Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 13s (2025)
https://www.jneonatalsurg.com

Integrating RF-GBDT: Optimizing Machine Learning Techniques for Diabetic Prediction Model

N. P. Jayasri[1], R. Aruna[2], S. Ravikumar[3], T. Thilagam[4]

[1]Research Scholar, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India

[2]Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India

[3]Associate Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India

[4]Assistant Professor(SG), Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India

*Corresponding Author:

Emial ID: thilaka28@gmail.com

.Cite this paper as: N. P. Jayasri, R. Aruna, S. Ravikumar, T. Thilagam, (2025) Integrating RF-GBDT: Optimizing Machine Learning Techniques for Diabetic Prediction Model. *Journal of Neonatal Surgery*, 14 (13s), 946-957.

ABSTRACT

Diabetes is a persistent medical condition that has seen a significant rise in incidence in recent years. Because precise datasets are necessary for early prognosis, this presents complications. Big data plays a significant role in predicting diabetes by examining enormous volumes of health-related information. Through advanced analytics, algorithms can identify patterns, risk factors, and correlations in data such as patient demographics, medical history, genetic markers, lifestyle factors, and biomarkers. This research introduces a diabetes prediction model, RF-GBDT classifier, tailored to identify potentially effective peptides against diabetes. Combining sequence data with the Random Forest – Optimized Gradient Boosting Decision Trees (GBDT) framework, RF-GBDT aims to improve the accuracy of antidiabetic peptide prediction. Results demonstrate the model's remarkable performance with an accuracy of 99.8% and an AUC of 95.2%. Furthermore, feature selection techniques streamline prediction times without compromising classifier accuracy. These findings, comparative to existing studies, affirm the efficacy of the proposed method, positioning it as a valuable adjunctive tool in diabetes diagnosis.

Keywords: *Diabetes diagnostics, artificial intelligence, machine learning, and big data analytics*

1. INTRODUCTION

In recent days, advancements in diabetes management have focused on personalized treatment plans and precision medicine tailored to individual patient needs. Furthermore, utilizing technology for real-time monitoring and predictive analytics in diabetes care such as artificial intelligence and continuous glucose monitoring systems has become more and more important [1]. In the healthcare industry, big data analytics is essential since it analyzes enormous volumes of patient data to find trends, risk factors, and connections pertaining to diabetes. Healthcare practitioners can create predictive models that precisely predict a person's risk of developing diabetes by utilizing advanced analytics techniques. This allows for proactive interventions and individualized treatment plans to lessen the effects of the disease and enhance patient outcomes [2].

Diabetes, a long-term illness marked by high blood sugar levels brought on by inadequate insulin synthesis or response, poses significant health risks, including nerve damage, heart disease, kidney failure, and stroke. While incurable, early awareness and intervention can mitigate its impact. Hyperglycemia, a common consequence of diabetes, adversely affects the cardiovascular system and impairs the function of vital organs such as the eyes, kidneys, and nerves. Early detection and management are crucial for preventing complications and maintaining overall health [19].

Manual diagnosis in healthcare is prone to inaccuracies, necessitating advanced automated systems for early disease detection. Data mining and machine learning algorithms offer efficient solutions for uncovering hidden patterns and improving diagnostic accuracy. The growing impact of diabetes has spurred the development of algorithms for extracting insights from large healthcare datasets, facilitating automated prediction and feature selection [19][20] [21].

Large datasets are used to create precise models that predict a person's risk of getting diabetes using machine learning methods, which are increasingly being applied to the prediction of diabetes [3]. This helps with early intervention and tailored treatment.

The major objectives of this work are depicted here.

- This article introduces a machine learning-based approach utilizing the PIMA diabetic dataset to address the pressing need for early diabetes detection in India.
- Evaluate the efficacy of the RF-GBDT classifier in predicting diabetes onset by leveraging big data analytics and diverse health-related datasets.
- Investigate the potential of RF-GBDT in identifying peptides with antidiabetic properties through the integration of sequence data and advanced machine learning techniques.
- Assess the impact of feature selection methods on prediction accuracy and computational efficiency within the RF-GBDT framework, highlighting its applicability as a valuable tool for early diabetes prognosis.

The remaining sections of this work are as follows: Section I provides the background in terms of the diabetes detection and prediction. The literature works based on diabetes prediction is shown in Section II. Section III defines material and methods for diabetic prediction. The implementation of prediction model is discussed in Section IV. The outcome and their discussion are shown in Section V. The paper is concluded in Section VI.

2. RELATED WORKS

Numerous techniques for diabetes prediction have been reported and proposed in recent years. Using the PIMA dataset, this study [4] uses machine learning to predict diabetes. Studies created to predict diabetes are typically grounded on deep learning or machine learning. a type 2 diabetes e-diagnosis solution for IoMT that uses interpretable machine learning models. It addresses trust difficulties in machine learning by evaluating J48, random forest, and Naïve Bayes decision tree models using the Pima Indians dataset. It concludes that random forest performs better with a larger feature set and that Naïve Bayes is more effective with finer features for binary classification.

A machine learning ensemble for predicting diabetes, integrating various methods like LnR, LR, KNN, NB, RF, SVM, and DT. Pre-processing steps including null removal, normalization, and label encoding were applied before evaluating on imbalanced datasets, mitigated through SMOTE. Results highlight random forest's superior performance, achieving 97% accuracy on a 2019 diabetes dataset and 80% on Pima Indian dataset, with notable improvements in false-negative detection rates on balanced data [5].

