Predicting Cardiac Complications of Myocardial Infarction Patients Using Machine Learning

Shriyansh Baidya
Lovejoy High School
Lucas, TX, USA
shribaidya9@gmail.com

Vibhuti Gupta
School of Applied Computational Sciences
Department of Computer Science & Data Science
Meharry Medical College
Nashville, TN, USA
vgupta@mmc.edu

***Abstract*—In the United States, heart disease is the leading cause of death, killing about 695,000 people each year. Myocardial infarction (MI) is a cardiac complication which occurs when blood flow to a portion of the heart decreases or halts, leading to damage in the heart muscle. Heart failure and Atrial fibrillation (AF) are closely associated with MI. Heart failure is a common complication of MI and a risk factor for AF. Machine learning (ML) and deep learning techniques have shown potential in predicting cardiovascular conditions. However, developing a simplified predictive model, along with a thorough feature analysis, is challenging due to various factors, including lifestyle, age, family history, medical conditions, and clinical variables for cardiac complications prediction. This paper aims to develop simplified models with comprehensive feature analysis and data preprocessing for predicting cardiac complications, such as heart failure and atrial fibrillation linked with MI, using a publicly available dataset of myocardial infarction patients. This will help the students and health care professionals understand various factors responsible for cardiac complications through a simplified workflow. By prioritizing interpretability, this paper illustrates how simpler models, like decision trees and logistic regression, can provide transparent decision-making processes while still maintaining a balance with accuracy. Additionally, this paper examines how age-specific factors affect heart failure and atrial fibrillation conditions. Overall this research focuses on making machine learning accessible and interpretable. Its goal is to equip students and non-experts with practical tools to understand how ML can be applied in healthcare, particularly for the cardiac complications prediction for patients having MI.**

***Index Terms*—machine learning, heart, myocardial infarction, health**

I. Introduction

Cardiovascular diseases, including myocardial infarction (MI), are among the leading causes of mortality worldwide. MI occurs when the blood flow to a part of the heart is blocked for an extended period, leading to damage or death of heart tissue. Early prediction of MI is crucial for preventing fatal outcomes, improving treatment, and reducing healthcare costs. Heart failure and Atrial fibrillation (AF) are closely associated with MI. Heart failure is a common complication of MI and a risk factor for AF.

Machine learning (ML) has revolutionized healthcare, particularly in the prediction and diagnosis of diseases. Various ML algorithms are well-suited for predicting outcomes based on large and complex datasets. By developing predictive models using machine learning, healthcare providers can make data-driven decisions to assess the risk of heart failure and AF and take preventive measures. This research explores the use of machine learning techniques to build a predictive model for predicting cardiac complications, such as heart failure and atrial fibrillation, using a publicly available dataset of myocardial infarction patients.

Cardiac complication prediction is challenging due to the multifactorial nature of the disease with various factors involved, including lifestyle, age, family history, medical conditions, and clinical variables. Machine learning (ML) and deep learning techniques have shown potential in predicting cardiac conditions. Several studies [1]–[5] have explored the use of ML models to predict cardiovascular conditions and associated outcomes. Although these studies show promising results however none of these studies have developed a simplified and interpretable framework to understand various factors associated with cardiac complications.

This paper aims to develop simplified models with comprehensive feature analysis and data preprocessing for predicting cardiac complications, such as heart failure and atrial fibrillation, using a publicly available dataset of myocardial infarction patients. Our major contributions in this paper are as follows:

- We develop simplified and interpretable machine learning models that predict heart failure and atrial fibrillation complications closely linked with myocardial infarction, utilizing clinical, physiological, and laboratory data.
- We performed a comprehensive feature analysis and data preprocessing to understand various factors involved in predicting cardiac complications.
- We extensively experiment on the given dataset and evaluate the performance using various metrics in predicting cardiac complications.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the dataset followed by the data description and proposed framework. Section 4 describes the experiments and results and Section 5 concludes the paper.

II. Related Work

Recent research on cardiac complications prediction associated with MI includes machine learning and deep learning methods. Jadhav et al. [1] demonstrated that decision trees

and random forests could be highly effective in predicting cardiovascular diseases by identifying key risk factors such as hypertension and cholesterol levels. Other research, such as the work by Agham et al. [2], explored the use of neural networks and deep learning to capture complex relationships in the data, which traditional models might overlook. Khera et al. [3], explored ML models to predict in-hospital mortality after acute myocardial infarction (AMI), comparing them with conventional logistic regression. Their findings suggested that although models like XGBoost and meta-classifiers offered improved risk resolution, the overall improvement in discrimination was modest.

Similarly, Than et al. [4] developed the Myocardial Ischemic Injury Index (MI3), which combined patient age, sex, and serial troponin levels using a gradient boosting algorithm to predict the likelihood of MI. Their model outperformed conventional approaches, demonstrating excellent calibration and area under the receiver operating characteristic curve (AUC). Chen et al. [5] applied ML models to classify MI presence and severity using clinical and paraclinical features. Their study highlighted that troponin levels had the strongest correlation with MI severity and that ML models, such as random forest and gradient boosting, achieved high accuracy in both classification and regression tasks.

III. Methodology

This section briefly describes our data description, data preprocessing methods, exploratory data analysis, and machine learning methods applied for the predictive models.

