Saputra *et al.*
Visual Computing for Industry, Biomedicine, and Art (2023) 6:16
https://doi.org/10.1186/s42492-023-00143-6

Visual Computing for Industry, Biomedicine, and Art

ORIGINAL ARTICLE

Open Access

Hyperparameter optimization for cardiovascular disease data-driven prognostic system

Jayson Saputra[1*], Cindy Lawrencya[1], Jecky Mitra Saini[1] and Suharjito Suharjito[1]

Abstract

Prediction and diagnosis of cardiovascular diseases (CVDs) based, among other things, on medical examinations and patient symptoms are the biggest challenges in medicine. About 17.9 million people die from CVDs annually, accounting for 31% of all deaths worldwide. With a timely prognosis and thorough consideration of the patient's medical history and lifestyle, it is possible to predict CVDs and take preventive measures to eliminate or control this life-threatening disease. In this study, we used various patient datasets from a major hospital in the United States as prognostic factors for CVD. The data was obtained by monitoring a total of 918 patients whose criteria for adults were 28-77 years old. In this study, we present a data mining modeling approach to analyze the performance, classification accuracy and number of clusters on Cardiovascular Disease Prognostic datasets in unsupervised machine learning (ML) using the Orange data mining software. Various techniques are then used to classify the model parameters, such as k-nearest neighbors, support vector machine, random forest, artificial neural network (ANN), naïve bayes, logistic regression, stochastic gradient descent (SGD), and AdaBoost. To determine the number of clusters, various unsupervised ML clustering methods were used, such as k-means, hierarchical, and density-based spatial clustering of applications with noise clustering. The results showed that the best model performance analysis and classification accuracy were SGD and ANN, both of which had a high score of 0.900 on Cardiovascular Disease Prognostic datasets. Based on the results of most clustering methods, such as k-means and hierarchical clustering, Cardiovascular Disease Prognostic datasets can be divided into two clusters. The prognostic accuracy of CVD depends on the accuracy of the proposed model in determining the diagnostic model. The more accurate the model, the better it can predict which patients are at risk for CVD.

Keywords Cardiovascular disease, Data-driven analytics, Data mining, Hyperparameter optimization, Orange data mining software, Prognostic system, Unsupervised machine learning

Introduction

Diseases related to the circulatory system impact the blood vessels and coronary arteries, and are prevalent globally. In developed nations, they are the primary cause of mortality in grown-ups. It is crucial to diagnose heart ailments with precision and timeliness by taking into account a patient's medical history and lifestyle. This approach enables accurate prognosis and the implementation of preventive measures to manage or eradicate these potentially fatal illnesses [1].

According to the 2013 Global Burden of Disease report by The Lancet, chronic illnesses pose the highest risk among all human ailments. Contributing factors comprise immoderate alcohol intake, hypertension, gender, and age. While these illnesses are widespread in affluent

*Correspondence:
Jayson Saputra
jayson@binus.ac.id
[1] Industrial Engineering Department, BINUS Graduate Program - Master of Industrial Engineering, Bina Nusantara University, Jakarta 11480, Indonesia

nations like the United States, where they account for 87% of fatalities, developing countries with lower and middle incomes require particular consideration due to the escalating incidence of chronic illnesses [2].

During the year 2020, the regions with the maximum age-adjusted rates of mortality caused by cardiovascular disease (CVD) were Eastern Europe, Central Asia, Oceania, North Africa, the Middle East, Central Europe, Sub-Saharan Africa, and South and Southeast Asia. Conversely, the regions with the minimum age-adjusted CVD mortality rates were high-income Asia-Pacific and North America, Latin America, Western Europe, and Australasia. Figure 1 indicates the age-standardized mortality rates per 100000 individuals affected by CVD across all countries [3].

The CVD fatality count in the United States declined from 1980 to 2010, but in recent times, it has escalated from 78454 in 2010 to 874613 in 2019. Figure 2 illustrates the patterns [3].

In 2019, coronary artery disease (41.3%) emerged as the primary reason for fatalities caused by CVD in the United States. Following this, stroke (17.2%), hypertension (11.7%), heart failure (HF, 9.9%), coronary heart disease (2.8%), and various other minor causes (17.3%) were observed [3].

CVDs make up 31% of the total fatalities globally, where 75% of the deaths occur in low- and middle-income nations. In wealthier nations, there is a higher occurrence and fatality rate of CVDs among individuals belonging to lower socioeconomic backgrounds [4]. Smoking, alcohol consumption, low fruit and vegetable intake, high salt intake, sedentary lifestyle, obesity, air pollution, genetic and metabolic factors, and other medical conditions are risk factors for CVD [5].

Forty percent of deaths in China are caused by CVD, a result of the aging population and a rise in stable metabolic risk factors. It's crucial to lower the prevalence of CVD through primary prevention, allocate more medical resources for emergency and critical care, and offer rehabilitation and secondary prevention services to decrease the chances of relapse, re-hospitalization, and disability in CVD survivors. In China, millions of individuals are affected by hypertension, dyslipidemia, diabetes, and vascular diseases, including myocardial infarction and stroke, are frequently diagnosed [6].

CVD continues to be the primary reason for illness and death across the globe, even with regional management measures [7]. The Morbidity and Mortality Conference has evolved into a valuable resource for surgeons to scrutinize complications and introduce changes to avert recurrence. Such insights can effectively curtail 'preventable' adverse outcomes among both novice and seasoned surgeons [8].

The role of gender in health and disease is gaining importance, yet there is a dearth of systematic gender research in the field of medicine. Women are at a greater

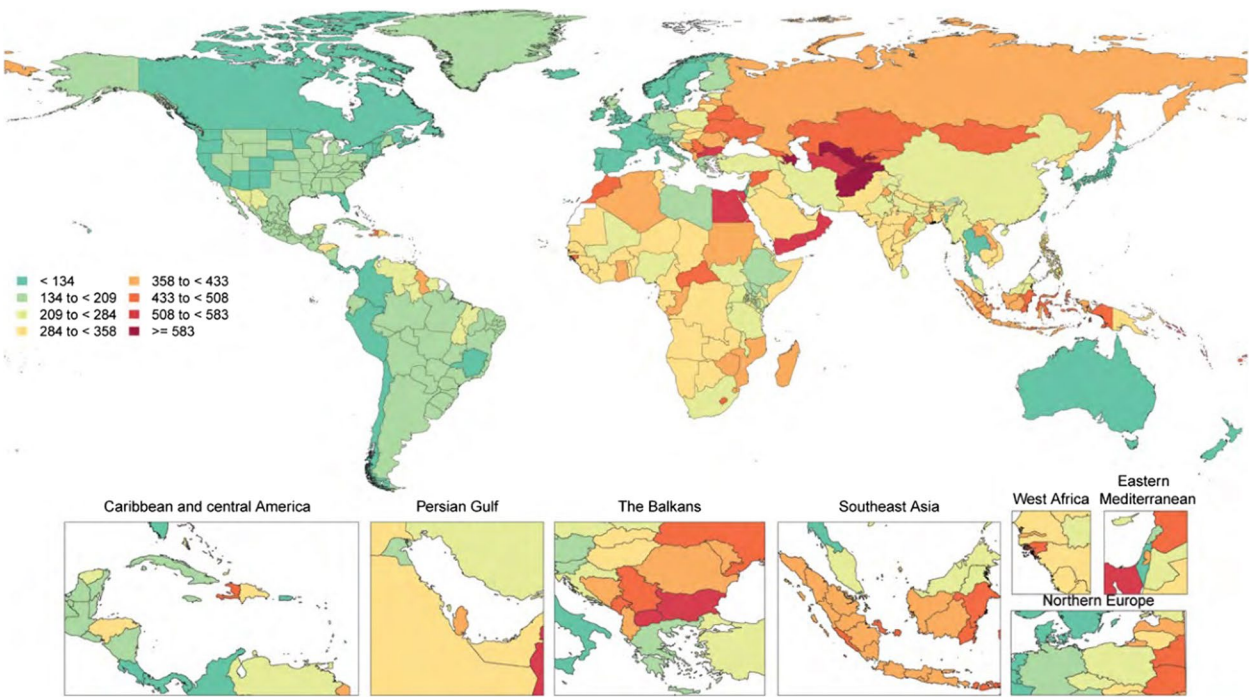


Fig. 1 CVD mortality rates are expected to increase significantly by 2020 [3]

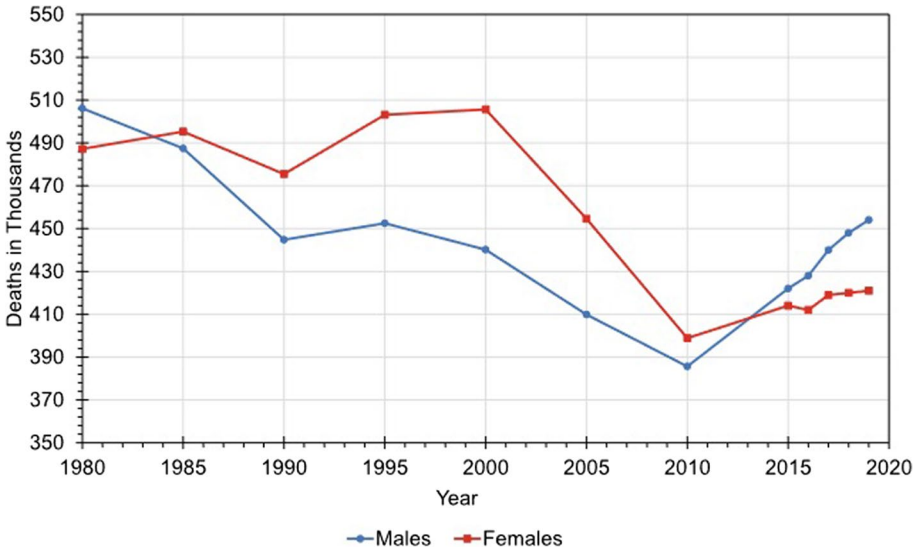


Fig. 2 Trends in CVD mortality for men and women in United States from 1980 to 2019 [3]

relative risk of suffering from CVD-related morbidity and mortality compared to men, primarily due to conventional factors such as obesity, hypercholesterolemia, hypertension, and diabetes, along with socio-economic and psychosocial factors, including depression. Additionally, depression amplifies the likelihood of CVD in women [9].

The coronavirus disease of 2019 (COVID-19) pandemic has altered the customary treatment for non-hospitalized individuals and those with sudden cardiac conditions, with the suspension of non-essential surgeries and a decrease in the effectiveness of current emergency medical services. In response to this crisis, novel methods like telehealth, online platforms, mobile apps, and artificial intelligence (AI) are being employed [10].

The pathophysiology of inflammation, blood clotting, and heart muscle damage linked with Severe Acute Respiratory Syndrome Coronavirus 2 can be evaluated by utilizing circulating biomarkers. Increased levels of cTn and NPs detected individuals with a higher probability of experiencing cardiovascular events while hospitalized, whereas increased levels of D-dimer detected individuals at risk of developing blood clotting issues [11].

People who have contracted COVID-19 are more prone to developing CVDs, such as disorders affecting the blood vessels in the brain, irregular heartbeats, heart diseases caused by reduced blood flow to the heart muscle, inflammation of the sac surrounding the heart, inflammation of the heart muscle, HF, and blood clots obstructing blood vessels. These hazards and difficulties are noticeable even in those who are not admitted to the hospital during the initial stage of the infection and intensify as per the level of care required during this phase [12]. Recognize and manage persons who have unaddressed or undetected risk factors for CVD in order to avert subsequent cardiovascular incidents resulting from the COVID-19 outbreak [13]. The outbreak of COVID-19 has resulted in a decrease in hospital admissions for all acute cardiovascular illnesses. However, there has been no alteration in hospital mortality rates except for acute aortic dissection, which has seen a rise [14].

In a vast population of unscreened COVID-19 patients across 30 medical facilities in Italy, impaired kidney function, heightened levels of C-reactive protein, and progressed age were notable indicators of mortality during hospitalization. These observations imply that pre-existing conditions, underlying illnesses, and clinical metrics may influence the likelihood of unfavourable outcomes and inpatient fatality in the European region [15].

In an explanatory analysis of 1099 instances of COVID-19, 24.9% of the individuals had concurrent ailments such as hypertension (15%), diabetes (7.4%), and coronary artery disease (2.5%). Aged patients (65 years and above) with concurrent ailments and acute respiratory distress syndrome are at an augmented peril of mortality. Numerous analyses have demonstrated a heightened vulnerability to Middle East respiratory syndrome (MERS)-CoV and human papillomavirus infections in patients with CVD, plausibly due to endothelial dysfunction, metabolic abnormalities, and the escalation of pro-inflammatory cytokines. CVD is a hazard factor for an unfavourable prognosis and significantly amplifies mortality from MERS. Various clinical analyses have

demonstrated that CVD is the most prevalent concurrent ailment in patients with COVID-19, and the CVD frequency is elevated in severe and fatal instances [16].

Cardiac insufficiency (CI) is a significant worldwide public health challenge, impacting 64 million individuals globally. The number of hospitalizations due to CI has increased by over three times in the last three decades and is linked to a high mortality rate. It places a substantial financial burden on public healthcare systems and has a noteworthy influence on the well-being of those affected [17].

The latest outbreak has hastened the acceptance of remote medical care in heart health and stimulated the progress of technological innovations like the metaverse. CardioVerse refers to the concept of incorporating the metaverse in cardiac medicine, which has multiple uses such as boosting medical consultations, aiding in heart-related procedures, and reforming medical learning. Despite the probable hindrances in different areas, the usage of unique tokens as safeguarding resources for patient information is emerging as a viable answer [18].

Timely identification of risk factors associated with infectious viral diseases can be crucial in preserving lives by efficiently allocating medical resources and prioritizing susceptible patients during national and global health crises [19]. Timely identification and preemptive measures against illnesses are crucial in averting their aggravation [20]. Managing contagious diseases is a primary concern for public health, and timely identification of infections is crucial to avert outbreaks and global health crises. Scientists are constructing frameworks for timely detection [21].

Knowledge regarding body position, alterations in posture, as well as active and stationary labour is crucial in comprehending the mechanical stresses and ergonomic principles [22]. Ecology and environment greatly rely on biodiversity information; however, various fields must collaborate to handle, exchange, and merge data in disease investigations [23].

CVDs exert a significant weight on the healthcare system, especially in the older population, owing to the presence of various coexisting conditions [24]. Chronic kidney disease (CKD) is the primary reason for mortality in individuals having CKD, where CVD is the primary cause of fatality [25].

Exercise is a crucial non-drug treatment for preventing and treating heart diseases. However, the impact of the length of physical activity on the risk factors related to heart health in grown-ups is not yet clear [26]. The intake of coffee has been demonstrated to have advantageous impacts on metabolic disorders, albeit it could upsurge lipid levels. Additionally, it diminishes the possibility of coronary artery disease, HF, heart arrhythmias, stroke, CVD, and death from all causes. The regular consumption of coffee and tea can be viewed as a component of a salubrious lifestyle and should not be disallowed for patients with CVD [27].

Factors that affect health, known as social determinants of health (SDoH), encompass economic, societal, ecological, and psychological elements. These determinants have a noteworthy effect on the health of individuals with CVD, as well as their outcomes, globally. To achieve health equity and tackle health disparities, SDoH involve determinants related to the framework, physicality, nutrition, and societal environment [28]. The economic condition of an individual is a factor that increases the risk for CVD, and a meagre family income can exacerbate the risk. The government should focus on reducing inequalities and enhancing cardiovascular results in underprivileged groups with low family incomes [29].

