

Enhancing Cardiovascular Disease Prediction: A Domain Knowledge-Based Feature Selection and Stacked Ensemble Machine Learning Approach

Zahiriddin Rustamov (✉ contact@zahiriddin.com)

United Arab Emirates University

Jaloliddin Rustamov

United Arab Emirates University

Nazar Zaki

United Arab Emirates University

Sherzod Turaev

United Arab Emirates University

Most Sarmin Sultana

University Malaya

Jeanne Ywei Tan

University Malaya

Vimala Balakrishnan

University Malaya

Research Article

Keywords: cardiovascular disease, domain knowledge, feature selection, machine learning, prediction, stacking ensemble

Posted Date: June 26th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3068941/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Cardiovascular diseases (CVDs) are prevalent disorders affecting the heart or blood arteries. Early disease detection significantly enhances survival prospects, thus emphasizing the necessity for accurate prediction methods. Emerging technologies, such as machine learning (ML), present promising avenues for more precise prediction of CVDs. However, a critical challenge lies in developing models that not only ensure optimal predictive performance but also conform to well-established domain knowledge, thereby enhancing their credibility. Single classifiers often fall short due to issues like overfitting and bias. In response, this study proposes a domain knowledge-based feature selection integrated with a stacking ensemble classifier. The Framingham Heart Study, UCI Heart Disease and UAE retrospective cohort study datasets were utilized for training and evaluation of the ML algorithms. The results indicate that the proposed domain knowledge-based feature selection performs on par with frequently adopted feature selection techniques. Moreover, the proposed stacked ensemble, in conjunction with domain knowledge-based feature selection, achieved the highest metrics with 89.66% accuracy, and 89.16% F1-score on the Framingham dataset. Similarly, the proposed method achieved an F1-score of 85.26% and 96.23% on the UCI Heart Disease and UAE datasets. Furthermore, this study employs explainable AI techniques to illuminate the decision-making process of the predictive models. Thus, the study establishes that domain knowledge-based feature selection promotes the credibility of ML models without compromising predictive performance.

1. Introduction

Cardiovascular diseases (CVDs) are disorders affecting the heart or involving constricted blood arteries. According to [1], CVD is the leading cause of death, with the highest annual death rate. Early disease detection increases the likelihood of survival. High blood pressure, excessive alcohol consumption, cholesterol, stress, smoking, diabetes, high body mass index (BMI), age, and obesity are among the risk factors for CVDs [2, 3]. The traditional way of estimating risk based on the individual experiences of doctors and patients is highly error-prone since the analysis relies on the doctor's knowledge and experience, which is not always precise. Furthermore, manually gaining insight from a large medical dataset is challenging and tedious. As a result, advanced machine learning (ML) algorithms emerged, which are capable of processing and analysing vast quantities of complex medical data and providing a more accurate prediction of diseases, including CVDs [4–6]. ML algorithms are typically trained to make accurate predictions by learning from the dataset provided in less time than manual methods. However, studies have demonstrated that ensemble learning methods often perform better than single-learning methods [7]. The objective of the ensemble learning methods is to overcome the weaknesses of single learners and consolidate their strengths by combining different single learners and aggregating their outputs under a combination rule. The main reason for the use of this method is that it achieves a more accurate and stable predictive performance than the results obtained through single classification techniques [4]. One of the significant challenges in CVD prediction is managing the vast number of available features, which complicates the task of discerning essential attributes. Additionally, the

presence of numerous features can potentially impair the performance of ML models. However, extensive global research has elucidated significant risk factors for CVDs. Consequently, the judicious application of well-established domain knowledge when developing ML algorithms can confer trustable and verifiable credibility. While these advancements are promising, there is a growing need to make AI and ML models more transparent and explainable, especially in healthcare. The advent of explainable AI provides an opportunity to understand the underlying decision-making processes of complex ML models. In the context of CVD prediction, explainable AI techniques like LIME and SHAP can help validate the model predictions against domain knowledge and improve the trust in the model's predictions.

There is no exact cause of CVDs, but risk factors increase the risk of getting them. Identifying those at the highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. Machine learning algorithms (i.e., supervised learning algorithms) have been widely used to detect the presence of CVDs. The problem with the current classifier for CVD prediction is the lack of model credibility and reliance on single classifiers [8]. The proposed methods for CVD prediction often do not align well with established domain knowledge, leading to a credibility deficit [9]. Moreover, due to factors such as overfitting and bias, the single classifier could not ensure optimum prediction [4]. Thus, this study proposes a stacking ensemble classifier built upon a domain knowledge-based feature selection, which combines several single classifiers to produce an optimal predictive model. The main contributions of this study are as follows:

- Comprehensive review and assessment of cardiovascular disease risk factors as highlighted in existing literature, to guide the development of credible and trustworthy models.
- Implementation and performance comparison of domain knowledge-based feature selection against popular feature selection techniques, thereby enhancing the reliability and performance of our predictive models for cardiovascular diseases.
- Development and performance evaluation of a stacking ensemble model, leveraging the predictive power of various heterogeneous algorithms, demonstrating its superiority against other ensemble models and single classifiers.
- Investigation of several sampling techniques for effectively managing an imbalanced dataset, with a particular focus on the integration of the polynom-fit-SMOTE and Tomek Links method.
- Extensive validation of the proposed methods on two publicly available cardiovascular disease datasets (UCI Heart Disease and UAE Retrospective Cohort Study), which underscores the generalizability and robustness of our proposed techniques.
- Incorporation of explainable AI techniques, namely SHAP and LIME, to ensure model interpretability and transparency, thereby enhancing the credibility of our predictions and fostering greater trust in ML-driven health prognosis.

