IJEASM

International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 3, March 2025

Uncovering Hidden Patterns for Diabetes Prediction: A Synergy of EDA and Ensemble Learning

Lokendra Argal[1] and Dr. Pinaki Ghosh[2]
[1]Research Scholar, Sanjeev Agrawal Global Educational (SAGE) University, Bhopal, India, 462022
[2]Professor, Sanjeev Agrawal Global Educational (SAGE) University, Bhopal, India, 462022
[1]lokendra.argal@gmail.com, [2]pinaki.g@sageuniversity.edu.in

Abstract: The rising incidence of diabetes presents a significant challenge to global healthcare systems, emphasizing the need for accurate and timely predictive strategies to support early diagnosis and intervention. This study leverages recent advancements in machine learning to develop a comprehensive predictive framework that integrates Light Gradient Boosting Machine (LGBM), K-Nearest Neighbors (KNN), and a Voting Classifier ensemble. Through detailed exploratory data analysis (EDA), we identified key features and patterns associated with diabetes, which informed model construction and interpretation. The results highlight the superior performance of the ensemble model, which outperformed individual algorithms in terms of accuracy and reliability. Furthermore, insights gained during the EDA process significantly enhanced feature selection and overall model effectiveness. This research demonstrates the potential of combining diverse machine learning approaches and data-driven insights to address the growing diabetes burden, offering a promising path toward improved healthcare outcomes through predictive analytics.

Keywords: *Diabetes prediction, KNN, LGBM, EDA, Healthcare, Ensemble Learning*

1. INTRODUCTION

Diabetes, also known as diabetes mellitus (DM), is a set of metabolic issues identified by high blood glucose levels over a prolonged duration of time. Symptoms of high glucose incorporate excessive urination, always feeling thirsty and increased hunger [1]. If not treated on time, diabetes can cause serious health issues in an individual such as diabetic ketoacidosis, hyperosmolar hyperglycaemic state, or even lead to death. This may lead to lifetime complications including cardiovascular ailment, brain stroke, kidney failure, ulcers in the foot, and eye complications etc. [2]. Diabetes is caused when the pancreas in the body is unable to generate insulin in enough quantity or when the cells and tissues in the body fail to utilize the insulin produced. Diabetes mellitus exists in three forms as explained below [3]:

- Diabetes Mellitus Type-1 is characterized by pancreas generating insulin less than what is required by the body, a condition also called "insulin-subordinate diabetes mellitus" (IDDM). People suffering from type-1 DM require external insulin dosage to make up for the less insulin produced by the pancreas [4].
- Diabetes Mellitus Type-2 is marked by the body resisting insulin as the body cells react differently to insulin than they normal would. This may ultimately lead to no insulin in the body. This is otherwise called "non-insulin subordinate diabetes mellitus" (NIDDM) or "adult starting diabetes". This type of diabetes is commonly found in people with high BMI or those who lead an inactive lifestyle [5].

Gestational diabetes is the third principle structure that is observed during pregnancy [6].

2. NEED OF EARLY PREDICTION OF DIABETES

Early prediction of diabetes is imperative due to its potential to prevent or delay the onset of the condition through proactive lifestyle changes, reducing the risk of debilitating complications such as cardiovascular disease and kidney failure [7], which not only improves individual health outcomes but also translates into significant cost savings for healthcare systems and enhances overall quality of life [8]. Moreover, early detection allows for targeted preventive interventions at the community level, ultimately reducing the burden of diabetes and promoting healthier populations [9]. Thus, emphasizing the importance of early prediction underscores its pivotal role in effective diabetes management and public health initiatives. Early prediction of diabetes is crucial for several reasons:

Preventive Measures- Early detection allows individuals to take proactive steps to prevent or delay the onset of diabetes [10]. Lifestyle changes such as adopting a healthier diet, increasing physical activity, and managing stress can significantly reduce the risk of developing diabetes.

