Sains Malaysiana 50(3)(2021): 753-768
*http://dx.doi.org/10.17576/jsm-2021-5003-17*

# Predicting 30-Day Mortality after an Acute Coronary Syndrome (ACS) using Machine Learning Methods for Feature Selection, Classification and Visualisation

(Meramalkan Kematian 30 Hari selepas Sindrom Koronari Akut (ACS) menggunakan Kaedah Pembelajaran Mesin untuk Pemilihan Ciri, Pengelasan dan Pemvisualan)

NANYONGA AZIIDA, SORAYYA MALEK*, FIRDAUS AZIZ, KHAIRUL SHAFIQ IBRAHIM & SAZZLI KASIM

ABSTRACT

*Hybrid combinations of feature selection, classification and visualisation using machine learning (ML) methods have the potential for enhanced understanding and 30-day mortality prediction of patients with cardiovascular disease using population-specific data. Identifying a feature selection method with a classifier algorithm that produces high performance in mortality studies is essential and has not been reported before. Feature selection methods such as Boruta, Random Forest (RF), Elastic Net (EN), Recursive Feature Elimination (RFE), learning vector quantization (LVQ), Genetic Algorithm (GA), Cluster Dendrogram (CD), Support Vector Machine (SVM) and Logistic Regression (LR) were combined with RF, SVM, LR, and EN classifiers for 30-day mortality prediction. ML models were constructed using 302 patients and 54 input variables from the Malaysian National Cardiovascular Disease Database. Validation of the best ML model was performed against Thrombolysis in Myocardial Infarction (TIMI) using an additional dataset of 102 patients. The Self-Organising Feature Map (SOM) was used to visualise mortality-related factors post-ACS. The performance of ML models using the area under the curve (AUC) ranged from 0.48 to 0.80. The best-performing model (AUC = 0.80) was a hybrid combination of the RF variable importance method, the sequential backward selection and the RF classifier using five predictors (age, triglyceride, creatinine, troponin, and total cholesterol). Comparison with TIMI using an additional dataset resulted in the best ML model outperforming the TIMI score (AUC = 0.75 vs. AUC = 0.60). The findings of this study will provide a basis for developing an online ML-based population-specific risk scoring calculator.*

*Keywords: Acute coronary syndrome; feature selection; hybrid model; machine learning; self-organising maps*

ABSTRAK

*Gabungan hibrid pemilihan ciri, pengelasan dan pemvisualan menggunakan kaedah pembelajaran mesin (ML) mempunyai potensi untuk pemahaman yang lebih baik untuk ramalan kematian pesakit bagi tempoh 30 hari dengan penyakit kardiovaskular menggunakan data penduduk yang khusus. Mengenal pasti ciri-ciri kaedah pemilihan dengan algoritma pengelas yang menghasilkan prestasi tinggi dalam kajian kematian adalah penting dan tidak pernah dilaporkan sebelum ini. Ciri-ciri kaedah pemilihan seperti 'Boruta', 'Random Forest' (RF), 'Elastic Net' (EN), 'Recursive Feature Elimination' (RFE), 'Learning Vector Quantization' (LVQ), 'Genetic Algorithm' (GA), 'Cluster Dendrogram' (CD), 'Support Vector Machine' (SVM) dan 'Logistic Regression' (LR) telah digabungkan dengan algoritma bagi pengelasan RF, SVM, LR dan EN bagi ramalan kematian bagi tempoh 30 hari. Model ML telah dibina menggunakan 302 pesakit dan 54 pemboleh ubah input dari Pangkalan Data Penyakit Kardiovaskular Kebangsaan Malaysia. Pengesahan terbaik model ML telah dijalankan dengan Trombolisis dalam Infarksi Miokardium (TIMI) menggunakan set data tambahan daripada 102 pesakit. Peta swaurus (SOM) telah digunakan untuk menggambarkan faktor yang berkaitan dengan kematian selepas ACS. Prestasi model diukur menggunakan kawasan di bawah lengkung (AUC) antara 0.48-0.80. Model terbaik mencatatkan (AUC = 0.80) adalah gabungan hibrid RF cara kepentingan berubah-ubah, pemilihan ke belakang berurutan dan pengelas RF menggunakan lima peramal (umur, trigliserida, kreatinin, troponin dan jumlah kolesterol). Model terbaik telah dibandingkan dengan TIMI menggunakan set data tambahan yang menyebabkan model ML mengatasi TIMI (AUC = 0.75 vs AUC = 0.60). Penemuan daripada kajian ini akan digunakan sebagai asas untuk membangunkan talian ML berdasarkan pengiraan pemarkahan risiko yang penduduk tertentu.*

*Kata kunci: Model hibrid; pembelajaran mesin; pemilihan ciri; peta swaurus sindrom koronari akut*

## INTRODUCTION

Heart attack or acute coronary syndrome (ACS) is the leading cause of mortality in the world (Castro-Dominguez et al. 2018). ACS is categorised into ST-elevation myocardial infarction (STEMI), non-ST elevation myocardial infarction (NSTEMI), and unstable angina (UA) (Torres et al. 2007). In Malaysia, 20-25% of all deaths are due to coronary artery disease (Hoo et al. 1969). Conventional predictive tools such as Thrombolysis in Myocardial Infarction (TIMI) score and the Global Registry of Acute Cardiac Events (GRACE) have limitations. Data may be lost due to fixed expectations on data performance and rigidness in the criteria for pre-selecting variables (Shouval et al. 2017). The TIMI risk score was developed from a North American population with limited participation from an Asian population. It is the most widely used risk predictor in Malaysia to predict ACS outcomes.

It is important to consider the significant population-specific mortality-related features of Malaysian ACS patients. It is to achieve a reliable, effective clinical diagnosis specifically tailored for the Malaysian population to reduce ACS-related mortality rate and monetary costs. The recognition of significant risk factors associated with Malaysian-specific mortality is pivotal as it allows for more accurate diagnosis, increases mortality rates, and reduces financial burdens.

Machine learning (ML) methods, consisting of automatic feature selection, allow the manipulation of large numbers of predictors and does not require underlying assumptions regarding the relationship between input and output features (Chen et al. 2012). Feature selection methods are categorised according to the ML classification algorithms used: embedded, filter, and wrapper methods. The filter method is a standalone feature selection method; the wrapper method uses data mining algorithms such as recursive feature elimination (RFE), sequential backward selection (SBS) and forward feature selection; and the embedded method is a combination of both filter and wrapper. Well-established examples for the wrapper and embedded methods include Random Forest (RF) and Elastic Net (EN). Achieving the highest accuracy of performance and selecting the smallest number of features are essential for optimising ML classification algorithms.

In mortality-related studies, RF, Naïve Bayes (NB), Logistic Regression (LR), Support vector machine (SVM) have been used for feature selection and prediction of mortality post-ACS and outperformed conventional methods such as TIMI and GRACE (Collazo et al. 2016; Motwani et al. 2016; Shouval et al. 2017; Steele et al. 2018; Wallert et al. 2017). Kohonen Self-Organising Map (SOM) allows the discovery of relationship and pattern in a dataset that leads to the discovery of knowledge through the visualisation of SOM maps. SOM has been used in the analysis of paediatric fracture (Kohonen et al. 2001; Malek et al. 2018; Tuckova et al. 2013). Application of SOM to discover and visualise the relationship between mortality-related variables is essential and has not been addressed in mortality-related studies.

The objective was to investigate the feasibility of various feature selection methods and ML classifiers to improve the deductive reasoning and performance for the prediction of 30-day mortality post-ACS using Malaysian-specific dataset. The use of different categories of feature selection methods to improve mortality prediction models has not been reported in the literature. We also compare the ML model with the conventional TIMI risk scoring method. We also propose SOM to visualise and identify the mortality-related factors post-ACS. The proposed study will further be used to develop an online population-specific ML-based mortality risk calculator.

## MATERIALS AND METHODS

The National Cardiovascular Disease-Acute Coronary Syndrome (NCVD-ACS) Registry records information about patients treated at participating institutions across Malaysia. The study cohort was drawn from registered patients admitted to the Coronary Care Unit (CCU), Universiti Teknologi MARA (UiTM) Sungai Buloh Hospital for ACS between 2014 and 2016. Fifty-four input variables were used based on the recommendation of the cardiologist. The hospital cardiologist agreed to diagnosis ACS based on clinical symptoms, electrocardiograms, biomarkers, and echocardiograms. Datasets from 302 patients with 54 variables were used without data imputation (11 continuous and 43 categorical; Table 1). Data were split for model training (70%) and testing (30%); similar datasets were used in the construction of all ML models (Hinde et al. 2003; Kuhn et al. 2008). An additional 102 datasets were used for comparing the best ML model found in this study with the conventional TIMI risk scoring method, a standard method, used in Malaysian hospitals. The present study was approved by the Institutional Review Board of the Universiti Teknologi MARA, which waived patients' informed consent. All data was then provided to the researchers for this study.

