Journal of Intelligent Systems and Internet of Things
Vol. 12, No. 01, PP. 08-19, 2024
DOI: https://doi.org/10.54216/JISIoT.120101

Outlier Management and its Impact on Diabetes Prediction: A Voting Ensemble Study

S. Phani Praveen[1], Kotte Sandeep[2], N. Raghavendra Sai[3], Aditi Sharma[4*,5], Jitendra Pandey[6], Vikas Chouhan[7]

[1]Department of CSE, PVP Siddhartha Institute of Technology, Vijayawada, A.P, India
[2]Department of IT, Dhanekula Institute of Engineering &Technology, Vijayawada, A.P, India
[3]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
[4]Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[5]IEEE Senior Member, Symbiosis International (Deemed University), Pune, India
[6]Middle East College, Knowledge Oasis Muscat, Oman,
[7]Canadian Institute for Cybersecurity, University of New Brunswick, Canada
Emails: phani.0713@gmail.com; kottesandeep@gmail.com; nallagatlaraghavendra@gmail.com; aditi.sharma@ieee.org; jitendra@mec.edu.om; vikas.chouhan@unb.ca

Abstract

The chronic metabolic disorder known as diabetes mellitus, which is defined by hyperglycemia, poses a significant threat to the health of people all over the world. The categorization is broken down into two primary categories: Type 1 and Type 2, with each category having its own unique causes and approaches to treatment. It is very necessary for the effective management of illnesses to have both the prompt detection and the exact prediction of outcomes. The applications of machine learning and data mining are becoming increasingly important as tools in this setting. The current research study analyses the usage of machine learning models, specifically Voting Ensembles, for the goal of predicting diabetes. Specifically, the researchers were interested in how accurate these models were. Using GridSearchCV, the Voting Ensemble, which consists of LightGBM, XGBoost, and AdaBoost, is fine-tuned to manage outliers. This may be done with or without the Interquartile Range (IQR) pre-processing. The results of a comparative analysis of performance, which is carried out, illustrate the benefits that are linked with outlier management. According to the findings, the Voting Ensemble model, when paired with IQR pre-processing, possesses greater accuracy, precision, and AUC score, which makes it more acceptable for predicting diabetes. Despite this, the strategy that does not use the IQR continues to be a workable and reasonable alternative. The current study emphasizes both the significance of outlier management within the area of healthcare analytics and the effect of data preparation procedures on the accuracy of prediction models. Both of these topics are brought up because of the relevance of the current work.

Received: August 17, 2023 Revised: November 11, 2023 Accepted: February 11, 2024

Keywords: GridSearchCV; Interquartile Range; Voting Ensemble model; LightGBM; XGBoost; AdaBoost.

1. INTRODUCTION

Diabetes mellitus, or diabetes, is a chronic metabolic disorder characterized by high blood glucose levels. This illness results from the body's decreased ability to produce or use insulin, a hormone that regulates blood glucose levels and promotes glucose absorption into cells for energy generation. Diabetes is mostly Type 1 and Type 2.

The immune system mistakenly targets and destroys pancreatic beta cells, which produce insulin, resulting in a decrease or complete lack of insulin synthesis and high blood glucose levels. Type 1 diabetes is managed with frequent insulin injections or insulin. Adults are most likely to develop type 2 diabetes.

This disorder is caused by insulin resistance or inadequate insulin synthesis. In early stages, the pancreas may compensate by producing more insulin. As time goes by, the pancreas may struggle to meet the body's needs, raising blood glucose levels. Lifestyle factors include obesity, sedentary lifestyle, and poor diets often cause Type 2 diabetes. Type 2 diabetes care includes lifestyle changes including diet and exercise, as well as oral medications or insulin in some cases [1].

Gestational diabetes affects blood glucose levels throughout pregnancy. Although it may not always show symptoms, it can harm maternal and fetal health. The well-being of all parties depends on quick identification and efficient management. Untreated diabetes can damage the kidneys, heart, circulatory system, blood vessels, eyes,

and brain. High thirst, frequent micturition, unaccounted-for weight loss, tiredness, and poor vision are typical diabetic symptoms [2].
Diabetes affects many people worldwide, making it a global health issue. Diabetes care requires lifestyle changes, medication management, and blood glucose monitoring. Early detection, public education, and diabetes prevention and management knowledge are crucial to addressing the complexities of this chronic disease and improving healthcare outcomes [3].

2. LITERATURE SURVEY

In 2017, Kavakiotis *et al*. [4] focuses on advanced technologies, health sciences, and machine learning have created new diabetes research options. This systematic review covers machine learning and data mining applications in diabetes prediction, diagnosis, complications, genetics, and healthcare management. On clinical datasets, supervised learning approaches, led by support vector machines, are widely used. As diabetes becomes a worldwide health issue, enormous data resources, electronic health records, and machine learning will lead to new ideas, deeper insights, and better diabetes research and treatment.
In 2018, Sarwar *et al*. [5] uses six machine learning algorithms to identify diabetes early in healthcare. Predictive analytics is used to help healthcare professionals make data-driven choices. On a PIMA Indian population dataset, the research tested SVM, KNN, LR, DT, RF, and NB. The maximum prediction accuracy was 77% for SVM and KNN, with a comprehensive feature significance analysis showing the relevance of plasma glucose concentration, body mass index, and age in diabetes prediction. Predictive analytics can revolutionize healthcare professionals' decision-making, and this research provides insights into algorithms for early diabetes prediction, with limitations and areas for future research and improvement.
In 2022, Ahmed *et al*. [6] introduces a new diabetes prediction approach utilizing integrated machine learning. The proposed framework uses the Support Vector Machine (SVM) and Artificial Neural Network (ANN) to assess a dataset and predict diabetes diagnostic findings. A 70:30 ratio has been maintained while partitioning the dataset into training and testing sets. These models' outcomes are input membership functions for a fuzzy model, which determines diabetes diagnosis. The combined machine learning model outperforms earlier methods with 94.87% prediction accuracy.
In 2020, Ljubic *et al*. [7] intended to develop a prediction model for 10 type 2 diabetes problems in 2020. Anticipating these concerns enables for focused and timely measures to stop or slow their growth. Healthcare Cost and Utilization Project State Inpatient Databases of California contributed 2003–2011 data for the study. Instead of Random Forest and Multilayer Perceptron, the research used sophisticated deep learning algorithms like LSTM and GRU. The research examined complications prediction based on the few hospitalizations between type 2 diabetes mellitus (DM2) diagnosis and issues diagnosis. Diagnostic experiments indicated the RNN GRU model functioned. Interesting, the model predicted 73% myocardial infarction and 83% chronic ischemic heart disease. Traditional models were 66%–76% accurate. The research found hospitalizations influence prediction accuracy. Four hospitalizations are usually better than two. Deep learning models needed a lot of training data, with 1000 patients yielding the best outcomes. Also, prediction accuracy dropped with time, with differences among tasks. The RNN GRU model processed electronic medical record data best, according to the study.
In 2020, Tripathi *et al*. [8] insufficient insulin production causes abnormal blood glucose levels in diabetes mellitus, a common and dangerous disease. If undiagnosed or ignored, this illness may damage vital organs including the kidneys, nerves, and eyes. Personalized healthcare has led to the application of machine learning for predictive analysis, allowing early illness diagnosis and symptom detection. This research uses machine learning classification techniques to create a diabetes prediction model. Key diabetes factors are identified and analyzed in this study. The model matches clinical results well and helps customize patient diagnoses. The Pima Indian Diabetes Database (PIDD) is analyzed using four classification methods: LDA, K-nearest neighbour (KNN), Support Vector Machine (SVM), and Random Forest (RF). The UCI machine learning repository provided the database. The test examines precision, recall, specificity, F-score, and accuracy. Random Forest (RF) outperforms other classifiers with 87.66% accuracy. Overall, the Random Forest (RF) classifier is the best choice for the model and shows promise for early diabetes prediction.
In 2022, Chang *et al*. [9] focuses on healthcare analytics may assist professionals and patients. Data analytics allows healthcare practitioners to discover and diagnose illnesses early, improving quality and patient outcomes. In this context, machine learning models are significant because they can identify trends in medical data and make predictions. Many healthcare fields use these models for illness diagnosis, prognosis, and therapy selection. To predict diabetes diagnosis, this study uses different machine learning algorithms. Comparing various models is the study's main goal to find the most efficient way. Prediction accuracy, precision, recall, and F1 score will be evaluated to achieve this. Random Forest, with 82.26% accuracy, stands out among the algorithms examined.
Machine learning and current technologies have transformed diabetes research in the ever-changing healthcare field. The above studies demonstrate the importance of predictive analytics and data mining in diabetes prognosis and treatment. This study shows that support vector machines and deep learning models may detect diabetes early. For accurate prediction, glucose levels, BMI, and age are crucial. These discoveries might improve healthcare quality, medical decision-making and patient diagnosis in addition to research. Diabetes diagnosis and treatment are improving as healthcare analytics develop. These research advance diabetes detection and treatment.