The remainder of this paper is organized as follows: Section 2 discusses the relevant literature on cardiovascular diseases, their risk factors, and the methods ML practitioners use to predict cardiovascular diseases. Next, Section 3 presents the methodology utilized in this study. Subsequently,

Section 4 discusses the results of this study. Finally, Section 5 concludes the study and highlights future works.

2. Literature Review

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. There exist different types of CVDs, which can have various causes on the body. CVDs are associated with risk factors that may cause CVDs, such as hypertension, diabetes, obesity, smoking, body mass index (BMI), and high cholesterol. Gender, race, age, and family history are among the risk factors individuals have no control over [10]. This study refers to gender as the binary sex categorization (i.e., male and female) designated at birth. Although there is no exact cause of CVDs, the most important is the timely identification of CVDs. According to [11], blood tests, electrocardiograms, Holter monitoring, echocardiograms, cardiac catheterisation, cardiac computerised tomography scans, and cardiac magnetic resonance imaging are among the common ways to determine if someone has CVD. However, the diagnostic test depends on the doctor's assessment of the individuals' condition. In addition, these tests are usually time-consuming, painful, or invasive. Computer-based diagnostic methods that use machine learning predictive models can be a non-invasive alternative, assisting doctors in making proper and objective diagnoses. As a result, such methods ease the experience of the patient [12].

One of the challenges in the development of ML algorithms is the process of feature selection, which is to accurately select a subset of features that provide the highest predictive power while reducing the complexity of the model. Although numerous feature selection methods are proposed and utilized for the task of CVD prediction, the reasoning of the selected features may not always conform with well-established domain knowledge. The interpretability of the methods used in training ML algorithms on well-established knowledge promotes the reliability and verifiability of the models. Ultimately, the developed models will be interpretable and credible as they conform to the well-established domain knowledge. For that reason, this study aims to analyse and assess the risk factors of cardiovascular diseases that are considered significant by the respective domain and use the acquired knowledge to perform feature selection on the selected datasets for the task of CVD prediction. The analysis of significant risk factors has been conducted from a small scale, such as in a university setting, to a larger scale, covering many countries globally.

[13] conducted a study to estimate the risk of developing cardiovascular diseases over ten years among the administrative staff at Zagazig University, Egypt. The study worked with 116 employees, of which 64 were males and 52 females. The authors studied the following risk factors: BMI, gender, waist circumference, physical activity, cigarette smoking, family history, diabetes, hypertension, blood cholesterol level, and high-density lipoprotein cholesterol (HDL-C). The findings show that participants with high CVD risk were mainly inactive, smokers with positive family history, diabetics, hypertensive, with high blood cholesterol levels, HDL-C, and central obesity. In a similar university setting, a study by [14] sought to determine the prevalence of modifiable risk factors among the University of Ghana students. The study was conducted on 120 students aged 25 to 45. The authors considered BMI, total triglycerides,

total cholesterol, low-density lipoprotein cholesterol (LDL-C), HDL-C, systolic blood pressure, diastolic blood pressure, and visceral fat as the risk factors for analysis. The study's results show that BMI, LDL-C, total triglycerides, total cholesterol, HDL-C, and visceral fats are all essential predictors of CVDs among the students studied. Exploring the rural and urban setting, [15] conducted a study to assess the prevalence, magnitude, and factors associated with the number of modifiable cardiovascular disease (CVD) risk factors. The study selected 2483 adults, aged 25 to 64, using a three-stage random sampling method from urban and rural Wolaita, Ethiopia. Age, gender, residence, education, occupation, wealth index, marital status, fruit consumption, vegetable consumption, sugar in beverages, and food and alcohol consumption were the risk factors under study. The findings demonstrate that urban residence, male gender, sugar in food and older age positively associated with the number of major modifiable CVD risk factors while being a farmer and walking as physical activity had a negative association. [16] conducted a study to assess the lifestyle risk factors and biomarkers associated with hypertension and coronary heart disease among patients attending cardiac clinics at Kilimanjaro Christian Medical Centre referral hospital in Tanzania. The authors studied Kilimanjaro Christian Medical Centre referral hospital between April to July 2008, adults aged 35 years or older. The study considered smoking, physical inactivity, alcohol intake, overweight, obesity, blood pressure, hyperglycemia, high alanine aminotransferase, high HDL-C, high LDL-C, and C-reactive protein as risk factors under study. The results of the study show that the most prevalent risk factors for hypertension and coronary heart diseases were alcohol intake (67%), high blood pressure (59%), physical inactivity (61%), obesity (39%), alanine aminotransferase (43%), HDL-C (79%), LDL-C (65%), C-reactive protein (78%), sodium (41%) and potassium (40%). Moreover, age, plasma glucose, alanine aminotransferase, and C-reactive protein were found to be independently and positively associated with hypertension and coronary heart diseases.