TABLE 1. Summary statistics of predictors used in this study. The table shows the values which are mean ± SD or median derived from SPSS; Uncorrected P-values are from Welch's t-tests if the variable is continuous, or Pearson Chi-square test if categorical

| | All cases (n =302) | Survivors (n = 279) | Non-survivors (n = 23) | p-value | Confident interval (CI = 95%) |
|---|---|---|---|---|---|
| Age (years) | 56.72 ± 11.7 | 56.5 ± 11.5 | 58.7 ± 14.3 | 0.001 | 55.40-58.06 |
| Age group | | | | | |
| 29-55 | 131 (43.3%) | 123 (44.1%) | 8 (34.78%) | | |
| 55-65 | 101 (33.44%) | 94 (33.69%) | 7 (30.43%) | | |
| Above 65 | 70 (23.17%) | 62 (22.22%) | 8 (34.78%) | | |
| Gender | | | | | |
| Male | 206 (68.2%) | 189 (67.7%) | 17 (73.9%) | 0. 001 | |
| Female | 96 (31.7%) | 90 (32.2%) | 6 (26.0%) | | |
| Ethnicity | | | | | |
| Malay | 159 (52.6%) | 146 (52.3%) | 13 (56.5%) | 0. 001 | |
| Chinese | 28 (9.27%) | 26 (9.31%) | 2 (8.69%) | | |
| Indian | 104 (34.43%) | 97 (34.7%) | 7 (30.43) | | |
| Others | 11 (3.64) | 10 (3.58%) | 1 (4.34%) | | |
| PCI type | 286 (94.7%) | 263 (94.2%) | 23 (100%) | 0. 001 | |
| Combined diabetes melitus and ACS subtypes | 122 (40.3%) | 112 (40.1%) | 10 (43.4%) | 0.612 | |
| Thrombolysis | 279 (92.3%) | 258 (92.4%) | 21 (91.3%) | 0. 001 | |
| Stk successful | 279 (92.3%) | 258 (92.4%) | 21 (91.3%) | 0. 001 | |
| LCL Simon Broome | 281 (93.0%) | 259 (92.8%) | 22 (95.6%) | 0. 001 | |
| TC Simon Broome | 290 (96.0%) | 269 (89.0%) | 21 (91.3%) | 0. 001 | |
| ACS subtype | | | | 0. 001 | |
| Unstable angina | 170 (56.29%) | 161 (57.7%) | 9 (39.1%) | | |
| NSTEMI | 77 (25.49) | 67 (24.01%) | 10 (43.47%) | | |
| STEMI | 51 (16.88%) | 47 (16.84%) | 4 (17.39%) | | |
| Others ACS Subtype | 4 (1.32) | 4 (1.43%) | 0 (0%) | | |
| Smoker | 232 (76.8%) | 212 (75.9%) | 20 (86.9%) | 0. 001 | |
| Ex-smoker | 270 (89.4%) | 248 (88.8%) | 22 (95.6%) | 0. 001 | |
| Hypertension | 230 (76.1%) | 210 (75.2%) | 20 (86.9%) | 0. 001 | |
| Alcohol | 292 (96.6%) | 270 (96.7%) | 22 (95.6%) | 0.001 | |
| Diabetes mellitus | 180 (59.6%) | 167 (59.8%) | 13 (56.5%) | 0. 001 | |
| Newly dm | 294 (97.3%) | 271 (97.1%) | 23 (100%) | 0.005 | |
| TC (mmol/L) | 4.89 ± 1.31 | 4.92 ± 1.30 | 4.52 ± 1.48 | 0. 001 | 4.74-5.04 |
| LDL (mmol/L) | 3.04 ± 1.13 | 3.07± 1.13 | 2.70 ± 1.11 | 0. 001 | 2.92-3.17 |
| HDL (mmol/L) | 1.02 ± 0.27 | 1.03 ± 0.26 | 0.94 ± 0.28 | 0.103 | 0.99-1.05 |
| Ticagrelor | 293 (97.0%) | 270 (96.7%) | 23 (100%) | 0. 001 | |

| | | | | | |
|---|---|---|---|---|---|
| FBS (mmol/L) | 7.98 ± 3.43 | 7.97 ± 3.37 | 8.13 ± 4.19 | 0. 001 | 7.59-8.37 |
| eGFR (mL/min) | 66.55 ± 34.11 | 67.33 ± 33.77 | 56.99 ± 37.39 | 0. 001 | 62.6-70.4 |
| ACE | 204 (67.5%) | 191 (68.4%) | 13 (56.5%) | 0. 001 | |
| CCB | 225 (74.5%) | 206 (73.8%) | 19 (82.6%) | 0. 001 | |
| Beta blockers | 221 (73.1%) | 206 (73.8%) | 15 (65.2%) | 0. 001 | |
| Treatment-modality | 277 (91.7%) | 256 (91.7%) | 21 (91.3%) | 0. 001 | |
| Clopidogrel | 235 (77.8%) | 221 (79.2%) | 14 (60.8%) | 0. 001 | |
| ARB | 283 (93.7%) | 260 (93.1%) | 23 (100%) | 0. 001 | |
| Statins | 292 (96.6%) | 269 (96.4%) | 23 (100%) | 0. 001 | |
| Statin dosage(mg) | 164 (54.3%) | 150 (53.7%) | 14 (60.8%) | 0. 001 | |
| Statin medication type | 147 (48.6%) | 141 (50.5%) | 6 (26.0%) | 0. 001 | |
| ASA | 276 (91.3%) | 255 (91.3%) | 21 (91.3%) | 0. 001 | |
| Nitrates | 182 (60.2%) | 170 (60.9%) | 12 (52.1%) | 0. 001 | |
| LM coronary angiogram | 293 (97.0%) | 272 (97.4%) | 21 (91.3%) | 0.003 | |
| LAD coronary angiogram | 267 (88.4%) | 247 (88.5%) | 20 (86.9%) | 0.001 | |
| LCx coronary angiogram | 271 (89.7%) | 251 (89.9%) | 20 (86.9%) | 0.019 | |
| RCA coronary angiogram | 271 (89.7%) | 252 (90.3%) | 19 (82.6%) | 0. 001 | |
| LAD stent | 290 (96.0%) | 268 (96.0%) | 22 (95.6%) | 0.001 | |
| LCx stent | 296 (98.0%) | 274 (98.2%) | 22 (95.6%) | 0.019 | |
| RCA stent | 297 (98.3%) | 275 (98.5%) | 22 (95.6%) | 0.025 | |
| Stroke | 284 (94.0%) | 266 (95.3%) | 18 (78.2%) | 0. 001 | |
| IHD | 178 (58.9%) | 165 (59.1%) | 13 (56.5%) | 0. 001 | |
| Fx IHD | 267 (88.4%) | 245 (87.8%) | 22 (95.6%) | 0. 001 | |
| CCF | 270 (89.4%) | 252 (90.3%) | 18 (78.2%) | 0. 001 | |
| COAD | 287 (90.0%) | 265 (94.9%) | 22 (95.6%) | 0. 001 | |
| BA | 282 (93.3%) | 262 (93.9%) | 20 (86.9%) | 0. 001 | |
| Obesity | 296 (98.0%) | 275 (98.5%) | 21 (91.3%) | 0.014 | |
| CKD | 274 (90.7%) | 255 (91.3%) | 19 (82.6%) | 0. 001 | |
| Dyslipidemia | 176 (58.2%) | 160 (57.3%) | 16 (69.5%) | 0. 001 | |
| HbA1c (mmol/mol) | 7.46 ± 2.23 | 7.49 ± 2.26 | 7.10 ± 1.85 | 0. 001 | 7.20-7.71 |
| Creatinine (µmol/L) | 103.44 ± 61.5 | 101.5 ± 57.63 | 126.2± 96.22 | 0. 001 | 96.47-110.4 |
| Troponin_1(ng/L) | 12.38 ± 135.1 | 4.60 ± 18.47 | 106.7 ± 485.3 | 0.112 | -2.91-27.69 |
| Tg (mg/dL mmol/L) | 1.85 ±1.20 | 1.84 ± 0.95 | 1.96 ±2.89 | 0. 001 | 1.71-1.98 |
| CK (u/L) | 361.3 ± 735.5 | 344.8 ± 698.9 | 562.2 ± 1087.4 | 0. 001 | 278.0-444.6 |

PCI: Percutaneous Coronary Intervention; ACS: Acute coronary syndrome; Stk: Streptokinase; HDL: High-density lipoprotein; Tg: Triglycerides; TC: Total cholesterol; CK: Creatine kinase; LDL: Low-density lipoprotein; eGFR: Estimated glomerular filtration rate; FBS: Fasting blood sugar; CCB: Calcium channel blockers; Newly dm: Newly diagnosed diabetes mellitus; CKD: Chronic kidney disease; ASA: Aspirin; BA: Bronchial asthma; COAD: Chronic obstructive airway disease; ARB: Angiotensin II receptor blockers; CCB: Calcium Channel Blocker; ACE: Angiotensin-Converting Enzyme; CCF: Congestive cardiac failure; IHD: Ischemic Heart Disease; Fx IHD: Fracture Ischemic Heart Disease; LM: Left main; LAD: Left anterior descending; LCx: Left–circumflex; RCA: Right coronary artery

A 10-fold cross-validation re-sampling with three repeats was used for model development of the training set to avoid over-fitting (Geisser 1993). Parameter tuning was carried out on all models. Later, the values of the parameters were selected based on a model that had performed higher than others. The hold-out, untouched test set was only used for validation, i.e. the final performance test of the developed models. This untouched set was predicted according to the class incidence as occurring in the clinical population. The area under the curve (AUC) was used as a predictive performance metric that is insensitive to class imbalances (Fawcett 2006). The sequential backwards selection (SBS) method was used to eliminate ranked variables in ascending order iteratively employing feature selection methods to improve model performance (Genuer et al. 2010). The SBS algorithm relies on significance as a sufficient condition to remove insignificant variables from the model (Dunkler et al. 2014). The variable that causes a significant increase in AUC in the testing dataset of the prediction model is considered necessary.