A. Data Collection

We utilized a publicly available dataset sourced from the UCI Machine Learning Repository at [6]. This dataset contains 1700 patient data and 124 features, including patient demographics, clinical history, physiological measurements, laboratory values (such as troponin levels), and outcomes associated with myocardial infarction. The outcomes include in-hospital mortality, MI severity, and other cardiac complications (i.e., atrial fibrillation, heart failure, Supraventricular tachycardia, Ventricular tachycardia, Ventricular fibrillation, Myocardial rupture, etc.).The dataset comprises both numerical and categorical variables, covering patient demographics, medical history, cardiovascular conditions, diagnostic tests, and interventions related to myocardial infarction (MI). These attributes can be systematically categorized as follows:

a) Demographic Attributes: Age: Age is a critical risk factor in cardiovascular diseases, including MI. Research has demonstrated a positive correlation between advancing age and MI prevalence due to cumulative vascular damage and reduced cardiac resilience. Gender: Gender differences influence MI risk, with men at higher risk overall, particularly under 60 years. Post-menopausal women, however, exhibit an increased MI risk due to estrogen depletion.

b) Medical History Attributes: Previous Myocardial Infarctions (INF_ANAM) Patients with a history of MI are at a heightened risk for subsequent infarctions, primarily due to existing damage to the cardiac musculature and compromised vascular function. Hypertension (GB): Hypertension is one of the strongest predictors of MI, imposing chronic stress on the heart and arteries, potentially leading to ischemia and MI. Chronic Heart Failure (ZSN): Chronic heart failure indicates existing cardiac insufficiency, a condition that complicates MI management and exacerbates patient outcomes. Arrhythmias and Atrial Fibrillation (e.g., nr11, nr03): Historical data on arrhythmias and fibrillations highlight structural and electrical anomalies within the heart, increasing susceptibility to MI and complicating recovery.

c) Angina and Coronary Disease Attributes: Exertional Angina (STENOK_AN): Angina experienced during exertion often reflects underlying ischemic episodes. Such episodes can lead to MI as the heart's demand for oxygen surpasses its supply. Functional Class of Angina (FK_STENOK): The functional class of angina, specifically Classes III and IV, indicates severe ischemia, correlating with increased MI risk. Coronary Heart Disease (IBS_POST): Recent diagnoses of coronary heart disease, particularly those involving unstable angina, serve as immediate precursors to MI, suggesting increased vulnerability in the days or weeks leading up to the event.

d) Cardiac and Blood Pressure Measurements: Systolic and Diastolic Blood Pressure (S_AD_KBRIG, D_AD_KBRIG): Both elevated and fluctuating blood pressure readings are pivotal indicators of cardiovascular stress, closely linked to MI risk. Heart Failure Severity and Duration (DLIT_AG): The chronicity and intensity of hypertension correlate with increased MI likelihood, as prolonged pressure on cardiac tissues causes progressive structural damage.

e) Laboratory Values and Electrolyte Levels: Serum Potassium and Sodium Levels (K_BLOOD, NA_BLOOD): Electrolyte imbalances significantly influence cardiac function; hypokalemia and hypernatremia, in particular, are associated with heightened MI risk. ALT, AST, and CPK Levels: Elevated alanine aminotransferase (ALT) and aspartate aminotransferase (AST) indicate possible hepatic strain, while creatine phosphokinase (CPK) is elevated in cases of muscle damage, including the myocardium during MI.

f) Electrocardiogram (ECG) Findings: ECG Patterns (e.g., ant_im, lat_im): Indicators such as QRS complex abnormalities in various leads suggest ischemia in specific heart regions (anterior, lateral), each corresponding to different MI locations. ECG Rhythm Variations (e.g., ritm_ecg_p_01, MP_TP_POST): ECG rhythms, especially arrhythmias, are monitored as they often accompany and exacerbate MI.

g) Medication and Intervention History: Medications (e.g., Beta-blockers, Calcium Channel Blockers, Anticoagulants): These medications are commonly administered post-MI to stabilize cardiac function, reduce thrombotic risk, and alleviate myocardial strain. Pain Management (Opioids and NSAIDs): Use of pain management drugs indicates pain severity during cardiac episodes, which is often correlated with ischemic intensity and myocardial stress.

h) Complications and Outcomes: Pulmonary Edema, Cardiogenic Shock, and Myocardial Rupture: Severe complications are tracked as they have significant impacts on patient survival and MI outcomes. Lethal Outcomes (LET_IS): This attribute categorizes mortality causes related to MI, such as cardiogenic shock and myocardial rupture, contributing to understanding fatal MI events.

These data attributes comprehensively capture critical factors associated with MI, offering insights into both risk prediction and outcome assessment. By integrating demographic, clinical, and diagnostic data, this dataset enables robust analyses that can deepen understanding of MI risk profiles, improve prognostic modeling, and inform targeted interventions.

B. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for analysis and model building, ensuring data quality, managing missing values, and transforming variables into suitable formats for machine learning algorithms. For this study, the dataset of 124 attributes related to myocardial infarction (MI) underwent the following preprocessing steps:

a) Identification and Extraction of Numerical Variables: Twelve variables in the dataset were identified as numerical, including age (AGE), systolic and diastolic blood pressure measurements taken at various treatment stages (S_AD_KBRIG, D_AD_KBRIG, S_AD_ORIT, D_AD_ORIT), and biochemical markers like serum potassium (K_BLOOD) and serum sodium (NA_BLOOD). These variables were extracted into a separate DataFrame to streamline the processing of numerical data.

