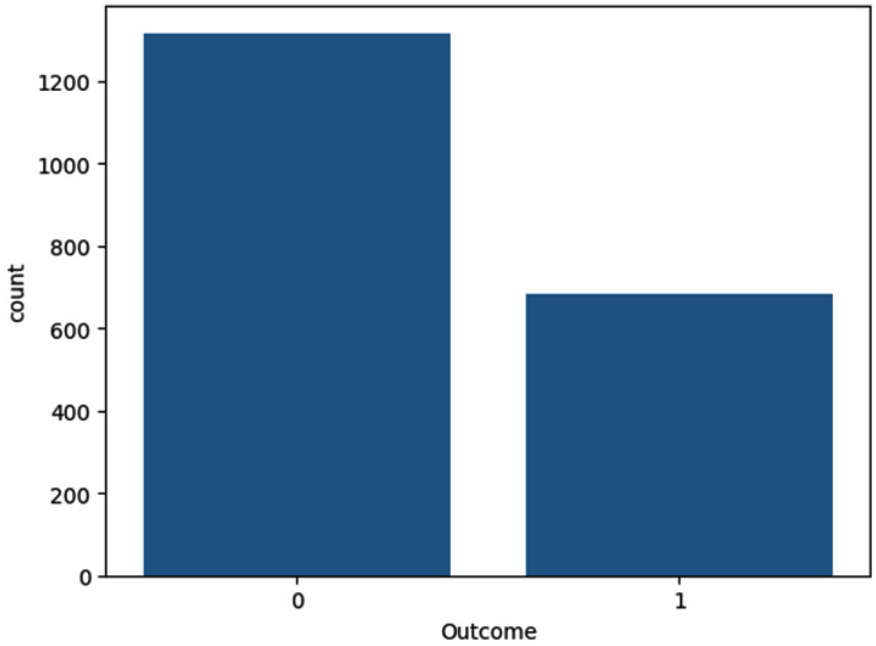

Fig. 7. Outcome

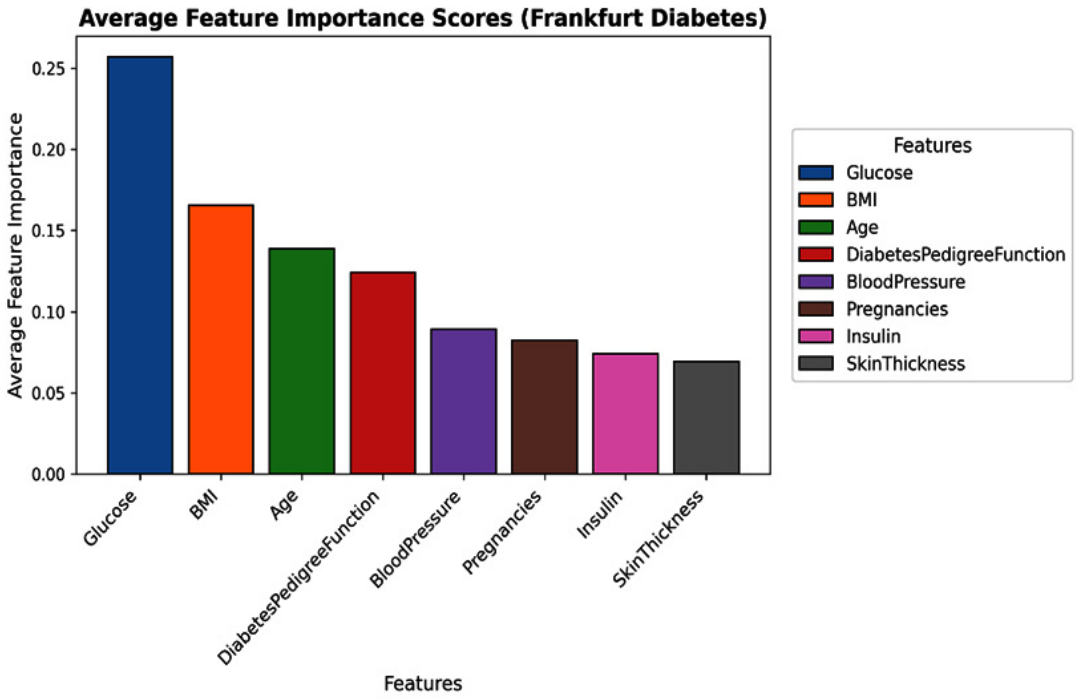

Fig. 8. Feature Importance

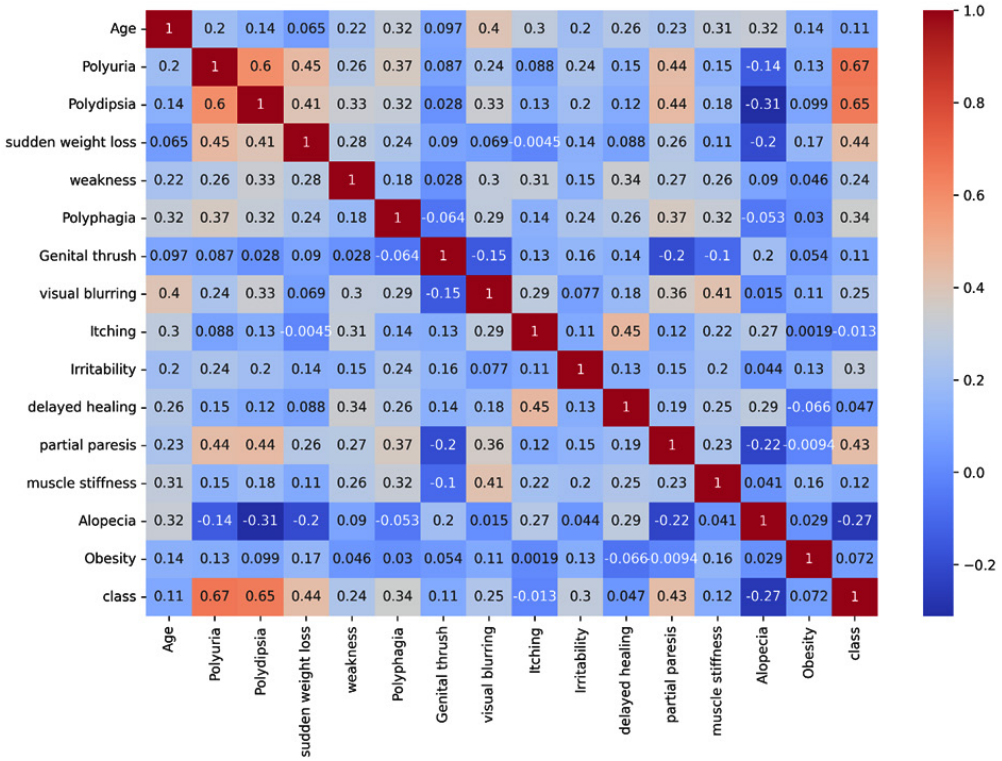

Fig. 9. Correlation Matrix

RESULTS

In this section, both datasets are discussed for diabetes prediction.

Frankfurt dataset

In the Frankfurt dataset, as shown in Fig.12, composed of numerical data, the highest accuracy was observed with the LightGBM algorithm, which achieved an accuracy of 98.75%. LightGBM's leaf-wise growth strategy and efficiency in handling large datasets enabled it to effectively capture complex patterns in the data, resulting in superior predictive accuracy and faster training times.

Sylhet dataset

The Sylhet dataset, consisting of categorical data, saw the highest accuracy with the CatBoost algorithm, which achieved an accuracy of 99.25%. CatBoost's advanced techniques, such as ordered boosting and efficient handling of categorical variables, enabled it to natively process categorical features with extensive preprocessing as shown in Fig. 13.

DISCUSSION

As shown in Table 1, comparing the ensemble models used in the analysis, Bagging techniques—especially the generic Bagging method with 95.75% accuracy—demonstrated strong performance in reducing variance by aggregating multiple learners. Random Forest (93%) and Extra Trees (91.33%) followed suit but showed slightly lower accuracy, possibly due