A hybrid combination of SBS and feature selection methods from embedded, filter, and wrapper methods was combined with classifier algorithms (Chandrashekar et al. 2014; Saeys et al. 2007). The filter method Cluster Dendrogram (CD) using Euclidean distance is a standalone feature selection method (Cox et al. 1958); CD calculates the correlation between the features in terms of Euclidean distance. The feature selection using CD was obtained by cutting the dendrogram at the desired level, where each connected component forms a cluster and the characteristics were selected based on the levels of each cluster. Wrapper method uses data mining algorithms such as Boruta (Kursa et al. 2010), RFE (Jafarian et al. 2011), LVQ (Hammer et al. 2002), GA (Holland et al. 1992), RF (Breiman 2001), and SVM with Radial Basis Function (RBF) (Vapnik 1998). The Boruta approach compares the importance of real predictor variables with those of random shadow variables using statistical testing and several RF runs. The goal of the RFE method is to find a minimal and best-performing set of variables by using the RFE feature importance function, and RFE tries to remove dependencies and collinearity that might occur in the model by recursively removing a small number of features at each iteration. The parameter settings for the GA in this study were experimentally determined by parameter optimisation. The final feature selection was conducted based on a population size of 275 and iteration of 200.

RF is an ensemble method that builds many decision trees randomly from bootstrapping samples, which are then clustered together by a classification method and-a by-product of RF-the variable importance. In this study, different values of mtry (mtry: 5, 7, 10, 15, and 20) and the number of trees (ntree: 500-4000) were used to determine the optimum RF model that produced the best results. The RF variable importance method was used to generate ranked variables that were then reduced using SBS iteratively.

For this analysis, SVM was implemented using the RBF kernel. The tuning parameters for SVM are the C parameter (cost), which regulates the margin width, and the gamma-parameter for the kernel calculation. In this study, SVM uses the Receiving Operating Characteristic (ROC) variable importance to select and rank important variables. In the case of two-class problems, a series of cut-offs were applied to predictor data to predict the class. Sensitivity and specificity were calculated for each cut-off and the ROC curve is computed. ROC is used as a measure of variable importance. The parameter tuning used was sigma 0.5 and cost 10 using a grid search for the SVM classifier.

The embedded method is a combination of both filter and wrapper, such as elastic net (EN) (Zou & Trevor 2005) and LR (Menard 2002). The function glm with the family binomial was used for constructing the LR model. EN optimises the coefficients until the change of the coefficients is less than the predetermined toleration value. Parameter tuning with maxit = 1000000, alpha = 0.5, lambda = 100 was done to improve model performance. The feature selection methods were then combined with RF, SVM, EN and LR classifier algorithms. These classifier algorithms were selected based on their higher predictive overall performance reported in previous mortality studies. In this study, we calculated the developed model predictive performance with a testing dataset that was not used for model development.

SOM was then used to ordinate ACS-related factors using features selected from the best model (Kohonen 2001). High dimension data is best visualised using SOM as it reduces the complexity of high dimension data by plotting data similarity in 1-dimensional or 2-dimensional maps through clustering techniques. Light colour represents clusters, whereas dark clusters represent cluster separators. SOM enables the discovery or identification of most relevant features or patterns through data reduction and projection. The quality of the map is measured by quantification and topological error. The Euclidean distance between the inputs was calculated and visualised as a distance matrix known as the U-matrix (unified distance matrix).

Statistical Package for Social Sciences (SPSS) program version 16.0 was used for statistical analysis, whereas the R package (Version 3.3.2) and the SOM toolbox in MATLAB VER. (R2013, Math Works) were used for model development. Figure 1 summarises the steps and algorithms used in this study.

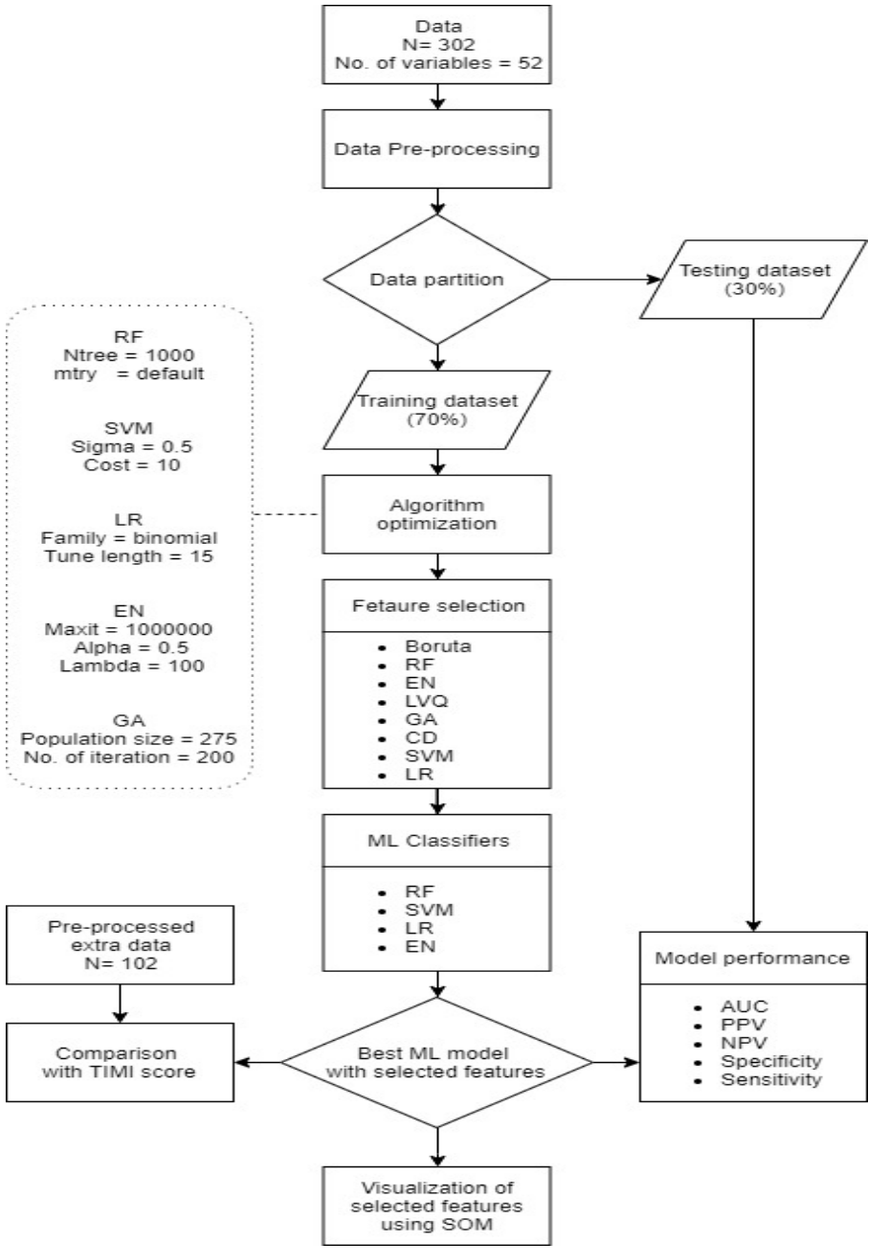


FIGURE 1. A summary of the steps and algorithms used

## Results and Discussion

Table 1 presents the characteristics of the patients used in this study. The mean age was 56.72 ± 11.7, and 67% were males. The majority (82%) of them have an ACS subtype of unstable angina and NSTEMI. Overall, the 30-day mortality rate was 7.6%. Predictor differences between survivors and non-survivors were significant and expected at 30 days in terms of age, gender, cardiovascular disease (CVD) diagnosis and severity, CVD risk factors, CVD comorbidities, biomarkers, and medicines (pb = 0.0001). Non-survivors are those associated with higher risk factors for CVD, such as smokers, history of hypertension, alcoholism, and diabetes. Among CVD biomarkers, there was no significant difference in troponin value between survivors and non-survivors.

RF, Boruta, RFE, LVQ, GA, CD, LR, EN, and SVM methods were used to rank predictor importance against 30-day mortality post-ACS. Figures 2 and 3 illustrate the ten most significant predictors chosen by each model out of a total of 54 predictors. Some predictors were significant across all models-Age, triglycerides (Tg), total cholesterol (TC), troponin, creatinine, estimated glomerular filtration rate (eGFR), high-density lipoprotein (HDL) and Creatine kinase (CK); whereas other predictors were model-specific-fasting blood sugar (FBS), Haemoglobin A1c (HbA1c), ethnicity statin medications, stroke, ACS subtype and treatment modularity. Overall, the different models chose a heterogeneous set of the most important predictors (cardiac variables, medications, and demographics).