Studies have also targeted specific genders to analyse the risk factors of CVDs. [17] compared the strength and causality of associations between significant risk factors for cardiovascular disease (CVD) and the four major CVDs: myocardial infarction, ischaemic stroke, heart failure, and atrial fibrillation. They studied 2322 male residents aged 50 years of Uppsala, Sweden, between 1970 and 1974. The study analysed the following risk factors: triglycerides, HDL-C, LDL-C, BMI, systolic blood pressure, diabetes, and smoking. The study's findings show that systolic blood pressure was the most critical risk factor since it was related to all CVDs examined. BMI was the second most important risk factor linked to all four CVDs, except ischaemic stroke. Smoking was a significant risk factor for ischaemic stroke and heart failure, while LDL-C was mainly related to myocardial infarction. Diabetes was a causal risk factor for incident myocardial infarction and heart failure. Neither HDL-C nor triglycerides were majorly essential risk factors in these multivariable models. On the other hand, [18] aimed to determine the prevalence of cardiovascular disease risk factors in women aged 30 to 60, covered by health centres in Ardabil, Iran. The study was conducted on 1006 women selected using multistage random sampling. The study analysed the following risk factors: systolic blood pressure, diastolic blood pressure, BMI, triglyceride, cholesterol, HDL-C, LDL-C, fasting blood sugar, and waist circumference. The study's results show that abdominal obesity (waist size above 88 cm), obesity and overweight, high triglycerides, high cholesterol, and LDL-C and HDL-C outside the acceptable range were the most critical risk factors for cardiovascular

disease among women. While [19] conducted a comparative study between both genders, aiming to estimate the 9-year incidence rate of CVD and deduce the risk factors associated with CVD among UAE nationals at high cardiovascular risk. In addition, they explored whether sex was a significant modifier of the risk factors associated with incident CVD in this population. The study was conducted on 977 subjects, of which 492 were women aged 18 years or older, who did not have a history of CVD, and who had one or more risk factors for CVD. The risk factors analysed in this study were age, smoking history, antidiabetic drugs, antihypertensive agents, lipid-lowering medication, BMI, systolic blood pressure, diastolic blood pressure, total cholesterol, HDL-C, TC/HDL-C ratio, LDL-C, triglycerides, HbA1c glycosylated hemoglobin, glomerular filtration rate (eGFR). The results of the study show that men had a greater risk of incident major CVD than women during this study's 9-year follow-up period. High systolic blood pressure, low eGFR, and an elevated HbA1c level were independent risk factors linked to major CVD in both sexes. Age and history of smoking among men and a high TC/HDL-C ratio in women were additional risk factors associated with incident major CVD. The findings were further solidified by [20], who conducted a study to assess the four risk score calculators linked to CVD for superiority and the most reliable CVD prognosis parameters. The study was conducted on 150 individuals aged between 30 and 79 years in India. The risk factors analysed were the following: age, systolic blood pressure, weight, diastolic blood pressure, white blood cell count, haemoglobin, height, neutrophil percentage, eosinophil percentage, total cholesterol, red blood cell count, monocyte percentage, LDL-C, HDL-C, cholesterol/HDL ratio, platelets (lack/cmm). The result of the study demonstrates that the most common risk factors for CVD were age, weight, total cholesterol, haemoglobin level, HDL, and LDL cholesterol. In addition, the study indicated that age, weight, total cholesterol, and haemoglobin level could be used as prognostic markers.

While the earlier studies targeted populations at a smaller scale, a study by [21] examined the association between self-reported risk factors and CVD presence among adults in Nevada, United States of America, between 2011 and 2017. The study was conducted on 5493 and 3764 subjects in 2011 and 2017, respectively. The study analysed the following risk factors: gender, age, alcohol intake, cigarette smoking, cholesterol, hypertension, BMI, and diabetes. The study's findings present that BMI (i.e., overweight, or obese) remained the most prevalent CVD risk factor. The second most common CVD risk factor was high cholesterol, followed by hypertension. Compared to females, males were 1.64 times more likely to have reported CVD in 2011, which increased to 1.92 in 2017. Compared to non-smokers, every day and former smokers were close to 2 times more likely in 2011 and 3 times more likely in 2017. Individuals with high cholesterol status were 2.67 times more likely to have reported CVD than those with normal levels in 2011. In 2011, individuals with hypertension were 3.74 times more likely to have reported CVD compared to those who did not have hypertension. This relationship increased its magnitude of risk to 6.18 times more likely in 2017. In 2011, individuals with diabetes were 2.90 times more likely to have reported CVD than those without the condition. The study by [22] also confirmed the results with obesity as a major driver of cardiovascular risk and associated strongly with other cardiovascular disease risk factors, such as hypertension, diabetes, and high cholesterol. Regarding longitudinal studies, [23] presented some of the most relevant findings regarding the causes of CVD proceeding from the Framingham Heart Study,

