BMC Medical Research Methodology

RESEARCH

Open Access

Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records

Ying Hu[1,2†], Hai Yan[3†], Ming Liu[2,6], Jing Gao[4], Lianhong Xie[4], Chunyu Zhang[1], Lili Wei[4], Yinging Ding[5*] and Hong Jiang[1,2*]

Abstract

Background Electronic medical records (EMR)-trained machine learning models have the potential in CVD risk prediction by integrating a range of medical data from patients, facilitate timely diagnosis and classification of CVDs. We tested the hypothesis that unsupervised ML approach utilizing EMR could be used to develop a new model for detecting prevalent CVD in clinical settings.

Methods We included 155,894 patients (aged ≥ 18 years) discharged between January 2014 and July 2022, from Xuhui Hospital, Shanghai, China, including 64,916 CVD cases and 90,979 non-CVD cases. K-means clustering was used to generate the clustering models with $k = 2$, 4, and 8 as predetermined number of clusters k = 2, 4, and 8. Bayesian theorem was used to estimate the models' predictive accuracy.

Results The overall predictive accuracy of the 2-, 4-, and 8-classification clustering models in the training set was 0.856, 0.8634, and 0.8506, respectively. Similarly, the predictive accuracy of the 2-, 4-, and 8-classification clustering models in the testing set was 0.8598, 0.8659, and 0.8525, respectively. After reducing from 19 dimensions to 2 dimensions by principal component analysis, significant separation was observed for CVD cases and non-CVD cases in both training and testing sets.

Conclusion Our findings indicate that the utilization of EMR data can support the development of a robust model for CVD detection through an unsupervised ML approach. Further investigation using longitudinal design is needed to refine the model for its applications in clinical settings.

Keywords Machine learning, K-means clustering, Bayesian theorem, Cardiovascular diseases, EMR

†Ying Hu and Hai Yan are co-first authors.

*Correspondence:
Yinging Ding
dingyy@fudan.edu.cn
Hong Jiang
jianghong_@fudan.edu.cn
[1]Department of Cardiology, National Clinical Research Center for Interventional Medicine, Shanghai Institute of Cardiovascular Diseases, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[2]Shanghai Engineering Research Center of AI Technology for Cardiopulmonary Diseases, Zhongshan Hospital, Fudan University, Shanghai 200032, China
[3]Department of General Surgery, Center for Bariatric and Hernia Surgery, Huashan Hospital, Fudan University, Shanghai 200040, China
[4]Shanghai Xuhui Central Hospital, Zhongshan-Xuhui Hospital, Fudan University, Shanghai 200031, China
[5]Department of Epidemiology, School of Public Health, and Key Laboratory of Public Health Safety of Ministry of Education, Fudan University, Shanghai 200032, China
[6]Department of Health Management Center, Zhongshan Hospital, Fudan University, Shanghai 200032, China

Introduction

Cardiovascular diseases (CVD) are the leading cause of death globally, accounting for approximately18 million deaths annually [1], and this number is expected to rise to 23.6 million by 2030. In China, two out of every five deaths are attributed to CVD, affecting an estimated 330 million people [2]. Traditional statistics-based prediction tools for future CVD [3], such as the Framingham Risk Score [4], Systematic Coronary Risk Evaluation [5] and QRISK scores [6, 7], are commonly used in primary prevention settings. However, these methods use a common set of risk factors and the overall accuracy remains unsatisfactory and limited application for early detection [3, 8]. Clinicians diagnose CVD by evaluating the clinical symptoms and signs of patients and using auxiliary diagnostic methods, such as blood tests and imaging (non-invasive and invasive) examinations. These procedures are expensive, time-consuming and often requires specialized expertise. Asymptomatic individuals may be overlooked during routine physical examinations or hospitalization for other unrelated diseases. An automated CVD detection tool that help identify high-risk individuals quickly and accurately is needed.

Machine learning (ML), a technique used to realize artificial intelligence, broadens the scope of traditional statistics by identifying nonlinear relationships and higher-order interactions among numerous variables. It can be categorized into supervised and unsupervised learning [8]. Supervised ML build models by associating a certain set of features with known outcomes (labeled data) to predict outcomes for new data, including naive Bayes, random forest, Logistic regression, support vector machines (SVM), K-Nearest Neighbor (KNN), artificial neural network [9] and genetic algorithm [10]. Unsupervised ML, on the other hand, focuses on identifying the underlying patterns in unlabeled data, including clustering, association and dimensionality reduction. Clustering analysis is a process that involves the identification of distinct subgroups within extensive and intricate data. K-means clustering is unsupervised approach to group objects into K number of clusters number of clusters based on their features. This technique ensures that each data point assigned to a specific cluster is in closer proximity to the centroid of the cluster compared to all other clusters [11]. Dimension reduction is a process of reducing high-dimensional data to a low-dimensional representation is achieved while preserving the inherent changes and structures in the original full-dimensional data. A recent study [12] employed unsupervised ML approach, specifically multiple kernel learning-based dimension reduction and K-means clustering, to combine echocardiographic data and clinical parameters to phenotype heart failure patients.

ML has been increasingly utilized to improve the accuracy and speed of CVD prediction and diagnosis [13]. Nevertheless, the majority of ML-based prediction models are built on community-based populations that share similar features [14–17], and the prevalence and severity of CVD may also affect the models' accuracy, limiting their clinical application [8]. Importantly, electronic medical records (EMR) as a digital version of paper records were initially introduced in hospitals to improve healthcare efficiency and promote patient care. EMR contain a wide variety of data, such as demographics, diagnoses, medications, laboratory and imaging tests. With the growing availability of rich and large sample size data recorded in EMR, there is growing interest to translate these data into clinical practices through the application of ongoing machine learning and AI advancements [18]. EMR-trained ML models have the potential in CVD risk prediction by integrating a range of medical data from patients, facilitate timely diagnosis and classification of CVDs [19]. Nevertheless, there have been limited study conducted on the EMR data for constructing CVD prediction models [20, 21].

Thus, using EMR data, we employed K-means clustering and Bayesian theorem to construct a model that can accurately identify the patients with high probability of having CVD in clinical settings. K-means clustering was utilized to generate the clustering models, and Bayesian theorem was utilized to estimate their predictive accuracy. Our work provides an example demonstrating the application of EMR-based ML to develop a prediction model for assessing the likelihood of having the CVD.

