Medical & Biological Engineering & Computing (2020) 58:3123–3140
https://doi.org/10.1007/s11517-020-02268-9

ORIGINAL ARTICLE

Feature selection and risk prediction for patients with coronary artery disease using data mining

Nashreen Md Idris[1] · Yin Kia Chiam[1] · Kasturi Dewi Varathan[2] · Wan Azman Wan Ahmad[3] · Kok Han Chee[3] · Yih Miin Liew[4]

Received: 16 September 2019 / Accepted: 8 September 2020 / Published online: 6 November 2020

Abstract
Coronary artery disease (CAD) is an important cause of mortality across the globe. Early risk prediction of CAD would be able to reduce the death rate by allowing early and targeted treatments. In healthcare, some studies applied data mining techniques and machine learning algorithms on the risk prediction of CAD using patient data collected by hospitals and medical centers. However, most of these studies used all the attributes in the datasets which might reduce the performance of prediction models due to data redundancy. The objective of this research is to identify significant features to build models for predicting the risk level of patients with CAD. In this research, significant features were selected using three methods (i.e., Chi-squared test, recursive feature elimination, and Embedded Decision Tree). Synthetic Minority Over-sampling Technique (SMOTE) oversampling technique was implemented to address the imbalanced dataset issue. The prediction models were built based on the identified significant features and eight machine learning algorithms, utilizing Acute Coronary Syndrome (ACS) datasets provided by National Cardiovascular Disease Database (NCVD) Malaysia. The prediction models were evaluated and compared using six performance evaluation metrics, and the top-performing models have achieved AUC more than 90%.

Keywords Data mining · Prediction model · Classification algorithms · Feature selection · Heart disease prediction · Coronary artery disease

1 Introduction

According to the World Health Organization, coronary artery disease (CAD) has stayed in the charts of the global top 10 causes of death for over the past 15 years [1]. Malaysia is not exempted as ischemic heart disease has remained the principal cause of death in 2016 among Malaysian by dominating 13.5% of the chart [2]. Patients with CAD may be presented clinically with stable angina or ACS. ACS is a spectrum of diseases ranging from unstable angina, non–ST-elevation myocardial infarction (NSTEMI) to ST-elevation myocardial infarction (STEMI) depending on the acuteness and severity of the coronary occlusion [3].

Conditions that have the potential to increase the risk of developing CAD are known as risk factors [4]. Examples of risk factor contributing to CAD are sociodemographic and socioeconomic factors (e.g., age, gender, occupation), genetic factors (e.g., family history of CAD), comorbidities, anthropometric measurement and biochemical markers (e.g., diabetes, hypertension, obesity), lifestyle factor (e.g., dietary intake, stress), and environment factors (e.g., passive smoking) [5]. In order to give effective treatment and a quality healthcare service, accurate and timely diagnosis of the patient and evaluation of risk level are required [6–8]. However, it remains a challenge for healthcare organizations to improve the assessment of risk level for patients with CAD amid a wide range of known and unknown risk factors.

✉ Yin Kia Chiam
yinkia@um.edu.my

[1] Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[2] Department of Information Systems, Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

[3] Department of Medicine, University Malaya Medical Centre, 50603 Kuala Lumpur, Malaysia

[4] Department of Biomedical Engineering, Faculty of Engineering, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Traditionally, CAD is diagnosed by doctors based on patients' symptoms, records, and risk factors, and the prediction was always a result of doctors' intuition and experience rather than a quantitative concoction of information extracted from the clinical data [9–11]. The judgement is subjective and interobserver variability is high. Objective and systematic prediction of the actual risk level of the patient is crucial to reduce such variability and to facilitate clinician in formulating appropriate treatment strategies for the patients.

In healthcare, data mining has been used for predicting disease and medication, as well as measuring the effectiveness of certain treatments given to patients [10, 12]. Applying data analytics in clinical data to improve the overall healthcare in Malaysia and to reduce the population mortality from CAD is still very much lacking [13], and to our knowledge, has not been implemented on the NCVD-ACS datasets gathered from hospitals in Malaysia. By implementing data analytics in clinical data of CAD patients, it can help the doctors to practice evidence-based medicine, and to make a better decision based on the clinical information.

Furthermore, healthcare datasets are not without issues such as missing values, irrelevant and redundant features that would lower the quality of a prediction model [14]. Significant research on feature selection method has therefore been carried out to overcome these issues by allowing the selection of significant attributes which are necessary for the prediction of the target class [14–16]. Keeping only significant attributes would also save time, effort and reduce the complexity of the machine learning algorithms [9, 16, 17].

In this research, we implemented three types of feature selection methods which are the filter method (Chi-squared test), wrapper method (Recursive Feature Elimination using Logistic Regression), and embedded method (Embedded Decision Tree using random forest algorithm) to select features that contributed the most to the risk prediction of patients with ACS from the NCVD-ACS dataset of a hospital in Malaysia. After selecting the significant features, classification models were developed using eight machine learning algorithms to predict the risk level based on Killip class: Logistic Regression (LR), Neural Network (NN), k-nearest neighbors algorithm (kNN), Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), Deep Learning (DL), and Vote (an ensemble technique with Naïve Bayes and Logistic Regression). The performance of prediction models was then tested against NCVD-ACS dataset from another hospital.

The rest of the paper is organized as follows. Section 2 describes the two NCVD-ACS datasets used in this research. Section 3 explains the methods for experiments involving data pre-processing, feature selection, classification modelling, using machine learning algorithms, and performance measure. Section 4 presents the selected features, performance results of the prediction models created using NCVD-ACS dataset from a hospital in Malaysia, and the evaluation conducted to validate the risk prediction models using NCVD-ACS dataset from another hospital in Malaysia. Section 5 provides the discussion on the performance comparison between various risk prediction models for both datasets and the significant features identified by the feature selection methods. Finally, the last section concludes the study and presents future work.

2 Dataset

The Malaysian National Cardiovascular Disease Database (NCVD)–ACS registry was established in 2006 and jointly published by the National Heart Association of Malaysia (NHAM) and the Ministry of Health Malaysia [18]. Over the past 10 years, the NCVD-ACS registry had collected and reported on 49,406 ACS patients, making it the biggest database on ACS in Malaysia [19]. In total, datasets from 2 hospitals were extracted and utilized in current research.

The first dataset consists of NCVD-ACS source data provider (SDP) from the University Malaya Medical Centre (UMMC) from the year 2006 to 2015. After cleaning the dataset, 65 features were selected, where 64 were the input attributes and 1 was the outcome or the predicted attribute (i.e., Killip class). These attributes feature the demographics status before ACS event, clinical presentation and examination, electrocardiography (ECG), baseline investigation results (obtained within 48 h from admission), clinical diagnosis at admission, pharmacological therapy given before admission, and in-hospital clinical outcomes. The description and type of attributes are presented in Appendix (Table 9). The second dataset used to evaluate the prediction models is NCVD-ACS SDP from Sultanah Aminah Hospital (SAH).

Killip class is the predicted attribute that identifies the risk level of hemodynamic measurement of a patient at the time of presentation. There are 4 Killip classes as stratified by severity: class I (lowest), class II, class III, and class IV (highest). In this research, only the information captured within 48 h of the first admission was used for risk prediction, whereas follow-up data after hospital discharge were excluded. To avoid overfitting during the training and testing phases, the tenfold cross-validation technique was implemented to randomly partitioned the dataset into 10 equal-sized subsamples.

3 Methods

This research adopted Cross-Industry Standard Process for Data Mining (CRISP-DM) to ensure the quality of experiment results, and this methodology consists of six phases in cyclic whereby several iterations were used to tune the final result towards the research goals [20]. Figure 1 shows an overview of the CRISP-DM methodology.

