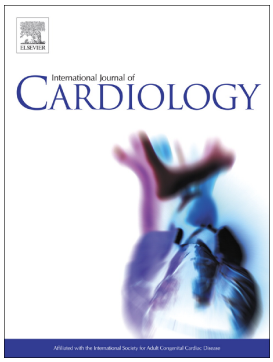


Machine learning to identify a composite indicator to predict cardiac death in ischemic heart disease

Alessandro Pingitore, Chenxiang Zhang, Cristina Vassalle, Paolo Ferragina, Patrizia Landi, Francesca Mastorci, Rosa Sicari, Alessandro Tommasi, Cesare Zavattari, Giuseppe Prencipe, Alina Sîrbu

PII: S0167-5273(24)00531-X

DOI: https://doi.org/10.1016/j.ijcard.2024.131981

Reference: IJCA 131981

To appear in: *International Journal of Cardiology*

Received date: 9 November 2023

Revised date: 13 March 2024

Accepted date: 17 March 2024

Please cite this article as: A. Pingitore, C. Zhang, C. Vassalle, et al., Machine learning to identify a composite indicator to predict cardiac death in ischemic heart disease, *International Journal of Cardiology* (2023), https://doi.org/10.1016/j.ijcard.2024.131981

Machine Learning to identify a composite indicator to predict cardiac death in Ischemic Heart Disease

Alessandro Pingitore[1], MD, PhD, Chenxiang Zhang, MSc[2], Cristina Vassalle[3], PhD, Paolo Ferragina, PhD[2], Patrizia Landi[1], Francesca Mastorci[1], PhD, Rosa Sicari[1], MD, PhD, Alessandro Tommasi, MSc[2], Cesare Zavattari, MSc[2], Giuseppe Prencipe, PhD*[2], Alina Sîrbu, PhD[2]

[1]Clinical Physiology Institute, CNR, Pisa, Italy
[2] Computer Science Department, University of Pisa, Pisa, Italy
[3] Fondazione CNR-Regione Toscana G Monasterio, Pisa, Italy

All authors take responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation.

Corresponding Author:
Giuseppe Prencipe, Associate Professor
Computer Science Department, University of Pisa, Pisa, Italy
TEL: +39-050-2213148
Email: giuseppe.prencipe@unipi.it

ABSTRACT

Background. Machine learning (ML) employs algorithms that learn from data, building models with the potential to predict events by aggregating a large number of variables and assessing their complex interactions. The aim of this study is to assess ML potential in identifying patients with ischemic heart disease (IHD) at high risk of cardiac death (CD).

Methods. 3987 (mean age 68 ± 11) hospitalized IHD patients were enrolled. We implemented and compared various ML models and their combination into ensembles. Model output constitutes a new *ML indicator* to be employed for stratification. Primary variable importance was assessed with ablation tests.

Results. An ensemble classifier combining three ML models achieved the best performance to predict CD (AUROC of 0.830, F1-macro of 0.726). ML indicator use through Cox survival analysis outperformed the 18 variables individually, producing a better stratification compared to standard multivariate analysis (improvement of ~20%). Patients in the low risk group defined through ML indicator had a significantly higher survival (88.8% versus 29.1%). The main variables identified were *Dyslipidemia*, *LVEF*, *Previous CABG*, *Diabetes*, *Previous Myocardial Infarction*, *Smoke*, *Documented resting or exertional ischemia*, with an AUROC of 0.791 and an F1-score of 0.674, lower than that of 18 variables. Both code and clinical data are freely available with this article.

Conclusion. ML may allow a faster, low-cost and reliable evaluation of IHD patient prognosis by inclusion of more predictors and identification of those more significant, improving outcome prediction towards the development of precision medicine in this clinical field.

Keywords: Machine Learning, Ischemic Heart Disease, Prognosis, Survival Analysis, Artificial Intelligence

INTRODUCTION

Ischemic Heart Disease (IHD) remains the most important cause of morbidity and mortality in the world (1). The objective to identify subjects at high risk of adverse events remains difficult due to IHD complexity, in which a large number of variables, clinical, anthropometric, socioeconomic, life-style and, also, cardiovascular imaging contribute to prognostic stratification. Artificial intelligence (AI) is the computer science field related to the capacity to perform tasks normally associated with cognitive abilities. Its subfield of Machine Learning (ML) employs algorithms and builds models learned from data, without specific encoding of knowledge. They provide the potential to predict events in different patient groups by aggregating a large number of variables and assessing their complex interactions (2). In the clinical setting, ML models have been used to identify predictors of events in IHD patients (3-9). Previous studies showed that ML could identify different variables predicting mortality at early and late follow up time. Aziz and al identified age, heart rate, Killip class, fasting blood glucose, prior primary percutaneous revascularization or pharmaco-invasive therapy and diuretics, as predictors for 30 days and 1 year mortality of AMI patients (6). Motwani et al showed that combining clinical and coronary computed tomographic angiography through ML improved prediction of overall mortality at 5 years in respect to include only coronary computed tomographic angiography data (7). D'Assenzo et al showed that ML identified different types of predictors in relation to the type of event considered (8). Okere et al showed that in-hospital length of stay and the mortality risk score, based on Elixhauser comorbidity measure, were predictors of 180-day in-hospital mortality using ML approach (9). In the study of Chiu, the fusion of six classifiers was used to construct and optimize the stacked set of second level classifiers, with an accuracy of 95 % in predicting mortality of HF patients at 3 days, 1 and 3

months and 1 year (10). Moreover, random Survival Forest (RSF) approaches, which are RF models trained to maximize stratification, were also used to identify, among a large number of variables, the predictors of cardiovascular events in a population-based observational cohort study (11). The authors used the RSF approach both to select event predictor variables (i.e., 20 for each event), and as a model for risk prediction. This approach of risk prediction provided better event prediction over standard risk scores, such as the Framingham score and the Framingham Cardiovascular score, and also the Cox proportional hazard regression model. Additionally, ML methods without missing value imputation can outperform expert variable selection with imputation of missing values, with better performances in terms of prediction and risk stratification (12). Thus, the hypothesis of the present study was that ML can identify IHD patients at high risk of long-term cardiac death (CD). Therefore, we conducted a retrospective analysis including patients with known IHD, considering clinical data in the ML analysis as potential predictors of CD occurring at long follow-up time (7 years).