Methods

Data source

The study obtained data from the electronic medical record (EMR) system and clinical laboratory information system (LIS) of Xuhui Central Hospital, an affiliate of Fudan University in China. The data consisted of diagnostic information and laboratory test results for adult patients who were discharged from January 2014 to July 2022. This study was performed in accordance with the guidelines of the Declaration of Helsinki. The study design was approved by the Ethics Committee of Shanghai Xuhui Central Hospital (approval no: 2023033), and the institutional review board waived the requirement to obtain the informed consent. The medical record number, gender, age and ICD-10 diagnostic information were extracted from the EMR system using SQL statements. A total of 155 894 patients were included.

The primary outcome of this study was determining the presence of CVD in each subject. CVD was defined based on the primary symptoms outlined in the International Classification of Diseases, 10th Revision (ICD-10) diagnostic information). These symptoms including

"coronary heart disease arrhythmia", "coronary artery insufficiency", "coronary heart disease", "coronary artery slow flow", "coronary artery bypass surgery status", "coronary artery stent thrombosis", "coronary artery stent implantation status", "coronary artery stenosis", "coronary artery fistula", "coronary atherosclerosis", "coronary atherosclerotic heart disease" [22]. Patients exhibiting the aforementioned symptoms were categorized as cases of CVD (n=64916) (Table 1), while the other patients who did not display these symptoms were classified as non-CVD cases (n=90979).

We searched the LIS system for various laboratory test results upon admission, including total cholesterol (TC), triglyceride (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), blood glucose, creatine kinase (CK), CK-MB isoenzyme (CK-MB), troponin (Tn), myoglobin (Mb), angiotensin (I/II), aldosterone, hemorheology, brain natriuretic peptide (BNP), glycosylated hemoglobin (GHB), homocysteine (HCY), tumor necrosis factor (TNF), interleukin, C-reactive protein (CRP), D-dimer, fibrinogen, creatinine, urea nitrogen, uric acid, glomerular filtration rate (GFR), plasma viscosity, erythrocyte aggregation index, hemoglobin, blood sodium, blood potassium, and other relevant test results.

Data preprocessing and variable selection

After data cleaning, the incomplete, incorrect, inaccurate, and irrelevant parts of 155 894 patients' data were identified and were replaced, modified, or deleted. Due to the inherent characteristics of the mining process, the vast majority of data attributes utilized within this method were of a quantitative type, specifically integer or real number data. The analysis eliminated gender as a variable due to its binary nature. The process of selecting predictor variables (features) was conducted by three medical experts with experience in the diagnosis of CVD selected the predictor variables (features) based on comprehensive review of relevant literature. Also, features with missing data in ≥20% of patients were removed, and features with missing data for <20% of the patients were subjected to multiple imputation. The features of these deletions included angiotensin, aldosterone, brain natriuretic peptide, homocysteine, free triiodothyronine, free tetraiodothyronine, and thyroid stimulating hormone.

The preliminary list focused on 15 variables that are clearly implicated in the pathogenesis of CVD [23], including blood lipids (TC, TG, HDL, LDL), cardiac markers (CK, CK-MB, Mb, Tn), renal function (creatinine, urea nitrogen, uric acid, GFR) and blood glucose markers (glucose, GHB). Four additional variables that have previously been associated with CVD but lack robust clinical evidence, were included in this study. These variables included coagulation markers such as D-dimer and fibrinogen as well as other biomarkers including hemoglobin, blood sodium, blood potassium). Finally, 19 features were selected as input for the ML algorithm. Table 2 shows the description of selected variables. Z-score normalization was used to standardize the numerical variables.

Table 1 Data overview of main CVD diseases

ICD-10 code	Main CVD symptoms	No. of patients
I25.104	Arrhythmia type of coronary heart disease	892
I24.800×001	Coronary insufficiency	11
I25.901	Coronary heart disease	186
I25.800×007	Slow coronary flow	1
Z95.101	Post coronary artery bypass graft status	57
I24.001	Coronary stent thrombosis	3
Z95.501	Post coronary stent implantation status	409
I25.101	Coronary stenosis	169
I25.800×005	Coronary artery fistula	7
I25.102	Atherosclerosis	1226
I25.103	Coronary atherosclerotic heart disease	61,951
I25.900 A	Coronary ischemia	1
Z98.800×403	Post coronary angiography	2
Z95.500×002	Post-coronary angioplasty status	1
Total		64,916

Statistical machine learning analysis

The entire dataset was randomly split into two non-overlapping sets: training set (90%, n=140304) and testing set (10%, n=15590). We ran our unsupervised ML algorithm on the training set first to generate the prediction model (i.e., create clusters), and then tested the models using the features of the testing set to assess their ability to accurately infer the class labels for the patients in the testing set. The estimation of the predictive accuracy of the clusters and models was afterwards conducted utilizing the Bayesian theorem. The dimensionality reduction approach of principal component analysis (PCA) was additionally employed to reduce the number of features from 19 to 2 dimensions in both the training and testing sets. This allowed for the visualization of the sample results projected onto the first two components [24]. The principal components are the continuous solutions derived from the discrete cluster membership markers for K-means clustering, PCA can serve as a tool to evaluate the 2-classification clustering model from a different angle [25]. The modeling process is depicted in Fig. 1.

K-means clustering and bayesian theorem

K-means clustering was used to classify the data-set into a fixed number (K) of distinct clusters. We selected k=2, 4, and 8 as predetermined number of clusters and iterated 1 million times to guarantee the stability of the results. The input of the model was a normalized vector of 19 parameters, and the output was whether CVD was