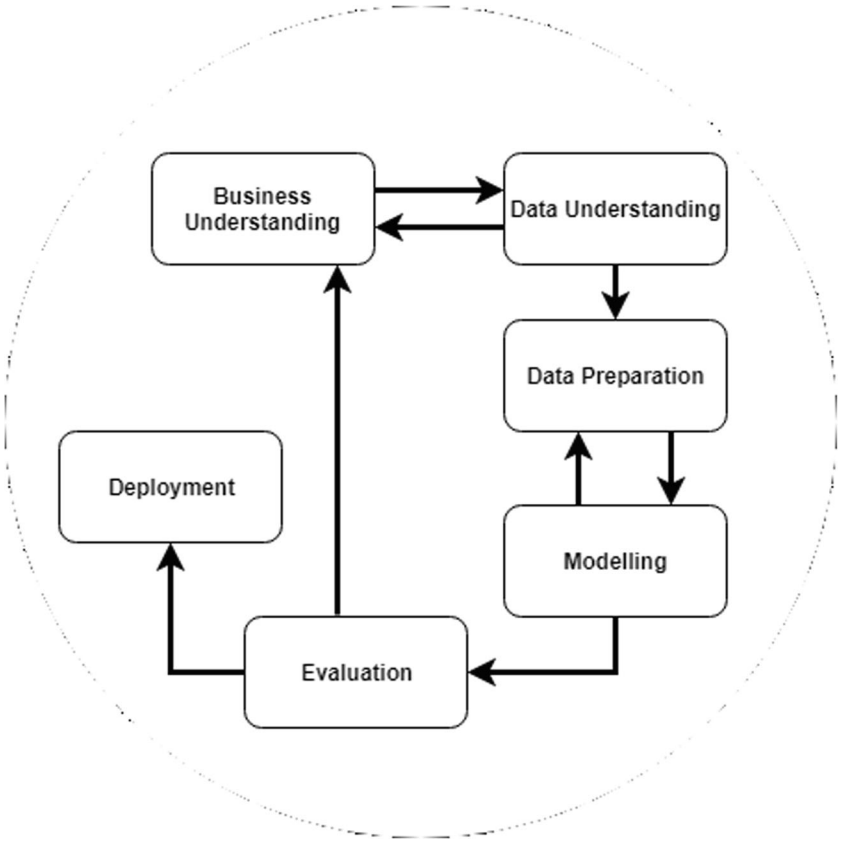

Fig. 1 CRISP-DM methodology (adapted from [20])

During business understanding phase, the goal of this research was defined, which is to identify significant features and machine learning algorithms for the building of classification models to predict CAD risk level. Next, the datasets were assessed for its quality in terms of percentage of the missing values (data understanding) and were pre-processed to become a clean dataset (data preparation). During modelling phase, experiments were conducted to select significant features using feature selection methods, and risk prediction models were built using the selected features and machine learning algorithms to exhibit the knowledge and classify the risk level of the patients with CAD. Subsequently, the performance of prediction models was measured using different metrics in the evaluation phase. Based on the initial results, a new iteration took place to improve the models with the selected algorithms and feature selection methods. Finally, the data mining results were reported, and the best performing risk prediction model was deployed in the deployment phase. Section 3.1 to Section 3.4 describes the data pre-processing, feature selection, classification modelling, and performance measure in more detail.

3.1 Data pre-processing and division

The datasets were stored in three different files, namely Patient, Patient Notification and Patient Outcome, which describes the patient's personal information, patient's medical observation and records, and outcome after receiving treatment, respectively. Originally, the Patient data file has 8 attributes, the Patient Notification data file has 184 attributes and Patient Outcome data file has 10 attributes.

Figure 2 illustrates the data pre-processing steps. During data cleaning, attributes with high missing values, noisy data, and outliers were identified. All the missing values and outliers were replaced by either mean or median. Next is the data integration phase in which all the files were merged.

During the data reduction phase, attributes that have 50% or more missing values, redundant and incorrect values, and irrelevant attributes were removed from the datasets. These attributes include treatment information after admission, time and date of admission, private information (e.g., name, patient ID), and attributes related to cardiac markers tests (e.g., CK, CK-MB, troponins). These attributes were removed after careful consideration and consultation with advisors of the NCVD committee formed by cardiologists. The merged dataset contains 5,100, observations and 65 attributes as the final cleaned dataset.

The predicted attribute "KillipClass" has four values, i.e., class I, class II, class III, and class IV based on the severity levels during the data transformation phase. The four Killip classes were further merged into two classes based on risk levels, i.e., the "Low Risk" class (by merging Killip classes I

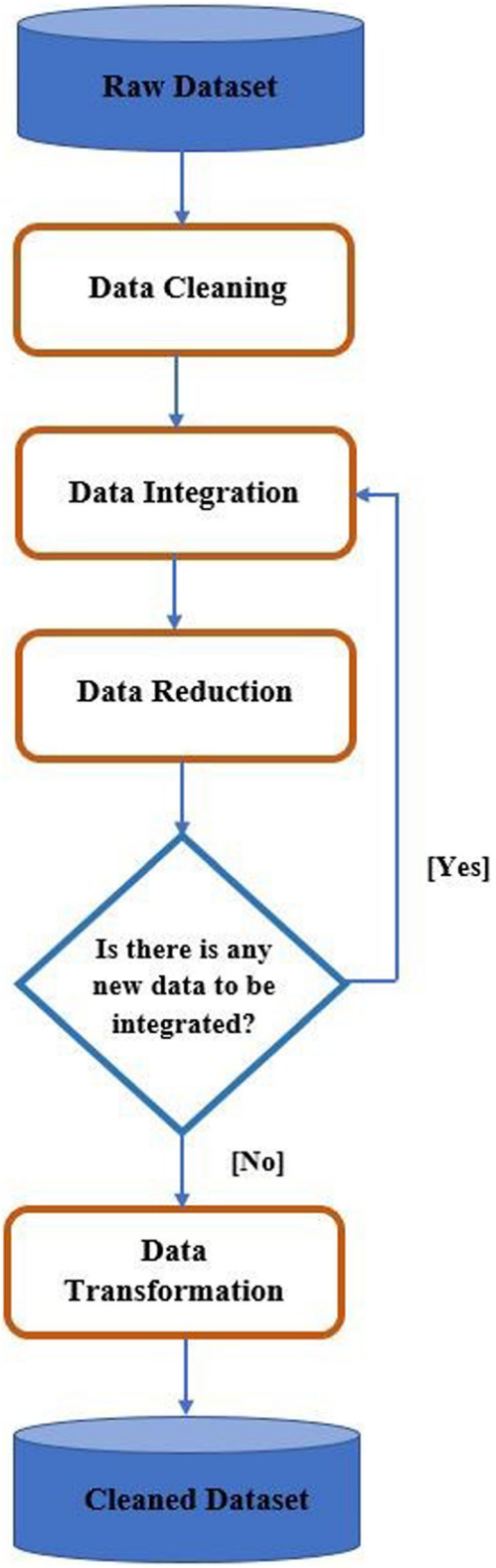

Fig. 2 Pre-processing steps

and II) and "High Risk" class (by merging Killip classes III and IV). Figures 3 and 4 show the distribution of the Killip classes (UMMC dataset) before and after the merging process.

To address the imbalanced dataset problem, Synthetic Minority Over-sampling Technique (SMOTE) [21] was implemented for both UMMC and SAH datasets. Additional "High Risk" class was synthetically created to match the amount of "Low Risk" class. Thus, the minority class, "High Risk" was oversampled to have the same number of observations as "Low Risk" class. The datasets were subsequently partitioned into 10 subsets for training and testing using the tenfold cross-validation technique.

3.2 Feature selection methods

The feature selection methods applied in this study are Chi-squared test (filter method), recursive feature elimination (RFE) using Logistic Regression (wrapper method) and Embedded Decision Tree (DT) using random forest algorithm (embedded method) [22, 23]. The Chi-squared test measures how strongly one attribute implies the other data, based on the available data using statistical calculation [24]. On the other hand, RFE works by recursively removing attributes and building a model based on remaining attributes [25]. Embedded DT using random forest constructs the set of attributes in the decision trees that form the reduced subset of attributes from the given datasets [24, 26]. The features selected by each method are reported and discussed in Section 4.1.

3.3 Classification modelling using machine learning algorithms

Classification models were built to predict the target variable which is the risk level based on Killip class. There are eight widely used supervised machine learning algorithms implemented for classification modelling in this research. These algorithms are LR, NN, kNN, DT, NB, SVM, DL, and Vote to get unbiased prediction outcomes. Experiments were conducted to build classification models for predicting the risk levels of patients with CAD (see Fig. 5) as follows:

- Eight (8) risk prediction models were built using the full features (64 features) and the eight machine learning algorithms;
- Forty-eight (48) risk prediction models were built using a subset of significant features selected by every feature selection method and the eight machine learning algorithms.

3.4 Performance measure

After building the risk prediction models, the performance of the classification models was measured in terms of accuracy, specificity, precision, recall, F1 score and Area Under the

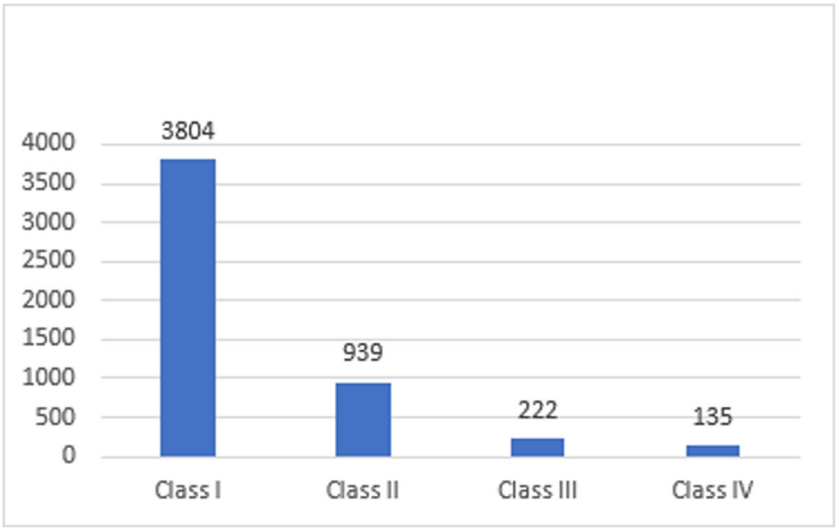

Fig. 3 Killip class distribution of the UMMC's NCVD-ACS dataset

Curve (AUC) – Receiver Operating Characteristics (ROC) curve. Among these six measures, AUC is one of the most essential evaluation metrics for comparing the performance of any classification model. Figure 6 shows the confusion matrix defined in this study.