METHODS

Patient Dataset

The study included 3987 (mean age 68 ± 11 years) IHD patients hospitalized (1977-2011) at the CNR Clinical Physiology Institute in Pisa, Italy (angina 35.24%, arrhythmias 4.44%, dyspnea 8.21%, documented myocardial ischemia 26.52%, acute heart failure 9.16%, syncope 1.67%, acute coronary syndrome 24.29, valvulopathy 1.48%), followed up for up to 10 years after hospital admission. IHD definition included one or more of the following conditions: 1) at least one coronary vessel with stenosis >= 75%; 2) acute myocardial infarction (MI); 3) previous coronary artery bypass surgery (CABG); 4) previous coronary intervention; 5) previous MI; 6)

post-ischemic dilated cardiomyopathy (IDC). Smoking habits, IHD family history, arterial hypertension, diabetes, obesity and dyslipidemia were coded in a dichotomized fashion (values 0/1; Smoking habits: 0-never smokers, 1-smokers for current/ex-smokers). All patients either suffered CD or survived for the whole observation period. Informed consent was obtained from each patient. The study complied with the Declaration of Helsinki regarding ethical conduct of research involving human subjects. All data (completely anonymous, evaluated as aggregated and not individually) were acquired in the context of institutional clinical assistance within clinical care purposes in a retrospectively collected modality from our Institution patient's dataset (*Image database),* including clinical characteristics, previous history, IHD risk factors, comorbidities, laboratory and instrumental results, pharmacological therapies, and post-discharge follow-up outcomes. Exclusion criteria: severe systemic diseases (e.g. neoplasia, acute or chronic inflammatory disease, immunological disease), non-CD during the observation period, refusal or inability to supply written Informed Consent (13-15). The data are available publicly on Github: https://github.com/orientino/ml4cad.

Follow-up

Follow-up data were obtained through review of the patient's record, telephone interview, personal communication with the patient's physician, or medical check. Death cause was derived from medical records or death certificates. CD definition required either significant arrhythmias, or cardiac arrest, or death attributable to congestive heart failure, or myocardial infarction, in the absence of any other precipitating factor.

ML Analysis

ML analysis aims to develop an indicator based on multiple clinical variables suitable to stratify patients at high and low risk of CD. This was achieved through: 1) building ML model to separate two target classes: patients who survived for 7 years from hospital admission, and those who underwent CD within the same period; besides a binary classification, this model also produces a probability of surviving for more than 7 years, which is the novel ML indicator; 2) evaluating the performance of the ML indicator using survival analysis; 3) identifying the most important clinical variables for this model, through single-variable and multi-variable ablation studies, obtaining a simplified ML indicator. The computational analysis was performed in Python, both data and analysis are publicly available at https://github.com/orientino/ml4cad.

ML Dataset Preparation

The processed dataset consists of 18 *independent variables*. We defined the binary *dependent variable* "survive7Y" indicating whether the patient survived for at least 7 years starting from the hospital admission. Even though follow-up was longer than 7 years, we chose the value "7 years" empirically, since we observed better predictive performance for it on the validation data. However, we also tested a 10-year threshold, with very similar results for both prediction and stratification (see the results section for the results of the 10-year models). Since the final dataset presented a class imbalance (84% of patients survived >7 years), we employed two techniques for reducing its effects: "class weights" and "dataset sampling" (17). Among the independent variables, creatinine and number of stenosed coronary vessels ("Vessels") contained missing values, replaced with 0 (after translating the "Vessels" values by 1).

ML Model Building

We employed a standard ML pipeline (2,3): we divided the dataset into 3 smaller datasets preserving the class ratio, with 60% of the patients placed in the training dataset (2391), 20% in the validation dataset (798) and 20% in the test dataset (798). We standardized all 3 datasets, by computing the scaling parameters (mean and standard deviation) on the training dataset to prevent data leakage. Standardisation was performed after replacing missing values with 0. This decreased original averages and shifted the distributions of the Creatinine and Vessels variables to the left. However, it was necessary, otherwise the missing values would have been equal to the mean of the distribution of the variables after standardisation. For our models, we needed missing values to be different from others as they include information on medical decisions (i.e. recording or not a certain variable, see also ref. 12). Then, we trained the ML models using the training dataset to predict the dependent variable "survive7Y". The models were evaluated using the validation dataset, while the test dataset was reserved exclusively to evaluate the final model with best results on the validation dataset. We used: logistic regression, support vector classifier, k-nearest neighbors, random forest (RF), adaboost, multilayer perceptron, gradient boosting, and extreme gradient boosting. For each model type, we optimized several model hyperparameters, and employed early stopping during training where possible, using 2-fold cross-validation on training data only. We include the complete definition of the hyperparameter search space as *Supplementary file 1*. To select the best hyperparameter combination for each model, given that in our case we train multiple models with a different number of hyperparameters to be optimized, we performed 5000 iterations of random search. Random search provides performance results comparable with a full grid search, but with much shorter running times (17,18), due to the fact that it does not explore all hyperparameter combinations but randomly selects a subset of combinations to test. We have also tested a higher number of samples (10000) but results were

very similar. To mitigate dataset imbalance, we applied "dataset sampling", which modifies the training dataset by simultaneously undersampling the largest class, and by oversampling the smallest class using SMOTE, Borderline SMOTE, or SVMSMOTE (16). We chose among these procedures during the hyperparameter optimisation phase. Lastly, the final proposed model was designed as an *ensemble* of a subset of ML models. The ensemble computes the output of all the models in the set and averages them to obtain a better output estimate. The combination of models chosen as the final model was based on the performance on the validation dataset. We evaluated the models using F1-macro, the area under the receiver operating characteristic curve (AUROC), precision and recall. We also performed calibration analysis, in order to evaluate how the probabilities that the models produce are aligned with the actual class labels. We report the Brier scores for all models, and we also compare probabilities with actual survival times.

ML Variable Importance

We estimated variable importance by using "*ablation tests*" (19), applying the trained predictive model (the ensemble model above) to patients where one or several variables are replaced with their mean value, to remove the information they hold, i.e., performing a variable knock-out. A new F1-score is computed, denoted by F1'. If F1' is lower than the original F1, then we can conclude that these variables are important for the prediction. Thus, we define a *variable importance* as the ratio F1/F1'. A value above 1 indicates an important variable; a value below 1 would indicate a noisy variable that does not help classification. Ablation tests have two main advantages: First, compared to feature selection methods that use rankings internal to the models (e.g., impurity-based feature importance for random forest), they are model agnostic, so they can

be used also with models that do not provide an internal importance of features (such as our ensemble models). Second, compared to methods that use an external ranking (such as statistical tests or variance based methods), they can be used for multivariable analysis, i.e., estimating the importance for a set of features simultaneously. Furthermore, they have the advantage that they do not make any assumption on the probability distribution of variables, unlike most statistical tests (20).

We performed two types of ablation analyses to assess variable importance: 1) single-variable analysis, knocking out one variable at a time; 2) multi-variable analysis, using hierarchical clustering (16) to identify groups of most similar variables, and then knocking out one group at a time. The *variable group importance* is calculated as the ratio between the original F1 score and the score after knocking out all the variables in the group, F1', similarly to the importance of single variables. Single- and multi-variable analysis were then combined to extract a set of important clinical variables for a reduced ML indicator, by considering the most important variable of each cluster.