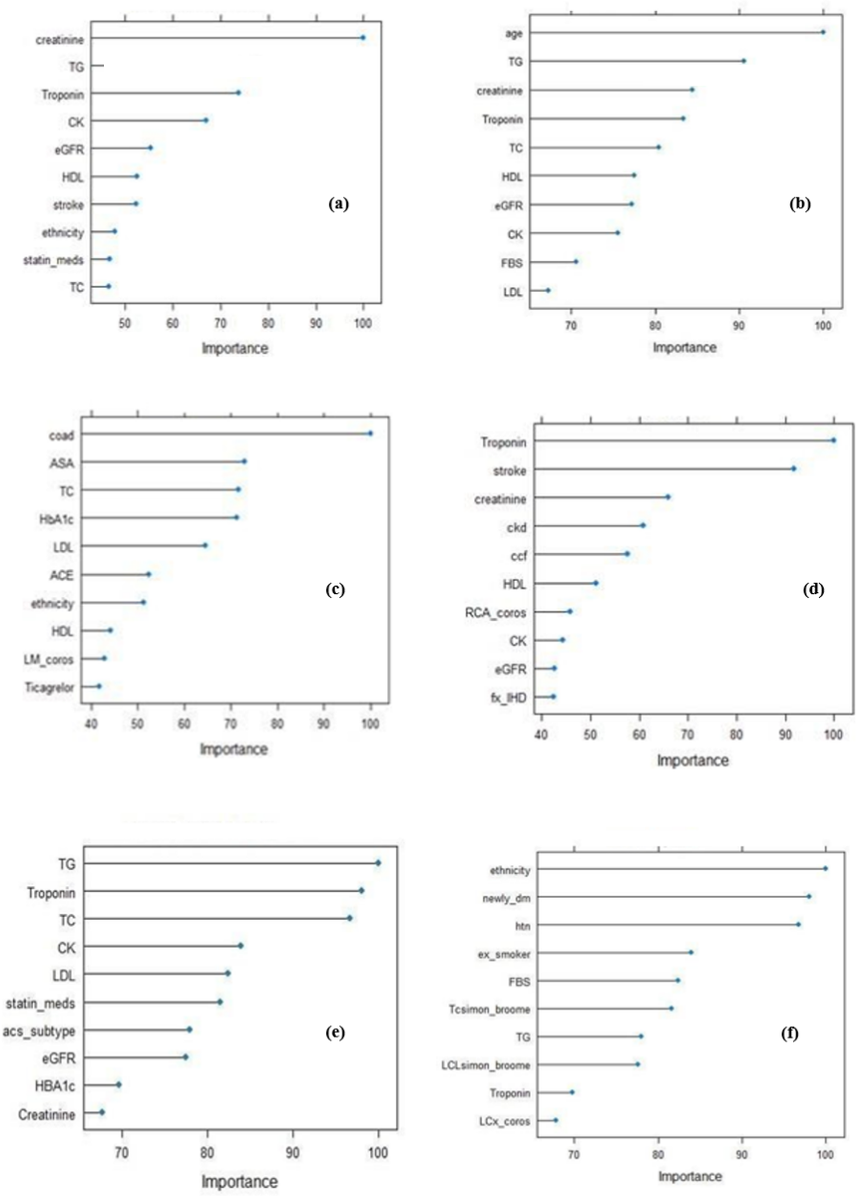


FIGURE 2. The graph of feature importance rank of the ten most important predictors chosen by (a) Learning Vector Quantification, (b) Random Forest, (c) Logistic Regression, (d) Elastic Net, (e) Support Vector Machine and (f) Genetic Algorithm

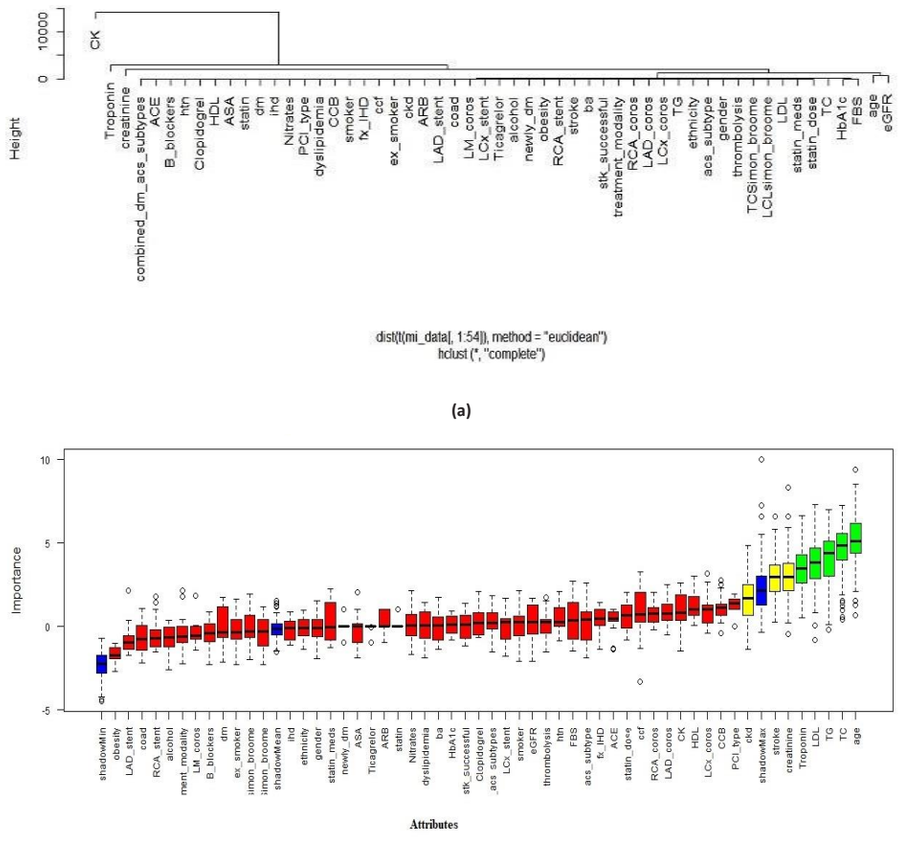


FIGURE 3. Illustration of all predictors that were selected by both Dendrogram (a) and Boruta (b)

SBS was used for feature reduction to improve the predictive performances of the ML model measured using AUC on the testing dataset, as illustrated in Figure 4. The AUC of 0.5 suggests no discrimination, 0.7 to 0.8 is measured as acceptable, 0.8 and above as excellent (Mandrekar et al. 2010).

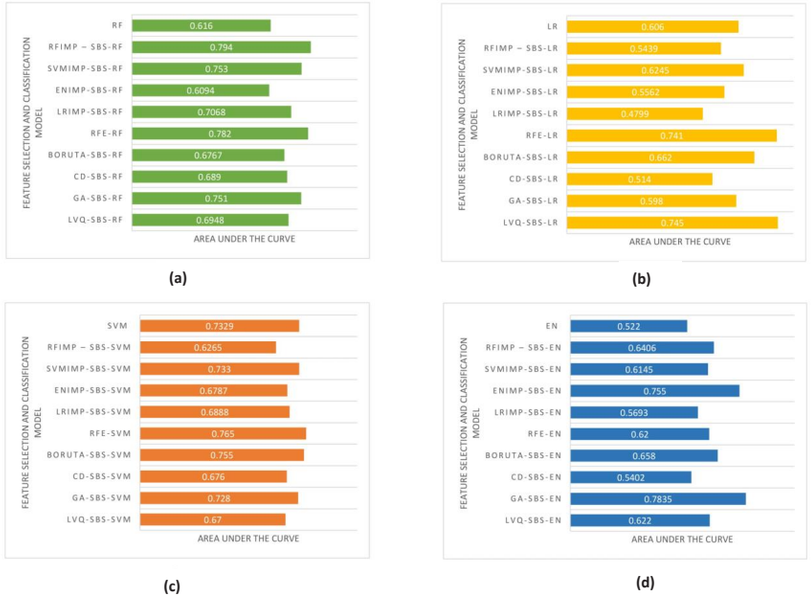


FIGURE 4.The graphs of feature selection and classifier combination performance. The predictive performance of 30 days' mortality prediction using different feature selection method and classifier: (a) Random Forest, (b) Logistic Regression, (c) Support Vector Machine and (d) Elastic Net

SVM reported the highest AUC (0.73), outperforming RF (0.62), EN (0.52), and LR (0.61) among the models developed using all 54 predictors. In this study, significant predictors with optimum ML performance identified across all ML models after SBS were age, Tg, TC, creatinine, CK, and troponin as presented in Table 2.