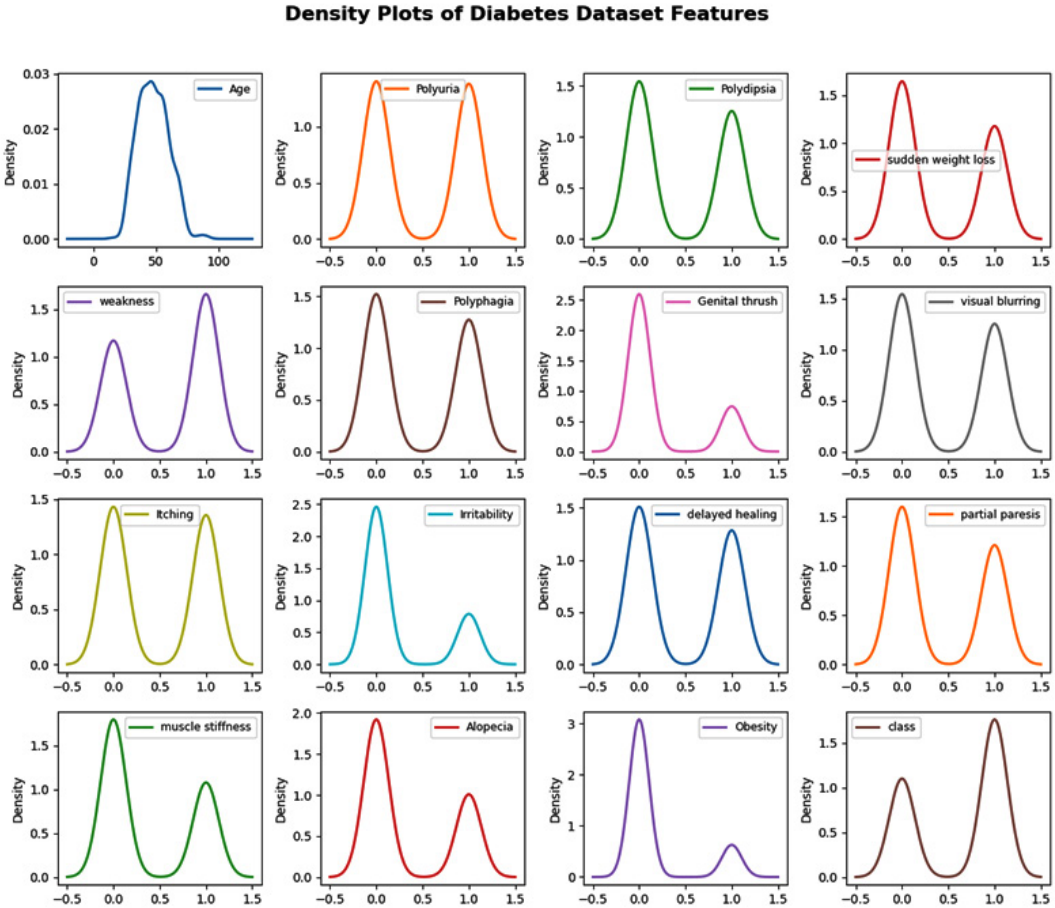

Fig. 10. Density Plot

to the way they introduce randomness. Boosting models, which aim to reduce bias, delivered varied results: while AdaBoost (90.75%), GBM (92%), and XGBoost (90.25%) offered solid predictive power, they were outperformed by LightGBM (98.75%), which emerged as the most accurate due to its highly efficient gradient boosting framework. CatBoost, though tailored for categorical features, lagged at 88.25%, indicating it might not have aligned optimally with the dataset. Overall, LightGBM's dominance suggests it best captured the underlying patterns while maintaining speed and scalability.

In the tabe 2, the performance comparison of Bagging stood out with the highest accuracy of 96.33%, outperforming both traditional Bagging (94.75%) and Random Forest (94%), suggesting that its extra randomness in split selection paid

