Predicting myocardial infarction (MI) complications and death using tree-based methods

Emma Billmyer, Julia Kancans, Esteban Lemus Wirtz, Ben Pedersen

1 Abstract

Accurately predicting complications and death following myocardial infarction (MI) is crucial for improving patient outcomes. This study aims to identify important predictors and compare the performance of different tree-based machine learning models for predicting MI complications and death using data from 1,700 patients admitted to the hospital for MI in Siberia, Russia. Four classification methods were applied: single decision tree, pruned tree, random forest, and boosting (AdaBoost). The models were trained on 70% of the data and tested on the remaining 30%. Model performance was evaluated using misclassification error rates and area under the ROC curve (AUC). A sensitivity analysis explored different approaches for handling missing data in decision tree packages and randomforest. Key predictors of MI complication and death identified included age, chronic heart failure, use of nitrates, angina pectoris, lateral MI, and white blood cell count. Random forest demonstrated the best performance, with the lowest cross-validated misclassification error (33.1% for complications, 7.4% for death) and highest AUC (0.748 for complications, 0.842 for death). Overall, this study highlights the potential of tree-based machine learning models, particularly random forest, for predicting adverse outcomes in MI patients and identifies important risk factors to guide clinical decision-making.

2 Background

Myocardial infarction (MI; colloquially “heart attack”) happens due to a lack of or decreased blood flow to the myocardium. It is usually caused by an underlying coronary artery disease and can lead to sudden death and serious complications (Ojha, 2023). Globally, MI has a prevalence of 3.8% among individuals aged <60 years and 9.5% among individuals aged >60 years (Salari et al., 2023). About 20% of MI patients die within a year of an MI event. More than half of them happen within 30 days of MI, highlighting the importance of predicting high risk patients. Understanding predictors of MI complications upon hospital admission is important to help improve mortality and reduce complications from happening (Ye et al., 2020).

While there has been established research on predicting MI among high-income populations, there is a lack of research that focuses on middle- and low-income countries. Additionally, current literature follows more traditional statistical methods such as logistic regression. It’s not clear if more complex methods, such as classification trees, random forest, and boosting, would improve traditional prediction methods.

Using data from a middle-income country and more novel regression models, we aimed to answer the following questions:

1) What are important predictors of death and MI complications following hospitalization for MI in a middle-income country?

2) Which models (i.e., classification tree, pruned tree, random forest, boosting) best predict death and MI complications?
    a) As a sensitivity analysis, method performance with missing data was also assessed.

2.1 Expected Results

Our original hypothesis was that the random forest and AdaBoosting models would yield lower test errors than the single tree & pruned tree. However, it was important to consider single tree methods due to their easy interpretability.

3 Methods

3.1 Data

Data was collected from 1700 patients admitted for MI to the Krasnoyarsk Interdistrict Clinical Hospital in Siberia, Russia. In addition to death, eleven MI complications were recorded (i.e., atrial fibrillation, supraventricular tachycardia, ventricular tachycardia, ventricular fibrillation, third-degree AV block, pulmonary edema, myocardial rupture, Dressler syndrome, chronic heart failure, relapse of the myocardial infarction, and post-infarction angina). The timing of the events was not noted in the database description (e.g., during the admission, within a year of hospital admissions, etc.).

In total, there were 101 features in the data. Potential features included measures taken upon admission to the hospital: demographic information (age and sex), medical history, clinical findings, laboratory results, and treatment information. Variables with greater than 10% missingness were dropped from the analysis, which left 86 features (**Appendix Table A1**). A complete case analysis was used for the primary analysis.

3.2 Statistical Analysis

The two outcomes of interest are binary variables for any MI complication and death. Any MI complication was defined as having at least one of the eleven complications listed above. For the model building process, the data was split into a training and a test dataset using a 70/30 split (70% used for training and 30% for testing). The training dataset was used to train four classification models using the following methods: single classification tree, pruned tree, random forest and boosting (AdaBoost). Pruned trees were chosen based on the tree size that minimized the cross validated tree deviance. Random forest was done with 200 classification trees. Adaboost was also done with 200 trees and a shrinkage of 0.01.

Models were summarized visually through tree plots (for single and pruned trees) and variable importance plots (for random forest and boosting). For the full tree and pruned tree, all selected variables were reported. The top 10 variables based off of the Gini index for random forest and relative influence for AdaBoost were reported in order to understand variable importance and identify common predictors across methods. We decided to report the top 10 variables for these two methods as variables after that had marginal contributions (as measured by Gini index/relative influence) to the models.

Models were then assessed using the testing data. Given the importance of accurate death and MI complication prediction, we used the area under the curve (AUC) from a receiver operating characteristic (ROC) curve in addition to misclassification rates. While misclassification focuses on the accuracy using

a single probability threshold (0.5 in our case), ROC analyses provide a more nuanced assessment since they test all possible thresholds and provide the test's sensitivity and specificity under each one. This helps provide a balanced assessment that focuses on finding methods with a low overall error as well as the accurate prediction of positive events (i.e. sensitivity and specificity). ROC curves were generated to provide a visual comparison of model performance across methods. Based on Mandrekar (2010), the threshold for AUC acceptability will be 0.7, with values lower than that considered not acceptable.

Given the challenging nature of combining multiple single or pruned trees, the presented models were trained using a single 70/30 split. The disadvantage is that a single split is more prone to "unlucky samples," highlighting the need to benchmark the robustness of our initial misclassification and AUC estimates through a method that is less prone to unlucky splits. Therefore, we obtained cross-validated misclassification error and AUC values using a 10-fold cross validation. Doing this yielded more stable misclassification estimates and AUC values that helped us have a trustworthy comparison of the methods and benchmark the original performance measures from the 70/30 split.