A. Motivation

Diabetes is a growing global health concern that inspired this investigation. Early detection and accurate forecasting of this disease are essential for effective management and mitigation. Machine learning models may improve diabetes diagnosis and management, improving healthcare outcomes.

B. Research Gap

In spite of the significant advancements that have been achieved in the area of machine learning applications for diabetes, there is still a lack of understanding about the relative effectiveness of different algorithms. A considerable gap in the existing body of research exists regarding the evaluation of the relative efficacy of many models for early diabetes prediction. This evaluation should take into account accuracy as well as any other pertinent performance criteria. The discovery of this disparity has the potential to not only guide medical professionals in the selection of algorithms that are best suited for the treatment of diabetes but also to promote the creation of further innovations in the critically important area of healthcare analytics.

3. Proposed Methodology

This research study demonstrates an innovative approach to diabetes risk assessment that makes use of machine learning techniques. Utilizing a GridSearchCV approach in order to optimize the hyperparameters of a Voting Ensemble model is the core component of the recommended methodology that we have provided. LightGBM, XGBoost, and AdaBoost [10][11] are the names of the three powerful machine learning algorithms that are included in the ensemble. This approach is new because it incorporates an Interquartile Range (IQR) data preparation strategy, which enables us to deal with any outliers that may be present in the dataset in an efficient manner. Through a comparison of the results produced using GridSearchCV with and without the use of the interquartile range (IQR) approach, the goal of this research is to illustrate the effect that removing outliers has on the predictive performance of our model. This will be accomplished by demonstrating the impact that removing outliers has on the predictive performance of our model. This in-depth inquiry will shed light on the potential benefits of using outlier management in healthcare analytics by offering insights into the effect of data preprocessing on the accuracy and reliability of diabetes prediction. Specifically, the investigation will focus on how diabetes prediction is affected by data preprocessing.

In this research study, we suggest an innovative method for diabetes prediction that makes use of machine learning strategies. The use of a GridSearchCV strategy, with the goal of optimizing the hyperparameters of a Voting Ensemble model, is the central component of the solution that we have developed. LightGBM, XGBoost, and AdaBoost are the names of the three potent machine learning algorithms that are included in this ensemble.

The addition of an Interquartile Range (IQR) data preparation strategy to handle outliers in the dataset is the novel aspect of our approach. This allows us to more effectively analyze the data. Our goal is to demonstrate the efficacy of outlier removal in improving the predictive performance of our model by contrasting the results that were obtained using GridSearchCV in combination with IQR and those that were obtained without IQR. This in-depth study will shed light on the potential advantages of outlier management in healthcare analytics while also providing insights into the influence that data preprocessing has on the accuracy and reliability of diabetes prediction.

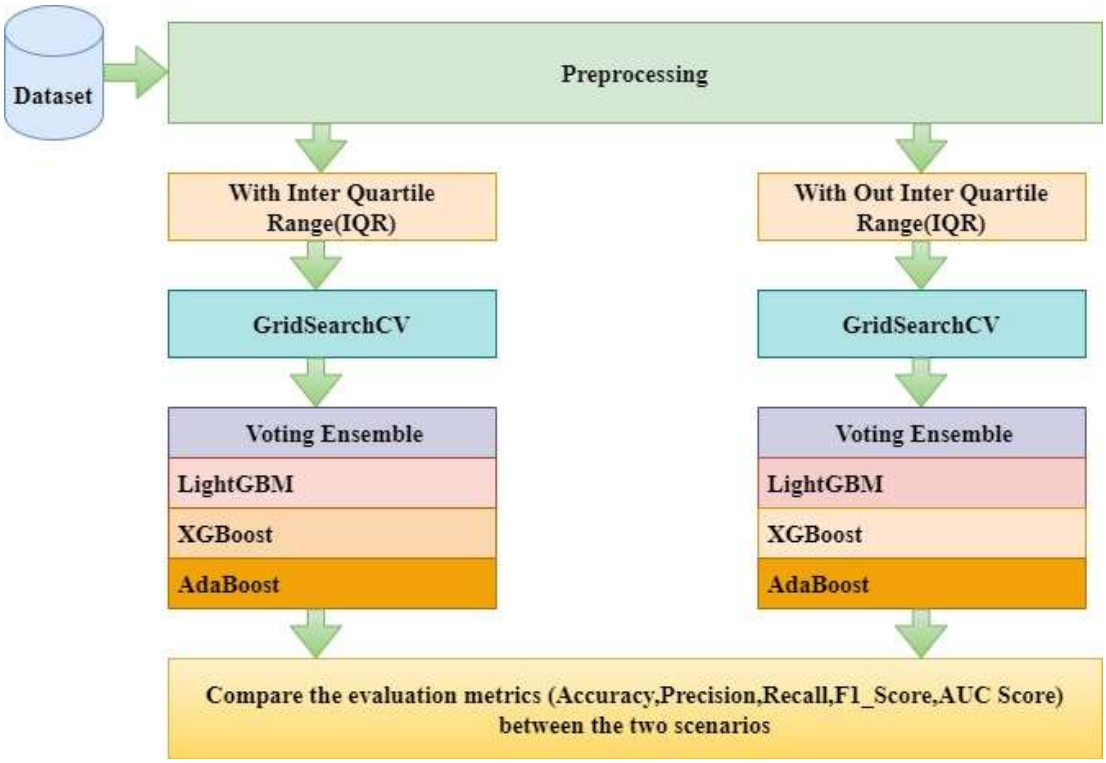

Figure 1: Comparing Voting Ensemble Approaches For Diabetes Prediction (Proposed Methodology)

3.1. Pre-processing

The primary objective of our study was to preprocess the diabetes dataset via the use of two essential techniques: data cleaning and the Interquartile Range (IQR) approach.