The demand for healthcare technology solutions has risen due to the increase in population and changes in lifestyle. Cancer prognosis can determine the likelihood of survival and indicate the seriousness of the illness as it pertains to the patient's future [30]. The medical industry is experiencing a surge in machine learning (ML) applications as they have the potential to accurately identify patterns in data. This capability can be leveraged to provide accurate diagnosis and prognosis of CVDs, leading to a reduction in misdiagnosis and improved patient care [31].

Forecasting and identification of cardiac ailments pose the greatest hurdles in the field of medicine and rely on facets such as medical evaluations and patient indications. ML methodologies play a pivotal and precise part in the prognosis of heart disease, and technological advancements have facilitated the amalgamation of machine communication with vast data utilities to handle unorganized and rapidly augmenting data [32]. The sole approach to acquiring significant insights in healthcare is through big data, and it is imperative to combine data from diverse origins to discover remedies [33]. The potential of big data to enhance healthcare services and financial returns is immense. Many industries, healthcare included, are making efforts to leverage this potential. By merging bio-medical and healthcare data, contemporary healthcare institutions can bring about a revolution in medical treatment and customized healthcare [34].

The extraction of valuable information from structured human-generated, computer-generated, and sensor data is known as data mining. This involves the collection, cleansing, processing, analysis, visualization, and interpretation of data, using sophisticated learning algorithms to identify patterns and relationships that can be applied in various fields. To achieve more straightforward and comprehensible outcomes, statistical, mathematical, and

ML approaches have been employed. Data mining is more than just a task; it encompasses the entire process of gathering, cleaning, processing, analyzing, visualizing, and interpreting data to extract valuable insights [35].

Data analysis is a crucial factor for achievement, as it enables the selection of appropriate data scrutiny methods and the formulation of data-driven products. Nonetheless, businesses frequently encounter a shortage of expertise or time to develop a comprehensive comprehension of data in data analysis. Familiarity with the attributes of the data is vital for devising data-driven products [36].

Unsupervised ML techniques have the potential to uncover risk determinants in patients with intricate clinical conditions like HF, which exhibits indications and manifestations of excess fluid. The worldwide occurrence and frequency of HF have surged, leading to a global outbreak. Novel approaches need to be explored to enhance the management of HF patients [37].

ML algorithms have the potential to enhance the diagnostic and prognostic capabilities of conventional regression methods, however, the outcomes are reliant on the data analysis software utilized [38]. ML techniques are employed for forecasting heart diseases, however, there exists variations in their parameters, aiding physicians in comprehending the information and executing the most suitable techniques [39]. ML is a crucial instrument in public health for recognizing and anticipating communities with higher chances of experiencing health consequences. Thus, it should be incorporated into medical education to direct and decipher scientific investigations [40].

The extraction of significant data from vast quantities of unprocessed information is known as data mining. This technique is utilized in various domains, such as scientific research. Orange employs segment-oriented visual programming to carry out data mining, AI, and inspection tasks. By linking pre-defined or user-provided components known as widgets, work forms are established. The process of constructing a data mining model involves activities such as reading, processing, visualizing, collecting data, and obtaining prediction models [41].

The medical industry is experiencing a transformation in decision-making procedures, thanks to the abundant digital information stored in hospitals, and the implementation of data mining and ML methods. While conventional ML techniques were previously utilized to forecast cancer survival rates, experts are currently transitioning towards deep learning and hybrid approaches to obtain a better understanding of survival prediction [42].

Modern health information systems are distinguished by their ability to rapidly grow and adapt to identify significant health trends and provide timely prevention support. ML-based systems can predict and diagnose heart diseases. Active learning techniques enhance classification accuracy by incorporating expert feedback from users with sparsely labelled data. The label-ranking classifier selection method employs hyperparameters optimized through network search and implements predictive modelling in the cardiac dataset scenario. Experimental evaluations were conducted to measure accuracy and F-score, with and without hyperparameter optimization. The optimized setting prioritized the selection method with respect to the F-score [43].

This study presents a technique for creating data mining models that investigates the performance, classification accuracy, and number of groupings in CVD predictive datasets using the Orange data mining software in unsupervised ML. Orange is a powerful instrument for analyzing and displaying data, identifying data trends, and improving performance. It delivers a user-friendly interface that can be adapted to various domains of research.

Literature review

The medical sector produces vast quantities of intricate information on patients, hospital assets, sickness diagnoses, digital health records, and medical apparatus. The potential of data mining applications is immense, with some of the most significant applications encompassing forecasting and identification of diseases, evaluating the efficacy of treatments, healthcare operations, prevention of fraud and abuse, customer relationship management, and the medical apparatus industry. An incorrect treatment selection can result in unfavorable consequences, such as patient mortality. Data mining can aid in forecasting and defining diseases within the field [44].

The healthcare sector has made significant progress, resulting in the accumulation of extensive healthcare data, such as electronic health records (EHRs), wearable sensors, and intelligent devices. This data holds undisclosed insights that can aid in making informed decisions. Extracting valuable information requires a thorough search of medical records, and open-source initiatives offer a wealth of data sources for diagnosing and predicting all illnesses [45]. Detecting abnormal sequences plays a crucial role in establishing and safeguarding contemporary health information technology (HIT) systems, ensuring a thorough account of the patient's condition and occurrences. Nevertheless, this can result in skewed data, intricate interconnections among events in sequences, and diminished complexity [46]. HIT usage poses a challenge in Sub-Saharan Africa, resulting in inadequate patient data. A dependable hospital patient database is crucial for delivering superior

healthcare and facilitating seamless communication between healthcare professionals [47].

Valuable information about individual patients and populations is stored in EHRs. The most frequent sources of unstructured EHR data are clinical text and images. Statistical algorithms like natural language processing, radiomics, deep learning, and ML are increasingly being utilized to analyze clinical texts and images. However, explaining and generalizing the outcomes of ML models in healthcare is a crucial and unresolved issue. To enhance the quality and access of unstructured data, developing ML methods that can produce clinically relevant synthetic data and de-identify clinical texts to speed up further research is a potential solution. This is achieved by creating privacy protection technologies such as pseudonymization [48].

Maiga et al. [49] conducted a comparison of ML algorithms for predicting CVDs using a dataset of 70000 medical records. The random forest (RF) model demonstrated impressive results with a classification accuracy of 73%, specificity of 65%, and sensitivity of 80%. These findings have significant implications for the medical industry as they could be utilized to predict the occurrence of CVD.

Peng et al. [50] created an XGBH model for predicting the risk of CVD using significant features extracted from 14832 CVD patients in Shanxi, China. Although it had a slightly lower precision and reduced efficacy in predicting CVD risk, it facilitated timely intervention and economical screening of high-risk patients.

Nouraei et al. [51] utilized three distinct unsupervised ML clustering methodologies on a combined data-set of patients affected by heart failure and preserved ejection fraction. The partitioning around medoids technique recognized six unique groups of patients with varying long-term results or mortality rates, whereas the other two clustering algorithms were subpar.

Detecting anomalies and irregularities in heart rate (HR) and other attributes can aid in comprehending the cause of the disease. The vast quantity of information produced by sensors in portable gadgets has irregularities that necessitate meticulous automation procedures for detection. Several techniques have been suggested to recognize these anomalies [52]. Ripan et al. [53] utilized five ML classification methods to construct prognostic models of results, which were authenticated exploiting customary cardiac datasets. They eliminated abnormalities and employed K-nearest neighbors (KNN), RF, support vector machine (SVM), naïve Bayes, and logistic regression (LR).

The utilization of ML algorithms can aid in the early detection and diagnosis of heart disease, leading to enhanced patient results. Additionally, they can assist patients in managing their condition and daily habits more effectively, ultimately increasing their likelihood of recuperation and survival. This is a positive indication that ML algorithms have the potential to identify illnesses sooner and enhance patient outcomes [54]. Magesh and Swarnalatha [55] proposed a method that utilizes Cleveland heart samples from the University of

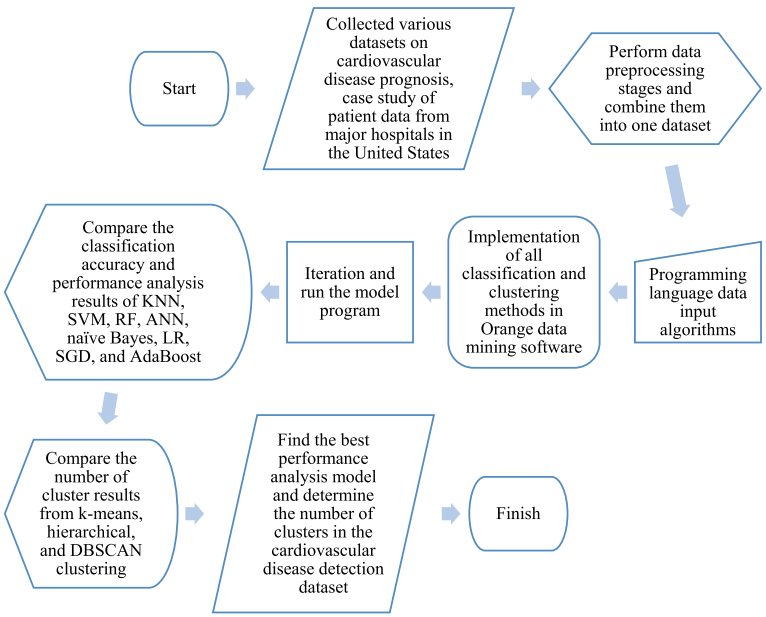


Fig. 3 Research flowchart

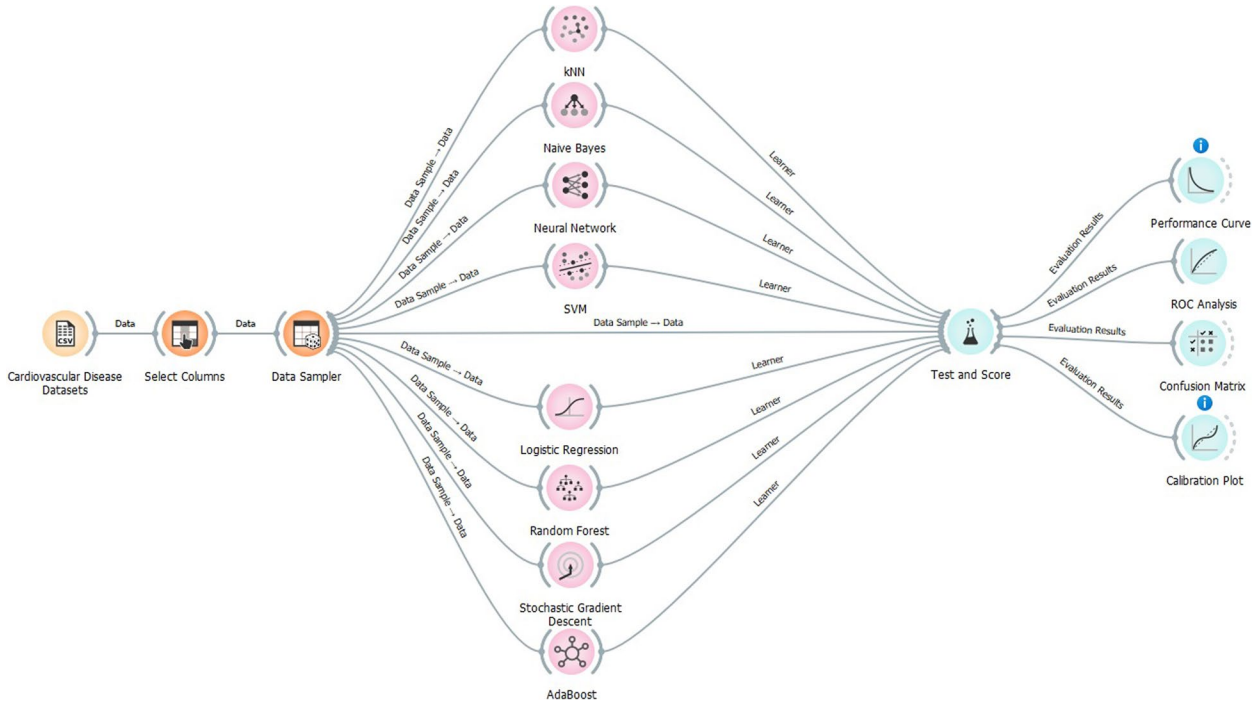


Fig. 4 Classification techniques used to classify Cardiovascular Disease Prognostic datasets

Test and Score - Orange

Cross validation
Number of folds: 20
Stratified
Cross validation by feature
Random sampling
Repeat train/test: 10
Training set size: 80 %
Stratified
Leave one out
Test on train data
Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.748	0.775	0.775	0.776	0.775
SVM	0.917	0.825	0.824	0.832	0.825
SGD	0.900	0.900	0.900	0.900	0.900
Random Forest	0.951	0.875	0.875	0.876	0.875
Neural Network	0.970	0.900	0.900	0.900	0.900
Naive Bayes	0.932	0.875	0.875	0.876	0.875
Logistic Regression	0.835	0.775	0.775	0.776	0.775
AdaBoost	0.800	0.800	0.800	0.800	0.800

Fig. 5 Test and score results

California, Irvine repository to predict CVD. The precision of the HF classifier was improved by 89.30% through the application of the cluster-based decision tree (DT) learning approach, leading to a significant reduction in the HF error rate from 23.30% to 9.70%.

Shrifan et al. [56] enhanced the accuracy and centroid convergence of k-means clustering through modifications. The newly suggested distance metric surpassed the majority of the literature, leading to an enhancement in the overall clustering accuracy for nine standard multivariate datasets to 80.57%.

The World Health Organization endeavoured to create, assess, and explicate updated models for determining the risk of CVD in low- and middle-income nations. Kaptoge et al. [57] observed significant discrepancies in the projected 10-year risk for a specific risk factor profile among different regions worldwide. After examining data from 79 countries, it was deduced that the percentage of people aged 40-64 years with estimated risk greater than 20% varied greatly, ranging from under 1% in Uganda to over 16% in Egypt.

Compare models by: Classification accuracy — Negligible diff.: 0.1

	kNN	SVM	SGD	Random Forest	Neural Network	Naive Bayes	Logistic Regression	AdaBoost
kNN		0.359	0.185	0.232	0.185	0.162	0.500	0.439
SVM	0.641		0.108	0.247	0.175	0.289	0.652	0.600
SGD	0.815	0.892		0.652	0.500	0.600	0.855	0.838
Random Forest	0.768	0.753	0.348		0.348	0.500	0.838	0.892
Neural Network	0.815	0.825	0.500	0.652		0.600	0.855	0.838
Naive Bayes	0.838	0.711	0.400	0.500	0.400		0.753	0.731
Logistic Regression	0.500	0.348	0.145	0.162	0.145	0.247		0.400
AdaBoost	0.561	0.400	0.162	0.108	0.162	0.269	0.600	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Fig. 6 Classification accuracy matrix

Nadakinamani et al. [58] proposed an ML-based CVD forecasting system that is extremely precise. The system's suitability was determined by assessing several metrics, and the random tree model produced excellent results, achieving a 100% accuracy rate, a 0.0011 mean absolute error, a 0.0231 root mean squared error, and a prediction time of only 0.01 s, the fastest of all models tested.

The platform is restricted to supervised ML algorithms, which confine the training dataset to labelled datasets. However, it supports unsupervised learning algorithms, enabling the platform to handle diverse types of training datasets. The selection parameters are set before the training process, which restricts the system options but permits users to modify them post-training to better align with their requirements and enhance efficiency [59]. Aggrawal and Pal [60] suggested a method for identifying mortality in cardiac patients receiving therapy using a sequential feature selection algorithm. To evaluate the accuracy of the selected feature selection (SFS) algorithm against the RF classifier, various ML algorithms such as linear discriminant analysis, RF, gradient boosting classifier (GBC), DT, SVM, and KNN were employed. According to the findings of the experiment, the SFS approach achieved an accuracy of 86.67%.