Avoidance of Complications- Diabetes, if left untreated or poorly managed, can lead to severe complications such as heart disease, stroke, kidney disease, blindness, and nerve damage. Early detection enables early intervention, which can help prevent or mitigate these complications [11].

Cost Savings- Early prediction and intervention can result in significant cost savings for individuals, healthcare systems, and society as a whole. Treating diabetes and its complications can be expensive, so early detection can help reduce healthcare costs by preventing or delaying the need for expensive medical treatments [12].

Improved Quality of Life- Early diagnosis allows individuals to start managing their condition effectively, leading to better control of blood sugar levels and overall health. This can result in a better quality of life, with fewer symptoms and complications associated with diabetes [13].

Public Health Impact- Early prediction of diabetes can have a broader public health impact by identifying individuals at risk within communities and implementing targeted prevention programs. This can help reduce the overall burden of diabetes on society and improve population health outcomes [14].

Overall, early prediction of diabetes is essential for promoting individual health, reducing healthcare costs, and improving public health outcomes.

3. ENSEMBLE LEARNING

Ensemble learning is a machine learning technique that combines multiple models to improve prediction accuracy and generalization performance [15]. It leverages the diversity of individual models to mitigate biases and errors, leading to more robust predictions [16]. Mathematically, the prediction of an ensemble model $F(x)$ is typically represented as a weighted combination of predictions from N base models:

$$\boldsymbol{F(x) = \sum_{i-1}^{N} w_i f_i(x)} \qquad \text{Eq. (1)}$$

Where, $f_i(x)$ represents the prediction of the i^{th} base model, and w_i denotes the weight assigned to the i^{th} model.

There are several types of ensembles learning methods, including bagging, boosting, and staking. Bagging (Bootstrap Aggregating) involves training multiple base models on bootstrap samples of the training data and averaging their predictions [17]. Boosting focuses on sequentially training weak learners to correct the errors of preceding models, leading to a strong final model. Stacking combines predictions from multiple base models using a meta-learner to produce the final prediction. Ensemble learning finds applications in various domains, including classification, regression, and anomaly detection. It is widely used in fields such as healthcare, finance, and marketing to improve decision-making and predictive accuracy [18].

Gradient Boosting

Gradient Boosting is a powerful machine learning technique used for building predictive models, particularly in regression and classification tasks [19]. It works by sequentially adding weak learners, typically decision trees, to improve the predictive performance of the model. The core idea behind Gradient Boosting is to optimize a loss function by minimizing the residuals (or gradients) of the loss with respect to the predicted values. Mathematically, the prediction of a Gradient Boosting model $F(x)$ for a given instance x is defined as:

$$F(x)=Ft-1(x)+\gamma ht(x) \qquad \text{Eq (2)}$$

Where $Ft-1(x)$ is the prediction of the model at iteration $t-1$, $h_t(x)$ is the weak learner (e.g., decision tree) added at iteration $\mathbf{t}$, and $\boldsymbol{\gamma}$ is the learning rate.

The algorithm proceeds in a forward stage-wise manner, where each weak learner is trained to minimize the residual error of the previous model. In each iteration, the weak learner is trained on the negative gradient of the loss function with respect to the predicted values [20]. The predictions of all weak learners are then aggregated to form the final prediction of the ensemble model. Gradient Boosting has gained popularity due to its flexibility, scalability, and ability to handle complex datasets [21]. It has been successfully applied in various domains, including healthcare, finance, and natural language processing.

4. LITERATURE REVIEW

Diabetes is a big health problem worldwide, and catching it early is really important. Recently, machine learning has been used more in healthcare to help doctors diagnose and predict diseases better. Diabetes happens when your body doesn't make enough insulin, which is needed to manage the sugar in your blood. Too much sugar in your blood can cause a lot of health issues in the long run. Type-2 diabetes is the most common type, sometimes called "Pima Indians' Diabetes". Smoking can make diabetes worse and cause problems with your heart, kidneys, and eyes. About 5.5% of the world's population has diabetes, and most have type 2. That number is expected to go up by 48% soon. Doctors can find diabetes by checking manually or using machines. Manual checks need trained professionals, but sometimes the early signs of diabetes are hard to spot. That's where machine learning comes in. With new technology, machines can help find diabetes early, which is really helpful for doctors.