Accuracy shows the overall performance of the model, which is the ratio of the number of correct predictions to the total number of predictions made.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Specificity is the ratio of correctly predicted cases as low risk to the total of low-risk cases.

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

Precision is the ratio of correctly predicted high risk to the total of high-risk cases

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

Recall shows the capability of a model to find the targeted cases within the dataset, which is the number of correctly predicted high-risk cases divided by the total of cases that are correctly predicted as high risk and cases that are incorrectly predicted as low risk.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

F1 score is the harmonic average between both metrics of precision and recall.

$$\text{F1} = 2 \times (\text{precision} \times \text{recall}/\text{precision} + \text{recall})$$

ROC is the probability curve, and AUC indicates the separability degree. It shows how well a model can differentiate between classes. The ROC curve is plotted against the FPR by

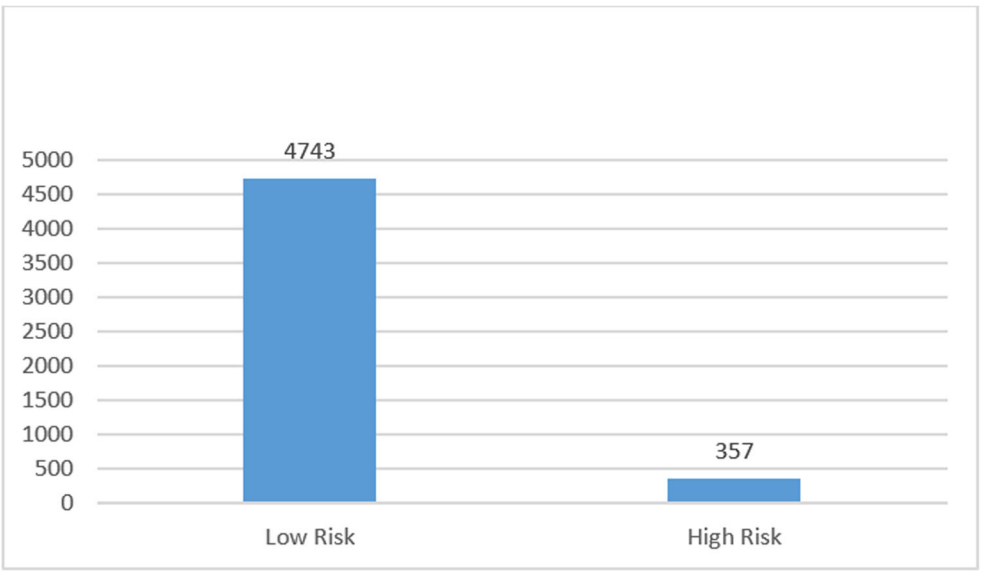

Fig. 4 Risk level distribution of the UMMC's NCVD-ACS dataset after merging the Killip classes

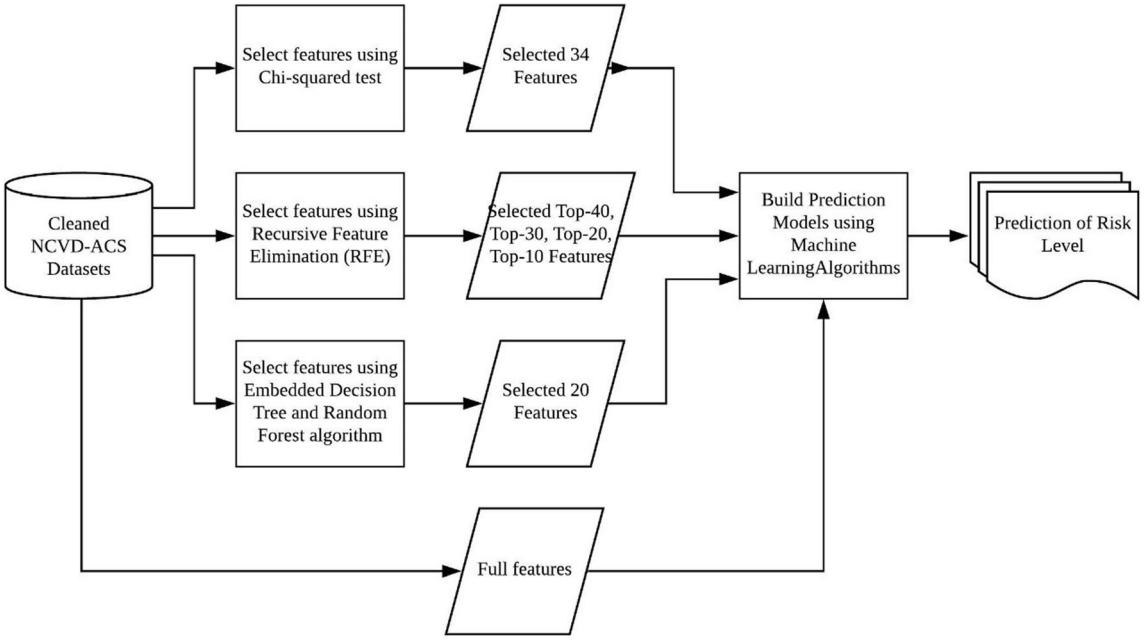

Fig. 5 Risk prediction models developed using significant features and machine learning algorithms

TPR, where TPR is on the *y*-axis and FPR on the *x*-axis. A model that has higher AUC is better at predicting patients with high risk and low risk.

4 Results

4.1 Selected features

During feature selection, irrelevant features were removed and significant features were selected to produce unbiased optimal results. The filter method, i.e., Chi-squared test, performed a statistical test to calculate the dependency of every feature variable with the target class (i.e., risk level based on Killip class). Out of the 64 features, 34 feature variables were identified as highly dependent features that can be used to predict the risk level (Table 1(a)). The embedded method, i.e., DT using random forest algorithm, measured the Gini impurity of each feature and resulted in 20 important feature variables for the classification of the risk level (Table 1(b)). On the other hand, RFE, the wrapper method was used to perform a greedy search that recursively eliminates and selects the best performing subset of features for risk prediction, ranging from top 40 to top 10 features (Table 1(c)).

	Predicted High Risk	Predicted Low Risk
Actual High Risk	TP	FN
Actual Low Risk	FP	TN

Fig. 6 Confusion matrix defined in the current study. True positive (TP) is the number of cases correctly identified as high risk; true negative (TN) is the number of cases correctly identified as low risk; false positive (FP) is the number of cases incorrectly identified as high risk; and false negative (FN) is the number of cases incorrectly identified as low risk.

4.2 Performance of classification models

Experiments were conducted to build and evaluate the performance of classification models for predicting the CAD risk levels, utilizing UMMC dataset through tenfold cross-validation technique. Table 2(a) and (b) show that all the models achieved the accuracy and precision of more than 73%. The highest accuracy (94.5%) and precision (94.5%) was achieved by the model of NN with Embedded DT features. The models built using full features obtained more than 84% in accuracy and precision. However, among these models, the model built using RFE-Top10 features attained accuracy and precision of less than 77% in risk prediction.

Table 2(c) and (d) present the specificity and recall rate. Among all the models, models developed using the kNN with full features obtained the lowest specificity rate (63%) and the highest recall rate (99.3%). On the other hand, Table 2(e) shows that most of the models gained more than 73% in F1 score, whereby the highest F1 score (94.5%) is achieved by

Table 1 Selection by the filter, embedded, and wrapper methods for CAD risk prediction