Ishaq et al. [61] employed nine categorization techniques: DT, AdaBoost, LR, stochastic gradient descent (SGD), RF, GBC, extra tree classifier (ETC), Gaussian naive bayes, and SVM. The issue of imbalanced classes was tackled by utilizing synthetic minority oversampling technique (SMOTE), and the RF was used to identify the most highly-ranked features to train the ML models. The experimental findings demonstrated that ETC outperformed the other models, achieving a SMOTE accuracy score of 0.9262 in predicting the survival of patients suffering from heart disease.

Healthcare experts frequently face difficulties in precisely forecasting heart ailments because of intricate tasks and concealed information, which necessitate contemplation and understanding [62]. Li et al. [63] suggested a ML technology-based system that is both efficient and precise in diagnosing heart ailments. The experimental findings indicate that the feature selection algorithm (FCMIM) proposed by Li et al., with high-level classifier support vectors, is suitable for creating intelligent cardiac detection systems. The diagnostic system (FCMIM-SVM) suggested by Li has demonstrated impressive accuracy in comparison to previously suggested methods and can be conveniently adopted in healthcare for the identification of heart diseases.

The outcomes of Oyeleye et al. [64] experiment demonstrated that by employing forward walking validation and linear regression, the autoregressive intregated moving average model can precisely anticipate the HR for all time spans, while other models are effective for time spans exceeding 1 min. This method of data analysis can be utilized to more accurately predict future HR with the help of accelerometers.

Mohammedqasem et al. [65] created an optimization system based on deep learning to enhance patient classification by processing unbalanced datasets. The system employs SMOTE and a feature-removal algorithm that operates recursively to identify the most efficient features. The experimental predictions demonstrated high consistency and appropriateness, reaching an accuracy level of up to 98% and 97% respectively.

The field of ML holds great potential in enhancing results by identifying prognostic models and categorizing innovative patient subpopulations. AI is permeating our routine activities via promotional algorithms, music and film preferences, and junk mail filtering, but its capacity to access intricate and multidimensional data is just as crucial in the medical domain. However, this has yet to be fully substantiated [66]. AI has the capability to recognize the ideal research specimens,

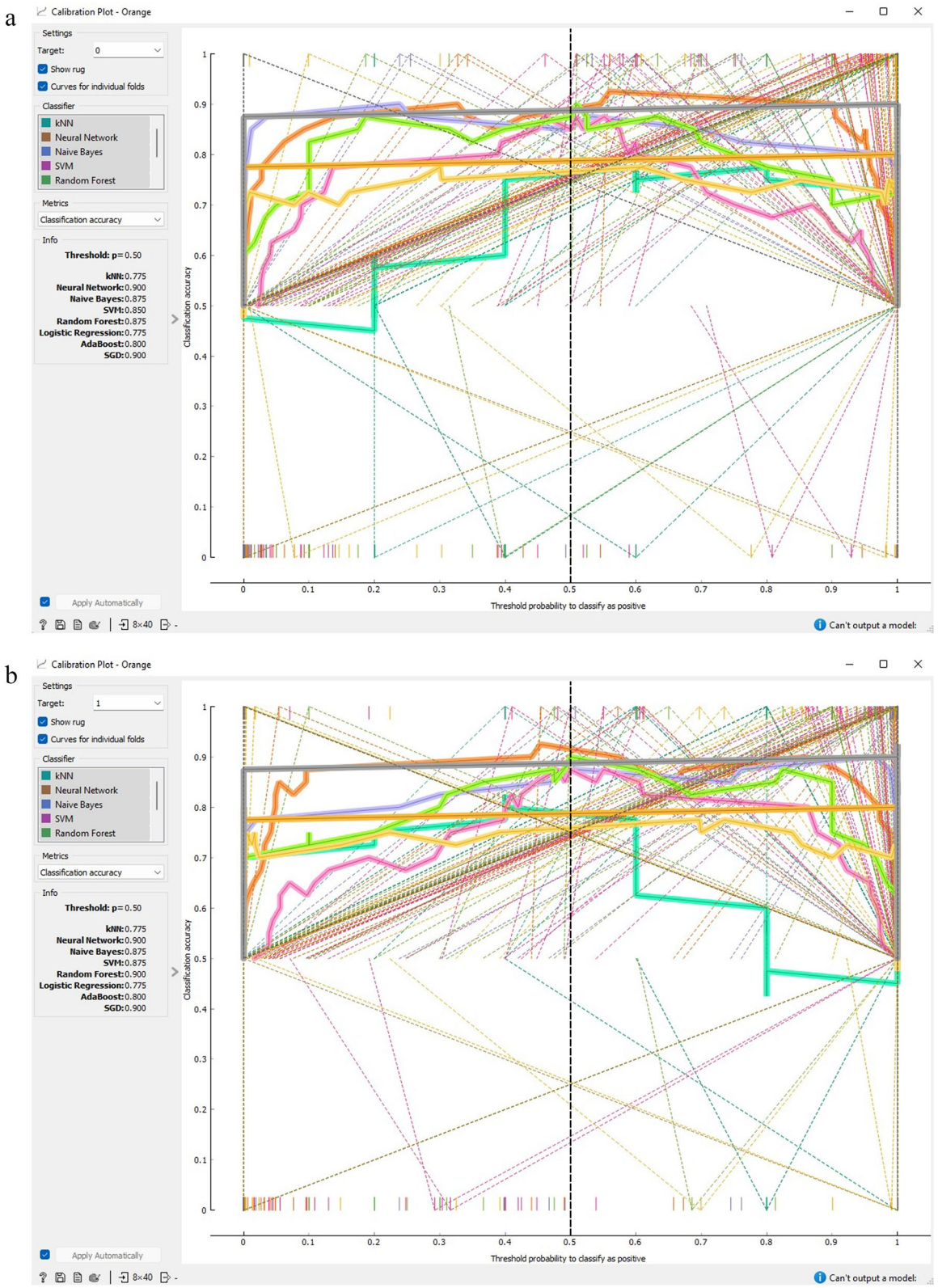
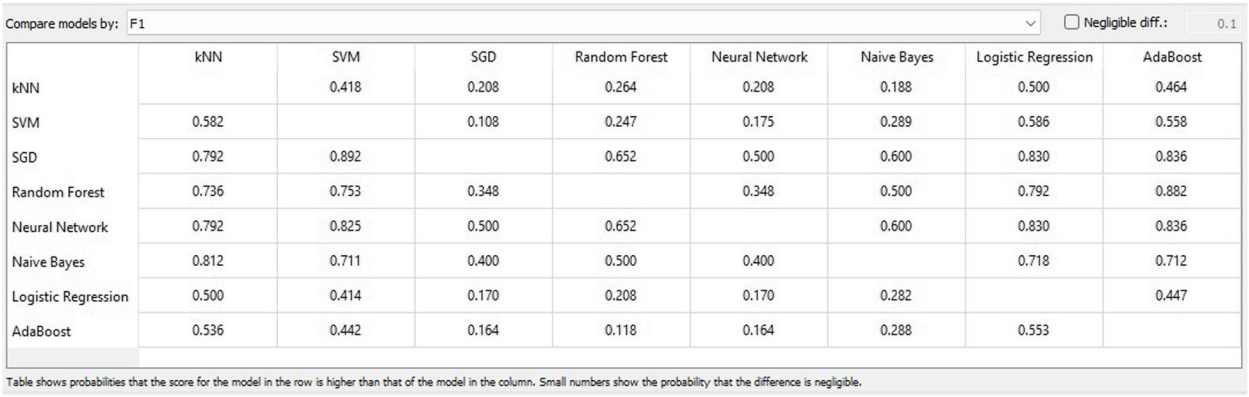


Fig. 7 Calibration plot based on classification accuracy CVD prognostic. **a** target = 0; **b** target = 1

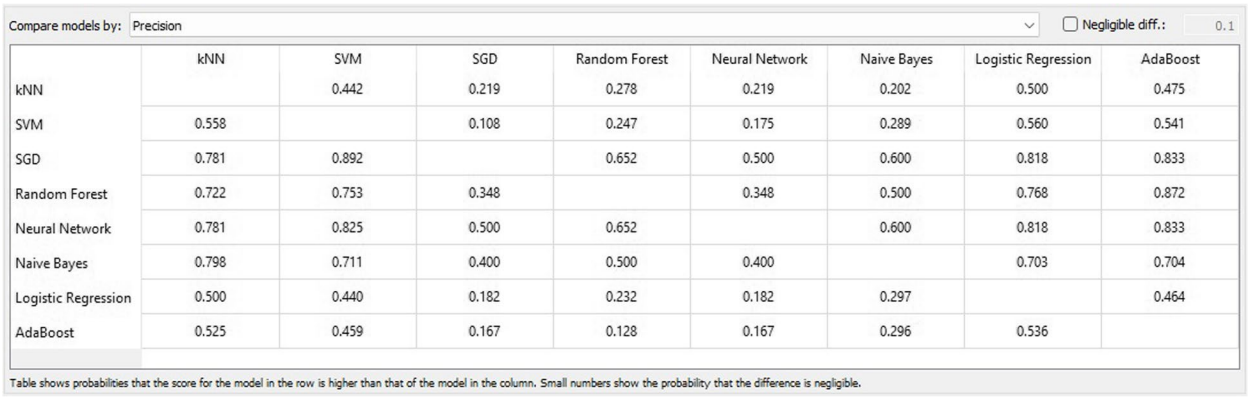


Compare models by: F1 Negligible diff.: 0.1

	kNN	SVM	SGD	Random Forest	Neural Network	Naive Bayes	Logistic Regression	AdaBoost
kNN		0.418	0.208	0.264	0.208	0.188	0.500	0.464
SVM	0.582		0.108	0.247	0.175	0.289	0.586	0.558
SGD	0.792	0.892		0.652	0.500	0.600	0.830	0.836
Random Forest	0.736	0.753	0.348		0.348	0.500	0.792	0.882
Neural Network	0.792	0.825	0.500	0.652		0.600	0.830	0.836
Naive Bayes	0.812	0.711	0.400	0.500	0.400		0.718	0.712
Logistic Regression	0.500	0.414	0.170	0.208	0.170	0.282		0.447
AdaBoost	0.536	0.442	0.164	0.118	0.164	0.288	0.553	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Fig. 8 F1 score matrix

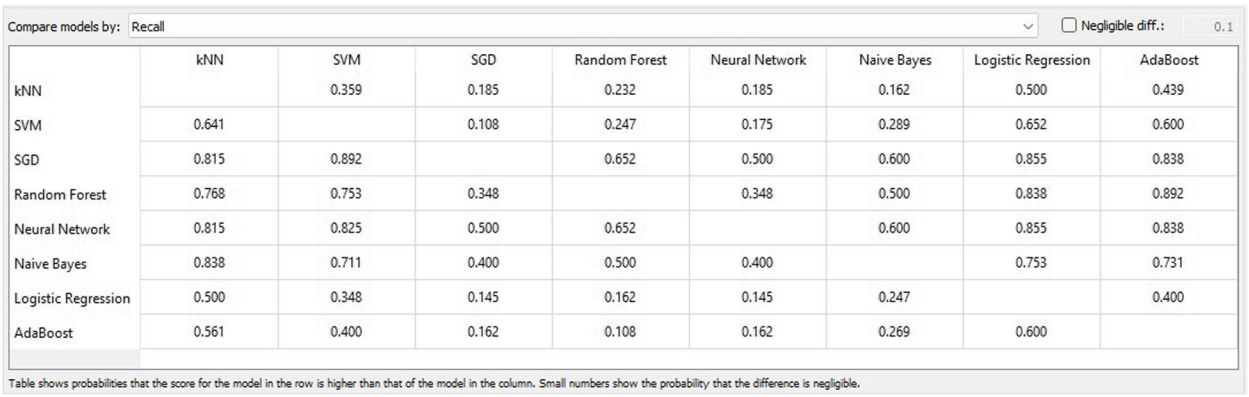


Compare models by: Precision Negligible diff.: 0.1

	kNN	SVM	SGD	Random Forest	Neural Network	Naive Bayes	Logistic Regression	AdaBoost
kNN		0.442	0.219	0.278	0.219	0.202	0.500	0.475
SVM	0.558		0.108	0.247	0.175	0.289	0.560	0.541
SGD	0.781	0.892		0.652	0.500	0.600	0.818	0.833
Random Forest	0.722	0.753	0.348		0.348	0.500	0.768	0.872
Neural Network	0.781	0.825	0.500	0.652		0.600	0.818	0.833
Naive Bayes	0.798	0.711	0.400	0.500	0.400		0.703	0.704
Logistic Regression	0.500	0.440	0.182	0.232	0.182	0.297		0.464
AdaBoost	0.525	0.459	0.167	0.128	0.167	0.296	0.536	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Fig. 9 Precision matrix



Compare models by: Recall Negligible diff.: 0.1

	kNN	SVM	SGD	Random Forest	Neural Network	Naive Bayes	Logistic Regression	AdaBoost
kNN		0.359	0.185	0.232	0.185	0.162	0.500	0.439
SVM	0.641		0.108	0.247	0.175	0.289	0.652	0.600
SGD	0.815	0.892		0.652	0.500	0.600	0.855	0.838
Random Forest	0.768	0.753	0.348		0.348	0.500	0.838	0.892
Neural Network	0.815	0.825	0.500	0.652		0.600	0.855	0.838
Naive Bayes	0.838	0.711	0.400	0.500	0.400		0.753	0.731
Logistic Regression	0.500	0.348	0.145	0.162	0.145	0.247		0.400
AdaBoost	0.561	0.400	0.162	0.108	0.162	0.269	0.600	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Fig. 10 Recall matrix

gather supplementary data points, assess continuous data from research participants, and eradicate data-related inaccuracies in overburdened healthcare systems [67]. The significance of AI in healthcare is growing, especially in the examination or anticipatory evaluation of medical information. Hypertensive patients were researched using Spark data analysis as a platform, and AI techniques were employed to pre-analyze the data for inconsistencies, duplication, inadequacy, disturbance, and inaccuracy [68].