Several studies have explored machine learning models for diabetes prediction and diagnosis. Shahid Mohammad Ganie et al. [1] assessed boosting algorithms, with gradient boosting achieving a high accuracy of 96%. Roshan Birjais et al. [2] focused on diagnosing diabetes, highlighting Gradient Boosting's superior predictive accuracy. M. Jishnu Sai et al. [3] proposed ensemble algorithms, emphasizing the LightGBM + k-NN + Adaboost ensemble for diabetes detection. Neha Prerna Tigga and Shruti Garg [4] implemented six classification methods, noting Random Forest's 94.10% accuracy. Varun Jaiswal et al. [5] discussed various machine learning techniques, emphasizing the shift towards higher reliability in diabetes prediction models. Raja Krishnamoorthi et al. [6] compared ML models, with Logistic Regression outperforming others. Md. Kamrul Hasan et al. [7] proposed a robust framework, achieving high sensitivity, specificity, and AUC values. KM Jyoti Rani et al. [8] developed an early diabetes prediction system, with the Decision Tree algorithm reaching 99% accuracy. Henock M. Deberneh and Intaek Kim [9] utilized ensemble models for predicting T2D occurrence. Ram D. Joshi and Chandra K. Dhakal [10] predicted type 2 diabetes in Pima Indian women using logistic regression and decision tree. Mitushi Soni and Sunita Varma [11] designed a Diabetes Prediction System with Random Forest showing higher accuracy. Leila Ismail et al. [12] evaluated 35 ML algorithms, identifying Bagging-LR as accurate for a balanced dataset. Luis Fregoso-Aparicio et al. [13] highlighted the relevance of dataset structure and recommended reporting multiple metrics. Md. Maniruzzaman et al. [14] employed LR and classifiers, achieving high ACC and AUC values. Leon Kopitar et al. [15] compared machine learning models, emphasizing interpretability and model calibration in clinical prediction. These studies collectively contribute to advancing the field of diabetes prediction and diagnosis through diverse methodologies and model comparisons. Jingyu Xue et al. [16] applied supervised machine-learning algorithms, including Support Vector Machine (SVM), Naive Bayes, and LightGBM, to a dataset of 520 diabetic and potential diabetic patients, revealing Support Vector Machine's superior performance. Ashima Singh et al. [17] proposed the eDiaPredict ensemble framework, incorporating XGBoost, Random Forest, Support Vector Machine, Neural Network, and Decision Tree for diabetes prediction, achieving an individual XGBoost accuracy of 92% and an ensemble accuracy of 95%, marking significant improvements. G. Geetha and K. Mohana Prasad [18] introduced the T2DDP model for type 2 diabetes prediction, utilizing Naïve Bayes and ensemble algorithms, reaching a remarkable 98% correctness rate. Norma Latif Fitriyani et al. [19] presented a Disease Prediction Model (DPM) for early detection of type 2 diabetes and hypertension, incorporating isolation forest, SMOTETomek, and ensemble approaches. The DPM outperformed other models, and a mobile application was developed for practical implementation. Md Abdur Rahim et al. [20] proposed a robust stacked ensemble method for diabetes prediction, utilizing SVM, KNN, NB, RF in base models, and Logistic Regression in the meta-model, achieving an accuracy of 94.17% on the PIMA Indian Diabetes Dataset. Sapna Singh and Sonali Gupta [21] employed bagging and boosting techniques on the Pima Indians dataset, with Random Forest attaining the highest accuracy at 82.46%, and AdaBoost achieving the highest recall at 75%. Random Forest emerged as the most accurate model for diabetes prediction [22-25].