3.3 Sensitivity Analysis

Dealing with missing data is a common problem in statistical analysis. While multiple imputation is a popular approach for handling missing data, it is not directly applicable to decision trees due to the inability to combine multiple decision trees coherently. Common solutions to missing data within classification trees include complete case analysis, where only observations with complete data are considered, and surrogate splits, which use alternative variables to split the data when the primary variable is missing. While our primary analyses were conducted using a complete case approach, we explored how the results would change if we employed methods to handle missing data, such as surrogate splits. For the decision trees, we show how two common R packages, tree() and rpart() handle missing data in decision trees and compare their results. Additionally, we explore the performance of random forest with two built in missing data options, usesurrogate =1 and usesurrogate = 2. A new 70/30 train-test split of the data that included observations with missing values was used in this analysis.

4 Results

4.1 Primary Analysis: Any Complication

Two variables were selected as common predictors of any MI complication in all four models: age and chronic heart failure (**Table 1**). Three additional variables were chosen by three methods including nitrates, angina pectoris and lateral MI (left ventricle). Overall, the test error rates for any complication were high across all the methods, all over 32%. For trees and pruned trees, the train error was also high (30.23% and 45.41%, respectively). The full tree is medium sized with eight terminal nodes (**Figure 1**). The pruned tree is smaller with only four terminal nodes, which makes its interpretation easier than the full tree. For example, based on the pruned tree, someone with stage I heart failure, who was given no liquid nitrates and is younger than 24.5 would be predicted to have no complications (**Figure 2**).

The 10 most important variables for random forest according to the mean gini index are similar to the top 10 variables according to the mean decrease in accuracy, but in slightly different orders (**Figure 3**). Boosting had approximately five variables it identified as having large relative influence scores (**Figure**

4). Age and presence of chronic heart failure had the highest relative influence scores (>30), followed by exertional angina pectoris, use of liquid nitrates, lateral MI (left ventricular).

4.2 Primary Analysis: Death

Three variables were selected as common predictors of death across all four methods: use of liquid nitrates, age, and lateral MI (left ventricle) (**Table 2**). An additional four variables were chosen by three out of four methods (white blood cell count, anterior MI (left ventricular), quantity of MIs, and functional class of angina pectoris). The test error rates were fairly low with misclassification rates of 9.91% for the tree and 7.12% for the other three methods (pruned tree, random forest, and boosting). The full tree for death is large with 23 total terminal nodes (**Figure 5**). It has a good mix of predicting both classes, but it is difficult to interpret due to its large size. While the pruned tree is smaller and easier to interpret with four terminal nodes, all of the terminal nodes in the pruned tree predict 'no death' (**Figure 6**).

Similar to any complication, the 10 most important variables in predicting death from the random forest according to the mean gini index are similar to the top 10 variables according to the mean decrease in accuracy but in slightly different orders (**Figure 7**). For death, boosting identified 15 variables with a relative influence greater than 0. The first six variables visually had significantly larger relative influence than the other variables (**Figure 8**). The three most important variables were use of liquid nitrates, lateral MI (left ventricle), and white blood cell count, followed by functional class of angina pectoris, quantity of MI, and anterior MI (left ventricle).

4.3 Cross-Validated Misclassification and AUC

Overall, the misclassification error and AUC results from the 10-fold cross validation were consistent with the ones obtained using a single 70/30 split (**Table 3**). For any complication, the cross validated (CV) misclassification rate ranged from 30-40% (except for AdaBoost which had ~50%), which follows the 70/30 split results closely. Similarly, the CV AUC ranges between 0.65-0.75, which follows the 70/30 split results. In terms of model comparison, random forest had the best performance in predicting any complication with the lowest CV misclassification (0.331) and the highest CV AUC (0.748), consistent with the single split results. The cross-validated performance measures for predicting death were also consistent with the initial 70/30 split results. The CV misclassification of death ranged from 7-8% and the CV AUC ranged from 0.65-0.84. Both performance measures follow a similar pattern as the 70/30 split results. Random forest had the best performance with the lowest CV misclassification (0.074) and the highest AUC (0.842), consistent with the single split results.

An interesting result to highlight is that random forest and AdaBoost yielded the same CV misclassification error for death (0.074) but different AUC values (0.842 and 0.726 respectively). The reason for this discrepancy is that both models predicted low probabilities of death, and when using a threshold of 0.5 to classify instances as death or non-death, both models made the same predictions. However, ROC curves test all possible threshold values and random forest tends to predict higher death probabilities than AdaBoost. When ROC curves start to test lower threshold values, random forest will start predicting more instances of deaths than AdaBoost, leading to better sensitivity (**Figure 9**). This particular case highlights the importance of choosing the right performance measures instead of solely focusing on misclassification.

4.4 Sensitivity Analysis

The tree package, used in our primary analysis for the decision tree, handles missing data by pushing missing values down the tree as far as possible. Observations are retained in the data until the tree splits on a variable for which that observation has a missing value. At that point, the observation is dropped from the analysis. The package does not provide options for using surrogate splits. **Table 4** presents the misclassification rates for any complication and death across the different missing data methods, separately. For death and any complication, the misclassification rate using the tree() method was 16.67% and 37.64%, respectively (**Table 4)**.

The Rpart package offers an alternative approach to handling missing values in decision trees through the use of surrogate variables. When an observation is missing the variable at a split, the algorithm looks for a surrogate variable. Surrogate splits are identified as splitting variables that lead to approximately the same division of cases going left/right as the original split. The Rpart package provides two options for handling missing data: usesurrogate = 1 and usesurrogate = 2. With usesurrogate = 1, if no good surrogates exist, the observation is not split. For usesurrogate = 2, if no good surrogates exist, the observation is sent in the majority direction. The option usesurrogate = 2 resulted in slightly lower misclassification rates than usesurrogate = 1 for both outcomes (any complication: 37.25% vs 35.10%; death: 15.49% vs 14.90%).