3.1.1. Data cleaning

Data cleaning is an essential process that plays a crucial role in maintaining the overall quality and integrity of a dataset. The process includes the identification and management of missing values, inconsistencies, and mistakes within the dataset. In the context of predicting diabetes, it is important to emphasize the significance of this particular stage in order to guarantee that the model is trained using precise and dependable data.

3.1.2. Interquartile Range (IQR)

The Interquartile Range (IQR) approach is a statistical technique often used in the identification and management of outliers within a dataset. The presence of outliers, which are data points that deviate greatly from the majority, has the potential to inject both noise and bias into the model. Through the use of the interquartile range (IQR) approach [12], outliers were successfully capping and then eliminated, leading to a more refined dataset. Our objective was to improve the data quality, minimize noise, and boost the accuracy and reliability of our diabetes prediction model by integrating data cleaning techniques with the interquartile range (IQR) approach. The inclusion of this preprocessing stage is crucial in guaranteeing that the model's forecasts are founded on data of superior quality, eventually resulting in enhanced healthcare outcomes.

Interquartile Range (IQR)

1. The Interquartile Range (IQR) is a statistical measure used to describe the spread or dispersion of a dataset. It is calculated as the difference between the upper

2. Determine the values of the first quartile (Q1) and the third quartile (Q3) for the given dataset.

3. The interquartile range (IQR) may be calculated by finding the difference between the third quartile (Q3) and the first quartile (Q1), expressed as IQR = Q3 - Q1.

4. The terms "lower bound" and "upper bound" are defined as follows:

4.1. The lower bound may be calculated by subtracting 1.5 times the interquartile range (IQR) from the first quartile (Q1). Similarly, the upper bound can be obtained by adding 1.5 times the IQR to the third quartile (Q3). In order to identify and address outliers, any data points that fall below the lower limit or above the upper bound are deemed outliers and are either eliminated from the dataset or adjusted to the corresponding bound.

The overview incorporates histograms and boxplots as visual aids to represent the distribution of data and identify the presence of outliers.

Before Capping (Uncapped Data):

The histograms illustrate the distribution of values for the chosen variables (Age, BMI, and Glucose) without any outlier handling. You may observe the early dispersion of data, including potential outliers.

Boxplots provide a visual representation of the statistical summary of each column, encompassing key measures such as the median, quartiles, and outliers. Outliers manifest as distinct data points that lie outside the range represented by the whiskers.

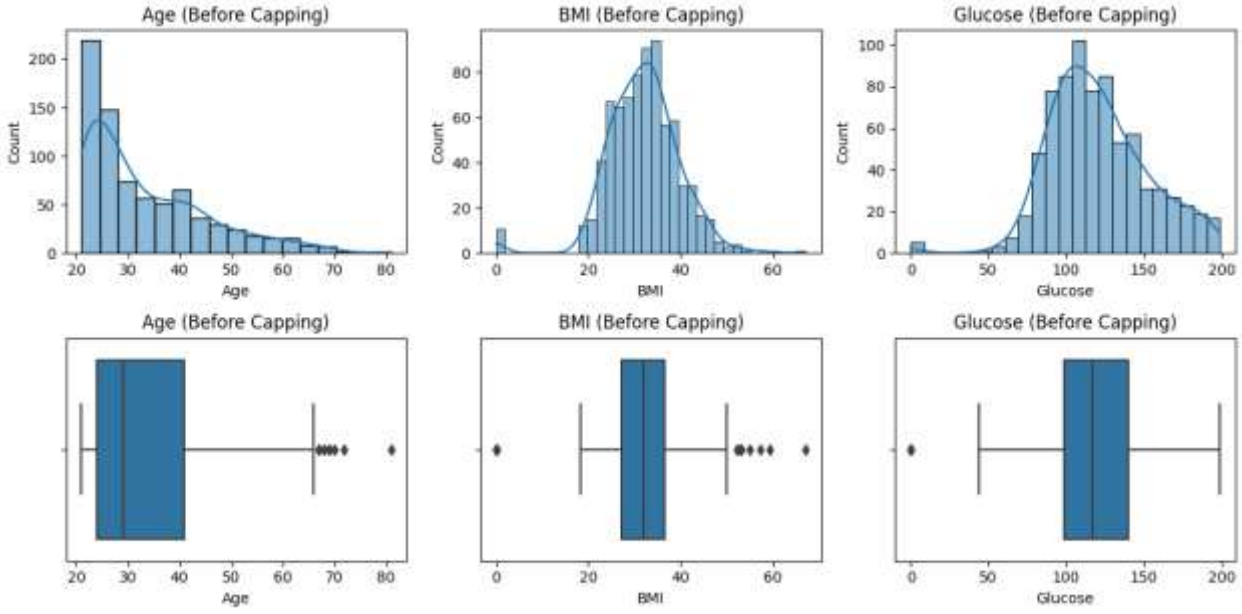

Figure 2: Before Capping (Age, BMI, Glucose)

After Capping (Capped Data):

The histograms now accurately represent the distribution of the data following the use of the capping strategy, which effectively eliminates or restricts the presence of severe outliers. The manner in which the data is limited within specific boundaries can be observed.

The boxplots exhibit a diminished occurrence of outliers subsequent to the application of capping. The data enclosed inside the whiskers now reflects the capped values, hence eliminating the inclusion of extreme data points.

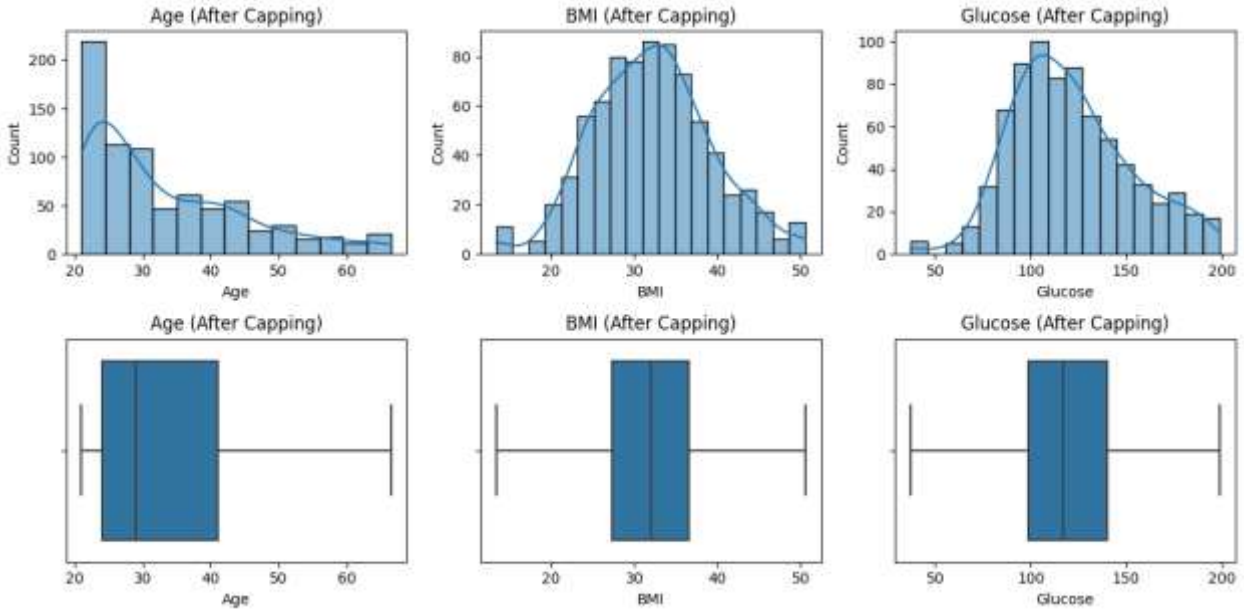

Figure 3: After Capping (Age, BMI, Glucose)

3.2. GridSearchCV

GridSearchCV, also known as Grid Search Cross-Validation, is a crucial machine learning approach for hyperparameter optimization. Model hyperparameters are configuration options that are not learnt from data yet affect performance. GridSearchCV methodically evaluates several hyperparameter combinations to find the best one [13].

