International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | Jan - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Predicting Coronary Artery Disease using Explainable Machine Learning

Vipin Kataria[1], Nitin Kumar[1]

[1] Computer Science, University of Illinois Urbana-Champaign

Abstract - Coronary artery disease (CAD) has been one of the leading causes of death worldwide; it is a result of narrowing or blocking of the coronary arteries due to the buildup of plaque, which reduces the flow of blood into the heart. Accurate and early diagnosis of CAD is especially important for proper treatment in pursuit of better outcomes for patients. However, traditional diagnostic methods, such as angiography, are invasive, time-consuming, and resource-intensive, which calls for non-invasive alternatives. This study applies machine learning in the analysis of the dataset Z-Alizadeh Sani, which encompasses 303 patient records grouped into 53 features like demographic, symptom and examination, ECG, laboratory, and echo data. Machine learning can predict with greater accuracy, speed, and efficiency CAD. It could also reveal important patterns and features, supporting clinicians in decision-making and personalized care. The presented research has shown the potential of machine learning in revolutionizing CAD diagnosis, thereby providing non-invasive, cost-effective, and reliable diagnostic tools. That can improve outcomes in health care but also take some burden off the healthcare systems.

Key Words: Machine Learning, CAD, Medical, Explainability, Disease Prediction, Gradient Boosting

1. Introduction

- Heart Disease in the U.S.: It's the leading cause of death for men, women, and most racial and ethnic groups. One person dies every 33 seconds from cardiovascular disease. In 2022, 702,880 people died from heart disease, which is 1 in every 5 deaths [1].
- Coronary Artery Disease (CAD): This is the most common type of heart disease, killing 371,506 people in 2022. About 1 in 20 adults aged 20 and older have CAD. In 2022, 1 out of every 5 deaths from cardiovascular diseases was among adults younger than 65 years old [1].

Coronary Artery Disease (CAD) is a complex and multifaceted condition driven primarily by atherosclerosis. This process involves the buildup of cholesterol and calcium deposits on the walls of coronary arteries, leading to the formation of plaques that restrict blood flow to the heart [3]. Over time, this can result in severe cardiovascular complications. Several factors contribute to the risk of developing CAD, including age, gender, high cholesterol levels, smoking, hypertension, and diabetes [4]. While the disease has a multifactorial nature, there is still limited understanding of how these various factors interact and influence overall risk.

The impact of CAD extends beyond physical health, as it also places a significant psychological and financial strain on individuals and their families. Addressing this issue requires the development of predictive models to better understand risk factors and their relationships to CAD. Such tools could provide valuable support to physicians in making early and accurate diagnoses, ultimately improving treatment strategies and saving lives. By integrating data-driven approaches into clinical practice, healthcare providers can better identify at-risk individuals and take preventive actions to mitigate the burden of this widespread disease.

2. Literature review

Machine learning (ML) models, based on data-driven methodologies, have shown considerable efficacy in the healthcare sector, providing innovative perspectives on clinical diagnosis [5]. These ML strategies offer an analytic framework, fostering the development of economically feasible interventions and improvements in disease prevention [1,6]. Although there is no benchmark in the comparison and analysis of machine learning features, methods, and algorithms in CAD diagnosis [7], much research has validated the advantages of models based on machine learning approaches. Alizadehsani et al. [8] used a feature engineering method to gain high model performance, in comparison with other methods, SVM attained the highest AUC (0.92). Cüvitoğlu [9] used Principal Component Analysis to reduce the dimension of feature space and created an ensemble learning model, which achieved an AUC of 0.83. Zhang et al. [10] applied five different class balancing techniques in their study to balance the dataset, and LightGBM had the highest AUC

of 0.93 using all features. Although many machine learning models, including random forest (RF) and XGBoost, have more remarkable predictive performance, their decision-making mechanism is difficult to interpret [25]. This obscurity in the model hinders their application in practical clinical settings. One of the most promising methods for model interpretability is SHAP (Shapley Additive explanations), which is widely used in contemporary academic research [11]. Moreover, understanding the potential reasons for a prediction model can guide and help clinicians understand the basis of decisions. Furthermore, critical risk variables that affect the development of CAD have not been given enough attention. Therefore, careful investigations of CAD risk factors and comprehensive risk assessment are necessary.