We also explored how random forest methods handle missing data. The randomforest package has two built-in options to deal with missing data. The first option is median (or mean) imputation. This option, called na.roughfix, is fast and replaces the missing values with the mean for continuous variables or median for categorical variables. The second option is a proximity base measure, which first imputes missing values using the na.roughfix option and then updates the missing values using the proximity matrix as a weight. Misclassification rates were similar across the two random forest options (any complication: 28.44% vs 27.72%; death: 8.56% vs 9.48%).

5 Discussion

5.1 Primary Analysis

Our paper shows that tree-based methods could help predict MI related complications and death, as most models have misclassification errors and AUCs below 0.5 (i.e. they perform better than guessing). However, many of the models did not reach the 0.7 AUC threshold considered as acceptable. Overall for both any complication and death, random forest performed best across all metrics: test error, CV error and AUC. Death and any complication had many common predictors such as use of liquid nitrates (dilate blood vessels to take pressure off of the heart), age, and MI that occur in the left ventricle (chamber of the heart that pumps oxygenated blood out to the rest of the body). This makes sense in context as all of the presence of any of these predictors may indicate a more ill patient. The 7.12% test error for pruned test, random forest and adaboost indicate that they may be correctly predicting all the non-deaths but not the deaths since the test set has 7.12% deaths. Although we had initially proposed to include trees for their interpretability, the results were mixed. The full tree from death was too large to be readable, and the pruned tree predicted that everyone would live, which is not helpful for purposes of classification. The tree from any complication was slightly more readable but still slightly difficult due to the different levels of some of the ordinal variables. The pruned tree only predicts no complication for 1 of the 4 nodes.

There are multiple possible reasons for the high error rates from any complication variable. It may be due to combining multiple outcomes into one composite outcome which may introduce more noise into the data. Similarly, another explanation may be that many of the complications did not have much of a differentiation (90/10) split so that may also make it difficult for the algorithm to distinguish between the two classes. The boosting function also gave warnings that a few of the variables have no variation which may also lead to problems with classification. In particular, adaboost had much higher errors than we expected for any complication. Boosting may have overfit the data more than any other method. Adaboost is also very sensitive to noisy data and outliers.

5.2 Sensitivity Analysis: Missing data

The results within the sensitivity analysis for missing data were mixed. For any complication, the methods that addressed missing data had lower misclassification rates than the complete case analysis. However, the same methods performed worse than the complete case analysis for death, particularly with the decision tree. This may be due to the differences between the training and testing samples used in the complete case and missing data assessment. After excluding data with missing observations, the rates of death dropped substantially from 15% to 7.4%, and any complication decreased slightly from 55% to 54%. It could be that the sample of individuals with missing data are systematically different from those with non-missing data, which is causing the decision trees to perform differently between the two methods. Overall, this is not an exhaustive list of handling missing data within decision trees and random forests. This remains a research topic of interest, with emerging new methodologies.

5.3 Limitations and Future Work

It is important to point out certain limitations in this analysis and next steps. We believe that instead of benchmarking our results against guessing, it should be benchmarked against the commonly used logistic regressions in the existing literature. Additionally, while our data included a comprehensive list of features and information, the data did not include a variable specifying the timing of the observed outcome or hospitalization length. For any MI complication, creating a composite outcome allowed us to have a balanced dataset, but it may have added too much heterogeneity, preventing the models from detecting single complication specific trends of high importance. We hypothesize this is the reason our predictions of any complication had a fairly high misclassification error and AUCs that are below the 0.7 acceptable threshold (Mandrekar, 2010). For death, we encountered unbalanced data, which led to some models having no sensitivity to predict death (i.e.always predicting survival). In this context, this is unacceptable as it means resources and care will not be allocated to patients with high death risk.

As for next steps, we propose to further investigate tree methods for the prediction of MI death and complications using other datasets with more balance, as well as testing other tree-based methods that are better suited to deal with unbalanced data. We propose to test 1) ensemble random forests (ERFs), which have previously outperformed random forest in rare event contexts (Siders et al., 2020); and 2) other Boosting methods such as CatBoost. We see the value in addressing missing data within these classification methods and suggest that future work considers methods to thoughtfully handle missing data in order to retain as much information as possible.

References

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315-1316.

Mechanic, O. J., Gavin, M., & Grossman, S. A. (2017). Acute myocardial infarction.

Ojha, N., & Dhamoon, A. S. (2022). Myocardial infarction. In *StatPearls [Internet]*. StatPearls Publishing.

Salari, N., Morddarvanjoghi, F., Abdolmaleki, A., Rasoulpoor, S., Khaleghi, A. A., Hezarkhani, L. A., ... & Mohammadi, M. (2023). The global prevalence of myocardial infarction: a systematic review and meta-analysis. *BMC Cardiovascular Disorders*, *23*(1), 206.

Siders, Z. A., Ducharme-Barth, N. D., Carvalho, F., Kobayashi, D., Martin, S., & Raynor, J. (2020). Ensemble Random Forests as a tool for modeling rare occurrences. Endang Spec Res 43: 183–197.

Ye, Q., Zhang, J., & Ma, L. (2020). Predictors of all-cause 1-year mortality in myocardial infarction patients. *Medicine*, *99*(29), e21288.