focusing on identifying risk factors, analysing their predictive capacity, and the consequences of these findings for prevention. They utilised the Framingham Heart Study data initially conducted in 1948 for biennial examinations, including 5209 healthy residents aged between 30 and 60 years. In 1971, 5124 sons and daughters (and their spouses) of the original participants were recruited for the Offspring Study. Finally, in 2002, 4095 participants were included in the Third-Generation cohort of the study. The risk factors studied were lipids (HDL-C, LDL-C, total serum cholesterol, and triglycerides), hypertension (diastolic blood pressure, systolic blood pressure), smoking, diabetes, physical inactivity, and BMI. The results of the study confirm that there is a strong relation between serum total cholesterol, LDL-C, and CVD. Moreover, isolated systolic hypertension was also a powerful predictor of cardiovascular disease. More importantly, the study demonstrated that systolic and diastolic blood pressure has a continuous, independent, graded, and positive association with cardiovascular outcomes. Diabetes is associated with 2 to 3 folds increase in the likelihood of developing CVDs. Obesity is associated with numerous comorbidities such as CVD, type 2 diabetes, and hypertension. Similarly, [24] conducted a study to evaluate the importance of risk factors in individuals and in the community using Cox regression analysis as relative risks and population-attributable risks, respectively. The study used The Copenhagen City Heart Study data which followed for 21 years a random sample of 5599 men and 6478 women aged 30 to 79 years at baseline in Rigshospitalet, Copenhagen, Denmark. Diabetes mellitus, hypertension, smoking, physical inactivity, alcohol intake, hypercholesterolemia, obesity, low or middle income, hypertriglyceridemia, and school education were the risk factors under study. The results of the study demonstrate that relative risks in men ranged from 1.69 to 1.20, with the highest risks for diabetes, hypertension, smoking, and physical inactivity. In women, relative risks ranged from 2.74 to 1.19, with the highest risks for diabetes, smoking, hypertension, and physical inactivity. Using the same data, [25] conducted a study to rank psychosocial and traditional risk factors of coronary heart disease (CHD) by importance. The study analysed the third examination of The Copenhagen City Heart Study, which happened from 1991 to 1994, with 8882 participants who did not have a history of CHD. The study studied the following risk factors: vital exhaustion, systolic blood pressure, total cholesterol, HDL-C, smoking, school education, family history of CVD, physical activity, alcohol intake, BMI, social network, life events during childhood, work-related life events and life events during adolescence. Forward selection, attributable population fraction, and gradient boosting machine were used for determining ranks. The importance of vital exhaustion for risk prediction was investigated by C-statistics and net reclassification improvement. The results of the study show that in men, the highest-ranking risk factors for coronary heart disease were vital exhaustion, systolic blood pressure, total cholesterol, HDL-C, and smoking. In women, smoking was significant, followed by vital exhaustion, HDL-C, and systolic blood pressure. [26] conducted a nation-wide study in China to quantify and compare the associations and population-attributable risks of common modifiable socioeconomic, psychosocial, lifestyle, and metabolic risk factors with the risks of CVD and all-cause mortality among adults of different age groups, with the group of adults aged 75 years and older as a significant population of interest. The study was conducted on 193,846 adults aged 40 years or older in China's Cardiometabolic Disease and Cancer Cohort Study from 2011 to 2016. The study selected the following risk factors for analysis: age, education, depression severity, cigarette smoking, alcohol intake, physical activity, healthy diet score,

sleep duration, BMI, waist circumference, diabetes, glucose profile, hypertension, systolic blood pressure, diastolic blood pressure and dyslipidaemia. The result of the study demonstrates that hypertension was the leading risk factor for CVD in the overall participants. Unhealthy diet, diabetes, low education, and inappropriate sleep duration contributed more than 10% each of the PAR% in at least one age group of < 75 years. In participants aged 75 or older, inappropriate sleep duration was the largest risk factor for CVD. Several specific risk factors showed significant interactions with age: inappropriate sleep duration was associated with greater hazards of CVD and mortality in older adults, while hypertension, diabetes, and chronic kidney disease were associated with greater hazards of CVD and mortality in younger adults. The Scottish Healthcare Improvement body released a guideline providing recommendations on estimating cardiovascular risk and interventions to reduce this risk in people with and without established CVD. The report suggests the three most important modifiable and causal cardiovascular risk factors: smoking, hypertension, and abnormal lipids known worldwide [27]. A global-scale study conducted by [28] aimed to report the derivation, validation, and illustration of the revised WHO cardiovascular disease risk prediction charts adapted to the circumstances of 21 global regions. They studied 376,177 individuals aged 40 to 80 without a known baseline history of CVD from 85 cohorts. During the ten-year follow-up, there were 19,333 incidents of cardiovascular events recorded. The study analysed the following risk factors: age, smoking status, systolic blood pressure, diabetes, total cholesterol, and BMI. The findings of the study show that in all regions evaluated, in the laboratory-based models, diabetes had a significant influence on the risk level across all age groups, particularly in those above 50 years. Moreover, smoking, hypertension, and high total cholesterol levels resulted in higher risk levels. Similarly, in the non-laboratory-based models, smoking and hypertension significantly increased the risk levels. Interestingly, the BMI did not significantly impact the risk level for individuals in the age groups below 50 years.

Table 1 shows the top 15 risk factors extracted from the analysed studies, along with the number of times the risk factor was studied, the number of times the risk factor was considered significant, and the percentage of the total against significant count. Although systolic blood pressure, diastolic blood pressure and hypertension are related, we opted to consider them as individual risk factors as they have also been reported separately in the discussed literature. We can observe that hypertension and age are among the most reported significant risk factors, with each study reporting them as essential for CVDs. Interestingly, the items measured in the lipid profile (i.e., total serum cholesterol, HDL-C, LDL-C, and triglycerides) have all been reported numerous times. Education, alcohol intake and diastolic blood pressure are among the risk factors that have been reported to be less significant compared to the times they have been studied, indicating these risk factors may not be as crucial when assessed individually. BMI and smoking are among the most studied risk factors based on the analysed studies, with a relatively high significance count indicating these risk factors may be critical for assessing CVDs. Ultimately, this study will consider these 15 risk factors for the feature selection task, ranked based on the order presented in the table.