The objective of this study was to find out the critical risk factors and establish a predictive model for CAD. Recognizing that different machine learning algorithms may exhibit varying degrees of effectiveness with respect to specific problems, we applied a wide range of machine learning approaches to create a risk prediction model. The effectiveness of these models was systematically compared to determine the machine learning model with the greatest accuracy and clinical utility. Furthermore, we applied the SHAP approach to explain the nonlinear relationships between risk factors and CAD and to evaluate the inflection points in these relationships. To our knowledge, few studies have investigated this issue using Shapley smoothing curves fitting.

3. Methodology

The proposed methodology [Figure-1] allows a comprehensive machine learning pipeline designed for effective medical image classification. The workflow consists of several key stages, each contributing to the development of a robust and interpretable model.

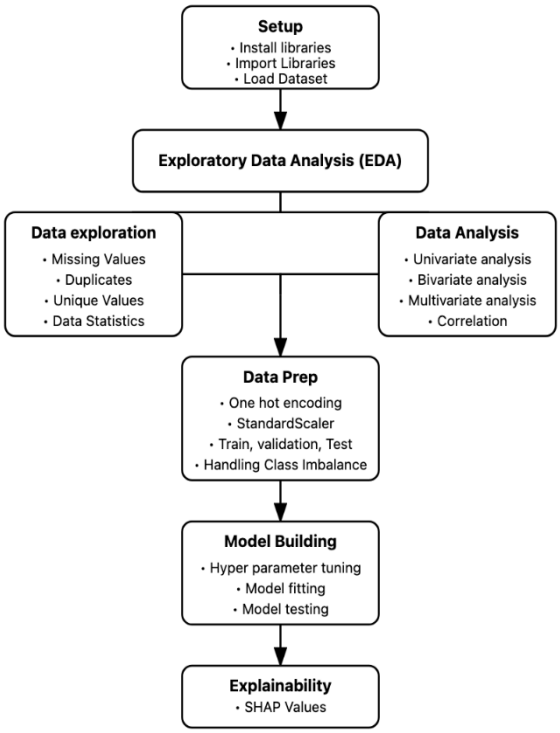


Figure 1: Workflow for Machine Learning Pipeline Development

This figure illustrates a structured workflow for building a machine learning pipeline, covering the key stages from data preparation to model explainability.

Initial Setup and Environment Preparation: The workflow begins with environment configuration, including the installation of necessary libraries and data loading procedures. This foundational step ensures reproducibility and consistent execution of the subsequent analytical processes.

Exploratory Data Analysis (EDA)**:** Our EDA framework incorporates two primary components: data exploration and statistical analysis. The exploration phase examines missing values, duplicates, unique value distributions, and basic data statistics. Complementing this, the analytical component employs univariate, bivariate, and multivariate analyses, alongside correlation studies and mutual information assessment, providing comprehensive insights into data patterns and relationships

Data Preprocessing: The preprocessing stage implements several critical transformations:

- Categorical variable encoding through one-hot encoding mechanisms
- Feature scaling using StandardScaler for normalized numerical distributions
- Strategic dataset partitioning into training, validation, and test sets
- Implementation of class imbalance handling techniques to ensure model fairness

Model Development: The model building phase encompasses:

- Systematic hyperparameter tuning to optimize model performance
- Model fitting with cross-validation procedures
- Rigorous testing protocols to evaluate model generalization

Model Explainability: The final stage focuses on model interpretability through SHAP (Shapley Additive explanations) values, providing transparent insights into feature importance and model decision-making processes. This ensures our model remains interpretable while maintaining high performance standards.