5. PROPOSED METHODOLOGY

The below given figure 1 shows the overall workflow of proposed work in pictorial representation. In initial phases the PIMA diabetes dataset has been read from PIMA dataset to experimental environment using CSV read function. Experiment was performed to implement proposed work on Kaggle cloud machine where PIMA repository is available to experiment. Description of PIMA was already given in section. After read of dataset to work with data exploratory data analysis was performed to check statistics of dataset values that can help in preprocessing dataset for our purpose. Issues null value, missing value, distribution of features must be seen before building model that may create

International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 3, March 2025

problem for good performance results in training and validation.

EDA help to check all given parameter, which are creating problems through graphical representation and features relation maps. Model training is the next step after preprocessing dataset and partitioning into Training and validation test. K-fold cross-validation is a method for evaluating predictive models by dividing the dataset into k folds. Each fold is used for training and validation, with performance metrics averaged to estimate the model's generalization performance. This method helps avoid overfitting and ensures a generalized model. To achieve this, the data set must be divided into three sets, despite the volume of the data. Hyperparameters are configuration variables set before model training, controlling the learning process and affecting model performance, accuracy, generalization, and other metrics. For hyperparameters tuning Grid Search method is used on cascaded model of Light gradient based and KNN. After reaching best performance model consider final version for testing and performance evaluation on test set. Below algorithm provides an idea for model.

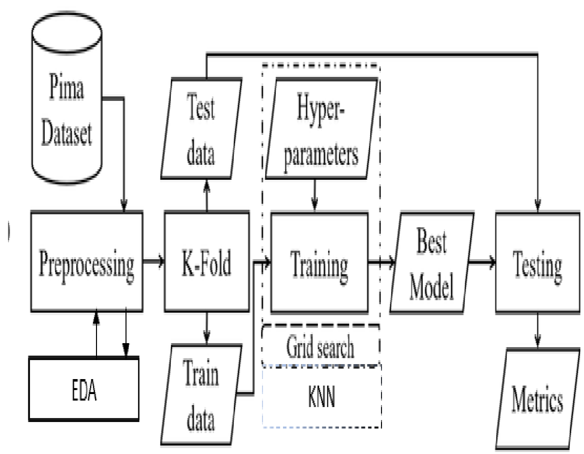

Fig. 1: Proposed Workflow

6. RESULTS & DISCUSSIONS

Available Dataset

The Pima (or Akimel O'odham, also spelled Akimel O'otham, "River People", formerly known as Pima) are a group of Native Americans living in an area consisting of what is now central and southern Arizona. The majority population of the surviving two bands of the Akimel O'odham is based in two reservations: the Keli Akimel O'otham on the Gila River Indian Community (GRIC) and the On'k Akimel O'odham on the Salt River Pima-Maricopa Indian Community (SRPMIC). Dataset description from Panda's is shown in figure 2.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies                 768 non-null int64
Glucose                     768 non-null int64
BloodPressure               768 non-null int64
SkinThickness               768 non-null int64
Insulin                     768 non-null int64
BMI                         768 non-null float64
DiabetesPedigreeFunction    768 non-null float64
Age                         768 non-null int64
Outcome                     768 non-null int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB


None
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 2: Dataset description from Panda's

The complete implementation work has been categorized in two parts in first phase dataset processing and analysis, to find issue and relation in column values. After Exploratory data analysis column value was replaces with best numeric values to maintain relations for class level. In second phase of experiment model training for light Gradient based model has been trained in two ways without KNN and with KNN on 5-fold hyperparameters tuning. Comparison on for predicting disease based on query types was evaluated on test results on fine-tuned model resulted from voting and grid search optimization method.

Task 1: -Exploratory Data Analysis

In statistical analysis results of missing value in perce65ntage among overall count has been shown in bar graph over total 652 data sample. Column field Insulin 48.7%, Skin thickness 29.56 %, Blood pressure, BMI 1.43 % and glucose 0. 65% have missing value and others have no missing values. These values are can be handled in using replaced with NaN. Figure-3 shows total numbers of sample having Missing Values in each column is given using bar graph.