Table 2 Dataset features description

No	Feature name	Type	Description	Upper threshold	Lower threshold
1	ID	Integer	Id number		
2	Date of Birth	Integer	Max = 23/04/1906; min = 22/04/1997		
3	Gender	Integer	Men: 1; women: 2		
4	Diagnosis	Text	Text description		
5	ICD-10 Code	Integer	ICD-10 number		
6	Total cholesterol (mmol/l)	Integer	CVD: MEAN = 3.9481; non-CVD: MEAN = 4.3562	5.7	/
7	Triglyceride (mmol/l)	Integer	CVD: MEAN = 1.3477; non-CVD: MEAN = 1.4385	1.8	/
8	High density lipoprotein (mmol/l)	Integer	CVD: MEAN = 1.1367; non-CVD: MEAN = 1.1070	/	0.8
9	Low density lipoprotein (mmol/l)	Integer	CVD: MEAN = 2.0539; non-CVD: MEAN = 2.3502	3.36	/
10	Urea nitrogen (mmol/l)	Integer	CVD: MEAN = 8.0624; non-CVD: MEAN = 6.7867	8.3	/
11	Creatinine (μmol/l)	Integer	CVD: MEAN = 92.5623; non-CVD: MEAN = 87.9544	116	/
12	Uric acid (μmol/l)	Integer	CVD: MEAN = 342.9268; non-CVD: MEAN = 317.6061	429	/
13	Glomerular filtration rate (ml/min)	Integer	CVD: MEAN = 66.3051; non-CVD: MEAN = 87.03	/	80
14	Blood glucose (mmol/l)	Integer	CVD: MEAN = 6.3598; non-CVD: MEAN = 6.1572	6.2	/
15	Glycosylated hemoglobin (%)	Integer	CVD: MEAN = 6.3972; non-CVD: MEAN = 6.4832	6.0	/
16	Creatine kinase (U/L)	Integer	CVD: MEAN = 92.2079; non-CVD: MEAN = 101.8716	134	/
17	Creatine kinase isoenzyme (ng/ml)	Integer	CVD: MEAN = 15.3261; non-CVD: MEAN = 16.2749	5.04	/
18	Troponin (ng/ml)	Integer	CVD: MEAN = 0.0826; non-CVD: MEAN = 0.0485	1.0	/
19	Myoglobin(ng/ml)	Integer	CVD: MEAN = 73.1981; non-CVD: MEAN = 68.5344	100	/
20	D-dimer (mg/L)	Integer	CVD: MEAN = 2.0789; non-CVD: MEAN = 2.1586	1.0	/
21	Fibrinogen (g/L)	Integer	CVD: MEAN = 3.6457; non-CVD: MEAN = 3.7269	4.0	/
22	Hemoglobin (g/L)	Integer	CVD: MEAN = 116.1850; non-CVD: MEAN = 118.77	150	/
23	Blood sodium (mmol/l)	Integer	CVD: MEAN = 140.6348; non-CVD: MEAN = 140.9027	145	136
24	Blood potassium (mmol/l)	Integer	CVD: MEAN = 4.0852; non-CVD: MEAN = 3.9647	5.33.5	3.5

present. We used the characteristics of K-means clustering to classify the disease, and classify the patients with or without CVD into two types for clustering. Ideally, patients with CVD should be clustered in several of the three clustering models of 2-, 4-, and 8-classification, while patients without CVD should be clustered in other clusters. However, in reality, it is impossible to achieve the ideal state. Our data only covered the major symptoms of patients who were diagnosed at a given time in the hospital, and they may only represent their occasional situation. Furthermore, not all of 19 features are strongly related to CVD pathogenesis. In practical situations, the more uneven the distribution of CVD and non-CVD ratios in each cluster, the better it is for the cluster to determine whether CVD is present. The more such clusters there are in the entire clustering model, the better it is for the entire clustering model to determine whether CVD is present.

The model was constructed to the accurate classification of patients, enabling to ascertain their disease status (i.e., CVD or non-CVD) with 100% probability. Therefore, after calculating the proportion of CVD in each cluster of the clustering model, the prediction accuracy of a single cluster in the three clustering models was calculated by using the inverse probability principle of Bayesian theorem, and then the overall prediction accuracy of the three clustering models was calculated by using Bayesian theorem. The predictive accuracy of the clustering model was determined by dividing the sum of the size of the bigger group in each cluster by the total number of samples. The specific method to calculate the accuracy by using Bayesian theorem was as follows:

The predictive probability for each cluster by the Bayesian theorem was:

$$P_n = \frac{X_n^{max}}{X_n^{all}}$$

X_n^{max}refers to the size of the bigger group (CVD cases or non-CVD cases) in the cluster, and X_n^{all}refers to the total number patients in the cluster.

The overall prediction accuracy (model performance) of our model by Bayesian theorem was:

$$P_{all} = \sum_{i=1}^{n} \frac{X_n^{max}}{X_n^{all}} \times \frac{X_n^{all}}{X_{all}^{all}} = \frac{\sum_{1}^{n} X_n^{max}}{X_{all}^{all}}$$

X_{all}^{all} refers to the number of all subjects in the sample.

This shows that the predictive accuracy of the clustering model is determined by dividing the sum of the size of the bigger group in each cluster by the total number of samples.

Model performance

The predictive probability of detecting the existence of CVD for a single cluster was calculated as the number of

