Predicting Heart Failure Complications Post-Myocardial Infarction

By

Ganesh Vannam & Nandhika Rajmanikandan

MA5790

December 2024

Abstract:

Myocardial Infarction (MI), or heart attack, occurs when blood flow to the heart is obstructed, causing damage to the heart muscle. It is a leading cause of morbidity and mortality worldwide, with complications such as chronic heart failure, cardiogenic shock, and arrhythmias significantly impacting patient outcomes. This study focuses on predicting chronic heart failure (ZSN) using a dataset collected in Krasnoyarsk between 1992 and 1995, consisting of 1,700 patients with 122 variables, including 14 continuous and 108 categorical predictors. Data preprocessing involved addressing 15,794 missing values using KNN imputation, removing near-zero variance features, and applying Box-Cox and Spatial Sign transformations to handle skewness and outliers, resulting in 81 features for analysis. Models were evaluated using stratified random sampling, 10-fold cross-validation, and the Kappa metric to address the imbalanced target variable. Partial Least Squares Discriminant Analysis (PLSDA) achieved the highest performance among linear models with a testing Kappa value of 0.3842, followed by Logistic Regression with a Kappa value of 0.357.

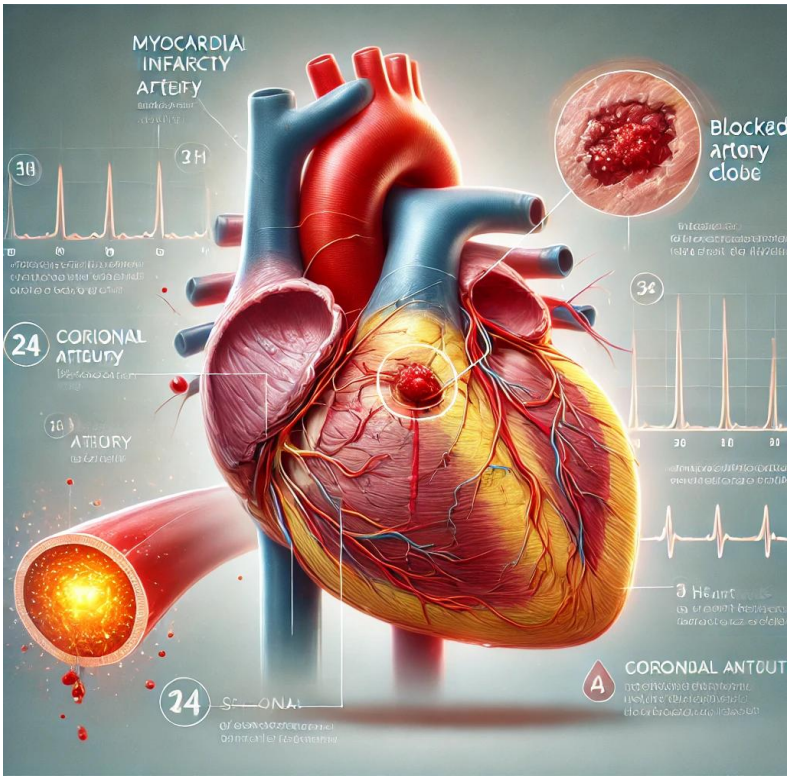


Table of Contents

1. Background

Myocardial Infarction (MI), commonly known as a heart attack, is a medical emergency caused by the sudden reduction or cessation of blood flow to the heart muscle, leading to tissue damage. This blockage typically results from the buildup of fatty deposits (plaque) in the coronary arteries, which may rupture and form a blood clot. MI remains one of the leading causes of death worldwide, particularly in urban populations exposed to chronic stress, sedentary lifestyles, and unbalanced diets. In the United States alone, over a million people suffer from MI annually, with 200,000–300,000 succumbing to acute events before receiving medical care.

The clinical presentation of MI varies widely. Typical symptoms include chest pain or discomfort, which may radiate to the arms, back, neck, or jaw. Other signs include shortness of breath, nausea, cold sweats, and fatigue. Women, older adults, and individuals with atypical presentations may lack chest pain but experience other symptoms, making early diagnosis challenging. Complications such as chronic heart failure, cardiogenic shock, arrhythmias, and cardiac arrest further complicate the disease course, often leading to long-term health challenges or mortality.

The variability in MI progression underscores the need for predictive models to assess patient risk and anticipate complications. Leveraging clinical data to identify high-risk patients can facilitate early interventions, improve patient outcomes, and reduce the burden on healthcare systems. This study utilizes a dataset collected in Krasnoyarsk from 1992 to 1995 to develop predictive models for chronic heart failure post-MI, contributing to advancements in personalized patient care.

Data source: **https://www.kaggle.com/datasets/rafatashrafjoy/myocardial-infarction-complications**

2. Goal of the Study

The primary goal of this study is to predict chronic heart failure (ZSN) complications in patients who have experienced myocardial infarction (MI) using clinical data collected during hospitalization. By identifying high-risk patients early, this research aims to enhance critical care management, facilitate timely preventive measures, and improve long-term patient outcomes. The study also seeks to evaluate and compare the performance of various linear and non-linear predictive models, providing insights into the most effective approaches for addressing the challenges posed by imbalanced datasets and diverse clinical variables.

3. Dataset Structure and Variable Description

The dataset used in this study contains clinical data related to myocardial infarction (MI) and its complications. It includes 122 variables, of which 14 are continuous and 108 are categorical. Our target variable is Chronic Heart Failure (ZSN). Below is a detailed description of the variables.

- **Chronic Heart Failure (ZSN):** Indicates if the patient had chronic heart failure (0: No, 1: Yes).
- **Age (AGE)**: The patient's age.
- **Gender (SEX)**: The patient's gender (0: Female, 1: Male).
- **Myocardial Infarctions (INF_ANAM)**: The number of myocardial infarctions in the patient's anamnesis (0: None, 1: One, 2: Two, 3: Three or more).

- **Exertional Angina Pectoris (STENOK_AN)**: History of exertional angina pectoris (0: Never, 1: Last year, 2: 1 year ago, 3: 2 years ago, 4: 3 years ago, 5: 4-5 years ago, 6: >5 years ago).
- **Functional Class of Angina (FK_STENOK)**: The functional class of angina pectoris in the last year (0: None, 1: I FC, 2: II FC, 3: III FC, 4: IV FC).
- **Coronary Heart Disease (IBS_POST)**: Presence of coronary heart disease (0: No, 1: Exertional angina, 2: Unstable angina).
- **Heredity on CHD (IBS_NASL)**: Family history of coronary heart disease (0: Not burdened, 1: Burdened).
- **Essential Hypertension (GB)**: Stage of essential hypertension (0: None, 1: Stage 1, 2: Stage 2, 3: Stage 3).
- **Symptomatic Hypertension (SIM_GIPERT)**: Presence of symptomatic hypertension (0: No, 1: Yes).
- **Duration of Hypertension (DLIT_AG)**: Duration of hypertension (0: None, 1: 1 year, 2: 2 years, 3: 3 years, 4: 4 years, 5: 5 years, 6: 6-10 years, 7: >10 years).
- **Chronic Heart Failure (ZSN_A)**: Presence of chronic heart failure (0: None, 1: Stage I, 2: Stage IIA (right), 3: Stage IIA (left), 4: Stage IIB).
- **Arrhythmia (nr11)**: History of arrhythmia (0: No, 1: Yes).
- **Premature Atrial Contractions (nr01)**: Presence of premature atrial contractions (0: No, 1: Yes).
- **Premature Ventricular Contractions (nr02)**: Presence of premature ventricular contractions (0: No, 1: Yes).
- **Paroxysms of Atrial Fibrillation (nr03)**: History of paroxysms of atrial fibrillation (0: No, 1: Yes).
- **Persistent Atrial Fibrillation (nr04)**: History of persistent atrial fibrillation (0: No, 1: Yes).
- **Ventricular Fibrillation (nr07)**: History of ventricular fibrillation (0: No, 1: Yes).
- **Ventricular Paroxysmal Tachycardia (nr08)**: History of ventricular paroxysmal tachycardia (0: No, 1: Yes).
- **First-Degree AV Block (np01)**: History of first-degree AV block (0: No, 1: Yes).
- **Third-Degree AV Block (np04)**: History of third-degree AV block (0: No, 1: Yes).
- **Left Bundle Branch Block (LBBB) (np05)**: Presence of LBBB (0: No, 1: Yes).
- **Incomplete LBBB (np07)**: Presence of incomplete LBBB (0: No, 1: Yes).
- **Complete LBBB (np08)**: Presence of complete LBBB (0: No, 1: Yes).
- **Incomplete Right Bundle Branch Block (RBBB) (np09)**: Presence of incomplete RBBB (0: No, 1: Yes).
- **Complete Right Bundle Branch Block (RBBB) (np10)**: Presence of complete RBBB (0: No, 1: Yes).
- **Diabetes Mellitus (endocr_01)**: Presence of diabetes mellitus (0: No, 1: Yes).
- **Obesity (endocr_02)**: Presence of obesity (0: No, 1: Yes).
- **Thyrotoxicosis (endocr_03)**: Presence of thyrotoxicosis (0: No, 1: Yes).
- **Chronic Bronchitis (zab_leg_01)**: History of chronic bronchitis (0: No, 1: Yes).
- **Obstructive Chronic Bronchitis (zab_leg_02)**: History of obstructive chronic bronchitis (0: No, 1: Yes).
- **Bronchial Asthma (zab_leg_03)**: History of bronchial asthma (0: No, 1: Yes).