The descriptive statistics of the data reveal notable patterns related to the patient cohort and specific health indicators. The mean age of 62 years suggests an older patient group, although the range extends from 26 to 92 years, reflecting diverse age profiles. Blood pressure values, including systolic and diastolic measures recorded by both the emergency team (S_AD_KBRIG and D_AD_KBRIG) and in intensive care (S_AD_ORIT and D_AD_ORIT), show high means and maximums, indicating patients with significant cardiac stress; however, the minimum values of zero in systolic readings suggest potential data entry errors. Biochemical markers like serum potassium (K_BLOOD) and sodium (NA_BLOOD) indicate instances of extreme electrolyte imbalances, with some values outside the normal ranges, highlighting cases of hypokalemia and hypernatremia that could complicate heart conditions. Furthermore, ALT (ALT_BLOOD), AST (AST_BLOOD), CPK (KFK_BLOOD), white blood cell counts (L_BLOOD), and ESR (ROE) levels include variations, with several outliers that suggest the presence of patients with active inflammation, muscle damage, or compromised liver function, all of which are relevant to myocardial infarction prognosis and severity. These observations underscore the complexity of the patient profiles and the presence of severe cardiovascular and systemic conditions.

b) Outlier Detection and Treatment: : Outlier detection was performed on numerical variables to identify extreme values that might distort model predictions. Attributes such as systolic and diastolic blood pressure (S_AD_KBRIG, D_AD_KBRIG) and serum potassium levels (K_BLOOD) were analyzed for extreme values using Z-score calculations, where values exceeding three standard deviations from the mean were flagged as potential outliers. Outliers were cross-referenced against known medical ranges to determine whether they should be excluded or retained, taking into account each variable's clinical relevance.

Several critical measurements, such as systolic and diastolic blood pressure recorded by both emergency and ICU teams (S_AD_KBRIG, D_AD_KBRIG, S_AD_ORIT, D_AD_ORIT), showed high mean and maximum values, which were consistent with patients experiencing significant cardiovascular strain typical in myocardial infarction (MI) cases. However, the minimum values of zero in systolic blood pressure readings were flagged for potential data inaccuracies, as zero values are clinically implausible and unlikely to occur outside of data entry errors. In these cases, entries with implausible zeros were removed to prevent skewing the analysis.

Attributes such as serum potassium (K_BLOOD) and sodium (NA_BLOOD) levels also displayed extreme outliers, indicating cases of severe hypokalemia or hypernatremia, both of which carry clinical significance in cardiac patients. Given the relevance of electrolyte imbalances in heart function, these outliers were retained if they fell within critical but realistic ranges, as they could be indicative of the severity of the cardiac event.

For additional clinical markers such as liver enzymes (ALT_BLOOD and AST_BLOOD), creatine phosphokinase (CPK or KFK_BLOOD), white blood cell counts (L_BLOOD), and erythrocyte sedimentation rate (ESR or ROE), outliers were also assessed against typical medical reference values. Elevated levels in these markers are common in cases of systemic inflammation or muscle damage, often associated with acute MI. Consequently, these outliers were retained in cases where they aligned with potential underlying cardiac complications, ensuring the model could learn from cases of heightened cardiac stress or inflammation.

Outliers not aligning with clinical plausibility or realistic ranges were excluded or flagged for correction, balancing data integrity with analytical accuracy. This selective approach ensured that the dataset reflected realistic medical conditions while preventing distortion from data errors.

c) Handling Missing Values: Median imputation was applied to handle missing values across numerical variables in the dataset, with a focus on variables prone to skewed distributions or outliers. This imputation method was selected to preserve the central tendency of the data without being influenced by extreme values, as mean imputation could have skewed results due to outliers. Variables such as systolic and diastolic blood pressure (S_AD_KBRIG, D_AD_KBRIG, S_AD_ORIT, D_AD_ORIT) and biochemical markers like serum potassium (K_BLOOD) and sodium (NA_BLOOD) were imputed with median values to ensure robustness against

the influence of outliers, which are prevalent in cardiac emergency datasets.

The approach ensured that imputed values reflected typical observations in the dataset while minimizing the effect of extreme readings in critical variables. Additionally, attributes like ALT (ALT_BLOOD), AST (AST_BLOOD), and CPK (KFK_BLOOD) were imputed using their respective medians to retain data consistency, especially where missing values occurred alongside variable measurements indicating cardiac stress or liver function abnormalities. This approach preserved data integrity, providing a stable baseline for analysis and avoiding distortions from extreme values while ensuring that medically relevant values remained consistent in their distributions.

d) Categorical Variables: Missing values in categorical attributes were managed by assigning them to separate categories, such as "unknown" or "not available," which was particularly useful for attributes with substantial missing proportions, like heredity on coronary heart disease (IBS_NASL). This imputation ensured that important categorical data with missing entries could still contribute to the model without bias.