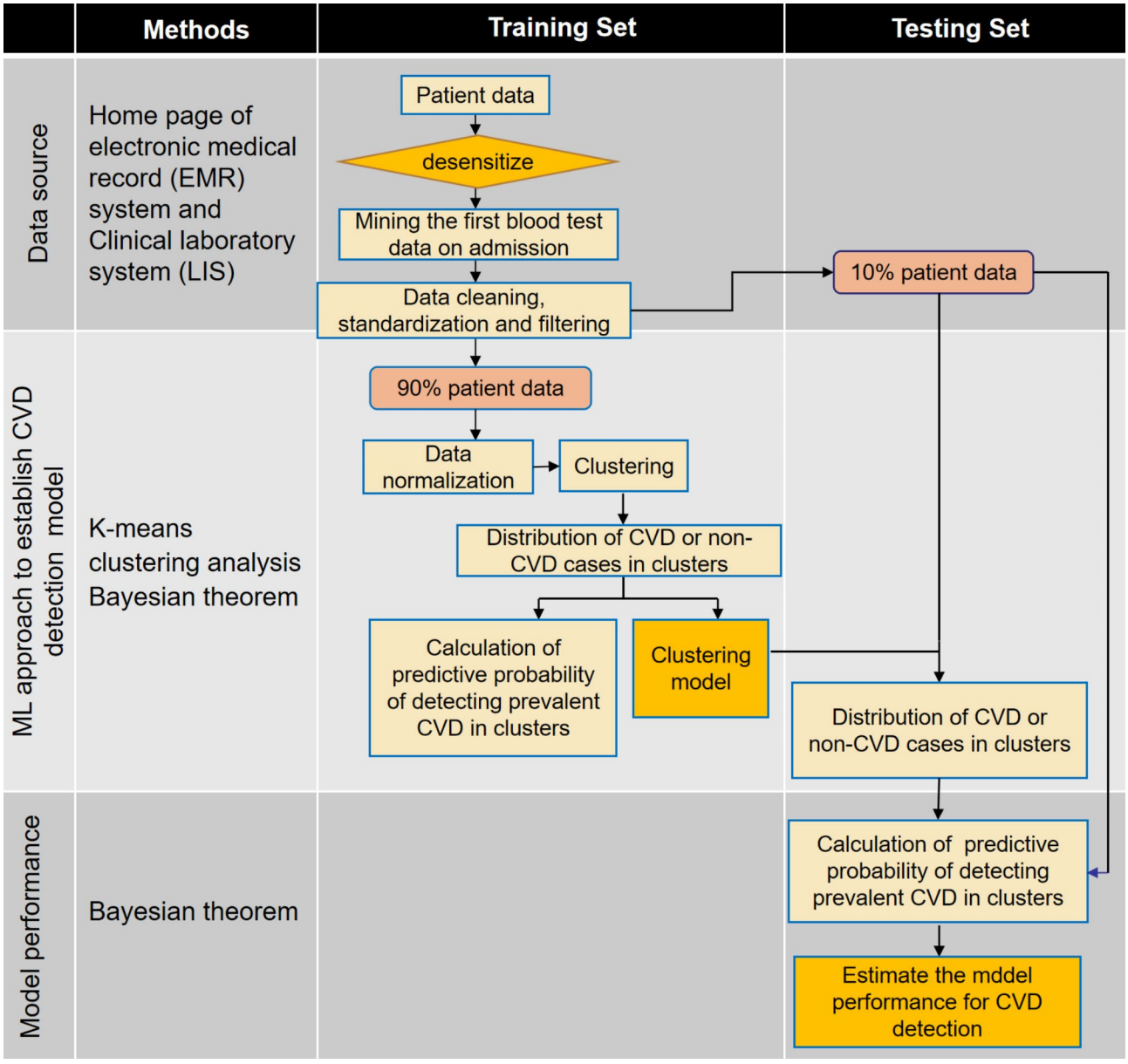

Fig. 1 Flow chart of ML approach to establish CVD detection model

patients with prevalent CVD divided by the total number of patients. The predictive probability of detecting prevalent CVD in each cluster was obtained from $k=2$, 4, and 8 classifications, respectively. After calculating the proportions of CVD and non-CVD cases in each cluster from $k=2$, 4, and 8 classifications, the predictive accuracy of each cluster was calculated by Bayesian theorem. We calculated the predictive accuracy (performance) of the overall model, which is equivalent to the predictive accuracy of all single clusters as shown above.

Comparisons of K-means clustering with other ML algorithms

We conducted a comparative experiment with three traditional ML methods to evaluate the performance of our K-means clustering approach. The models included in this comparison were SVM, K-Nearest Neighbor (KNN), and Logistic regression. After establishing the models, we calculated area under the curve (AUC) of the models separately. Finally, we plotted the Receiver Operating Characteristic (ROC) curve. AUC served as the main indicator of model performance.

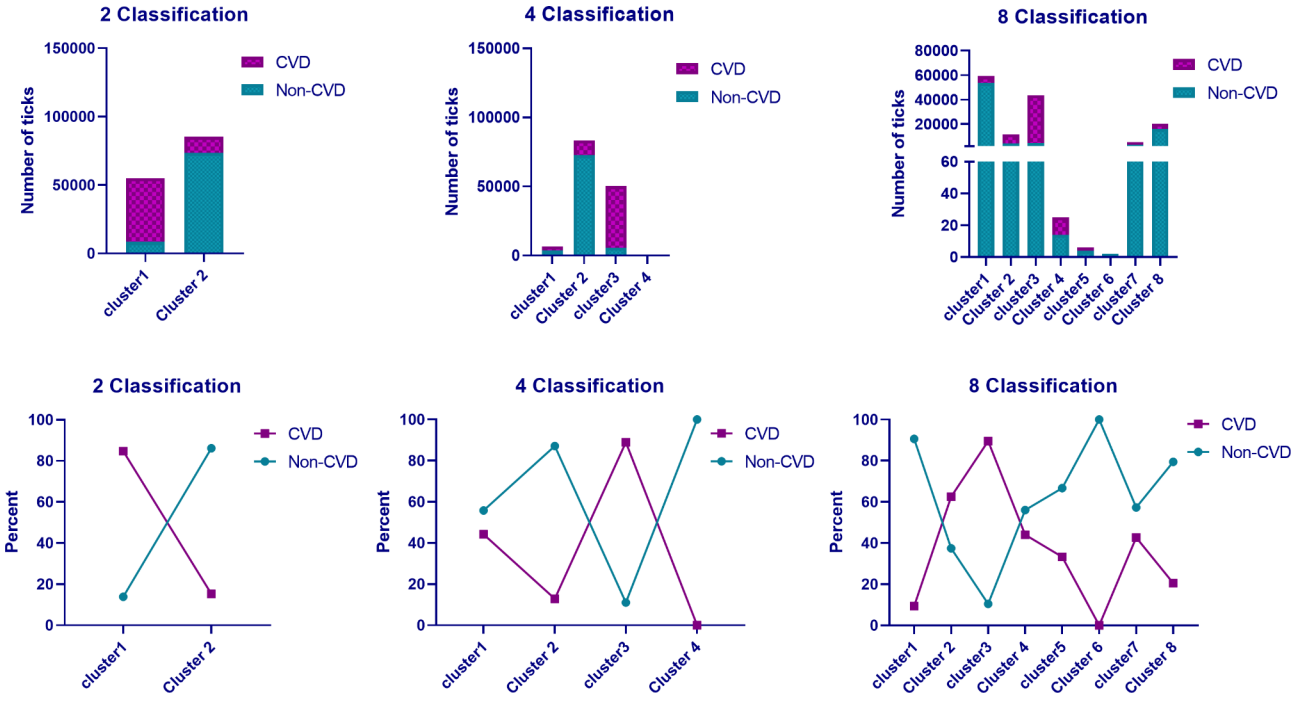

Fig. 2 Distribution of CVD and non-CVD cases in each cluster with different predetermined number of clusters in the training set

Table 3 Distribution of CVD and non-CVD cases in each cluster with different predetermined number of clusters in the training set

Predetermined No. clusters	Presence of CVD	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Total
k = 2	Non-CVD	8402	73,478							81,880
	CVD	46,619	11,805							58,424
	Total	55,021	85,283							140,304
	% of CVD	0.8473	0.1384							
k = 4	Non-CVD	3650	72,690	5538	2					81,880
	CVD	2889	10,743	44,792	0					58,424
	Total	6539	83,433	50,330	2					140,304
	% of CVD	0.4418	0.1288	0.8900	0					
k = 8	Non-CVD	53,631	4367	4550	14	4	2	3025	16,287	81,880
	CVD	5554	7286	39,100	11	2	0	2255	4216	58,424
	Total	59,185	11,653	43,650	25	6	2	5280	20,503	140,304
	% of CVD	0.0938	0.6252	0.8958	0.4400	0.3333	0	0.4271	0.2056	

Results

Characteristics of study subjects

Of 155 894 patients included, we filtered out 41.64% who already experienced a CVD outcome (during or before baseline). The remaining patients (90 979) did not experience any CVD outcome. Coronary atherosclerotic heart disease was the most common CVD (61 951 patients), followed by atherosclerosis (1 226 patients) and arrhythmia type of coronary heart disease (892 patients). Table 1 shows the number of patients according to different CVD symptoms.