- **Chronic Pneumonia (zab_leg_04)**: History of chronic pneumonia (0: No, 1: Yes).
- **Pulmonary Tuberculosis (zab_leg_06)**: History of pulmonary tuberculosis (0: No, 1: Yes).
- **Systolic BP (S_AD_KBRIG)**: Systolic blood pressure according to the Emergency Cardiology Team (mmHg).
- **Diastolic BP (D_AD_KBRIG)**: Diastolic blood pressure according to the Emergency Cardiology Team (mmHg).
- **Systolic BP (S_AD_ORIT)**: Systolic blood pressure according to the Intensive Care Unit (mmHg).
- **Diastolic BP (D_AD_ORIT)**: Diastolic blood pressure according to the Intensive Care Unit (mmHg).
- **Pulmonary Edema (O_L_POST)**: Presence of pulmonary edema at ICU admission (0: No, 1: Yes).
- **Cardiogenic Shock (K_SH_POST)**: Presence of cardiogenic shock at ICU admission (0: No, 1: Yes).
- **Record ID (ID)**: Unique identifier for each record.
- **Paroxysms of Atrial Fibrillation at ICU Admission (MP_TP_POST)**: Paroxysms of atrial fibrillation at the time of admission to intensive care (0: No, 1: Yes).
- **Paroxysms of Supraventricular Tachycardia at ICU Admission (SVT_POST)**: Paroxysms of supraventricular tachycardia at the time of admission to intensive care (0: No, 1: Yes).
- **Paroxysms of Ventricular Tachycardia at ICU Admission (GT_POST)**: Paroxysms of ventricular tachycardia at the time of admission to intensive care (0: No, 1: Yes).
- **Ventricular Fibrillation at ICU Admission (FIB_G_POST)**: Ventricular fibrillation at the time of admission to intensive care (0: No, 1: Yes).
- **Anterior Myocardial Infarction (ECG changes in V1-V4) (ant_im)**: Presence of anterior myocardial infarction (0: No infarct, 1: No changes, 2: QR-complex, 3: Qr-complex, 4: QS-complex).
- **Lateral Myocardial Infarction (ECG changes in V5-V6, I, AVL) (lat_im)**: Presence of lateral myocardial infarction (0: No infarct, 1: No changes, 2: QR-complex, 3: Qr-complex, 4: QS-complex).
- **Inferior Myocardial Infarction (ECG changes in III, AVF, II) (inf_im)**: Presence of inferior myocardial infarction (0: No infarct, 1: No changes, 2: QR-complex, 3: Qr-complex, 4: QS-complex).
- **Posterior Myocardial Infarction (ECG changes in V7-V9) (post_im)**: Presence of posterior myocardial infarction (0: No infarct, 1: No changes, 2: QR-complex, 3: Qr-complex, 4: QS-complex).
- **Right Ventricular Myocardial Infarction (IM_PG_P)**: Presence of right ventricular myocardial infarction (0: No, 1: Yes).
- **ECG Rhythm at Admission - Sinus Rhythm (ritm_ecg_p_01)**: ECG rhythm at the time of admission to hospital (0: No, 1: Yes).
- **ECG Rhythm at Admission - Atrial Fibrillation (ritm_ecg_p_02)**: ECG rhythm at the time of admission to hospital (0: No, 1: Yes).
- **ECG Rhythm at Admission - Atrial Rhythm (ritm_ecg_p_04)**: ECG rhythm at the time of admission to hospital (0: No, 1: Yes).

- **ECG Rhythm at Admission - Idioventricular Rhythm (ritm_ecg_p_06)**: ECG rhythm at the time of admission to hospital (0: No, 1: Yes).
- **ECG Rhythm at Admission - Tachycardia (ritm_ecg_p_07)**: ECG rhythm at the time of admission to hospital with heart rate above 90 (0: No, 1: Yes).
- **ECG Rhythm at Admission - Bradycardia (ritm_ecg_p_08)**: ECG rhythm at the time of admission to hospital with heart rate below 60 (0: No, 1: Yes).
- **Premature Atrial Contractions at Admission (n_r_ecg_p_01)**: Presence of premature atrial contractions on ECG at the time of admission (0: No, 1: Yes).
- **Frequent Premature Atrial Contractions at Admission (n_r_ecg_p_02)**: Presence of frequent premature atrial contractions on ECG at the time of admission (0: No, 1: Yes).
- **Premature Ventricular Contractions at Admission (n_r_ecg_p_03)**: Presence of premature ventricular contractions on ECG at the time of admission (0: No, 1: Yes).
- **Frequent Premature Ventricular Contractions at Admission (n_r_ecg_p_04)**: Presence of frequent premature ventricular contractions on ECG at the time of admission (0: No, 1: Yes).
- **Paroxysms of Atrial Fibrillation at Admission (n_r_ecg_p_05)**: Presence of paroxysms of atrial fibrillation on ECG at the time of admission (0: No, 1: Yes).
- **Persistent Atrial Fibrillation at Admission (n_r_ecg_p_06)**: Presence of persistent atrial fibrillation on ECG at the time of admission (0: No, 1: Yes).
- **Paroxysms of Supraventricular Tachycardia at Admission (n_r_ecg_p_08)**: Presence of paroxysms of supraventricular tachycardia on ECG at the time of admission (0: No, 1: Yes).
- **Paroxysms of Ventricular Tachycardia at Admission (n_r_ecg_p_09)**: Presence of paroxysms of ventricular tachycardia on ECG at the time of admission (0: No, 1: Yes).
- **Ventricular Fibrillation at Admission (n_r_ecg_p_10)**: Presence of ventricular fibrillation on ECG at the time of admission (0: No, 1: Yes).
- **Sinoatrial Block at Admission (n_p_ecg_p_01)**: Presence of sinoatrial block on ECG at the time of admission (0: No, 1: Yes).
- **First-Degree AV Block at Admission (n_p_ecg_p_03)**: Presence of first-degree AV block on ECG at the time of admission (0: No, 1: Yes).
- **Type 1 Second-Degree AV Block (Mobitz I/Wenckebach) at Admission (n_p_ecg_p_04)**: Presence of Type 1 second-degree AV block (Mobitz I) on ECG at the time of admission (0: No, 1: Yes).
- **Type 2 Second-Degree AV Block (Mobitz II/Hay) at Admission (n_p_ecg_p_05)**: Presence of Type 2 second-degree AV block (Mobitz II) on ECG at the time of admission (0: No, 1: Yes).
- **Third-Degree AV Block at Admission (n_p_ecg_p_06)**: Presence of third-degree AV block on ECG at the time of admission (0: No, 1: Yes).
- **LBBB (Anterior Branch) at Admission (n_p_ecg_p_07)**: Presence of LBBB (anterior branch) on ECG at the time of admission (0: No, 1: Yes).
- **LBBB (Posterior Branch) at Admission (n_p_ecg_p_08)**: Presence of LBBB (posterior branch) on ECG at the time of admission (0: No, 1: Yes).
- **Incomplete LBBB at Admission (n_p_ecg_p_09)**: Presence of incomplete LBBB on ECG at the time of admission (0: No, 1: Yes).