Table 1
Highest 15 risk factors of cardiovascular diseases based on the domain knowledge.

Risk Factor	Total Count	Significant Count	Significance Ratio %
Hypertension	7	7	100
Age	6	6	100
Cholesterol	11	10	90.91
Diabetes	8	7	87.50
LDL-C	7	6	85.71
HDL-C	9	7	77.78
Smoking	11	8	72.73
Gender	3	2	66.67
BMI	13	8	61.54
Systolic Blood Pressure	9	5	55.56
Physical Activity	6	3	50.00
Triglyceride	6	2	33.33
Education	4	1	25.00
Alcohol Intake	6	1	16.67
Diastolic Blood Pressure	6	1	16.67

Figure 1 shows the distribution of the studies considered based on their respective continents, with most of the studies being based on Africa, followed by Asia, Europe, and North America. South America and Oceania have the least number of analyzed studies. The study by [28] consisted of data from 21 global regions, which led to the influx of the total number of studies per continent. Nevertheless, the studies considered from the domain have been conducted on a global scale to ensure the risk factors are not region-based.

Machine learning (ML) is one of the most rapidly evolving fields of artificial intelligence, capable of analysing massive amounts of data from various fields, such as medicine. As a result, researchers have proposed numerous approaches to predict the presence of CVDs based on various risk factors. Similarly, different datasets have been adopted for the training and validation of the proposed approaches, namely UCI Cleveland as the most frequently used dataset for training and benchmarking [4, 29]. However, the Cleveland dataset comprises 303 instances and does not contain many of the discussed risk factors associated with CVDs. This section will discuss the various research proposed by ML practitioners to

predict the presence of CVDs, including the used datasets, feature selection methods, evaluation metrics, and achieved results.

[30] used support vector machines (SVM) for heart disease diagnosis using two feature selection methods: Mean Fisher-based Feature Selection Algorithm and Accuracy-based Feature Selection Algorithm. The study utilized three datasets, UCI Cleveland, Hungarian, and Switzerland, with 709 data points. The feature subset consists of nine features: chest pain type, heart rate, exercise-induced angina, ST depression induced by exercise relative to rest, number of major vessels coloured by fluoroscopy, defect types, age, gender, and slope of the peak exercise ST segment. The model achieved an accuracy of 83.1%. [31] extended the dataset to utilize the four datasets from the UCI Heart Disease and aimed to use ML techniques to diagnose human heart diseases. The study uses the UCI Heart Disease, which consists of UCI Cleveland, Hungarian, Switzerland, and Long Beach datasets and contains 1025 samples. The study used four different machine learning algorithms: logistic regression (LR), k-nearest neighbours (KNN), SVM, and eXtreme gradient boosting (XGB). The best model was XGB, achieving an accuracy of 99.03%, precision of 99%, recall of 99%, and F1-score of 99%. Using the same dataset, [32] proposed a study for heart disease prediction using a combination of machine learning and deep learning. The study used nine features: age, gender, exercise-induced angina, ST depression induced by exercise relative to rest, number of major vessels coloured by fluoroscopy, resting blood pressure, chest pain type, fasting blood sugar, and heart rate. Seven machine learning models were trained and evaluated: LR, KNN, decision tree (DT), random forest (RF), SVM, XGB, and deep learning (DL). The best performance was achieved with DL with an accuracy of 94.2%, specificity of 83.1%, and sensitivity of 82.3%. Similarly, [33] presented a study using ensemble learning based on a hybrid DL model to predict the early stage of heart disease. The study uses CDC Heart Disease and UCI Heart Disease datasets, which have a total of 57373 and 1025 samples, respectively. The study uses recursive feature elimination (RFE) to select the most critical features: ten features for the CDC Heart Disease dataset (BMI, stroke, physical health, mental health, walking difficulty, age, race, diabetes, general health condition, and sleep duration) and six features for UCI Heart Disease dataset (age, chest pain type, heart rate, ST depression induced by exercise relative to rest, number of major vessels coloured by fluoroscopy, and defect types). The study uses nine different ML algorithms: SVM, LR, naive bayes (NB), DT, RF, KNN, CNN-LSTM, CNN-GRU and stacking ensemble. The best model was stacking ensemble with SVM, achieving an accuracy of 78.81%, precision of 78.1%, recall of 78.81% and F1-score of 78.81% for the CDC Heart Disease dataset. As for the UCI Heart Disease dataset, the best model was also the stacking ensemble achieving an accuracy of 97.17%, precision of 97.42%, recall of 97.17%, and F1-score of 97.15%.

[34] aimed to use ML algorithms with a feature selection method to predict heart diseases. The study uses two datasets: UCI Cleveland (330 samples) and StatLog (270 samples). The study used a feature selection technique, the chi-squared test (χ^2), to select the most important six features: heart rate, ST depression induced by exercise relative to rest, number of significant vessels coloured by fluoroscopy, type of chest pain, exercise-induced angina, and cholesterol. The study uses SVM as the learning algorithm. The results show an accuracy of 89.70%, specificity of 89.70%, sensitivity of 89.74% and F1-