Predictive probability of each cluster in 2-, 4-, and 8-classification clustering models

K-means clustering was used to classify the patients in the training set, with 2, 4, and 8 chosen as the predetermined number of clusters. As shown in Fig. 2; Table 3. In the *2-classification clustering model*, the predictive probability of detecting prevalent CVD in clusters 1 and 2 were 0.8473 and 0.1384, respectively. In the *4-classification clustering model*, the predictive probability of detecting prevalent CVD in clusters 1, 2, 3 and 4 were 0.4418, 0.1288, 0.8899 and 0, respectively. In the *8-classification clustering model*, the predictive probability of detecting prevalent CVD in clusters 1, 2, 3, 4, 5, 6, 7 and 8 were 0.0938, 0.6252, 0.8958, 0.4400, 0.3333, 0, 0.4271, and 0.2056, respectively. For each clustering model, the

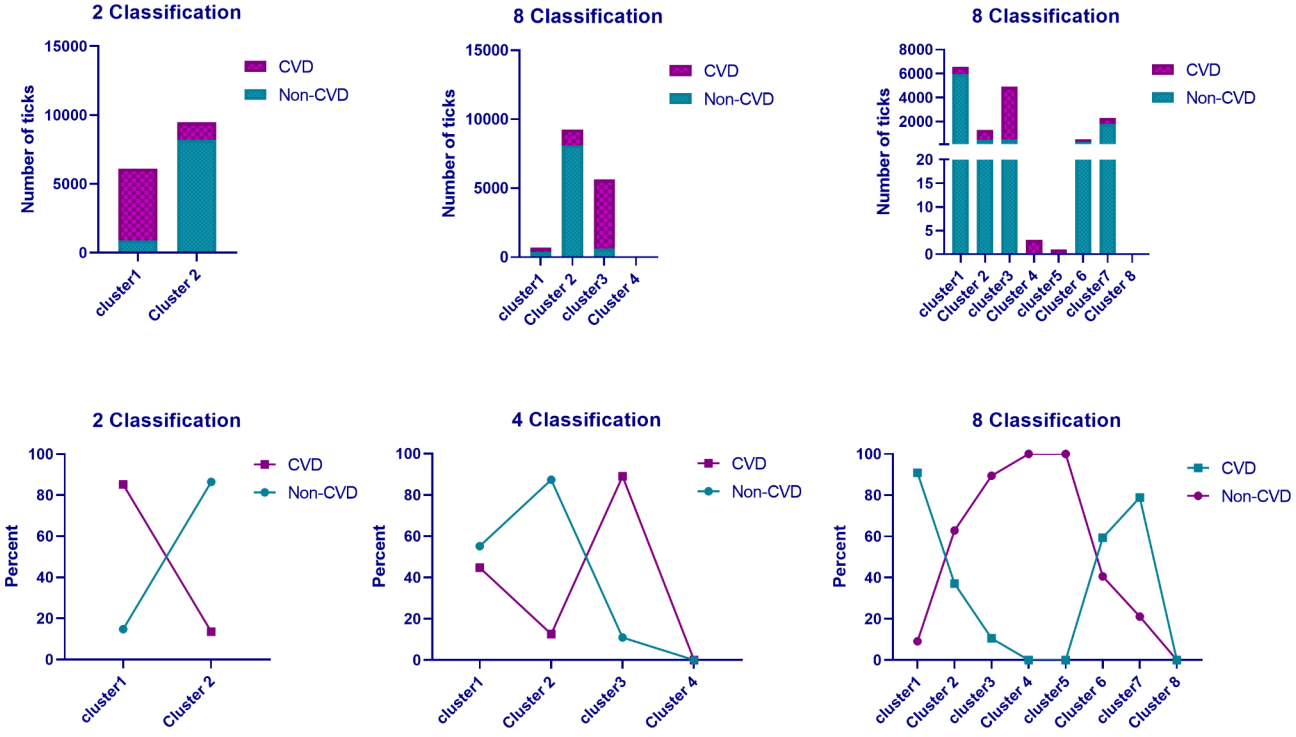

Fig. 3 Distribution of CVD and non-CVD cases in each cluster with different predetermined number of clusters in the testing set

Table 4 Distribution of CVD and non-CVD cases in each cluster with different predetermined number of clusters in the testing set

Predetermined No. clusters	Presence of CVD	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Total
k = 2	Non-CVD	907	8191							9098
	CVD	5213	1279							6492
	Total	6120	9470							15,590
	% of CVD	0.8518	0.1351							
k = 4	Non-CVD	377	8105	616	0					9098
	CVD	306	1169	5017	0					6492
	Total	683	9274	5633	0					15,590
	% of CVD	0.4480	0.1261	0.8906	0					
k = 8	Non-CVD	5978	476	518	0	0	311	1815	0	9098
	CVD	603	806	4381	3	1	213	485	0	6492
	Total	6581	1282	4899	3	1	524	2300	0	15,590
	% of CVD	0.0916	0.6287	0.8943	1.0000	1.0000	0.4065	0.2109		

cluster with the highest probability was the one most likely to have prevalent CVD.

The clustering models were further evaluated in the testing set. As shown in Fig. 3; Table 4, in the *2-classification clustering model*, the predictive probability of detecting prevalent CVD in clusters 1 and 2 were 0.8518 and 0.1351, respectively. In the *4-classification clustering model*, the predictive probability of detecting prevalent CVD in clusters 1, 2, 3 and 4 were 0.4480, 0.1261, 0.8906, and 0, respectively. In the *8-classification clustering model*, the predictive probability of detecting prevalent CVD in clusters 1, 2, 3, 4, 5, 6, 7 and 8 were 0.0916, 0.6287, 0.8943, 1, 1, 0, 0.4065 and 0.2109, respectively.

It should be noted that in the 4- and 8-clustering models, two clusters accounting for the majority of the total samples provided the main information needed to determine whether or not CVD was present, whereas other clusters accounting for a relatively small proportion of the overall samples provided minimal information.

Model performance of 2-, 4-, and 8- classification clustering models

Bayesian theorem was used to assess the 2-, 4-, and 8-classification clustering models' predictive accuracy as the model performance. The overall predictive accuracy of the 2-, 4-, and 8-classification clustering models in the training set was 0.856, 0.8634, and 0.8506, respectively,

Table 5 Comparative model performance in the testing sets

Model	ACC
KNN	0.8461
Logistic Regression	0.7992
SVM	0.8194
K-mean	0.8634

while the predictive accuracy of the 2-, 4-, and 8-classification clustering models in the testing set was 0.8598, 0.8659, and 0.8525, respectively (Table 5). Here, all values from the testing and evaluation sets were similar and above 0.85, showing that the models had good performance in detecting the CVD.