The technique comprises specifying hyperparameter values, creating a grid or matrix with all possible combinations, and using cross-validation. Machine learning and statistical modeling employ cross-validation extensively. It divides a dataset into folds.

The model is trained repeatedly on one fold and assessed on the remaining data. It provides a complete evaluation of the model's performance and generalization. The above approach is repeated to ensure a complete evaluation of the model's efficacy across hyperparameter settings.

Cross-validation results are used to calculate accuracy, F1 score, and mean squared error for each combination. Find the best hyperparameters for model performance. The performance indicator depends on the problem and evaluation criteria. Hyperparameter tuning is hard and time-consuming, but GridSearchCV [14] automates it. This makes it vital for enhancing model prediction accuracy and ensuring optimal performance. Fine-tuning models is essential to machine learning because it unlocks their full potential.

GridSearchCV

1. The GridSearchCV is a method used in machine learning for hyperparameter tuning. It systematically chooses a machine learning model, such as LightGBM, XGBoost, or AdaBoost, and establishes a hyperparameter grid. This grid should outline the hyperparameters that will be optimized and the range of values that they might potentially assume.

2. The dataset should be partitioned into separate training and validation sets, often employing cross-validation to ensure robustness.

3. The model should be trained on the training data for each combination of hyperparameters.

4. Assess the performance of the model on the validation set by employing a designated scoring metric, such as accuracy or F1 score.

5. The process of repeating steps 3 and 4 should be conducted for all possible combinations of hyperparameters.

6. Choose the hyperparameters that result in optimal performance on the validation set.

3.3. Combine the Models into a Voting Ensemble for Improved Diabetes Performance Prediction

This research study presents a unique methodology for predicting diabetes by leveraging machine learning methodologies. The core aspect of our suggested methodology is around the utilization of a GridSearchCV technique to effectively tune the hyperparameters of a Voting Ensemble model. This ensemble comprises three robust machine learning methods, namely LightGBM, XGBoost, and AdaBoost.

3.3.1. LightGBM

Microsoft developed LightGBM, a powerful gradient boosting algorithm. The tool's speed and efficiency have made it essential for managing large datasets and complex tasks in the machine learning and data science sectors. Managing categorical attributes directly is a LightGBM strength[15]. LightGBM does not require one-hot encoding for categorical input, unlike other machine learning methods. This saves memory and speeds training. Histogram-based learning is also used by LightGBM. Binning ongoing features reduces memory utilization and boosts system efficiency.

The framework uses gradient boosting to combine weak models, commonly decision trees, to create a strong prediction model. Using leaf-wise development, LightGBM is unique [16]. Selecting the leaf with the largest delta loss helps the tree develops, preventing overfitting and speeding training. LightGBM [16] can do distributed computing and optimize the model with many adjustable parameters. This method is suitable for classification, regression, ranking, and recommendation systems because to its versatility.

In this regard, LightGBM is a quick and accurate machine learning framework. Its unique traits make it ideal for complex large dataset challenges. It is used in healthcare, finance, and other industries that value efficiency and performance.

3.3.2. XGBoost

XGBoost, or Extreme Gradient Boosting, is an efficient and versatile machine learning algorithm. The system excels in expected accuracy and model durability. Tianqi Chen's XGBoost technique is famous in machine learning challenges and practical applications. Ensemble learning[17,18] uses predictions from several models, primarily decision trees, to create a robust and accurate model like this one. XGBoost optimizes gradient boosting, handles missing data, uses regularization, and uses parallel processing, making it beneficial. The suggested method uses "pruning" to reduce overfitting and improve model generalization. Shallow trees and robust regularization help XGBoost [19] balance performance and model complexity.

Additionally, XGBoost gives a significance score that quantifies feature value, aiding feature selection and interpretability. The system supports classification, regression, and ranking, proving its adaptability in a variety of applications. The speed, accuracy, and ability to efficiently handle massive datasets have made XGBoost a top choice for data scientists and machine learning practitioners, advancing high-performance prediction models.

3.3.3. AdaBoost

Adaptive Boosting, or AdaBoost, is a machine learning ensemble approach that combines weak learners into a robust and accurate prediction model[20]. Robert E. Schapire and Yoav Freund introduced this idea.

The core of AdaBoost is iteratively training weak models, usually decision trees or stumps. This training method emphasizes data items identified erroneously in previous cycles. The adaptive procedure in this scenario weights training examples differently, emphasizing those with higher classification difficulty. Thus, AdaBoost algorithm uses these weak learners' predictions to build a robust ensemble model with improved prediction accuracy through collective decision-making.

AdaBoost excels in circumstances when other algorithms fail [21]. It is used in face identification, text classification, and object recognition. The system's ability to learn from its mistakes during training makes it useful for machine learning classification and regression applications.

Voting Ensemble with LightGBM, XGBoost, and AdaBoost

1. The proposed approach involves the use of a voting ensemble technique that combines three powerful machine learning algorithms, namely LightGBM, XGBoost, and AdaBoost.
2. The dataset was utilized to train separate machine learning models, namely LightGBM, XGBoost, and AdaBoost. Each model was trained with its own distinct set of hyperparameters.
3. The combination of predictions from many models may be achieved by employing a weighted average approach. The weights may be determined by evaluating the performance of each model on the validation set.
4. The forecast of the ensemble model is determined by calculating the weighted average of the predictions generated by each individual model.
5. The utilization of an ensemble methodology capitalizes on the individual strengths of each model in order to enhance the overall accuracy of predictions.

6. In the training phase, the weights for each model can be calculated using approaches like as cross-validation or by considering their individual performance.

7. The Voting Ensemble is a technique that amalgamates the predictions generated by many models, so possibly mitigating the issue of overfitting and augmenting the model's ability to generalize.

4. Result and Discussion

4.1. Performance metrics

A common statistic is accuracy, which measures the ratio of successfully predicted instances to the total number of examples in a dataset. The measure assesses model performance broadly.
Precision, or Positive Predictive Value, is the ratio of genuine positive forecasts to the total number of positive predictions. Minimizing false positives is crucial.

The percentage of correct positive forecasts to the total number of positive cases is called recall, sensitivity, or true positive rate. When accuracy is crucial, false negatives must be minimized. The harmonic mean of accuracy and recall is the F1 Score. The metric trades accuracy and recall, making it suitable for datasets with an uneven class distribution.
The Receiver Operating Characteristic Area Under the Curve (ROC AUC) measures a binary classification model's ability to differentiate positive and negative classes. Larger numbers indicate better model performance in the region below the Receiver Operating Characteristic (ROC) curve.
These metrics are essential for assessing machine learning models, especially in binary classification problems. Each measure has a certain purpose, and prioritizing one depends on the situation. To reduce false negatives in medical diagnosis, excellent recall is essential. In spam detection, accuracy is prioritized to reduce false positives.