score of 89.70%. [12] presented a hybrid cost-sensitive ensemble model for the prediction of heart disease. The authors used three datasets: StatLog, UCI Cleveland, and UCI Hungarian, with 270, 303, and 294 instances, respectively. The missing data was handled using a group-based mean imputation. The study used the relief feature selection technique to select six, eight, and five features from each dataset. The authors evaluated the performance of five ML algorithms: RF, SVM, ensemble learning machine (ELM), KNN, and cost-sensitive ensemble. The results showed that the cost-sensitive ensemble achieved an accuracy of 94.44%, precision of 92.59%, and recall of 92.15% on the StatLog dataset. The accuracy on the UCI Cleveland and Hungarian datasets was 93.83% and 89.47%, respectively. [35] utilized a different kind of feature selection technique and aimed at using ML algorithms to predict the risk of cardiovascular disease. The study uses a combination of two datasets: UCI Heart Disease and StatLog, consisting of 1190 samples. The study handles missing data by replacing it with the mean value and uses the Least Absolute Shrinkage and Selection Operator (LASSO) feature selection method to select the most important 11 features: age, gender, resting blood pressure, serum cholesterol, defect types, the slope of the peak exercise ST segment, fasting blood sugar, resting electrocardiographic results, ST depression induced by exercise relative to rest, maximum heart rate and chest pain type. The study uses 14 different machine learning algorithms: RF, NB, DT, KNN, SVM, LR, gradient boosting (GB), adaptive boosting (AdaBoost), DT bagging method, RF bagging method, KNN bagging method, AdaBoost boosting method and GB boosting method. The best model was RF, achieving an accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100%. The authors argue that some of the models in the related literature suffer from high-dimensional data. However, the definition of high-dimensional data was not provided. According to [36], high dimensional data is when p (number of features) $\geq N$ (number of samples), which is not the case for the studies referenced. The authors did not mention checking for duplicate observations in the StatLog, and UCI Cleveland datasets, as the datasets share similar observations. On the other hand, [37] made use of a Shapley Additive exPlanations (SHAP) feature importance technique to select the ten most important features: the slope of the peak exercise ST segment, chest pain type, gender, ST depression induced by exercise relative to rest, cholesterol, exercise-induced angina, age, fasting blood sugar, resting blood pressure, and heart rate. The study used 11 algorithms: KNN, SVM, LR, RF, ExtraTrees, GB, XGB, light GBM (LGBM), CatBoost, multilayer perceptron (MLP), and stacking ensemble. The best model was the stacking ensemble, achieving an accuracy of 89.86%, precision of 92.5%, recall of 90.24%, and F1-score of 91.36%. [38] proposed ML algorithms to predict human heart diseases. The study uses three datasets: UCI Cleveland (303 samples), Z-Alizadeh Sani (303 samples), and StatLog Heart (270 samples). The study uses nine different ML algorithms: AdaBoost, LR, ExtraTrees, NB, SVM, linear discriminant analysis (LDA), classification and regression trees (CART), RF, and XGB. The best model was SVM, achieving an accuracy of 96.72%, precision of 96%, recall of 96%, and F1-score of 96%. However, it is not clear whether the datasets have been combined or evaluated separately. Additionally, the StatLog dataset duplicates the UCI Cleveland dataset, which may introduce potential bias in the model. [39] adopted an additional dataset, UCI Hungarian, to improve and propose an intelligent detection system for coronary heart disease using a two-tier classifier ensemble with a total of 971 observations. The authors used correlation-based feature selection and seven machine learning algorithms: RF, GB, XGB, DT, CART, random tree, and a stacking ensemble to combine the outputs from

these models. The results show that stacking ensemble achieved the best performance with F1-score of 96.6%, accuracy of 98.13%, and AUC of 98.7% for the Z-Alizadeh Sani dataset, accuracy of 93.55%, F1-score of 91.67% and AUC of 93.42% for the StatLog dataset, 85.71% accuracy, 86.49% F1-score, and 85.86% AUC for the Cleveland dataset, and 91.18% accuracy, 90.91% F1-score, and 92.98% AUC for the Hungarian dataset.

Although the UCI Heart Disease dataset is among the most used benchmarking dataset for CVDs, the dataset does not contain many risk factors associated with CVDs. Therefore, this study adopted the Framingham Heart Study dataset as the primary dataset for the analysis and prediction of CVDs. [40] uses the Framingham dataset, which contains 7458 samples, to predict heart diseases. The study handles missing data by replacing them with a mean value. The study also used an over-sampling method to balance the data by creating six copies of the minority class. The most important features were selected using covariance, a measure of how much two random variables change together. The study uses four ML algorithms: artificial neural network (ANN), SVM, DT, and gaussian. The best model was the DT, achieving an accuracy of 92.59%, precision of 86.7%, recall of 100%, and F1-score of 94.08%. However, the adopted over-sampling method may have led to overfitting in the model as the technique created six copies of the minority class. On the other hand, [41] adopted the Random Over-sampling Examples technique (ROSE) to balance the data. The most important features selected using a multiple logistic regression method resulted in a subset of eight features: gender, age, cigarette smoking per day, prevalent stroke, prevalent hypertension, total serum cholesterol, systolic blood pressure, and glucose. The study used five ML algorithms: DT, RF, SVM, ANN, and LR. The best model was boosted DT, achieving an accuracy of 85%, recall of 5%, specificity of 99%, precision of 57%, and an AUC of 53%. While [42] used the Synthetic Minority Over-sampling Technique (SMOTE) to balance the data. The most important features were selected using the k highest score (SelectKBest) technique to remove irrelevant features and improve the model's performance. The study uses six different ML algorithms: LR, KNN, CART, NB, SVM, and RF. The best model was RF, achieving a ROC-AUC score of 94.63%. [7] adopted an additional dataset, UCI Cleveland, to propose an improved ensemble learning approach to predict heart disease risk. The study uses eight different ML algorithms: KNN, LR, LDA, SVM, CART, GB, RF, and randomized DT ensemble (RDTE). The best model was the RDTE, achieving an accuracy of 93%, precision of 96%, sensitivity of 91%, and F1-score of 93% for the Cleveland dataset. Likewise, RDTE achieved an accuracy of 91%, precision of 92%, sensitivity of 90%, and F1-score of 91% for the Framingham dataset. [43] adopted two additional datasets: Faisalabad Institute (299 samples) and South African Heart (462 samples), to identify the main risk factors of CVDs using ML algorithms. The study uses RF feature importance to select the four most important features: the Faisalabad Institute dataset (serum creatinine, ejection fraction, age, and platelets), Framingham dataset (systolic blood pressure, BMI, age, and total cholesterol), UCI Cleveland (chest pain type, heart rate, number of vessels coloured by fluoroscopy, and exercise relative to rest), and South African Heart dataset (tobacco, age, LDL, and adiposity). The study uses ten ML algorithms (AdaBoost, CatBoost, DT, GB, KNN, LGBM, LR, RF, SVM, and XGBRF) to predict CVD risk. The best results were obtained from DT with a 76.59% accuracy on the Faisalabad Institute dataset, LGBM with an accuracy of 83.10% on the Framingham dataset, LR with an accuracy of 82.42%