A diabetes prediction model using machine learning methods assessed by precision, recall, and F1-measure is presented in this paper [6]. LR, NB, and KNN obtained accuracies of 90%, 79%, and 69%, respectively, using the PIDD dataset. LR outperforms other algorithms in terms of accuracy and emerges as the best efficient algorithm for diabetes prediction.

This study [7] introduced a novel approach employing deep learning for early diabetes detection. By transforming numerical data from the PIMA dataset into images based on feature importance, CNN models like ResNet18 and ResNet50 are utilized. Three classification strategies, including direct CNN classification, fusion of ResNet deep features with SVM, and direct SVM classification of fusion features, highlight the efficacy of diabetes image representation in early diagnosis.

This work [8] uses the Pima Indian dataset in conjunction using a confidential dataset of Bangladeshi women patients to build a self-contained diabetes prediction system. Using mutual information for feature selection, class imbalance is addressed using an XGBoost classifier that uses the ADASYN algorithm in conjunction with SMOTE. It achieves 81% accuracy, 0.81 F1 coefficient, and an AUC of 0.84. Moreover, domain adaption strategies and explainable AI approaches like SHAP and LIME are combined to create a flexible system that is accessed via an Android smartphone application and a website framework.

The goal of this work [9] is to build use a classifier model to forecast diabetes the WEKA tool. The study makes use of Naive Bayes, Support Vector Machine, Random Forest, and the Simple CART technique. Based on performance outcomes, finding the most effective algorithm for predicting diabetes disease is the aim. The experimental results of each method on the dataset were analyzed in detail. Notably, Support Vector Machine outperformed the others and predicted the illness with the highest accuracy possible.

This study [10] uses data mining techniques to present a diabetes prediction model. Naive Bayes, Support Vector Machine (SVM), Random Forest, and Logistic Regression are the four methods that are employed. The model is trained using Python, and it is evaluated using an actual dataset from Kaggle. Performance analysis includes sensitivity analysis, accuracy metrics, and confusion matrix evaluation. Notably, compared to other data mining approaches, logistic regression shows a high accuracy of 82.46%.

Recent literature explores diabetes prediction using the Pima Indian Diabetes (PID) dataset from the UCI ML Repository, encompassing data from 768 patients and nine attributes. Various machine learning (ML) algorithms, including Logistic Regression (LR) and Support Vector Machine (SVM), have been investigated for their efficacy. Notably, LR and SVM models exhibit promising performance in diabetes prognosis. Additionally, neural network (NN) models with dual hidden layers, optimized through varying epochs, attain notable accuracy, exemplified by an 88.6% success rate [16].

This [17] study addresses diabetes prediction using machine learning on the Pima Indians dataset and proprietary data, employing a semisupervised model with gradient boosting and SMOTE for class imbalance. Among ten evaluated algorithms, XGBoost with SMOTE achieved notable accuracy (97.4% on private data, 83.1% on combined datasets). Explainable AI techniques like SHAP and a mobile app for instant predictions enhance accessibility, providing innovative insights for early diabetes detection and management in Saudi Arabia.

Highlighting the critical need for accurate diabetes prediction given its widespread impact and substantial healthcare expenses, [18] utilizes ML methods, including a semisupervised model and SMOTE, achieving high accuracies (97.4% and 83.1% on private and combined datasets). Employing explainable AI techniques like SHAP enhances interpretability, and a mobile app aids accessible diabetes prediction, providing valuable insights for early detection and management, especially in Saudi Arabia, advancing ML-driven diabetic prognosis.

In order to forecast the emergence of diabetes in the future, this inductive study [19] investigates the association between individual MetS risk variables and diabetes mellitus in a non-conservative scenario. The study refutes the widely held notion that low HDL levels are associated with diabetes by using logistic regression analysis to show a positive connection, especially in women. The study also suggests using Naïve Bayes and J48 decision trees to forecast the onset of diabetes. With a 79% ROC performance, Naïve Bayes with K-medoids under-sampling is shown to be the most successful of them. These results point to the need for more research on the pathophysiological importance of HDL and how it contributes to the onset of diabetes.

3. MATERIALS AND METHODS

The operational processes and utilizing many machine learning methods to developing the suggested autonomous diabetes prediction system are outlined in this section. Figure 1 presents the various phases of this work. In order to handle unbalanced class concerns and correct any relevant disparities, the dataset was first collected and preprocessed. This included replacing null occurrences with mean values. The dataset was then split using the holdout validation procedure into the training and test sets. After that, a range of classification methods were used in order to determine which classification algorithm would work best for this dataset. The step-by-step process of the RF-GBDT Diabetic Prediction model is shown in Figure 1.