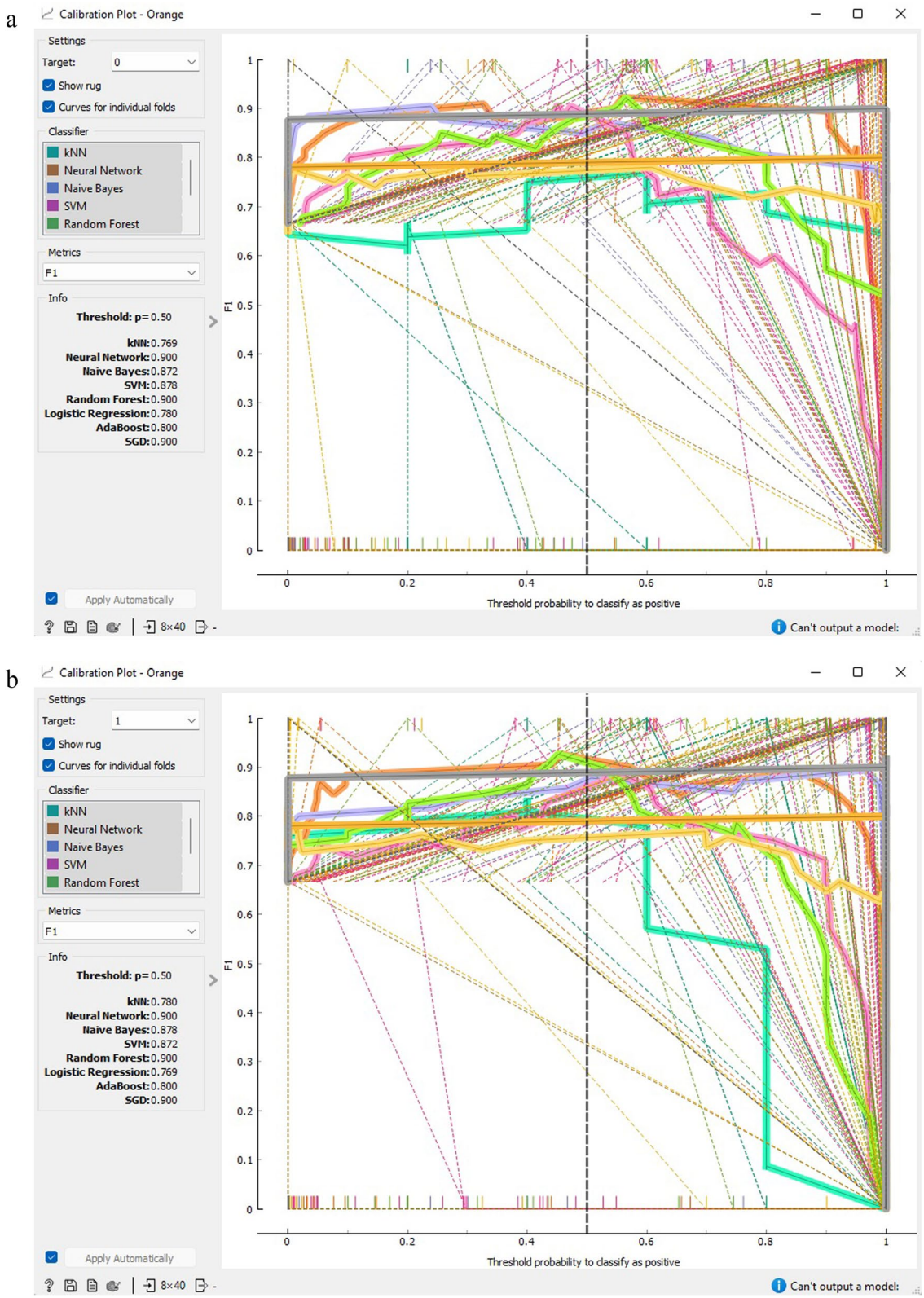


Fig. 11 Calibration plot based on F1 score CVD prognostic. **a** target = 0; **b** target = 1

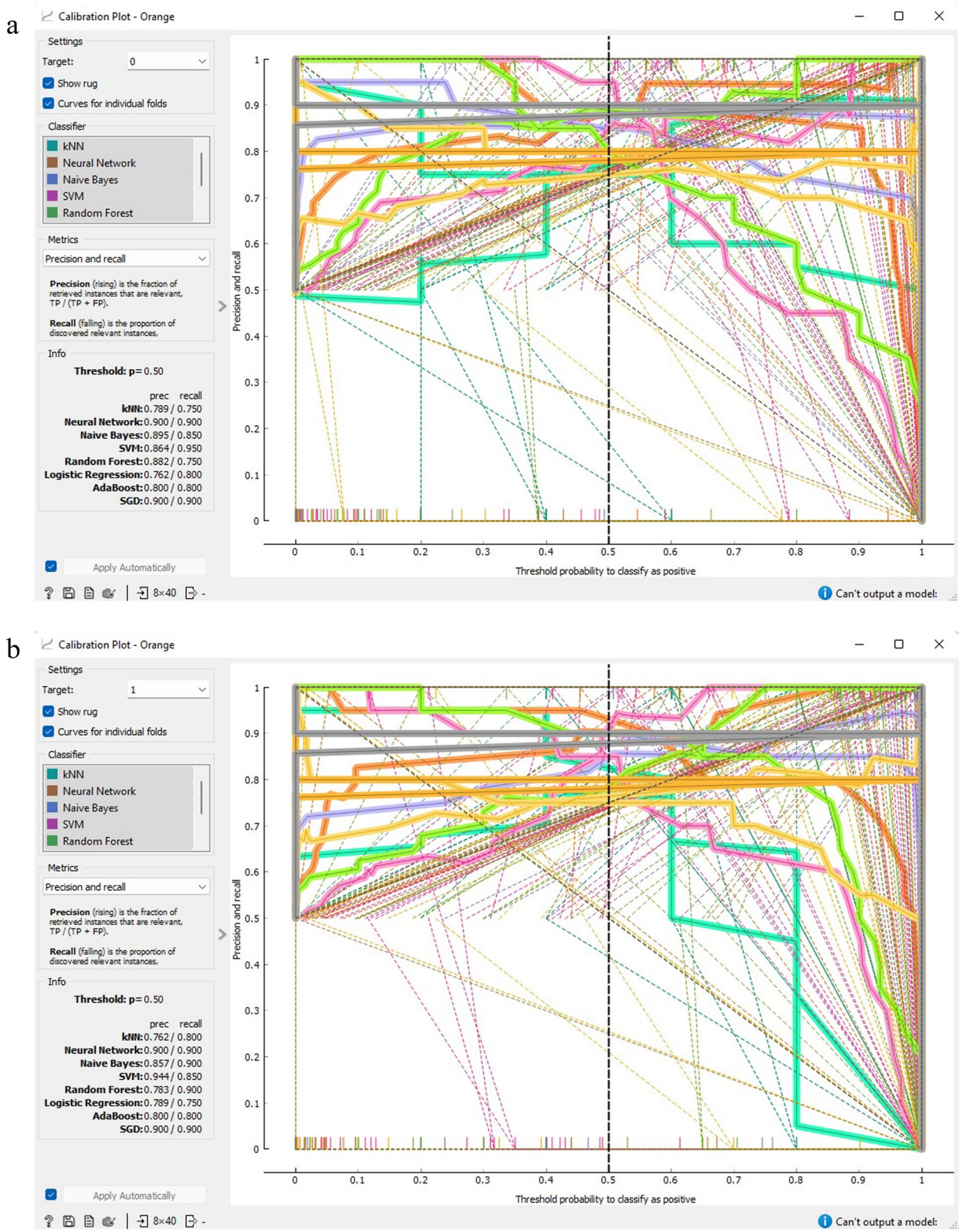
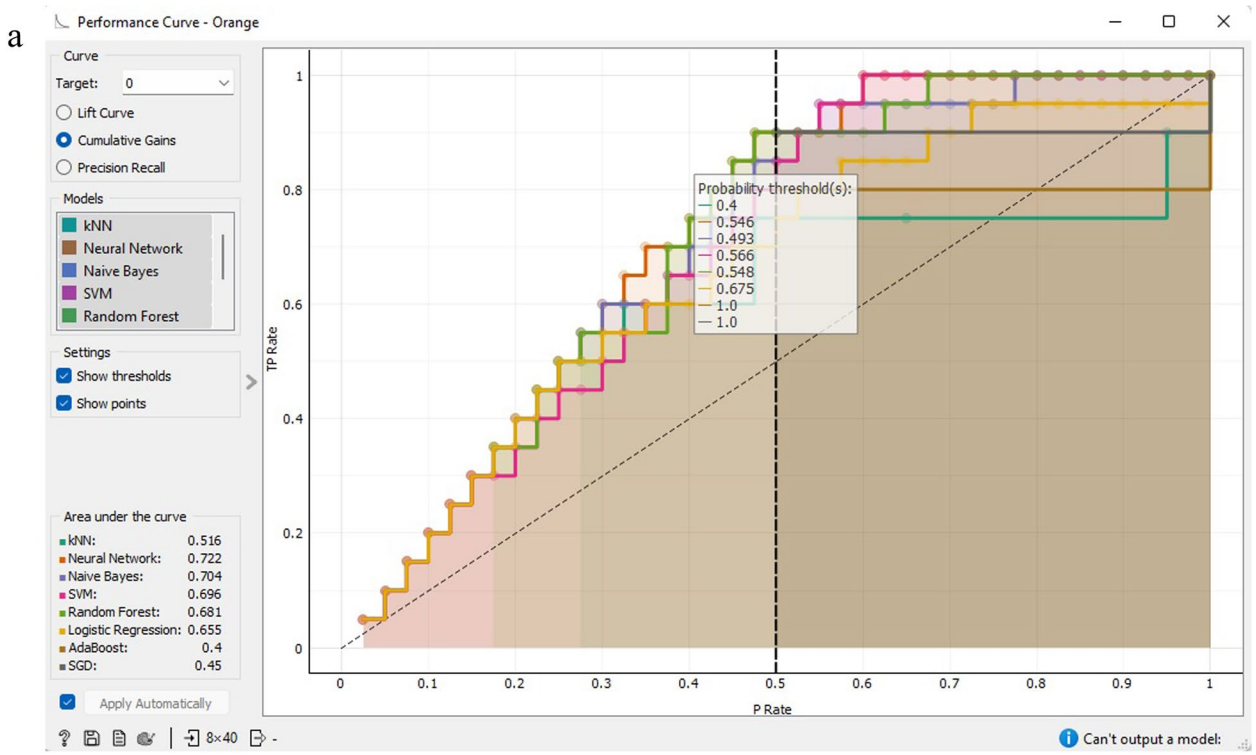


Fig. 12 Calibration plot based on precision and recall CVD prognostic. **a** target = 0; **b** target = 1



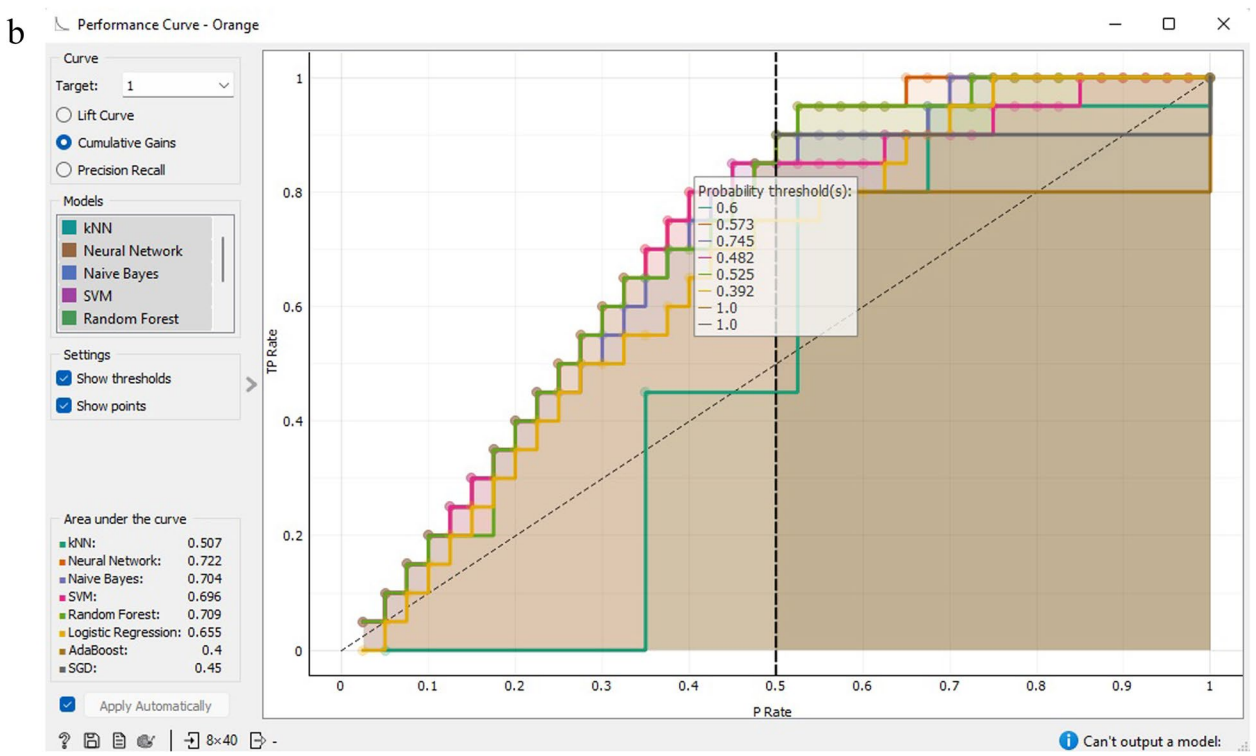


Fig. 13 Performance curve analysis CVD prognostic. **a** target = 0; **b** target = 1

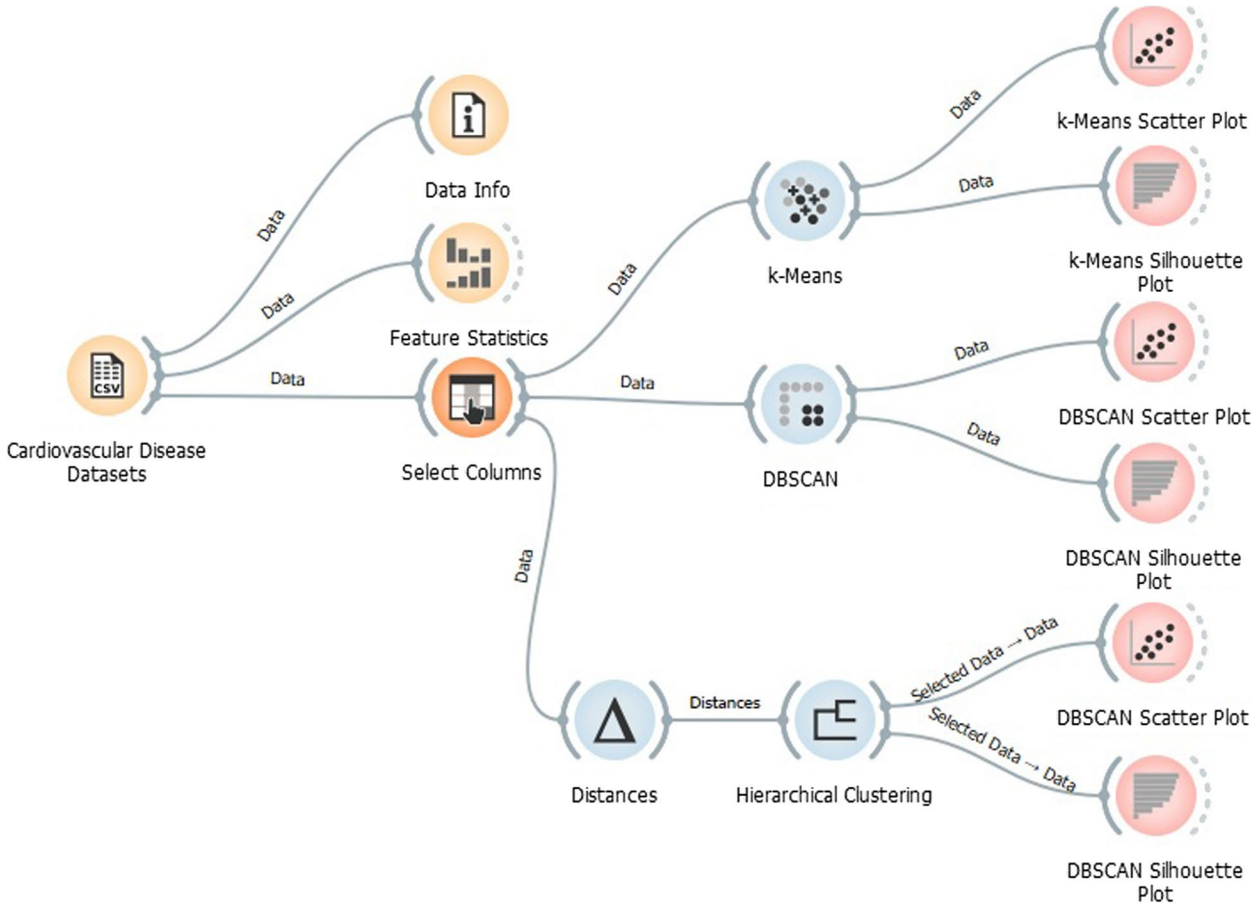


Fig. 14 Clustering techniques are used to analyze Cardiovascular Disease Prognostic datasets