and 73.38% on the Cleveland and South African Heart datasets, respectively. [44] instead opted to adopt a Cardiovascular Disease Kaggle dataset along with the Framingham Heart Study and UCI Cleveland datasets to propose a boosting ensemble framework. The dataset includes 70000, 7192, and 303 records, respectively. The study adopted the SMOTE sampling technique to balance the imbalanced dataset. Moreover, the study used a feature selection method, SelectKBest, to select the top five features: systolic blood pressure, age, total cholesterol, cigarette smoking, and diastolic blood pressure. The framework was tested using four different ML models: DT, KNN, LR, and boosting ensemble. The best performance was achieved using the boosting ensemble with an accuracy of 99.1%, 98.0%, and 95.5% for the three datasets. Lastly, [45] adopted the four datasets under the UCI Heart Disease (920 samples) together with the Framingham dataset (7192 samples). The study used two techniques for handling missing data, Multiple Imputation by Chained Equations (MICE) and SMOTE. The study uses eight different machine learning algorithms: XGB, AdaBoost, GB, ExtraTrees, LGBM, SGDC, nuSVM, and stacking ensemble. The best model was the stacking ensemble, achieving an accuracy of 95.83%, specificity of 94.59%, precision of 97%, recall of 95% and F1-score of 96% for the first dataset. Similarly, for the second dataset, the best model was also the stacking ensemble, achieving accuracy of 90.24%, specificity of 87.51%, precision of 92%, recall of 88% and F1-score of 90%.

Ensemble learning is a general meta-approach to machine learning as it looks for the best predictive performance using a combination of methods to achieve the best accuracy. The use of single classifiers may not be able to provide the best outcomes. Hence, the combination of the algorithms will combine all the strengths of the models and bring out improved accuracy [18]. Numerous other research has been conducted focusing on ensemble learning techniques. However, the dataset adopted is rarely the Framingham dataset. [4] conducted a systematic mapping study for ensemble classification methods in cardiovascular disease. The study analysed 351 papers ranging from 2000 to 2019. DTs, ANNs, SVMs, KNNs, LR and NB, were among the frequently used single classifiers to create ensembles. Majority voting was the dominant ensemble technique, with over 300 papers that adopted it. Methods such as boosting, bagging, and stacking are in the minority. Most of the existing research on the Framingham Heart Study dataset focused on improving the accuracy of single classifiers. However, the missing values and the imbalance in the target attribute have not been paid much attention to, resulting in proposed approaches with misleading high accuracy scores. Moreover, in the current literature, the stacking ensemble method is among the rarely adopted ensemble techniques for predicting CVDs. Hence, this study aims to use the stacking ensemble method to produce an optimal predictive model for the prediction of CVDs. The reason for selecting the stacking ensemble method is that stacking is the most effective when using low-correlated base models. As a result, the base models' diversity helps achieve more robust and better predictive models. Table 2 summarises the literature review, outlining the used datasets, sampling methods to handle data imbalance, feature selection techniques, best-performing classifiers, and their respective accuracies. We can observe that k highest score (SelectKBest) and the LASSO methods are among the frequently adopted techniques for feature selection. Hence, this study will adopt these two techniques to benchmark against the proposed domain knowledge-based feature selection. Please refer to Table S1 for a comprehensive summary of the literature review.

Table 2
Summary of the literature review.