- **Complete LBBB at Admission (n_p_ecg_p_10)**: Presence of complete LBBB on ECG at the time of admission (0: No, 1: Yes).
- **Incomplete RBBB at Admission (n_p_ecg_p_11)**: Presence of incomplete RBBB on ECG at the time of admission (0: No, 1: Yes).
- **Complete RBBB at Admission (n_p_ecg_p_12)**: Presence of complete RBBB on ECG at the time of admission (0: No, 1: Yes).
- **Fibrinolytic Therapy (Celiasum 750k IU) (fibr_ter_01)**: Administration of fibrinolytic therapy by Celiasum 750k IU (0: No, 1: Yes).
- **Fibrinolytic Therapy (Celiasum 1m IU) (fibr_ter_02)**: Administration of fibrinolytic therapy by Celiasum 1m IU (0: No, 1: Yes).
- **Fibrinolytic Therapy (Celiasum 3m IU) (fibr_ter_03)**: Administration of fibrinolytic therapy by Celiasum 3m IU (0: No, 1: Yes).
- **Fibrinolytic Therapy (Streptase) (fibr_ter_05)**: Administration of fibrinolytic therapy by Streptase (0: No, 1: Yes).
- **Fibrinolytic Therapy (Celiasum 500k IU) (fibr_ter_06)**: Administration of fibrinolytic therapy by Celiasum 500k IU (0: No, 1: Yes)
- **Fibrinolytic Therapy by Celiasum 250k IU (fibr_ter_07):** Indicates if fibrinolytic therapy with Celiasum 250k IU was administered (0: No, 1: Yes).
- **Fibrinolytic Therapy by Streptodecase 1.5m IU (fibr_ter_08):** Indicates if fibrinolytic therapy with Streptodecase 1.5m IU was administered (0: No, 1: Yes).
- **Hypokalemia (< 4 mmol/L) (GIPO_K):** Indicates if the patient has hypokalemia (0: No, 1: Yes).
- **Serum Potassium Content (K_BLOOD):** The patient's serum potassium content in mmol/L.
- **Increase of Sodium in Serum (>150 mmol/L) (GIPER_Na):** Indicates if there is an increase of sodium in serum (0: No, 1: Yes).
- **Serum Sodium Content (Na_BLOOD):** The patient's serum sodium content in mmol/L.
- **Serum AlAT Content (ALT_BLOOD):** The patient's serum AlAT content in IU/L.
- **Serum AsAT Content (AST_BLOOD):** The patient's serum AsAT content in IU/L.
- **Serum CPK Content (KFK_BLOOD):** The patient's serum CPK content in IU/L.
- **White Blood Cell Count (L_BLOOD):** The patient's white blood cell count in billions per liter.
- **ESR (Erythrocyte Sedimentation Rate) (ROE):** The patient's erythrocyte sedimentation rate in mm.
- **Time Elapsed from the Beginning of the Attack of CHD to the Hospital (TIME_B_S):** Indicates the time elapsed from the onset of coronary heart disease (0: Less than 2 hours, 1: 2-4 hours, 2: 4-6 hours, 3: 6-8 hours, 4: 8-12 hours, 5: 12-24 hours, 6: More than 1 day, 7: More than 2 days, 8: More than 3 days).
- **Relapse of Pain in the First Hours of the Hospital Period (R_AB_1_n):** Indicates if there was a relapse of pain in the first hours (0: No, 1: Only one, 2: 2 times, 3: 3 or more times).
- **Relapse of Pain in the Second Day of the Hospital Period (R_AB_2_n):** Indicates if there was a relapse of pain on the second day (0: No, 1: Only one, 2: 2 times, 3: 3 or more times).

- **Relapse of Pain in the Third Day of the Hospital Period (R_AB_3_n):** Indicates if there was a relapse of pain on the third day (0: No, 1: Only one, 2: 2 times, 3: 3 or more times).
- **Use of Opioid Drugs by the Emergency Cardiology Team (NA_KB):** Indicates if opioid drugs were used by the Emergency Cardiology Team (0: No, 1: Yes).
- **Use of NSAIDs by the Emergency Cardiology Team (NOT_NA_KB):** Indicates if NSAIDs were used by the Emergency Cardiology Team (0: No, 1: Yes).
- **Use of Opioid Drugs in the ICU in the First Hours of the Hospital Period (NA_R_1_n):** Indicates if opioid drugs were used in the ICU during the first hours (0: No, 1: Once, 2: Twice, 3: Three times, 4: Four times).
- **Use of Liquid Nitrates in the ICU (NITR_S):** Indicates if liquid nitrates were used in the ICU (0: No, 1: Yes)
- **Use of Lidocaine by the Emergency Cardiology Team (LID_KB):** Indicates if lidocaine was used by the Emergency Cardiology Team (0: No, 1: Yes).
- **Use of Liquid Nitrates in the ICU (NITR_S):** Indicates if liquid nitrates were used in the ICU (0: No, 1: Yes)
- **Use of Opioid Drugs in the ICU in the First Hours of the Hospital Period (NA_R_1_n):** Indicates if opioid drugs were used in the ICU during the first hours (0: No, 1: Once, 2: Twice, 3: Three times, 4: Four times).
- **Use of Opioid Drugs in the ICU in the Second Day of the Hospital Period (NA_R_2_n):** Indicates if opioid drugs were used in the ICU on the second day (0: No, 1: Once, 2: Twice, 3: Three times).
- **Use of Opioid Drugs in the ICU in the Third Day of the Hospital Period (NA_R_3_n):** Indicates if opioid drugs were used in the ICU on the third day (0: No, 1: Once, 2: Twice).
- **Use of NSAIDs in the ICU in the First Hours of the Hospital Period (NOT_NA_1_n):** Indicates if NSAIDs were used in the ICU during the first hours (0: No, 1: Once, 2: Twice, 3: Three times, 4: Four or more times).
- **Use of NSAIDs in the ICU in the Second Day of the Hospital Period (NOT_NA_2_n):** Indicates if NSAIDs were used in the ICU on the second day (0: No, 1: Once, 2: Twice, 3: Three times). **Use of NSAIDs in the ICU in the Third Day of the Hospital Period (NOT_NA_3_n):** Indicates if NSAIDs were used in the ICU on the third day (0: No, 1: Once, 2: Twice).
- **Use of Lidocaine in the ICU (LID_S_n):** Indicates if lidocaine was used in the ICU (0: No, 1: Yes).
- **Use of Beta-Blockers in the ICU (B_BLOK_S_n):** Indicates if beta-blockers were used in the ICU (0: No, 1: Yes).
- **Use of Calcium Channel Blockers in the ICU (ANT_CA_S_n):** Indicates if calcium channel blockers were used in the ICU (0: No, 1: Yes).
- **Use of Anticoagulants (Heparin) in the ICU (GEPAR_S_n):** Indicates if anticoagulants (heparin) were used in the ICU (0: No, 1: Yes).
- **Use of Acetylsalicylic Acid in the ICU (ASP_S_n):** Indicates if acetylsalicylic acid was used in the ICU (0: No, 1: Yes).
- **Use of Ticlid in the ICU (TIKL_S_n):** Indicates if Ticlid was used in the ICU (0: No, 1: Yes).

- **Use of Trental in the ICU (TRENT_S_n):** Indicates if Trental was used in the ICU (0: No, 1: Yes).
- **Atrial Fibrillation (FIBR_PREDS):** Indicates if the patient had atrial fibrillation (0: No, 1: Yes).
- **Supraventricular Tachycardia (PREDS_TAH):** Indicates if the patient had supraventricular tachycardia (0: No, 1: Yes).
- **Ventricular Tachycardia (JELUD_TAH):** Indicates if the patient had ventricular tachycardia (0: No, 1: Yes).
- **Ventricular Fibrillation (FIBR_JELUD):** Indicates if the patient had ventricular fibrillation (0: No, 1: Yes).
- **Third-Degree AV Block (A_V_BLOK):** Indicates if the patient had a third-degree AV block (0: No, 1: Yes).
- **Pulmonary Edema (OTEK_LANC):** Indicates if the patient had pulmonary edema (0: No, 1: Yes).
- **Myocardial Rupture (RAZRIV):** Indicates if the patient had a myocardial rupture (0: No, 1: Yes).
- **Dressler Syndrome (DRESSLER):** Indicates if the patient had Dressler syndrome (0: No, 1: Yes).
- **Relapse of the Myocardial Infarction (REC_IM):** Indicates if the patient had a relapse of myocardial infarction (0: No, 1: Yes).
- **Post-Infarction Angina (P_IM_STEN):** Indicates if the patient had post-infarction angina (0: No, 1: Yes).
- **Lethal Outcome (Cause) (LET_IS):** Indicates the cause of lethal outcome (0: Unknown, 1: Cardiogenic shock, 2: Pulmonary edema, 3: Myocardial rupture, 4: Progress of congestive heart failure, 5: Thromboembolism, 6: Asystole, 7: Ventricular fibrillation)