Clustering visualization

Because predictive accuracy was not dependent on the number of classifications as above showed, 2-classification clustering model is simplified and thus optimal. PCA was conducted to reduce 19 dimensions (features) down to two dimensions. PCA plots of the samples projected onto the first two principal components in the training and testing sets are shown in Figs. 4 and 5, respectively.

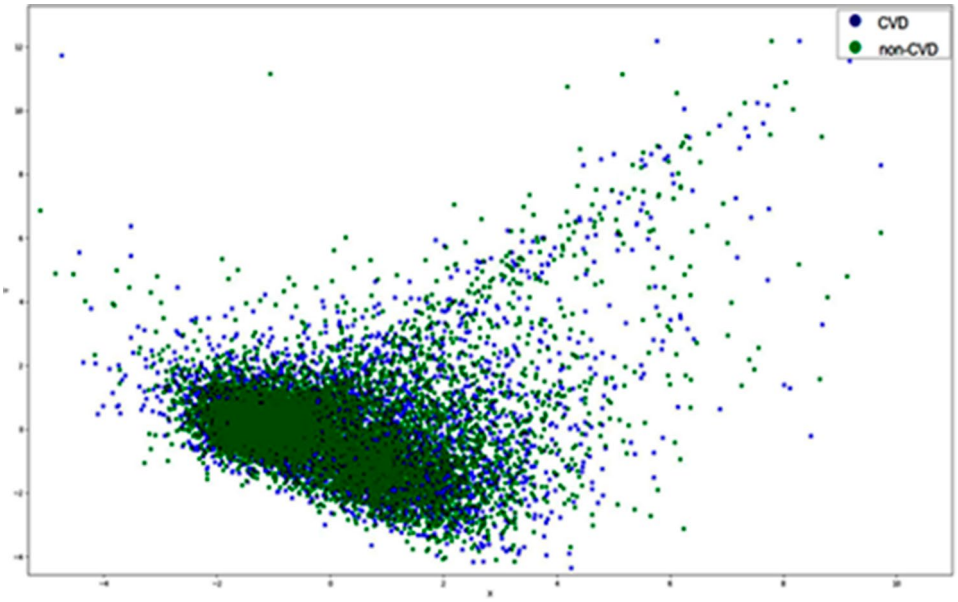

Fig. 4 Principal component analysis (PCA) of the training set. PCA plot with samples plotted in two dimensions using their projections on the first two principal components

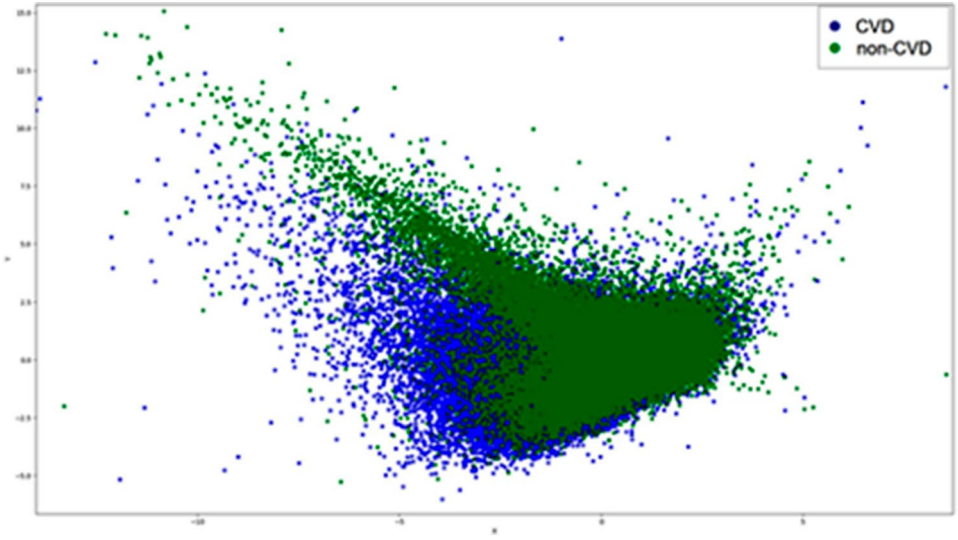

Fig. 5 Principal component analysis (PCA) of the testing set. PCA plot with samples plotted in two dimensions using their projections on the first two principal components

Table 6 Model performance with different predetermined number of clusters in the training and testing sets, respectively

Predetermined No. of clusters	Model performance	
	Training set	Testing set
$k=2$	0.8560	0.8598
$k=4$	0.8634	0.8659
$k=8$	0.8506	0.8525

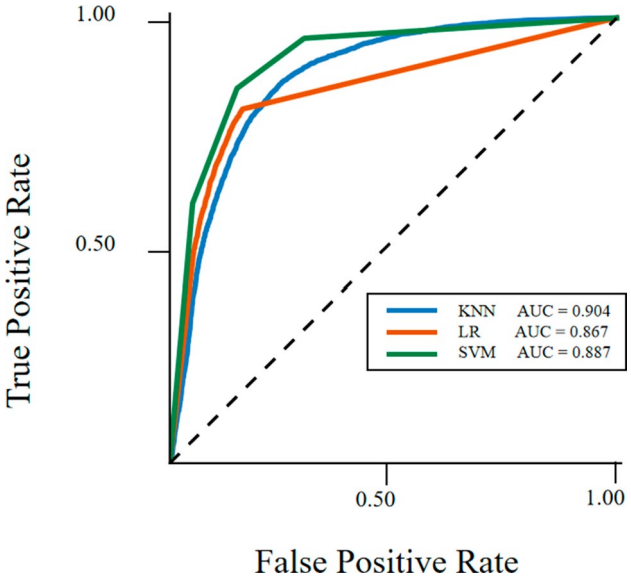

Fig. 6 ROC Curve for KNN, Logistic regression and SVM models

Significant separation was observed for CVD cases and non-CVD cases in both training and testing sets.

Performance of other models

The evaluation of models of KNN, SVM and Logistic regression was based on the testing set, and the results are presented in Table 6. The predictive accuracy for each model was as follows: K-means clustering achieved the highest accuracy of 0.8598, followed by KNN with a predictive accuracy of 0.846, SVM with a predictive accuracy of 0.819, and Logistic regression with a predictive accuracy of 0.7992 (Fig. 6).