Tables and Figures

Table 1: Important variables and test error for predicting any MI complication across classification methods

Variables	Description	Tree	Pruned Tree	Random Forest	ADABoost
	Test Error	**38.39 %**	**40.56 %**	**32.82 %**	**51.39 %**
AGE	Age	3	3	1	1
ZSN_A	Presence of chronic heart failure in the anamnesis	1	1	6	2
NITR_S	Use of liquid nitrates in the ICU	2	2		4
STENOK_AN	Exertional angina pectoris in the anamnesis	5		3	3
lat_im	Presence of a lateral myocardial infarction (left ventricular)	4		5	5
ant_im	Presence of an anterior myocardial infarction (left ventricular)			7	6
zab_leg_01	Chronic bronchitis in the anamnesis	6			8
inf_im	Presence of an inferior myocardial infarction (left ventricular)			8	7
INF_ANAM	Quantity of myocardial infarctions in the anamnesis			10	10
L_BLOOD	White blood cell count (billions per liter)			2	
TIME_B_S	Time elapsed from the beginning of the attack of CHD to the hospital			4	
GB	Presence of essential hypertension			9	
n_p_ecg_p_12	Complete RBBB on ECG at the time of admission to hospital				9

Note:

Numbers indicate how important variables were. Order does not matter for tree and pruned tree. Variables not selected were (additional description found in the appendix): SEX, FK_STENOK, IBS_POST, SIM_GIPERT, nr11, nr01, nr02, nr03, nr04, nr07, nr08, np01, np04, np05, np07, np08, np09, np10, endocr_01, endocr_02, endocr_03, zab_leg_02, zab_leg_03, zab_leg_04, zab_leg_06, O_L_POST, K_SH_POST, MP_TP_POST, SVT_POST, GT_POST, FIB_G_POST, post_im, IM_PG_P, ritm_ecg_p_01, ritm_ecg_p_02, ritm_ecg_p_04, ritm_ecg_p_06, ritm_ecg_p_07, ritm_ecg_p_08, n_r_ecg_p_01, n_r_ecg_p_02, n_r_ecg_p_03, n_r_ecg_p_04, n_r_ecg_p_05, n_r_ecg_p_06, n_r_ecg_p_08, n_r_ecg_p_09, n_r_ecg_p_10, n_p_ecg_p_01, n_p_ecg_p_03, n_p_ecg_p_04, n_p_ecg_p_05, n_p_ecg_p_06, n_p_ecg_p_07, n_p_ecg_p_08, n_p_ecg_p_09, n_p_ecg_p_10, n_p_ecg_p_11, fibr_ter_01, fibr_ter_02, fibr_ter_03, fibr_ter_05, fibr_ter_06, fibr_ter_07, fibr_ter_08, NA_R_3_n, LID_S_n, B_BLOK_S_n, ANT_CA_S_n, GEPAR_S_n, ASP_S_n, TIKL_S_n, TRENT_S_n

Table 2: Important variables and test error for predicting death after MI across classification methods

Predictors selected for MI death across 4 methods

Variables	Description	Tree	Pruned Tree	Random Forest	ADABoost
	Test Error	**9.91 %**	**7.12 %**	**7.12 %**	**7.12 %**
lat_im	Presence of a lateral myocardial infarction (left ventricular)	2	2	2	2
NITR_S	Use of liquid nitrates in the ICU	1	1	10	1
AGE	Age	13	3	1	10
L_BLOOD	White blood cell count (billions per liter)	3		3	3
ant_im	Presence of an anterior myocardial infarction (left ventricular)	5		6	6
INF_ANAM	Quantity of myocardial infarctions in the anamnesis	8		8	5
FK_STENOK	Functional class of angina pectoris in the last year	20		9	4
STENOK_AN	Exertional angina pectoris in the anamnesis	7		4	
inf_im	Presence of an inferior myocardial infarction (left ventricular)	4		7	
TIME_B_S	Time elapsed from the beginning of the attack of CHD to the hospital	14		5	
n_p_ecg_p_12	Complete RBBB on ECG at the time of admission to hospital	16			9
np08	Complete LBBB in the anamnesis	6			
ZSN_A	Presence of chronic heart failure in the anamnesis				7
endocr_01	Diabetes mellitus in the anamnesis				8
IM_PG_P	Presence of a right ventricular myocardial infarction	9			
n_p_ecg_p_08	LBBB (posterior branch) on ECG at the time of admission to hospital	10			
LID_S_n	Use of lidocaine in the ICU	11			
zab_leg_02	Obstructive chronic bronchitis in the anamnesis	12			
post_im	Presence of a posterior myocardial infarction (left ventricular)	15			
IBS_POST	Coronary heart disease in recent weeks, days before admission to hospital	17			
ritm_ecg_p_01	ECG rhythm at the time of admission to hospital – sinus (with a heart rate 60-90)	18			
GB	Presence of essential hypertension	19			
n_r_ecg_p_04	Frequent premature ventricular contractions on ECG at the time of admission to hospital	21			
n_r_ecg_p_03	Premature ventricular contractions on ECG at the time of admission to hospital	22			
MP_TP_POST	Paroxysms of atrial fibrillation at the time of admission to intensive care unit	23			

Note:

Numbers indicate how important variables were. Order does not matter for tree and pruned tree. Variables not selected were (additional description found in the appendix): SEX, SIM_GIPERT, nr11, nr01, nr02, nr03, nr04, nr07, nr08, np01, np04, np05, np07, np09, np10, endocr_02, endocr_03, zab_leg_01, zab_leg_03, zab_leg_04, zab_leg_06, O_L_POST, K_SH_POST, SVT_POST, GT_POST, FIB_G_POST, ritm_ecg_p_02, ritm_ecg_p_04, ritm_ecg_p_06, ritm_ecg_p_07, ritm_ecg_p_08, n_r_ecg_p_01, n_r_ecg_p_02, n_r_ecg_p_05, n_r_ecg_p_06, n_r_ecg_p_08, n_r_ecg_p_09, n_r_ecg_p_10, n_p_ecg_p_01, n_p_ecg_p_03, n_p_ecg_p_04, n_p_ecg_p_05, n_p_ecg_p_06, n_p_ecg_p_07, n_p_ecg_p_09, n_p_ecg_p_10, n_p_ecg_p_11, fibr_ter_01, fibr_ter_02, fibr_ter_03, fibr_ter_05, fibr_ter_06, fibr_ter_07, fibr_ter_08, NA_R_3_n, B_BLOK_S_n, ANT_CA_S_n, GEPAR_S_n, ASP_S_n, TIKL_S_n, TRENT_S_n, death