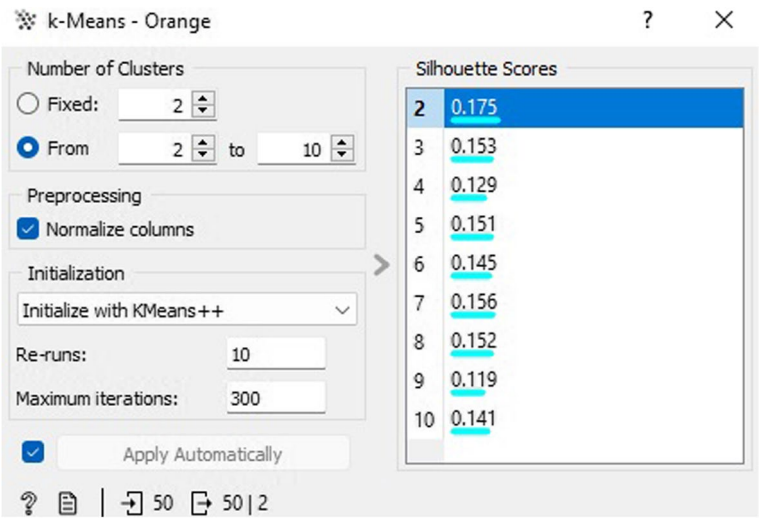


Fig. 15 K-means clustering

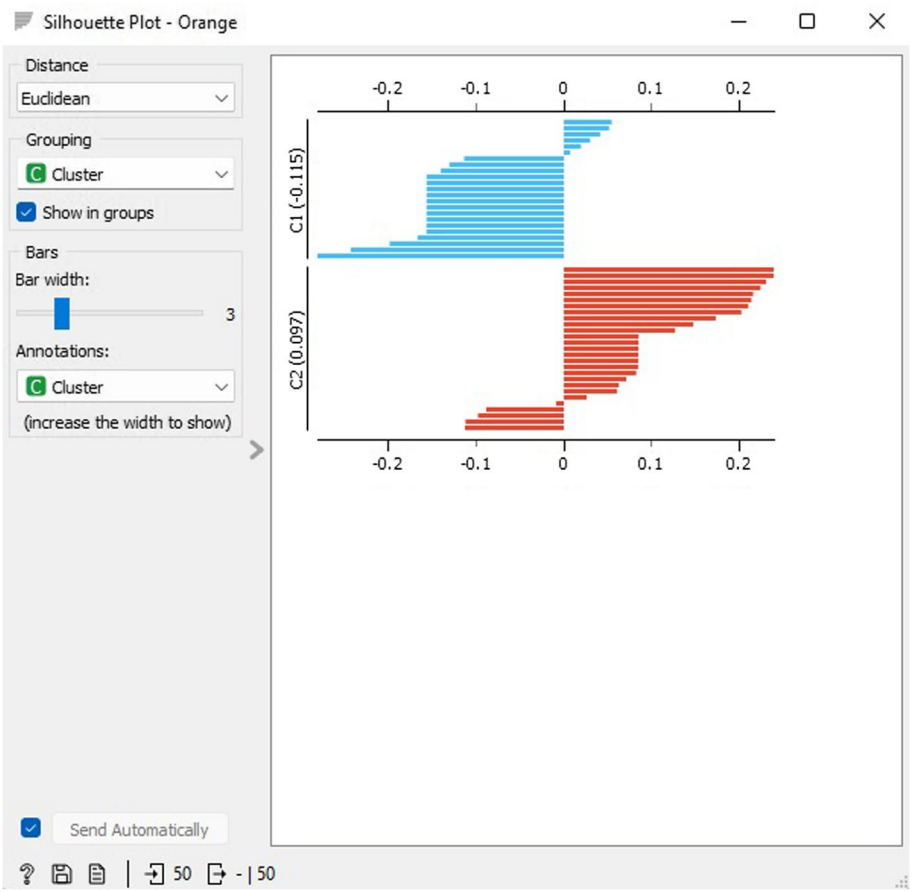


Fig. 16 K-means clustering silhouette scores

Velu et al. [69] suggested utilizing a technique based on ML to anticipate liver complications by analyzing the outcomes of liver function tests that were conducted during medical check-ups. The system encompasses an interface that medical professionals can utilize to obtain patient data. The patients' liver function test outcomes were assessed to identify whether they had liver disease by examining the blood levels of enzymes and proteins that are specific to liver function tests.

The use of AI in electrocardiography (ECG) is a prime example of how AI is transforming cardiovascular medicine. Advanced AI techniques, including convolutional neural networks using deep learning, have made it possible to interpret ECGs quickly and accurately, similar to how humans would. This has allowed for the detection of signals and patterns that would have otherwise gone unnoticed by human interpreters. By utilizing extensive digital ECGs that come with comprehensive clinical data, AI models have been developed to identify left ventricular dysfunction, silent atrial fibrillation (previously undetected and asymptomatic), hypertrophic cardiomyopathy, and other phenotypes such as age, sex, and race. As mobile and wearable ECG technologies become increasingly available, the clinical and population-level implications of AI-based ECG phenotyping are still unfolding [70].

The gaps in this study compared to existing or previous research are as follows:

a. The use of Orange data mining software for unsupervised ML clustering of CVD datasets has not been explored before.
b. The topic of achieving optimal model analysis and classification accuracy across all classification types has seldom been explored.
c. Investigations that strive to identify the quantity of clusters utilizing diverse clustering techniques within a single dataset are infrequent.

This research centered on utilizing data mining through the Orange data mining software to examine

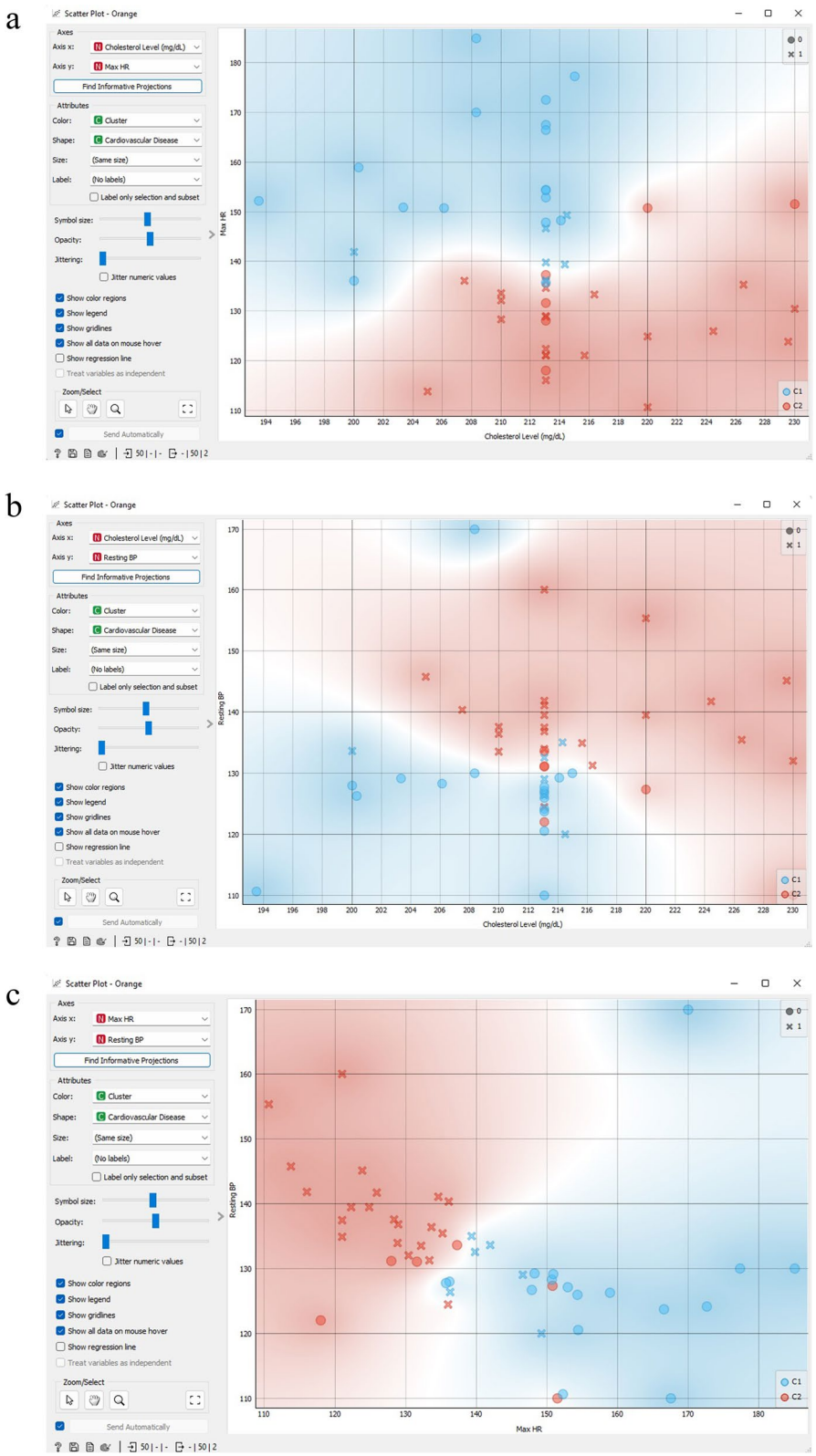


Fig. 17 Scatter plot of k-means cluster analysis. **a** K-means clustering scatter plot correlations between cholesterol level and maximum HR attributes; **b** K-means clustering scatter plot correlations between cholesterol level and resting blood pressure attributes; **c** K-means clustering scatter plot correlations between maximum HR and resting blood pressure attributes

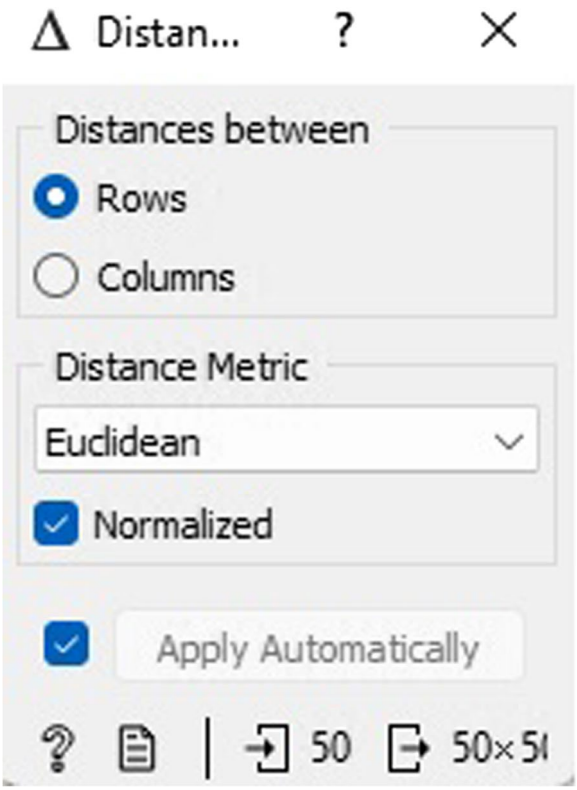


Fig. 18 Distances

the precision of classification, number of clusters, and overall performance of CVD prognostic datasets in unsupervised ML, drawing from insights gleaned from prior scientific literature reviews.

This paper is organized as follows: Methods section outlines the research methodology, Results and discussion section presents the results and discussion, Conclusions section summarizes the research conclusions and provides suggestions for future studies of unsupervised ML in the CVD dataset.

Methods

Figure 3 presents a flowchart of the research stages.

For this research, we utilized diverse datasets on CVD prognosis in a case study of patient information from prominent hospitals in the United States. The participants were individuals aged between 28-77 years. We procured varied datasets from Kaggle, which assembles public information from websites, like frequent visitors, without compromising personal data. The data comprised observational findings of 918 patients from one of the most notable hospitals in the United States. The datasets had 18 characteristics, out of which 2 were categorical and 16 were numerical. The clinical parameters that were available in the dataset (18 attributes) included age, gender, resting blood pressure, maximum HR, old peak, creatine phosphokinase, ejection fraction, platelet count, serum creatinine, serum sodium, time, systolic and diastolic blood pressure, HR (bpm), cholesterol level, low density lipoprotein (LDL) level, high density lipoprotein (HDL) level, and CVD prognosis.

The initial stage of data processing employs a tool for imputation that calculates the average frequency of missing data attributes. After that, a range of classification techniques were employed to model parameters, including KNN, SVM, RF, artificial neural network (ANN), naïve Bayes, LR, SGD, and AdaBoost, to identify the most effective performance analysis and assess classification accuracy. Subsequently, we utilized various unsupervised ML clustering methods, such as k-means, hierarchical, and density-based spatial clustering of applications with noise (DBSCAN) clustering, to determine the number of clusters for CVD patients. The Orange data mining software was utilized for all analyses.

Results and discussion

Performance analysis and classification accuracy on Cardiovascular Disease Prognostic datasets

Figure 4 shows the overall layout view of the classification techniques on Cardiovascular Disease Prognostic datasets.

Based on the outcomes and evaluations presented in Fig. 5, it is evident that SGD and ANN are the most efficient techniques for categorizing CVD predictive information.

The accuracy of classification refers to the percentage of correctly classified instances, whereas the accuracy of words pertains to the degree of proximity between a group of measurements and their actual values. The accuracy matrix for classification can be observed in Fig. 6.

Figure 7 demonstrates that SGD and ANN models outperform alternative methods of classification with regards to prognostic data for CVD. The calibration graph charts the anticipated probabilities of the classifier in relation to class probabilities and can serve as a means of verifying if the classifier is excessively hopeful or despondent. Additionally, the tool can exhibit a calibrated model where the user can set their own probability threshold.