4.2. Performance Assessments

4.2.1 Voting Ensemble Classifier with IQR result analysis

The results of the implementation of the Voting Ensemble Classifier with IQR preprocessing indicate that the model has an adequate degree of accuracy in its predictions. The model displays an ability to predict outcomes with an accuracy of around 78.73% by correctly identifying the examples contained in the dataset at the same percentage.
In addition, the model has a precision rate of 72.72%, which indicates that it achieves a high degree of accuracy when predicting favorable outcomes (about 72.72%). This observation demonstrates how accurate the model's positive predictions may be by highlighting its ability to generate them. Nevertheless, the sensitivity of the model, which is assessed by a recall rate of 55.55%, shows that it is only capable of detecting 55.55% of the real positive events, potentially missing a considerable proportion of them in the process. This is because the recall rate of 55.55% is how the sensitivity of the model is calculated. The F1 score of 62.99% demonstrates, however, that the model successfully strikes a balance between the accuracy and the recall of its predictions. The area under the curve (AUC) score, which was calculated to be 83.68%, demonstrates how well the model is able to reliably differentiate between positive and negative classes. In a nutshell, the model demonstrates an outstanding level of accuracy and precision. However, there is room for improvement in terms of recollection, particularly if the identification of positive occurrences is the most important thing to focus on in the specific context.

Table 1: Voting Ensemble Classifier with IQR Performance Metrics

Voting Ensemble Classifier with IQR Results	
Metrics	***Values***
Accuracy	78.73
Precision	72.72
Recall	55.55
f1_score	62.99
AUC Score	83.68

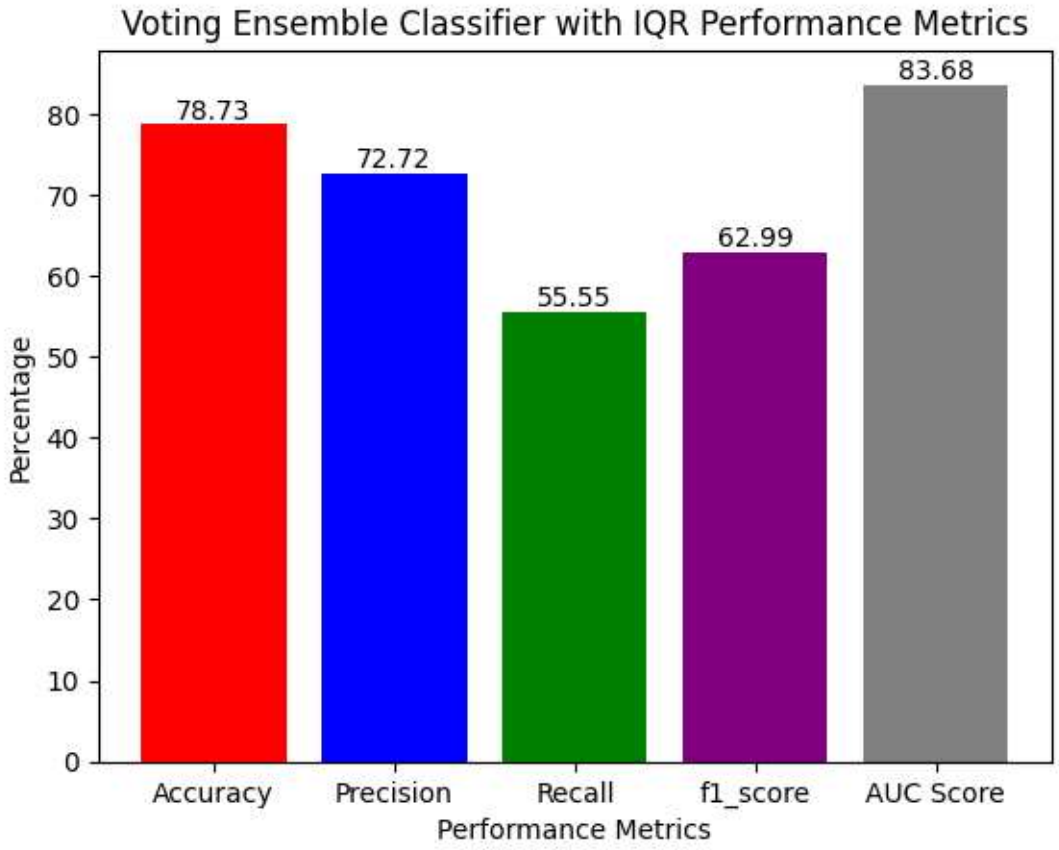

Figure 4: Bar Graph shows Voting Ensemble Classifier with IQR Performance Metrics