To make the categorical data suitable for analysis, one-hot encoding was applied to nominal categorical variables, converting each category into a distinct binary indicator. This transformation was applied to variables such as gender (SEX) and various medical history indicators, preserving the full detail of categorical distinctions. For ordinal categorical variables, such as functional class of angina (FK_STENOK) and hypertension stage (GB), integer encoding was used to retain their ordered relationships. This ensured that the relational structure of severity or progression remained intact, providing a model-ready dataset where the hierarchy of categories was respected. This encoding strategy facilitated an accurate integration of categorical data into the model, maintaining interpretability and robustness in the machine learning process.

C. Machine Learning Models

To predict myocardial infarction-related outcomes such as atrial fibrillation, and heart failure, several machine learning models were applied, each selected for specific advantages in handling the dataset's blend of categorical and numerical variables, as well as for interpretability and predictive performance.

- Logistic Regression (LR): Logistic regression was chosen as a baseline model for its simplicity, interpretability, and suitability for binary classification tasks. It effectively models the probability of outcomes like atrial fibrillation or chronic heart failure based on predictors such as age, blood pressure, and medical history. The coefficients derived from logistic regression allow for straightforward interpretation of feature importance, helping identify critical risk factors associated with myocardial infarction (MI).
- Decision Tree Model: The decision tree model was utilized for its capability to create interpretable, rule-based classifications. By partitioning the data based on feature values, decision trees provide a clear structure to understand how attributes like blood pressure or history of arrhythmias contribute to MI outcomes. Additionally, decision trees handle both categorical and numerical features and are resilient to missing data.
- Random Forest Classifier: Random forest, an ensemble method, was employed for its strong predictive power and robustness against overfitting. It constructs multiple decision trees, each trained on different data subsets, and aggregates their results, improving stability and accuracy. This model is particularly valuable in identifying important features through feature importance scores, aiding in the assessment of significant MI-related risk factors. Additionally, random forest handles both outliers and missing data effectively, making it well-suited for complex clinical datasets.
- Gradient Boosting Classifier: Gradient boosting, particularly implementations like XGBoost and LightGBM, was used for its ability to model complex, non-linear relationships. By iteratively training an ensemble of weak learners, gradient boosting minimizes errors and captures interactions between variables. This high-accuracy model is advantageous in medical datasets where intricate patterns often underlie patient outcomes. Its advanced feature handling and ability to manage categorical and numerical data interactions make it a robust choice for predictive analysis in myocardial infarction datasets.

D. Feature Selection

In preparing the data for modeling, two target variables—presence of atrial fibrillation (FIBR_PREDS) and chronic heart failure (ZSN)—were separated from the feature set. These binary outcome variables indicate whether atrial fibrillation or chronic heart failure was observed, and they serve as the primary targets for predicting myocardial infarction-related complications.

For categorical feature transformation, one-hot encoding was applied to nominal variables, converting each category into distinct binary columns. This encoding preserved each category's unique information without introducing artificial ordinal relationships that could mislead models. For example, variables like gender (SEX) and various medication indicators (e.g., LID_S_n for lidocaine use) were expanded into binary columns for each possible category.

Ordinal variables, which represent ranked information (e.g., functional class of angina (FK_STENOK) and stages of hypertension (GB)), were treated differently; they were numerically encoded to maintain their inherent order and rank. This approach retained meaningful relationships within these categories, ensuring that ordinal data contributes effectively to the modeling process without arbitrary rankings. The encoding strategy enabled full utilization of categorical data in machine learning algorithms while minimizing bias from unintended category hierarchies.

E. Exploratory Data Analysis (EDA)

In the univariate analysis, histograms were generated for each numerical feature to assess distributions, central tendencies, and outliers. Key variables such as age, blood pressure, and serum biochemical markers were visualized to better understand the demographic and clinical characteristics of the dataset. The histogram for age shows a concentration around the mean of 62 years, with a range from 26 to 92 years as shown in Figure 1. This indicates a focus on older patients, typically at higher risk for cardiac conditions. Most ages are clustered between 50 and 70, representing a high-risk age group for myocardial infarction (MI). Figure 2 shows the gender distribution in the dataset. There are 63% males and 37% females in the data which shows that more MI patients belong to male as compared to females. Figures 3 and 4

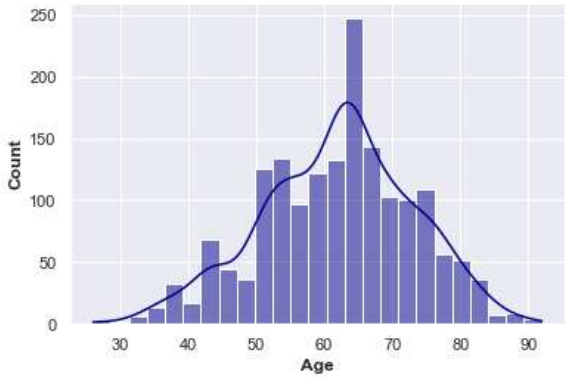

Fig. 1. Age distribution

shows the histograms of systolic (S_AD_KBRIG) and diastolic (D_AD_KBRIG) blood pressure measured by the emergency team reveal wide ranges with peaks in the hypertensive range (e.g., 120–160 mmHg for systolic blood pressure). Values of zero in systolic readings suggest potential inaccuracies, as non-zero blood pressure is expected under clinical conditions. Blood pressure levels recorded in the ICU (S_AD_ORIT and D_AD_ORIT) also show elevated peaks, reflecting the presence of cardiac stress in ICU-admitted patients.