The advantage of multi-variable ablation tests is that they can enable simultaneous knockout of correlated variables, removing thus completely the signal contained in that group. In single-variable tests, even if we knock out one variable, if there is another correlated variable then the model could still be able to perform well, by employing the information from the correlated variable.

Survival Analysis

We conducted survival analysis using the Kaplan-Meier curves on the test dataset, considering a CD event in the first 7 years. Firstly, we investigated the quality of stratification obtained by

using the model output, i.e. the model's probability of survival over 7 years, using a threshold to divide the dataset into two parts. A low threshold on the probability will divide the patients into a very small homogeneous high-risk group and a large heterogeneous low risk group. As the threshold increases, the high-risk group grows but at the same time it becomes more heterogeneous, with patients with higher survival, while the low-risk group becomes smaller. In general, even for higher thresholds, the low-risk group is much smaller, since the median probability value on the test dataset is 0.695. Here, we used a threshold of 0.6 that allowed us to maintain a good proportion between the two groups without sacrificing too much the homogeneity of the smaller high-risk group. , A different Kaplan-Meier estimator was used to fit on each group: by plotting the estimator's results it was possible to visualize the quality of the stratification. Lastly, univariate and multivariate Cox Regression has also been performed, obtaining a Concordance value (C-index) for the predictor variables producing a quantitative comparison. To evaluate model performance, we again divided the patients into training and test datasets, and we report the average performance in 5-fold cross validation. We used the lifelines Python library (21) for survival analysis.

RESULTS

Patients

Clinical characteristics of the 3987 patients are summarized in Table 1. Patients with CD were older, had higher incidence of diabetes, atrial fibrillation, previous CABG and previous MI, IDC, reduced LVEF and a higher number of stenosed coronary vessels.

ML Indicator to predict CD risk

We trained various ML classifiers with several parameters, combining the various models to obtain an optimal ensemble. Out of all possible combinations, the best performing ensemble (*validation* F1-macro=0.685 and AUC=0.817) was the one combining LR, RF and adaboost. This integrates the two top performing models based on the validation F1 score with the model in third position in terms of AUC. Table 2 shows standard classification performance metrics on the internal validation dataset and on the external test dataset, for each individual model and for the ensemble model; Figure 1 shows ROC curves for the models. The ensemble classifier achieved the best performance (AUROC of 0.830, F1-macro of 0.726 , precision of 0.705 and recall of 0.762; Figure 1).

Table 2 also shows results of calibration analysis on test data, including Brier scores for all models. We note that all models have very low Brier scores, indicating that 7-year survival probabilities are well aligned with the two classes. Although classification performance is best for the ensemble methods, calibration results are best for the RF model. However, the ensemble model still shows good calibration. To investigate this in more detail, Figure 2 compares the survival probabilities provided by the model on test data with the actual survival in years. We note how as actual survival times increase, the distribution of model output values shifts towards larger values, as required. Also, as we move further from the 7-year threshold, probabilities become more extreme, i.e. closer to 0 or 1, indicating that the model is more confident in these patients. For censored patients, where we do not have an exact survival value, we observe that the model generally assigns large probabilities, again as expected, since we did not observe a CD event in these patients.

The performance of the new ML indicator through Cox survival analysis was compared against univariate and multivariate survival analysis performed on the original clinical variables. Table 3

shows the performance of Cox univariate regression for all variables individually, along with the performance of traditional multivariate Cox regression: not only the ML indicator outperformed the 18 variables individually, but it also produced a better stratification compared to standard multivariate analysis (improvement in C-index of ~15%).

Figure 3a shows the Kaplan-Meier curves for the ML Indicator, the high-risk group was significantly separated from the low-risk one by using a ML indicator cut-off of 0.6.

In this study we trained our model by considering events after 7 years of hospitalization. The threshold was chosen based on performance on validation data, however results were very similar for other thresholds. For the 10-year threshold we obtained an AUROC of 0.828 and an F1-macro score of 0.726 for the ensemble model, very close to that of 7 years (Table 2). For Cox regression, the C-index with the ML indicator with 18 variables was 0.814 for 10 years, and 0.816 for 7 years. Thus, the values are comparable and still larger than that obtained with Cox regression on the original variables, even multivariate, suggesting that the performance of our analysis does not depend a lot on the chosen threshold.

Primary variable selection and a simplified indicator

To select a subset of variables that maximizes the model performance, we performed single- and multi-variable ablation tests for feature selection. Table 4 shows the ranking of variables, and corresponding importance in single-variable ablation tests, for the 18 clinical variables employed in our study. Top variables were dyslipidemia, LVEF, Diabetes, previous MI, also showing very low p-values in Table 3 (univariate Cox Regression). Ranking obtained with ML methods agrees with Cox analysis: top variables in Table 4 are also at the top (or at least statistically significant) in Table 3. However, the opposite is not true; i.e., some significant variables in univariate Cox

Regression are ranked low by our method: Age, IDC, and angiography show little contribution to the predictive performance but have very low p-values at Univariate Cox Analysis. Among these, age maintains a low p-value also at Multivariate Cox Analysis. All this suggests that the ML model finds and uses its relationship with other related variables (e.g., creatinine, smoke and gender, see clustering results below), processing its significance in connection with other related biomarkers. Thus, assigning importance based on ML methods could have better ability to account for complex relations over traditional statistical analysis. Almost all variables provided a contribution, with values generally above one, albeit at times the contribution was very low (importance close to 1, Table 4). The only variable reducing the predictive power of the ML model is angina (importance<1), also not significant at Cox regression. Besides single-variable importance analysis, ML methods also have the advantage of enabling meaningful multivariable analysis, through the combination of clustering and ablation tests. Clinical variables were first clustered into meaningful groups and then ablation tests assigned a predictive importance to each group. Figure 4 shows the 7 variable clusters, which group together variables correlated from a statistical as well as clinical viewpoints, thus validating the overall clustering procedure. Table 4 shows the 7 cluster ranking, and their importance in the predictive model, with a good agreement with results presented in Table 4, (most important variables being those related to dyslipidemia, LVEF, IDC, and CABG). All clusters contribute to the prediction (values>1), suggesting the importance of using all available data. We note that some importance values were much closer to 1 in single-variable ablation tests. This is probably because in single-variable tests the model can employ information from a correlated variable to maintain a high performance. However, in multivariable tests that is not possible, explaining the generally higher importance values. When the single-variable and multivariable analysis were combined, a 7-variable-combined indicator

(dyslipidemia, LVEF, previous CABG, diabetes, previous MI, smoke, and documented resting or exertional ischemia) was obtained. ML prediction power achieved AUROC of 0.791,F1-score of 0.674, precision of 0.656 and recall of 0.719, lower than the model that combines all 18 variables (Table 2). Univariate Cox regression using the simplified ML indicator obtained an overall C-index of 0.77 , while multivariate Cox regression using the same 7 clinical variables resulted in a C-index of 0.74. Again, the simplified ML indicator produces a more meaningful combination with respect to the 7 variables, (better risk stratification), being also superior to all individual variables (see Table 3 for comparison). However, as seen before for AUROC and F1-score, the simplified indicator did not outperform the full version, which combines all 18 variables. Figure 3b shows Kaplan-Meier curves, where the 2 groups separated according to the threshold of 0.6 were significantly different.