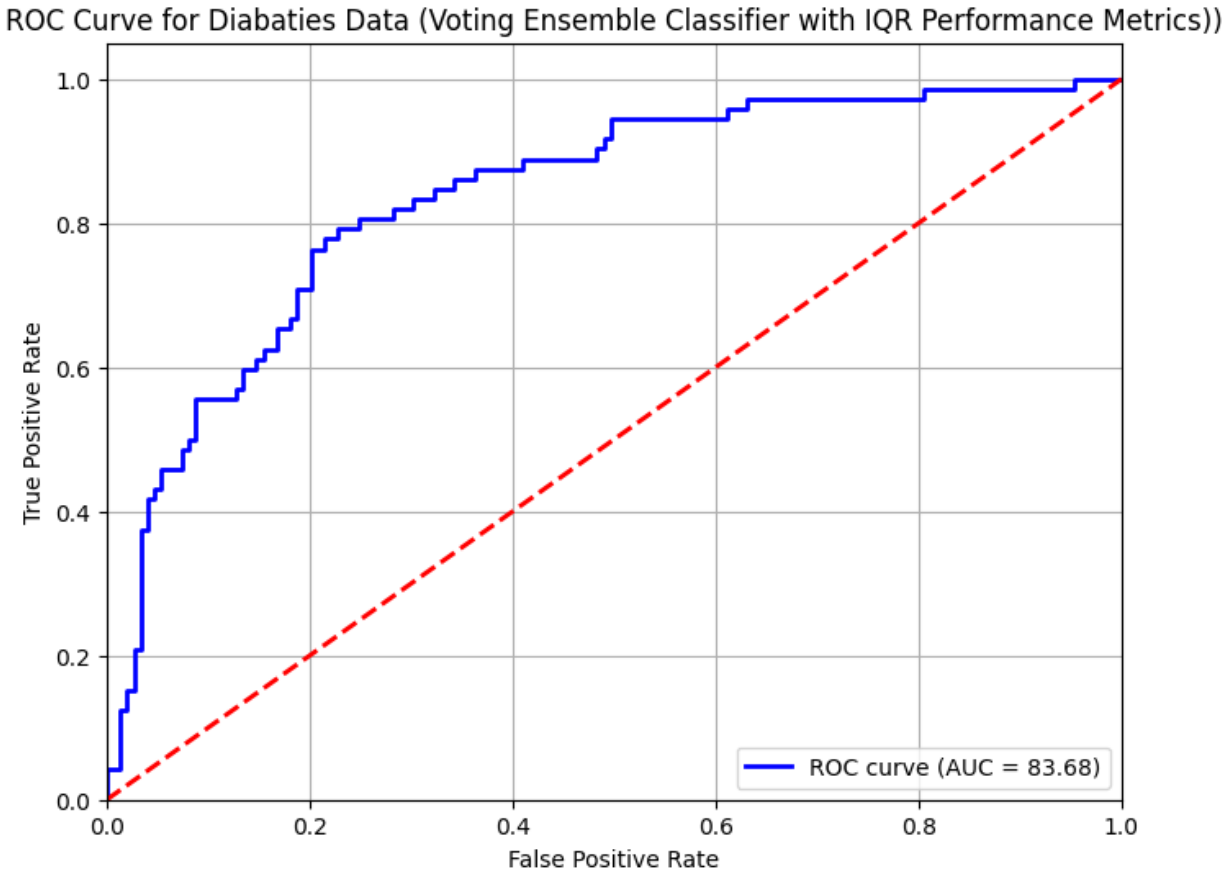

Figure 5: ROC Curve of Voting Ensemble Classifier with IQR Performance Metrics

4.2.2 Voting Ensemble Classifier without IQR result analysis

The results of the Voting Ensemble Classifier, when not subjected to IQR preprocessing, produce a model with prediction skills that are sufficient. The fact that an accuracy of 76.19% was attained indicates that about 76.19% of the instances contained within the dataset were correctly classified, hence validating the overall efficiency of

its predictive powers. The accuracy, which was tested at 65.82%, indicating that the model is approximately capable of properly predicting positive outcomes. This may be inferred from the fact that the accuracy was assessed. This lends credence to the notion that the model is capable of exhibiting a remarkable skill in producing accurate positive predictions. In addition, the recall rate, which is also commonly referred to as sensitivity, demonstrates that the model is capable of recognizing positive scenarios due to the fact that it is able to recognize 65.0% of the real positive occurrences. This suggests that the model is sensitive. The F1 score, which is now at 65.40%, displays a harmonic equilibrium between accuracy and recall, which emphasizes the model's capacity to deliver both exact and comprehensive positive predictions. In other words, the score demonstrates a harmonious balance between accuracy and recall. The area under the receiver operating curve (AUC) score that was calculated for the model was 81.44%, which indicates that it successfully differentiates between positive and negative classifications. To summarize, the predictive power of the model is rather high since it demonstrates an appropriate level of both accuracy and recall. In addition, the skill with which it can differentiate between positive and negative categories is glaringly visible. However, there is a slight deviation from the outcomes achieved by IQR preprocessing in terms of the accuracy and precision of the results acquired by this method.

Table 2: Voting Ensemble Classifier without IQR Performance Metrics

Voting Ensemble Classifier without IQR Results	
Metrics	***Values***
Accuracy	76.19
Precision	65.82
Recall	65.0
f1_score	65.40
AUC Score	81.44

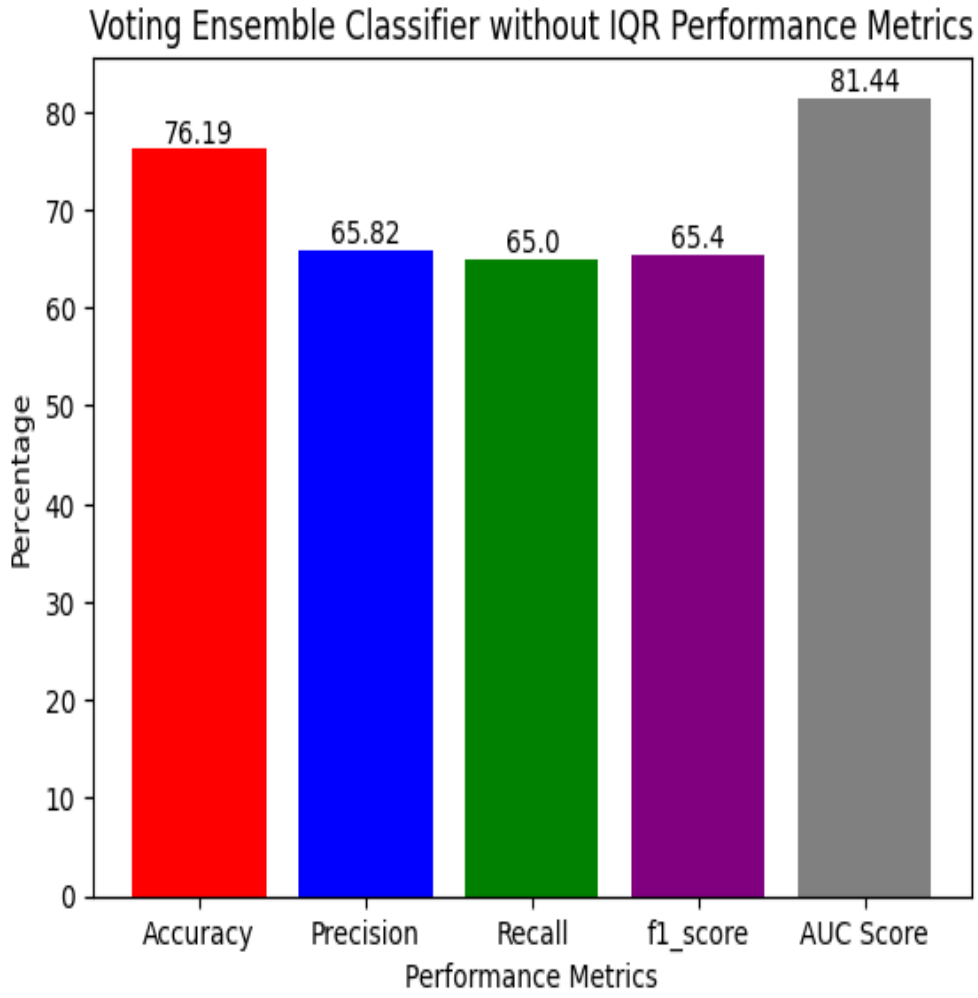

Figure 6: Bar Graph shows Voting Ensemble Classifier without IQR Performance Metrics