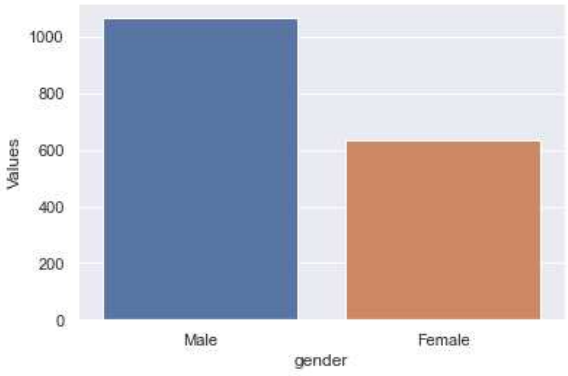

Fig. 2. Gender distribution

Figure 5 and 6 shows the histograms for the blood chemistry for the target variable FIBR_PREDS and ZSN. The markers such as potassium (K_BLOOD), sodium (NA_BLOOD), and liver enzymes (ALT_BLOOD and AST_BLOOD) show wide distributions with values outside typical ranges, signaling possible electrolyte imbalances and systemic inflammation, which are common in acute MI cases. Elevated levels in creatine phosphokinase (KFK_BLOOD) and white blood cell counts (L_BLOOD) suggest tissue damage and inflammation, commonly associated with myocardial infarction. While most values fall within normal ranges, significant outliers are present, highlighting cases of extreme physiological distress.

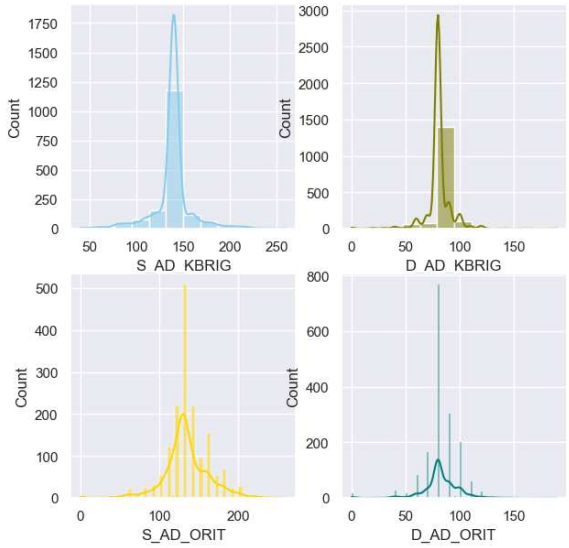

Fig. 3. Histograms for Systolic and Diastolic Blood Pressure for the target variable (FIBR_PREDS)

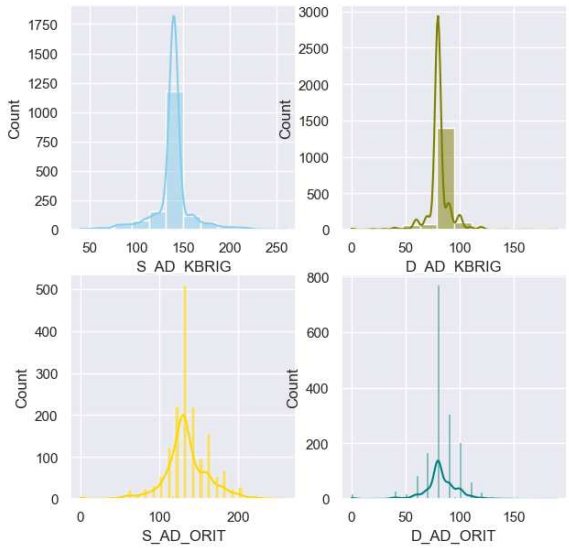

Fig. 4. Histograms for Systolic and Diastolic Blood Pressure for the target variable (ZSN)

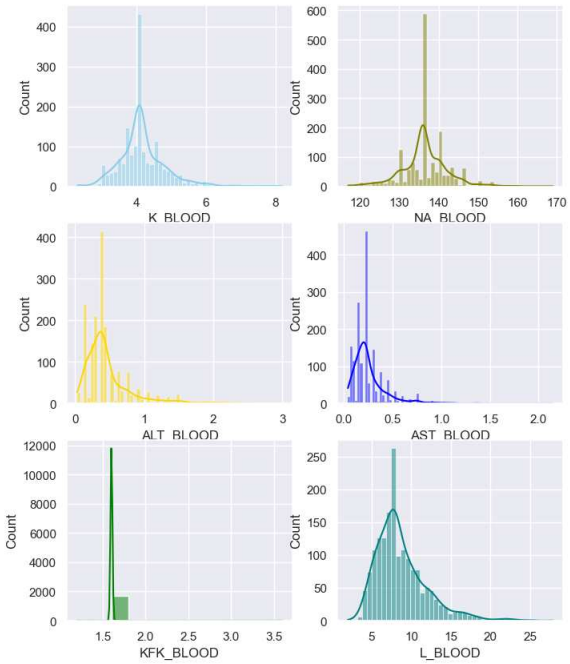

Fig. 5. Histograms for the Blood Chemistry for the target variable (FIBR_PREDS)