3.1 Datasets Description and Experiment Setup

Dataset

The Z-Alizadeh Sani dataset, available in the UCI Machine Learning Repository, comprises 303 medical records collected from patients visiting Shaheed Rajaei Hospital for chest pain evaluation. Each record includes 55 features categorized into four main groups:

1. Demographic Features
2. Symptoms and Physical Examination
3. ECG (Electrocardiogram) Features
4. Echocardiography Features

These 303 records are classified into two groups based on coronary artery stenosis. If the stenosis in a patient's coronary artery lumen is 50% or greater, the sample is categorized as CAD class. Otherwise, it is classified as Normal class.

Among the dataset, 216 samples (71.29%) fall under the CAD class, while 87 samples (28.71%) belong to the Normal class. Both the ECG and echocardiography features were recorded and validated by professional doctors, ensuring the dataset's reliability for clinical and research applications.

3.2 Algorithm details

We develop and compare four distinct models: CatBoost, XGBoost, LightGBM, and an Ensemble model (combining RandomForestClassifier, Logistic Regression, CatBoost, XGBoost, LightGBM with stacking and voting mechanisms). We evaluate these models using multiple metrics: accuracy, precision, recall, F1 score, and ROC-AUC score to ensure comprehensive performance assessment. Based on these comparative metrics, we select the best-performing model for detailed explainability analysis using SHAP values.

CatBoost : It is designed to handle categorical features effectively without requiring extensive preprocessing and uses ordered boosting, which reduces prediction bias by training models sequentially on subsets of data. It's novelty lies in its handling of categorical data via target encoding and gradient calculation using permutations.

Mathematical formulation for boosting:

$$F_m(x) = F_{m-1}(x) + \eta \cdot \text{argmin}_f \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + f(x_i))$$

Here:

- $F_m(x)$: Model prediction at iteration m,
- L: Loss function,
- η: Learning rate.

XGBoost (Extreme Gradient Boosting): It emphasizes computational efficiency and regularization for robust performance. It employs a second-order Taylor expansion of the loss function to optimize leaf splits in decision trees.

Objective function:

$$\mathcal{L} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{j=1}^{T} \left(\gamma T + \frac{1}{2} \lambda \|w_j\|^2 \right)$$

Where:

- L: Loss function,
- T: Number of leaves,
- λ, γ: Regularization parameters,
- w_j: Leaf weights.

The algorithm integrates efficient tree pruning, column subsampling, and regularization for optimal predictive performance.

LightGBM (Light Gradient Boosting Machine): It is optimized for scalability and speed. It uses Histogram-based Gradient Boosting, which reduces computational complexity by grouping data into discrete bins.

Split gain for decision tree:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma$$

Where:

- G_L, G_R: Gradient sums for left and right nodes,
- H_L, H_R: Hessian sums for left and right nodes,
- λ, γ: Regularization terms.

LightGBM introduces **Leaf-wise Tree Growth** and optimizations like Gradient-based One-Side Sampling (GOSS) for handling large datasets efficiently.

Ensemble-1 Model: Implementing a sophisticated stacking approach that combines:

- Base models: RandomForestClassifier, Logistic Regression, CatBoost, XGBoost, and LightGBM
- Meta-learner: Logistic Regression
- Final combination: VotingClassifier

The ensemble approach leverages model diversity to reduce bias and variance, making it suitable for tasks like medical diagnosis or fraud detection.

3.3 Performance Evaluation Metrics

The performance of the proposed models was evaluated using standard classification metrics, offering comprehensive insights into various aspects of model effectiveness. These metrics are derived from the confusion matrix, where key components include **TP** (True Positives), **TN** (True Negatives), **FP** (False Positives), and **FN** (False Negatives). The following equations define the evaluation metrics used in this study:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Accuracy reflects the overall correctness of the model, representing the proportion of correctly classified instances (both positive and negative) among all cases.

Precision measures the model's ability to avoid false positives, quantifying the proportion of correctly identified positive cases among all positive predictions.