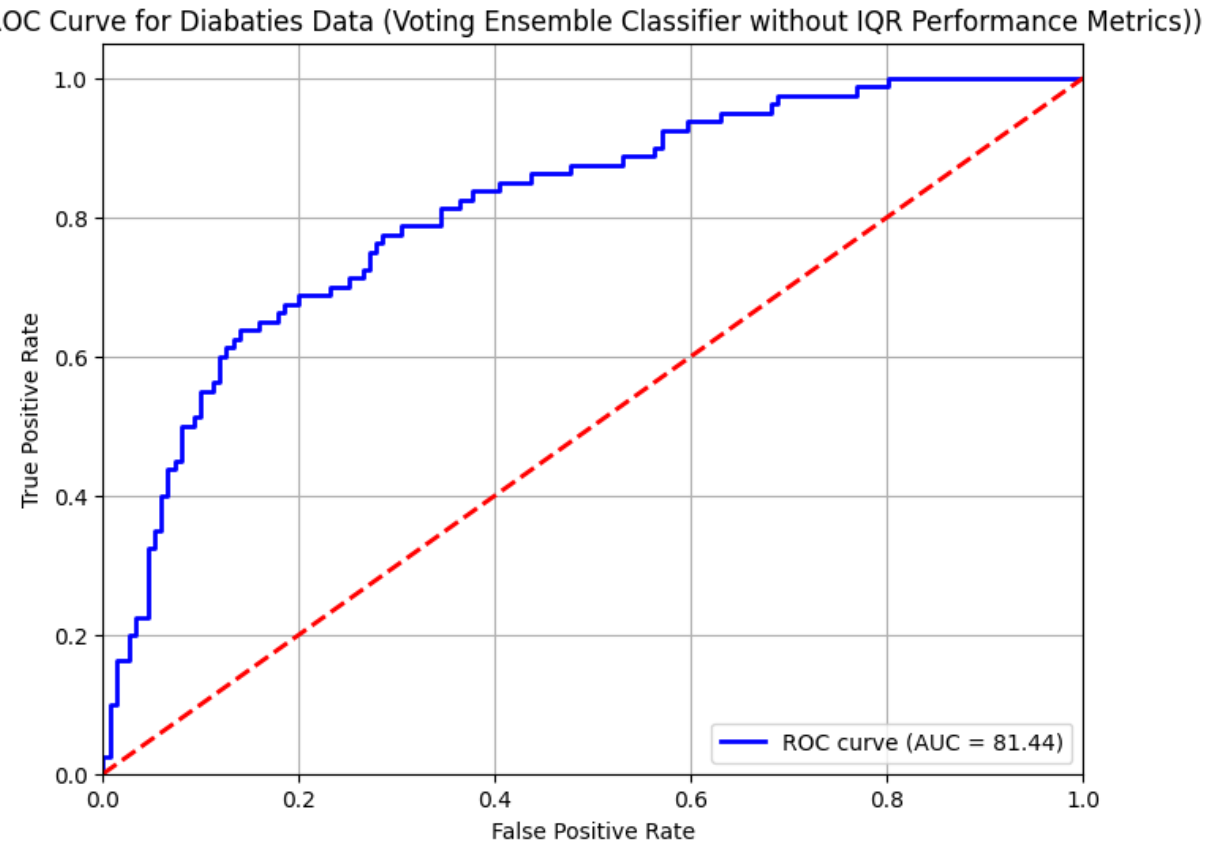

Figure 7: ROC curve of Voting Ensemble Classifier without IQR Performance Metrics

4.2.3 Performance Comparison of Voting Ensemble Classifier with IQR and Without IQR

Comparing two voting ensemble models with and without IQR preprocessing reveals diabetes diagnosis prediction effectiveness. The "Voting with IQR" model is more accurate, at 78.83%. This means it can appropriately label roughly 79% of dataset occurrences, making it the more reliable diabetes predictor. A accuracy rate of 72.72% indicates the model's capacity to make reliable positive predictions, which is crucial in healthcare. However, its recall (sensitivity) rate of 55.55% may miss some real positives. Despite these challenges, it earns a 62.99% F1 score. The model's 83.68% AUC score also shows its accuracy in distinguishing diabetics from non-diabetics.

However, "Voting without IQR" performs well with 76.19% accuracy. Its accuracy and recall rates are good, earning a 65.40% F1 grade. The model's AUC score of 81.44% shows its discrimination across classes; however it is less robust than the IQR-preprocessed model.

In conclusion, the "Voting with IQR" ensemble has greater accuracy and balance in diabetes prediction, while the "Voting without IQR" ensemble is still a feasible option. The former method has better accuracy, while the latter offers a good trade-off between precision and recall, depending on a healthcare application.

Table 3: Performance Comparison of Voting Ensemble Classifier with IQR and Without IQR

Algorithm	***Accuracy***	***Precision***	***Recall***	***f1_score***	***AUC Score***
Voting with IQR	78.83	72.72	55.55	62.99	83.68
Voting without IQR	76.19	65.82	65.0	65.40	81.44

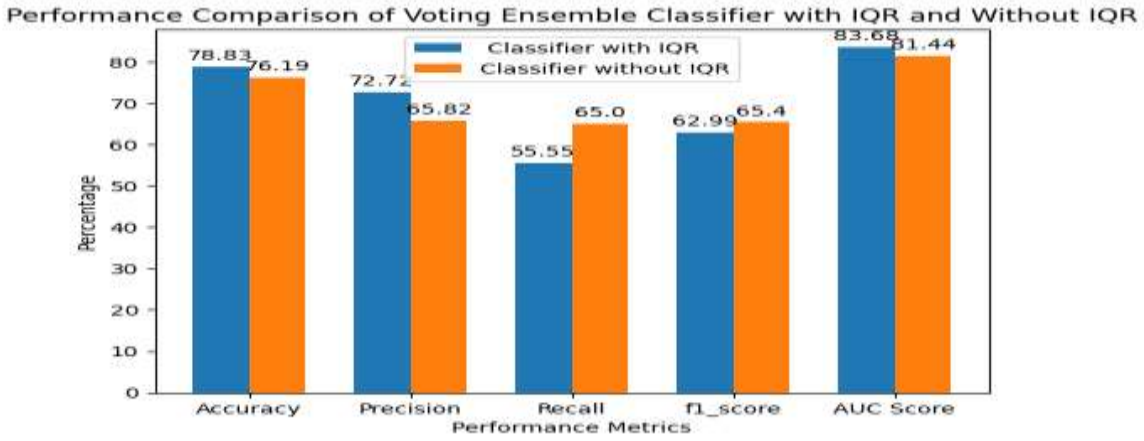

Figure 8: Bar Graph shows Performance Comparison of Voting Ensemble Classifier with IQR and Without IQR

5. Conclusion

Diabetes is a serious worldwide public health concern that has to be identified with pinpoint accuracy and in a timely manner in order to support successful treatment and control efforts. This paper highlights the efficacy of Voting Ensembles, LightGBM, XGBoost, and AdaBoost algorithms, which are some of the machine learning approaches that have shown promise in the field of diabetes prediction. The use of GridSearchCV for the goal of hyperparameter optimization, in conjunction with the use of Interquartile Range (IQR) preprocessing, has shown notable improvements in both the accuracy of predictions and the robustness of the model. GridSearchCV was developed by Microsoft Research and is available as a free download from the company's website. The Voting Ensemble model, which was preprocessed via the Interquartile Range (IQR) approach, displayed improved performance by achieving high levels of accuracy and precision, as well as a high Area Under the Curve (AUC) score. The fact that the method that does not employ the interquartile range (IQR) nonetheless produces good results demonstrates both its applicability and its efficiency. The utilization of the interquartile range (IQR) preprocessing method has served as an illustration of the relevance of outlier control in the field of healthcare analytics. The findings of this research not only provide an important contribution to the advancement of diabetes diagnosis, but they also highlight the crucial role that machine learning plays in the applications that are used in the healthcare industry. It is necessary to do further research in order to investigate datasets that are more extensive and to take into account the practical application of results in order to improve the outcomes for patients and the quality of therapy.

References

[1] Jaiswal, Varun, Anjli Negi, and Tarun Pal. “A Review on Current Advances in Machine Learning Based Diabetes Prediction.” Primary Care Diabetes 15, no. 3 (June 2021): 435–43. https://doi.org/10.1016/j.pcd.2021.02.005.

[2] Krishna, T. B. M., Praveen, S. P., Ahmed, S., & Srinivasu, P. N. (2023). Software-driven secure framework for mobile healthcare applications in IoMT. Intelligent Decision Technologies, 17(2), 377-393.

[3] “Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey.” Tehnicki Vjesnik - Technical Gazette 26, no. 3 (June 2019). https://doi.org/10.17559/tv-20190421122826.