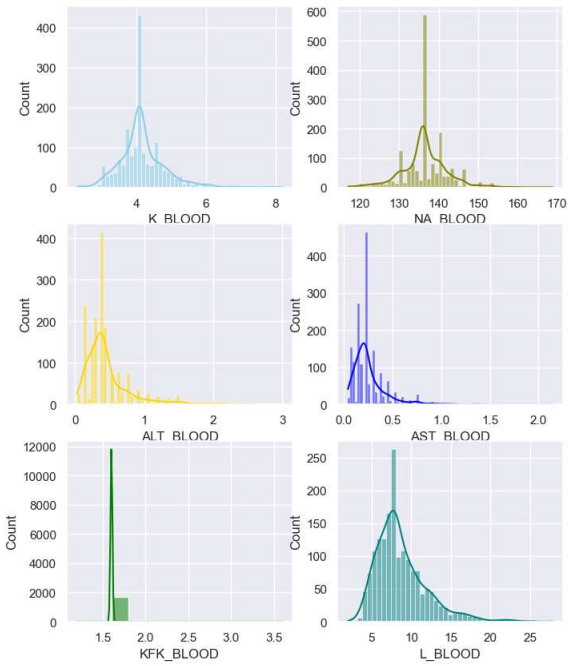

Fig. 6. Histograms for the Blood Chemistry for the target variable (ZSN)

Scatter plots and Pearson correlation coefficients were used to investigate relationships between numerical features, revealing underlying trends as shown in figures 7 to 10. Scatter plots between age and systolic blood pressure (both in the ER and ICU settings) show a dispersed pattern with no strong linear relationship, though clustering around common blood pressure values suggests general age-group trends. Scatter plots between serum potassium and sodium levels against blood pressure values show weak associations, indicating that these markers vary widely within the population. However, their outlier values contribute valuable insights into patient conditions with abnormal electrolyte levels. The Pearson coefficients indicate moderate correlations between systolic and diastolic blood pressure across different treatment stages, particularly between ER and ICU values, suggesting consistency in patient profiles through stages of care. Age has low correlation values with most other health indicators, emphasizing that while age is critical in risk assessment, it is independent of other physiological metrics in the dataset.

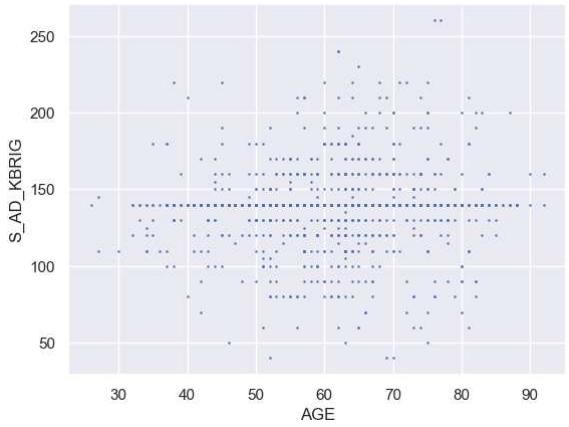

Fig. 7. Scatter plot for the Age and Systolic Blood Pressure for the target variable (FIBR_PREDS)

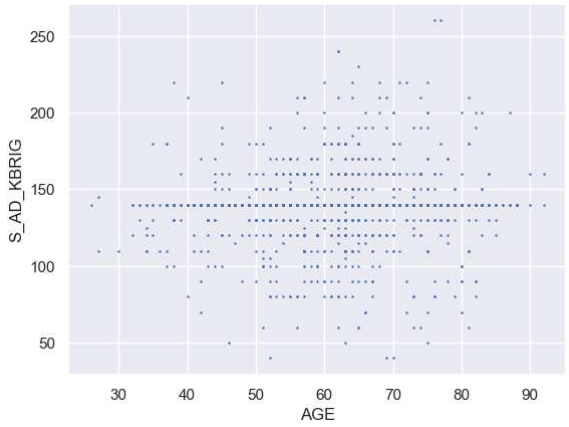

Fig. 8. Scatter plot for the Age and Systolic Blood Pressure for the target variable (ZSN)

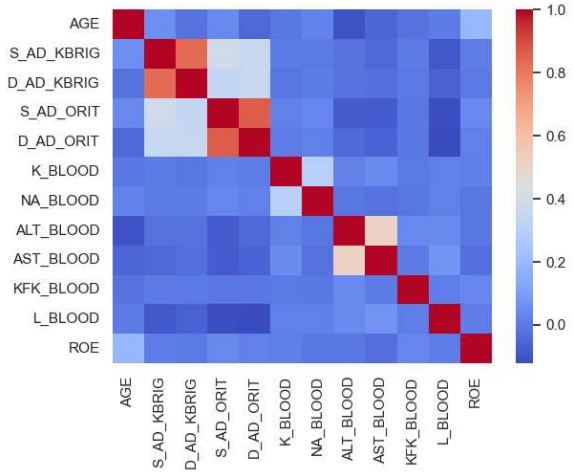

Fig. 9. Pearson correlation coefficients between numerical variables for the target variable (FIBR_PREDS)