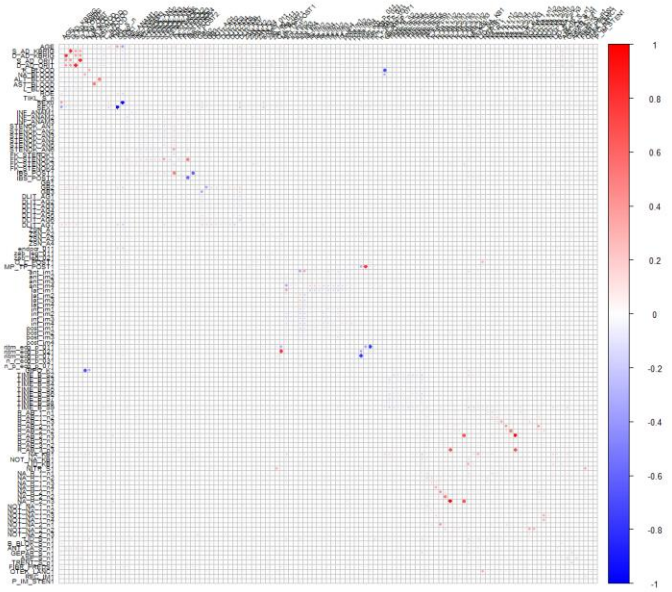
4. Preprocessing of the predictors:

Before splitting the data and building models, the predictors underwent preprocessing tailored to the specific modeling requirements. Initially, the dataset contained 15,974 missing values across the predictors. These were addressed using KNN imputation with K=5 Three predictors with over 95% missing data were excluded from the study, along with the variable ID, which was deemed irrelevant. Following these adjustments, the dataset included 119 variables: 11 continuous and 108 categorical. Among the categorical variables, 62 degenerate variables (those with insufficient variability) were identified and removed. Dummy variable encoding was then applied to the categorical variables, resulting in 115 total variables. A second inspection revealed additional degenerate variables, which were subsequently removed, leaving a final set of 81 predictors for analysis.

A. Correlation:

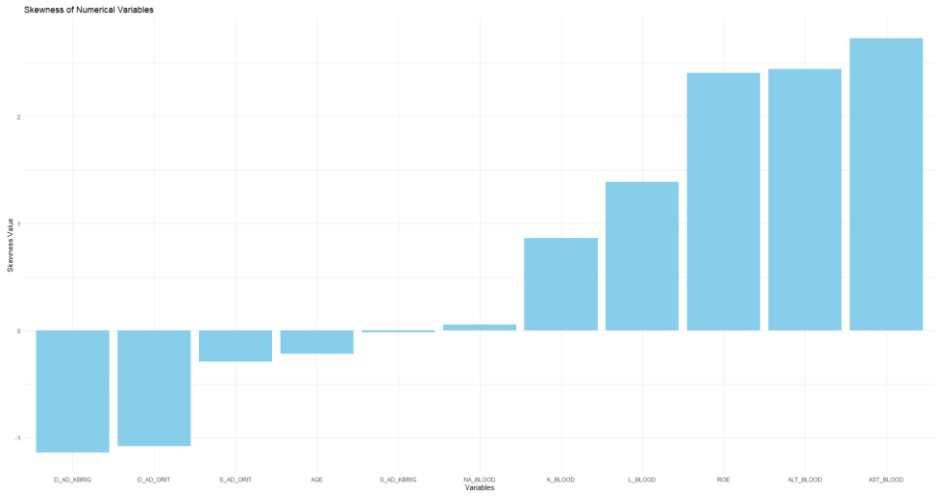
The remaining predictors were examined for high correlation, using a cut-off value of 75%. A correlation plot was generated, where the red gradient represents low to high positive correlations, while the value gradient represents low to high negative correlations. Six predictors were identified as highly correlated. These variables will be excluded during model building for models that are not designed to handle highly

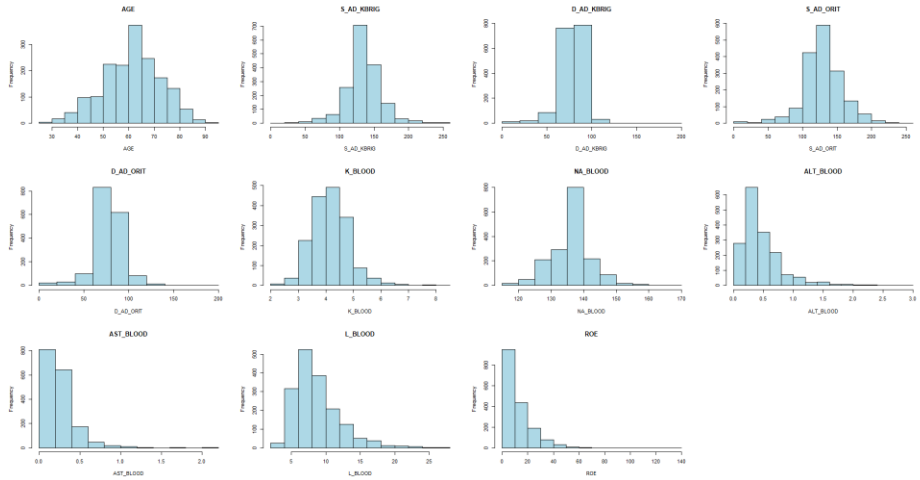
correlated predictors.

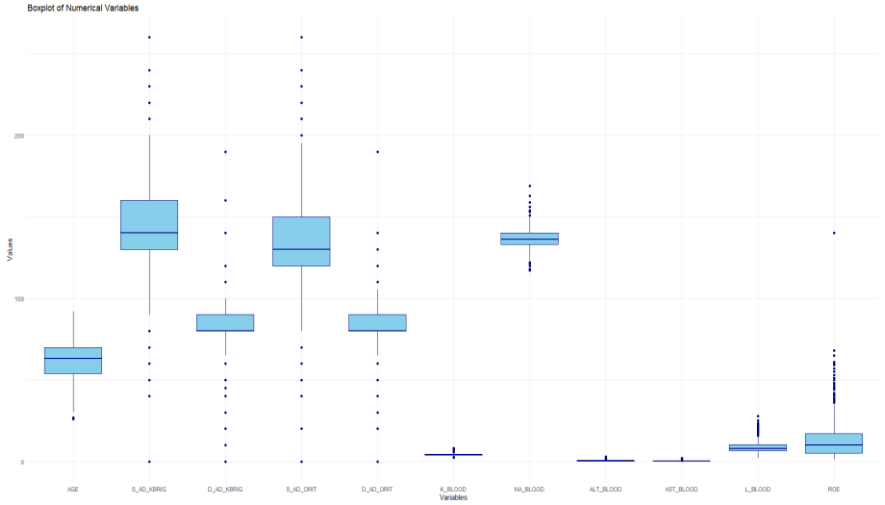


B. Transformations:

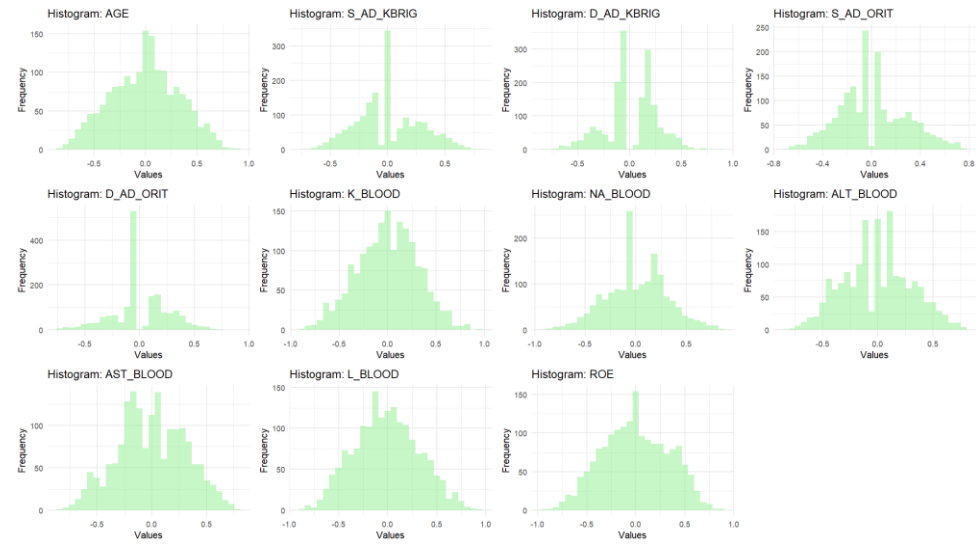
The distributions of 12 continuous variables are inspected. The histogram and boxplots are used to represent the distributions. The distributions of the continuous variables are generally right skewed, with most of the data concentrated in the lower to middle ranges, and a long tail extending towards higher values. A few variables show a more uniform distribution, but the majority exhibit peaks at certain values, indicating some clustering in specific ranges. There are also some outliers present in the upper ranges of several variables. Overall, these distributions suggest that most of the observations are concentrated in the lower to mid-range values, with fewer occurrences at the higher end. The skewness values of the numerical variables are also represented as follows:

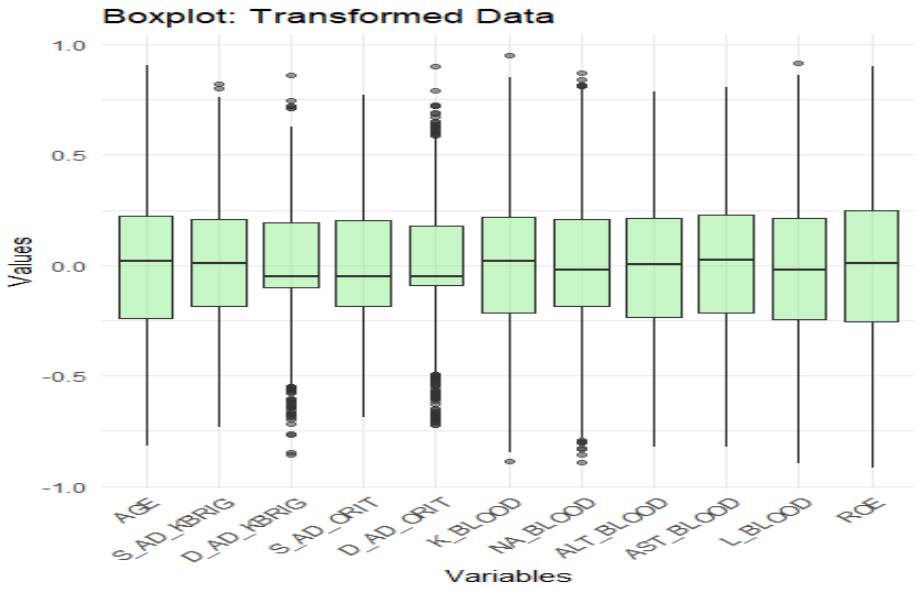






To address skewness and outliers, Box-Cox and Spatial Sign transformations were applied. The resulting histograms and boxplots indicate that the distributions have become more normalized, and the outliers have been effectively managed. However, since the Spatial Sign transformation alters the data scale, boxplots may not fully reflect the transformation. These adjusted predictors will be utilized in neural networks and other models that are sensitive to outliers





5.Splitting of the Data:

The target variable indicates whether a person is at risk of developing complications from chronic heart attack or not. Its distribution is imbalanced, as it is naturally expected that fewer individuals are at risk of complications compared to those who are not. This variable has two possible outcomes, making it binary. To address the imbalance, stratified splitting was used, ensuring proportional representation of both outcomes. The dataset was divided into a training set (80%) and a testing set (20%).

6.Model Building:

In this analysis, we will develop a comprehensive suite of predictive models to evaluate and compare their performance. These models are categorized as follows:

Linear Models:

- Logistic Regression
- Linear Discriminant Analysis (LDA)
- Partial Least Squares Discriminant Analysis (PLSDA)
- Penalized Models

Non-Linear Models:

- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Mixture Discriminant Analysis (MDA)
- Flexible Discriminant Analysis (FDA)
- Neural Network
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Naïve Bayes

Each model will be trained and evaluated using the dataset to assess its effectiveness in achieving the desired predictive outcomes. A 10-fold cross-validation method is applied for resampling and optimizing the model's hyperparameters. The evaluation metric selected for this study is the Kappa value, as it accounts for the imbalance in the target variable's distribution and provides an unbiased measure of model performance.

A. Linear Models:

The Linear models that have been built here involve a common preprocessing which is centering and scaling. Additionally, for the Logistic Regression and Linear Discriminant Analysis model, highly correlated features have been removed with a cut-off of 75%, whereas the Partial Least Squares Discriminant Analysis and Penalized Models has the capability to handle correlated predictors. The results are tabulated as follows:

Models	Best Tuning Parameter	Training Kappa	Testing Kappa
LR	No	0.29960	0.357
LDA	No	0.302857	0.3492
PLSDA	**ncomp=4**	**0.2976**	**0.3842**
Penalized	Alpha=1 & lambda=0.0003231	0.3478	0.2482

The Partial Least Squares Discriminant Analysis (PLSDA) and Logistic Regression models achieved the highest testing Kappa values, 0.3842 and 0.357, respectively. PLSDA has the added advantage of performing feature selection through its components. Given its higher Kappa value and feature selection capability, we consider PLSDA the best model. Logistic Regression, while effective, does not offer the same feature selection benefit. PLSDA's ability to handle complex datasets with many predictors makes it a better fit for this study. Therefore, we will proceed with PLSDA as the preferred model. .

B. Non - Linear Models:

The non-linear also undergo a common centering and scaling preprocessing, except for the Naive Bayes model. The high correlated predictors are removed for the neural networks, Quadratic Discriminant Analysis, Mixture Discriminant Analysis and Naive Bayes. The neural network model is tested for two cases: with spatial Sign and without spatial Sign. The difference in the kappa value with spatial Sign was slightly higher than the latter. The Box-Cox transformation is being applied on the predictors only in the Naive Bayes model. The results are tabulated as follows:

Models	Best Tuning Parameter	Training Kappa	Testing Kappa
QDA	No	0.2841	0.2551
RDA	Gamma=0.7 & Lambda=0.9	0.3457	0.3263
MDA	**Subclasses=4**	**0.30168**	**0.3514**
FDA	Degree=1 & nprune=20	0.3251	0.3127
NN	Size=3 & decay=0.00023	0.2784	0.2032
NN with Spatial Sign	Size = 5, decay = 0.04216	0.29043	0.2745
KNN	K=3	0.1018	0.1566
SVM	Sigma=0.0068 & c=0.25	0.2961	0.2773
NB	No	0.1803	0.1569

Mixture Discriminant Analysis (MDA) emerged as the best-performing non-linear model; however, its testing Kappa value was lower than that of Partial Least Squares Discriminant Analysis (PLSDA) and Logistic Regression, which were the best overall models due to their higher testing Kappa values.

Summary:

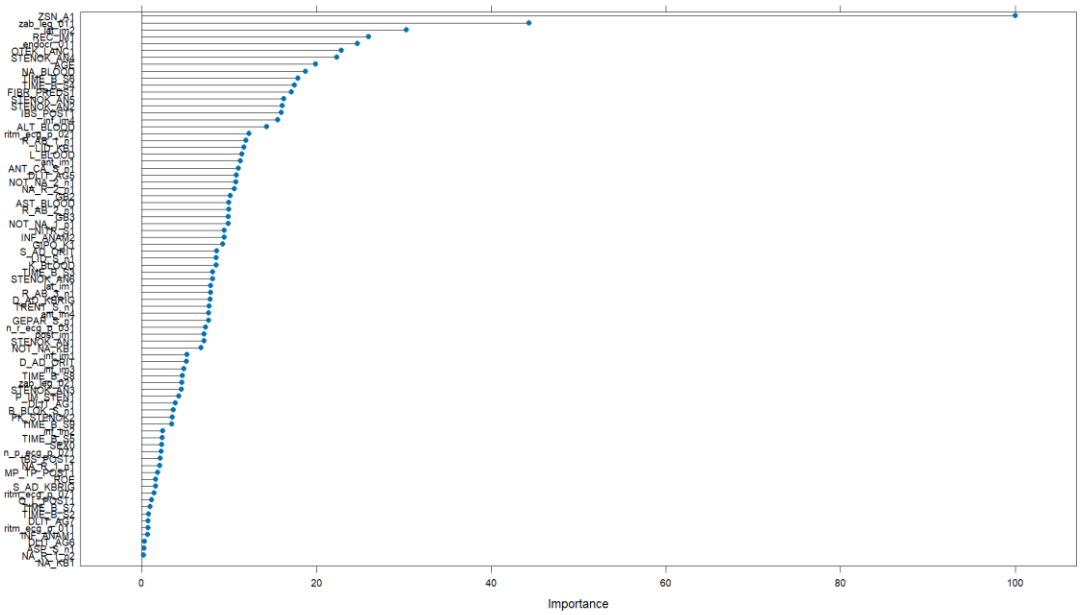
In summary, Partial Least Squares Discriminant Analysis (PLSDA) is the best-performing model, followed by Logistic Regression. While PLSDA achieved a Kappa value of 0.3842, its performance may be limited by the higher number of categorical variables, which can be harder to model and interpret. Despite this, PLSDA's ability to perform feature selection and handle complex data makes it the preferred model.