DISCUSSION

In this study we applied ML tools for CD risk stratification in a large dataset of IHD patients. The main results can be summarized as follows:

1) the proposed ML model was able to predict a CD with AUROC of 0.830, F1-macro of 0.726, precision of 0.705 and recall of 0.762, with a calibration concordance index of 0.143 .

2) the ensemble method, consisting of an aggregation of logistic regression, RF and Adaboost models, had the higher prognostic stratification capability, superior to standard survival models;

3) the most important variables at single-variable ablation analysis were dyslipidemia, LVEF, diabetes, previous MI and paroxysmal or chronic atrial fibrillation;

4) considering the 7 clusters identified at multi-variable ablation analysis, the prognostic predictive weight of our indicator was superior to the standard survival models using the same 7 variables, but significantly lower when compared to that obtained by our same indicator, including all the 18 variables of the study.

These results show the effectiveness of ML-based models to identify IHD patients at high risk of CD, and to identify the major risk factors through ablation or analysis of feature importance, as demonstrated here. ML, in particular the use of ensemble methods for risk stratification, has been previously applied in different settings of cardiac diseases, and mainly in patients with AMI (5-10,12,18,22-31), which however are largely different from our study in terms of patient population, type of outcomes, follow-up lengths, and variables.

We focused on longer term CD prediction (7 years), and obtained AUROC values superior to those found in the literature for long-term event prediction. Furthermore, we employed the output of ML models as a new *ML indicator* for stratification, evidencing improved patient stratification compared to existing clinical variables. Our method, however, does not build the model having survival analysis as a goal, but for a different, related, classification task (predicting the events), and then employs the output as a variable in the survival analysis, similar to other recent works (5,25,26,29). This approach has the advantage of employing any ML model, and not only RSF.

Furthermore, we employed the new integrated model to identify the best predictive variables. Unlike the above-mentioned works, we used a *posteriori* method of variable ranking (ablation tests) that enabled us to evaluate how important the contribution of each variable was. Then, we performed multivariable analysis through clustering and ablation to identify important groups of

variables, which has not been done in previous works. Finally, we identified 7 important factors that were integrated into a simplified ML indicator. Similarly to previous studies (11,27), the indicator combining 18 variables showed better predictive performance in comparison to that including only 7 variables, as the addition of more parameters increases the possibility that some of them contain prognostic value helpful to increase model performance. This result highlights the importance of processing as many variables as possible, in the prognostic prediction analysis, allowing for a more personalized and systemic approach to individual patients. This is definitely aligned with *personalized* or *precision medicine*, and suggests that all variables can enter ML analysis, potentially increasing prognostic stratification power, without the risk of overfitting or undercutting the input data (12). While adding more variables in a model generally augments the risk of overfitting, the fact that the number of patients available is starting to grow, due to the increased digitisation of healthcare, reduces this risk. Furthermore, ML pipelines include various levels of cross validation that allow to avoid overfitting and ensure model generalization. In this way, all variables can be included without having to give up on any useful information through data undercutting. Our study brings further support in this direction, with the integrated ML indicator able to outperform other survival models. Furthermore, ML makes minimal assumptions about the data-generating systems and the probability distributions underlying the processes being measured. Therefore, it is more effective compared to classical statistical methods in presence of a large number of data and variables gathered without a carefully controlled experimental design, and in the presence of complicated nonlinear interactions, as in our study population (32).

Study limitations

ML methods are considered as “black-box methods”, meaning that they provide a predictive capacity without explaining *why* certain results are obtained. The ranking of variables that we provided is a first step towards improving the explainability of our ML indicator. The simplified indicator also contributes to explainability and could be easier to adopt as the number of included variables is smaller. However, this simplification reduced the predictive and stratification power. The study is retrospective, with data gathered several years ago, thus lacking new variables, evidencing how new integrated variables obtained from ML analysis need to be continuously updated with new predisposing parameters.

By excluding no-CD in the training phase, we may have introduced a selection bias, not reflecting the clinical reality of patient outcomes. However, as stated, the model can be used on all deaths, both CD and no-CD. This approach incurs an obvious limitation, that is, some no-CD deaths could be attributed to a cardiovascular cause, and that is the reason why many authors do recommend all-cause deaths as an end-point (33). However, our decision to exclude those deaths was taken to ensure a better understanding of how the AI-model works, because we did not want to query the model on end points that could not be predicted upfront, such as cancer and/or accidental events. It is the scope of our future research to develop the model into a more comprehensive assessment of outcome.

CONCLUSION

ML indicator, including all the available clinical variables, produced a higher stratification compared to the standard approach (improvement of ~20% in survival analysis). ML approaches allow for accurate, reliable and low cost prognostic and risk stratification models in IHD patients, favoring the development of precision medicine.

Conflict of Interest: there is no conflict of interest to declare by all authors.

Acknowledgments: this study was supported by the Tuscan Region through the FESR project IPOTERI: Innovazione nel POtenziamento e TElemonitoraggio della RIabilitazione post operatoria and by the National Recovery and Resilience Plan (NRPP), Mission 4, Component 2 Investment 1.4 R&D "Innovation Ecosystems", project Tuscany Health Ecosystem (THE).

REFERENCES

1) Roth G, Mensah GA, Johnson CO et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019 Update From the GBD 2019 Study. J AM Coll Cardiol 2020; 76:2982-3021
2) Lj, H. Statistics versus machine learning. Nat Methods. 2018;15:233
3) Lin A, Kolossavary M, Motwani M et al. Artificial intelligence in cardiovascular imaging for risk stratification in coronary artery disease. Radiology: Cardiovascular Imaging 2021;3:e200512
4) Haq IU, Chhatwal K, Sanaka K, Xu B. Artificial Intelligence in Cardiovascular Medicine: Current Insights and Future Prospects. Vasc Health Risk Manag. 2022;12;18:517-528.
5) Myers, P.D., Scirica, B.M. and Stultz, C.M., 2017. Machine learning improves risk stratification after acute coronary syndrome. Scientific reports, 7(1), p.12692
6) Aziz F, Malek S, Ibrahim KS, et al. (2021) Short and long-term mortality prediction after an acute ST-elevation myocardial infarction (STEMI) in Asians: A machine learning approach. PLoS ONE 16(8): e0254894.