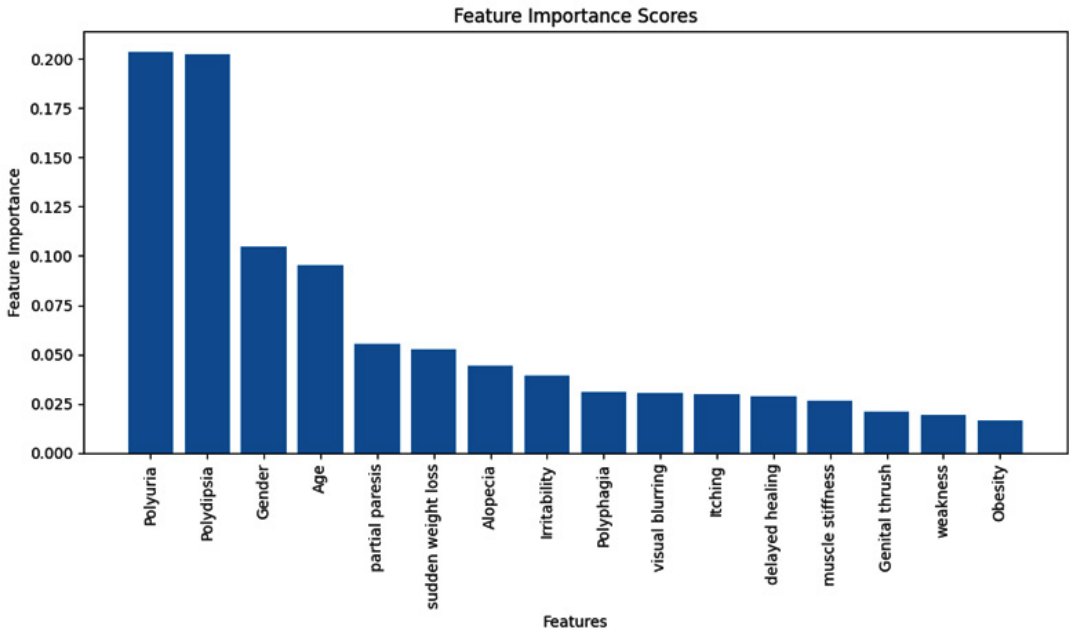

Fig. 11. Feature Importance

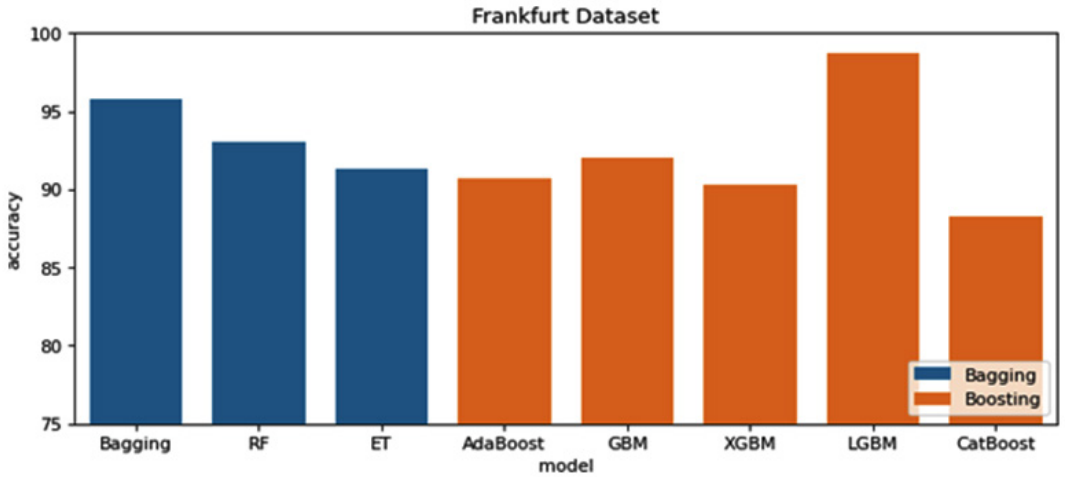

Fig. 12. Comparison of models for Frankfurt Dataset

off. Among boosting models, CatBoost achieved the top accuracy at 99.25%, followed closely by the state-of-the-art [26 27 28] Random Forest with 99%, indicating a unique configuration or enhancement in the boosting process. LightGBM (93.75%) and AdaBoost (92.75%) provided competitive but slightly lower results, while GBM (90%) and XGBoost (90.25%) trailed behind. The fields of artificial intelligence and healthcare are coming together, and using machine models to predict and diagnose diabetes is a huge trend [29, 30, 31]

It is recommended that statistical analysis, like what is done with Statistical Significance Testing SPSS, be added to the algorithm's results and findings to make them more reliable. In the analysis of the Sylhet dataset, a paired t-test showed that CatBoost's accuracy (99.25%) was statistically significantly better than Extra Trees' accuracy (96.33%) ($p < 0.01$)." LightGBM (98.75%) was also found to be statistically better than the standard Bagging method (95.75%) on the Frankfurt dataset ($p < 0.05$).