International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 3, March 2025

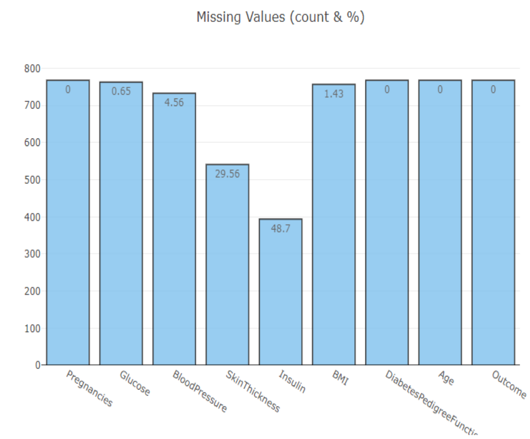

Figure 3: Missing Value in PIMA

To check the effect of missing value change to NaN, This NaN will be consider for data analysis but not effective to consider in prediction as it dilutes standard measures mean for a column. Now as NaN cannot be considered for numerical analysis to need to change in numeric value to variable mean value can to place in place of NaN. The correlation Matrix looks as below in figure 4. Individual correlation with target class is shown with differed column in Figure 5 on Scerum, Plasma, skin thickness, Blood Pressure, BMI, age, pregnancy, and diabetes Pedigree Function respectively.

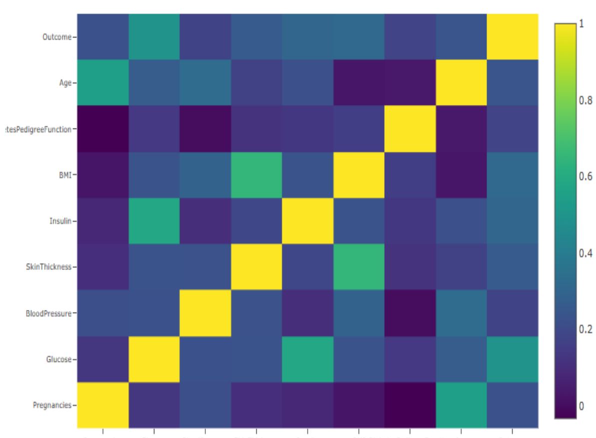

Figure 4: Correlation after NaN replacement

The NaN can be replaced by mean or made in of the variable. The same kind of correlation has been found in the experimentation which is shown in figure 4. After all NaN replacement with each of its median or Mean value now all missing data will be change to all **"0"** as a level of bar plot means no NaN remaining. Bar Chart after preprocessing has been given in Figure 5.

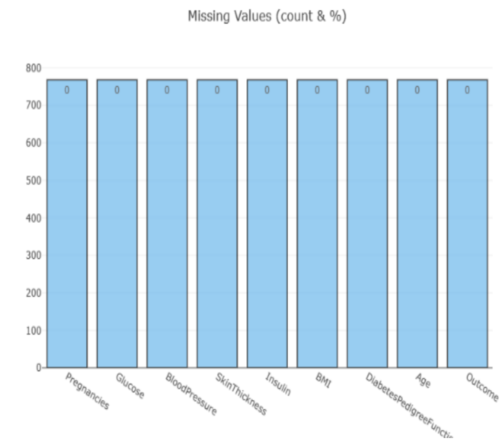

Figure 5: After NaN replacement Bar Chart

There are different types of symptoms that relate the seriousness of diabetes in patient and reflected in human body. So, one features dependent to other and must consider along with. There are some feature relations has been designed and try to judge them based on distribution of samples like, the chance of diabetes in a patient is possible due to age, or in hormonal changes like pregnancies. Standard Scalar and Label Encoder help estimators in standardization of a dataset to ensure features behave like standard normally distributed data, such as Gaussian with 0 mean and unit variance. This involves independent centering and scaling on each feature. Encode labels with value between 0 and n_classes-1. Use of standard scalar and label encoding correlates features in batter way to improve different class of PIMA and also help out to logically connect and re-design new attribute.