Feature selection method	Number of features	Features
(a) Filter method (Chi-squared test)	34 selected features	cAnginaPast2Wk, cHeartFail, ECGAbnormTypeBBB, ACEIPre, TotalDayStay, PtOutcome, ASAPre, HeartRate, ACSStratum, cPremCVD, cRenal, ECGAbnormTypeTWave, cDMm, ECGAbnormTypeSTElev2, PtAgeAtNotification, DayCCU, StatinPre, cDM_OHA, BBPre, ECGAbnormLocationAL, DiureticPre, ADPAPre, cDM_NonPharmaTherapy, TG, ECGAbnormLocationLL, cPVascular, cLung, DayICU, OralHypoglyPre, ECGAbnormLocationRV, ECGAbnormTypeSTDep, cDM_Insulin, cCerebrovascular, ECGAbnormLocationTP
(b) Embedded method (Embedded DT using the random forest)	20 selected features	PtOutcome, PtAgeAtNotification, PtRace, TotalDayStay, BMI, SmokingStatus, TG, DayCCU, ACSStratum, cHeartFail, cDys, cAnginaPast2Wk, StatusAspirinUse, ECGAbnormTypeSTDep, ECGAbnormLocationLL, cHPT, ECGAbnormTypeTWave, cDM, ECGAbnormLocationAL, ECGAbnormLocationIL
(c) Wrapper method (RFE)	Top 40 features	cDM_NonPharmaTherapy, cAnginaPast2Wk, cHeartFail cPVascular, ECGAbnormTypeBBB, ECGAbnormLocationRV, ACEIPre, DiureticPre, TotalDayStay, PtOutcome, ASAPre, GPRIPre, HeartRate, ACSStratum, cPremCVD, cAnginaMT2Wk, cRenal, BBPre, LMWHPre, ECGAbnormTypeTWave, cLung, ADPAPre, cDM, AnginaEpisodeNo, ECGAbnormLocationTP, ECGAbnormLocationLL, BleedingEpisodeCriteria, LipidLAPre, cCerebrovascular, ECGAbnormTypeSTElev2, HeparinPre, TG, BMI, ECGAbnormLocationAL, LVEF, StatinPre, TC, PtAgeAtNotification, HDLC, LDLC
	Top 30 features	cDM_NonPharmaTherapy, cAnginaPast2Wk, cHeartFail, cPVascular, ECGAbnormTypeBBB, ECGAbnormLocationRV, ACEIPre, DiureticPre, TotalDayStay, PtOutcome, ASAPre, GPRIPre, HeartRate, ACSStratum, cPremCVD, cAnginaMT2Wk, cRenal, BBPre, LMWHPre, ECGAbnormTypeTWave, cLung, ADPAPre, cDM, AnginaEpisodeNo, ECGAbnormLocationTP, ECGAbnormLocationLL, BleedingEpisodeCriteria, LipidLAPre, cCerebrovascular, ECGAbnormTypeSTElev2
	Top 20 features	cDM_NonPharmaTherapy, cAnginaPast2Wk, cHeartFail, cPVascular, ECGAbnormTypeBBB, ECGAbnormLocationRV, ACEIPre, DiureticPre, TotalDayStay, PtOutcome, ASAPre, GPRIPre, HeartRate, ACSStratum, cPremCVD, cAnginaMT2Wk, cRenal, BBPre, LMWHPre, ECGAbnormTypeTWave
	Top 10 features	cDM_NonPharmaTherapy, cAnginaPast2Wk, cHeartFail, cPVascular, ECGAbnormTypeBBB, ECGAbnormLocationRV, ACEIPre, DiureticPre, TotalDayStay, PtOutcome

NN with Embedded DT features. In Table 2(f), AUC rates range from 74.1 to 97.6%, where kNN model with features selected Embedded DT is the most capable model of distinguishing between high-risk and low-risk classes. Overall, the models developed using full features are excellent

Table 2 Accuracy, precision, specificity, recall, F1-score, and AUC achieved using the full features and selected features on UMMC dataset using tenfold cross-validation

Performance metrics	Machine learning algorithms	Full features (64 features)	Chi-squared test (34 features)	Embedded DT (20 features)	RFE-Top40 (40 features)	RFE-Top30 (30 features)	RFE-Top20 (20 features)	RFE-Top10 (10 features)
(a) Accuracy	LR	90.5	85.1	83.5	86.4	80.1	77.6	74.1
	NN	94.1	93.6	94.5	93.7	86.6	82.1	76.4
	kNN	81.2	85.5	85.5	83.8	83.0	81.1	76.3
	DT	92.3	91.3	91.6	91.1	85.7	81.4	76.4
	NB	84.5	80.4	81.5	82.7	78.8	77.4	73.4
	SVM	90.8	84.8	83.4	86.4	79.9	78.4	74.1
	DL	90.2	84.3	83.2	86.1	79.7	77.8	73.8
	Vote	87.8	82.8	83.4	84.1	79.2	77.6	74.4
(b) Precision	LR	90.5	85.1	83.5	86.5	80.2	77.6	74.6
	NN	94.1	93.6	94.5	93.8	86.7	82.7	76.6
	kNN	85.9	88.3	88.3	87.3	84.2	82.3	76.5
	DT	92.4	91.5	91.9	91.3	86.1	81.5	76.5
	NB	84.8	80.6	81.5	83.3	79.1	77.7	74.7
	SVM	90.8	84.8	83.6	86.7	79.9	78.4	74.9
	DL	90.3	84.3	83.2	86.2	79.7	77.8	74.4
	Vote	88.0	82.8	82.7	84.3	79.3	77.6	75.1
(c) Specificity	LR	88.7	83.6	81.7	83.4	79.2	76.2	81.0
	NN	94.6	92.2	93.9	92.2	84.1	75.1	81.3
	kNN	63.0	72.1	72.0	68.5	73.8	71.5	79.6
	DT	89.6	88.0	87.6	87.4	81.1	78.6	80.0
	NB	79.8	76.3	81.3	76.5	74.4	72.0	85.1
	SVM	89.3	82.9	80.2	82.4	79.4	80.5	83.2
	DL	88.3	83.2	81.2	83.2	78.5	76.2	81.8
	Vote	85.0	80.5	82.1	80.3	76.6	77.2	82.7
(d) Recall	LR	92.2	86.7	85.2	89.3	81.1	78.9	67.2
	NN	93.7	95.0	95.1	95.2	89.0	89.0	71.5
	kNN	99.3	99.0	99.0	99.1	92.2	90.7	73.1
	DT	94.9	94.7	95.6	94.8	90.4	84.1	72.8
	NB	89.1	84.5	81.8	89.0	83.3	82.9	61.6
	SVM	92.3	86.6	86.7	90.5	80.3	76.2	64.9
	DL	92.1	85.3	85.1	89.1	80.8	79.3	65.8
	Vote	90.7	85.1	83.4	88.0	81.7	77.9	66.0
(e) F1 score	LR	90.5	85.1	83.5	86.3	80.1	77.6	74.0
	NN	94.1	93.6	94.5	93.7	86.6	82.0	76.3
	kNN	80.5	85.3	85.2	83.4	82.9	80.9	76.3
	DT	92.2	91.3	91.6	91.1	85.7	81.3	76.4
	NB	84.4	80.4	81.5	82.7	78.8	77.3	73.0
	SVM	90.8	84.8	83.4	86.4	79.9	78.3	73.8
	DL	90.2	84.3	83.2	86.1	79.7	77.8	73.6
	Vote	87.8	82.8	82.7	84.1	79.2	77.6	74.2
(f) AUC	LR	95.5	92.1	91.1	93.0	89.2	87.0	80.3
	NN	96.6	96.6	96.6	96.4	93.1	91.2	82.6
	kNN	96.6	97.2	97.6	96.9	93.1	91.2	82.5
	DT	94.6	94.3	94.3	94.4	91.8	89.8	82.0
	NB	92.8	89.1	89.8	91.0	87.7	85.7	79.2
	SVM	90.8	84.8	83.4	86.4	79.9	78.4	74.1

Table 2 (continued)

Performance metrics	Machine learning algorithms	Full features (64 features)	Chi-squared test (34 features)	Embedded DT (20 features)	RFE-Top40 (40 features)	RFE-Top30 (30 features)	RFE-Top20 (20 features)	RFE-Top10 (10 features)
	DL	95.7	91.8	90.8	92.8	89.0	86.8	80.6
	Vote	95.4	91.6	90.8	92.8	88.9	86.7	79.9

at differentiating high-risk and low-risk patients with the AUC more than 90% while models developed using RFE-Top10 features are displays the lower AUC (i.e range from 74.1 to 82.6%).

4.3 Performance on unseen data from other center

This section describes the evaluation of the prediction models using the NCVD-ACS dataset from another hospital, i.e., SAH in Malaysia. The objective of this evaluation is to validate the findings of the selected features and the performance of the prediction models.

Table 3 shows the comparison of UMMC and SAH datasets. Same pre-processing was applied to SAH's dataset, whereby both final cleaned datasets have 65 attributes in total. As compared to UMMC dataset, there are 1,015 low-risk patients and 200 high-risk patients in SAH dataset. There are only a total of 1215 instances in the final cleaned SAH dataset.

Table 4 (a) and (b) show the accuracy and precision of the risk prediction models. All the models achieved accuracy range from 77.7 to 89.7% and precision range from 79.4 to 89.7%. Same as UMMC dataset, the highest accuracy (89.7%) and precision (89.7%) was obtained by the model developed using NN with Embedded DT features. The models built using NB with RFE-Top10 features and NB with Embedded DT features obtained the lowest precision of 79.4%. On the other hand, the risk prediction models achieved specificity rates range from 62.4 to 89.9% (Table 4(c)). Table 4(d) shows that the recall values of all models are almost similar as compared to UMMC dataset. The highest recall rate (98.4%) was gained by the model built by kNN with full features. In Table 4(e), the F1 scores of all models range between 77.3 and 89.7%. The model, i.e., NN with Embedded DT, obtained the highest F1 score of 89.7%. Table 4(f) shows AUC rates range from 79.2 to 95%, where the model developed using kNN with Embedded DT features achieving the highest AUC.