7) Motwani M. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European Heart Journal (2017) 38, 500–507

8) D'Assenzo F. et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. Lancet 2021;397:199-207

9) Okere AN, Sanogo V, Alqhtani H, Diaby V. Identification of risk factors of 30-day readmission and 180-day in-hospital mortality, and its corresponding relative importance in patients with Ischemic heart disease: a machine learning approach. Expert Rev Pharmacoecon Outcomes Res. 2021;21:1043-1048

10) Chiu CC, Wu CM, Chien TN, et al. Applying an Improved Stacking Ensemble Model to Predict the Mortality of ICU Patients with Heart Failure. J Clin Med. 2022;11:646

11) Ambale-Venkatesh B, Yang X, Wu CO et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121:1092-1101

12) Steele AJ, Denaxas SC, Shah AD, et al. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PloSONE. 2018;13:e0202344

13) Pingitore A, Carpeggiani C. The Need for Open-access Structured Data in Cardiology Research. Biomed Data J. 2015; 1: 36-38

14) Iervasi G, Molinaro S, Landi P et al. Association Between Increased Mortality and Mild Thyroid Dysfunction in Cardiac Patients. Arch Intern Med. 2007;167:1526-1532.

15) Pingitore A, Picano E, Varga A et al. Prognostic value of pharmacological stress echocardiography in patients with known or suspected coronary artery disease: a

prospective, large scale, multicenter, head-to-head comparison between dipyridamole and dobutamine test. J Am Coll Cardiol 1999;34: 1769-7

16) Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms. 2009;3:4-21

17) Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research. 2012;13: 281-305

18) Anggoro, D.A. and Mukti, S.S. Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure. International Journal of Intelligent Engineering & Systems, 2021;14:

19) Meyes R, Lu M, Waubert de Puiseau C, Meisen T. Ablation Studies in Artificial Neural Networks. 2019

20) Verma, J. P., & Abdel-Salam, A. S. G. *Testing statistical assumptions in research*. John Wiley & Sons. 2019

21. Davidson-Pilon. Lifelines: survival analysis in Python. Journal of Open Source Software-2019;4:1317

22. Goldman O, Raphaeli O, Goldman E, Leshno M. Improvement in the prediction of coronary heart disease risk by using artificial intelligence networks. Qual Manag Health Care. 2021;30:244-250

23. Morgan DJ, Bame B, Zimand P et al. Assessment of Machine Learning vs Standard Prediction Rules for Predicting Hospital Readmissions. JAMA Netw Open. 2019;2:e190348

24. Bergamini M, Iora PH, Rocha TAH, et al. Mapping risk of ischemic heart disease using machine learning in a Brazilian state. PLoS ONE 15(12): e0243558.

25. Liu N et al. Utilizing machine learning dimensionality reduction for risk stratification of chest pain patients in the emergency department. BMC Medical Research Methodology (2021) 21:74

26. Kwon L Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. PLoS ONE. 2019;14:e0224502

27. Hadanny A. Predicting 30-day mortality after ST elevation myocardial infarction: Machine learning- based random forest and its external validation using two independent nationwide datasets. Journal of Cardiology. 2021:78:439–446

28. Beaulieu-Jones, B.K., Yuan, W., Brat, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ digital medicine, 2021;4:62

29. Gibson, W.J., Nafee, T., Travis, R., et al. Machine learning versus traditional risk stratification methods in acute coronary syndrome: a pooled randomized clinical trial analysis. J Thromb Thrombolysis. 2020;49:1-9

30. Kasim, S., Malek, S., Song, C., et al. 2022. In-hospital mortality risk stratification of Asian ACS patients with artificial intelligence algorithm. Plos one. 2022;17:p-e.0278944

31. Panchavati, S., Lam, C., Zelin, et al. Retrospective validation of a machine learning clinical decision support tool for myocardial infarction risk stratification. Health Technol Lett. 2021;8:139-147

32. Bzdok D. Statistics versus machine learning. Nat Methods. 2018;15:233–234

33. Lauer MS, Blackstone EH, Young JB, Topol EJ. Cause of death in clinical research. Time for a reassessment? J AM Coll Cardiol. 1999;34:618-620

FIGURE TITLES AND CAPTIONS

Figure 1. AUROC Curves for all the ML Models.

On the left the curves for each individual model, while on the right we superimpose the ensemble model that shows best performance. AUROC: area under the receiver operating characteristic curve; logistic regression (LR), support vector classifier (SVC); k-nearest neighbors (KNN); random forest (RF), adaboost; multilayer perceptron (MLP); gradient boosting (GB); and extreme gradient boosting (XGB).

Figure 2. Calibration analysis for the ensemble model.

The figure displays the survival probabilities generated by the ML model with 18 variables, for different actual survival times. We note that probabilities increase as survival time increases, as required.

Figure 3. Kaplan-Meyer curves for the ML Indicator and for the simplified ML indicator.

Figure 3a: we stratify using the threshold 0.6 on the probability of surviving more than 7 years estimated by the model. The curves show the fraction of patients that survive for a given number of years, separately for those under the threshold (blue line) and those over the threshold (red line). Patients classified as low risk by our model survive much longer, with a final survival at 88.8% for the red curve and 29.1% for the green curve. Ischemic Heart Disease (IHD).

Figure 3b: We stratify using the thresholds 0.6 on the probability of surviving more than 7 years estimated by the simplified model. The final survival values for the two groups are 89.1% for the low risk group (red curve) and 44.3% for the high risk group (blue curve). Ischemic Heart Disease (IHD).

Figure 4. Clustering clinical variables.

Dendrogram produced by the clustering algorithm, which grouped all the variables into 7 clusters containing related variables. Myocardial Infarction (MI); coronary by-pass grafting (CABG); left

ventricular ejection fraction (LVEF); coronary stenosed vessels (Vessels); previous coronary intervention (PCI).

TABLES (each on a separate page)

Table 1. Table describing the patients and the 18 variables considered. P-values correspond to a univariate chi-square analysis on CVD death/no CVD[1] death for discrete variables, and a Wilcoxon rank-sum test for continuous variables (Age, LVEF[2], Vessels, Creatinine). A total of 3987 patients are included, of which 757 suffered CVD death over the observation period.