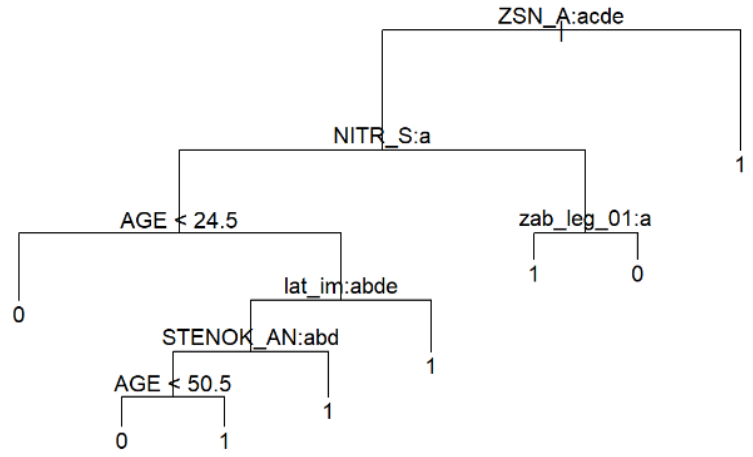
Table 3. 10-Fold cross-validated misclassification error and AUC for predicting any MI complication and death

	Tree	Pruned Tree	Random Forest	AdaBoost
Any complication				
Error	0.386	0.392	0.331	0.536
AUC	0.675	0.661	0.748	0.693
Death				
Error	0.088	0.075	0.074	0.074
AUC	0.646	0.655	0.842	0.726

Table 4. Misclassification rates across different missing data handling methods in decision trees and random forest

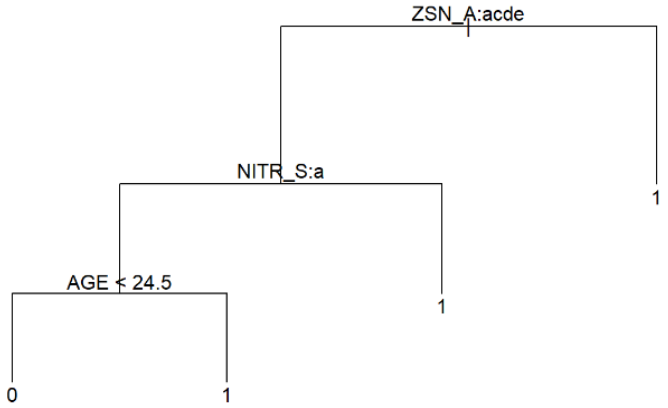
	Any complication	Death
Decision Tree		
Complete Case Analysis	38.39%	9.91%
Tree()	37.64%	16.67%
Rpart() with surrogate splits 1	37.25%	15.49%
Rpart() with surrogate splits 2	35.10%	14.90%
Random Forest		
Complete Case Analysis	32.20%	7.12%
Na.roughfix	28.44%	8.56%
proximity based measure	27.72%	9.48%

Figure 1: Full decision tree model for predicting any MI complication



Note: Terminal nodes ending in 0 predict no complication, and terminal nodes ending in 1 predict any complication

Figure 2: Pruned decision tree for predicting any MI complication



Note: Terminal nodes ending in 0 predict no complication, and terminal nodes ending in 1 predict any complication

Figure 3: Top 20 most important predictors of any MI complication from the random forest model