Table 3 Predicted attribute class distribution of UMMC and SAH datasets

Comparison category	UMMC dataset		SAH dataset	
No of attributes	65		65	
Distribution of Killip class	Low risk	High risk	Low risk	High risk
	4743	357	1015	200
Total number of instances	5100		1215	

5 Discussion

5.1 Performance of risk prediction models

The prediction models were built based on the full set of features and selected features. Overall, the accuracy of all the models using UMMC dataset (73.4 to 94.5%) is higher than SAH dataset (77.7% to 89.7%). The recall rate of all the models using SAH dataset is almost similar (range from 65.5 to 98.4%) to UMMC dataset (range from 61.6 to 99.3%).

Nevertheless, accuracy does not guarantee the performance results obtained is acceptable as it may be biased to dominant class if the dataset is imbalanced. Due to the imbalanced distribution of classes, SMOTE was implemented to overcome this issue and used six different performance evaluation metrics (accuracy, precision, specificity, recall, F1 score, and AUC score) for comparison of classification model's performance evaluation. Additionally, we had listed the top five models for each performance metrics as presented in Table 5. The name of each model is represented by the combinations of machine learning (ML) algorithms and set of features. A subset of features selected by each feature selection method (FSM) is represented using the name of the FSM. Based on the top five models, we counted the frequency of each model appeared as the top five for each performance metric. The models that have appeared as the top three of the six performance metrics are listed in Table 6 for both datasets.

According to Table 6, the NN model with Embedded DT features is identified as one of the top five models for all six performance metrics when evaluated using UMMC dataset. On the other hand, this model has appeared as top models for four performance metrics when it was evaluated using SAH dataset. Meanwhile, for both UMMC and SAH dataset, there are two models that have a frequency of 5 over 6; these models were developed using NN with RFE-Top40 features and NN with full features.

Table 4 Accuracy, precision, specificity, recall, F1-score, and AUC achieved using the full features and selected features on unseen SAH dataset

Performance metrics	Machine learning algorithms	Full features (64 features)	Chi-squared test (34 features)	Embedded DT (20 features)	RFE-Top40 (40 features)	RFE-Top30 (30 features)	RFE-Top20 (20 features)	RFE-Top10 (10 features)
(a) Accuracy	LR	85.9	82.1	81.7	86.1	82.4	82.5	79.0
	NN	89.6	89.5	89.7	89.6	85.5	83.3	79.8
	kNN	80.4	85.0	83.3	83.6	84.8	82.8	79.8
	DT	86.1	86.9	85.4	85.6	84.4	82.6	79.7
	NB	83.3	80.1	79.3	83.3	79.3	79.7	77.7
	SVM	86.1	83.2	80.4	86.2	80.9	79.7	79.2
	DL	85.4	82.2	81.3	85.6	82.4	82.2	79.2
	Vote	85.2	82.2	79.9	84.9	80.3	80.5	78.1
(b) Precision	LR	85.9	82.1	81.8	86.1	82.5	82.7	79.7
	NN	89.6	89.5	89.7	89.6	85.7	83.3	80.2
	kNN	84.9	87.1	86.5	85.5	85.0	82.8	80.3
	DT	86.2	87.1	85.6	85.8	84.5	82.6	80.4
	NB	83.4	80.2	79.4	83.3	79.9	80.9	79.4
	SVM	86.1	83.3	80.6	86.3	81.5	80.3	80.0
	DL	85.4	82.2	81.3	85.6	82.6	82.4	79.8
	Vote	85.2	82.3	80.1	84.9	80.8	81.2	79.8
(c) Specificity	LR	86.5	83.3	83.9	85.5	85.9	86.5	86.5
	NN	88.6	87.7	87.7	88.6	82.3	83.3	85.9
	kNN	62.4	73.3	68.3	71.9	80.4	82.1	86.5
	DT	83.1	83.5	81.4	82.2	82.9	81.6	87.4
	NB	84.4	82.3	82.6	83.4	86.3	89.7	89.9
	SVM	86.8	86.1	84.0	83.1	87.9	86.9	87.2
	DL	85.5	83.5	82.8	85.2	86.2	86.1	86.4
	Vote	86.2	84.4	83.9	85.3	86.5	87.8	89.9
(d) Recall	**LR**	85.2	80.9	79.5	86.7	78.8	78.5	71.5
	NN	90.5	91.2	91.6	90.6	88.8	83.4	73.6
	kNN	98.4	96.7	98.2	95.3	89.2	83.5	73.1
	DT	89.2	90.2	89.4	89.1	86.0	83.6	72.0
	NB	82.3	78.0	76.1	83.1	72.2	69.8	65.5
	SVM	85.4	80.3	76.7	89.3	74.0	72.4	71.2
	DL	85.2	80.8	79.8	86.0	78.6	78.3	72.0
	Vote	84.1	80.0	75.9	84.4	74.1	73.2	66.4
(e) F1 score	LR	85.9	82.1	81.7	86.1	82.3	82.5	78.9
	NN	89.6	89.5	89.7	89.6	85.5	83.3	79.7
	kNN	79.7	84.8	82.9	83.4	84.7	82.8	79.7
	DT	86.1	86.9	85.3	85.6	84.4	82.6	79.6
	NB	83.3	80.1	79.3	83.3	79.2	79.5	77.3
	SVM	86.1	83.2	80.4	86.1	80.8	79.5	79.1
	DL	85.4	82.2	81.3	85.6	82.4	82.2	79.1
	Vote	85.2	82.2	79.9	84.9	80.2	80.4	77.8
(f) AUC	LR	92.2	90.6	90.5	92.2	89.4	88.3	84.0
	NN	93.8	93.5	93.2	93.5	91.4	89.6	84.5
	kNN	94.5	94.7	95.0	93.5	93.1	90.1	84.4
	DT	89.7	91.3	91.5	90.1	89.7	87.1	82.8
	NB	91.8	89.8	89.1	91.3	88.3	86.5	81.1
	SVM	86.1	83.2	80.4	86.2	80.9	79.7	79.2
	DL	92.2	90.5	90.1	92.0	89.3	88.3	84.3
	Vote	93.0	90.8	90.3	92.3	89.7	88.2	83.9

The AUC score also can be used as a measurement to evaluate the performance of a classification model as this metric is useful and informative to assess the capability of a model to identify between classes. The AUC score range of models developed using RFE-Top10 features is the lowest range for both datasets where the range is from 74.1 to 82.6% for UMMC dataset and 79.2 to 84.5% for SAH dataset. Based on the AUC scores, the model developed using kNN with Embedded DT features has achieved the highest AUC score for both datasets where the score is 97.6% for UMMC dataset and 95% for SAH dataset. In brief, this makes the model developed using kNN with Embedded DT features identified as the best performing model in terms of AUC score.

From the machine learning algorithm perspective, NN and kNN demonstrate consistent performance as they commonly appeared as one of the top 5 ML algorithms for all performance metrics in UMMC and SAH datasets. Thus, in terms of frequency shown in Table 6, there are four models

Table 5 Performance analysis of top five classification models for risk prediction

Performance metrics	UMMC dataset		SAH	
	Model	Value	Model	Value
(a) Accuracy	Embedded DT + NN	94.5	Embedded DT + NN	89.7
	Full features + NN	94.1	Full features + NN	89.6
	RFE-Top40 + NN	93.7	RFE-Top40 + NN Chi-squared + NN	89.5
	Chi-squared + NN	93.6	Chi-squared + DT	86.9
	Full features + DT	92.3	RFE-Top40 + SVM	86.2
(b) Precision	Embedded DT + NN	94.5	Embedded DT + NN	89.7
	Full features + NN	94.1	Full features + NN	89.6
	RFE-Top40 + NN	93.8	RFE-Top40 + NN Chi-squared + NN	89.5
	Chi-squared + NN	93.6	Chi-squared + kNN	87.1
	Full features + DT	92.4	Chi-squared + DT Embedded DT + kNN	86.5
(c) Specificity	Full features + NN	94.6	RFE-Top10 + NB	89.9
	Embedded DT + NN	93.9	RFE-Top10 + Vote RFE-Top20 + NB	89.7
	Chi-squared + NN	92.2	RFE-Top40 + NN	88.6
	RFE-Top40 + NN Full features + DT	89.6	Full features + NN RFE-Top30 + SVM	87.9
	Full features + SVM	89.3	RFE-Top20 + Vote	87.8
(d) Recall	Full features + kNN	99.3	Full features + kNN	98.4
	RFE-Top40 + kNN	99.1	Embedded DT + kNN	98.2
	Chi-squared + kNN	99.0	Chi-squared + kNN	96.7
	Embedded DT + kNN Embedded DT + DT	95.6	RFE-Top40 + kNN	95.3
	Embedded DT + NN	95.1	Embedded DT + NN	91.6
(e) F1 score	Embedded DT + NN	94.5	Embedded DT + NN	89.7
	Full features + NN	94.1	RFE-Top40 + NN	89.6
	RFE-Top40 + NN	93.7	Full features + NN Chi-squared + DT	86.9
	Chi-squared + NN	93.6	RFE-Top40 + LR	86.1
	Full features + DT	92.2	RFE-Top40 + SVM Full features + DT Full features + SVM Full features + LR	85.9
(f) AUC	Embedded DT + kNN	97.6	Embedded DT + kNN	95.0
	Chi-squared + kNN	97.2	Chi-squared + kNN	94.7
	RFE-Top40 + kNN	96.9	Full features + kNN	94.5
	Full features + NN	96.6	Full features + NN	93.8
	Full features + kNN Embedded DT + NN Chi-squared + NN RFE-Top40 + NN	96.4	Chi-squared + NN RFE-Top40 + NN RFE-Top40 + kNN	93.5

developed using NN that have appeared as top models for both datasets making NN as the best performing ML algorithm in this study. On the other hand, models developed using LR, NB, SVM, DL, and Vote showed a slightly poor performance than other models when using the features selected by RFE-Top30, RFE-Top20, RFE-Top10, and Embedded