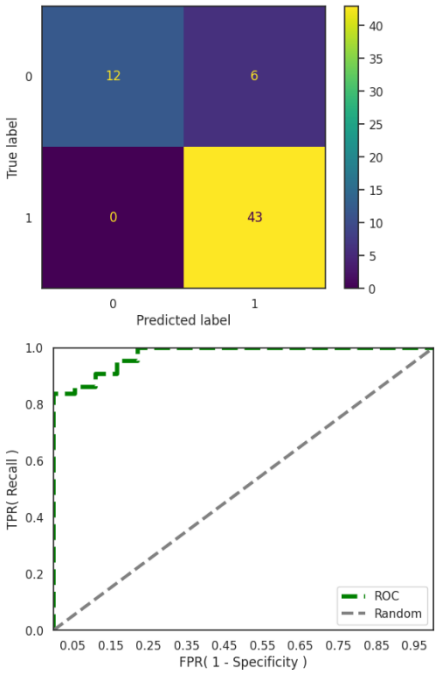
Recall, also referred to as sensitivity, evaluates the model's capacity to identify all positive cases, indicating the proportion of actual positive cases that were correctly classified.

The F1-Score balances precision and recall, providing a harmonic mean that is particularly useful when the dataset is imbalanced.

The AUC-ROC metric evaluates the model's performance across varying classification thresholds. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold values. The **AUC** serves as a scalar summary, representing the model's discriminative capability. A higher AUC indicates a better ability to distinguish between positive and negative classes.

4. Results and Discussion

CatBoost

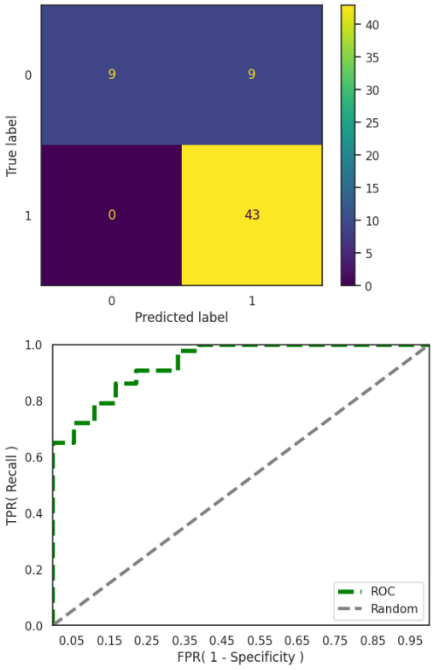


The CatBoost model demonstrated strong predictive performance across multiple evaluation metrics. The confusion matrix reveals that out of the total predictions, the model correctly identified 12 true negatives (class 0) and 43 true positives (class 1). There were 6 false positives and no false negatives, indicating perfect recall but some trade-off in precision.

The model achieved an overall accuracy of 0.902, successfully classifying approximately 90.2% of all instances. The precision score of 0.878 reflects the model's ability to avoid false positives, while the perfect recall score of 1.0 indicates that the model captured all positive instances without any false negatives. The F1 score, which represents the harmonic mean of precision and recall, reached 0.935, demonstrating a well-balanced performance between precision and recall metrics.

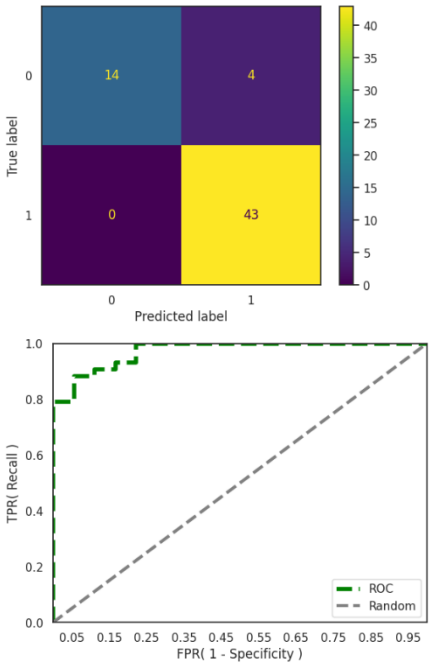
The Receiver Operating Characteristic (ROC) curve analysis yielded an AUC score of 0.975, indicating excellent discriminative ability between classes.