TABLE 2. The selected variables for each of the feature selection method proposed in this study. List of all the variables used for classification and its performance results are in AUC

| Feature selection and classification model | Variable selected |
|---|---|
| RANDOM FOREST | |
| RF | All variables |
| RFVarImp-SBS-RF | Age, TC, Tg, Troponin, Creatinine |
| SVMVarimp - SBS - RF | TC, Tg, Troponin, CK |
| ENVarimp-SBS - RF | Troponin, Creatinine, Stroke, CKD |
| LRVarImp-SBS-RF | TC, HDL, HbA1c, LDL, COAD, ASA, ACE, ethnicity, LM coronary angiogram, Ticagrelor, alcohol, Statin |
| RFE-SBS - RF | Age, TC, Tg, Troponin, Stroke |
| Boruta-SBS-RF | Age, TC, Tg, Troponin, LDL |
| CD-SBS- RF | Age, Troponin, Creatinine, eGFR, CK, HbA1c, FBS |
| LVQ- SBS- RF | Tg, Troponin, Creatinine, HDL, Stroke, eGFR, CK, Ethnicity |
| GA- SBS- RF | Tg, Troponin, Ethnicity, FBS, Newly dm, Hypertension, Ex-smoker, TC Simon Broome, LCL Simon Broome, LCx coronary angiogram |
| SUPPORT VECTOR MACHINE | |
| SVM | All variables |
| RFVarImp-SBS-SVM | Age, TC, Tg, creatinine, Troponin, HDL, eGFR, CK |
| SVMVarImp – SBS – SVM | TC,Tg, Troponin, creatinine, eGFR, CK, LDL, statin_meds, ACS_SUBTYPE, HbA1c |
| ENVarImp-SBS – SVM | Troponin, Creatinine, Stroke, CKD |
| LRVarImp-SBS- SVM | TC, HDL, HbA1c, LDL, COAD, ASA, ACE, ethnicity, LM coronary angiogram, Ticagrelor, alcohol, Statin |
| RFE-SBS – SVM | Age, TC, Tg |
| Boruta -SBS-SVM | Age, TC, Tg |
| CD-SBS- SVM | Age, Troponin, Creatinine, eGFR, CK, HbA1c, FBS |
| LVQ- SBS- SVM | Tg, Troponin, Creatinine, HDL, Stroke, eGFR, CK, Ethnicity |
| GA- SBS- SVM | Tg, Troponin, Ethnicity, FBS, Newly dm, Hypertension, Ex-smoker, TC Simon Broome, LCL Simon Broome, LCx coronary angiogram |
| ELASTIC NET | |
| EN | All variables |
| RFVarImp-SBS-EN | Age, TC, Tg, Troponin, creatinine, HDL, eGFR, CK |
| SVMVarImp – SBS – EN | TC, Tg, Troponin, CK |
| ENVarImp-SBS – EN | Troponin, Creatinine, Stroke, CKD |
| LRVarImp-SBS- EN | TC, HbA1c, COAD, ASA |
| RFE-SBS – EN | Age, TC, Tg |
| Boruta -SBS- EN | Age, TC, Tg |

| | |
|---|---|
| CD-SBS- EN | Age, Troponin, Creatinine, eGFR, CK, HbA1c, FBS |
| LVQ- SBS- EN | Tg, Troponin, Creatinine, HDL, Stroke, eGFR, CK, Ethnicity |
| GA- SBS- EN | Tg, Troponin, Ethnicity, FBS |
| LOGISTIC REGRESSION | |
| LR | All variables |
| RFVarImp-SBS-LR | Age, TC, Tg, Troponin, Creatinine |
| SVMVarimp – SBS – LR | TC, Tg, Troponin, CK |
| ENVarImp-SBS –LR | Troponin, creatinine, HDL, CK, Stroke, CKD, CCF, RCA coronary angiogram |
| LRVarImp-SBS-LR | TC, HbA1c, COAD, ASA |
| RFE-SBS – LR | Age, TC, Tg |
| Boruta -SBS- LR | Age, TC, Tg |
| CD-SBS- LR | Age, Troponin, Creatinine, eGFR, CK, HbA1c, FBS |
| LVQ- SBS- LR | Tg, Troponin, Creatinine, Stroke |
| GA- SBS- LR | Tg, Troponin, Ethnicity, FBS |

RFVarImp: Random forest variable importance ranking; SVMVarImp: Support vector machine variable importance ranking; ENVarImp: Elastic net variable importance ranking; LRVarImp: Logistic regression variable importance ranking; SBS: sequential backward selection

Table 3 provides additional results on model performance. The best-performing models were RFVarImp-SBS-RF model (AUC = 0.79; Five predictors: age, TC, Tg, troponin, and creatinine) and RFE-RF model (AUC = 0.78; Five predictors: age, TC, Tg, troponin, and stroke). There were no significant differences in performance between the two models ($p > 0.05$). Combinations of feature selection methods with classifiers such as LR, EN reported the lowest performance value. The sensitivity and specificity in Table 3 are based on the 0.5 cut-off point for comparison. The AUC provides model performance on average for various cut-off points, but at 0.5 cut-off point, some lower AUC models perform better. However, higher AUC models are preferred in mortality risk studies.

TABLE 3. Additional performance measures each machine learning model

| Model | Sense/Spec | PPV/NPV | Detection rate | Detection incidence | AUC | Accuracy (95% CI) |
|---|---|---|---|---|---|---|
| RF | 0.17/0.98 | 0.33/0.94 | 0.01 | 0.03 | 0.62 | 0.92(0.84,0.97) |
| RFimp - SBS-RF | 0.33/0.94 | 0.28/0.95 | 0.02 | 0.08 | 0.79 | 0.89 (0.82,0.95) |
| SVMimp-SBS-RF | 0.17/0.87 | 0.08/0.93 | 0.01 | 0.13 | 0.75 | 0.82(0.72,0.89) |
| ENimp-SBS-RF | 0.33/0.92 | 0.22/0.95 | 0.02 | 0.10 | 0.61 | 0.88(0.79,0.94) |
| LRimp-SBS-RF | 0.17/0.90 | 0.11/0.94 | 0.01 | 0.10 | 0.71 | 0.85(0.76,0.92) |
| RFE-RF | 0.50/0.87 | 0.23/0.96 | 0.03 | 0.15 | 0.78 | 0.85(0.76, 0.92) |
| Boruta -SBS-RF | 0.00/0.95 | 0.00/0.93 | 0.00 | 0.04 | 0.68 | 0.89(0.80,0.94) |
| CD-SBS-RF | 0.33/0.95 | 0.33/0.95 | 0.02 | 0.07 | 0.69 | 0.91 (0.83,0.96) |
| LVQ-SBS-RF | 0.17/0.90 | 0.11/0.94 | 0.01 | 0.10 | 0.70 | 0.53 (0.76,0.92) |
| GA-SBS-RF | 0.17/0.84 | 0.07/0.93 | 0.01 | 0.16 | 0.75 | 0.80(0.70,0.87) |

| | | | | | | |
|---|---|---|---|---|---|---|
| SVM | 0.00/0.99 | 0.00/0.93 | 0.01 | 0.00 | 0.73 | 0.92(0.84,0.97) |
| RFimp-SBS-SVM | 0.50/0.67 | 0.10/0.95 | 0.03 | 0.34 | 0.63 | 0.66(0.55,0.76) |
| SVMimp-SBS-SVM | 0.00/.987 | 0.00/0.93 | 0.00 | 0.01 | 0.73 | 0.92(0.84,0.97) |
| ENimp-SBS-SVM | 0.50/0.84 | 0.19/0.96 | 0.03 | 0.18 | 0.68 | 0.82(0.72,0.89) |
| LRimp-SBS-SVM | 0.17/0.8 | 0.06/0.93 | 0.01 | 0.19 | 0.69 | 0.76(0.66,0.85) |
| RFE-SVM | 0.17/0.67 | 0.04/0.92 | 0.01 | 0.31 | 0.77 | 0.64(0.53, 0.74) |
| Boruta -SBS-SVM | 0.17/0.64 | 0.03/0.91 | 0.01 | 0.35 | 0.76 | 0.61(0.50,0.71) |
| CD-SBS-SVM | 0.17/0.69 | 0.04/0.92 | 0.01 | 0.30 | 0.68 | 0.65(0.54,0.75) |
| LVQ-SBS-SVM | 0.17/0.84 | 0.07/0.93 | 0.01 | 0.16 | 0.67 | 0.80 (0.70,0.86) |
| GA-SBS-SVM | 0.00/0.53 | 0.00/0.88 | 0.00 | 0.44 | 0.73 | 0.80(0.70,0.87) |
| EN | 0.17/0.82 | 0.06/0.93 | 0.01 | 0.18 | 0.52 | 0.76(0.67,0.85) |
| RFimp-SBS-EN | 0.67/0.53 | 0.09/0.96 | 0.04 | 0.48 | 0.64 | 0.54(0.43,0.65) |
| SVMimp-SBS-EN | 0.00/0.70 | 0.00/0.91 | 0.00 | 0.28 | 0.76 | 0.65(0.54,0.75) |
| ENimp-SBS-EN | 0.50/0.75 | 0.13/0.95 | 0.03 | 0.27 | 0.62 | 0.73(0.63,0.82) |
| LRimp-SBS-EN | 0.50/0.64 | 0.09/0.95 | 0.03 | 0.37 | 0.57 | 0.63(0.52,0.73) |
| RFE-EN | 0.50/0.60 | 0.08/0.94 | 0.03 | 0.40 | 0.62 | 0.60(0.49,0.70) |
| Boruta -SBS-EN | 1.00/0.24 | 0.09/1.00 | 0.07 | 0.78 | 0.66 | 0.29(0.20,0.40) |
| CD-SBS-EN | 0.50/0.54 | 0.07/0.94 | 0.03 | 0.46 | 0.54 | 0.54(0.43,0.65) |
| LVQ-SBS-EN | 0.50/0.76 | 0.13/0.95 | 0.03 | 0.26 | 0.62 | 0.74(0.64,0.83) |
| GA-SBS-EN | 0.00/0.94 | 0.00/0.92 | 0.00 | 0.06 | 0.79 | 0.87(0.78,0.93) |
| LR | 0.33/0.81 | 0.11/0.94 | 0.02 | 0.20 | 0.61 | 0.78(0.67,0.85) |
| RFimp-SBS-LR | 0.29/0.76 | 0.09/0.93 | 0.02 | 0.24 | 0.55 | 0.72(0.62,0.81) |
| SVMimp-SBS-LR | 0.50/0.72 | 0.12/0.95 | 0.03 | 0.29 | 0.56 | 0.71(0.61,0.80) |
| ENimp-SBS-LR | 0.17/0.87 | 0.08/0.94 | 0.01 | 0.13 | 0.48 | 0.82(0.72,0.89) |
| LR-SBS-LR | 0.17/0.59 | 0.03/0.91 | 0.01 | 0.39 | 0.63 | 0.56(0.45,0.66) |
| RFE-LR | 0.00 /0.80 | 0.00/0.92 | 0.00 | 0.19 | 0.74 | 0.74 (0.64,0.83) |
| Boruta -SBS-LR | 0.83/0.54 | 0.12/0.98 | 0.06 | 0.48 | 0.66 | 0.56(0.45,0.67) |
| CD-SBS-LR | 0.17/0.76 | 0.05/0.93 | 0.01 | 0.24 | 0.51 | 0.72(0.61,0.81) |
| LVQ-SBS-LR | 0.00/0.92 | 0.00/0.93 | 0.00 | 0.08 | 0.75 | 0.85(0.76,0.92) |
| GA-SBS-LR | 0.57/0.60 | 0.11/0.94 | 0.04 | 0.41 | 0.60 | 0.60(0.49,0.70) |