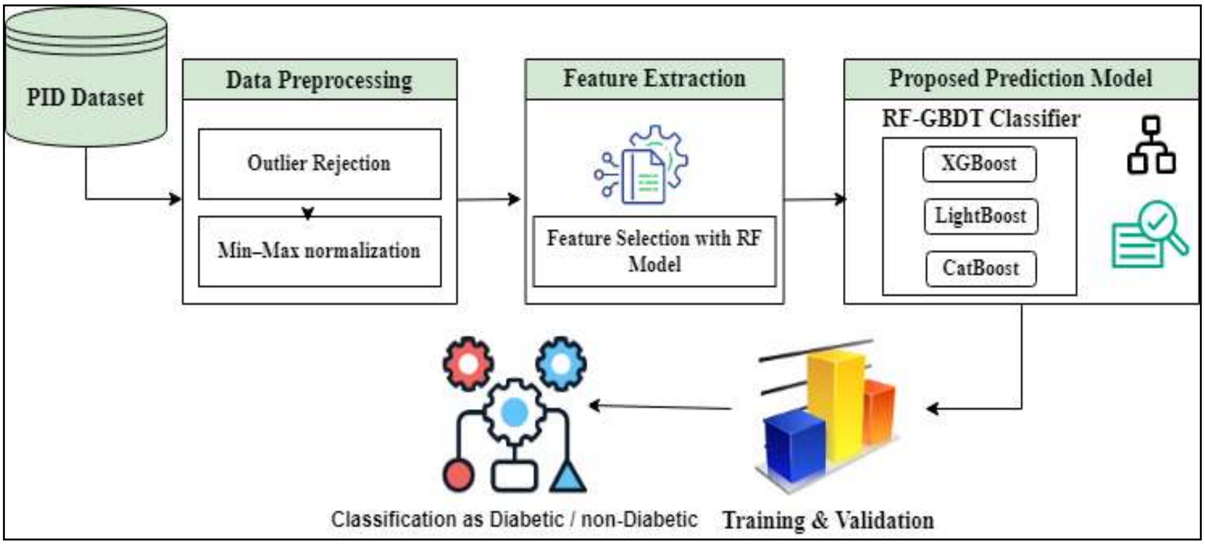

Figure 1 The architecture of the proposed system

Dataset

The Pima Indian dataset comprises medical data aimed at studying diabetes prevalence among the Pima Indian community in Phoenix, Arizona. It contains 768 records with 8 numerical attributes, including glucose levels, blood pressure, and body mass index, alongside a binary variable indicating the presence of diabetes. This dataset is widely used in machine learning and statistical modeling for binary classification tasks to predict the onset of diabetes based on the provided medical attributes. Figure 2 displays the correlation between characteristics from the PID dataset as a Heapmap.

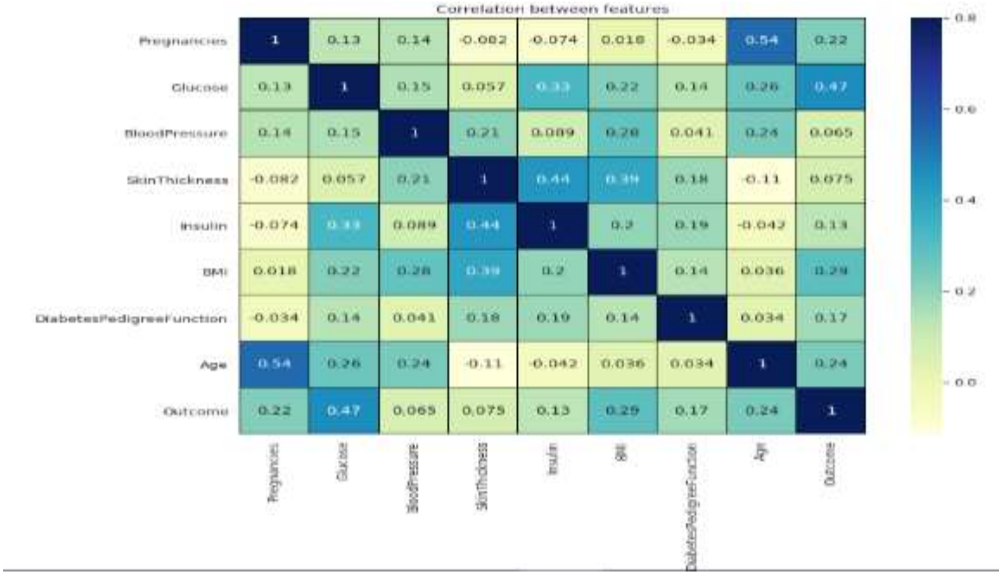

Figure 2 Correlation among the attributes of the diabetes dataset.

Data Preprocessing

It is typical practice in most data analysis studies to substitute the mean of the related characteristic (the column mean value for tabular data, for example) for missing values. Nevertheless, it is more efficient to use the median of the related attributes for small datasets rather than the mean [12]. As a result, the median was employed in this study as a fundamental method to fill in the missing feature values. The median was computed independently for patients with and without diabetes to guarantee that the substituted values were more representative.

To mitigate the impact of varying intervals among feature values on the machine learning model, the features are normalized utilizing the Min-Max Normalization technique [13]. This process ensures that all feature values are bounded within the range of 0 and 1. The data is scaled to a uniform range using the following equation:

$$X_{normalized} = \frac{\mathrm{X} - X_{min}}{X_{max} - X_{min}}$$

where X_{max} and X_{min} stand for the highest and lowest values, respectively, in each feature column. The histograms of the diabetic dataset's characteristics are displayed in Figure 3.