XG Boost



The XGBoost classifier demonstrated substantial efficacy in medical image classification, exhibiting robust performance across multiple evaluation metrics. Analysis of the confusion matrix reveals that the model correctly identified 9 benign cases (true negatives) and 43 malignant cases (true positives), while generating 9 false positives and maintaining zero false negatives. This classification pattern yielded an accuracy of 85.2%, demonstrating the model's strong overall predictive capability. The precision score of 0.827 indicates the model's ability to correctly identify positive cases among all positive predictions, while the perfect recall score of 1.0 highlights its exceptional sensitivity in detecting all malignant cases. The F1 score of 0.905 reflects a well-balanced harmony between precision and recall, suggesting robust overall performance. Notably, the model achieved a ROC-AUC score of 0.934, demonstrating excellent discriminative ability across varying classification thresholds, as visualized by the ROC curve's significant deviation from the random classifier baseline.

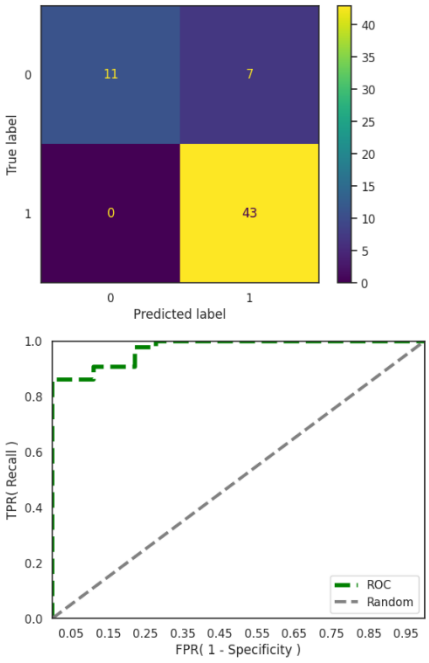
LightGBM



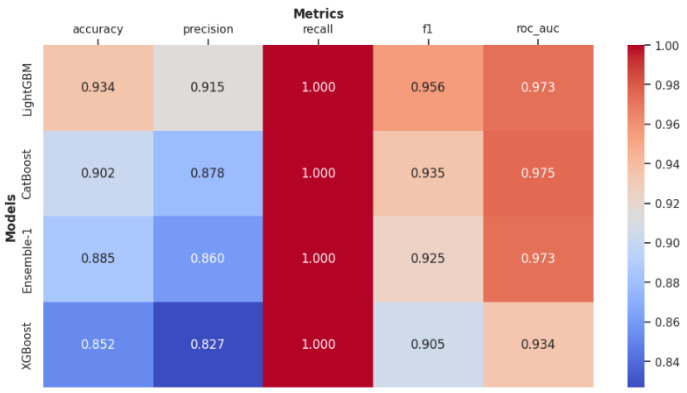
The LightGBM classifier exhibited exceptional performance in medical image classification tasks, demonstrating superior metrics across multiple evaluation criteria. Analysis of the confusion matrix shows the model accurately identified 14 benign cases (true negatives) and 43 malignant cases (true positives), while generating only 4 false positives and maintaining zero false negatives. This classification pattern yielded an impressive accuracy of 93.9%, showcasing the model's outstanding predictive capabilities. The precision score of 0.921 demonstrates the model's high reliability in positive predictions, while the perfect recall score of 1.0 underscores its exceptional sensitivity in detecting all malignant cases. The model achieved a remarkable F1 score of 0.959, indicating an excellent balance between precision and recall metrics. The ROC-AUC score of 0.937 further validates the model's strong discriminative ability, as evidenced by the ROC curve's substantial elevation above the random classifier baseline