Overall, this comparison highlights that while both Bagging and Boosting techniques offer robust performance, the Boosting variants—especially CatBoost and the augmented Random Forest—delivered superior accuracy in this particular setup, likely due to better handling of data complexity or optimized parameters. This capability allowed CatBoost to effectively capture complex patterns in the data, resulting in superior predictive accuracy and robustness as compared to the state-of-the-art method. [26 27 28]

Table 1. Comparison of Bagging and Boosting Methods

Method	Model	Accuracy (%)
Bagging	Bagging	95.75
Bagging	RF	93
Bagging	ET	91.33
Boosting	AdaBoost	90.75
Boosting	GBM	92
Boosting	XGBM	90.25
Boosting	LGBM	98.75
Boosting	CatBoost	88.25

Table 2. Comparison of Bagging and Boosting Methods

Method	Model	Accuracy (%)
Bagging	Bagging	94.75
Bagging	RF	94
Bagging	ET	96.33
Boosting	AdaBoost	92.75
Boosting	GBM	90
Boosting	XGBM	90.25
Boosting	LGBM	93.75
Boosting[23]	Random Forest	99
Boosting	CatBoost	99.25

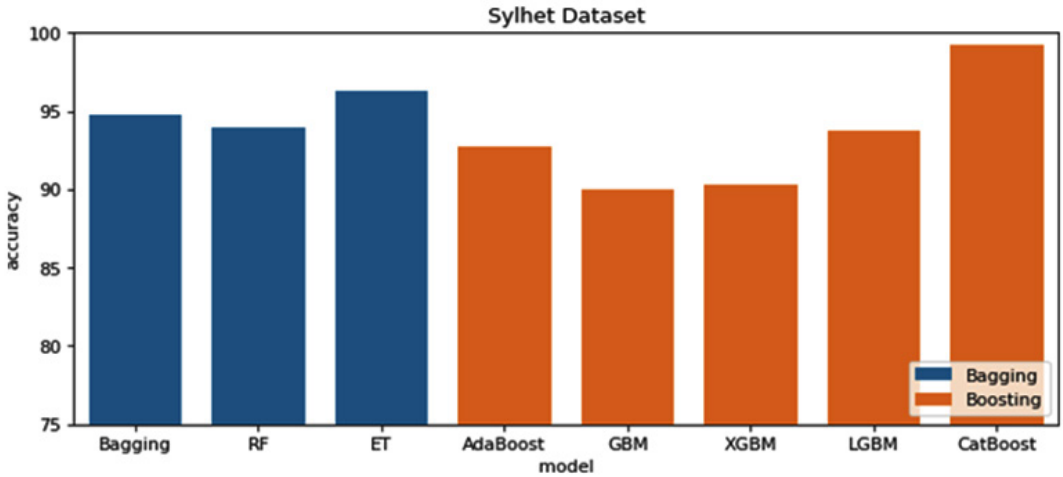

Fig. 13. Comparison of models for the Sylhet Dataset

CONCLUSION

Early detection of diabetes is crucial as it significantly prevents the progression of the disease and its associated complications. By identifying early-stage diabetes, timely interventions and lifestyle changes can be implemented to manage blood sugar levels and reduce the risk of severe health issues. In this research, two distinct datasets were utilized for diabetes detection: the Sylhet dataset, which consists of categorical data for prediabetes detection, and the Frankfurt dataset, composed of numerical values for further diabetes diagnosis and confirmation. Both datasets underwent thorough preprocessing, including feature importance analysis to identify highly correlated parameters and outlier removal using the Interquartile Range (IQR) method. The diabetes detection process was enhanced through the use of ensemble learning algorithms.

