International Journal of Computer Applications (0975 - 8887)
Volume 186 - No.49, November 2024

Gradient Edge: Advancing Predictive Modelling with Enhanced Gradient Boosting: A Multi-Dataset Approach

Samuel Benny Varghese
Student, MSc Computer Science
Mary Matha Arts and Science College
Mananthavady, Wayanad, Kerala

Eldho K.J.,PhD
Assistant Professor
Mary Matha Arts and Science College
Mananthavady, Wayanad, Kerala

ABSTRACT

Using gradient boosting, it assesses the effectiveness of many machine learning algorithms on three datasets: water potability, diabetes, and heart disease. Its goal is to evaluate these models' ability to forecast various environmental and health events. Since every dataset presents unique difficulties, a thorough understanding of the algorithms' advantages and disadvantages is possible. The goal of this study's introduction of Enhanced Gradient Boosting is to advance the field by improving prediction accuracy. The study's enhanced method will overcome the drawbacks of the traditional Gradient Boosting approach in managing the complexities present in diverse datasets. Standard Gradient Boosting performs poorly on the water potability dataset, particularly when it comes to class separation. For example, class 0 and class 1 model precisions are equivalent to 0.66 and 0.64, respectively, with recall rates of 0.93 and 0.20 and F1-scores of 0.78 and 0.31. The typical Gradient Boosting model performed well, with an accuracy of 0.66, on both the diabetes and heart disease datasets. On the other hand, the new method outperformed the old one, particularly when handling the noisy water potability data. An organized method that includes the following phases is used to construct the Enhanced Gradient Boosting model: data collecting, data preprocessing, EDA, data splitting, and scaling. This end-to-end method shows how creative algorithm creation combined with comparison analysis may lead to tailored machine learning solutions. These findings show how well the algorithm performed on the two datasets and highlight how versatile Gradient Boosting is for solving various prediction issues. The discovered results have great significance not only for applications related to medical diagnostics and environmental monitoring, but also for paving the way for future advancements within the machine learning research framework.

KEYWORDS

Gradient Boosting, Machine Learning, Predictive Modelling, Healthcare Diagnostics, Environmental Monitoring

1 INTRODUCTION

Predictive modeling in the environmental and health sciences is facilitated by machine learning methods. This study will compare the efficacy of support vector machines, decision trees, and gradient boosting on three distinct datasets: diabetes, heart disease, and water potability. Understanding how different algorithms handle the unique issues that each dataset presents and those that impact the model's performance is necessary for effective predictive analytics. Given its exceptional accuracy for the datasets connected to diabetes and heart disease, gradient boosting is one of the more outstanding methods that have been explored. Gradient Boosting performed more accurately than Random Forest and Decision Trees on the diabetes dataset. This will effectively demonstrate how it manages complex healthcare data with the presence of both class imbalances and high-dimensional features. Gradient Boosting demonstrated exceptional performance in the heart disease dataset, demonstrating its resilience in medical diagnosis. Based on the diabetes and heart disease datasets,[10] gradient boosting performs exceptionally well as an ensemble technique. It is renowned for how effectively it manages complex datasets. In the diabetes dataset, gradient boosting produced the greatest accuracy compared to Random Forest and Decision Trees. That should suggest that it handled high-dimensional health care data and class imbalances well. Gradient Boosting, on the other hand, demonstrated identical performance in the heart disease dataset when taking into account SVM's outstanding performance, which also demonstrated its adaptability and dependability in medical diagnosis. This will effectively demonstrate how it manages complex healthcare data with the presence of both class imbalances and high-dimensional features.[7] Gradient Boosting demonstrated exceptional performance in the heart disease dataset, demonstrating its resilience in medical diagnosis. Based on the diabetes and heart disease datasets, gradient boosting performs exceptionally well as an ensemble technique. It is renowned for how effectively it manages complex datasets. In the diabetes dataset, gradient boosting produced the greatest accuracy compared to Random Forest and Decision Trees. That should suggest that it handled high-dimensional health care data and class imbalances well.

Gradient Boosting, on the other hand, demonstrated identical performance in the heart disease dataset when taking into account SVM's outstanding performance, which also demonstrated its adaptability and dependability in datssets. This paper highlights the amazing potential of gradient boosting in predictive modeling while acknowledging the heterogeneity in its effectiveness across different data sets, providing insightful information about both the technique's advantages and disadvantages. One of the major concerns in today's environmental management and public health is safeguarding and preserving the safety and purity of the water. Although everyone needs clean drinking water, millions of people worldwide are at serious danger for health problems due to contaminated water sources. As a result, it is crucial to have extremely precise and efficient water quality monitoring systems. Using the most recent machine learning techniques to evaluate and forecast water potability is perhaps the most promising approach. This will improve the ability to identify polluted water sources and enable timely remedies. In this paper, we investigated the usefulness of a Gradient Boosting Classifier in predicting the potability of water using an extensive dataset of water quality variables. Each characteristic in the dataset has a specific significance in evaluating the water's pH, hardness, sediment, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity. Using this machine learning model, a tool that can reliably identify whether water samples are drinkable or not will be constructed. Preprocessing to handle missing data and feature scaling, using GridSearchCV to adjust hyperparameters, and evaluating the model's performance using metrics like accuracy, confusion matrix, and classification report are some of the crucial components. The Gradient Boosting Classifier was used because it is a resilient method for handling complicated data interactions and may improve prediction accuracy through repeated learning. By providing a solid basis for the machine learning-based assessment of water quality, this research advances the field. The results of this study might be applied to improve water management strategies, bolster public health initiatives for pollution early detection, and guarantee safer water supplies.

2 LITERATURE REVIEW

Bai, George; Chandra,[2] Rohitash introduce a novel framework for Bayesian ensemble learning, combining fundamental neural networks with gradient boosting in a synergistic way, and utilizing Langevin gradient-based proposal distributions for improved model training. Although the variance-bias trade-off focuses on how Bayesian techniques combined with gradient boosting can improve prediction accuracy and more adeptly handle uncertainty, the approach they proposed is based on very basic considerations, such as ensemble learning methods themselves. The framework uses Markov Chain Monte Carlo sampling with enhanced Langevin gradients to address the model convergence and performance in time-series prediction and classification. These contributions, which expand on the integration of gradient boosting with Bayesian inference and provide fresh perspectives with practical machine learning applications, mark a significant advancement toward predictive modeling. This will not only address the shortcomings of the conventional methods, but it will also improve the prediction's resilience and accuracy, thereby paving the way for the development of machine learning-based solutions that are more dependable and efficient.

Mohapatra et al. [9]introduce a novel method for diagnosing CVD using ML. First, given that cardiovascular diseases (CVDs) are one of the leading causes of death worldwide, the authors discuss the critical need for faster and more accurate diagnostic techniques. They provide a prediction method for automated diagnosis that makes use of electronic health information and the stacking methodology. As a result, our method divided the output of several classifiers into two tiers, producing a robust model with 92 percentage accuracy, exceptional sensitivity, and precision. This study demonstrated the value of machine learning in the field of health, particularly in terms of improved prediction made possible by studies of large datasets such as electronic health records. By adding layers to a multi-layer model, this stacking ensemble approach enhances the performance of conventional classifiers, making it helpful for early CVD detection and intervention. Even though the study's results are encouraging, more conversation about the difficulties in using these models in clinical settings, such as data security and system integration, is needed. Overall, the work is a significant contribution to the discipline as it identifies a potential avenue for applying cutting-edge, cutting-edge machine learning approaches to diagnose cardiac conditions more quickly and accurately.

Kawarkhe et al.[5] introduce A novel method of predicting diabetes through the use of ensemble learning classifiers." High blood glucose levels are a hallmark of diabetes, a chronic illness brought on by either insufficient insulin synthesis or an insufficient body response to insulin. Its data is scarce, and the presence of outliers makes diagnosing the illness difficult. In response to these difficulties, the author suggests a model that uses ensemble classifiers on raw preprocessed data, including CatBoost, LDA, Random Forest, Gradient Boosting Classifiers, and Logistic Regression. This study used ensemble learning to integrate many classifiers and produced extremely accurate diabetes predictions, with a 90.62 percent accuracy rate. AUC and ROC Curves are the performance metrics that were employed to test the findings. In addition to improving prediction results, this might offer promise for better clinical studies that detect diabetes early. Even though this model has a lot of potential, more debate may provide more information about the difficulties in implementing such systems in actual healthcare settings, such as privacy and data integration. Overall, this study offers valuable information on the prognosis against diabetes and establishes a solid framework for additional study and real-world diagnostic applications.

Alawee et al. [1]Introduce machine learning techniques to do research on solar distillation technologies' productivity forecasts. The convex tubular sun still and the conventional tubular sun still are the two types of solar distillation systems that are taken into consideration in this study. The performances of both systems were examined in this work, and the efficient estimation of distillate production was achieved by applying gradient boosting, a sophisticated machine learning technique. The results indicate that CTSS was significantly more productive than TSS, suggesting that it may be a practical method for desalination. Given the gradient boosting model's R-squared value of 0.99 from the CTSS model and its RMSE value of 1.2 percentage and CVRMSE of 4 percentage, the study's gradient boosting model demonstrated amazing predictive power. In comparison, the R-squared value of 0.86, RMSE of 58.2%, and CVRMSE of 29.3% were found in the TSS model. Given that the CTSS model's R-squared value was 0.99, its RMSE was 1.2%, and its CVRMSE was 4%, the gradient boosting model that was constructed showed some really impressive prediction accuracy. In contrast, the TSS model produced an R-square value of 0.86 as well as RMSE and CVRMSE values of 58.2% and 29.3%, respectively. The current work will make a substantial contribution to the body of knowledge about solar still improvement.

Abdul Haq et al.[3] Introduced a method based on hybrid deep learning models to enhance aquaculture water quality prediction. In

order to forecast water quality in aquaculture systems, the current study evaluates the application of a hybrid deep learning technique that blends CNN with LSTM and GRU networks. The research examines how two distinct hybrid models perform on two distinct water quality datasets and shows how changes in hyperparameters affect performance. The investigation's findings demonstrate that the hybrid CNN-LSTM model outperforms a number of other models, including baseline LSTM and GRU, CNN, and their attention-based variations, in terms of prediction accuracy and computation time. This study combines the temporal dependency learning of LSTM and GRU with the feature extraction capacity of CNN to perform better in the job of water characteristics prediction. This work has effectively demonstrated how new aquaculture productivity gains may be inspired by hybrid deep learning approaches, which will greatly aid in the implementation of sustainable and successful aquaculture management practices.

Shihua Luo and Tianxin [4] Chen from Jiangxi University of Finance and Economics explore to forecast the amount of hot metal silicon in the blast furnace system, the two sophisticated gradient boosting algorithms—Extreme Gradient Boosting and Light Gradient Boosting Machine—will be discussed in the upcoming article. Predicting the silicon content of hot metal, one of its most crucial constituents, is essential for assessing both the overall performance and thermal status of this intricate industrial process. Through a comparison of the two most competitive machine learning algorithms, XGBoost and LightGBM, with their benchmark conventional techniques, lasso, random forest, support vector machine, and GBDT, this study's machine learning algorithms increase the accuracy of silicon content estimates. XGBoost and LightGBM are GBDT derivative algorithms that enhance prediction performance through the incorporation of regularization and effective computing techniques. The study's empirical findings demonstrated that XGBoost and LightGBM significantly outperformed the conventional algorithms. More significantly, this result also implied the efficacy of these algorithms in directing and determining the state of the blast furnace because the R-squared of the boosts was greater than 0.7 on the training set for two real-world blast furnace systems. The application of two novel gradient-boosting enhancements—XGBoost and LightGBM—to the prediction of hot metal silicon content, demonstrating their efficacy in an entirely new industrial setting, is what makes the current study unique. The findings offer further understanding of how to enhance blast furnace performance and suggest potential fixes for the shortcomings of the models that have been looked into thus far. Future study may involve investigating the applicability of these algorithms on other industrial prediction challenges and further validating these results using additional data.

Many areas advanced by the prospect of providing ML-based predictive prediction. Novel approaches that fused neural networks with gradient boosting or bridge generalization and Bayesian techniques enhanced model durability and accuracy with respect to uncertainty. These advancements address the shortcomings of traditional models and enhance expected performance. [8]Although the ensemble classifiers have been used to check for sparse data and outliers in the prediction of diabetes, new methods have recently validated their use in the identification and diagnosis of disorders like cardiovascular disease by stacking the classifiers with Electronic Health Records (EHR), which automates and improves the accuracy of diagnosis. The possibility of offering ML-based predictive prediction accelerated several fields. New methods that combined neural networks with bridge generalization, gradient boosting, or Bayesian procedures improved the robustness and uncertainty-awareness of the model. These improvements improve predicted performance and solve the drawbacks of conventional models. By stacking the classifiers with Electronic Health Records (EHR), which automates and increases diagnosis accuracy, new techniques have validated the use of ensemble classifiers in the identification and diagnosis of disorders like cardiovascular disease, even though they have been used to check for sparse data and outliers in the diabetes prediction. [11]Modern techniques for predicting the risk of heart disease have led to other approaches that have achieved excellent accuracy through model performance and different data processing procedures. These are welcome advancements; the rise of machine learning in several domains may be attributed to more precise, efficient, and practical solutions. These methods are accompanied by the capacity to deal with complex issues, to offer perceptive data as they become older, and to be essential in both research and real-world applications.

3 PROBLEM FORMULATION

The goal of this research is to develop a Gradient Boosting Classifier that can use a dataset including the most important water quality characteristics to identify water potability. The dataset comprises the following crucial elements for evaluating water safety: conductivity, organic carbon, trihalomethanes, pH, hardness, solids, chloramines, sulfate, and turbidity. In this regard, the study's goal was to create a reliable prediction model that could identify whether or not water samples were drinkable, facilitating the early detection and efficient handling of problems with water quality. StandardScaler was utilized to normalize the features in the research data using feature scaling. The mean value was used to impute missing data. Since a gradient boosting classifier can predict a result that may express intricate interactions between factors, it was selected for the task. Given its effectiveness in managing complicated interactions and forecasting high performance, a Gradient Boosting Classifier was selected. GridSearchCV was used to do hyperparameter tweaking, which maximized model performance. The number of estimators, learning rate, and tree depth were the pre-set optimum parameters. Metrics like the confusion matrix, accuracy score, and classification report were used to evaluate the performance of the model. The enhanced GridSearchCV demonstrated exceptional forecast precision for the potability of water and, as a result, might be useful in applications involving the monitoring of water quality. As a result, by offering a trustworthy method for assessing water potability, the study advances the field of environmental data analysis and advances the development of water safety protocols and decision-making.

4 METHODOLOGY

This method explains the application of Gradient Boosting in a machine learning model to anticipate water potability, based on a variety of quality variables. Important characteristics including pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity are included in the publicly accessible dataset. The approach will comprise data collection and preparation, data separation and scaling, EDA, and model creation. Data must first be gathered and preprocessed in order to deal with missing values. EDA then aids in the distribution and relationships within the data. The test and training sets are then divided, and feature scaling is used to keep everything consistent. Afterwards, the Gradient Boosting model is developed and its hyperparameters are optimally tuned using GridSearchCV. Assessment of performance: This guarantees that the final model predicts the potability of water

accurately and consistently. The model's robustness and efficiency are guaranteed by the method's completeness.

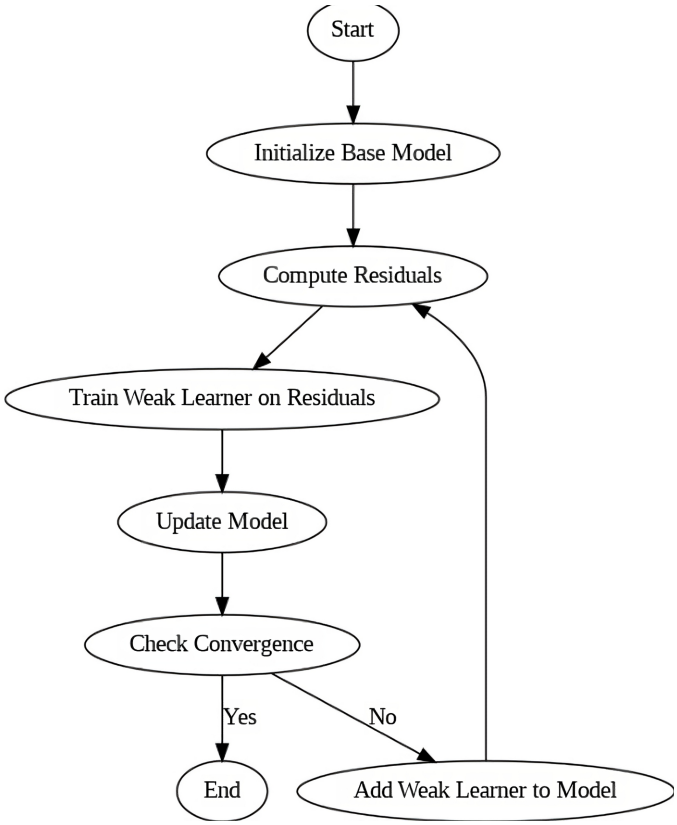

Fig. 1. Methodolgy of Gradient Boosting

The data utilized is freely accessible to the public and includes a target variable that indicates whether or not the water is drinkable along with a variety of water quality markers. Water safety is characterized by a number of important factors, including pH, hardness, and solids. In the pre-processing stage, missing values are imputed with the corresponding feature mean to ensure data integrity is preserved. Feature selection has been limited to the most pertinent variables that are constrained by the dataset in an effort to improve performance while reducing noise. This is a crucial step toward effective analysis and modeling as it also harmonizes the data so that every attribute is on the same scale. To understand each characteristic's tendency and dispersion, the exploratory data analysis first computes the summary statistics for each attribute. To find trends or anomalies of any sort, a number of metrics are computed, including mean, median, standard deviation, and range. Distribution Analysis: To detect skewness and locate outliers, this stage presents a density plot or histogram that shows the overall distribution of each attribute. Correlation Analysis: This method helps to find multicollinearity by examining the correlations between attributes using a correlation matrix. Additionally, we examine the potability target variable's distribution to look for any class imbalances that should be corrected when the model is being trained. Using an 80/20 ratio, we divided the dataset into training and test subsets, with the former being retained for performance evaluation and the latter being utilized to train the model. This divide helps ensure that the model performs well when applied to new, untested data. By subtracting the mean and scaling the features to unit variance, the 'StandardScaler' standardizes the data. In the event that scale disparities exist, the goal of this phase is to guarantee that no feature significantly affects the model's results.

This Gradient Boosting Classifier has a fixed random seed set in order to make it repeatable. This method was selected because to its propensity for managing complicated data and its capacity to create predictive models by iteratively adding decision trees and fixing errors from previous iterations. GridSearchCV will be used to tune the hyperparameters. It accomplishes this by experimenting with different setups of the optimal parameters, including max depth, learning rate, and n estimators. GridSearchCV will use cross-validation to ensure that the optimal hyperparameters are chosen and that they function on various subsets of data. Metrics including accuracy, precision, recall, F1-score, and a confusion matrix are used to measure the system's performance and reliability in water potability prediction after it has been trained on the optimal hyperparameters of the training data and tested on a test set.

4.1 gradient boosting mathematical framework

4.1.1Initialization Begin with an initial prediction $F_0(x)$, which is typically the average of the target values:

$$F_0(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

where γ represents the initial prediction and L is the loss function.

4.1.2Iterative Updates For each boosting round m from 1 to M:

—**Compute Residuals:** Determine the negative gradient of the loss function with respect to the current predictions $F_{m-1}(x)$:

$$r_{i,m} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \tag{2}$$

—**Train Weak Learner:** Fit a weak learner $h_m(x)$ to these residuals:

$$h_m(x) = \text{argmin}_h \sum_{i=1}^{n} [r_{i,m} - h(x_i)]^2 \tag{3}$$

—**Update Model:** Enhance the model with the new weak learner scaled by a learning rate η:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{4}$$

4.1.3Final Prediction The final model after M iterations is:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \eta \cdot h_m(x) \tag{5}$$

4.1.4Loss Function for Classification In binary classification, the log-loss function L is defined as:

$$L(y_i, p_i) = -\left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] \tag{6}$$

where $p_i = \sigma(F_m(x_i))$ and σ denotes the sigmoid function.

5 RESULTS AND DISCUSSIONS

The enhanced Gradient Boosting method achieved an overall accuracy of 66.16% on the dataset, showing moderate predictive strength. While this accuracy offers a reasonable level of predictive capacity, a detailed look at class-specific performance metrics provides valuable insights for future refinement.

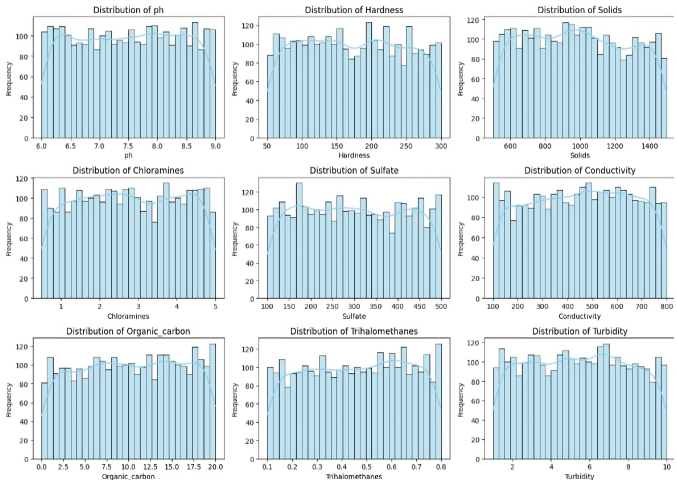

Fig. 2. Dataset Discription

5.1 class-specific performance analysis

The model's performance on class 0 was strong, with an F1-score of 0.78, a recall of 0.93, and a precision of 0.66. Given that class 0 represents the majority in the dataset, the model accurately classifies most instances from this class.
In contrast, class 1 performance was comparatively weaker, as reflected by an F1-score of 0.31, recall of 0.20, and a precision of 0.64. The low recall indicates difficulty in identifying class 1 instances, as shown in the confusion matrix with 195 false negatives and 27 false positives. This imbalance reveals a tendency for the model to underpredict class 1, resulting in a high number of missed cases, which negatively affects the recall score.

5.2 challenges posed by class imbalance

The macro-average F1-score of 0.54 reflects the disparity in model performance between classes. The lower performance on class 1, as compared to class 0, suggests that this imbalance may stem from the dataset's class distribution. Although the enhanced Gradient Boosting model improves recognition of minority classes compared to traditional methods, the data's class imbalance remains a substantial obstacle.

5.3 future improvements for imbalanced data

To address these limitations, techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or assigning class weights during training could be considered. Such methods could potentially improve the model's ability to recognize minority class instances, thereby reducing false negatives and improving recall for class 1.
In summary, while the enhanced Gradient Boosting method provides improvements, further adjustments to account for data imbalance could boost its robustness, leading to a more balanced performance across classes.
In addition to the stated total accuracy of 66.16%, further analysis may be performed to demonstrate how class imbalance presents a particularly challenging scenario for this model. Precision-recall curves and ROC-AUC scores show that the model can handle class 1 nearly perfectly, but it has a very high true positive rate for class 0. This may be inferred from class 1's poor accuracy and recall ratings, which indicate a high rate of misclassification and a high

Table 1. Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.66	0.93	0.78	412
1	0.64	0.20	0.31	244
Accuracy			**0.66**	**656**
Macro avg	0.65	0.57	0.54	656
Weighted avg	0.66	0.66	0.60	656

quantity of false negatives. Features that contribute most to class 0 may not be as effective in class 1, according to feature importance analysis, hence feature engineering may need to be expanded.

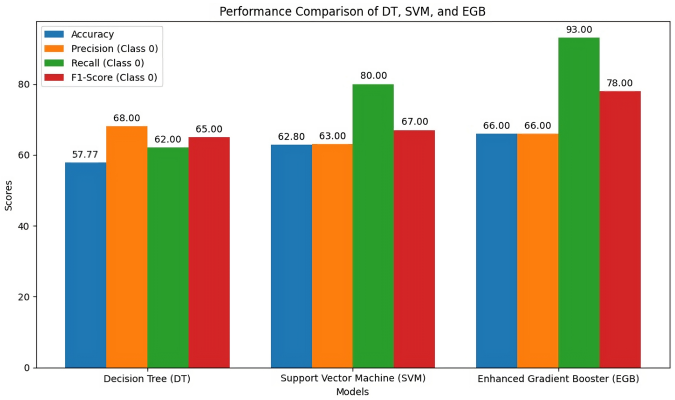

Fig. 3. Permformance

Variability in performance metrics was also reflected in cross-validation findings, highlighting the necessity of using reliable approaches during the validation process. Furthermore, it is imperative to include techniques like SMOTE for dataset balance and refine feature selection to make the model sensitive to minority classes in order to work on improving these performance difficulties. In order to ensure the model's overall efficacy and generalizability, future research should focus on these tactics.

6 CONCLUSION

This study compares several machine learning algorithms: it focuses on Gradient Boosting, using water potability, diabetes, and heart disease datasets. The investigation shows that while the typical Gradient Boosting model performs rather well on the datasets related to diabetes and heart disease, it underperforms on the water potability dataset. Particularly on complicated datasets like the water potability dataset, the Enhanced Gradient Boosting approach considerably increases prediction accuracy while successfully resolving the performance constraints associated with water potability. This study's rigorous methodology, which includes exploratory data analysis, model creation, and data preparation, offers a broad framework for testing and refining machine learning models. These findings emphasize once more that customizing algorithms to fit certain data features is the only way to get the best results. To maximize forecast accuracy and dependability, these tailored solutions are essential in fields like environmental monitoring and medical diagnostics. This study's key finding is that, although conventional

approaches could be effective in some situations, they require advancements like those provided by Enhanced Gradient Boosting in order to get better results on more challenging datasets.
As a result, future research will expand the range of datasets on which the Enhanced Gradient Boosting technique may be applied. The two primary areas where changes may be made are feature engineering, which would allow the model's performance and quality to be enhanced by finding and adding pertinent features, and hyperparameter optimization, which would allow for improved model parameter choices. An enhanced interpretability of the model under consideration facilitates a more profound examination of the decision-making process of algorithms, an essential aspect of fine-tuning and enhancing those models with greater knowledge. An in-depth examination of the algorithm's robustness and flexibility will be provided by extending this research to include more intricate datasets and a variety of machine learning models. It may lead to the development of fresh, adaptable, and efficient machine learning solutions for a variety of prediction problems, with enormous promise for the environmental sciences, health, and many other fields. This emphasizes the concept once more: by exploring these paths, future studies will be in a better position to further the development of reliable and adaptable predictive modeling systems.

7 REFERENCES

[1] Wissam H Alawee, Luttfi A Al-Haddad, Ali Basem, Dheyaa J Jasim, Hasan Sh Majdi, and Abbas J Sultan. Forecasting sustainable water production in convex tubular solar stills using gradient boosting analysis. *Desalination and Water Treatment*, 318:100344, 2024.

[2] George Bai and Rohitash Chandra. Gradient boosting bayesian neural networks via langevin mcmc. *Neurocomputing*, 558:126726, 2023.

[3] KP Rasheed Abdul Haq and VP Harigovindan. Water quality prediction for smart aquaculture using hybrid deep learning models. *Ieee Access*, 10:60078–60098, 2022.

[4] Liangjun Jiang, Zhenhua Xia, Ronghui Zhu, Haimei Gong, Jing Wang, Juan Li, and Lei Wang. Diabetes risk prediction model based on community follow-up data using machine learning. *Preventive Medicine Reports*, 35:102358, 2023.

[5] Madhuri Kawarkhe and Parminder Kaur. Prediction of diabetes using diverse ensemble learning classifiers. *Procedia Computer Science*, 235:403–413, 2024.

[6] Wen Li, Wei Wang, and Wenjun Huo. Regboost: a gradient boosted multivariate regression algorithm. *International Journal of Crowd Science*, 4(1):60–72, 2020.

[7] Shihua Luo and Tianxin Chen. Two derivative algorithms of gradient boosting decision tree for silicon content in blast furnace system prediction. *IEEE Access*, 8:196112–196122, 2020.

[8] G Manikandan, B Pragadeesh, V Manojkumar, AL Karthikeyan, R Manikandan, and Amir H Gandomi. Classification models combined with boruta feature selection for heart disease prediction. *Informatics in Medicine Unlocked*, 44:101442, 2024.

[9] Subasish Mohapatra, Sushree Maneesha, Prashanta Kumar Patra, and Subhadarshini Mohanty. Heart diseases prediction based on stacking classifiers model. *Procedia Computer Science*, 218:1621–1630, 2023.

[10] Temidayo Oluwatosin Omotehinwa, David Opeoluwa Oyewola, and Ervin Gubin Moung. Optimizing the light gradient-boosting machine algorithm for an efficient early detection of coronary heart disease. *Informatics and Health*, 1(2):70–81, 2024.

[11] FE Sghiouer, A Nahli, H Bouka, and M Chlaida. Analysis of the drought effects on the physicochemical and bacteriological quality of the inaouene river water (taza, morocco). *Scientific African*, page e02328, 2024.