Theoretically, for a survival model, for one new patient at the time of the first ACS, the best ML model RFVarImp-SBS-RF (5 predictors), the average mortality risk is reduced to 4.9% (NPV). If the model outcome is non-survival, the average risk of a patient being deceased is increased to 28.3% (PPV), which represents an average risk ratio of 5.9 (ratios of non-survivors to survivors) by this model. Additional dataset of 102 patients including predictors: age, TC, Tg, troponin, and creatinine were tested and compared to the TIMI score. The additional test performed for the comparative purpose for the ML model recorded an AUC value of 0.75 vs. the TIMI score with an AUC value of 0.60.

The SOM map was used to investigate and visualise the association between predictors and mortality for the best model (RFVarImp-SBS-RF). SOM map performance reported quantisation and topographic error of 0.150 and 0.056, respectively. Figure 5 illustrates the coloured scale of the U-matrix cluster map symbolising vector distances (predictors are vector elements). The blue colour represents the minimum distance (create clusters of vectors with similar features), whereas the red colour represents the maximum distance (the vectors are dissimilar). Component planes are represented by the predictors by warm colours corresponding to high mean values and vice versa. Non-survivals are explained in older patients (> 65 years of age) with a high troponin value (>11.4 ng/L). Survivors are categorised: Older patients (>65 years of age) who survived post-ACS, with the lower troponin value (~0.5 ng/L) and higher creatinine value (>138 μmol/L); younger patients (< 55 years of age) with lower creatinine value (<138 μmol/L) irrespective of Tg, TC and troponin values.

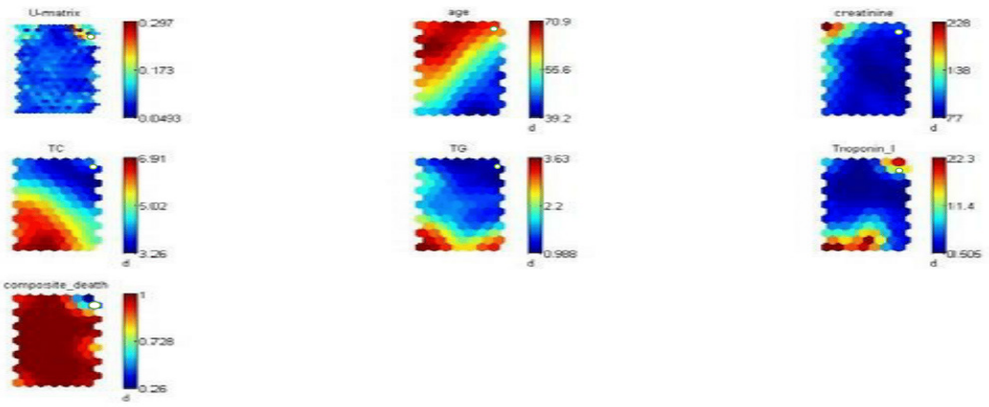


FIGURE 5. The SOM represents the association of selected variables with post ACS mortality

In the case of mortality prediction, AUC is an important performance measure as it helps to evaluate a model's performance, regardless of the decision boundary chosen. Although some models in this study showed a better result on sensitivity and specificity at the decision boundary of 0.5, lower AUC was reported for overall model performance. High performances of AUC > 0.75 for a testing dataset, which was not used for model development, were reported for 8 out of 36 models in this study. RFVarImp-SBS-RF classifier model and RFE-RF model performance in this study using five predictors were similar to the ML model performance recorded in other mortality-related studies (Shouval et al. 2017; Steele et al. 2018; Tuckova 2013; Wallert et al. 2017). The best model, RFVarImp-SBS-RF, when tested against TIMI risk scores using an additional dataset, outperformed TIMI (AUC = 0.75 vs. AUC = 0.65), which is consistent with findings from population-specific mortality studies comparing TIMI and ML (Shouval et al. 2017; Wallert et al. 2017).

The hybrid combination of SVM and RF resulted in higher performance compared to LR and EN models. Combination of a feature selection method with classification algorithms reported higher performances in the literature (Mokeddem et al. 2013; Perez-Riverol et al. 2017; Sonawane et al. 2014). The RFE feature selection method used in various clinical datasets performs better when combined with the ML classifier, especially SVM

and RF (Chopra et al. 2017; Lin et al. 2017; Perez-Riverol et al. 2017; Yang et al. 2017). Boruta, GA, CD and RFE have been reported to achieve higher performance when combined with ML classifiers such as SVM and RF (Alalyan et al. 2019; Chopra et al. 2017; Galili et al. 2015; Prokashgoswami et al. 2013; Zhang et al. 2017).

In this study, all models improved with parameter optimisation *via* feature selection. The application of feature selection algorithms improves model performance using a reasonable number of predictors by reducing predictor's dimensionality (Yang et al. 2018). RF and SVM with standard implementation and default parameters outperformed LR due to parameter optimisation that enables the algorithm to adjust the data information to increase the predictive model performance (Couronné et al. 2018; Fernández-Delgado et al. 2014; Huang et al. 2016; Liu et al. 2017).

The univariate analysis indicates the relationship between the predictor selected from ML algorithms and the outcome that is vague to the clinician. The univariate analysis (Table 1) also illustrates the significance of the selected variables. The best ML model in this study selected five predictors: age, Tg, creatinine, troponin, and TC. Tg, troponin, and TC were among the high-ranking variables chosen for all models. In this study, various feature selection algorithms selected a different combination of predictors for 30-day mortality predictions post-ACS. Model-specific predictors that were highly ranked include age, HbA1c, FBS, CK, eGFR, creatinine, ethnicity, ACS subtype, and stroke history. Levels of FBS and HbA1c, especially in non-diabetics, are related to increased risk of ACS-related mortality (Liang et al. 2016). Glucose levels support the relationship between hyperglycaemia and increased risk of mortality in patients with STEMI in the Asian population (Johansson et al. 2017). Risk factors leading to worse post-MI outcomes include comorbid diabetes, hypertension, older age, reduced renal function, and stroke history (Wu et al. 2018).

We have also demonstrated that clinical data can be visualised in a 2-dimensional representation using SOM to understand the association of predictors with mortality. This allows a clinician to place a new patient within the context of previous or similar cases if there is confidence in the original training data. The results of the SOM technique prove its ability to perform with the lowest quantisation and topographic errors (Zhou 2010).

Non-survivals from the SOM map were related to higher troponin levels in older patients. Meanwhile, survival in the same age group was related to lower levels of troponin, Tg, TC, and creatinine. Cardiac troponin levels have been an independent predictor of all-cause mortality. The prognostic importance of troponin must be recognised with the patient's age. Higher mortality has been associated with patients age 65 years and older with troponin < 0.01 ng/mL for troponin I and T (Cheng et al. 2015). Younger patients (< 55 years of age) with low levels of creatinine, regardless of Tg, TC, and troponin, have been associated with survival, indicating that older age and creatinine levels had a considerable unfavourable consequence on mortality (Marenzi et al. 2015).