The findings demonstrate that CatBoost performed exceptionally well on the Sylhet dataset with an accuracy of 99.25% due to its advanced techniques, such as ordered boosting and efficient handling of categorical variables. Similarly, LightGBM exhibited superior performance on the Frankfurt dataset with 98.75% accuracy, leveraging its leaf-wise growth strategy to capture complex patterns in the numerical data effectively. These ensemble learning techniques not only improved the accuracy of predictions but also showcased their robustness and efficiency compared to traditional methods. The advantages of ensemble learning were evident in the study. Bagging was used to reduce variance, while boosting methods helped minimize bias. By combining multiple models, the ensemble methods provided a significant increase in prediction accuracy. The comparisons made in this study underline the importance of selecting the appropriate algorithm based on the nature of the dataset. The results from this study indicate that ensemble classifiers are superior to individual classifiers. The ensemble-based approaches, such as Catboost and LightGBM, significantly enhance the accuracy of diabetes detection. Future work could explore the integration of additional datasets, further optimization of models, and real-world applications to enhance the robustness and applicability of these findings.

ACKNOWLEDGEMENT

The author is extremely grateful to the Department of E&TC Engineering at the MIT Academy of Engineering in Pune, India, for granting permission to conduct this research. The author is also deeply appreciative of the Department of Semiconductor Engineering at the School of Engineering, Management, and Research, D. Y.

Funding Source

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The author(s) do not have any conflict of interest.

Data Availability

The following two datasets were employed in the research experiments of this article: This was gathered through direct questionnaires administered to patients of the Sylhet Diabetes Hospital in Sylhet, Bangladesh, and subsequently approved by a physician. The Frankfurt dataset was compiled from 2000 individuals at the Frankfurt Hospital in Germany.

Ethics Statement

This research did not involve human participants, animal subjects, or any material that requires ethical approval.

Informed Consent Statement

This study did not involve human participants, and therefore, informed consent was not required.

Clinical Trial Registration

This research does not involve any clinical trials.

Permission to reproduce material from other sources

Not Applicable.

Author Contributions

Smita Kulkarni: Research Framework, Methodological Approach, Manuscript Drafting; Dnyanda Hire: Data Collection; Data Evaluation; Priya Charles: Reviewing, and Data Editing; Shweta Suryawanshi: Support & Facilities, Supervision of Work.