Table 6 Top three models that have appeared as the top five models among all performance evaluation metrics

Dataset	Combination of features + ML	Frequency
UMMC	Embedded DT + NN	6
	Full Feature + NN RFE-Top40 + NN Chi-squared + NN	5
	Full Features + DT	4
SAH	RFE-Top40 + NN Full Features + NN	5
	Embedded DT + NN	4
	Chi-squared + NN Chi-squared + DT Chi-squared + kNN Embedded DT + kNN	3

DT for both datasets. Furthermore, SVM-based models have the lowest range in AUC score for both datasets where the range is from 74.1 to 90.8% for UMMC dataset and 79.2 to 86.1% for SAH dataset. Overall, all the models managed to get scores more than 60% for all of the performance metrics which means that their performances are quite good.

While from feature selection perspective, models that have used 20 features selected by Embedded DT features has achieved the highest AUC score and also high performance among all evaluation metrics. On the other hand, the models adopting the top 10 significant features selected by RFE has obtained the lowest score range in all of the performance metrics except for specificity for both datasets. Among all the subset of features, the models developed using top 40 features selected by RFE and any machine learning algorithm has shown more stable performance for both datasets.

These results suggest that the models created using significant features could perform better than or similar to the models built with the full set of features. Besides, most of the models performed better with selected features than full features. This shows that the implementation of feature selection methods is worthwhile to improve the performance of the models for risk prediction. Removing the redundant features reduces the processing time and complexity of the models, and also improves the quality of the models.

5.2 Performance benchmarking of the proposed models

This section shows the comparison of the proposed models with the performance of existing studies for predicting risk or presence of heart disease. The benchmarking method was used to assess whether the best performing model has achieved acceptable performance as compared to other models proposed by other studies. Table 7 presents the benchmarking of performance results of the best performing

Table 7 Benchmarking of the proposed models against performance achieved by existing studies

Source	ML algorithm used	Accuracy	Specificity	F1 score	AUC
Chaurasia and Pal [27]	Decision Tree (CART)	83.94%	–	–	–
Subanya and Rajalaxmi [28]	SVM	86.76%	–	–	–
Ismaeel et al. [29]	Extreme Learning Machine	86.50%	–	–	–
El-Bialy et al. [30]	Decision Tree	78.54%	–	–	–
Nahar et al. [31]	Naïve Bayes (computerized feature selection process (CFS))	86%	–	58.8%	93.7%
Verma et al. [32]	Decision Tree	80.68%	–	–	–
Wiharto et al. [33]	Neural Network	86.3%	88.2%	–	92.1%
Paul et al. [34]	Adaptive Weighted Fuzzy system ensemble	92.31%	Approx. 90	–	–
Amin et al. [11]	Vote with Naïve Bayes and Logistic Regression	87.41%	–	–	–
Ali et al. [34]	SVM	92.22%	100%		
Reddy et al. [36]	Adaptive genetic algorithm	90.00%	90%	–	–
Latha and Jeeva [37]	Ensemble learning	85.48%	–	–	–
Proposed model	Neural Network with Embedded DT features	94.5% (UMMC); 89.7% (SAH)	93.9% (UMMC); 87.7% (SAH)	94.5% (UMMC); 89.7% (SAH)	96.6% (UMMC); 93.5% (SAH)

model in this study, NN with Embedded DT features against the results obtained by recent studies that are using the UCI machine learning repository. Most of the existing studies reported the accuracy results only. According to Table 7, the performance of the proposed model is better than the existing studies. Besides NN with Embedded DT features, the top models reported in Section 5.1 has achieved an accuracy of more than 86% and AUC more than 90%. This benchmarking proved that the proposed prediction models in this research have acceptable performance for predicting risk as compared to the existing studies that have conducted heart disease prediction.

Most of the previous studies related to heart disease prediction focus on predicting the presence of heart disease in a patient, rather than predicting the risk. Additionally, these studies always used publicly available datasets which have a small number of attributes and records (e.g., UCI heart disease datasets only have 14 attributes and around 300 instances). This study focuses on predicting the risk level of patients with heart disease. Machine learning algorithms are not one-size-fits-all, and the results found in other studies cannot be used to decide on a classifier that will perform the best with minimal error for the Malaysia heart disease datasets for risk prediction. This study is the first study using Malaysia's NCVD datasets to predict the risk level of ACS patients in Malaysia. The datasets have a larger number of instances (UMMC-5100; SAH-1215) and attributes (65). As a result, this study tried several machine learning algorithms that are suitable for this research problem and then compare the performance results using different performance metrics. Furthermore, since this study focuses on risk prediction; more features are added to explore the significant features that can be used for risk prediction, especially in Malaysia context. This makes a novel contribution of this study .

5.3 Analysis of significant features

There are 11 features selected by all three feature selection methods (Chi-squared test, RFE, and embedded DT) as the target variables for risk prediction based on the Killip class as shown in Table 8. These 11 features are cAnginaPast2Wk, cHeartFail, PtOutcome, TotalDayStay, cDM, ECGAbnormTypeTWave, ECGAbnormLocationAL, ECGAbnormLocationLL, ACSStratum, PtAgeAtNotification, and TG. These features were grouped as the following types of clinical information:

- Demographics: PtAgeAtNotification;
- Status before ACS event: cAnginaPast2Wk, cHeartFail, cDM;
- ECG: ECGAbnormTypeTWave, ECGAbnormLocationAL, ECGAbnormLocationLL;
- Clinical diagnosis at admission: ACSStratum;
- Baseline investigations (obtained within 48 h from admission): TG;
- In-hospital clinical outcomes: TotalDayStay, PtOutcome.

As illustrated in Fig. 7, among the 11 features, there are 4 common features selected 6 times by all the feature selection methods, including RFE-Top40, RFE-Top30, RFE-Top20, RFE-Top10. These features are

Table 8 List of features selected by Chi-squared test, RFE, and Embedded DT

No.	Selected feature	Description	Feature selection method
1.	cAnginaPast2Wk	The patient has new onset angina (less than 2 weeks)	Chi-squared test, RFE-Top40, RFE-Top30, RFE-Top20, RFE-Top10, Embedded DT
2.	cHeartFail	The patient has a history of heart failure	Chi-squared test, RFE-Top40, RFE-Top30, RFE-Top20, RFE-Top10, Embedded DT
3.	PtOutcome	Patient outcome	Chi-squared test, RFE-Top40, RFE-Top30, RFE-Top20, RFE-Top10, Embedded DT
4.	TotalDayStay	Total number of overnight stays	Chi-squared test, RFE-Top40, RFE-Top30, RFE-Top20, RFE-Top10, Embedded DT
5.	cDM	ACS comorbidities: Diabetes	Chi-squared test, RFE-Top40, RFE-Top30, Embedded DT
6.	ECGAbnormTypeTWave	ECG abnormalities type: T-wave inversion ≥ 1 mm (0.1 mV)	Chi-squared test, RFE-Top40, RFE-Top30, RFE-Top20, Embedded DT
7.	ECGAbnormLocationAL	ECG abnormalities location: anterior leads	Chi-squared test, RFE-Top40, Embedded DT
8.	ECGAbnormLocationLL	ECG abnormalities location: lateral leads	Chi-squared test, RFE-Top40, RFE-Top30, Embedded DT
9.	ACSStratum	Acute coronary syndrome stratum	Chi-squared test, RFE-Top40, RFE-Top30, RFE-Top20, Embedded DT
10.	PtAgeAtNotification	Age at notification	Chi-squared test, RFE-Top40, Embedded DT
11.	TG	Lipid profile (fasting): triglyceride	Chi-squared test, RFE-Top40, Embedded DT