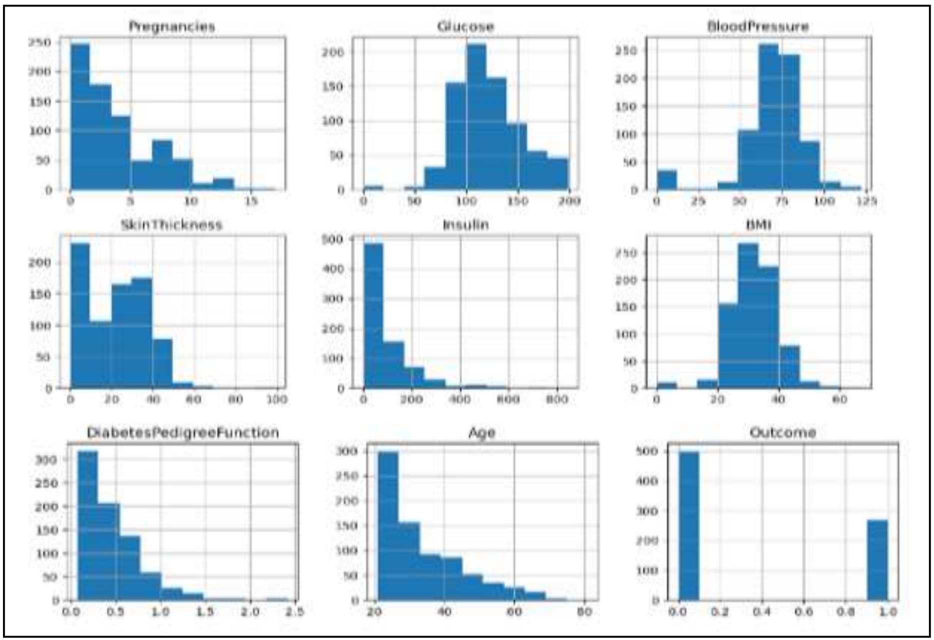

Figure 3 Feature histograms from the diabetes dataset.

Feature Selection

Multiple decision trees are combined in Random Forest (RF), an ensemble learning approach that improves classification performance. A random selection of characteristics is chosen for splitting at each node of a decision tree in RF, which is trained using a bootstrap sample of the dataset. The total of all the decision trees' predictions is used to produce the final forecast [22].

The probability p of selecting a feature at each node in a decision tree within the RF algorithm is determined by:

$$\mathrm{p} = \frac{1}{\sqrt{m}}$$

where m is the total number of features. This simple probability-based feature selection mechanism ensures that each tree in the forest uses a diverse set of features, reducing the correlation between trees and improving overall model performance [23].

The Random Forest (RF) approach was created by Breiman (2001) based on the concept of building numerous decision trees using random feature subspaces and subsets of sample data. RF improves generalization performance by averaging these distinct trees' predictions, which lowers variance.

RF is known for its scalability to large datasets and robustness to irrelevant features. Its ability to effectively handle multidimensional feature spaces makes it suitable for a wide range of classification tasks, while its resistance to irrelevant features helps maintain highperformance even in noisy datasets[24]. The Table 1 describes the count, mean, std, min and max value of each variable using RF.

Table 1: Description of count, mean, std, min and max value of each variable using RF

Statistic	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
Pregnancies	768	3.85	3.37	0	1	3	6	17
Glucose	768	120.89	31.97	0	99	117	140.25	199
Blood Pressure	768	69.11	19.36	0	62	72	80	122
Skin Thickness	768	20.54	15.95	0	0	23	32	99
Insulin	768	79.8	115.24	0	0	30.5	127.25	846
BMI	768	31.99	7.88	0	27.3	32	36.6	67.1
Diabetes Pedigree Function	768	0.47	0.33	0.08	0.24	0.37	0.63	2.42
Age	768	33.24	11.76	21	24	29	41	81
Outcome	768	0.35	0.48	0	0	0	1	1

Prediction model

The prediction models used in this study are explained in more detail in this section. In order to create models that forecast the likelihood of diabetes development, we tackled the problem of estimating the conditional probability. $\mathcal{X}$ will represent the input space, and Y will represent the output space, with $\mathcal{Y} = \{0,1\}$. The dimension of the input space, d, corresponds to the number of input variables, $\mathcal{X}$.

Assumed to be the training data are:$\mathcal{D}_{\text{train}} = \{(\vec{x}_i, y_i)\}_{i=1}^{N}$The independent generators of originate from the same joint probability distribution (which is unknown)$p(\vec{x}, y)$. Making use of the training data, our goal is to build models to estimate the conditional probability$p(y = 1 \mid \vec{x})$. Our prediction models were DP-GBDT models.

Gradient Boosting Decision Tree: Gradient Boosting Decision Tree (GBDT) is a popular machine learning algorithm renowned for its effectiveness. By gradually adding weak learners, it functions as an ensemble learning technique that repeatedly reduces the loss function. GBDT uses decision trees $T(\vec{x}; \vec{\theta})$ as its weak learners to do this.

$$T(\vec{x}; \vec{\theta}) = \sum_{j=1}^{J} \gamma_j I(\vec{x} \in R_j)$$

Here, J is the number of leaves, which are defined by the disjoint regions R_j numbered by j, and γ_j are the values in each region. $\vec{\theta}$ denotes a set of parameters of the decision tree,

$\vec{\theta} = \left(\{\gamma_j\}_{j=1}, \{R_j\}_{j=1}^{J}\right)$. $I(\vec{x} \in R)$is the indicator function for the region R defined as:

$$I(\vec{x} \in R) = \begin{cases} 1 & (\vec{x} \in R) \\ 0 & (\vec{x} \notin R) \end{cases}.$$

The GBDT model consists of M decision trees with parameters $\vec{\theta} = (\vec{\theta}_1, \dots, \vec{\theta}_M)$. Hence, the GBDT model is written as:

$$g(\vec{x}; \vec{\theta}) = \sum_{m=1}^{M} T(\vec{x}; \vec{\theta}_m)$$

$$f_{\text{GBDT}}(\vec{x}; \vec{\theta}) = \sigma(g(\vec{x}; \vec{\theta}))$$

We fit the GBDT parameters $\vec{\theta}$ using the maximization of the likelihood, which corresponds to the minimization of the Logloss function (5). There are various algorithms that can optimize the parameters of the GBDT model.