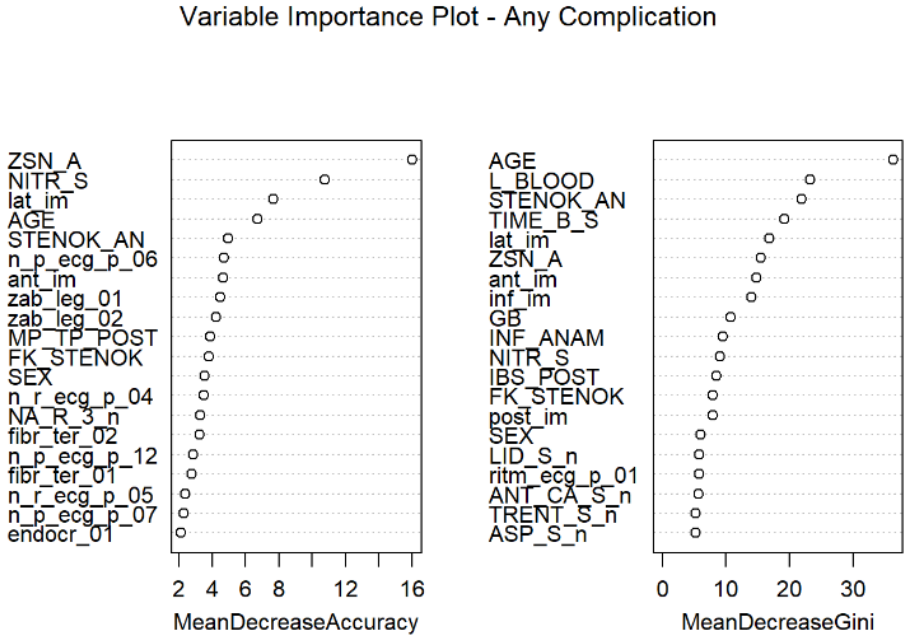
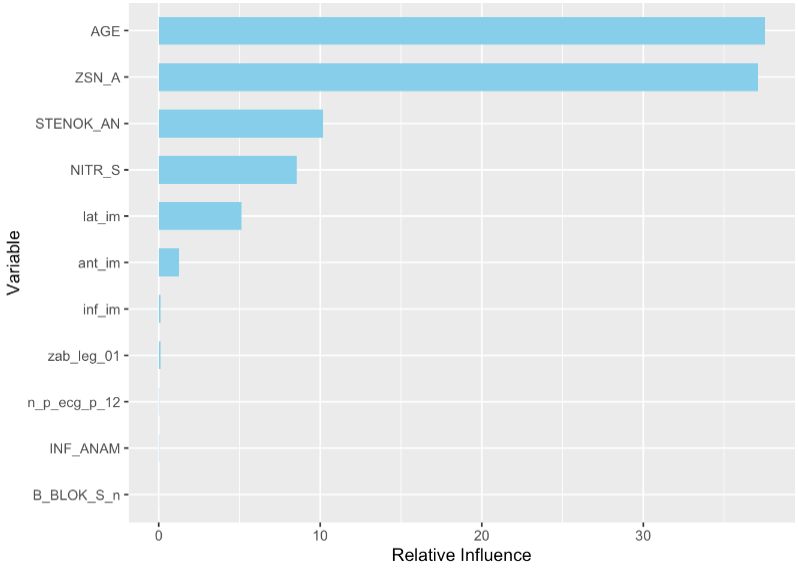
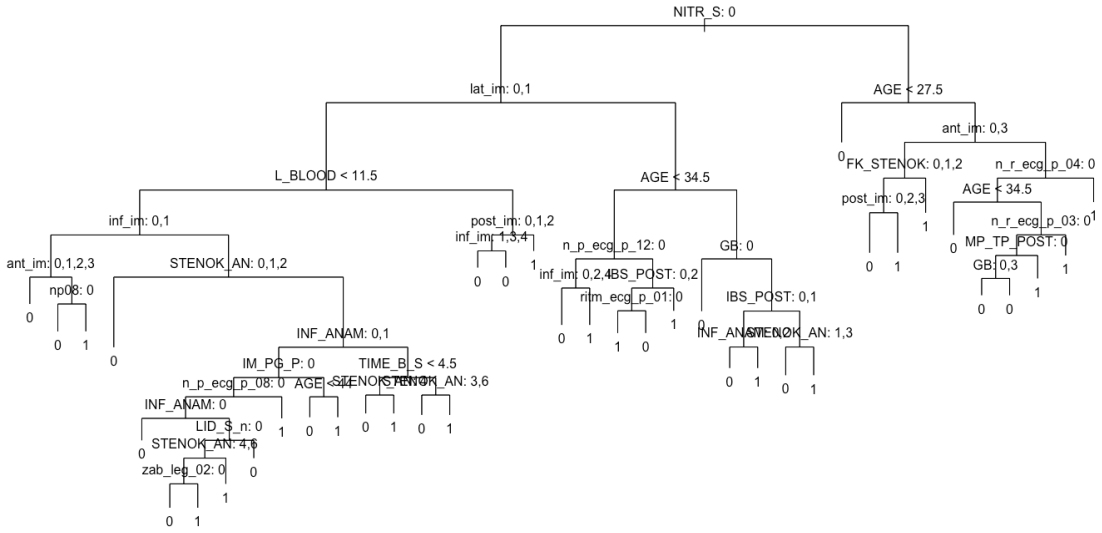


Figure 4: Top 11 most important variables for predicting any MI complication based on AdaBoost relative influence



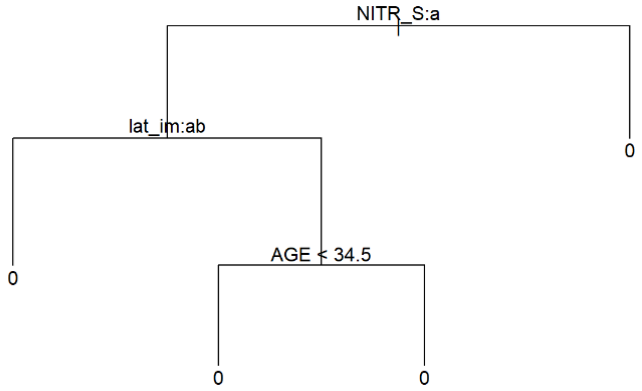
Note: Only variables with a relative influence greater than 0 are reported.

Figure 5: Full decision tree model for predicting death after myocardial infarction



Note: Terminal nodes ending in 0 predict no death, and terminal nodes ending in 1 predict death

Figure 6: Pruned decision tree for predicting death after MI



Note: Terminal nodes ending in 0 predict no death, and terminal nodes ending in 1 predict death. All notes in the pruned tree predict no death.