Figure 8 displays the F1 score chart which links the models of the attribute categorization methodology. This is a harmonic assessment of precision and recall,

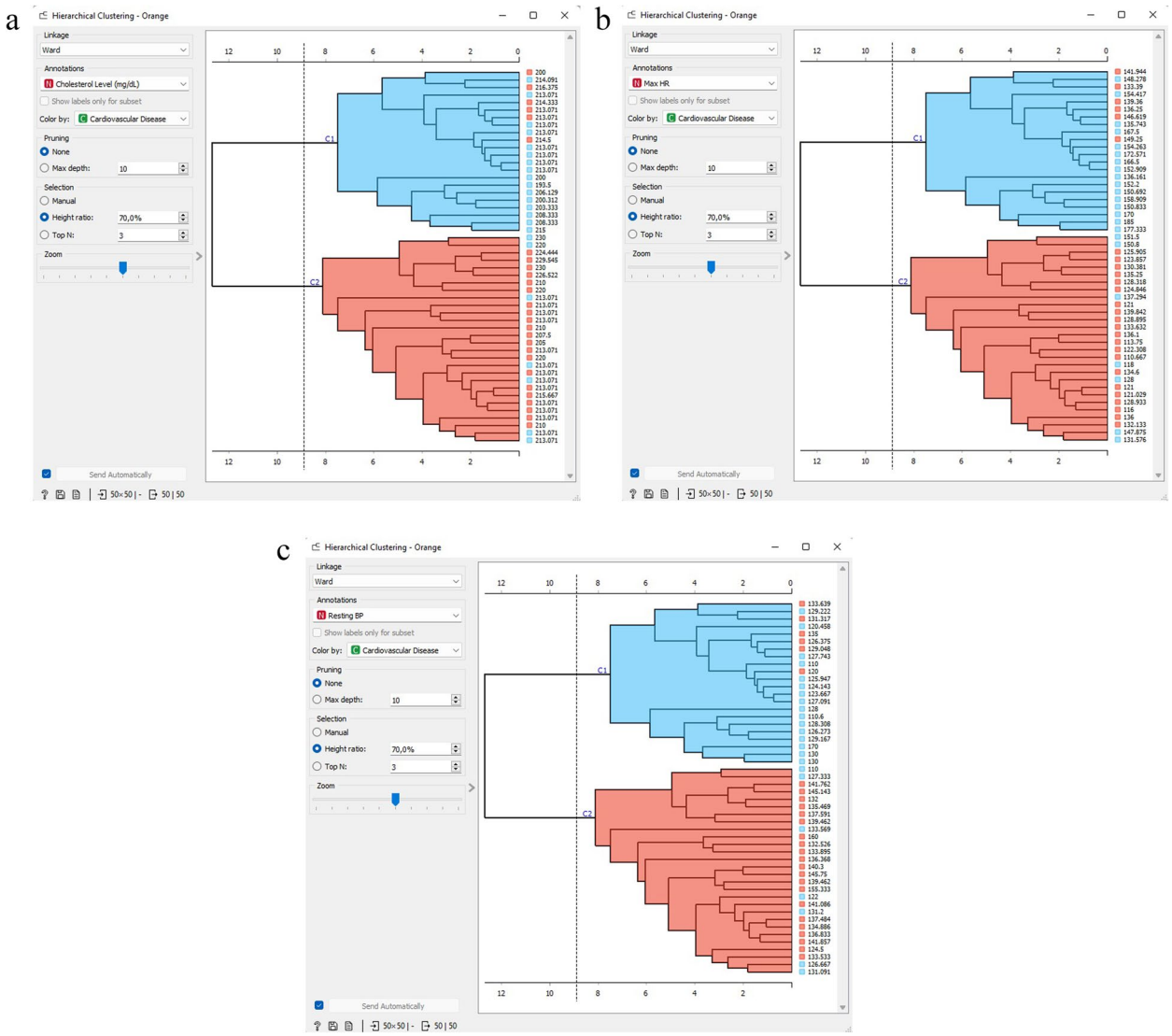


Fig. 19 Hierarchical clustering. **a** Hierarchical clustering based on the cholesterol level attribute; **b** Hierarchical clustering based on the maximum HR attribute; **c** Hierarchical clustering based on the resting blood pressure attribute

signifying both precision and recall in a single measure. The maximum attainable score was 1, indicating impeccable precision and recall, while the minimum was 0.

In the fields of ML, object detection, classification, pattern recognition, and information retrieval, precision and recall serve as performance metrics for data extracted from a sample space, corpus, or collection. Precision, also referred to as positive predictive value, represents the ratio of retrieved instances that are relevant, as depicted in Fig. 9. Sensitivity, or recall, represents the ratio of relevant instances that are retrieved. Consequently, both precision and recall are founded on relevance, as illustrated in Fig. 10.

Figures 11 and 12 demonstrate that both SGD and ANN models outperform alternative classification techniques for prognostic data related to CVD. The calibration curve depicts the anticipated probabilities of the classifier in relation to the class probabilities, enabling the assessment of whether the classifier is overly optimistic or pessimistic. The tool also presents a calibrated model that allows the user to establish their own probability threshold.

Figure 13 demonstrates that the ANN framework exhibited the most optimal performance for categorizing

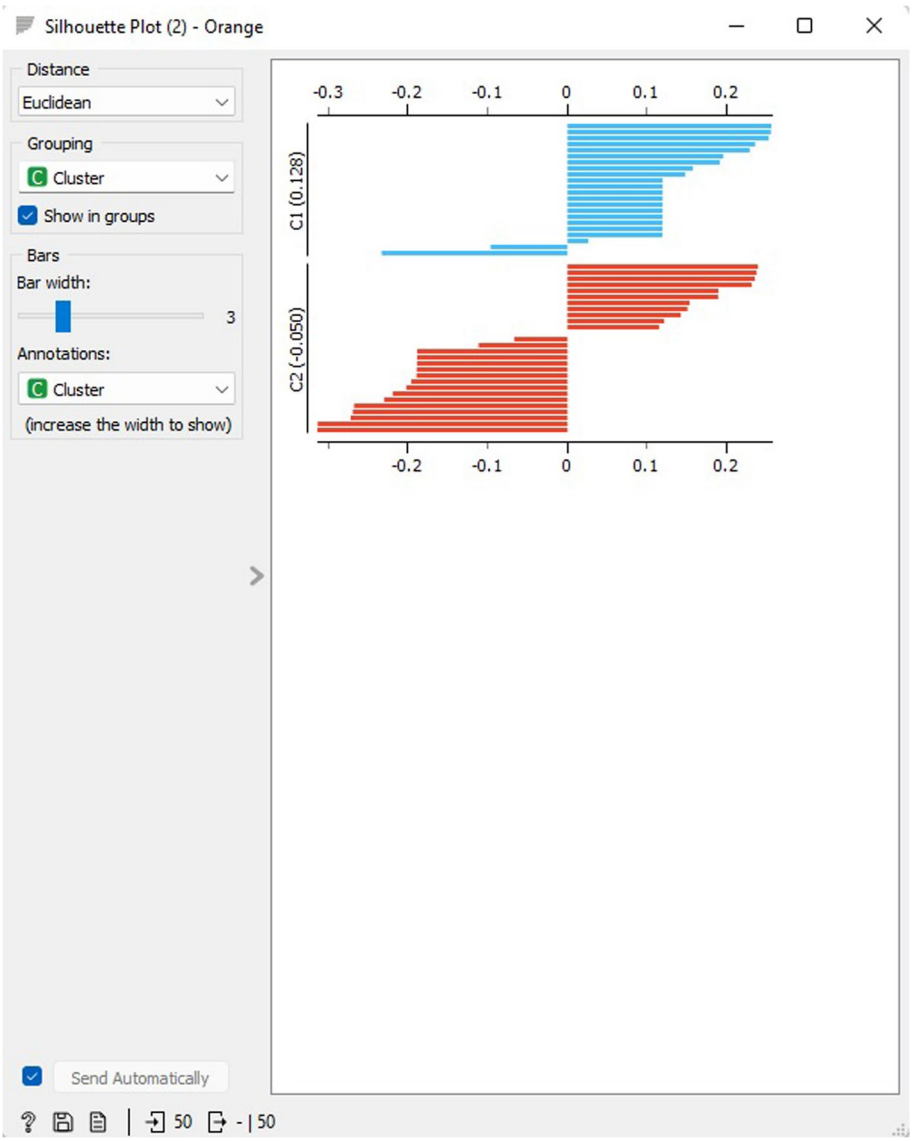


Fig. 20 Hierarchical clustering silhouette scores

CVD prognostic data. The performance graph represents the ratio of accurate positive data instances in comparison to the classifier's threshold, while the cumulative return diagram showcases the correlation between actual positive cases and the support. The greater the region between the curve and the baseline (dashed line), the more exceptional the model.

Determining the number of clusters on Cardiovascular Disease Prognostic datasets

Figure 14 shows an overall layout view of the clustering techniques on Cardiovascular Disease Prognostic datasets.

(1) K-means clustering on Cardiovascular Disease Prognostic datasets

In the illustration depicted as Fig. 15, the CVD prognostic dataset is partitioned into two clusters using the k-means cluster technique, with a silhouette score of 0.175. The widget algorithm utilizes k-means clustering to process the data and generates an updated dataset that includes the cluster label as a meta-attribute. Additionally, the widget presents the silhouette points of the group outcomes for various k values. A higher silhouette score indicates superior grouping.

Figure 16 displays the k-means clustering silhouette outcomes for two groupings. Cluster 1 (C1) has an

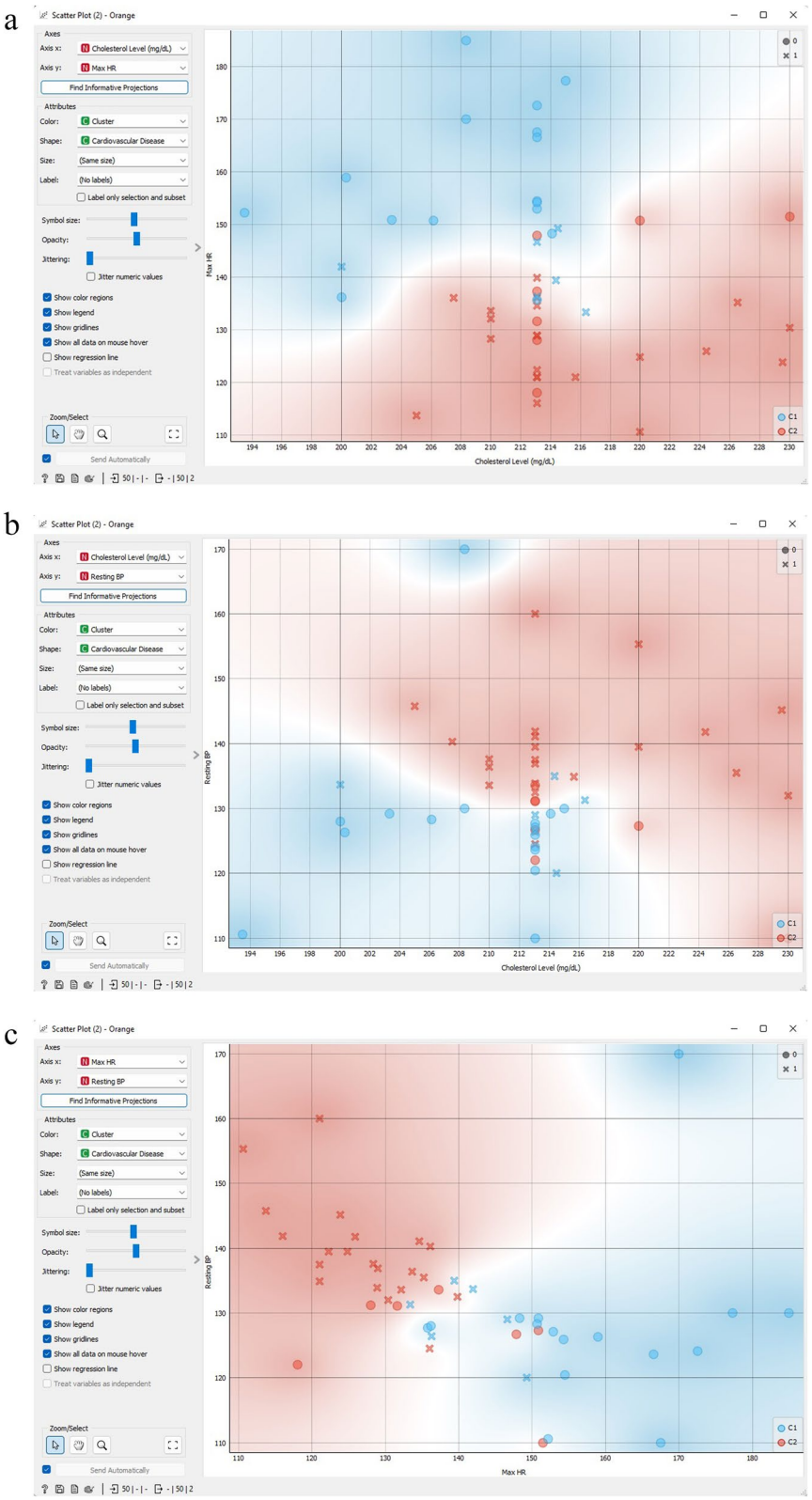


Fig. 21 Hierarchical cluster scatter diagram that illustrates the correlation between cholesterol level, maximum HR, and resting blood pressure. **a** Hierarchical clustering scatter plot correlations between cholesterol level and maximum HR attributes; **b** Hierarchical clustering scatter plot correlations between cholesterol level and resting blood pressure attributes; **c** Hierarchical clustering scatter plot correlations between maximum HR and resting blood pressure attributes

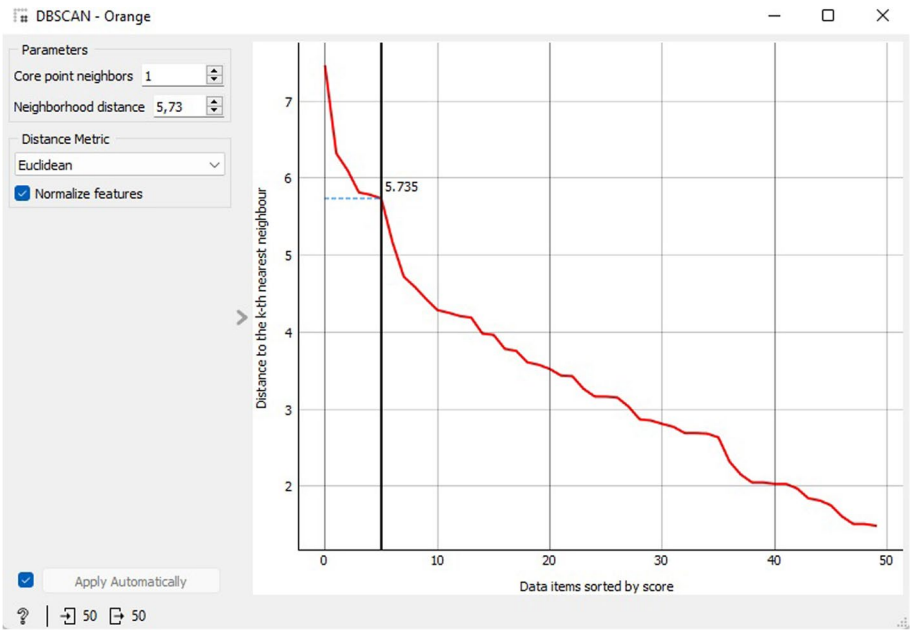


Fig. 22 DBSCAN clustering

average value of -0.115 and cluster 2 (C2) has an average value of 0.097. The Silhouette Graph widget presents a pictorial representation of the uniformity of data groupings and enables users to visually evaluate the grouping quality. The silhouette value signifies how comparable an item is to its grouping in comparison to other groupings, and events with a silhouette value close to 1 suggest that the data point is in proximity to the center of the grouping, while events with a silhouette value close to 0 are situated at the boundary of the two groupings.

Figure 17 exhibits a scatter plot of k-means cluster analysis that depicts the correlation among cholesterol level, maximum HR, and resting blood pressure. The scatter-plot tool showcases a two-dimensional scatter, where information is exhibited as a set of dots with x-axis and y-axis attribute values. On the widget's left-hand side, several chart features, including dot hue, magnitude and shape, axis headings, maximum dot size, and jitter, can be modified.

(2) Hierarchical clustering on Cardiovascular Disease Prognostic datasets

In the CVD prognostic dataset, Fig. 18 illustrates the normalized distances between rows and columns. The objective of normalization was to guarantee impartial treatment of individual attributes, and it was executed per column.

Figure 19 depicts the outcomes of the hierarchical clustering, which were segregated into two clusters. The distance tool computes the hierarchical clustering of diverse object categories from the distance array and exhibits the associated dendrogram.

The outcomes of the hierarchical cluster silhouette analysis for two clusters are presented in Fig. 20, with C1 exhibiting an average score of 0.128 and C2 exhibiting an average score of -0.050. The Silhouette Plot widget enables users to assess the quality of the data clusters visually and depicts the consistency of the clusters. The silhouette score indicates how comparable an object is to its cluster in contrast to other clusters, and objects with a silhouette score near 1 suggest that the data point is located in the center of the cluster, while objects with a silhouette score near 0 are found at the boundary of the two clusters.