[4] Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. “Machine Learning and Data Mining Methods in Diabetes Research.” Computational and Structural Biotechnology Journal 15 (2017): 104–16. https://doi.org/10.1016/j.csbj.2016.12.005.

[5] Phani Praveen, S., Hasan Ali, M., Musa Jaber, M., Buddhi, D., Prakash, C., Rani, D. R., & Thirugnanam, T. (2023). IoT-Enabled Healthcare Data Analysis in Virtual Hospital Systems Using Industry 4.0 Smart Manufacturing. International Journal of Pattern Recognition and Artificial Intelligence, 37(02), 2356002.

[6] Ashish Dixit,R. P. Aggarwal,B. K. Sharma,Aditi Sharma. "Safeguarding Digital Essence: A Sub-band DCT Neural Watermarking Paradigm Leveraging GRNN and CNN for Unyielding Image Protection and Identification." Journal of Intelligent Systems and Internet of Things, Vol. 10, No. 1, 2023 ,PP. 33-47.

[7] Neyda Hernández Bandera,Jenny M. Moya Arizaga,Enrique Rodríguez Reyes. "Assessment and prediction of Chronic Kidney using an improved neutrosophic artificial intelligence model." International Journal of Neutrosophic Science, Vol. 21, No. 1, 2023 ,PP. 174-183.

[8] Doaa Sami Khafaga,Abdelhameed Ibrahim,S. K. Towfek,Nima Khodadadi. "Data Mining Techniques in Predictive Medicine: An Application in hemodynamic prediction for abdominal aortic aneurysm disease." Journal of Artificial Intelligence and Metaheuristics, Vol. 5, No. 1, 2023 ,PP. 29-37.

[9] Ahmed, Usama, Ghassan F. Issa, Muhammad Adnan Khan, Shabib Aftab, Muhammad Farhan Khan, Raed A. T. Said, Taher M. Ghazal, and Munir Ahmad. “Prediction of Diabetes Empowered With Fused Machine Learning.” IEEE Access 10 (2022): 8529–38. https://doi.org/10.1109/access.2022.3142097.
[10] Reem Atassi,Aditi Sharma. "An Efficient and Secured Triple-Layered Wireless Sensor Network with Machine Learning Techniques." International Journal of Wireless and Ad Hoc Communication, Vol. 6, No. 2, 2023 ,PP. 08-17.
[11] Gajender Kumar,Vinod Patidar,Prolay Biswas,Mukta Patel,Chaur Singh Rajput,Anita Venugopal,Aditi Sharma. "IOT enabled Intelligent featured imaging Bone Fractured Detection System." Journal of Intelligent Systems and Internet of Things, Vol. 9, No. 2, 2023 ,PP. 08-22.
[12] Ljubic, Branimir, Ameen Abdel Hai, Marija Stanojevic, Wilson Diaz, Daniel Polimac, Martin Pavlovski, and Zoran Obradovic. “Predicting Complications of Diabetes Mellitus Using Advanced Machine Learning Algorithms.” Journal of the American Medical Informatics Association 27, no. 9 (September 1, 2020): 1343–51. https://doi.org/10.1093/jamia/ocaa120.
[13] A. Yuva Krishna,K. Ravi Kiran,N. Raghavendra Sai,Aditi Sharma,S. Phani Praveen,Jitendra Pandey. (2023). Ant Colony Optimized XGBoost for Early Diabetes Detection: A Hybrid Approach in Machine Learning. Journal of Intelligent Systems and Internet of Things, 10 (2), 76-89.
[14] Chang, Victor, Meghana Ashok Ganatra, Karl Hall, Lewis Golightly, and Qianwen Ariel Xu. “An Assessment of Machine Learning Models and Algorithms for Early Prediction and Diagnosis of Diabetes Using Health Indicators.” Healthcare Analytics 2 (November 2022): 100118. https://doi.org/10.1016/j.health.2022.100118.
[15] K. Arava, C. Paritala, V. Shariff, S. P. Praveen and A. Madhuri, "A Generalized Model for Identifying Fake Digital Images through the Application of Deep Learning," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 1144-1147, doi: 10.1109/ICESC54411.2022.9885341.
[16] Reddy, Shiva, Nilambar Sethi, R. Rajender, and Gadiraju Mahesh. “Forecasting Diabetes Correlated Non-Alcoholic Fatty Liver Disease by Exploiting Naïve Bayes Tree.” ICST Transactions on Scalable Information Systems, July 13, 2018, 173975. https://doi.org/10.4108/eai.29-4-2022.173975.
[17] S. P. Praveen, S. Sindhura, P. N. Srinivasu and S. Ahmed, "Combining CNNs and Bi-LSTMs for Enhanced Network Intrusion Detection: A Deep Learning Approach," 2023 3rd International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 2023, pp. 261-268, doi: 10.1109/ICCIT58132.2023.10273871.
[18] Ashish Patel,Richa Mishra ,Aditi Sharma. "Maize Plant Leaf Disease Classification Using Supervised Machine Learning Algorithms." Fusion: Practice and Applications, Vol. 13, No. 2, 2023 ,PP. 08-21.
[19] B. V. Marrapu, K. Y. N. Raju, M. J. Chowdary, H. Vempati and S. Phani Praveen, "Automating the Creation of Machine Learning Algorithms using basic Math," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 866-871, doi: 10.1109/ICSSIT53264.2022.9716270.
[20] Mangkunegara, Iis Setiawan, and Purwono Purwono. “Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV.” 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), June 16, 2022. https://doi.org/10.1109/cyberneticscom55287.2022.9865624.
[21] Shehadeh, Ali, Odey Alshboul, Rabia Emhamed Al Mamlook, and Ola Hamedat. “Machine Learning Models for Predicting the Residual Value of Heavy Construction Equipment: An Evaluation of Modified Decision Tree, LightGBM, and XGBoost Regression.” Automation in Construction 129 (September 2021): 103827. https://doi.org/10.1016/j.autcon.2021.103827.
[22] Rufo, Derara Duba, Taye Girma Debelee, Achim Ibenthal, and Worku Gachena Negera. “Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM).” Diagnostics 11, no. 9 (September 19, 2021): 1714. https://doi.org/10.3390/diagnostics11091714.
[23] Sirisha, U., & Chandana, B. S. (2023). Privacy preserving image encryption with optimal deep transfer learning based accident severity classification model. *Sensors*, *23*(1), 519.
[24] Sirisha, U., & Chandana, B. S. (2023). Utilizing a Hybrid Model for Human Injury Severity Analysis in Traffic Accidents. *Traitement du Signal*, *40*(5).
[25] B. Narasimha Swamy,Rajeswari Nakka,Aditi Sharma,S. Phani Praveen,Venkata Nagaraju Thatha,Kumar Gautam. (2023). An Ensemble Learning Approach for detection of Chronic Kidney Disease (CKD). Journal of Intelligent Systems and Internet of Things, 10 (2), 38-48.
[26] Bahad, Pritika, and Preeti Saxena. “Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics.” International Conference on Intelligent Computing and Smart Communication 2019, December 20, 2019, 235–44. https://doi.org/10.1007/978-981-15-0633-8_22.
[27] Haq, Amin Ul, Jian Ping Li, Jalaluddin Khan, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad, Ghufran Ahmad Khan, and Amjad Ali. “Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data.” Sensors 20, no. 9 (May 6, 2020): 2649. https://doi.org/10.3390/s20092649.