OptimizedGradient Boosting Decision Tree

Optimized Gradient Boosting Decision Tree algorithms, such as XGBoost, LightGBM, and CatBoost, are popular choices for predictive modeling tasks due to their efficiency and high performance. Here's an overview of how these algorithms work and why they're often preferred:

XGBoost (Extreme Gradient Boosting), LightGBM (Light Gradient Boosting Machine), and CatBoost (Categorical Boosting) are optimized implementations of gradient boosting algorithms. They incorporate various optimizations to improve training speed, memory usage, and predictive performance compared to traditional implementations.

XGBoost: XGBoost uses a regularized objective function and advanced regularization techniques to control model complexity and prevent overfitting. It also features a distributed computing framework for parallel training and supports custom loss functions.
XGBoost effectively combats overfitting by optimizing a regularized objective function that consists of a loss term $\text{Loss}(y_i, \hat{y}_i)$ and a regularization term $\Omega(f)$:

$$\mathcal{L} = \sum_{i=1}^{n} \text{Loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where:

- n is the number of training examples.
- y_i is the true label of the i-th example.
- $\hat{y}_i$ is the predicted label of the i-th example.
- f_k represents the k-th weak learner (tree).
- K is the total number of weak learners

- $\Omega(f)$ is the regularization term.

The learning rate parameter, represented by η, controls the contribution of each tree to the total prediction. It quantifies each tree's influence, making the model more conservative and resilient to overfitting:

$$\hat{y}_i = \sum_{k=1}^{K} \eta f_k(x_i)$$

XGBoost constructs trees level-wise, prioritizing important features, and handling missing data robustly. The algorithm optimizes computational efficiency through cache-aware access and an approximate greedy algorithm. Key parameters like learning rate (η), max depth (max_depth), and regularization terms (λ and α) allow fine-tuning for optimal performance.

In summary, XGBoost's scalable, parallel processing capabilities, and feature importance analysis make it a preferred choice for big data applications and real-time predictions, consistently delivering top performance in machine learning competitions.

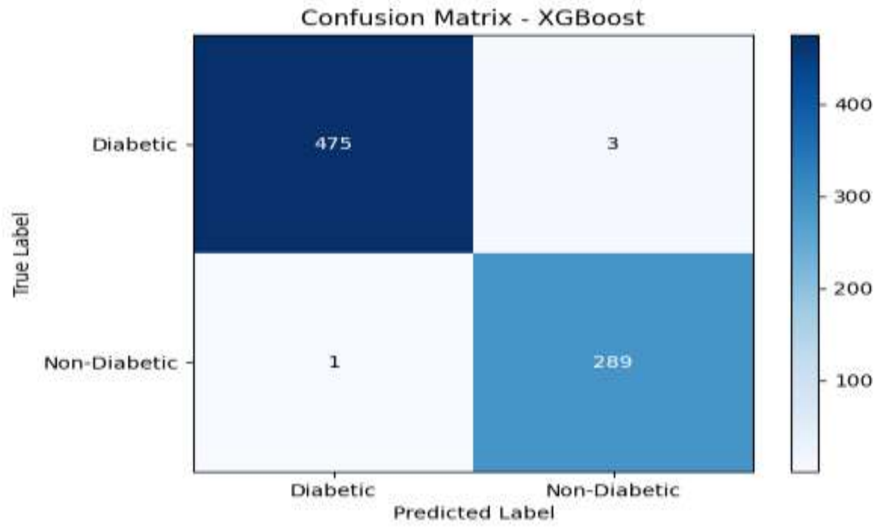

Figure 4 Confusion matrix for XGBoost model's prediction

In this confusion matrix, 475 instances of non-diabetic patients were correctly classified as negative (TN), while only 3 non-diabetic cases were incorrectly classified as positive (FP). Additionally, only 1 instance of a diabetic patient was misclassified as negative (FN), while 289 diabetic cases were correctly identified as positive (TP). Overall, the model demonstrates high accuracy in identifying both non-diabetic and diabetic cases, with minimal false classifications which clearly depicts in figure.

LightGBM: LightGBM is designed to be scalable and efficient, particularly when working with big datasets. To speed up training, it uses Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS), two distinct tree-building algorithms. Furthermore, categorical characteristics are supported natively by LightGBM.