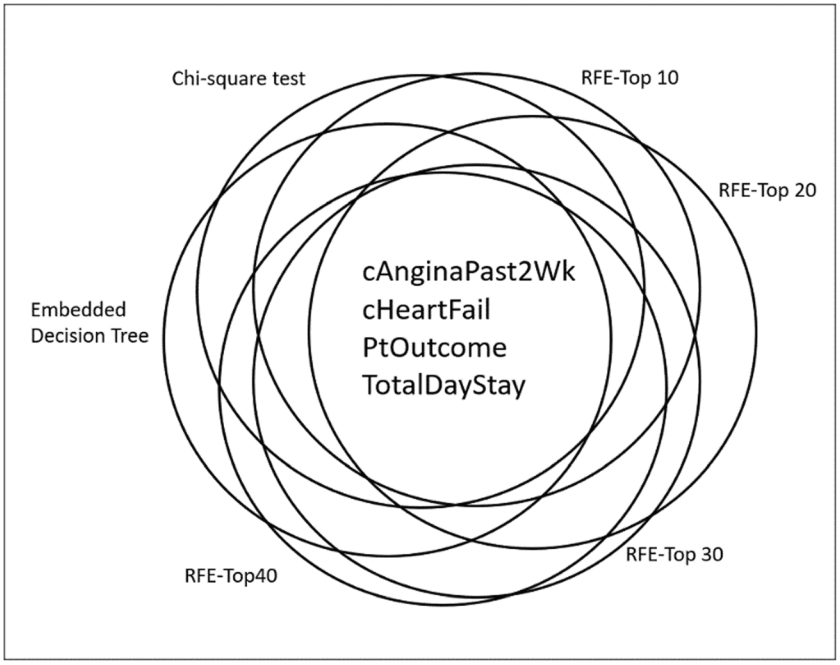

Fig. 7 Common features selected by all feature selection methods for risk prediction

cAnginaPast2Wk, cHeartFail, PtOutcome, and TotalDayStay. These four features represent whether the patient has a recent onset of angina (within the past 2 weeks), has a history of heart failure, patient outcome, and a total number of overnight stays. On the other hand, there are 12 features that are not selected by any of the feature selection methods which we assumed are less significant in predicting the risk level for NCVD-ACS dataset. These features are cMI, cCAP, ECGAbnormTypeSTElev1, ARBPre, CalcAntagonistPre, InsulinPre, AntiArrPre, PtSex, BPSys, BPDias, FBG, and HbA1c.

6 Conclusion and future work

In this research, classification models were developed with significant features and machine learning algorithms to predict the risk level based on Killip class. Real-life NCVD-ACS datasets provided by NCVD Malaysia were used in the experiments to select the significant features and develop risk prediction models. To conclude, there are four common features, including cAnginaPast2Wk, cHeartFail, PtOutcome, and TotalDayStay which are selected by all the feature selection methods and are assumed the most significant attributes in predicting CAD risk levels. The NN-based risk prediction model with features selected by Embedded DT was identified as the best performing model which appeared as top models for both datasets. In terms of AUC scores, all the top models have achieved more than 90% and the best performing model is kNN model with Embedded DT features. Based on the experiment results, the two machine learning algorithms that demonstrate good and consistent performances for both datasets are NN and kNN. Among the three feature selection methods, models developed using 20 features selected by embedded DT have achieved highest accuracy, precision, F1 score, AUC while the subset of top 40 features selected by RFE produces models with more stable performance. Nevertheless, this research can incorporate other techniques to handle the imbalanced datasets problems in the future and improve the quality of the proposed models for risk prediction. Additionally, the study may be extended to evaluate different types of machine learning algorithms such as unsupervised techniques for CAD risk prediction.

Acknowledgments This work was supported by the Ministry of Education Malaysia (Higher Education)'s Fundamental Research Grant Scheme (FRGS), Project Code: FP057-2017A. The authors would like to thank the Governance Board member of the Malaysian National Cardiovascular Disease Database (NCVD) Registry for providing us with the Acute Coronary Syndrome (ACS) dataset to be used for the research. Acknowledgement also goes to the Ministry of Health Malaysia and National Heart Association of Malaysia for funding the NCVD Registry database. Thanks to all the NCVD investigators and to all source data providers for their contribution to this registry.

Appendix

Table 9 Description of attributes from NCVD-ACS Dataset

Attribute name	Description	Data type and value
Demographics		
PtSex	Gender	Nominal–male, female
PtRace	Race/ethnic group	Nominal–Malay, Chinese, Indian, other races
PtAgeAtNotification	Age at notification	Numerical–age in year
Status before ACS event:		
Smoking status	Smoking status	Nominal–never, former (quit > 30 days), current (any tobacco use within last 30 days)
StatusAspirin Use	Status of aspirin used	Nominal–used < 7 days previously, used ≥ 7 days previously, Never
cDys	ACS comorbidities–dyslipidemia	Nominal–Yes/No
cHPT	ACS comorbidities–hypertension	Nominal–Yes/No
cDM	ACS comorbidities–diabetes	Nominal–Yes/No
cDM_OHA	Diabetes-OHA	Nominal–Yes/No
cDM_Insulin	Diabetes-Insulin	Nominal–Yes/No
cDM_NonPharmaTherapy	Diabetes–Non-pharmacology therapy/diet therapy	Nominal–Yes/No
cPremCVD	Family history of premature cardiovascular disease	Nominal–Yes/No
cMI	Myocardial infarction history	Nominal–Yes/No
cCAP	Documented CAD	Nominal–Yes/No
cAnginaMT2Wk	Chronic angina (≥ 2 weeks)	Nominal–Yes/No
cAnginaPast2Wk	New onset angina (< 2 weeks)	Nominal–Yes/No
cHeartFail	History of heart failure	Nominal–Yes/No
cLung	Chronic lung disease	Nominal–Yes/No
cRenal	Chronic renal disease	Nominal–Yes/No
cCerebrovascular	Cerebrovascular disease	Nominal–Yes/No
cPVascular	Peripheral vascular disease	Nominal–Yes/No
Clinical presentation and examination:		
AnginaEpisodeNo	Number of distinct episodes of angina in the past 24 h	Numerical
HeartRate	Heart rate at presentation	Numerical–beats/minute
BPSys	Blood pressure–systolic	Numerical–mmHg
BPDias	Blood pressure–diastolic	Numerical–mmHg
BMI	Body mass index (BMI)	Numerical
Electrocardiography (ECG):		
ECGAbnormTypeSTElev1	ECG abnormalities type: ST-segment elevation ≥ 1 mm	Nominal–True/False
ECGAbnormTypeSTElev2	ECG abnormalities type: ST-segment elevation ≥ 2 mm	Nominal–True/False
ECGAbnormTypeSTDep	ECG abnormalities type: ST-segment depression ≥ 0.5 mm	Nominal–True/False
ECGAbnormTypeTWave	ECG abnormalities type: *T*-wave inversion ≥ 1 mm	Nominal–True/False
ECGAbnormTypeBBB	ECG abnormalities type: bundle branch block (BBB)	Nominal–True/False
ECGAbnormLocationIL	ECG abnormalities location: Inferior leads	Nominal – True/False
ECGAbnormLocationAL	ECG abnormalities location: anterior leads	Nominal–True/False

Table 9 (continued)

Attribute name	Description	Data type and value
ECGAbnormLocationLL	ECG abnormalities location: lateral leads	Nominal–True/False
ECGAbnormLocationTP	ECG abnormalities location: true posterior	Nominal–True/False
ECGAbnormLocationRV	ECG abnormalities location: right ventricle	Nominal–True/False
Baseline investigations (obtained within 48 h from admission):		
TC	Lipid profile (fasting): total cholesterol	Numerical–mmol/L
HDLC	Lipid Profile (fasting): HDL-C	Numerical–mmol/L
LDLC	Lipid profile (fasting): LDL-C	Numerical–mmol/L
TG	Lipid profile fasting): triglyceride	Numerical–mmol/L
FBG	Fasting blood glucose	Numerical–mmol/L
HbA1c	Hemoglobin A1c or glycated hemoglobin test for diabetes	Numerical
LVEF	Left ventricular ejection fraction in %	Numerical–%
Clinical diagnosis at admission:		
ACSStratum	Acute coronary syndrome stratum	Nominal–STEM, NSTEMI, unstable angina
Pharmacological therapy given before admission:		
ASAPre	ASA given before admission	Nominal–Yes/No
ADPAPre	Pharmacological therapy	Nominal–Yes/No
GPRIPre	GP receptor inhibitor	Nominal–Yes/No
HeparinPre	Unfrac heparin	Nominal–Yes/No
LMWHPre	Low-molecular-weight heparin	Nominal–Yes/No
BBPre	Beta blocker	Nominal–Yes/No
ACEIPre	ACE inhibitor	Nominal–Yes/No
ARBPre	Angiotensin II receptor blocker	Nominal–Yes/No
StatinPre	Statin	Nominal–Yes/No
LipidLAPre	Other lipid lowering agent	Nominal–Yes/No
DiureticPre	Diuretics	Nominal–Yes/No
CalcAntagonistPre	Calcium antagonist	Nominal–Yes/No
OralHypoglyPre	Oral hypoglycemic agent	Nominal–Yes/No
InsulinPre	Insulin	Nominal–Yes/No
AntiArrPre	Anti-arrhythmic agent	Nominal–Yes/No
In-Hospital clinical outcomes:		
DayCCU	Number of overnight stays at CCU (in days)	Numerical
DayICU	Number of overnight stays at ICU/CICU (in days)	Numerical
TotalDayStay	Total number of overnight stays	Numerical
BleedingEpisodeCriteria	Bleeding complication (TIMI criteria)	Nominal–*Major* (Any intracranial bleed or other bleeding: 5 g/dL Hb drop), *Minor* (Non-CNS bleeding with 3–5 g/dL Hb drop), *Minimal* (Non-CNS bleeding, non-overt bleeding, < 3 g/dL Hb drop), *None*
PtOutcome	Patient outcome	Nominal–discharged, died
Target class attribute:		
Killip class	Killip classification code	Nominal–class I (no clinical signs of heart failure), class II (rales or crackles in the lungs, an S3, and elevated jugular venous pressure), class III (frank acute pulmonary oedema), class IV (cardiogenic shock or hypotension, and evidence of peripheral vasoconstriction)