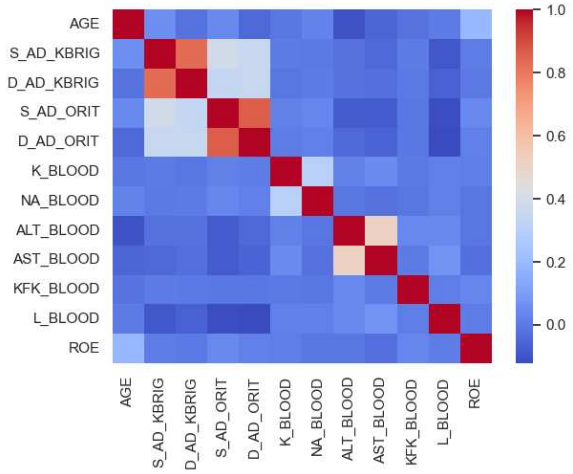

Fig. 10. Pearson correlation coefficients between numerical variables for the target variable (ZSN)

IV. Experiments and Results

The dataset comprises 1,700 entries with 124 features. Initial preprocessing involved substantial cleaning and separating categorical features, followed by one-hot encoding to generate 158 dummy variables from the original categorical features. To prevent multicollinearity, one category was removed from each feature group to retain only $K-1$ indicators per categorical variable. For analysis, two target variables were considered:

- Fibrillation Prediction (FIBR_PREDS): This binary classification target indicates whether fibrillation was present (0 or 1) in a patient.
- Heart Failure (ZSN): This target variable, also binary, denotes the presence of heart failure complications post-myocardial infarction.

The dataset was divided into an 80-20 split for training and test sets, providing sufficient data for training and unbiased evaluation. Numerical features, including age, blood pressure readings, and biochemical markers, were standardized using StandardScaler to ensure uniform feature scales and prevent any single feature from disproportionately impacting the model. Two primary models were chosen: Logistic Regression and Random Forest, both suited for classification tasks. Logistic Regression is valued for its interpretability, while Random Forest provides robustness and higher accuracy in complex feature spaces. Hyperparameters were tuned using GridSearchCV and RandomizedSearchCV: For Logistic Regression, regularization parameters were optimized. For Random Forest, parameters like the number of estimators (trees) and maximum tree depth were adjusted. The models were evaluated using metrics such as accuracy, mean absolute error (MAE), mean squared error (MSE), and R^2 scores. For additional clustering-based analysis, Silhouette Score was applied where applicable.

A. Experimental results

Based on the two target variables, Fibrillation Prediction (FIBR_PREDS) and Heart Failure (ZSN), the model evaluation results presented in the research highlight the performance of four predictive models on the myocardial infarction complications dataset. The evaluation criteria include training and testing accuracy to determine each model's generalizability. Tables 1 and 2 summarize the prediction results for atrial fibrillation and heart failure for the dataset.

TABLE I
Atrial Fibrillation Prediction Results

ML Classifiers	Training Accuracy (%)	Testing Accuracy (%)
Logistic Regression	92.9	83.8
Decision Tree	91.5	85.6
Random Forest	91	85.9
XGBoost	100	86.8
Multilayer Perceptron	100	91

As shown in tables 1 and 2, the Logistic Regression model showed consistent generalizability across both targets, with a slight decrease in accuracy from training to testing, signaling potential for further fine-tuning. The Decision Tree model achieved moderate stability across both targets, maintaining slightly better test accuracy than Logistic Regression. Random Forest demonstrated robust performance and good generalizability across both targets, with competitive test accuracies.

TABLE II
Heart Failure Prediction Results

ML Classifiers	Training Accuracy (%)	Testing Accuracy (%)
Logistic Regression	82.3	77.1
Decision Tree	83.5	79.1
Random Forest	81	78.4
XGBoost	92	83.5
Multilayer Perceptron	100	84

Multilayer perception achieved the highest testing accuracy for trail fibrillation prediction however XGBoost outperforms in the heart failure prediction. Further tuning

could help improve generalizability for Fibrillation Prediction (FIBR_PREDS). Overall, the evaluation results suggest that XGBoost and Multilayer perceptron provide the most accurate predictions for myocardial infarction complications in this dataset.

B. Feature Importance Plots

We have further examined the most important features in the training dataset using the best performing machine learning algorithms. For both Atrial fibrillation and heart failure prediction, we used feature importance plots to illustrate the top 10 features in the training data.

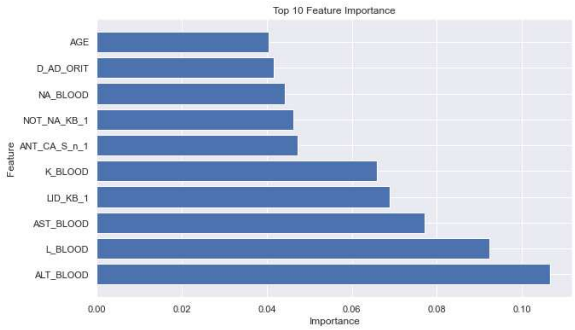

Fig. 11. Feature Importance plot for FIBR_PREDS

Figure 11 shows the feature importance plot using XGBoost algorithm on atrial fibrillation prediction. As shown in Figure 11, blood based clinical features are the most critical in predicting the atrial fibrillation with the most important feature being ALT_BLOOD which is the serum ALT content in blood. Furthermore, L_BLOOD and AST_BLOOD also plays a critical role for atrial fibrillation.