Figure 7: Top 20 most important predictors of death after MI from the random forest model

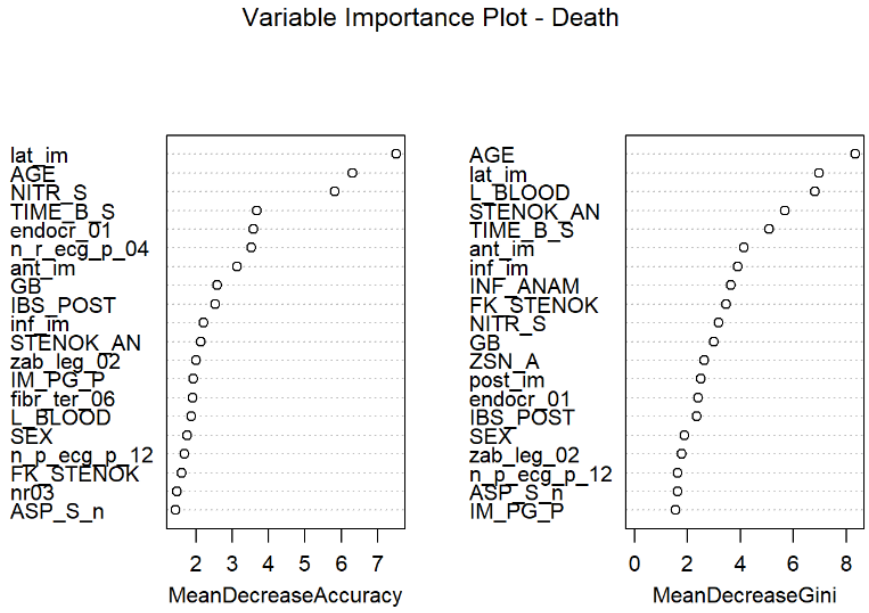
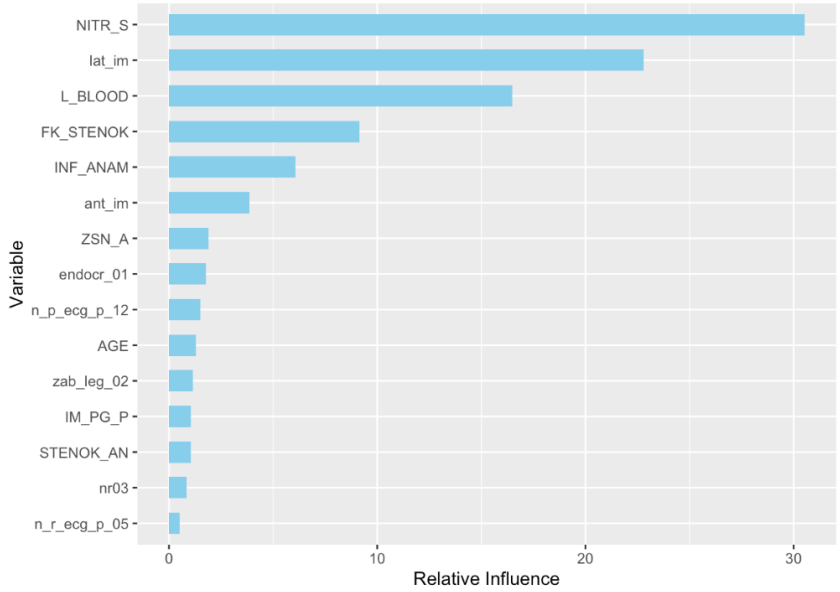
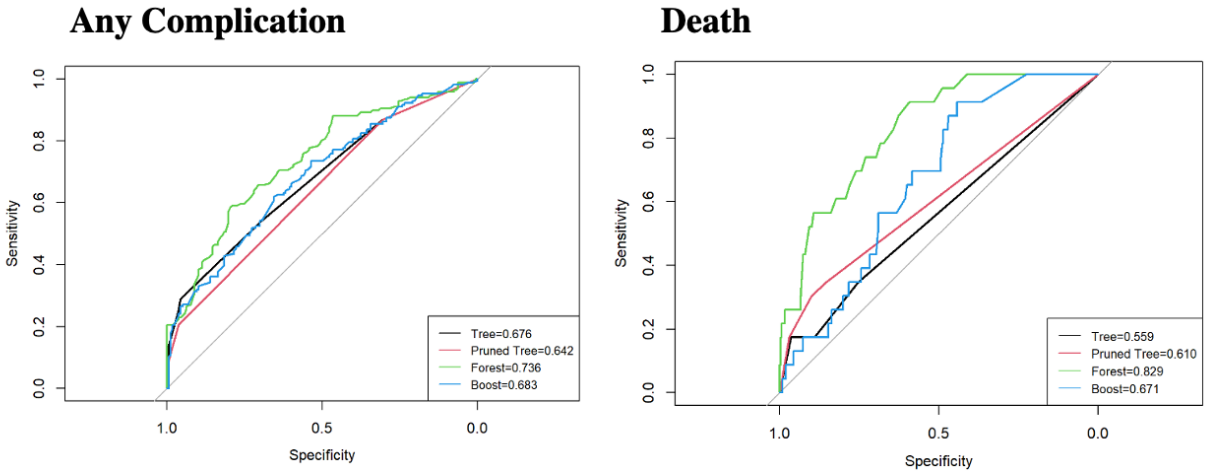


Figure 8: Top 15 most important variables for predicting death after MI based on AdaBoost relative influence



Note: Only variables with a relative influence greater than 0 are reported.