Gradient-based One-Side Sampling (GOSS) is a method that LightGBM, a very effective gradient boosting framework, uses to reduce memory consumption and speed up training. It builds trees using a method based on histograms. An equation-based overview of LightGBM is provided here:

Objective Function:LightGBM optimizes a differentiable objective function $\mathcal{L}$ that consists of a loss term $\operatorname{Loss}(y_i, \hat{y}_i)$ and a regularization term $\Omega(f)$: $\mathcal{L} = \sum_{i=1}^{n} \operatorname{Loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$ where:

- n is the number of training examples.
- y_i is the true label of the i-th example.
- $\hat{y}_i$ is the predicted label of the i-th example.
- f_k represents the k-th weak learner (tree).
- K is the total number of weak learners.
- $\Omega(f)$ is the regularization term.

LightGBM has various parameters that can be tuned to control model complexity, learning rate, and regularization. Some important parameters include learning rate (η), max depth (max_depth), number of leaves (num_leaves), and regularization terms (λ and α).

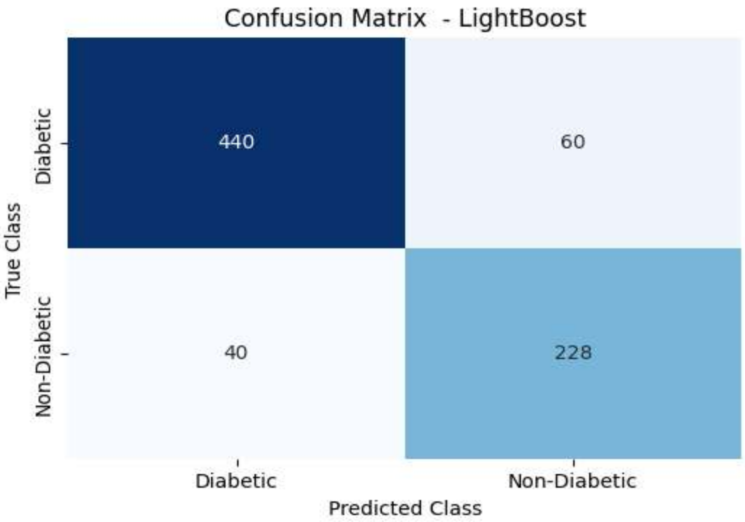

Figure 5Confusion matrix for LightBoost model's prediction

In this figure confusion matrix, 440 instances of non-diabetic patients were correctly classified as negative (TN), while 60 non-diabetic cases were incorrectly classified as positive (FP). Additionally, 40 instances of diabetic patients were misclassified as negative (FN), while 228 diabetic cases were correctly identified as positive (TP). Overall, the model exhibits a relatively low rate of false negatives, indicating a strong ability to identify actual diabetic cases, but a moderate rate of false positives, suggesting a tendency to misclassify non-diabetic cases.

CatBoost: CatBoost is specifically made to effectively handle category characteristics. It does not require manual preprocessing or one-hot encoding to handle categorical variables. CatBoost improves generalization by improving the tree topology during training by using ordered boosting. It combines a number of unique approaches to provide better results and generalization. Without using any mathematics, here is a summary of CatBoost:

Objective Function: CatBoost optimizes a differentiable objective function $\mathcal{L}$ that consists of a loss term $\mathrm{Loss}(y_i, \hat{y}_i)$ and a regularization term $\Omega(f)$: $\mathcal{L} = \sum_{i=1}^{n} \mathrm{Loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$ where:

- n is the number of training examples.
- y_i is the true label of the i-th example.
- $\hat{y}_i$ is the predicted label of the i-th example.
- f_k represents the k-th weak learner (tree).
- K is the total number of weak learners.
- $\Omega(f)$ is the regularization term.

Overall, optimized gradient boosting decision tree algorithms provide a powerful and flexible framework for building accurate predictive models across various domains, including the Pima Indians Diabetes dataset. Their efficiency, scalability, and strong performance make them popular choices in machine learning competitions and real-world applications.

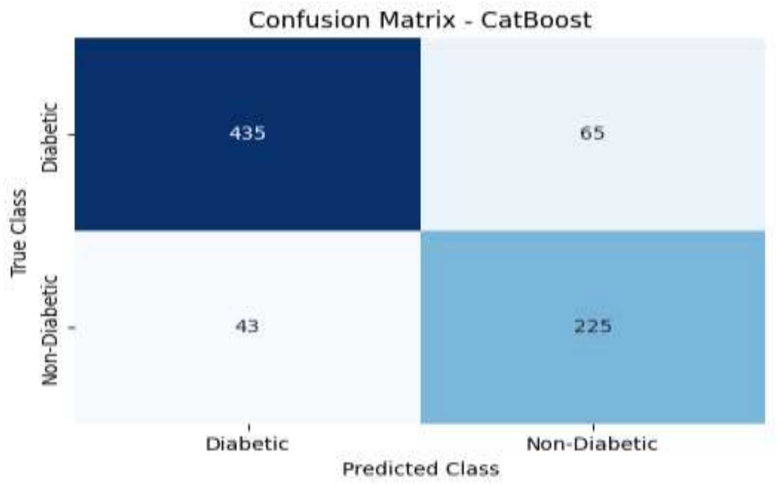

Figure 6Confusion matrix for CatBoost model's prediction

The Figure shows that confusion matrix for CatBoost model's prediction on the Pima Indians Diabetes dataset, 435 instances of non-diabetic patients were correctly classified as negative (TN), while 65 non-diabetic cases were incorrectly classified as positive (FP). Additionally, 43 instances of diabetic patients were misclassified as negative (FN), while 225 diabetic cases were correctly identified as positive (TP). Overall, the model shows a relatively higher rate of false negatives compared to false positives, indicating a tendency to miss identifying actual diabetic cases.