High troponin level and age have also been associated with non-survival. Ageing is a significant predictor of mortality in ACS and an independent risk factor for adverse outcomes post-ACS (Engberding et al. 2017). Age has been selected as a factor that affects mortality post-STEMI by ML models in previous studies (Shouval et al. 2017; Wallert et al. 2017). Older patients usually have more complex cardiovascular disease, more comorbidities, and more atypical clinical presentation (Engberding et al. 2017). Elderly patients had a higher burden of cardiovascular risk factors in this study. Compared to younger patients, more elderly patients had a history of diabetes mellitus, hypertension, dyslipidaemia, stroke, ischemic heart disease, chronic obstructive airways disease, bronchial asthma, and chronic kidney disease. However, what is interesting to note is that creatinine in the younger population is a better predictor of survival compared to TC, Tg, or troponin. Troponin, which is a risk marker for survivors in the Caucasian population, has been tested to be a better predictor compared to other comorbid diseases. From SOM, we discovered that troponin is identified as a risk predictor for this population, generating a new hypothesis that the choice of risk predictors should be population specific.

This present study chose age, creatinine, and cardiac markers for the best model similar to GRACE and TIMI. Variables related to medical history, although not selected during feature selections, were considered insignificant. The data on the vital signs at admission and the ECG findings were not available for the current study, which is considered a limitation and will be implemented for future research.

TIMI scores were derived from Western Caucasian cohort. These models are population-specific and may not be capable of taking into account nuances related to a specific region. In this study, TIMI comparisons were made only on the best-selected model due to the limited dataset that allows this comparison. ML using additional datasets outperformed TIMI risk score in this study. It is important to recognise the significant features affecting the population-specific mortality rate in ACS patients in order to achieve a reliable and valid clinical diagnosis. Prediction of the future health status of a patient may be a significant part of the medical sector, as it can promote early detection of diseases, effective treatment, disease prevention, and patients with high-risk can also be distinguished, and appropriate measures can be taken.

ML, especially RF, can handle a noisy dataset. RF has a built-in variable importance method that enables numerous input variables without having to delete certain variables for reduced dimensionality. Variable importance generates RF scores by measuring the increase in the prediction error (Kesavaraj et al. 2013; Zhang et al. 2011).

The limited number of datasets was the limitation of this study. However, ML methods allow classifications involving a high dimensional dataset (p) with a low number of samples (n) (Shaikhina et al. 2019).

The study allows the evaluation of a hybrid combination of various feature selection methods (filter, wrapper, and embedded) with predictive machine learning methods that enable the development of a population-specific module that outperforms the conventional TIMI risk score method. The models developed may be useful as complementary decision support tools used in conjunction with the traditional risk scoring method for improving patient health. The application of the SOM method allows the visualisation of the association among various mortality risk factors that have not been reported in any other mortality-related studies. The high performance of the RF model was achieved by dimensionality reduction of the variables using a feature selection method that enables model interpretation using the SOM method from a clinical point of view.

Future work will look into larger datasets obtained from the Malaysian National Cardiovascular Disease Registry that can be used to for continuous external model validation and improvement using the most recent high-quality data with the possibility of developing models that focus on a specific type of ACS such as STEMI and NSTEMI mortality. This work will be used as a basis for selecting the best combination of feature selection and ML classifiers for the online population-specific risk calculator.

## CONCLUSION

In conclusion, we have demonstrated the ability to apply a hybrid combination of methodologies for feature selection and 30-day mortality predictions in ACS patients. A combination of applying RF variable importance along with the SBS technique for variable ranking and selection technique proves that it is an effective method for selecting significant variables and prediction. It is evident from this work that it is possible to create a compressed data representation that can be used as a tool where the abundance of data obscures the straightforward diagnostic reasoning in the ACS-related mortality study. A conclusion is drawn that the use of such a map in conjunction with the presentation of ACS-related mortality can be a useful screening mechanism for detecting patients with a high risk of ACS. At this stage, it is not possible to claim that the results of this study are universally applicable. Since the study is based on limited clinical data, more data is needed to improve and validate the system. Once developed, it has great potential as a risk assessment tool that enables a consensus to be achieved between clinicians on the risk stratification of ACS patients. Only then can we develop a useful tool for placing a patient within a clinical context, thereby allowing a consensus to be achieved between clinicians and assessing the particular risk to the patient.

## ACKNOWLEDGEMENTS

The research is supported and funded by University of Malaya (Project number: GPF013B-2018).

## REFERENCES

Alalyan, F., Zamzami, N. & Bouguila, N. 2019. Model-based hierarchical clustering for categorical data. In *IEEE 28th International Symposium on Industrial Electronics (ISIE)*. Vancouver, Canada: IEEE. pp. 1424-1429. doi: 10.1109/ISIE.2019.8781307.

Breiman, L. 2001. Using iterated bagging to debias regressions. *Machine Learning* 45(3): 261-277. https://doi.org/10.1023/a:1017934522171.

Castro-Dominguez, Y., Dharmarajan, K. & Mcnamara, R.L. 2018. Predicting death after acute myocardial infarction. *Trends in Cardiovascular Medicine* 28(2): 102-109. https://doi.org/10.1016/j.tcm.2017.07.011.

Chandrashekar, G. & Sahin, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40(1): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chen, X. & Ishwaran, H. 2012. Random forests for genomic data analysis. *Genomics* 99(6): 323-329. https://doi.org/10.1016/j.ygeno.2012.04.003.

Cheng, J.M., Helming, A.M., Vark, L.C.V., Corstiaan, I.K., Uil, A.D., Jewbali, L.S., van Geuns, R., Zijlstra, F., van Domburg, R.T., Boersma, E. & Akkherhuis, K.M. 2015. A simple risk chart for initial risk assessment of 30-day mortality in patients with cardiogenic shock from ST-elevation myocardial infarction. *European Heart Journal: Acute Cardiovascular Care* 5(2): 101-107. https://doi.org/10.1177/2048872615568966.

Chopra, A., Dimri, A. & Pradhan, T. 2017. Prediction of factors affecting amlodipine induced pedal edema and its classification. In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Udupi, India: DBLP. pp. 1684-1689. https://doi.org/10.1109/icacci.2017.8126085.

Collazo, R.A., Pessôa, L.A.M., Bahiense, L., Pereira, B.D.B., Reis, A.F.D. & Silva, N.S.E. 2016. A comparative study between artificial neural network and support vector machine for acute coronary syndrome prognosis. *Pesquisa Operacional* 36(2): 321-343. https://doi.org/10.1590/0101-7438.2016.036.02.0321.

Couronné, R., Probst, P. & Boulesteix, A. 2018. Random forest versus logistic regression: A large-scale benchmark

experiment. *BMC Bioinformatics* 19(1): 270. https://doi.org/10.1186/s12859-018-2264-5.

Cox, D.R. 1958. Two further applications of a model for binary regression. *Biometrika* 45(3-4): 562-565. https://doi.org/10.1093/biomet/45.3-4.562.

Dunkler, D., Plischke, M., Leffondré, K. & Heinze, G. 2014. Augmented backward elimination: A pragmatic and purposeful way to develop statistical models. *PLoS ONE* 9(11): e113677. https://doi.org/10.1371/journal.pone.0113677.

Engberding, N. & Wenger, N.K. 2017. Acute coronary syndromes in the elderly. *F1000Research* 6: 1791. https://doi.org/10.12688/f1000research.11064.1.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27(8): 861-874. https://doi.org/10.1016/j.patrec.2005.10.010.

Fernández-Delgado, M., Eva, C., Senén, B. & Dinani, A. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15: 3133-3181.

Galili, Tal. 2015. Dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31(22): 3718-3720. https://doi.org/10.1093/bioinformatics/btv428.

Geisser, S. 1993. *Predictive Inference: An Introduction*. London: Chapman and Hall. http://dx.doi.org/10.1007/978-1-4899-4467-2.

Genuer, R., Poggi, J. & Tuleau-Malot, C. 2010. Variable selection using random forests. *Pattern Recognition Letters* 31(14): 2225-2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Hammer, B. & Villmann, T. 2002. Generalized relevance learning vector quantization. *Neural Networks* 15(8-9): 1059-1068. https://doi.org/10.1016/s0893-6080(02)00079-5.

Hinde, C.J. 2003. Extracting causal nets from databases. In *Developments in Applied Artificial Intelligence Lecture Notes in Computer Science, IEA/AIE 2003, Lecture Notes in Computer Science*. pp. 166-175. https://doi.org/10.1007/3-540-45034-3_17.

Holland, J.H. 1992. Genetic algorithms. *Scientific American* 267(1): 66-72. https://doi.org/10.1038/scientificamerican0792-66.

Hoo, F.K., Boo, Y.L., Foo, Y.L., Mohd, S., Lim, S. & Ching, S.M. 1969. Acute coronary syndrome in young adults from a Malaysian tertiary care centre. *Pakistan Journal of Medical Sciences* 32(4): 841-845. https://doi.org/10.12669/pjms.324.9689.

Huang, B.F.F. & Boutros, P.C. 2016. The parameter sensitivity of random forests. *BMC Bioinformatics* 17: 331. https://doi.org/10.1186/s12859-016-1228-x.

Jafarian, A., Ngom, A. & Rueda, L. 2011. A novel recursive feature subset selection algorithm. In *IEEE 11th International Conference on Bioinformatics and Bioengineering*. Taichung, Taiwan: IEEE. pp. 78-83. https://doi.org/10.1109/bibe.2011.19.