Figure 21 exhibits a hierarchical cluster scatter diagram that illustrates the correlation between cholesterol level, maximum HR, and resting blood pressure. The scatter-plot gadget presents a two-dimensional scatter, where the data is portrayed as a set of dots with x-axis and y-axis characteristic values. On the widget's left-hand side, different chart features, such as point hue, magnitude and form, axis headings, maximum point size, and jitter, can be modified.

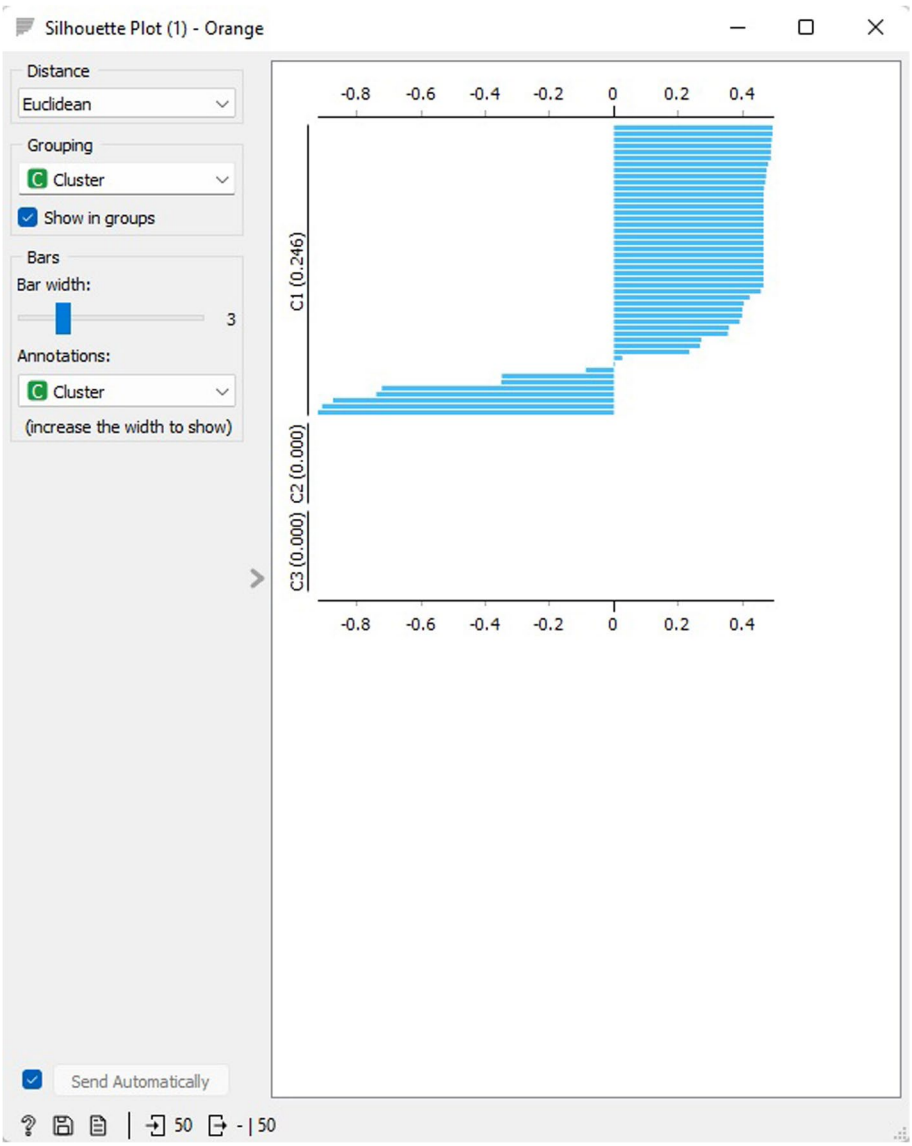


Fig. 23 DBSCAN clustering silhouette scores

(3) DBSCAN clustering on Cardiovascular Disease Prognostic datasets

Figure 22 illustrates the results obtained from the application of DBSCAN. The optimal number of clusters for the CVD prognostic datasets under DBSCAN was segregated into 3 clusters, where the core point neighbors were 1, neighborhood distance was 5.73, and the distance metric was Euclidean. The widget utilizes the DBSCAN clustering algorithm on the data, resulting in a fresh dataset with group identities as meta-attributes. Additionally, it exhibits an ordered chart depicting the k-th nearest neighbor distances, provided the k-values pertain to the core point neighbors. The chart exhibits the distance to the k-th nearest neighbor, ascertained by selecting the core point's neighborhood. The correct inclusive range can be chosen by shifting the black slider to the left or right.

The DBSCAN cluster silhouette results for 3 clusters are presented in Fig. 23. C1 displays an average score of 0.246, while C2 and C3 both have an average score of 0.000. The Silhouette Plot widget offers a visual representation of the consistency of data clusters, enabling users to evaluate the quality of the clusters. The silhouette score indicates the similarity between an

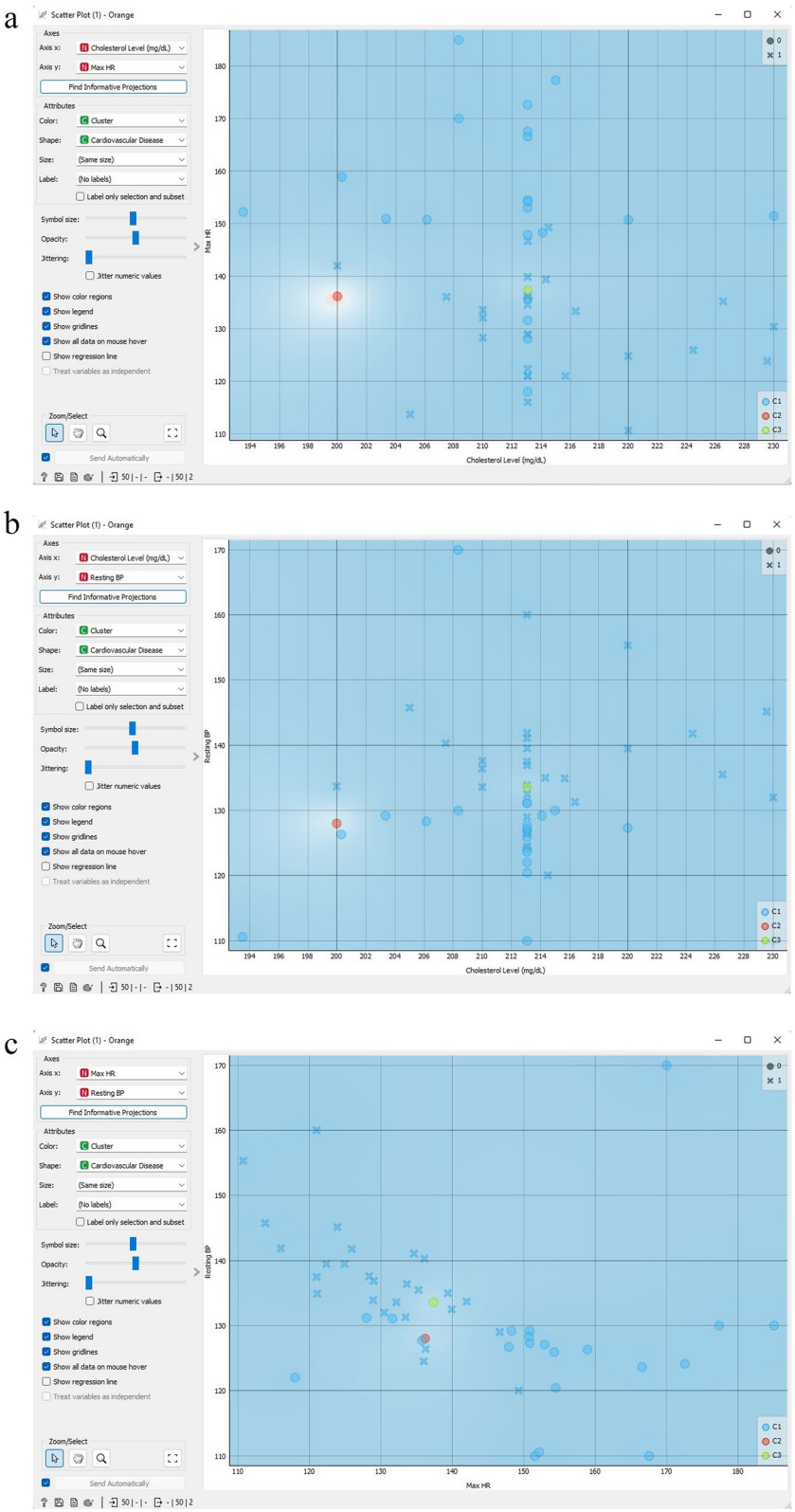


Fig. 24 The scatter plot of DBSCAN cluster that demonstrates the correlation among cholesterol level, maximum HR, and resting blood pressure. **a** DBSCAN clustering scatter plot correlations between cholesterol level and maximum HR attributes; **b** DBSCAN clustering scatter plot correlations between cholesterol level and resting blood pressure attributes; **c** DBSCAN clustering scatter plot correlations between maximum HR and resting blood pressure attributes

Table 1 Comparative results of k-means, hierarchical, and DBSCAN clustering in the CVD prognostic datasets

Clustering methods	K-means clustering	Hierarchical clustering	DBSCAN clustering
The best number of clusters	2	2	3
Distance metric	Euclidean	Euclidean	Euclidean
Silhouette scores	C1 = -0.115 C2 = 0.097	C1 = 0.128 C2 = -0.050	C1 = 0.246 C2 = 0.000 C3 = 0.000

object and its cluster in comparison to other clusters. Scores close to 1 suggest that the data event is positioned close to the center of the cluster, whereas scores close to 0 indicate that the data event is situated at the border of the three clusters.

Figure 24 presents the scatter plot of DBSCAN cluster that demonstrates the correlation among cholesterol level, maximum HR, and resting blood pressure. The scatter-plot tool exhibits a scatter with two dimensions, where the data is exhibited as a group of points with x-axis and y-axis characteristic values. The widget's left side allows the customization of different chart characteristics, including point color, size, and shape, axis headings, maximum point size, and jitter.

Table 1 indicates that the CVD predictive datasets can be categorized into two groups, determined by the outcomes of various clustering techniques, including k-means clustering and Hierarchical clustering.

Conclusions

This research used various datasets on CVD prognosis in a case study of patient information from prominent hospitals in the United States. The participants were individuals aged between 28-77 years. We procured varied datasets from Kaggle, which assembles public information from websites, like frequent visitors, without compromising personal data. The data comprised observational findings of 918 patients from one of the most notable hospitals in the United States. The datasets had 18 characteristics, out of which 2 were categorical and 16 were numerical. The clinical parameters that were available in the dataset (18 attributes) included age, gender, resting blood pressure, maximum HR, old peak, creatine phosphokinase, ejection fraction, platelet count, serum creatinine, serum sodium, time, systolic and diastolic blood pressure, HR, cholesterol level, LDL level, HDL level, and CVD prognosis.

The initial stage of data processing employs a tool for imputation that calculates the average frequency of missing data attributes. After that, a range of classification techniques were employed to model parameters, including KNN, SVM, RF, ANN, naïve Bayes, LR, SGD, and AdaBoost, to identify the most effective performance analysis and assess classification accuracy. Subsequently, we utilized various unsupervised ML clustering methods, such as k-means, hierarchical, and DBSCAN clustering, to determine the number of clusters for CVD patients. The Orange data mining software was utilized for all analyses.

The results showed that the most outstanding performance analysis and classification accuracy for CVD prognosis datasets were observed with SGD and ANN. The CVD prognosis datasets were able to be segregated into two clusters through clustering techniques like k-means and hierarchical clustering. The precision of the suggested model in determining the diagnostic model is crucial for the accuracy of CVD prognosis. The better the model's accuracy, the more reliable it becomes in predicting the patients who are susceptible to CVD.

Prognostic systems for CVDs are valuable in the upkeep and surveillance of patient populations, as well as in the reduction of mortality rates. These systems can serve as a significant tool in raising awareness about personal health and in the early detection and prevention of CVDs. The precision of CVDs prognostic diagnosis relies on the model's precision in determining the diagnostic model. Therefore, the more precise the model, the more accurate the prediction of patients who may be at risk of developing CVDs.

In this context, we suggest concepts for additional investigation of an unsupervised ML CVDs prognosis dataset.

a. Future researchers can explore alternative distance measures employed in different clustering techniques, like Fuzzy c-means and k-medoids clustering, and enhance the operational efficiency of modified k-means by diminishing the time complexity of Cardiovascular Disease Prognostic Rules.
b. Future researchers ought to employ a metaheuristic-based feature selection approach that takes features as input and organizes the original dataset population based on its features. The goal is to determine the least number of features that produce the least amount of error in

classifying samples and forecasting CVDs patients. Heightening the accuracy of the input data for the ML method should enable the learning model to recognize precise patterns for the diagnosis and prognosis of CVDs.

c. Future researchers ought to employ methods like Ant Colony Optimization Algorithms and Particle Swarm Optimization to enhance model efficacy. To attain superior forecast accuracy, they should adopt hybrid and ensemble models, and explore novel research opportunities in this domain through the application of predictive data mining techniques in medical diagnosis.

Abbreviations

CVD	Cardiovascular disease
COVID-19	Coronavirus disease of 2019
MERS	Middle East respiratory syndrome
KNN	K-nearest neighbors
SVM	Support vector machine
RF	Random forest
ANN	Artificial neural network
LR	Logistic regression
SGD	Stochastic gradient descent
DBSCAN	Density-based spatial clustering of applications with noise
CI	Cardiac insufficiency
HF	Heart failure
CKD	Chronic kidney disease
SDoH	Social determinants of health
ML	Machine learning
HIT	Health information technology
EHR	Electronic health records
GBC	Gradient boosting classifier
DT	Decision tree
SFS	Selected feature selection
ETC	Extra tree classifier
SMOTE	Synthetic minority oversampling technique
FCMIM	Fuzzy c-means index matrix
HR	Heart rate
AI	Artificial intelligence
ECG	Electrocardiography
LDL	Low density lipoprotein
HDL	High density lipoprotein
C	Cluster

Acknowledgements
Not applicable.

Authors' contributions
All authors contributed to the conception of the study and commented on previous versions of the manuscript. JS contributed to the conceptualization, writing-original draft preparation, writing-review and editing, data curation, methodology, investigation, software, formal analysis, and validation; CL contributed to the Software and formal analyses; JMS contributed to the data curation, methodology, validation, and investigation; S contributed to the supervision and project administration. All the authors have read and approved the final version of this manuscript.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing financial interests or personal relationships that may have influenced the work reported in this study.