	Count (#)/ Mean±SD	**Percent. (%)**	**CVD Death (n=757; 18.99%)**	**No CVD Death (n=3230; 81.01%)**	**p-value**
Sex, male/female	3058/929	76.7%/ 23.3%	566/191; 74.8%/25.2%	2492/738; 77.1%/22.9%	0.1776
Age (years)	68±11		74±10	67±11	< 0.0001
Angina	2676	67.1%	473; 62%	2203; 68%	0.0029
Previous CABG[3]	589	14.8%	193; 25.4%	396; 12.2%	< 0.0001
Previous PCI[4]	966	24.2%	167; 22.0%	799; 24.7%	0.1337
PMI[5]	1836	46.0%	465; 61.4%	1371; 42.4%	< 0.0001
AMI[6]	661	16.5%	139; 18.3%	522; 16.1%	0.1581
LVEF	49.58±12.18		40.28±14.19	51.75±10.54	<

[1] CVD: cardiovascular diseases
[2] LVEF: left ventricular ejection fraction
[3] CABG: coronary by-pass grafting
[4] PCI: previous coronary intervention
[5] PMI: previous myocardial infarction
[6] AMI: acute myocardial infarction

					0.0001
Resting or exertional ischemia	2501	62.7%	390; 51.5%	2111; 65.3%	< 0.0001
Post ischemic DCM[7]	746	18.7%	290; 38.3%	456; 14.1%	< 0.0001
History of smoke	1903	47.7%	330; 43.5%	1573; 48.6%	0.0127
History of diabetes	1096	27.4%	267; 35.2%	829; 25.6%	< 0.0001
History of hypertension	2503	62.7%	474; 62.6%	2029; 62.8%	0.9508
History of dyslipidemia	3228	80.9%	487; 64.3%	2741; 84.8%	< 0.0001
Paroxysmal or chronic - AF[8]	475	11.9%	170; 22.4%	305; 9.4%	< 0.0001
Creatinine	1.20±0.67		1.5±1.02	1.14±0.55	<0.0001
Angiography	3193	80.0%	493; 65.1%	2700; 83.5%	<0.0001
Vessels	1.74±1		2.15±1.02	1.66±0.98	0.0094

Table 2. Building the ML Indicator: comparison of the model performance for different model types. The best model is reported in the last row, while the best result per metric is marked with *.

[7] DCM: dilated cardiomyopathy
[8] AF: atrial fibrillation

Model	**Val. AUROC**	**Val. F1-macro**	**Test AUROC**	**Test F1-macro**	**Test Precision-macro**	**Test Recall-macro**	Test Brier Score
LR[9]	0.808	0.672	0.820	0.704	0.683	0.750	0.139
SVC[10]	0.781	0.668	0.790	0.683	0.664	0.737	0.211
k-NN[11]	0.708	0.607	0.758	0.660	0.646	0.727	0.166
RF[12]	0.805	0.680	0.826	0.698	0.695	0.700	***0.119**
Adaboost	0.799	0.680	0.786	0.663	0.647	0.704	0.236
MLP[13]	0.789	0.668	0.802	0.708	0701	0.717	0.120
GB[14]	0.811	0.664	0.815	0.704	0.683	0.747	0.137
XGB[15]	0.810	0.676	0.820	0.708	0.686	0.759	0.139
Ensemble (LR, RF, Adaboost)	***0.817**	***0.685**	***0.830**	***0.726**	***0.705**	***0.762**	0.143

[9] LR: logistic regression
[10] SVC: support vector classifier
[11] K-NN: k-nearest neighbors
[12] RF: random forest
[13] MLP: multilayer perceptron
[14] GB: gradient boosting
[15] XGB: extreme gradient boosting

Table 3. Univariate and multivariate Cox regression: performance in terms of C-index for standard clinical variables (univariate and multivariate) and for the novel ML[16] Indicator (univariate, because treated as a single composite variable containing all the information of other variables).

	Variable	**C-index**
Univariate Cox regression	**ML Indicator (18 variables)**	**0.82**
	LVEF[17]	0.75
	Age	0.69
	Post-ischemic Dilated Cardiomyopathy	0.62
	Dyslipidemia	0.62
	Angiography	0.59
	Previous CABG[18]	0.56
	Documented resting or exertional ischemia	0.58
	Paroxysmal or chronic atrial fibrillation	0.54
	Previous Myocardial Infarction	0.54
	Diabetes	0.53
	Creatinina	0.60
	Gender	0.55
	Acute Myocardial Infarction	0.54
	Smoke	0.51
	Hypertension	0.50

[16] ML: machine learning
[17] LVEF: left ventricular ejection fraction
[18] CABG: coronary by-pass grafting

	Angina	0.51
	Previous PCI[19]	0.49
	Vessels	0.48
Multivariate Cox Regression (18 variables)		0.71

[19] PCI: previous coronary intervention

Table 4. Single-variable ablation tests for variable ranking. The Importance column quantifies the increase in performance (in terms of F1 score, see "Methods") when the variable is included in the model. The higher the value, the more important the variable.

Variable	Importance
Dyslipidemia	1.141
LVEF[20]	1.106
Diabetes	1.056
PMI[21]	1.052
Paroxysmal or chronic atrial fibrillation	1.044
Previous CABG[22]	1.041
Smoke	1.022
Vessels	1.021
Creatinina	1.018
Gender	1.016
Documented resting or exertional ischemia	1.016
Previous PCI[23]	1.013
Angiography	1.013
Post-ischemic Dilated Cardiomyopathy	1.011
Hypertension	1.011
AMI[24]	1.005
Age	1.001
Angina	0.995

[20] LVEF: left ventricular ejection fraction
[21] PMI: previous myocardial infarction
[22] CABG: coronary by-pass grafting
[23] PCI: previous coronary intervention
[24] AMI: acute myocardial infarction

Table 5. Multi-variable ablation tests for variable ranking. The Importance column quantifies the increase in performance (in terms of F1 score, see "Methods") when the group of variables is included in the model. The higher the value, the more important the group of variables.