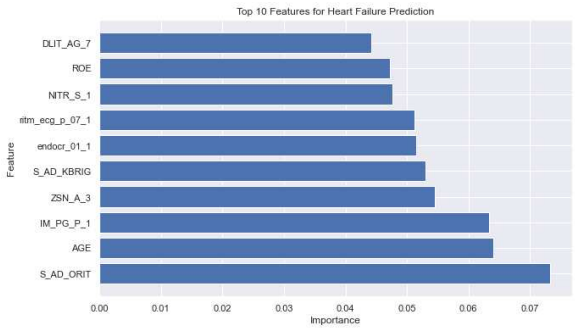

Fig. 12. Feature Importance plot for Heart Failure

Figure 12 shows the feature importance plot using XGBoost algorithm on heart failure prediction. As shown in Figure 12, S_AD_ORIT which represents the systolic blood pressure measured in ICU unit is the most important feature responsible for heart failure. Moeover, IM_PG_P representing the presence/absence of a right ventricular myocardial infarction and Age are also critical factors for heart failure.

C. Shapley Plots

Shapley values are widely used to explain the black-box models to interpret the results. We have used Shapley to illustrate the multilayer perception model for both atrial fibrillation and heart failure prediction.

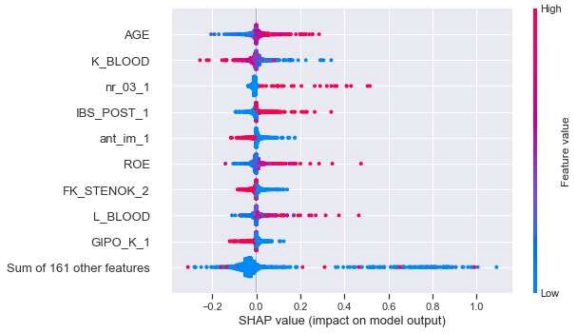

Fig. 13. Shapley plot for Atrial Fibrillation

Figure 13 shows the shapley plot using multilayer perceptron algorithm on atrial fibrillation prediction. As shown in Figure 13, age and K_BLOOD (capturing serum potassium content in blood) are most important features for the prediction of atrial fibrillation.

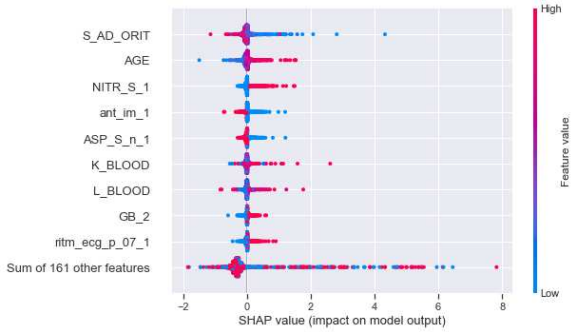

Fig. 14. Shapley plot for Heart Failure

Figure 14 shows the shapley plot using multilayer perceptron algorithm on heart failure prediction. As shown in Figure 14, age and S_AD_ORIT (capturing systolic blood pressure in ICU) are most important features for the prediction of heart failure.

Overall, these results aligns with the feature importance plots and clinically relevant features of predicting the atrial fibrillation and heart failure conditions for myocardial infraction patients.

V. Conclusion

This paper effectively demonstrates a predictive model for assessing myocardial infarction complications, specifically focusing on fibrillation and heart failure conditions. Through a meticulous preprocessing pipeline—including handling missing values, encoding categorical data, and standardizing numerical variables—the data was prepared for robust model

training. These results indicate that the models effectively capture significant patterns within the dataset and perform well in predicting complications. Atrial fibrillation and heart failure prediction achieved strong performance across models, clustering analysis further indicated potential patient subgrouping based on silhouette scores. This suggests that clustering could assist in identifying patient subgroups with specific risks, aiding in tailored clinical interventions. Future research will aim to overcome these limitations by integrating additional clinical variables and longitudinal data to improve predictive accuracy and robustness. Additionally, exploring ensemble methods or deep learning models may better capture complex data relationships. Expanding the model's use to other cardiovascular conditions and incorporating more risk factors could further increase its clinical utility, promoting comprehensive and individualized patient care.

ACKNOWLEDGMENT

This project was supported, by grant from the U.S. National Science Foundation (IIS2334391). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of NSF.

REFERENCES

[1] S. Jadhav et al. Machine learning in cardiovascular disease prediction. *Journal of Medical Systems*, 44(2):27–34, 2020.

[2] A. Agham et al. Neural networks in healthcare: Predicting myocardial infarction using deep learning techniques. *IEEE Access*, 8(7):12312–12318, 2021.

[3] R. Khera, J. Haimovich, N. C. Hurley, et al. Use of machine learning models to predict death after acute myocardial infarction. *Canadian Journal of Cardiology*, 2021.

[4] M. P. Than, J. W. Pickering, Y. Sandoval, et al. Machine learning to predict the likelihood of acute myocardial infarction. *Circulation*, 140:899–909, 2019.

[5] Z. Chen, J. Shi, T. Pommier, et al. Prediction of myocardial infarction from patient features with machine learning. *Frontiers in Cardiovascular Medicine*, 9:754609, 2022.

[6] S. Golovenkin, V. Shulman, D. Rossiev, P. Shesternya, S. Nikulina, Y. Orlova, and V. Voino-Yasenetsky. Myocardial infarction complications. UCI Machine Learning Repository, 2020. Dataset.