Johansson, S., Rosengren, A., Young, K. & Jennings, E. 2017. Mortality and morbidity trends after the first year in survivors of acute myocardial infarction: A systematic review. *BMC Cardiovascular Disorders* 17(1): 53. https://doi.org/10.1186/s12872-017-0482-9.

Kesavaraj, G. & Sukumaran, S. 2013. A study on classification techniques in data mining. In *Fourth International Conference on Computing, Communications and Networking Technologies.* Tiruchengode, India: IEEE. pp. 1-7. https://doi.org/10.1109/icccnt.2013.6726842.

Kohonen, T. 2001. Self-organizing maps. In *Springer Series in Information Sciences*. Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-56927-2.

Kohonen, T. 2001. Learning vector quantization. In *Self-Organizing Maps Springer Series in Information Sciences*. Berlin, Germany: Springer. pp. 245-261. https://doi.org/10.1007/978-3-642-56927-2_6.

Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28(5): 1-26. https://doi.org/10.18637/jss.v028.i05.

Kursa, M.B. & Rudnicki, W.R. 2010. Feature selection with the boruta package. *Journal of Statistical Software* 36(11): 1-13. https://doi.org/10.18637/jss.v036.i11.

Liang, H., Guo, Y.C., Chen, L.M., Li, M., Han, W.Z., Zhang, X. & Jiang, S.L. 2016. Relationship between fasting glucose levels and in-hospital mortality in Chinese patients with acute myocardial infarction and diabetes mellitus: A retrospective cohort study. *BMC Cardiovascular Disorders* 16: 156. https://doi.org/10.1186/s12872-016-0331-2.

Lin, X., Li, C., Zhang, Y., Su, B., Fan, M. & Wei, H. 2017. Selecting feature subsets based on svm-rfe and the overlapping ratio with applications in bioinformatics. *Molecules* 23(1): 52. https://doi.org/10.3390/molecules23010052.

Liu, C.H., Bryan, B.P.C., Little, D.A. & Cardoso, A. 2017. Generalising random forest parameter optimisation to include stability and cost. *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science* 10536: 102-113. https://doi.org/10.1007/978-3-319-71273-4_9.

Malek, S., Gunalan, R., Kedija, S.Y., Lau, C.F., Mogeeb, A.A., Milow, M.P., Lee, S.A. & Saw, A. 2018. Random forest and self-organizing maps application for analysis of pediatric fracture healing time of the lower limb. *Neurocomputing* 272: 55-62. https://doi.org/10.1016/j.neucom.2017.05.094.

Mandrekar, J.N. 2010. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* 5(9): 1315-1316. https://doi.org/10.1097/jto.0b013e3181ec173d.

Marenzi, G., Cabiati, A., Cosentino, N., Assanelli, E., Milazzo, V., Rubino, M., Lauri, G., Morpurgo, M., Moltrasio, M., Marana, I., Metrio, M.D., Bonomi, A., Veglia, F. & Bartorelli, A. 2015. Prognostic significance of serum creatinine and its change patterns in patients with acute coronary syndromes. *American Heart Journal* 169(3): 363-370. https://doi.org/10.1016/j.ahj.2014.11.019.

Menard, S. 2002. *Applied Logistic Regression Analysis*. 2nd ed. USA: SAGE Publishing. https://doi.org/10.4135/9781412983433.

Mokeddem, S., Atmani, B. & Mokaddem, M. 2013. Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. In *Computer Science & Information Technology (CS & IT)*. Dubai, UAE: DDBM. pp. 41-52. https://doi.org/10.5121/csit.2013.3305.

Motwani, M., Dey, D., Berman, D.S., Germano, G., Achenbach, S., Al-Mallah, M.H., Andreini, D., Budoff, M.J., Cademartini, F., Callister, T.Q., Chang, H.J., Chinnaiyan, K., Chow, B.J.W., Cury, B.C., Delago, A., Gomez, M., Gransar, H., Hadamitzky, M., Hausleiter, J., Hindoyan, N., Feuchtner, G., Kaufmann, P.A., Kim, Y.J., Leipsic, J., Lin, F.Y., Maffei, E., Marques, H., Pantone, G., Raff, G., Rubinshtein, R., Shaw, L.J., Stehli, J., Villines, T.C., Duniing, A., Min, J.K. & Slomka, P.J. 2016. 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* 38(7): 500-507. https://doi.org/10.1093/eurheartj/ehw188.

Perez-Riverol, Y., Kuhn, M., Vizcaíno, J.A., Hitz, M. & Audain, E. 2017. Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS ONE* 12(12): e0189875. https://doi.org/10.1371/journal.pone.0189875.

Prokashgoswami, J. & Mahanta, A.J. 2013. Categorical data clustering based on an alternative data representation technique. *International Journal of Computer Applications* 72(5): 7-12. https://doi.org/10.5120/12488-8301.

Saeys, Y., Inza, I. & Larranaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19): 2507-2517. https://doi.org/10.1093/bioinformatics/btm344.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R. & Khovanova, N. 2019. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control* 52: 456-462. https://doi.org/10.1016/j.bspc.2017.01.012.

Shouval, R., Hadanny, A., Shlomo, N., Iakobishvili, Z., Unger, R., Zahger, D., Alcalai, R., Atar, S., Gottlieb, S., Matetzky, S., Goldenberg, I. & Beigel, R. 2017. Machine learning for prediction of 30-day mortality after ST elevation myocardial infraction: An acute coronary syndrome Israeli survey data mining study. *International Journal of Cardiology* 246: 7-13. https://doi.org/10.1016/j.ijcard.2017.05.067.

Sonawane, J.S. & Patil, D.R. 2014. Prediction of heart disease using learning vector quantization algorithm. In *Conference on IT in Business, Industry and Government (CSIBIG)*. Indore, India: IEEE Xplore. https://doi.org/10.1109/csibig.2014.7056973.

Steele, A.J., Denaxas, S.C., Shah, A.D., Hemingway, H. & Luscombe, N.M. 2018. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS ONE* 13(8): e0202344. https://doi.org/10.1371/journal.pone.0202344.

Torres, M. & Moayedi, S. 2007. Evaluation of the acutely dyspneic elderly patient. *Clinics in Geriatric Medicine* 23(2): 307-325. https://doi.org/10.1016/j.cger.2007.01.007.

Tuckova, J. 2013. The possibility of kohonen self-organizing map applications in medicine. In *IEEE 11th International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics*. France: IEEE. pp. 1-6. https://doi.org/10.1109/ecmsm.2013.6648946.

Vapnik, V. 1998. The support vector method of function estimation. In *Nonlinear Modeling.* Boston, MA: Springer. pp. 55-85. https://doi.org/10.1007/978-1-4615-5703-6_3.

Wallert, J., Tomasoni, M., Madison, G. & Held, C. 2017. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Medical Informatics and Decision Making* 17(1): 99. https://doi.org/10.1186/s12911-017-0500-y.

Wu, C., Singh, A., Collins, B., Fatima, A., Qamar, A., Gupta, A., Hainer, J., Klein, J., Jarolim, P., Carli, M.D., Nasir, K., Bhatt, D.L. & Blankstein, R. 2018. Causes of troponin elevation and associated mortality in young patients. *The American Journal of Medicine* 131(3): 284-292. https://doi.org/10.1016/j.amjmed.2017.10.026.

Yang, J., Li, X., Chen, T., Li, Y., Xie, G. & Yang, Y. 2018. Machine learning models to predict in-hospital mortality for ST-elevation myocardial infraction: From China acute myocardial infarction (cami) registry. *Journal of the American College of Cardiology* 71(11): A236. https://doi.org/10.1016/s0735-1097(18)30777-0.

Yang, X. 2017. Identification of risk genes associated with myocardial infarction based on the recursive feature elimination algorithm and support vector machine classifier. *Molecular Medicine Reports* 17(1): 1555-1560. https://doi.org/10.3892/mmr.2017.8044.

Zhang, L. & Lin, X. 2011. Some considerations of classification for high dimension low-sample size data. *Statistical Methods in Medical Research* 22(5): 537-550. https://doi.org/10.1177/0962280211428387.

Zhang, Z., Murtagh, F., Poucke, S.V., Lin, S. & Lan, P. 2017. Hierarchical cluster analysis in clinical research with heterogeneous study population: Highlighting its visualization with R. *Annals of Translational Medicine* 5(4): 75. https://doi.org/10.21037/atm.2017.02.05.

Zhou, X. 2010. Enhancement of topology preservation of self-organizing map. *Journal of Computer Applications* 29(12): 3256-3258. https://doi.org/10.3724/sp.j.1087.2009.03256.

Zou, H. & Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Nanyonga Aziida, Sorayya Malek* & Firdaus Aziz
Bioinformatics Division
Institute of Biological Sciences
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

Khairul Shafiq Ibrahim & Sazzli Kasim
Department of Cardiology
Faculty of Medicine
Universiti Teknologi MARA (UiTM)
Sungai Buloh Campus, Jalan Hospital
47000 Sungai Buloh, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: sorayya@um.edu.my

Received: 23 December 2019
Accepted: 26 August 2020