Logistic Regression, with a Kappa value of 0.357, performed significantly better than most non-linear models, likely because the classes are linearly separable. Although it lacks PLSDA's feature selection, its simplicity and effectiveness in linear problems make it a strong alternative. Overall, PLSDA is the top choice, with Logistic Regression as a reliable backup.

Appendix 1: Important Predictors of Top Two Models

Both models indicate that the ZSN_A is the most important predictor, which is apparent because it is closely related to cardiovascular risk.

Twenty most important variables for Logistic Regression:

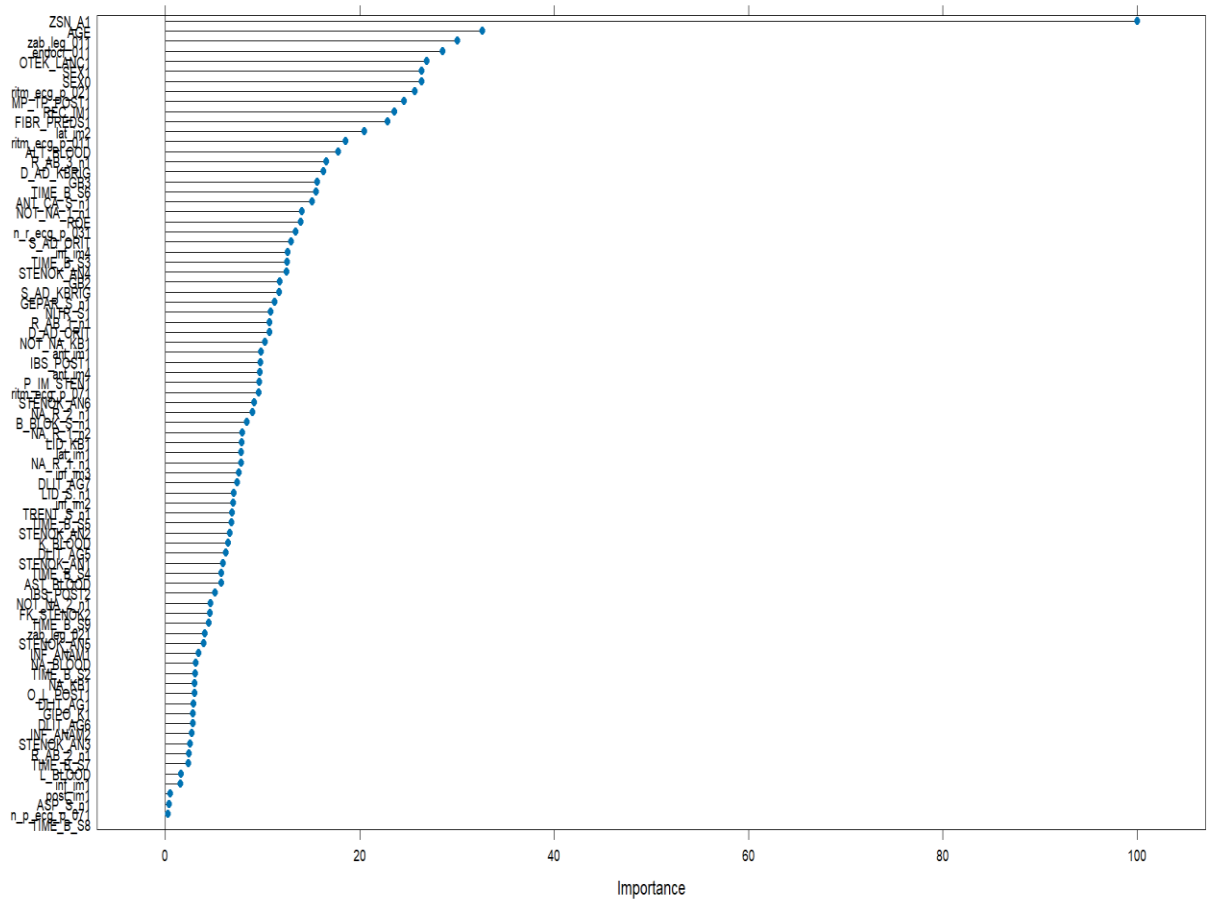


glm variable importance

only 20 most important variables shown (out of 80)

	Overall
ZSN_A1	100.00
zab_leg_011	44.32
lat_im2	30.28
REC_IM1	25.95
endocr_011	24.72
OTEK_LANC1	22.86
STENOK_AN4	22.35
AGE	19.92
NA_BLOOD	18.75
TIME_B_S6	17.90
TIME_B_S4	17.46
FIBR_PREDS1	17.12
STENOK_AN5	16.24
STENOK_AN2	16.12
IBS_POST1	15.98
inf_im4	15.59
ALT_BLOOD	14.28
ritm_ecg_p_021	12.26
R_AB_1_n1	11.91
LID_KB1	11.68

Top Twenty Important Predictors for PLSDA Model:



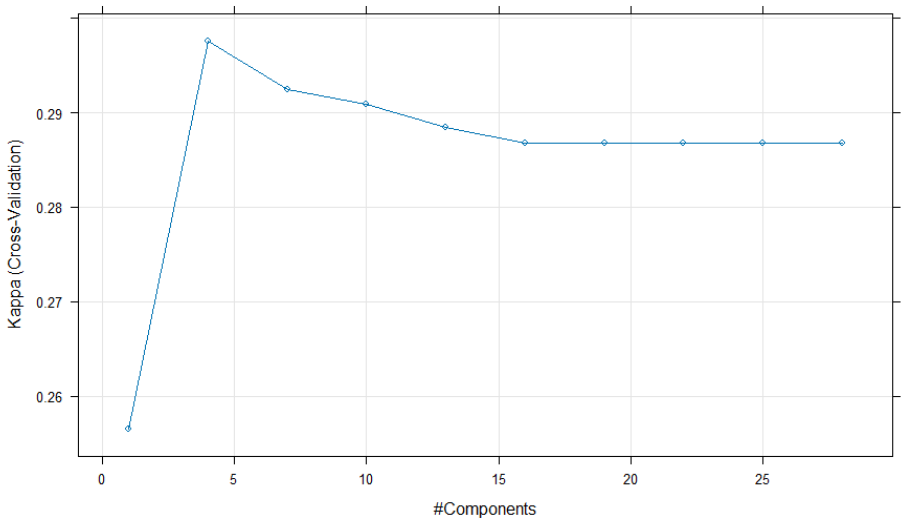
pls variable importance

only 20 most important variables shown (out of 81)