Discussion

In this study, the data retrieved from the EMR was employed to construct a CVD detection model using unsupervised ML algorithm and subsequently assessed its predictive accuracy using the Bayesian theorem. Our study confirms the efficacy of unsupervised ML as a new approach for identifying individuals at high-risk of having CVD by utilizing routine blood tests conducted during physical examinations or hospitalization for other medical conditions. This can assist healthcare providers in assessing the necessity for additional health examinations or appropriate treatment, thereby facilitating early detection of CVD and reducing unnecessary medical expenses.

Unsupervised clustering algorithms, which need no labeling the input data, have proven to be useful in disease detection, diagnosis and classification [26]. In a recent work, hierarchical clustering analysis was used to evaluate numerous clinical variables and discovered new clinical phenotypes of atrial fibrillation [27]. The other study utilized K-means clustering to detect the varied etiology and prognosis of heart failure with preserved ejection fraction [28]. Our investigation showed that by extracting information from underutilized EMR data, the K-means clustering models surpassed the performance of SVM, KNN and Logistic regression models, with a predictive accuracy of over 85% in both the training and testing sets. Our findings suggest that unsupervised ML approach may yield novel tools in the detection of CVD with high accuracy. Furthermore, since the patient's data may be obtained from the EMR without the necessity of gathering additional health information in the context of limited medical expenditures, the adoption of this strategy is simple and efficient.

Various CVD guidelines recommend different CVD risk prediction tools. The most commonly used tool is Framingham risk score, which incorporate age, sex, diabetes, smoking, systemic blood pressure, and body mass index [29]. The QRISK2 scores, which is another frequently used prediction tool, incorporate many factors such as age, gender, race, blood pressure, diabetes, family history of coronary heart disease, chronic renal disease, blood lipids, rheumatoid arthritis, medication use, weight, smoking, etc [30]. However, ML-based prediction models often incorporate a diverse array of variables. An ML-based model for CVD prediction was developed using a dataset from the UK BioBank, which consisted of 423,604 CVD-free patients. The model was built using 473 variables [31]. However, due to the lack of a solid pathological basis and the inability of professionals to recognize it, this condition is rarely used in clinical settings. The 19 variables in our selection from EMR data was chosen based on their clinical significance. Specifically, TC, TG, HDL, and LDL are key components of blood lipid profiles. Glucose and GHB are linked to diabetes, whereas creatinine, urea nitrogen, urea nitrogen, and GFR are associated with chronic kidney disease. Mb, Tn, and CK-MB are important in diagnosing coronary heart disease since their levels are typically elevated in those with acute coronary syndrome. The current guidelines incorporate these variables, but do not include D-dimer, fibrinogen, hemoglobin, blood sodium, and blood potassium [32, 33]. It has been noted that the coagulation indicators D-dimer and fibrinogen exhibit an elevation during thromboembolism. During CVD events, the blood's coagulation status is shown to

be hypercoagulable as a result of activation of coagulation mechanisms [34]. Hemoglobin as a indicator for blood viscosity, and an increase in blood viscosity has been linked to CVD events [35]. Elevated sodium levels have an direct influence on the progression of hypertension, which is considered a notable risk factor for ischemic heart disease, stroke, and others [36]. According to previous reports, serum potassium levels was associated with CVD events and mortality [37]. Collectively, we believe that these variables may possess some pathological foundations that contribute to the development of CVD. Therefore, our model may serve as a useful model in assessing the likelihood of having the CVD.

Several limitations should be acknowledged. First, this was a cross-sectional analysis of input features and prevalent CVD status recorded in EMR, the temporal order of causality could not be determined. Second, this was a single institution, our models should be externally validated. In addition, we focused on variables that are often recorded in EMR, other major CVD risk factors such as BMI and family history of CVD were not incorporated in analysis as they are not consistently recorded in EHR, however, the prediction accuracy as estimated by Bayesian theory was deemed satisfactory, and thus findings should not be severely affected.

In conclusion, this study demonstrates the application of a ML approach that integrates K-means clustering and Bayesian theorem with EMR data to develop an automated model for evaluating the likelihood of having the CVD. Additional longitudinal investigations including more characteristics (e.g., comorbidities, medication use, and CVD events) across several institutions are needed to improve the model's accuracy and facilitate its potential implementation applications in clinical context.

Acknowledgements
Not applicable.

Author contributions
Ying Hu and Hong Jiang contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ming Liu, Jing Gao and Lianhong Xie. Data curation were managed by Lili Wei and Chunyu Zhang. The first draft of the manuscript was written by Ying Hu and Hai Yan. The review and editing were completed by Hong Jiang and Yinging Ding. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by Shanghai Aging and Maternal and Child Health Research Project (No.2020YJZX0141); Clinical Special Project of Shanghai Municipal Health Commission, China(No.202040083). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
The datasets used and analyzed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study was performed in accordance with the guidelines of the Declaration of Helsinki. The study design was approved by the Ethics Committee of Shanghai Xuhui Central Hospital (approval no.: 2023033), and the institutional review board waived the requirement to obtain the informed consent.

Consent for publication
All authors have approved for its publication.

Competing interests
The authors declare no competing interests.

Received: 18 October 2023 / Accepted: 25 November 2024
Published online: 19 December 2024

References

1. RuaN Y, Guo Y, Zheng Y, et al. Cardiovascular disease (CVD) and associated risk factors among older adults in six low-and middle-income countries: results from SAGE Wave 1. BMC Public Health. 2018;18(1):778. https://doi.org/10.1186/s12889-018-5653-9.
2. Summary of China Cardiovascular Health and Diseases Report 2020. Chin Circulation J. 2021;36(06):521–45. https://doi.org/10.3969/j.issn.1000-3614.2021.06.001.
3. Dimopoulos A C, Nikolaidou M, Caballero F F, et al. Machine learning methodologies versus cardiovascular risk scores in predicting disease risk. BMC Med Res Methodol. 2018;18(1):179. https://doi.org/10.1186/s12874-018-0644-1.
4. Greenland P, Alpert J S, Beller G A, et al. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice guidelines. Circulation. 2010;122(25):e584–636. https://doi.org/10.1016/j.jacc.2010.09.001.
5. Piepoli M F, Hoes A W, Agewalls S, et al. 2016 European guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J. 2016;37(29):2315–81. https://doi.org/10.1093/eurheartj/ehw106.
6. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336(7659):1475–82. https://doi.org/10.1136/bmj.39609.449676.25.
7. Hippisley-Cox J, Coupland C. Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. Bmj, 2017, 357(j2099). https://doi.org/10.1136/bmj.j2099.
8. Shu S, Ren J. Clinical application of machine learning-based Artificial Intelligence in the diagnosis, prediction, and classification of Cardiovascular diseases. Circ J. 2021;85(9):1416–25. https://doi.org/10.1253/circj.CJ-20-1121.
9. Trayanova N A, Popescu D M, SHADE JK. Machine learning in Arrhythmia and Electrophysiology. Circ Res. 2021;128(4):544–66. https://doi.org/10.1161/CIRCRESAHA.120.317872.
10. Ordikhani M, Saniee Abadeh M, Prugger C, et al. An evolutionary machine learning algorithm for cardiovascular disease risk prediction. PLoS ONE. 2022;17(7):e0271723. https://doi.org/10.1371/journal.pone.0271723.
11. Dalmaijer E S, Nord C L, Astle D E. BMC Bioinformatics. 2022;23(1):205. https://doi.org/10.1186/s12859-022-04675-1. Statistical power for cluster analysis [J].
12. Cikes M, Sanchez-Martinez S, Claggett B, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail. 2019;21(1):74–85. https://doi.org/10.1002/ejhf.1333.
13. Gautam N, Mueller J, Alqaisi O, Gandhi T, Malkawi A, Tarun T, Alturkmani HJ, Zulqarnain MA, Pontone G, Al'Aref SJ. Machine Learning in Cardiovascular Risk Prediction and Precision Preventive approaches. Curr Atheroscler Rep. 2023;25(12):1069–81. https://doi.org/10.1007/s11883-023-01174-3.