Task 2:

In task 2 it involves in prediction model development and comparison. The first model LGBM with 5- fold was implemented on pre-processed data set after task 1. After task 1 all the features are normalised and correlated with each other that was shown in correlation graph. Now the propose model was implemented having interest in LGBM with KNN using grid search and voting method for optimising hyper parameter values. Using these hyperparameters optimization, the proposed model performs well. Results are checked on accuracy, precision, recall, F1 score and ROC curve. Cumulative graphs for both implemented prediction graphs are given in Figure 6 and figure 7.

After execution, **accuracy 0.8958, precision 0.8561, recall 0.8433 and F1 Score 0.8496** was seen as an output. Confusion matrix, ROC curve based on supporting value, and precision recall curve for LGBM is given in Figure 6.

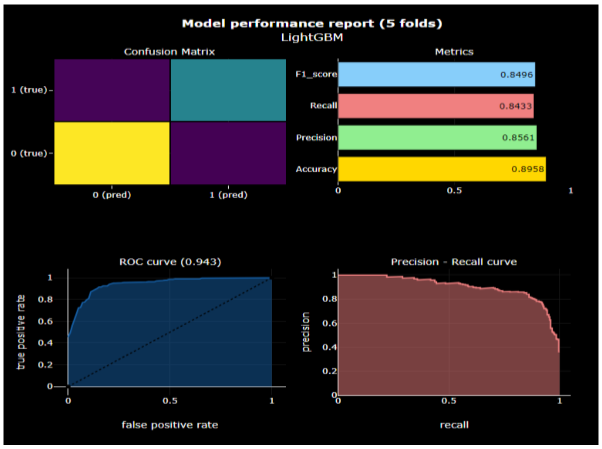

Figure 6: LGBM model performance graph

Threshold plot check cross regions for all performance parameter to judge chance of improvement over taken classifier based on discrimination threshold. To draw and analyze performance of LGBM all parameter line was drawn and distribution threshold taken 0.43 as given in Figure 7.

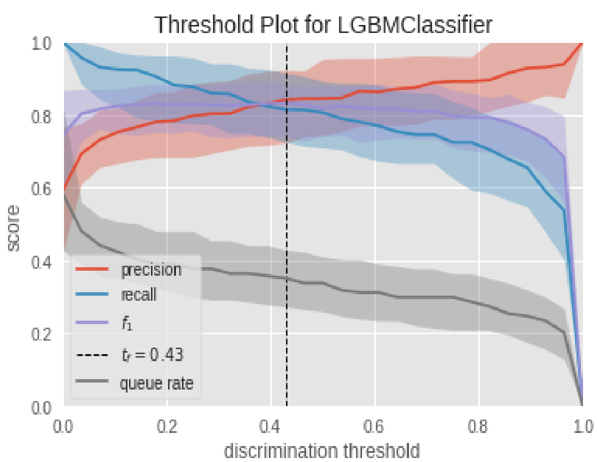

Figure 7: Threshold plot for LGBM classifier

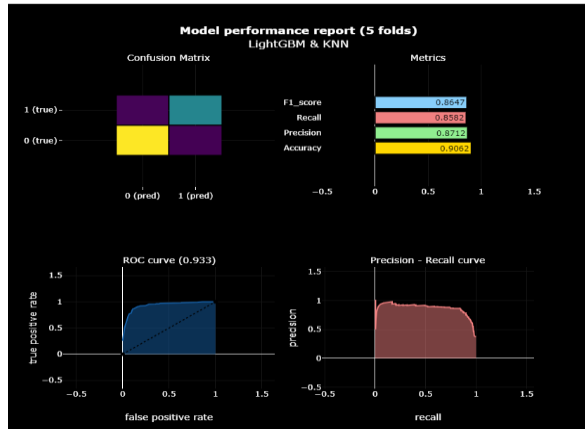

Figure 8: Proposed model performance graph

The cross point of all lines lies left region mean need to improve for improving positive class sample. This cross point must be on the line or right section approached. After executions, accuracy 0.9062, precision 0.8712, recall 0.8582 and F1 Score 0.8647 was seen as an output. Confusion matrix, ROC curve based on supporting value, and precision recall curve for LGBM is given in Figure 8.