References

1. WHO (2018) The top 10 causes of death. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 20 June 2018
2. Department of Statistics Malaysia (2018) Statistics on causes of death, Malaysia, 2018. Available via DOSM. https://www.dosm.gov.my/v1/index.php. Accessed 30 Jan 2019
3. Sweis RN, Jivan A (2018) Overview of coronary artery disease (CAD). MSD Manual Professional Version, Available: https://www.msdmanuals.com/professional/cardiovascular-disorders/coronary-artery-disease/overview-of-coronary-artery-disease. Accessed 20 Feb 2019
4. Hajar R (2017) Risk factors for coronary artery disease: historical perspectives. Heart Views 18(3):109–114
5. Rashid NA, Nawi AM, Khadijah S (2019) Exploratory analysis of traditional risk factors of ischemic heart disease (IHD) among predominantly Malay Malaysian women. BMC Public Health 19(4): 545
6. Narain R, Saxena S, Goyal AK (2016) Cardiovascular risk prediction: A comparative study of framingham and quantum neural network based approach. Patient Prefer Adherence 10:1259–1270. https://doi.org/10.2147/PPA.S108203
7. Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L (2017) An analytical method for diseases prediction using machine learning techniques. Comput Chem Eng 106:212–223
8. Katus H, Ziegler A, Ekinci O, Giannitsis E, Stough WG, Achenbach S, ..., Crea F (2017) Early diagnosis of acute coronary syndrome. Eur Heart J 38(41):3049–3055
9. Esfandiari N, Babavalian MR, Moghadam AME, Tabar VK (2014) Knowledge discovery in medicine: Current issue and future trend. Expert Syst Appl 41(9):4434–4463
10. Mohan S, Thirumalai C, Srivastava G (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 7:81542–81554
11. Amin MS, Chiam YK, Varathan KD (2019) Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform 36:82–93
12. Mathan K, Kumar PM, Panchatcharam P, Manogaran G, Varadharajan R (2018) A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Des Autom Embed Syst 22:225–242
13. Jaafar J, Atwell E, Johnson O, Clamp S, Wan Ahmad WA (2013) Evaluation of machine learning techniques in predicting acute coronary syndrome outcome. In: Research and Development in Intelligent Systems XXX. Springer, pp 321–333
14. Sun S (2015) An innovative intelligent system based on automatic diagnostic feature extraction for diagnosing heart diseases. Knowl-Based Syst 75:224–238
15. Nahar J, Imam T, Tickle KS, Chen YPP (2013) Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Syst Appl 40(1):96–104
16. Shilaskar S, Ghatol A (2013) Feature selection for medical diagnosis: evaluation for cardiovascular diseases. Expert Syst Appl 40(10):4146–4153
17. Phyu TZ, Oo NN (2016) Performance comparison of feature selection methods. In: MATEC Web of Conferences 42(06002). EDP Sciences, pp 1–4
18. Chin SP, Jeyaindran S, Azhari R, Wan Azman WA, Omar I, Robaayah Z, Sim KH (2008) Acute coronary syndrome (ACS) registry-leading the charge for National Cardiovascular Disease (NCVD) Database. Med J Malaysia 63(Suppl C):29–36
19. Wan Ahmad WA (ed) (2017) Annual report of the NCVD-ACS registry, 2014-2015. National Heart Association of Malaysia. Available: https://www.malaysianheart.org/?p=ncvd&a=1250
20. Wirth R, Hipp J (2000) CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Manchester, UK, pp 29–39
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
22. Jovic A, Brkic K, Bogunovic N (2015) A review of feature selection methods with applications. In: 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, pp 1200–1205
23. Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: A review. Egypt Inform J 19(3):179–189
24. Han J, Pei J, Kamber M (2012) Data mining: concepts and techniques, 3rd edn. Elsevier
25. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28
26. Breiman L (2001) Random forests. Mach Learn 45:5–32
27. Chaurasia V, Pal S (2013) Early prediction of heart diseases using data mining techniques. Carib J SciTech 1:208–217
28. Subanya B, Rajalaxmi R R (2014) Feature selection using Artificial Bee Colony for cardiovascular disease classification. In: International Conference on Electronics and Communication Systems (ICECS), pp. 1–6
29. Ismaeel S, Miri A, Sadeghian A, Chourishi D (2015) An extreme learning machine (ELM) predictor for electric arc furnaces' vi vharacteristics. In: IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), New York, pp 329–334
30. El-Bialy R, Salamay MA, Karam OH, Khalifa ME (2015) Feature analysis of coronary artery heart disease data sets. Proc Comput Sci 65:459–468
31. Nahar J, Imam T, Tickle K S, Garcia-Alonso D (2015) Medical knowledge based data mining for cardiac stress test diagnostics. In: 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE). IEEE, pp 1–7
32. Verma L, Srivastava S, Negi PC (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 40(7):1–7
33. Wiharto W, Kusnanto H, Herianto H (2017) Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis. Int J Electr Comput Eng (IJECE) 7(2): 1023–1031
34. Paul AK, Shill PC, Rabin MRI, Murase K (2018) Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. Appl Intell 48(7):1739–1756
35. Ali L, Khan S U, Arshad M, Ali S, Anwar M (2019) A multi-model framework for evalu-ating type of speech samples having complementary information about Parkin-son's disease. In: 2019 international conference on electrical, communication, and computer engineering (ICECCE). IEEE, pp 1–5
36. Reddy GT, Reddy MPK, Lakshmanna K, Rajput DS, Kaluri R, Srivastava G (2019) Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evol Intell 1–12
37. Latha CBC, Jeeva SC (2019) Improving the accuracy of prediction of heart dis-ease risk based on ensemble classification techniques. Inform Med Unlocked 16:100203

Nashreen Md Idris received her BCS(SE) degree from the Universiti Malaya in 2017. She is currently a master student at the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia. Her research interests include data mining, healthcare data analytics and software engineering.

Yin Kia Chiam received the Ph.D. in Computer Science and Engineering from the University of New South Wales, Australia. She is currently a senior lecturer at the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia. She has published in several reputable journals and conferences both locally and internationally. Her research focus is in the field of Software Engineering and Data Mining.

Kasturi Dewi Varathan is currently a senior lecturer at the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia. She holds a Ph.D. in Computer Science from the National University of Malaysia. Her main research focus is in the field of Data Analytics and Information Retrieval. Dr.Kasturi has published in several high-rank journals and conferences. She has served as visiting scientist to the Information Systems Research Group of the Faculty of Informatics, University of Lugano, Switzerland as well as research fellow at the Institute of Visual Informatics, National University of Malaysia. She is also a recipient of the prestigious Leadership in Innovation Fellowship award by the United Kingdom and Malaysian Government.

Wan Azman Wan Ahmad is currently Professor of Medicine and Cardiology Special Grade A at the Department of Medicine, Faculty of Medicine, Universiti Malaya, Malaysia. His devotion to cardiology and his passion to push the frontiers of cardiology is well known and proven in the research he does and in his numerous affiliation and involvement in international and national professional bodies. He has been awarded eleven Fellowships including Fellow of the European Society of Cardiology (FESC) and the Fellowship of the American College of Cardiology (FACC). He has been given numerous awards including the University of Malaya Eminent Scholar Award in 2013. He is the current President of the National Heart Association of Malaysia and Past Chairman of the Interventional Cardiovascular Society of Malaysia. To date he has been a member of 86 Advisory boards and 16 Clinical Practice Guideline. He has been the Primary Investigator of more than 60 studies. He has published more than 100 scientific papers, 253 abstracts and proceedings, 33 book chapters and 34 books/guidelines/reports.

Kok Han Chee is currently the Professor in Cardiology of Universiti Malaya, Malaysia. His main research interests included atrial fibrillation and heart diseases in diabetes mellitus. He has been involved in multiple clinical trials using novel anticoagulants for stroke prevention and novel drugs in diabetes mellitus to prevent cardiac death. He is also a council member in the National Heart Association of Malaysia.

Yih Miin Liew received the PhD degree from the University of Western Australia, Perth, Australia. She is currently a senior lecturer at the Department of Biomedical Engineering, Faculty of Engineering, in Universiti Malaya, Malaysia. She is active in medical imaging and image processing research for healthcare.