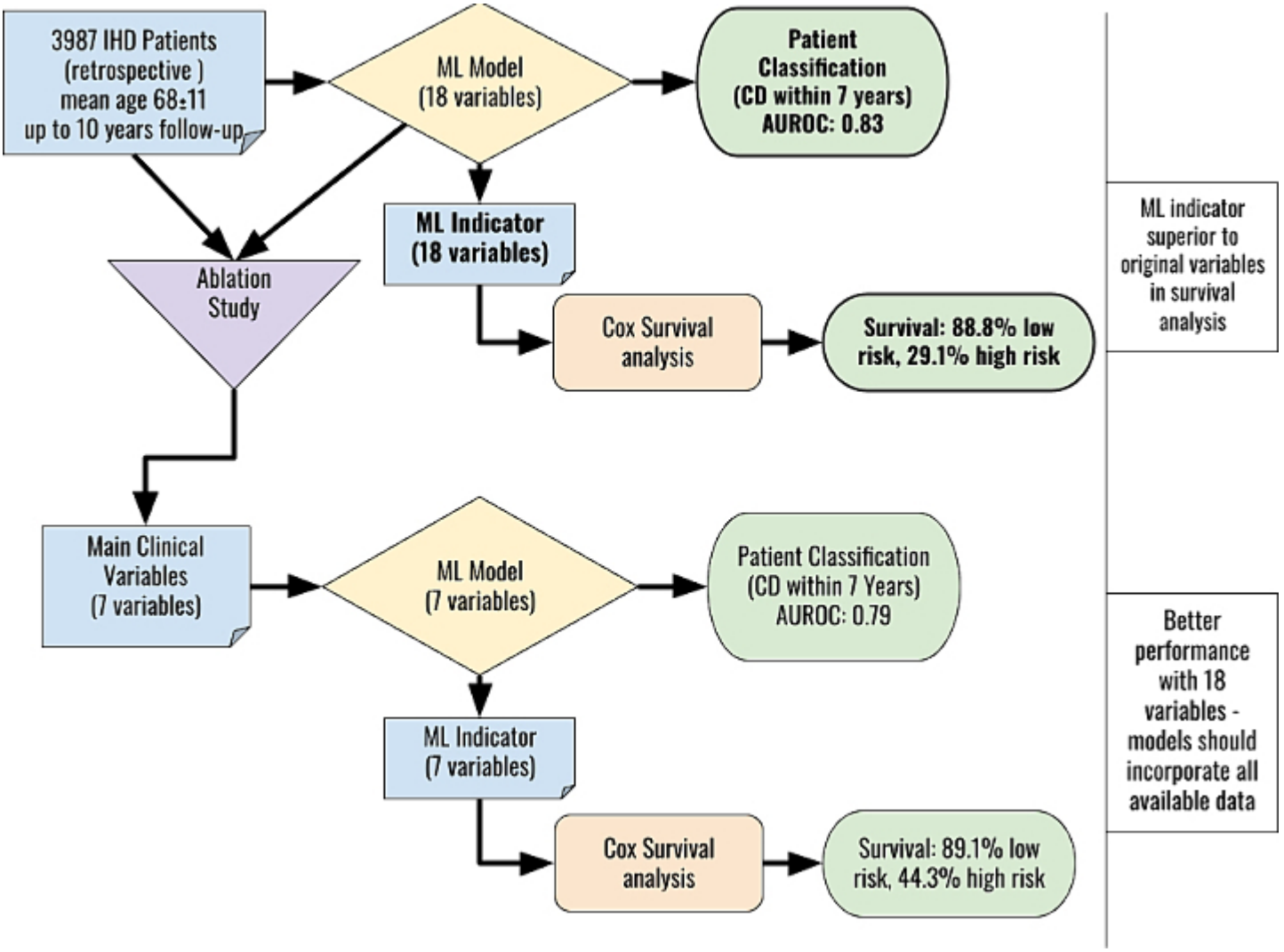
Cluster	Variables	Importance
1	Dyslipidemia, Paroxysmal or chronic atrial fibrillation	1.243
2	LVEF[25], Post-ischemic Dilated Cardiomyopathy	1.101
3	Previous CABG[26], Angiography, Vessels	1.081
4	Diabetes, Hypertension	1.080
5	Previous PCI[27], Previous Myocardial Infarction	1.061
6	Gender, Age, Smoke, Creatinine	1.053
7	Angina, AMI[28], Documented resting or exertional ischemia	1.034

[25] LVEF: left ventricular ejection fraction
[26] CABG: coronary by-pass grafting
[27] PCI: previous coronary intervention
[28] AMI: acute myocardial infarction

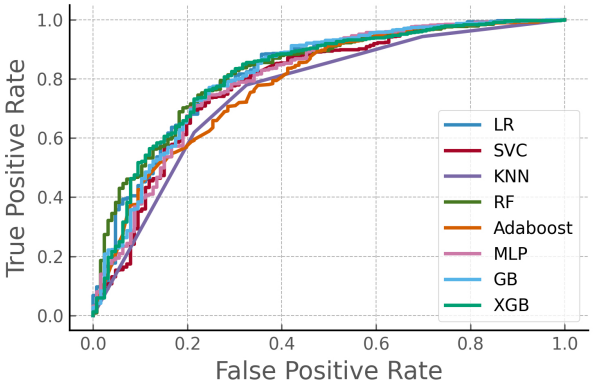
Graphical abstract

Highlights

Machine Learning improves cardiac death prediction patients with ischemic heart disease



Graphics Abstract


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LR
SVC
KNN
RF
Adaboost
MLP
GB
XGB

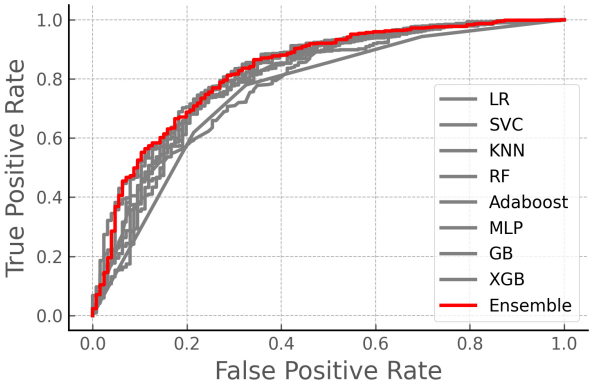

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LR
SVC
KNN
RF
Adaboost
MLP
GB
XGB
Ensemble