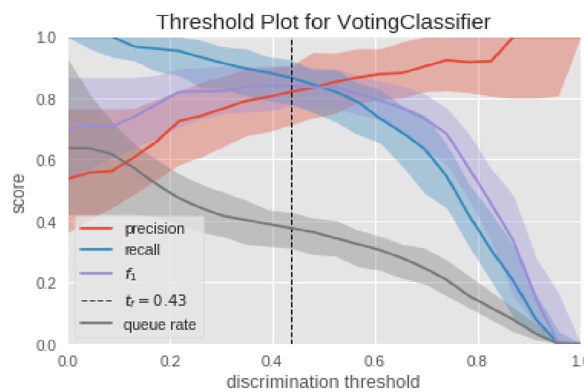

Figure 9: Threshold Plot for voting classifier effect on proposed work

Threshold plot check cross regions for all performance parameter to judge chance of improvement over taken classifier based on discrimination threshold. To draw and analyse performance of proposed work, all parameter line was drawn and distribution threshold taken 0.43 as given in Figure 9. Comparing with Figure 7, Figure 9 has cross point of parameters in right section showing proposed work batter comparative to previous.

Results of both the comparative model has been given in Table 1 It is clear from table after 5-fold in implemented prediction models, mean value of proposed model is comparatively batter in on taken parameters. Proposed model results 90.6 % accuracy and 0.864 F1- score which was an effect of hyperparameters optimization. Moreover, threshold plot shows overall threshold value that can separate positive and negative classes as target (healthy and diabetic).

Table 1 : Performance comparison on parameters

Fold / Model	LGBM					Proposed LGBM +KNN				
	accuracy	precision	recall	F1-Score	Roc Curve	accuracy	precision	recall	F1-Score	Roc Curve
1	0.903	0.915	0.796	0.851	0.945	0.896	0.896	0.796	0.843	0.922
2	0.864	0.789	0.833	0.811	0.926	0.877	0.797	0.87	0.832	0.918
3	0.896	0.865	0.833	0.849	0.949	0.916	0.902	0.852	0.876	0.937
4	0.889	0.846	0.83	0.838	0.944	0.902	0.88	0.83	0.854	0.94
5	0.928	0.875	0.925	0.899	0.972	0.941	0.893	0.943	0.917	0.953
mean	0.896	0.858	0.844	0.85	0.947	0.906	0.873	0.858	0.865	0.934

To provide a condensed and understandable depiction of specific kinds of relationships among components in diverse systems, threshold graphs can be seen. The cross point of different parameter line on threshold graph experiences the behaviour on threshold 0.43 shown as dashed line. The cross point of precision, recall and F1-score lies in left side of threshold 0.43 for LGBM while lies in right side of graph for proposed model. Moving right shows all three parameters accurately seems to lies higher site of discrimination to optimize model and proposed model is better in consideration for validation. Area under the curve bounded by cross region is also skewed means precise prediction at time of validation for available query.