Figure 9: Receiver operating characteristic (ROC) curves comparing classification methods for predicting any complication and death



Appendix

Table A1: Data dictionary for the 86 features included in the classification analyses

Variables	Description
AGE	Age
SEX	Gender
INF_ANAM	Quantity of myocardial infarctions in the anamnesis
STENOK_AN	Exertional angina pectoris in the anamnesis
FK_STENOK	Functional class of angina pectoris in the last year
IBS_POST	Coronary heart disease in recent weeks, days before admission to hospital
GB	Presence of essential hypertension
SIM_GIPERT	Symptomatic hypertension
ZSN_A	Presence of chronic heart failure in the anamnesis
nr11	Observing of arrhythmia in the anamnesis
nr01	Premature atrial contractions in the anamnesis
nr02	Premature ventricular contractions in the anamnesis
nr03	Paroxysms of atrial fibrillation in the anamnesis
nr04	Persistent form of atrial fibrillation in the anamnesis
nr07	Ventricular fibrillation in the anamnesis
nr08	Ventricular paroxysmal tachycardia in the anamnesis
np01	First-degree AV block in the anamnesis
np04	Third-degree AV block in the anamnesis
np05	LBBB (anterior branch) in the anamnesis
np07	Incomplete LBBB in the anamnesis
np08	Complete LBBB in the anamnesis
np09	Incomplete RBBB in the anamnesis
np10	Complete RBBB in the anamnesis
endocr_01	Diabetes mellitus in the anamnesis
endocr_02	Obesity in the anamnesis
endocr_03	Thyrotoxicosis in the anamnesis
zab_leg_01	Chronic bronchitis in the anamnesis
zab_leg_02	Obstructive chronic bronchitis in the anamnesis
zab_leg_03	Bronchial asthma in the anamnesis
zab_leg_06	Pulmonary tuberculosis in the anamnesis
O_L_POST	Pulmonary edema at the time of admission to intensive care unit
K_SH_POST	Cardiogenic shock at the time of admission to intensive care unit
MP_TP_POST	Paroxysms of atrial fibrillation at the time of admission to intensive care unit
SVT_POST	Paroxysms of supraventricular tachycardia at the time of admission to intensive care unit
GT_POST	Paroxysms of ventricular tachycardia at the time of admission to intensive care unit
FIB_G_POST	Ventricular fibrillation at the time of admission to intensive care unit
ant_im	Presence of an anterior myocardial infarction (left ventricular)
lat_im	Presence of a lateral myocardial infarction (left ventricular)
inf_im	Presence of an inferior myocardial infarction (left ventricular)
post_im	Presence of a posterior myocardial infarction (left ventricular)
IM_PG_P	Presence of a right ventricular myocardial infarction
ritm_ecg_p_01	ECG rhythm at the time of admission to hospital – sinus (with a heart rate 60-90)
ritm_ecg_p_02	ECG rhythm at the time of admission to hospital – atrial fibrillation
ritm_ecg_p_04	ECG rhythm at the time of admission to hospital – atrial
ritm_ecg_p_06	ECG rhythm at the time of admission to hospital – idioventricular
ritm_ecg_p_07	ECG rhythm at the time of admission to hospital – sinus with a heart rate above 90
ritm_ecg_p_08	ECG rhythm at the time of admission to hospital – sinus with a heart rate below 60
n_r_ecg_p_01	Premature atrial contractions on ECG at the time of admission to hospital
n_r_ecg_p_02	Frequent premature atrial contractions on ECG at the time of admission to hospital
n_r_ecg_p_03	Premature ventricular contractions on ECG at the time of admission to hospital
n_r_ecg_p_04	Frequent premature ventricular contractions on ECG at the time of admission to hospital
n_r_ecg_p_05	Paroxysms of atrial fibrillation on ECG at the time of admission to hospital
n_r_ecg_p_06	Persistent form of atrial fibrillation on ECG at the time of admission to hospital
n_r_ecg_p_08	Paroxysms of supraventricular tachycardia on ECG at the time of admission to hospital

Variables	Description
n_r_ecg_p_09	Paroxysms of ventricular tachycardia on ECG at the time of admission to hospital
n_r_ecg_p_10	Ventricular fibrillation on ECG at the time of admission to hospital
n_p_ecg_p_01	Sinoatrial block on ECG at the time of admission to hospital
n_p_ecg_p_03	First-degree AV block on ECG at the time of admission to hospital
n_p_ecg_p_04	Type 1 Second-degree AV block (Mobitz I/Wenckebach) on ECG at the time of admission to hospital
n_p_ecg_p_05	Type 2 Second-degree AV block (Mobitz II/Hay) on ECG at the time of admission to hospital
n_p_ecg_p_06	Third-degree AV block on ECG at the time of admission to hospital
n_p_ecg_p_07	LBBB (anterior branch) on ECG at the time of admission to hospital
n_p_ecg_p_08	LBBB (posterior branch) on ECG at the time of admission to hospital
n_p_ecg_p_09	Incomplete LBBB on ECG at the time of admission to hospital
n_p_ecg_p_10	Complete LBBB on ECG at the time of admission to hospital
n_p_ecg_p_11	Incomplete RBBB on ECG at the time of admission to hospital
n_p_ecg_p_12	Complete RBBB on ECG at the time of admission to hospital
fibr_ter_01	Fibrinolytic therapy by Celiasum 750k IU
fibr_ter_02	Fibrinolytic therapy by Celiasum 1m IU
fibr_ter_03	Fibrinolytic therapy by Celiasum 3m IU
fibr_ter_05	Fibrinolytic therapy by Streptase
fibr_ter_06	Fibrinolytic therapy by Celiasum 500k IU
fibr_ter_07	Fibrinolytic therapy by Celiasum 250k IU
fibr_ter_08	Fibrinolytic therapy by Streptodecase 1.5m IU
L_BLOOD	White blood cell count (billions per liter)
TIME_B_S	Time elapsed from the beginning of the attack of CHD to the hospital
NITR_S	Use of liquid nitrates in the ICU
NA_R_3_n	Use of opioid drugs in the ICU in the first hours of the hospital period
LID_S_n	Use of lidocaine in the ICU
B_BLOK_S_n	Use of beta-blockers in the ICU
ANT_CA_S_n	Use of calcium channel blockers in the ICU
GEPAR_S_n	Use of a anticoagulants (heparin) in the ICU
ASP_S_n	Use of acetylsalicylic acid in the ICU
TIKL_S_n	Use of Ticlid in the ICU
TRENT_S_n	Use of Trental in the ICU