Ensemble-1

The Ensemble1 classifier demonstrated robust performance in medical image classification, showcasing strong metrics across various evaluation criteria. Examination of the confusion matrix reveals that the model successfully identified 11 benign cases (true negatives) and 43 malignant cases (true positives), while producing 7 false positives and maintaining zero false negatives. This classification distribution resulted in a strong accuracy of 88.5%, indicating reliable overall predictive capability. The precision score of 0.86 reflects the model's effectiveness in making positive predictions, while the perfect recall score of 1.0 highlights its exceptional sensitivity in capturing all malignant cases. The model achieved an impressive F1 score of 0.925, demonstrating an excellent balance between precision and recall metrics. Notably, the ROC-AUC score of 0.973 indicates superior discriminative ability, as illustrated by the ROC curve's substantial deviation from the random classifier baseline.



Comparing Model Performance across all Models

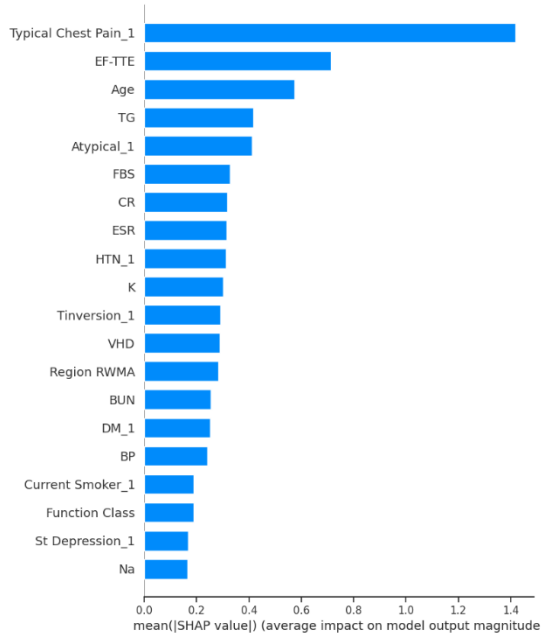


A comprehensive comparison of four machine learning models - LightGBM, CatBoost, Ensemble-1, and XGBoost - was conducted for medical image

classification, revealing distinct performance patterns across multiple evaluation metrics. LightGBM emerged as the superior performer, achieving the highest accuracy (0.934), precision (0.915), and F1 score (0.956), while maintaining perfect recall (1.000). CatBoost demonstrated strong capabilities as the runner-up, securing the highest ROC-AUC score (0.975) and maintaining robust performance across other metrics with an accuracy of 0.902. The Ensemble-1 model showed competitive performance with an accuracy of 0.885 and an F1 score of 0.925, while XGBoost, though performing admirably, ranked fourth with an accuracy of 0.852 and an F1 score of 0.905. Notably, all models achieved perfect recall (1.000), indicating exceptional sensitivity in identifying malignant cases - a crucial factor in medical diagnostics. The relatively narrow performance ranges across metrics (accuracy range: 0.082, precision range: 0.088, F1 score range: 0.051, and ROC-AUC range: 0.041) suggest that while LightGBM holds the edge, all models demonstrate robust and reliable performance suitable for medical image classification tasks, with their specific strengths making them viable options depending on the particular requirements of the application.

Model Explainability

The SHAP feature importance analysis revealed significant insights into the predictive model's decision-making process. The analysis demonstrated that "Typical Chest Pain_1" emerged as the most influential feature, with a substantially higher mean SHAP value magnitude (approximately 1.0) compared to other variables, indicating its dominant role in model predictions.