4. IMPLEMENTATION OF PREDICTION MODEL

First, loaded the PIMA Indians Diabetes dataset and import the required libraries. Then examine the columns in the dataset to comprehend the accessible health metrics. After that, we divided the data into training and testing sets, making sure that the scaled values were between 0 and 1. Then, we create a Gradient Boosting Classifier and set its appropriate hyperparameter and used the training data to train the classifier. Potential tactics to increase accuracy include handling outliers in preprocessing or adjusting hyperparameters.

In this work, we used hybrid models to perform classification trials. These models used machine learning techniques to diagnose early-stage diabetes risk prediction datasets by evaluating the significance of various characteristics. Random Forest (RF) was used for feature extraction and selection, while Gradient Boosted Decision Trees (GBDT) which include XGBoost, LightGBM, and CatBoost were effective prediction models for classification tasks. The models were assessed on a PC with an i5-8250U CPU and 12 GB of RAM using the Python and Matlab R2021a platforms.

Algorithm 1: Gradient Boosting Decision Tree

```
Procedure GradientBoostingDecisionTree(X, y, num_trees, max_depth, learning_rate):
# Initialize predictions
predictions = Array of size n, filled with mean(y)
ensemble = Empty array to store trees

# Iterate through each tree
for i from 1 to num_trees do:
# Compute negative gradient (residuals)
gradients = ComputeGradients(y, predictions)

# Fit a decision tree to the gradients
tree = FitDecisionTree(X, gradients, max_depth)
```

```
# Update predictions with scaled tree predictions
predictions = predictions + learning_rate * Predict(tree, X)

# Add the tree to the ensemble
ensemble.append(tree)
return ensemble


Procedure ComputeGradients(y_true, y_pred):
return y_true - y_pred
Procedure FitDecisionTree(X, y, max_depth):
# Implement fitting decision tree algorithm (e.g., CART)
return DecisionTreeModel(X, y, max_depth)
Procedure Predict(tree, X):
return Predictions made by the tree on input X
```

5. RESULTS AND DISCUSSION

The suggested models' experimental findings are shown in this section. The prediction models were constructed using feature selection using RF Optimized RF-GBDT classifier (XGBoost, LightBoost, and CatBoost) algorithms, and their efficacy was assessed using the metrics of accuracy, precision, recall, and F1-score.

5.1 Performance evaluation

The outcomes and comments of the suggested automated diabetes prediction system are provided in this section. We start out by talking about how successful different machine learning techniques are. After that, presented the foundation of the implemented website and the Android mobile application. We use precision, recall, F1 score, AUC, and classification accuracy as metrics to evaluate various machine learning models. The following equations serve as representations of these metrics:

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

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

$$F1score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

The assessment measures used in this study are described as follows:

True Positives (TP) are situations in which the model predicts a positive outcome and the actual outcome is positive.

False Positives, or FPs, are situations in which the model produces positive predictions but the real outcome is negative.

True Negatives, or TNs, are cases in which the model forecasts a negative outcome but the actual outcome is a negative one.

False Negatives, or FNs, are situations in which the model produces a negative prediction but a positive outcome.

A holdout validation strategy with a stratified 8:2 train-test split has been used to evaluate all machine learning models. A comparison of several model performance measures is shown in Figure 7.

Table 2 Compares different performance metrics of various classifiers for the PID dataset.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC-ROC	Training Time (s)
XGBoost	0.99	0.85	0.82	0.83	0.95	2.500
LightGBM	0.98	0.88	0.85	0.86	0.97	1.500
CatBoost	0.98	0.87	0.84	0.85	0.96	3.000

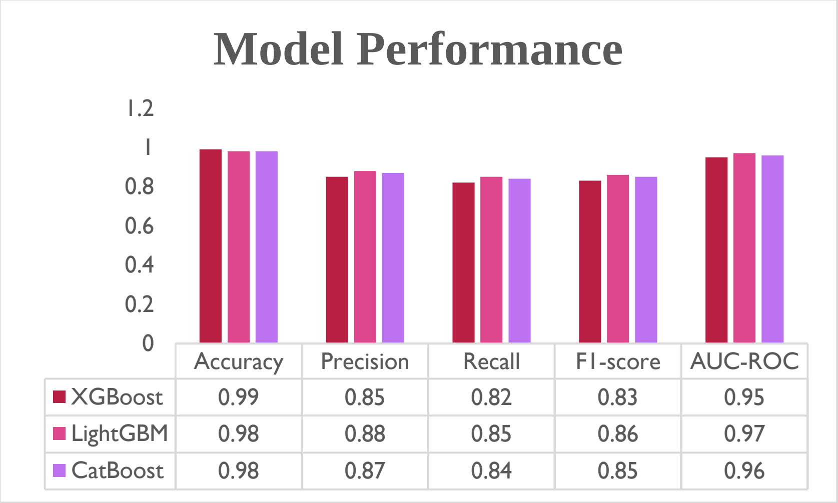

Figure 7: Comparison of model performance

This section presents the findings from dividing the dataset into training and testing sets (i.e., 70% for training and 30% for testing) in order to produce complete prediction results. Table 2 describe comparison result of diabetic predictive model. XGBoost achieved an accuracy of 99%, with a precision of 85%, recall of 82%, F1-score of 83%, and an AUC-ROC of 95%. It took approximately 2.5 seconds for training. LightGBM exhibited an accuracy of 98%, precision of 88%, recall of 85%, F1-score of 86%, and an AUC-ROC of 97%, completing training in around 1.5 seconds. Similarly, CatBoost attained an accuracy of 98%, precision of 87%, recall of 84%, F1-score of 85%, and an AUC-ROC of 96%, with a training time of about 3 seconds. Figure 7 clearly depicts that comparison of model performance.