14. Song H, Koh Y, Rhee T M, et al. Prediction of incident atherosclerotic cardiovascular disease with polygenic risk of metabolic disease: analysis of 3 prospective cohort studies in Korea. Atherosclerosis. 2022;348:16–24. https://doi.org/10.1016/j.atherosclerosis.2022.03.021.
15. Klooster C C V, Bhatt D L, Steg P G, et al. Predicting 10-year risk of recurrent cardiovascular events and cardiovascular interventions in patients with established cardiovascular disease: results from UCC-SMART and REACH. Int J Cardiol. 2021;325:140–8. https://doi.org/10.1016/j.ijcard.2020.09.053.
16. Lu P, Guo S, Zhang H, et al. Research on improved depth Belief Network-based prediction of Cardiovascular diseases. J Healthc Eng. 2018. https://doi.org/10.1155/2018/8954878. 2018(8954878.
17. Li Y, Sperrin M, Ashcroft DM, Van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. BMJ. 2020;371:m3919. https://doi.org/10.1136/bmj.m3919.
18. Tang AS, Woldemariam SR, Miramontes S, et al. Harnessing EHR data for health research. Nat Med. 2024;30:1847–55. https://doi.org/10.1038/s41591-024-03074-8.
19. Ward A, Sarraju A, Chung S, Li J, Harrington R, Heidenreich P, Palaniappan L, Scheinker D, Rodriguez F. Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. NPJ Digit Med. 2020;3:125. https://doi.org/10.1038/s41746-020-00331-1.
20. Qiu Y, Wang W, Wu C, et al. A risk factor attention-based model for cardiovascular disease prediction. BMC Bioinformatics. 2022;23(Suppl 8):425. https://doi.org/10.1186/s12859-022-04963-w.
21. Li Q, Campan A, Ren A, Eid WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. Int J Med Inf. 2022;163:104786.
22. Meng H, Ruan J, Yan Z, et al. New Progress in early diagnosis of atherosclerosis. Int J Mol Sci. 2022;23(16):8939. https://doi.org/10.3390/ijms23168939.
23. Francula-Zaninovic S, Nola I A. Management of Measurable Variable Cardiovascular Disease' risk factors. Curr Cardiol Rev. 2018;14(3):153–63. https://doi.org/10.2174/1573403X14666180222102312.
24. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26:303–4. https://doi.org/10.1038/nbt0308-303.
25. Ding C, He X. K-means Clustering via Principal Component Analysis. Proceedings of the 21 st International Conference on Machine Learning, Banff, Canada, 2004.
26. Frades I, Matthiesen R. Overview on techniques in cluster analysis. Methods Mol Biol. 2010;593:81–107. https://doi.org/10.1007/978-1-60327-194-3_5.
27. Inohara T, Shrader P, Pieper K, et al. Association of of Atrial Fibrillation Clinical Phenotypes with treatment patterns and outcomes: a Multicenter Registry Study. JAMA Cardiol. 2018;3(1):54–63. https://doi.org/10.1001/jamacardio.2017.4665.
28. Harada D, Asanoi H, Noto T, et al. Different pathophysiology and outcomes of heart failure with preserved ejection Fraction Stratified by K-Means clustering. Front Cardiovasc Med. 2020;7(607760). https://doi.org/10.3389/fcvm.2020.607760.
29. Petruzzo M, Reia A, Maniscalco G T, et al. The Framingham cardiovascular risk score and 5-year progression of multiple sclerosis. Eur J Neurol. 2021;28(3):893–900. https://doi.org/10.1111/ene.14608.
30. Brunström M, Andersson J, Eliasson M, et al. [SCORE2 - an updated model for cardiovascular risk prediction]. Lakartidningen. 2021;118:21164.
31. Alaa A M, Bolton T, Di Angelantonio E, et al. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS ONE. 2019;14(5):e0213653. https://doi.org/10.1371/journal.pone.0213653.
32. Virani Ss, Newby L K, Arnold S V, et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA Guideline for the management of patients with chronic coronary disease: a report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice guidelines. Circulation. 2023. https://doi.org/10.1161/CIR.0000000000001168.
33. Knuuti J, Wijns W. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J. 2020;41(3):407–77. https://doi.org/10.1093/eurheartj/ehz425.
34. LIndahl B. Acute coronary syndrome - the present and future role of biomarkers. Clin Chem Lab Med. 2013;51(9):1699–706. https://doi.org/10.1515/cclm-2013-0074.
35. Canaud B, Rodriguez A. Whole-blood viscosity increases significantly in small arteries and capillaries in hemodiafiltration. Does acute hemorheological change trigger cardiovascular risk events in hemodialysis patient?. Hemodial Int. 2010;14(4):433–40. https://doi.org/10.1111/j.1542-4758.2010.00496.x.
36. Zhou B, Perel P, Mensah G A, et al. Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension. Nat Rev Cardiol. 2021;18(11):785–802. https://doi.org/10.1038/s41569-021-00559-8.
37. Liu S, Zhao D, Wang M, et al. Association of Serum Potassium Levels with Mortality and Cardiovascular events: findings from the Chinese multi-provincial cohort study. J Gen Intern Med. 2022;37(10):2446–53. https://doi.org/10.1007/s11606-021-07111-x.

Publisher's note