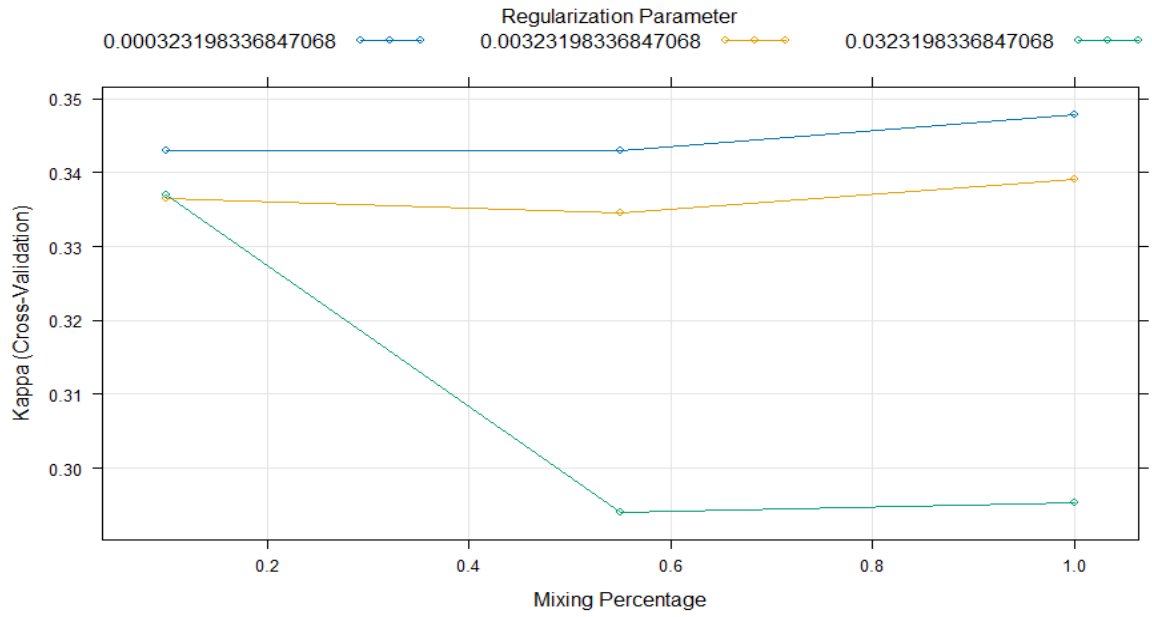
	Overall
ZSN_A1	100.00
AGE	32.65
zab_leg_011	30.04
endocr_011	28.55
OTEK_LANC1	26.91
SEX1	26.37
SEX0	26.37
ritm_ecg_p_021	25.68
MP_TP_POST1	24.58
REC_IM1	23.58
FIBR_PREDS1	22.87
lat_im2	20.50
ritm_ecg_p_011	18.55
ALT_BLOOD	17.78
R_AB_3_n1	16.58
D_AD_KBRIG	16.29
GB3	15.64
TIME_B_S6	15.51
ANT_CA_S_n1	15.13
NOT_NA_1_n1	14.08

Appendix 2: Linear Models:

Partial Least Square Discriminant Analysis: The optimal number of components that the model selected is to be 4 as shown below:

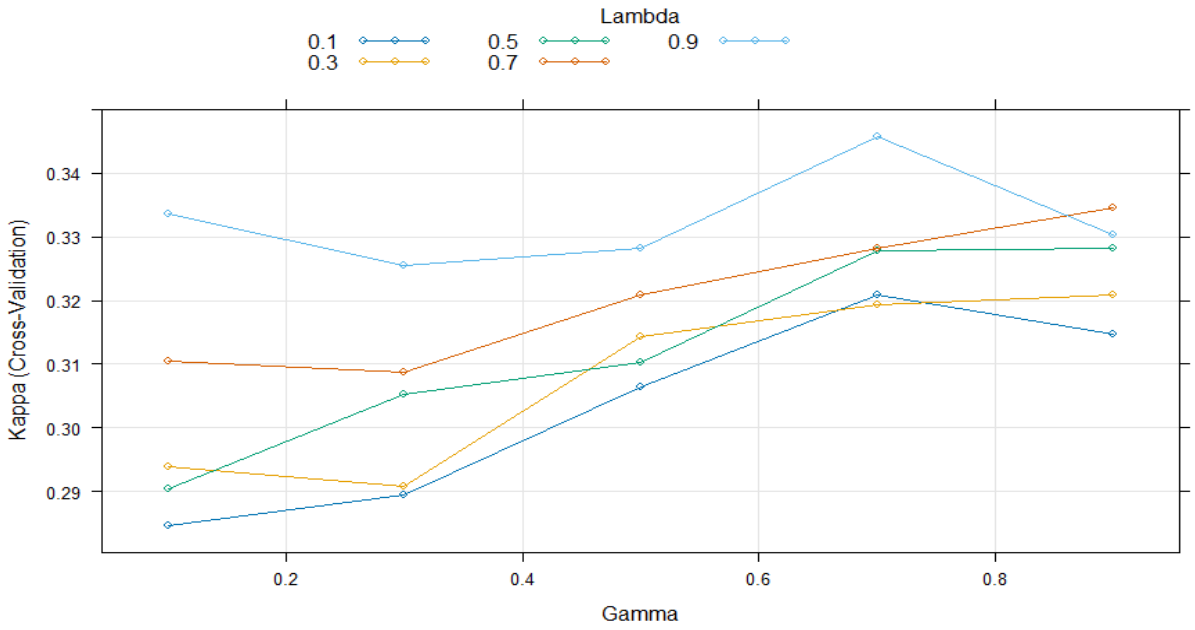


Penalized Models: The best hyperparameters for the penalized model is shown as follows:

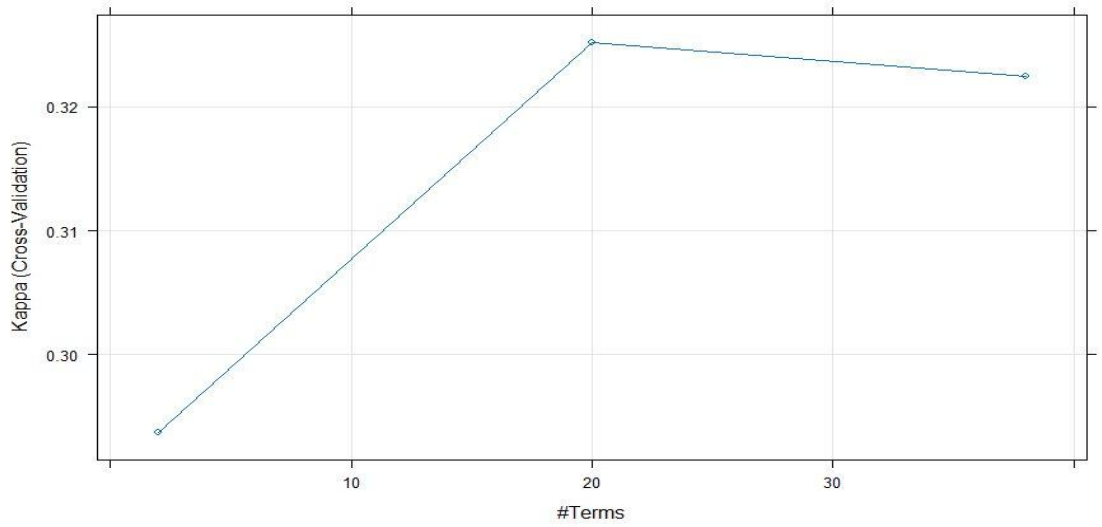


Appendix 3: Non-Linear Models:

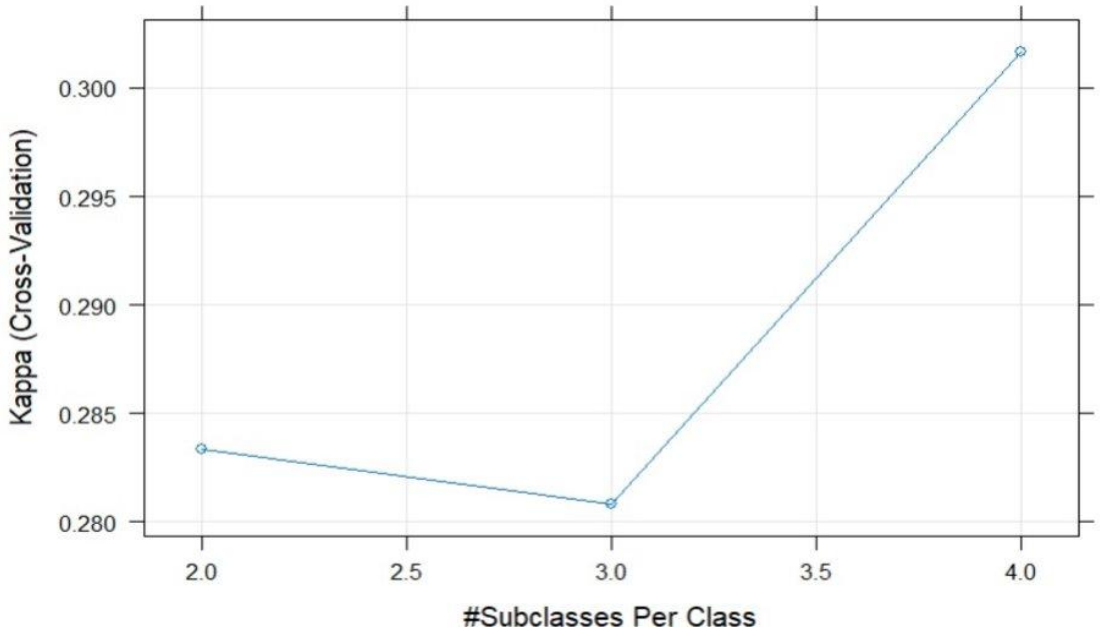
Regularized Discriminant Analysis (RDA): The best model was found to have a gamma value of 0.7 and lambda value of 0.9 which indicates that it is leaning towards QDA model. The hyperparameter plot is as follows:



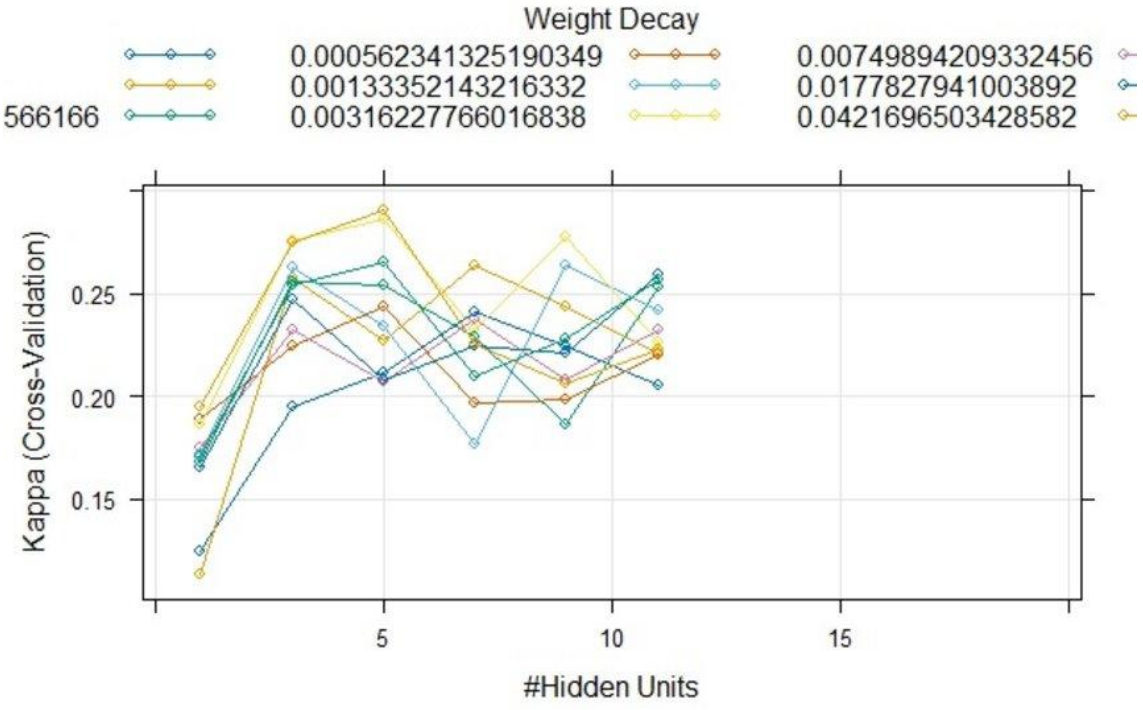
Flexible Discriminant Analysis: The flexible discriminant analysis is capable of feature selection, and it has the degree which represents the interactions between the predictors. The tuning plot is as follows:



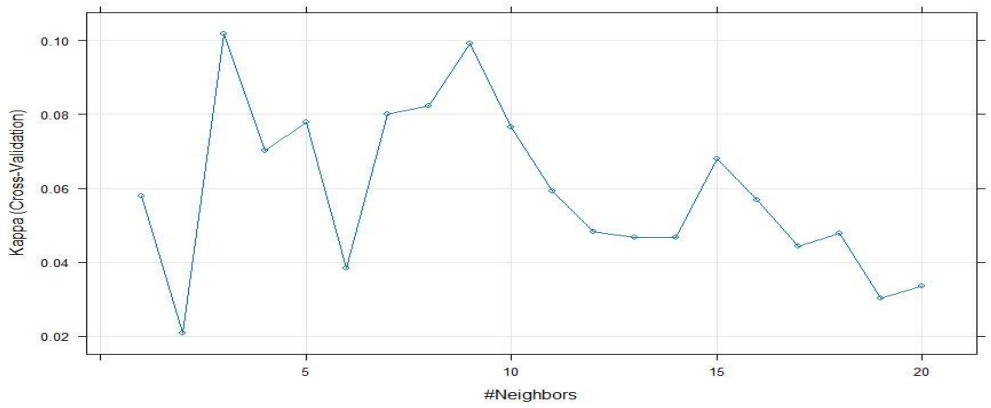
Mixture Discriminant Analysis: The mixture discriminant analysis achieved the best kappa values among all the non-linear models. The number of subclasses for the best model is four. The plot is as follows:



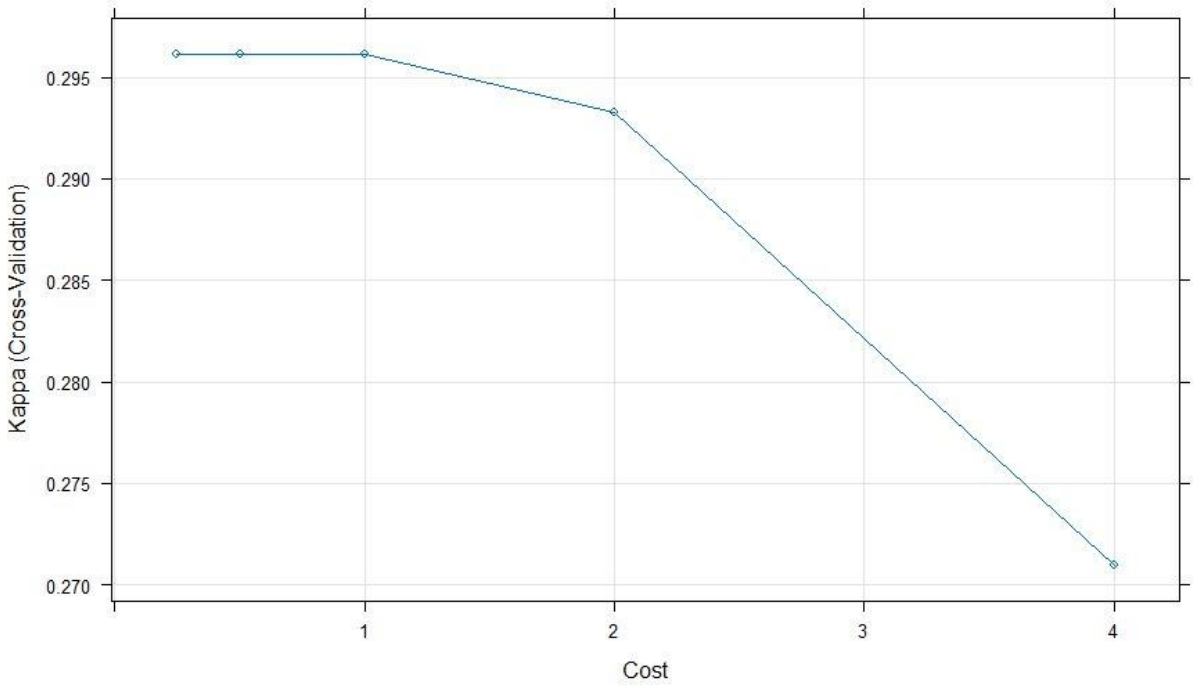
Neural Network: The neural network uses two hyperparameters which are decay and size. The hyperparameter plots for neural network after performing spatial Sign transformation is presented here as follows:



KNN: The KNN model has only a hyperparameter which is the number of K Neighbors. Here, a K value of 3 gives the best model as follows:



Support Vector Machine: The SVM model has two hyperparameters. The Sigma has been held constant at a value of 0.006848174 and the best Cost was 0.25. The plot is shown as follows:


Kappa (Cross-Validation)
0.295
0.290
0.285
0.280
0.275
0.270
1
2
3
4
Cost