Author(s)	Dataset	Sampling	Feature Selection	Classifier	Accuracy
[30]	Cleveland	-	Mean Fisher & Accuracy-based	SVM	83.10%
	Hungarian				83.90%
	Switzerland				92.10%
	UCI Heart Disease*				84.20%
[31]	UCI Heart Disease	-	-	XGBoost	99.03%
[32]	Cleveland	-	LASSO	Deep Learning	94.20%
[33]	UCI Heart Disease	-	Recursive Feature Elimination	Stacking Ensemble	97.17%
	CDC Heart Disease	-			78.81%
[34]	Cleveland	-	Chi-Squared Test	SVM	89.47%
	StatLog				89.70%
[12]	Cleveland	-	Relief	Cost-Sensitive Ensemble	-
	Hungarian				
	StatLog				
[35]	UCI Heart Disease & StatLog	-	LASSO	Random Forest	100.0%
[37]	UCI Heart Disease & StatLog	-	Gradient Boosting Decision Tree + SHAP	Stacking Ensemble	89.86%
[38]	Cleveland, StatLog & Z-Alizadeh Sani	-	-	SVM	96.72%
[39]	Z-Alizadeh Sani	-	Particle Swarm Optimization	Stacking Ensemble	98.13%
	StatLog				93.55%
	Cleveland				85.71%
	Hungarian				91.18%
[40]	Framingham	Over	Covariance	Decision Tree	92.59%
[41]	Framingham	ROSE	Multiple Logistic Regression	Boosted Decision Tree	85.00%
[42]	Framingham	SMOTE	SelectKBest	Random Forest	94.63%
[7]	Framingham	-	-	Randomized	91.00%

Author(s)	Dataset	Sampling	Feature Selection	Classifier	Accuracy
	Cleveland			Decision Tree	93.00%
[43]	Faisalabad Institute	-	Random Forest	Decision Tree	76.59%
	Framingham			LGBM	83.10%
	Cleveland			SVM	81.16%
	South African Heart			Decision Tree	72.51%
[44]	Framingham	SMOTE	SelectKBest	Boosting Ensemble	99.10%
	Cleveland				98.00%
	CVD Kaggle				95.50%
[45]	UCI Heart Disease	SMOTE	-	Stacking Ensemble	95.83%
	Framingham				90.24%

* Excluding the Long Beach dataset.

Figure 2 shows the sum of features utilized for training the ML algorithms. Since many studies utilise the UCI Heart Disease dataset, many of the features are specific to that dataset. Nevertheless, age, type of chest pain, ST depression induced by exercise relative to rest, number of major vessels, and gender are among the most utilized features for CVD prediction. Comparing the acquired results to the domain knowledge in Table 1, we can observe an overlap between features: age, gender, systolic blood pressure, total cholesterol, smoking, and BMI.

Figure 3 presents the total number of algorithms used in the ML studies. SVM, LR, KNN and RF are the most consistently utilized algorithms. Tree-based classifiers (i.e., RF, DT, GB, XGB and AdaBoost) have also been utilized frequently. The reason for that could be due to the high level of interpretability of decision trees [9].

The UCI Heart Disease, Framingham Heart Study, and StatLog Heart datasets are among the most utilized for predicting CVDs, as shown in Fig. 4. The UCI Cleveland has been widely known as a benchmark dataset for the task of CVD prediction [29, 32].

3. Methodology

Figure 5 shows the proposed research methodology flowchart for this study.

3.1. Description of Dataset

This study utilizes three publicly available datasets: The Framingham Heart Study, UCI Heart Disease, and United Arab Emirates (UAE) Retrospective Cohort Study. The Framingham dataset will be considered the primary dataset for conducting the experiments. In contrast, the latter datasets will be regarded as validation datasets to assess the generalization capabilities of the methods proposed.

3.1.1. Framingham Heart Study

The Framingham Heart Study dataset is used in this research to train machine learning algorithms. The dataset was collected from 1948 to 1972, where each participant was examined thrice for the outcome of the following events: angina, myocardial infarction, thrombotic infarction or cerebral hemorrhage, or death. The Framingham dataset was selected as the primary dataset as it offers a comprehensive set of features related to patient details, including demographic, behavioural, medical history, and current medical conditions. The dataset contains 11,627 rows and 39 variables. However, not all the variables are relevant to our study. Hence, we will manually select the most relevant attributes from the dataset for predicting the presence of CVDs. Table 3 shows 17 out of 39 attributes considered relevant for our study. We will perform feature selection on these attributes to maintain the performance of the models while reducing the computational time and model complexity.

Table 3
Description of selected features of the Framingham Heart Study dataset.

Feature	Description	Data Type	Values or Range
gender	Sex of participants	Nominal	1 = Men 2 = Women
age	Age at exam (years)	Discrete	32–81
education	Highest education attained	Nominal	1 = 0–11 years 2 = High School Diploma, GED 3 = College, Vocational School 4 = BS, BA degree or more
smoker	Participant cigarette smoking at exam	Nominal	0 = Not a smoker 1 = Smoker
cigpday	Number of cigarettes smoked per day	Discrete	0–90
sysbp	Systolic blood pressure (mmHg)	Continuous	83.5–295
diabp	Diastolic blood pressure (mmHg)	Continuous	30–150
hyperten	Hypertensive; defined by either Systolic is ≥ 140 mmHg or Diastolic ≥ 90 mmHg	Nominal	0 = Free of disease 1 = Prevalent disease
bpmeds	Use of anti-hypertensive medication at exam	Nominal	0 = Not currently used 1 = Current use
hdlc	High-density lipoprotein cholesterol (mg/dl)	Continuous	10–189
ldlc	Low-density lipoprotein cholesterol (mg/dl)	Continuous	20–565
totchol	Serum total cholesterol (mg/dl)	Continuous	107–696
bmi	Body Mass Index (kg/m^2)	Continuous	14.43–56.8
glucose	Casual serum glucose (mg/dl)	Continuous	39–478

Feature	Description	Data Type	Values or Range
diabetic	Presence of diabetes	Nominal	0 = