Prediction of Myocardial Infarction Complications using Gradient Boosting

Gamal Saad Mohamed Khamis
Department of Computer Science, Science College, Northern Border University, Arar, Saudi Arabia
gamal.khamees@nbu.edu.sa (corresponding author)

Zakariya M. S. Mohammed
Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia
zakariya.mohammed@gmail.com

Sultan Munadi Alanazi
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
sultan.alanazi@nbu.edu.sa

Ashraf F. A. Mahmoud
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
ashraf.abubaker@nbu.edu.sa

Faroug A. Abdalla
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
faroug.abdalla@nbu.edu.sa

Sana Abdelaziz Bkheet
Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia
sana.bkheet@nbu.edu.sa

Received: 23 September 2024 | Revised: 18 October 2024 | Accepted: 26 October 2024

 DOI: https://doi.org/10.48084/etasr.9076

ABSTRACT

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, representing a significant public health challenge. Myocardial Infarction (MI), a severe manifestation of CVDs, contributes substantially to these fatalities. Machine learning holds great promise for predicting MI. This study explores the potential of Gradient Boosting (GB) techniques for this purpose, explicitly focusing on CatBoost, LightGBM, XGBoost, and XGBoost Random Forest. The study leverages GB's embedded feature selection, missing-value handling, and hyperparameter tuning capabilities. Performance was evaluated using multiple metrics: Area Under the Curve (AUC), classification accuracy, F1 score, precision, recall, and Matthews Correlation Coefficient (MCC). A probabilistic comparison matrix was used to assess the relative performance of the GB models. The results demonstrate the superiority of CatBoost, achieving a classification accuracy of 94.9%, an AUC of 0.992, a recall of 94.9%, and an MCC of 0.82. The probabilistic comparison further confirms CatBoost's superior performance. These findings contribute to MI prediction, highlighting the predictive potential of the CatBoost algorithm and ultimately aiding the fight against MI to achieve better patient outcomes.

Keywords-GB; myocardial infarction; prediction; machine learning

I. INTRODUCTION

Cardiovascular Diseases (CVDs) encompass various conditions that affect the heart and blood vessels. These diseases are the leading cause of death worldwide, accounting for an estimated 17.9 million deaths annually, which represents 31% of all global deaths [1]. Myocardial Infarction (MI), also known as heart attack, is a type of CVD that occurs when blood flow to the heart is blocked, causing damage to the heart muscle. It is a life-threatening emergency and a significant public health burden. The high prevalence and mortality rates

associated with CVDs, including MI, underscore the critical need for early detection and effective management strategies to mitigate their impact. MI contributes to a substantial proportion of these fatalities, with heart attacks and strokes accounting for more than 80% of CVD deaths [2]. Early detection and treatment of MI are critical to improving survival rates and patient outcomes. In this light, metabolomics has been recognized as a promising approach for diagnosing MI, allowing a comprehensive characterization of metabolites that reflect the underlying pathophysiological processes [3]. Machine Learning (ML) techniques, particularly ensemble classifiers [4], have shown considerable promise in analyzing intricate patterns in complex metabolomics data. Gradient Boosting (GB) is a robust ML algorithm that combines weak predictive models to form a strong predictor. It is well-suited for the high dimensionality and collinearity typical of metabolomics data [5]. sGB ML classifiers are powerful tools for analyzing complex metabolomics data. These classifiers offer insights into the metabolic perturbations during MI and hold promise for developing diagnostic and prognostic biomarkers [6]. Despite the promising potential of GB classifiers in MI prediction, a comprehensive comparison of their performance and the impact of different preprocessing methods and hyperparameter tuning has yet to be performed. This study aims to address this gap by systematically evaluating the effectiveness of various GB techniques in predicting MI.

Recent studies have employed GB ML classifiers to dig into metabolomics data, revealing biomarkers and metabolic profiles associated with acute myocardial ischemia. In [6], metabolomics data combined with ML algorithms, such as GB, achieved better performance in human samples. Similarly, in [7], GB classifiers were used to target metabolome profiling as a diagnostic approach for CVDs. In [8], desorption electrospray ionization mass spectrometry imaging was combined with GB tree algorithms to identify segments of the infarcted myocardium. In [9], extreme GB classifiers were optimized for metabolomics methods to predict coronary artery disease. In [10], ML models were used, specifically extreme GB (XGBoost), to predict mortality post-acute MI, demonstrating the predictive power of this approach. In [11], GB was applied to compute a score that reflects an individual's probability for a Type 1 MI diagnosis, emphasizing the method's diagnostic capabilities. In [12], a model incorporated a light GB machine and XGBoost to predict the risk of late cardiogenic shock in patients with ST-segment elevation myocardial infarction, highlighting the prognostic potential of these classifiers. These studies underscore the versatility of GB classifiers in providing valuable insights across a spectrum of clinical applications related to MI, from diagnosis to risk assessment and outcome prediction. Integrating such ML techniques with metabolomics data is a testament to the interdisciplinary approach required to address complex medical challenges.

GB classifiers have established their effectiveness in classification tasks, including prominent variants such as XGBoost, LightGBM, and CatBoost. Preprocessing techniques fine-tune their performance. Many studies have shown that preprocessing steps such as discretization, feature selection, and missing value imputation play pivotal roles in classifier efficiency and accuracy [13-15]. The study in [17] built on the foundation laid by [16] by comparing various ML algorithms to predict MI complications using the same dataset. The results showed that Random Forest (RF) performed particularly well, complementing this study's exploration of GB algorithms for MI prediction. The strong performance of tree-based ensemble methods highlights their potential as a powerful class of algorithms for MI prediction tasks.

Although ML has been applied to predict heart attacks, previous works have some fundamental limitations. Much of the focus has been on refining specific aspects, such as feature selection and technical enhancements to GB. This narrow focus may restrict how broadly their insights can be generalized. Significantly, not all of them compared multiple versions of the GB algorithm or explained how they adjusted the model hyperparameters for optimal performance. This study builds on these previous works by directly comparing different GB algorithms to predict myocardial complications using an MI dataset. Given this context, this study aims to compare the performance of GB techniques in predicting MI. It also evaluates how different embedded preprocessing methods and hyperparameter tuning influence the predictive success of each algorithm on MI data, to determine the most effective approach for this critical application. This study improves our understanding of MI prediction by highlighting the promising predictive capabilities of the CatBoost algorithm. Thus, it supports efforts to improve patient outcomes in the battle against this condition.

Table I compares studies using different CVD datasets. Comparisons highlight the strong and consistent performance of gradient-boosting algorithms across various CVD and MI prediction tasks. This highlights the robustness of GB compared to algorithms like SVM and linear regression. This study contributes to this topic by demonstrating the algorithms' effectiveness with the MI dataset and highlighting avenues for future research, building on the strengths of GB and the promise shown by other top-performing algorithms.

TABLE I. PERFORMANCE AND RESULTS EVALUATION

Study	Algorithms	Evaluation accuracy	Topic
[7]	GB Tree	80%	CVD diagnostics
	SVM	80%	
	RF	91%	
	Linear Regression	74%	
	Multi-Layer Perceptron	80%	
[8]	GB tree ensemble	97%	Recognition of MI
[9]	XGBoost	74%	Predict obstructive coronary artery disease
[16]	Multi-Layer Perceptron	90%	MI complications
	Naive Bayes (NB)	79%	
	Decision Tree (DT)	88%	
[4]	lazy.IBk	100%	CVD prediction
	Decision Table/Naive Bayes (DTNB)	86%	
	Multi-Objective Evolutionary (MOE) fuzzy classifier	82%	
[5]	Light GB Machine (LightGBM)	97%	Predicting metabolite–disease associations

II. DATASET AND METHODS

A. Dataset and Data Preprocessing

This study used a comprehensive dataset comprising 1700 entries, compiled at the Krasnoyarsk Interdistrict Clinical Hospital in Russia to examine and predict the potential outcomes of MI [18]. It encompasses 124 distinct variables, with 111 details on the patient's demographic background, prior medical conditions, complications observed upon hospital admission, ECG findings, and subsequent medical interventions. The other 12 variables are divided to document various complications that occurred at four distinct intervals: (a) upon hospital admission, (b) 24 hours post-admission, (c) 48 hours post-admission, and (d) 72 hours post-admission. The dataset features descriptions are available in [18]. GB is a powerful ensemble learning technique that builds models sequentially, typically used for regression and classification tasks. It combines multiple weak learners, usually decision trees, to create a robust predictive model. The preprocessing techniques embedded within GB help improve the model's performance and handle various data types more effectively. Some essential embedded preprocessing techniques follow.

1) Handling Missing Values - Imputation

GB algorithms such as XGBoost and LightGBM can handle missing values internally by finding the best imputation strategy during training [19, 20].

2) Feature Importance – Automatic Feature Selection

GB calculates feature importance scores, allowing users to understand which features are most influential in making predictions. This can help select the most relevant features and remove redundant ones [21].

3) Handling Categorical Variables:

- One-Hot Encoding: Some implementations, such as CatBoost, automatically handle categorical variables without requiring manual one-hot encoding [22].
- Target Encoding: CatBoost also supports target encoding, replacing categorical variables with the mean target value for each category [22].

4) Regularization

- Shrinkage: GB uses a technique called shrinkage (or learning rate), which involves scaling the contribution of each tree. This acts as a form of regularization to prevent overfitting [21].
- Subsampling: Using a subset of data to train each tree introduces randomness and reduces overfitting [23].

5) Tree Pruning

Max depth and Min samples: Parameters such as maximum tree depth, minimum samples per leaf, and minimum samples to split help control the complexity of each decision tree, thus preventing overfitting [19].

6) Gradient-Based Optimization

Loss function optimization: GB optimizes a specified loss function (e.g., mean squared error for regression, log-loss for classification) by sequentially adding trees that minimize the residuals (errors) from the previous trees [3].

7) Handling Imbalanced Data

Class weights can be adjusted for classification tasks on imbalanced datasets to give greater importance to minority classes [19].

8) Feature Engineering

GB can inherently capture interaction terms between features by building trees that split on multiple features [21]. These embedded preprocessing techniques make GB a robust and versatile method for many machine-learning tasks. Each implementation may have unique features and optimizations, but the core principles remain similar.

B. Gradient Boosting (GB) Algorithms

GB constitutes a robust ensemble methodology within ML, constructing models incrementally and facilitating the refinement of arbitrary differentiable loss functions. The core idea is to create a robust predictive model by progressively combining several simpler models, such as decision trees. Here is a straightforward breakdown of the process:

- Start with a basic model: Start with a simple model that estimates the target values initially. This initial guess is typically the average of the target values.

$$F(x) = \arg min \sum i = 1^N \, L(y_i, c \,) \quad (1)$$

where L is the loss function, y_i are the actual target values, and c is a constant.

- Calculate errors: Evaluate the initial model's performance by computing the differences (errors) between the target and predicted values:

$$rim \, = \, y_i - F_{m-1}(x_i) \quad (2)$$

where rim are the residuals (errors) at iteration m, y_i are the actual values, and $F_{m-1}(x_i)$ are the predictions from the model at the previous iteration.

- Build a new model to address errors: Create a new simple model $(h_m(x))^{\,n}$ to predict the identified errors (3). Fit this new model to the residuals and determine the optimal weight (p_m) for the latest model:

$$p_m = argmin \, p \sum_{i=1}^{N} \, L(y_i, F_{m-1}(x_i) + ph_m(x_i)) \quad (3)$$

- Update the original model: Refine the original model by incorporating the new model's predictions:

$$F_m(x) = F_{m-1}(x) + p_m h_m(x) \quad (4)$$

This step helps in correcting errors made by the initial model.

- Repeat the process (4): Continuously repeat the cycle of calculating errors, building new models to predict those errors, and updating the model. Each iteration improves the model's accuracy.

In essence, GB iteratively enhances model's performance. Each new model explicitly targets the previous models' errors, progressively strengthening the overall predictive capability.

1) LightGBM

LightGBM, short for Light Gradient Boosting Machine, is esteemed for its swift processing, scalability, and outstanding performance, all built on decision tree algorithms. It is utilized in various ML fields, including ranking and classification. LightGBM enhances the GB technique by incorporating efficient versions of two novel methods: Gradient-based One-Sided Sampling (GOSS) and Exclusive Feature Bundling (EFB) [20]. The following hyperparameters were tuned for LightGBM in this study: (i) Number of trees: Specify how many GB trees will be included. (ii) Learning rate: Step size shrinkage is used to prevent overfitting. (iii) Replicable training: Fix the random seed, which enables the replicability of the results. (iv) Limit depth of individual trees: Specify the maximum depth of the particular tree. (v) Do not split a subset smaller than a value: controls the minimum number of data points (samples) a leaf node must have after a split. (vi) Fraction of training instances: Specify the percentage of the training instances for fitting the individual tree.

2) XGBoost

Short for Extreme Gradient Boosting, XGBoost is an advanced GB model that excels in efficiency, adaptability, and portability. By offering parallel tree boosting, sometimes called GBDT or GBM, XGBoost stands out in its ability to tackle various data science challenges swiftly and precisely. The XGBoost algorithm has been instrumental in numerous Kaggle competition triumphs [19]. The following hyperparameters were tuned for XGBoost in this study: (i) Number of trees, (ii) Learning rate, (iii) Replicable training, (iv) Limit the depth of individual trees, (v) Fraction of training instances. (vi) Regularization (Lambda), (vii) Fraction of training instances, (viii) Fraction of features for each tree: Specify the percentage of features to use when constructing each tree, (ix) Fraction of features for each level: Specify the percentage of features for each level, and (x) Fraction of features for each split: Specify the percentage of features for each split.

3) XGBRF

XGBRF represents a sophisticated ML paradigm that synthesizes XGBoost with RF, thereby creating a robust ensemble model. This synthesis is designed to capitalize on the respective strengths of each algorithm: XGBoost's capacity for high predictive accuracy and RF's ability to reduce variance. The XGBRF approach builds on the unique capabilities of XGBoost and Random Forest. Although there is a scarcity of direct studies on XGBRF, the proposition suggests coalescing an autonomous RF with XGBoost in conducting classification or regression analyses. This could enhance the model's equilibrium between bias and variance, improving its predictive prowess [24]. The fusion of RF's resilience with the predictive precision of XGBoost can precipitate considerable progress in domains where advanced predictive modeling is imperative, encompassing finance, healthcare, and bioinformatics. Although XGBoost and XGBRFB share many common hyperparameters, XGBRF adds additional parameters specific to the RF approach. This integration allows XGBRF to leverage the benefits of both GB and RF, providing more flexibility and control over the model's behavior and performance.

4) CatBoost

CatBoost emerges as a distinguished ML algorithm, acclaimed for its adept handling of categorical variables and strong resilience against overfitting. Developed by Yandex research experts, CatBoost is an open-source GB framework that demonstrably enhances efficiency and accuracy in classification and regression [22]. Intrinsic to CatBoost is its ability to autonomously process categorical variables, obviating the need for the extensive data preprocessing that other algorithms typically demand. This characteristic significantly enhances CatBoost's utility in environments with large and complex categorical data [25]. Its wide applicability is further illustrated by its deployment across various sectors, manifesting both versatility and operational effectiveness. In particular, it has been adopted to predict reference evapotranspiration within humid zones, a vital process for administering water resources and tailoring irrigation programs [26]. Such utilization underscores CatBoost's relevance in environmental studies and its contribution to the sustainable governance of natural endowments. Within the financial sector, CatBoost has outperformed traditional ML techniques in bankruptcy prediction tasks, and its interpretable nature has been particularly lauded for providing insight into its predictive rationale [27].

Additionally, merging CatBoost with feature selection methods has yielded fruitful results, such as precise estimates of aboveground biomass, a testament to the algorithm's ability to efficiently manage voluminous datasets populated with many variables [28]. The following hyperparameters were tuned for CatBoost in this study: (i) Number of trees, (ii) Learning rate, (iii) Replicable training. (iv) Limit the depth of individual trees, (v) Fraction of training instances, and (vi) Regularization (Lambda).

C. Performance Metrics

The models were assessed using accuracy, precision, recall, F1-score, and ROC-AUC to comprehensively evaluate their performance in predicting MI. These metrics were also used in [29] for heart disease prediction. This study aimed to discern the most effective GB algorithm for predicting MI. The findings provide valuable insights for healthcare professionals in the early detection and treatment of heart disease.

III. EXPERIMENT AND RESULTS

A. Experiments

The experiments aimed to evaluate the effectiveness of different GB algorithms (Lite GB, XGB, XGBRF, and CatBoosting) in predicting MI. Each model was trained on the dataset with adjusted hyperparameters. GB embeds default preprocessing, executed in the following order:

- Removes instances with unknown target values
- Continues categorical variables (with one-hot-encoding)

- Removes empty columns
- Imputes missing values with mean values
- Automatic Feature Selection.

Figure 1 outlines the methodology framework, showing the initial model implementation with hyperparameter tuning and preprocessing and the comparative analysis to evaluate each model's performance.

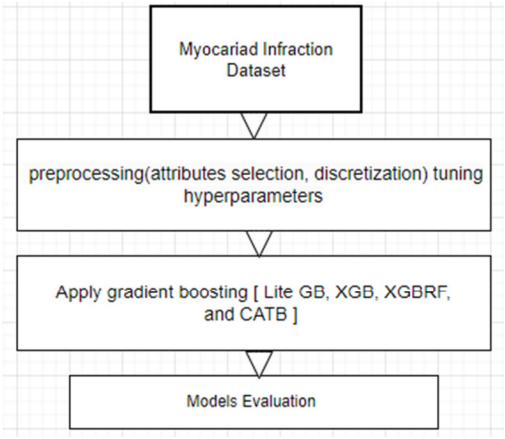

Fig. 1. The method framework.

B. Results

To identify the best GB model for predicting MI, the performance metrics of CatBoost, LightGBM, XGBoost, and XGBRF were calculated. Table II shows the hyperparameters used and Table III shows the performance evaluation results.

CATBoost emerged as the best performer in most evaluation metrics. Its impressive accuracy of 94.9%, precision of 93.8%, recall of 94.9%, F1 score of 93.6%, and AUC of 0.992 demonstrate its exceptional ability to distinguish between MI and non-MI samples based on metabolomics data. These results underscore the high effectiveness of CATBoost for this specific prediction task. LightGBM closely trailed CATBoost, showcasing robust performance. With an accuracy of 94.6%, precision of 94.1%, recall of 94.9%, F1 score of 94.1%, and AUC of 0.988, LightGBM proves to be a highly promising alternative. While its metrics are slightly lower than those of CATBoost, the minor differences suggest that LightGBM could be a viable option depending on specific implementation considerations. XGBoost's performance was comparable to LightGBM's, with an accuracy of 94.6%, precision of 93.6%, recall of 94.9%, F1 score of 94.1%, and an AUC of 0.99. This similarity indicates that both models are equally effective for MI prediction based on metabolomics data. XGBoost and LightGBM might hinge on factors such as computational efficiency or ease of hyperparameter tuning. The XGBRF model performed lower than the other GB variants. However, its accuracy of 94.2%, precision of 93.6%, recall of 94.2%, F1 score of 93.8%, and AUC of 0.985 are still respectable. The slightly reduced performance compared to the other GB models suggests that incorporating RF elements may not provide additional benefits for this problem. This highlights the importance of carefully evaluating different model architectures and selecting the most suitable approach for a specific task.

TABLE II. MODELS' HYPERPARAMETERS

Learning rate	LightGBM - CATB = 0.1 and XGB - XGBRF=0.3
Number Of Trees	100
Limit depth of individual tree	3
Fraction of training instances	1
Regularization (Lambda)	XGB-XGBRT-CATB =1

TABLE III. EVALUATION RESULTS FOR CATBOOST, LIGHTGBM, XGB, AND XGBRF

	ROC-AUC	**Accuracy**	**F1 score**	**Precision**	**Recall**	**MCC**
CatBoost	0.992	94.9%	93.6%	93.8%	94.9%	0.82
LightGBM	0.988	94.6%	94.1%	94.1%	94.9%	0.815
XGB	0.99	94.6%	94.1%	93.6%	94.9%	0.812
XGBRF	0.985	94.2%	93.8%	93.6%	94.2%	0.798

The probabilistic comparison matrix evaluates the relative performance of GB models, providing a detailed pairwise comparison by evaluating the likelihood that the row model outperforms the column model across multiple metrics. The values in the table estimate the probability that the model's performance in the row is superior to that of the model in the column. Specifically, the comparison is made using three key evaluation metrics, including classification accuracy, recall, ROC-AUC. A probability greater than 0.5 suggests that the model in the row performs better than the model in the column for the given metric. The higher the probability, the more substantial the evidence that the model in the row is superior. The results showed that CatBoost outperformed the other GB algorithms. probabilistic model comparison BY CLASSIFICATION ACCURACY

	CatBoost	**LightGB**	**XGB**	**XGBRF**
CatBoost	-	0.805	0.793	0.929
LightGB	0.195	-	0.558	0.866
XGB	0.207	0.442	-	0.858
XGBRF	0.071	0.134	0.142	-

TABLE IV. PROBABILISTIC MODEL COMPARISON BY RECALL

	CatBoost	**LightGB**	**XGB**	**XGBRF**
CatBoost	-	0.805	0.793	0.929
LightGB	0.195	-	0.558	0.866
XGB	0.207	0.442	-	0.858
XGBRF	0.071	0.134	0.142	-

TABLE V. PROBABILISTIC MODEL COMPARISON BY ROC-AUC

	CatBoost	**LightGB**	**XGB**	**XGBRF**
CatBoost	-	0.919	0.979	0.975
LightGB	0.081	-	0.587	0.941
XGB	0.021	0.413	-	0.939
XGBRF	0.025	0.059	0.061	-

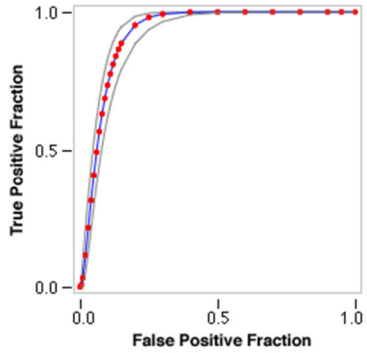

Fig. 2. ROC curve of CatBoost classification model.

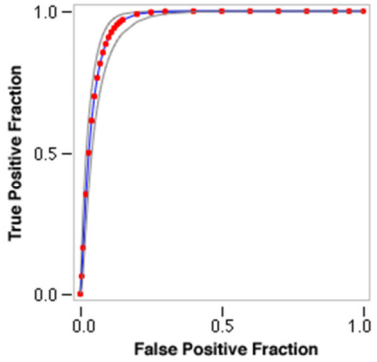

Fig. 3. ROC curve of LightGBM classification model.

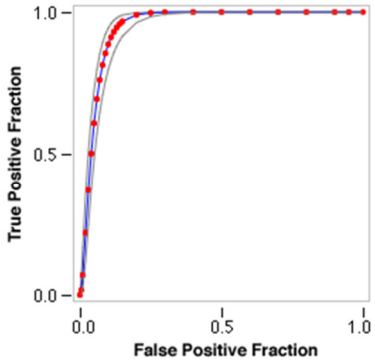

Fig. 4. ROC curve of XGBRF classification model.

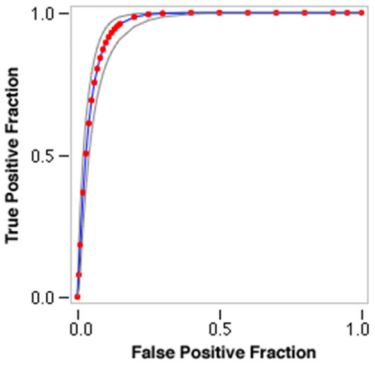

Fig. 5. ROC curve of XGB classification model.

IV. DISCUSSION

This study demonstrated the effectiveness of GB algorithms, namely CatBoost, LightGBM, XGB, and XGBRF, in predicting MI from metabolomics data. CatBoost emerged as the top-performing model, but all GB variants showed robust performance. These findings contribute to the growing body of literature applying ML to MI prediction from complex clinical datasets and underscore the potential of GB algorithms for this task. A key strength of GB methods is their embedded feature selection and handling of missing values, reducing the need for separate preprocessing steps. This allows for more streamlined model development compared to algorithms requiring extensive preprocessing. Comparison with previous works on the MI dataset highlights the value of diverse computational approaches for extracting insights. In [30], trajectory analysis was used to reveal disease heterogeneity and longitudinal progression patterns. Trajectory analysis in ML involves examining sequences of data points that represent movement or progression over time or space. These insights could enhance future ML models by incorporating features that capture disease progression trajectories.

In [17], RF performed well in predicting MI complications, aligning with the current study's findings on the effectiveness of tree-based ensemble methods (GB). The strong performance of tree-based ensemble methods across both studies highlights their promise in MI prediction tasks. In terms of evaluation metrics, this study found that CatBoost achieved the highest accuracy (0.949), closely followed by LightGBM (0.946) and XGBoost (0.946). In [16], the RF model achieved an accuracy of 0.963 in predicting MI complications. The high accuracy values in both studies underscore the potential of ML to achieve accurate MI predictions.

V. CONCLUSION

This study supports the continued exploration of advanced computational methods for analyzing complex clinical datasets such as the MI dataset. Future work can build on these findings by integrating insights from trajectory analysis, ML, and potentially other approaches to improve the accuracy of predictive models and patient outcomes. The MI dataset, which has been explored by only a few studies to date, remains a valuable resource for developing and refining predictive models for this critical disease. Future work could also explore ensembling multiple algorithms to achieve even better predictive performance. Although GB algorithms have built-in methods to help address imbalanced classes, further studies are essential.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2024-2267-01."

REFERENCES

[1] "Cardiovascular diseases," *World Health Organization*. https://www.who.int/health-topics/cardiovascular-diseases.

[2] "Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report," *World Heart Federation*. https://world-heart-

federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/.

[3] A. Surendran, M. Aliani, and A. Ravandi, "Metabolomic characterization of myocardial ischemia-reperfusion injury in ST-segment elevation myocardial infarction patients undergoing percutaneous coronary intervention," *Scientific Reports*, vol. 9, no. 1, Aug. 2019, Art. no. 11742, https://doi.org/10.1038/s41598-019-48227-9.

[4] S. M. Alanazi and G. S. M. Khamis, "Optimizing Machine Learning Classifiers for Enhanced Cardiovascular Disease Prediction," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12911–12917, Feb. 2024, https://doi.org/10.48084/etasr.6684.

[5] C. Zhang, X. Lei, and L. Liu, "Predicting Metabolite–Disease Associations Based on LightGBM Model," *Frontiers in Genetics*, vol. 12, Apr. 2021, https://doi.org/10.3389/fgene.2021.660275.

[6] J. Cao *et al.*, "Combined metabolomics and machine learning algorithms to explore metabolic biomarkers for diagnosis of acute myocardial ischemia," *International Journal of Legal Medicine*, vol. 137, no. 1, pp. 169–180, Jan. 2023, https://doi.org/10.1007/s00414-022-02816-y.

[7] N. E. Moskaleva *et al.*, "Target Metabolome Profiling-Based Machine Learning as a Diagnostic Approach for Cardiovascular Diseases in Adults," *Metabolites*, vol. 12, no. 12, Dec. 2022, Art. no. 1185, https://doi.org/10.3390/metabo12121185.

[8] K. Margulis, Z. Zhou, Q. Fang, R. E. Sievers, R. J. Lee, and R. N. Zare, "Combining Desorption Electrospray Ionization Mass Spectrometry Imaging and Machine Learning for Molecular Recognition of Myocardial Infarction," *Analytical Chemistry*, vol. 90, no. 20, pp. 12198–12206, Oct. 2018, https://doi.org/10.1021/acs.analchem.8b03410.

[9] E. Panteris *et al.*, "Machine Learning Algorithm to Predict Obstructive Coronary Artery Disease: Insights from the CorLipid Trial," *Metabolites*, vol. 12, no. 9, Sep. 2022, Art. no. 816, https://doi.org/10.3390/metabo12090816.

[10] R. Khera *et al.*, "Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction," *JAMA Cardiology*, vol. 6, no. 6, pp. 633–641, Jun. 2021, https://doi.org/10.1001/jamacardio.2021.0122.

[11] M. P. Than *et al.*, "Machine Learning to Predict the Likelihood of Acute Myocardial Infarction," *Circulation*, vol. 140, no. 11, pp. 899–909, Sep. 2019, https://doi.org/10.1161/CIRCULATIONAHA.119.041980.

[12] Z. Bai *et al.*, "Development of a machine learning model to predict the risk of late cardiogenic shock in patients with ST-segment elevation myocardial infarction," *Annals of Translational Medicine*, vol. 9, no. 14, Jul. 2021, Art. no. 1162, https://doi.org/10.21037/atm-21-2905.

[13] L. Devos, W. Meert, and J. Davis, "Fast GB Decision Trees with Bit-Level Data Structures," in *Machine Learning and Knowledge Discovery in Databases*, Würzburg, Germany, 2020, pp. 590–606, https://doi.org/10.1007/978-3-030-46150-8_35.

[14] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "GB Feature Selection With Machine Learning Classifiers for Intrusion Detection on Power Grids," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 1104–1116, Mar. 2021, https://doi.org/10.1109/TNSM.2020.3032618.

[15] G. Madhu, B. L. Bharadwaj, G. Nagachandrika, and K. S. Vardhan, "A Novel Algorithm for Missing Data Imputation on Machine Learning," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, Nov. 2019, pp. 173–177, https://doi.org/10.1109/ICSSIT46314.2019.8987895.

[16] S. E. Golovenkin *et al.*, "Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data," *GigaScience*, vol. 9, no. 11, Nov. 2020, Art. no. giaa128, https://doi.org/10.1093/gigascience/giaa128.

[17] A. Satty, M. M. Y. Salih, A. A. Hassaballa, E. A. E. Gumma, A. Abdallah, and G. S. M. Khamis, "Comparative Analysis of Machine Learning Algorithms for Investigating Myocardial Infarction Complications," *Engineering, Technology & Applied Science Research*, vol. 14, no. 1, pp. 12775–12779, Feb. 2024, https://doi.org/10.48084/etasr.6691.

[18] S. E. Golovenkin *et al.*, "Myocardial infarction complications," *UCI Machine Learning Repository*, 2020.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[20] G. Ke *et al.*, "LightGBM: A Highly Efficient GB Decision Tree," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[21] J. H. Friedman, "Greedy function approximation: A GB machine.," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, https://doi.org/10.1214/aos/1013203451.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[23] J. H. Friedman, "Stochastic GB," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, Feb. 2002, https://doi.org/10.1016/S0167-9473(01)00065-2.

[24] V. Kanaparthi, "Credit Risk Prediction using Ensemble Machine Learning Algorithms," in *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, Apr. 2023, pp. 41–47, https://doi.org/10.1109/ICICT57646.2023.10134486.

[25] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *Journal of Big Data*, vol. 7, no. 1, Nov. 2020, Art. no. 94, https://doi.org/10.1186/s40537-020-00369-8.

[26] G. Huang *et al.*, "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *Journal of Hydrology*, vol. 574, pp. 1029–1041, Jul. 2019, https://doi.org/10.1016/j.jhydrol.2019.04.085.

[27] S. B. Jabeur, C. Gharib, S. Mefteh-Wali, and W. B. Arfi, "CatBoost model and artificial intelligence techniques for corporate failure prediction," *Technological Forecasting and Social Change*, vol. 166, May 2021, Art. no. 120658, https://doi.org/10.1016/j.techfore.2021.120658.

[28] M. Luo *et al.*, "Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass," *Forests*, vol. 12, no. 2, Feb. 2021, Art. no. 216, https://doi.org/10.3390/f12020216.

[29] P. Anuradha and V. K. David, "Feature Selection and Prediction of Heart diseases using GB Algorithms," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, Mar. 2021, pp. 711–717, https://doi.org/10.1109/ICAIS50930.2021.9395819.

[30] Q. X. Song *et al.*, "The machine learning model based on trajectory analysis of ribonucleic acid test results predicts the necessity of quarantine in recurrently positive patients with SARS-CoV-2 infection," *Frontiers in Public Health*, vol. 10, Nov. 2022, https://doi.org/10.3389/fpubh.2022.1011277.