7. CONCLUSION

Diabetes is spreading rapidly, making early diagnosis and treatment more crucial than ever. It occurs when blood sugar levels rise too high due to the body's inability to use glucose effectively. Since glucose, derived from the food we eat, is essential for energy, imbalances—whether too high or too low—can lead to serious health complications. Therefore, predicting elevated blood sugar levels is vital for timely intervention. In this experimental study, we applied the Light Gradient Boosting Machine (LGBM) and K-Nearest Neighbors (KNN) algorithms, along with a Voting Classifier ensemble technique, to predict the presence of diabetes. We began with exploratory data analysis (EDA) to understand the dataset's feature distribution, uncover correlations, and detect patterns relevant to diabetes prediction. The insights gained from EDA informed our feature selection and model development. By leveraging the individual strengths of LGBM in handling complex data and KNN in capturing localized patterns, we built effective predictive models. The Voting Classifier combined predictions from both models, harnessing their diversity to boost overall performance. Our results showed that the ensemble model surpassed the individual models in both accuracy and robustness. This study highlights the effectiveness of ensemble learning in diabetes prediction and underscores the importance of integrating multiple approaches to improve predictive outcomes. Future work can further explore and refine ensemble methods—particularly those combining LGBM and KNN—for more advanced and accurate diabetes prediction systems.

REFERENCES

[1] Ganie, Shahid Mohammad, et al. "An ensemble learning approach for diabetes prediction using boosting techniques." Frontiers in Genetics 14 (2023): 1252159.

[2] Birjais, Roshan, et al. "Prediction and diagnosis of future diabetes risk: a machine learning approach." SN Applied Sciences 1 (2019): 1-8.

[3] Sai, M. Jishnu, et al. "An ensemble of Light Gradient Boosting Machine and adaptive boosting for prediction of type-2 diabetes." International Journal of Computational Intelligence Systems 16.1 (2023): 14.

[4] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." Procedia Computer Science 167 (2020): 706-716.

[5] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15.3 (2021): 435-443.

[6] Engineering, Journal Of Healthcare. "Retracted: A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques." Journal of healthcare engineering 2023 (2023): 9872970.

[7] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.

[8] Rani, K. J. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305.

[9] Deberneh, Henock M., and Intaek Kim. "Prediction of type 2 diabetes based on machine learning algorithm." International journal of environmental research and public health 18.6 (2021): 3317.

[10] Joshi, Ram D., and Chandra K. Dhakal. "Predicting type 2 diabetes using logistic regression and machine learning approaches." International journal of environmental research and public health 18.14 (2021): 7346.

[11] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (IJERT) 9.9 (2020): 921-925.

[12] Ismail, Leila, et al. "Type 2 diabetes with artificial intelligence machine learning: methods and evaluation." Archives of Computational Methods in Engineering 29.1 (2022): 313-333.

[13] Fregoso-Aparicio, Luis, et al. "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review." Diabetology & metabolic syndrome 13.1 (2021): 148.

[14] Maniruzzaman, Md, et al. "Classification and prediction of diabetes disease using machine learning paradigm." Health information science and systems 8 (2020): 1-14.

[15] Kopitar, Leon, et al. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." Scientific reports 10.1 (2020): 11981.

[16] Xue, Jingyu, Fanchao Min, and Fengying Ma. "Research on diabetes prediction method based on machine learning." Journal of Physics: Conference Series. Vol. 1684. No. 1. IOP Publishing, 2020.

[17] Singh, Ashima, et al. "eDiaPredict: an ensemble-based framework for diabetes prediction." ACM Transactions on Multimidia Computing Communications and Applications 17.2s (2021): 1-26.

[18] Geetha, G., and K. Mohana Prasad. "An hybrid ensemble machine learning approach to predict type 2 diabetes mellitus." Webology 18.Special Issue 02 (2021): 311-331.

[19] Fitriyani, Norma Latif, et al. "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension." Ieee Access 7 (2019): 144777-144789.

[20] Rahim, Md Abdur, et al. "Stacked ensemble-based type-2 diabetes prediction using machine learning techniques." Annals of Emerging Technologies in Computing (AETiC) 7.1 (2023): 30-39.

[21] Singh, Sapna, and Sonali Gupta. "Prediction of Diabetes Using Ensemble Learning Model." Machine Intelligence and Soft Computing: Proceedings of ICMISC 2020. Springer Singapore, 2021.

[22] Singh, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.

[23] Nagendra Singh, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[24] Singh, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[25] Naiyer, Vaseem, et al. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.