Figure 1

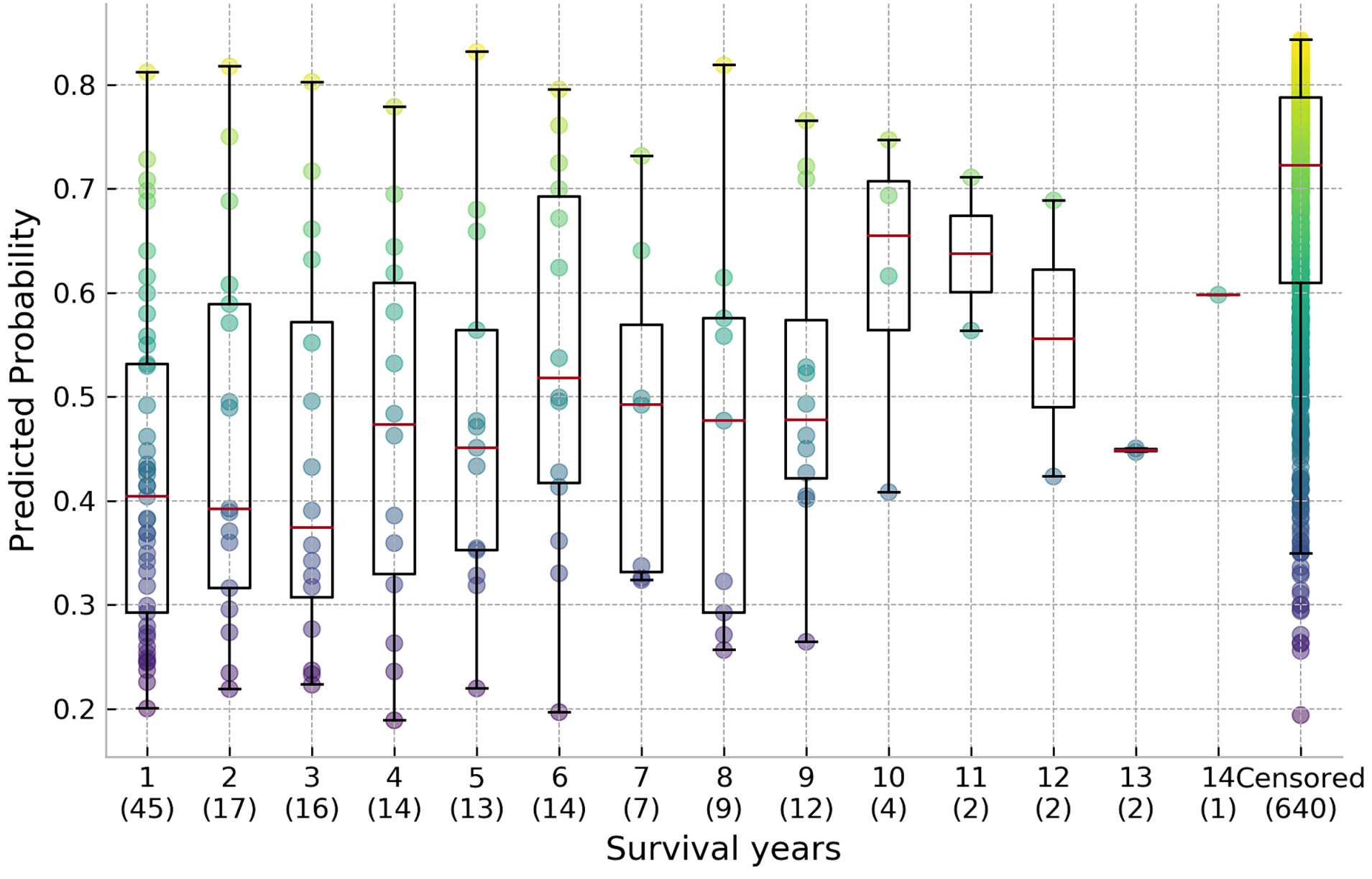
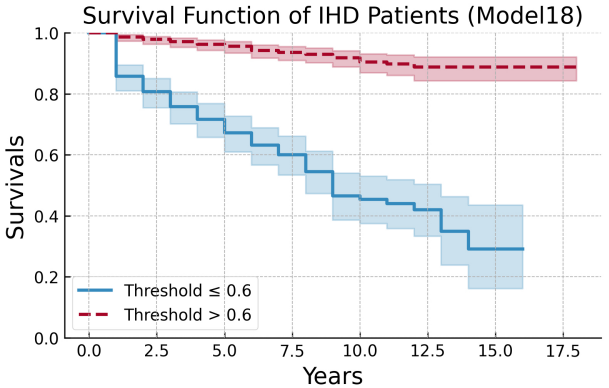

Predicted Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
1 (45)
2 (17)
3 (16)
4 (14)
5 (13)
6 (14)
7 (7)
8 (9)
9 (12)
10 (4)
11 (2)
12 (2)
13 (2)
14 (1)
Censored (640)
Survival years

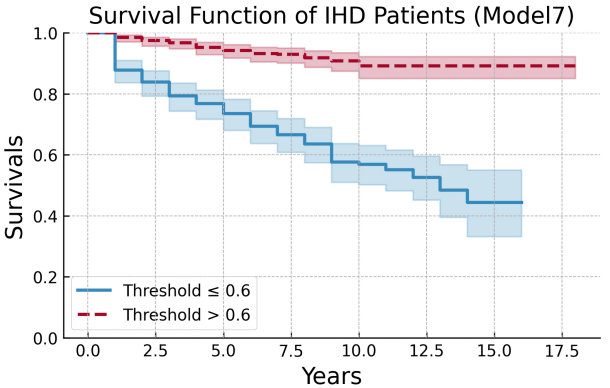
Figure 2

(a)



Threshold ≤ 0.6								
At risk	268	198	133	76	31	12	1	0
Censored	0	19	52	96	124	141	149	150
Events	0	51	83	96	113	115	118	118
Threshold > 0.6								
At risk	530	488	398	302	141	42	3	1
Censored	0	31	110	198	351	448	487	489
Events	0	11	22	30	38	40	40	40

(b)



Threshold ≤ 0.6								
At risk	312	244	180	128	64	25	1	0
Censored	0	18	53	89	136	171	192	193
Events	0	50	79	95	112	116	119	119
Threshold > 0.6								
At risk	486	442	351	250	108	29	3	1
Censored	0	32	109	205	339	418	444	446
Events	0	12	26	31	39	39	39	39

Figure 3

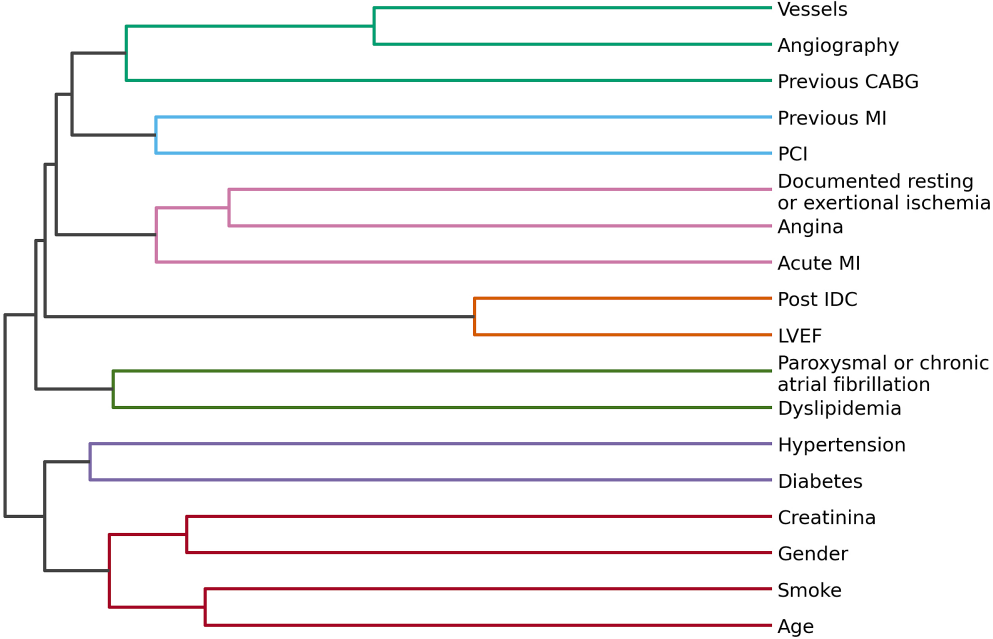

Vessels
Angiography
Previous CABG
Previous MI
PCI
Documented resting
or exertional ischemia
Angina
Acute MI
Post IDC
LVEF
Paroxysmal or chronic
atrial fibrillation
Dyslipidemia
Hypertension
Diabetes
Creatinina
Gender
Smoke
Age

Figure 4