Received: 15 March 2023 Accepted: 4 July 2023
Published online: 01 August 2023

References

1. Nanehkaran YA, Licai Z, Chen JD, Jamel AAM, Shengnan Z, Navaei YD et al (2022) Anomaly detection in heart disease using a density-based unsupervised approach. Wireless Commun Mobile Comput 2022:6913043. https://doi.org/10.1155/2022/6913043
2. Shorewala V (2021) Early detection of coronary heart disease using ensemble techniques. Inf Med Unlocked 26:100655. https://doi.org/10.1016/j.imu.2021.100655
3. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS et al (2022) Heart disease and stroke statistics - 2022 update: a report from the American heart association. Circulation 145(8):e153-e639. https://doi.org/10.1161/CIR.0000000000001052
4. Zhao Y, Wood EP, Mirin N, Cook SH, Chunara R (2021) Social determinants in machine learning cardiovascular disease prediction models: a systematic review. Am J Prev Med 61(4):596-605. https://doi.org/10.1016/j.amepre.2021.04.016
5. Şahin B, İlgün G (2022) Risk factors of deaths related to cardiovascular diseases in World Health Organization (WHO) member countries. Health Soc Care Community 30(1):73-80. https://doi.org/10.1111/hsc.13156
6. The Writing Committee of the Report on Cardiovascular Health and Diseases in China (2022) Report on cardiovascular health and diseases in China 2021: an updated summary. Biomed Environ Sci 35(7):573-603. https://doi.org/10.3967/bes2022.079
7. Faghy MA, Yates J, Hills AP, Jayasinghe S, Da Luz Goulart C, Arena R et al (2023) Cardiovascular disease prevention and management in the COVID-19 era and beyond: an international perspective. Prog Cardiovasc Dis 76:102-111. https://doi.org/10.1016/j.pcad.2023.01.004
8. Dao Trong P, Olivares A, El Damaty A, Unterberg A (2023) Adverse events in neurosurgery: a comprehensive single-center analysis of a prospectively compiled database. Acta Neurochir 165(3):585-593. https://doi.org/10.1007/s00701-022-05462-w
9. Möller-Leimkühler AM (2022) Gender differences in cardiovascular disease and comorbid depression. Dialogues Clin Neurosci 9(1):71-83. https://doi.org/10.31887/DCNS.2007.9.1/ammoeller
10. Boukhris M, Hillani A, Moroni F, Annabi MS, Addad F, Ribeiro MH et al (2020) Cardiovascular implications of the COVID-19 pandemic: a global perspective. Can J Cardiol 36(7):1068-1080. https://doi.org/10.1016/j.cjca.2020.05.018
11. Bhatt AS, Daniels LB, De Lemos J, Goodrich E, Bohula EA, Morrow DA (2023) Multi-marker risk assessment in patients hospitalized with COVID-19: results from the American heart association COVID-19 cardiovascular disease registry. Am Heart J 258:149-156. https://doi.org/10.1016/j.ahj.2022.12.014
12. Xie Y, Xu E, Bowe B, Al-Aly Z (2022) Long-term cardiovascular outcomes of COVID-19. Nat Med 28(3):583-590. https://doi.org/10.1038/s41591-022-01689-3
13. Dale CE, Takhar R, Carragher R, Katsoulis M, Torabi F, Duffield S et al (2023) The impact of the COVID-19 pandemic on cardiovascular disease prevention and management. Nat Med 29:219-225. https://doi.org/10.1038/s41591-022-02158-7
14. Yamamoto T, Harada K, Yoshino H, Nakamura M, Kobayashi Y, Yoshikawa T et al (2023) Impact of the COVID-19 pandemic on incidence and mortality of emergency cardiovascular diseases in Tokyo. J Cardiol 82(2):134-139. https://doi.org/10.1016/j.jjcc.2023.01.001
15. Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N et al (2020) Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the Multicentre Italian CORIST study.

Nutr Metab Cardiovasc Dis 30(11):1899-1913. https://doi.org/10.1016/j.numecd.2020.07.031
16. Li MY, Dong YL, Wang HJ, Guo WN, Zhou HF, Zhang ZL et al (2020) Cardiovascular disease potentially contributes to the progression and poor prognosis of COVID-19. Nutr Metab Cardiovasc Dis 30(7):1061-1067. https://doi.org/10.1016/j.numecd.2020.04.013
17. Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G et al (2021) Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC Heart Failure 8(1):106-115. https://doi.org/10.1002/ehf2.13073
18. Skalidis I, Muller O, Fournier S (2022) CardioVerse: the cardiovascular medicine in the era of Metaverse. Trends Cardiovasc Med 2022 May 11:S1050-1738(22)00071-8. https://doi.org/10.1016/j.tcm.2022.05.004
19. Davazdahemami B, Zolbanin HM, Delen D (2022) An explanatory analytics framework for early detection of chronic risk factors in pandemics. Healthcare Anal 2:100020. https://doi.org/10.1016/j.health.2022.100020
20. Alanazi R (2022) Identification and prediction of chronic diseases using machine learning approach. J Healthcare Eng 2022:2826127. https://doi.org/10.1155/2022/2826127
21. Alqaissi EY, Alotaibi FS, Ramzan MS (2022) Modern machine-learning predictive models for diagnosing infectious diseases. Comput Math Methods Med 2022:6902321. https://doi.org/10.1155/2022/6902321
22. Muşat EC, Borz SA (2022) Learning from acceleration data to differentiate the posture, dynamic and static work of the back: an experimental setup. Healthcare 10(5):916. https://doi.org/10.3390/healthcare10050916
23. Astorga F, Groom Q, Shimabukuro PHF, Manguin S, Noesgaard D, Orrell T et al (2023) Biodiversity data supports research on human infectious diseases: global trends, challenges, and opportunities. One Health 16:100484. https://doi.org/10.1016/j.onehlt.2023.100484
24. Ciumărnean L, Milaciu MV, Negrean V, Orășan OH, Vesa SC, Sălăgean O et al (2022) Cardiovascular risk factors and physical activity for the prevention of cardiovascular diseases in the elderly. Int J Environ Res Public Health 19(1):207. https://doi.org/10.3390/ijerph19010207
25. Matsushita K, Jassal SK, Sang YY, Ballew SH, Grams ME, Surapaneni A et al (2020) Incorporating kidney disease measures into cardiovascular risk prediction: development and validation in 9 million adults from 72 datasets. eClinicalMedicine 27:100552. https://doi.org/10.1016/j.eclinm.2020.100552
26. Liu HX, Liu SQ, Wang K, Zhang TR, Yin L, Liang JQ et al (2022) Time-dependent effects of physical activity on cardiovascular risk factors in adults: a systematic review. Int J Environ Res Public Health 19(21):14194. https://doi.org/10.3390/ijerph192114194
27. Chieng D, Kistler PM (2022) Coffee and tea on cardiovascular disease (CVD) prevention. Trends Cardiovasc Med 32(7):399-405. https://doi.org/10.1016/j.tcm.2021.08.004
28. Powell-Wiley TM, Baumer Y, Baah FO, Baez AS, Farmer N, Mahlobo CT et al (2022) Social determinants of cardiovascular disease. Circ Res 130(5):782-799. https://doi.org/10.1161/CIRCRESAHA.121.319811
29. Minhas AMK, Jain V, Li M, Ariss RW, Fudim M, Michos ED et al (2023) Family income and cardiovascular disease risk in american adults. Sci Rep 13(1):279. https://doi.org/10.1038/s41598-023-27474-x
30. Mehbodniya A, Khan IR, Chakraborty S, Karthik M, Mehta K, Ali L et al (2022) Data mining in employee healthcare detection using intelligence techniques for industry development. J Healthcare Eng 2022:6462657. https://doi.org/10.1155/2022/6462657
31. Bhatt CM, Patel P, Ghetia T, Mazzeo PL (2023) Effective heart disease prediction using machine learning techniques. Algorithms 16(2):88. https://doi.org/10.3390/a16020088
32. Indrakumari R, Poongodi T, Jena SR (2020) Heart disease prediction using exploratory data analysis. Proc Comput Sci 173:130-139. https://doi.org/10.1016/j.procs.2020.06.017
33. Jayasri NP, Aruna R (2022) Big data analytics in health care by data mining and classification techniques. ICT Express 8(2):250-257. https://doi.org/10.1016/j.icte.2021.07.001
34. Dash S, Shakyawar SK, Sharma M, Kaushik S (2019) Big data in healthcare: management, analysis and future prospects. J Big Data 6(1):54. https://doi.org/10.1186/s40537-019-0217-0
35. Tougui I, Jilbab A, El Mhamdi J (2020) Heart disease classification using data mining tools and machine learning techniques. Health Technol 10(5):1137-1144. https://doi.org/10.1007/s12553-020-00438-1
36. Panzner M, Von Enzberg S, Meyer M, Dumitrescu R (2022) Characterization of usage data with the help of data classifications. J Knowl Econ. https://doi.org/10.1007/s13132-022-01081-z
37. Mpanya D, Celik T, Klug E, Ntsinjana H (2023) Clustering of heart failure phenotypes in johannesburg using unsupervised machine learning. Appl Sci 13(3):1509. https://doi.org/10.3390/app13031509
38. Beunza JJ, Puertas E, García-Ovejero E, Villalba G, Condes E, Koleva G et al (2019) Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). J Biomed Inf 97:103257. https://doi.org/10.1016/j.jbi.2019.103257
39. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R et al (2020) Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep 10(1):16057. https://doi.org/10.1038/s41598-020-72685-1
40. Alanazi A (2022) Using machine learning for healthcare challenges and opportunities. Inf Med Unlocked 30:100924. https://doi.org/10.1016/j.imu.2022.100924
41. Thange U, Shukla VK, Punhani R, Grobbelaar W (2021) Analyzing COVID-19 dataset through data mining tool "orange". In: Proceedings of the 2021 2nd international conference on computation, automation and knowledge management, IEEE, Dubai, 19-21 January 2021. https://doi.org/10.1109/ICCAKM50778.2021.9357754
42. Kaur I, Doja MN, Ahmad T (2022) Data mining and machine learning in cancer survival research: an overview and future recommendations. J Biomed Inf 128:104026. https://doi.org/10.1016/j.jbi.2022.104026
43. El-Hasnony IM, Elzeki OM, Alshehri A, Salem H (2022) Multi-label active learning-based machine learning model for heart disease prediction. Sensors 22(3):1184. https://doi.org/10.3390/s22031184
44. Ghorbani R, Ghousi R (2019) Predictive data mining approaches in medical diagnosis: a review of some diseases prediction. Int J Data Network Sci 3(2):47-70. https://doi.org/10.5267/j.ijdns.2019.1.003
45. Raykar SS, Shet VN (2020) Cognitive analysis of data mining tools application in health care services. In: Proceedings of the 2020 international conference on emerging trends in information technology and engineering, IEEE, Vellore, 24-25 February 2020. https://doi.org/10.1109/ic-ETITE47903.2020.442
46. Niu HR, Omitaomu OA, Langston MA, Olama M, Ozmen O, Klasky HB et al (2022) Detecting anomalous sequences in electronic health records using higher-order tensor networks. J Biomed Inf 135:104219. https://doi.org/10.1016/j.jbi.2022.104219
47. Jeong J, Kim YJ, Kong SY, Do Shin S, Ro YS, Wi DH et al (2022) Monitoring of characteristics of the patients visiting an emergency center in Cameroon through the development of hospital patient database. Afr J Emerg Med 12(1):77-84. https://doi.org/10.1016/j.afjem.2021.12.002
48. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A et al (2021) Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Comput Stat 13(6):e1549. https://doi.org/10.1002/wics.1549
49. Maiga J, Hungilo GG, Pranowo (2019) Comparison of machine learning models in prediction of cardiovascular disease using health record data. In: Proceedings of the 2019 International conference on informatics, multimedia, cyber and information system, IEEE, Jakarta, 24-25 October 2019. https://doi.org/10.1109/ICIMCIS48181.2019.8985205
50. Peng MX, Hou F, Cheng ZX, Shen TT, Liu KX, Zhao C et al (2023) A cardiovascular disease risk score model based on high contribution characteristics. Appl Sci 13(2):893. https://doi.org/10.3390/app13020893
51. Nouraei H, Nouraei H, Rabkin SW (2022) Comparison of unsupervised machine learning approaches for cluster analysis to define subgroups of heart failure with preserved ejection fraction with different outcomes. Bioengineering 9(4):175. https://doi.org/10.3390/bioengineering9040175
52. Sunny JS, Patro CPK, Karnani K, Pingle SC, Lin F, Anekoji M et al (2022) Anomaly detection framework for wearables data: a perspective review on data concepts, data analysis algorithms and prospects. Sensors 22(3):756. https://doi.org/10.3390/s22030756
53. Ripan RC, Sarker IH, Hossain SMM, Anwar MM, Nowrozy R, Hoque MM et al (2021) A data-driven heart disease prediction model through K-means clustering-based anomaly detection. SN Comput Sci 2(2):112. https://doi.org/10.1007/s42979-021-00518-7

54. Dalal S, Goel P, Onyema EM, Alharbi A, Mahmoud A, Algarni MA et al (2023) Application of machine learning for cardiovascular disease risk prediction. Comput Intell Neurosci 2023:9418666. https://doi.org/10.1155/2023/9418666
55. Magesh G, Swarnalatha P (2021) Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. Evol Intell 14:583-593. https://doi.org/10.1007/s12065-019-00336-0
56. Shrifan NHMM, Akbar MF, Isa NAM (2022) An adaptive outlier removal aided K-means clustering algorithm. J King Saud Univ Comput Inf Sci 34(8):6365-6376. https://doi.org/10.1016/j.jksuci.2021.07.003
57. Kaptoge S, Pennells L, De Bacquer D, Cooney MT, Kavousi M, Stevens G et al (2019) World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Global Health 7(10):e1332-e1345. https://doi.org/10.1016/S2214-109X(19)30318-3
58. Nadakinamani RG, Reyana A, Kautish S, Vibith AS, Gupta Y, Abdelwahab SF et al (2022) Clinical data analysis for prediction of cardiovascular disease using machine learning techniques. Comput Intell Neurosci 2022:2973324. https://doi.org/10.1155/2022/2973324
59. Ketkar Y, Gawade S (2022). A decision support system for selecting the most suitable machine learning in healthcare using user parameters and requirements. Healthcare Anal 2:100117. https://doi.org/10.1016/j.health.2022.100117
60. Aggrawal R, Pal S (2020) Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. SN Comput Sci 1(6):344. https://doi.org/10.1007/s42979-020-00370-1
61. Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V et al (2021) Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access 9:39707-39716. https://doi.org/10.1109/ACCESS.2021.3064084
62. Powar A, Shilvant S, Pawar V, Parab V, Shetgaonkar P, Aswale S (2019) Data mining & artificial intelligence techniques for prediction of heart disorders: a survey. In: Proceedings of the 2019 international conference on vision towards emerging trends in communication and networking, IEEE, Vellore, 30-31 March 2019. https://doi.org/10.1109/ViTECoN.2019.8899547
63. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A (2020) Heart disease identification method using machine learning classification in E-healthcare. IEEE Access 8:107562-107582. https://doi.org/10.1109/ACCESS.2020.3001149
64. Oyeleye M, Chen TH, Titarenko S, Antoniou G (2022) A predictive analysis of heart rates using machine learning techniques. Int J Environ Res Public Health 19(4):2417. https://doi.org/10.3390/ijerph19042417
65. Mohammedqasem R, Mohammedqasim H, Ata O (2022) Real-time data of COVID-19 detection with IoT sensor tracking using artificial neural network. Comput Electr Eng 100:107971. https://doi.org/10.1016/j.compeleceng.2022.107971
66. Ashton JJ, Young A, Johnson MJ, Beattie RM (2023) Using machine learning to impact on long-term clinical care: principles, challenges, and practicalities. Pediatr Res 93(2):324-333. https://doi.org/10.1038/s41390-022-02194-6
67. Javaid M, Haleem A, Singh RP, Suman R, Rab S (2022) Significance of machine learning in healthcare: features, pillars and applications. Int J Intell Networks 3:58-73. https://doi.org/10.1016/j.ijin.2022.05.002
68. Li B, Ding S, Song GL, Li JJ, Zhang Q (2019) Computer-aided diagnosis and clinical trials of cardiovascular diseases based on artificial intelligence technologies for risk-early warning model. J Med Syst 43(7):228. https://doi.org/10.1007/s10916-019-1346-x
69. Velu SR, Ravi V, Tabianan K (2022) Data mining in predicting liver patients using classification model. Health Technol 12(6):1211-1235. https://doi.org/10.1007/s12553-022-00713-3
70. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA (2021) Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 18(7):465-478. https://doi.org/10.1038/s41569-020-00503-2

Publisher's note