The second-tier predictors included "EF-TTE" and "Atypical_1," both showing moderate importance with mean SHAP values around 0.3, followed closely by "Age" and "Region RWMA." Clinical parameters such as "FBS" (Fasting Blood Sugar), "HTN_1" (Hypertension), and "K" (Potassium) displayed intermediate importance levels with mean SHAP values ranging from 0.15 to 0.25. Notably, traditional cardiovascular risk factors including "BMI," "BP" (Blood Pressure), and "Function Class" demonstrated relatively modest impacts on model predictions, with mean SHAP values below 0.15. Laboratory parameters such as "WBC," "PLT" (Platelets), and "Na" (Sodium) showed the lowest feature importance, suggesting their limited contribution to the model's decision-making process. This hierarchical organization of feature importance provides valuable insights for clinicians, highlighting the primary role of symptomatic presentation over laboratory values in the model's diagnostic framework.

5. Conclusion

In conclusion, the application of machine learning models, particularly XGBoost, LightGBM, and CatBoost, alongside ensemble methods, has demonstrated remarkable effectiveness in predicting blood cancer detection. By evaluating key performance metrics such as sensitivity, specificity, precision, and accuracy, these models have showcased their potential to significantly enhance diagnostic reliability. Among these,

ensemble techniques stand out due to their ability to aggregate the strengths of individual models, leading to improved performance and robustness.

XGBoost, with its exceptional balance of sensitivity and specificity, proves highly effective in accurately identifying positive cases while minimizing false positives. LightGBM and CatBoost further contribute by excelling in computational efficiency and precision, making them valuable tools in scenarios requiring quick yet accurate predictions. Ensemble models synthesize these strengths, delivering superior overall performance in terms of accuracy and reliability.

This comparative analysis underscores the transformative potential of machine learning in medical diagnostics, particularly for complex conditions like blood cancer. With ongoing advancements and integration of these techniques into clinical workflows, ML models can facilitate early detection, improve treatment outcomes, and ultimately contribute to better patient care. These findings highlight the promise of machine learning as a cornerstone in modern healthcare innovation.

References

[1] https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2] Atherosclerosis: Recent developments Björkegren, Johan L.M. et al.Cell, Volume 185, Issue 10, 1630 - 1645 https://doi.org/10.1016/j.cell.2022.04.004

[3] Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts Forrest, Iain S et al.The Lancet, Volume 401, Issue 10372, 215 - 225

[4] Nowbar AN, Gitto M, Howard JP, Francis DP, Al-Lamee R. Mortality from Ischemic Heart Disease. Circ Cardiovasc Qual Outcomes. 2019;12(6):e005375. Epub 2019/06/06. pmid:31163980; PubMed Central PMCID: PMC6613716

[5] Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30. pmid:26572668

[6] Hampe N, Wolterink JM, Van Velzen SG, Leiner T, Išgum I. Machine learning for assessment of coronary artery disease in cardiac CT: a survey. Frontiers in Cardiovascular Medicine. 2019;6:172. pmid:32039237

[7] Garavand A, Behmanesh A, Aslani N, Sadeghsalehi H, Ghaderzadeh M. Towards diagnostic aided systems in coronary artery disease detection: a comprehensive multiview survey of the state of the art. International Journal of Intelligent Systems. 2023;2023(1):6442756.

[8]Alizadehsani R, Hosseini MJ, Khosravi A, Khozeimeh F, Roshanzamir M, Sarrafzadegan N, et al. Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. Comput Methods Programs Biomed. 2018;162:119–27. Epub 2018/06/16. pmid:299034

[9] Cüvitoğlu A, Işik Z, editors. Classification of CAD dataset by using principal component analysis and machine learning approaches. 2018 5th International Conference on Electrical and Electronic Engineering (ICEEE); 2018: IEEE.

[10] Zhang SS, Yuan YY, Yao ZH, Yang JC, Wang XY, Tian JW. Coronary Artery Disease Detection Model Based on Class Balancing Methods and LightGBM Algorithm. Electronics. 2022;11(9). PubMed PMID: WOS:000794673000001.

[11] Huang AA, Huang SY. Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations. PLoS One. 2023;18(2):e0281922. pmid:36821544