REFERENCES

1. Ansari A, Kadam B, Barve S, Chikmurge D. Enhanced biased weights Adaboost algorithm for diabetes detection on imbalanced dataset. *14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. July 2023:1-6.
2. Singh D, Khandelwal A, Bhandari P, Barve S, Chikmurge D. Predicting lung cancer using XGBoost and other ensemble learning models. *14th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2023:1-6. Available from: https://api.semanticscholar.org/CorpusID:265407215
3. Liu, Yan, Zhiyu Zhang, Huazhu Song, Renjie Li, and Kaituo Mi. An improved stacking model for predicting myocardial infarction risk in imbalanced data. *Health Information Science and Systems*. 2025: 13, no. 1 : 16.
4. Shen, Kanghui, Jing Liu, and Jingya Li. Research on Multi-Label Disease Classification Based on the LGB-CatBoost Model. *In 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE). IEEE* 2025: pp. 2208-2212.
5. Katlariwala, Soham Biren, Vaibhav C. Gandhi, Nirav Patel, Devendra Parmar, and Ayushi Desai. TriBoost and Beyond: Advanced Machine Learning Approaches for Diabetes Risk Prediction. *In 2025 International Conference on Electronics and Renewable Systems (ICEARS),* IEEE, 2025: pp. 1780-1785.
6. Nagulpelli S, Chavan A, Kandalkar A, Kulkarni S. AI-based health management system. In: Kulkarni AJ, Mirjalili S, Udgata SK, eds. *Intelligent Systems and Applications*. Singapore: Springer Nature Singapore; 2023:379-389.
7. Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J King Saud Univ Comput Inf Sci*. 2023:35(2):757-774.
8. Reddy G, Bhattacharya S, Ramakrishnan SS, et al. An ensemble-based machine learning model for diabetic retinopathy classification. 2020.
9. Aamir Z, Murtza I. Pre-diabetic diagnosis from habitual and medical features using ensemble classification. *J Comput Biomed Inform*. 2023;5(1):283-294. Available from: https://www.jcbi.org/index.php/Main/article/view/205
10. Kaul K, Tarr JM, Ahmad SI, Kohner EM, Chibber R. *Introduction to Diabetes Mellitus*. New York, NY: Springer New York; 2013:1-11.
11. Nnamoko N, Hussain A, England D. Predicting diabetes onset: An ensemble supervised learning approach. *2018 IEEE Congress on Evolutionary Computation (CEC)*. 2018:1-7.
12. Gunasekar G, Prasad K. An hybrid ensemble machine learning approach to predict type 2 diabetes mellitus. *Webology*. 2021:18:311-331.
13. Saihood Q, Sonuç E. A practical framework for early detection of diabetes using ensemble machine learning models. *Turk J Electr Eng Comput Sci*. 2023:31:722-738.
14. Alam U, Asghar O, Azmi S, Malik RA. Chapter 15 – general aspects of diabetes mellitus. In: Zochodne DW, Malik RA, eds. *Diabetes and the Nervous System. Handbook of Clinical Neurology*. Vol 126. Elsevier; 2014:211-222.
15. Machoke M, Mbelwa J, Agbinya J, Sam A. Performance comparison of ensemble learning and supervised algorithms in classifying multi-label network traffic flow. *Eng Technol Appl Sci Res*. 2022:12: 8667-8674.
16. R K, Geetha P, E R, Ar K. Prediction of diabetes and cholesterol diseases based on ensemble learning techniques. 2022: 9:491.
17. Mandal S, Mandal JK. An ensemble of LightGBM and AdaBoost for Type-2 diabetes prediction. *Int J Comput Intell Syst*. 2023:16(1):184.
18. Patel N, Shah R. Ensemble learning with boosting techniques for diabetes prediction. *Front Genet*. 2023:14:1252159.
19. Gupta M, Sinha A. Prediction of diabetes using diverse ensemble classifiers. *Procedia Comput Sci*. 2024: 225:200-209.
20. Reddy TK, Kumar V. Optimizing diabetes prediction by handling data imbalance using SMOTE and ensemble models. *Appl Comput Syst*. 2024:5(2):45-56.
21. Yadav R, Singh A. CatBoost ensemble approach for early-stage risk prediction of diabetes. *IEEE Conference on Artificial Intelligence in Healthcare*. 2023:88-93.
22. Sharma D, Mehta P. Early and advanced-stage diagnosis of diabetes using a two-stage boosting model with CatBoost and LightGBM.. *IEEE International Conference on Smart Health Analytics*. 2025.
23. Feng Q, Li Z. Enhanced bagging ensemble for time-series prediction of diabetes trends. *arXiv preprint. 2025. Available from: https://arxiv.org/abs/2506.13786*
24. Abdulrahman M, et al. A systematic review of AI techniques in diabetes prediction and diagnosis. *arXiv preprint. 2024. Available from: https://arxiv.org/abs/2412.14736*
25. Chen Y, Zhao L. CatBoost-based insulin resistance prediction using multi-feature clinical datasets. *arXiv preprint. 2025. Available from: https://arxiv.org/abs/2503.05119*

26. Vakil V, Pachchigar S, Chavda C, Soni S. Explainable predictions of different machine learning algorithms used to predict early-stage diabetes. *arXiv preprint. 2021. Available from: https://arxiv.org/abs/2111.09939*
27. Zhang, Wangyouchen, Zhenhua Xia, Guoqing Cai, Junhao Wang, and Xutao Dong. Enhancing diabetes risk prediction through focal active learning and machine learning models. 2025: vlo 20, no. 7: e0327120.
28. Wu, Jingjing, Qingqing Zeng, Sijie Gui, Zhuolan Li, Wanyu Miao, Mi Zeng, Manyi Wang, Li Hu, and Guqing Zeng. Construction and evaluation of prediction model for postoperative re-fractures in elderly patients with hip fractures. *International Journal of Medical Informatics* 2025: 195: 105738.
29. Sun, Qi, Xin Cheng, Kuo Han, Yichao Sun, He Ren, and Ping Li. Machine learning-based assessment of diabetes risk. *Applied Intelligence*: 2025:55, no. 2: 106.
30. Das, Dip, Sourav Kumar, Md Amir Hussain, and B. Ramachandra Reddy. Diabetes Prediction using Ensemble Learning Techniques. *Procedia Computer Science*, 2025: 258: 3155-3164.
31. Jaiswal, Amit Kumar, Nirmalya Basu, and Ajay Pal Singh. Exploring AI Classifiers to Identify Gestational Diabetes Mellitus During Pregnancy. *In International Conference on Data Mining and Information Security, Springer Nature Singapore*, 2024; 3-11.