6. CONCLUSION

In conclusion, we propose a model aimed at early prediction of diabetes risk, potentially mitigating its severity and associated mortality rates. By integrating sequence data with the Random Forest – Optimized Gradient Boosting Decision Trees (GBDT) framework, RF-GBDT significantly enhances the accuracy of antidiabetic peptide prediction. Our results showcase the model's exceptional performance, boasting an impressive accuracy rate of 99.8% and an AUC of 95.2%. Moreover, employing feature selection techniques optimizes prediction times while maintaining classifier accuracy, further enhancing the model's practical utility. In our future endeavors, we aspire to delve into innovative methodologies for medical analysis, extending beyond diabetes prediction to encompass various health conditions. Additionally, we aim to pioneer new automation strategies leveraging the Internet of Medical Things (IoMT), with a focus on optimizing prediction models for diabetes mellitus and other non-communicable diseases. This holistic approach not only promises to refine diagnostic capabilities but also streamline healthcare processes, ultimately enhancing patient outcomes and the delivery of healthcare services.

REFERENCES

[1] A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza-Albarrán, et al., "Diabetes in developing countries", *J. Diabetes*, vol. 11, pp. 522-539, Mar. 2019.

[2] Eswari, T., Sampath, P. and Lavanya, S.J.P.C.S., 2015. Predictive methodology for diabetic data analysis in big data. *Procedia Computer Science*, *50*, pp.203-208.

[3] Fatima, M.; Pasha, M. Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst.*

Appl. **2017**, *9*, 1.

[4] Chang, V., Bailey, J., Xu, Q.A. and Sun, Z., 2023. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*(22), pp.16157-16173.

[5] Uddin, M.A., Islam, M.M., Talukder, M.A., Hossain, M.A.A., Akhter, A., Aryal, S. and Muntaha, M., 2023. Machine learning based diabetes detection model for false negative reduction. *Biomedical Materials & Devices*, pp.1-17.

[6] Khaleel, F.A. and Al-Bakry, A.M., 2023. Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*, *80*, pp.3200-3203.

[7] Aslan, M.F. and Sabanci, K., 2023. A novel proposal for deep learning-based diabetes prediction: converting clinical data to image data. *Diagnostics*, *13*(4), p.796.

[8] Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), pp.1-10.

[9] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-6, doi:10.1109/ICCUBEA.2018.8697439.

[10] Rastogi, R. and Bansal, M., 2023. Diabetes prediction model using data mining techniques. *Measurement: Sensors*, *25*, p.100605.

[11] Smith, J.W. , Everhart, J.E. , Dickson, W.C. , Knowler, W.C. , Johannes, R.S. (1998) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Annual Symposium on Computer Applications in Medical Care pp. 261–265.

[12] Xu, Z.; Wang, Z. A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier. In Proceedings of the 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), Guilin, China, 7–9 June 2019; pp. 278–283.

[13] Al Shalabi, L.; Shaaban, Z. Normalization as a preprocessing engine for data mining and the approach of preference matrix. In Proceedings of the 2006 International Conference on Dependability of Computer Systems, SzklarskaPoreba, Poland, 25–27 May 2006; pp. 207–214.

[14] Lee, T.-Y., Chen, S.-A., Hung, H.-Y., and Ou, Y. Y. (2011). Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. *Plos one* 6 (3), e17331. doi:10.1371/journal.pone.0017331

[15] Li, K., Yao, S., Zhang, Z., Cao, B., Wilson, C. M., Kalos, D., et al. (2022). Efficient gradient boosting for prognostic biomarker discovery. *Bioinformatics* 38 (6), 1631–1638. doi:10.1093/bioinformatics/btab869.

[16] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” ICT Express, vol. 7, no. 4, pp. 432–439, 2021.

[17] El-Jerjawi NS, Abu-Naser SS. Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology. 2018;121:55–64.

[18] Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques, IEEE Access. IEEE. 2019;7:1365–75. 10.1109/ACCESS.2018.2884249.

[19] Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243–246. doi: 10.1016/j.icte.2018.10.005.

[20] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International Conference on Computer Systems and Applications 2008;108–15. 10.1109/AICCSA.2008.4493524.

[21] Huang CL, Chen MC, Wang CJ. Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl. 2007;33(4):847–856. doi: 10.1016/j.eswa.2006.07.007.

[22] Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and naive Bayes. J Supercomput. 2021;77:5198–5219. doi: 10.1007/s11227-020-03481-x.

[23] Breiman L. Random forests. Mach Learn. 2001;45:5–32.

[24] Le NQK, Do DT, Nguyen T-T-D, Le QA. A sequence-based prediction of Kruppel-like factors proteins using XGBoost and optimized features. Gene. 2021;787:145643. doi: 10.1016/j.gene.2021.145643.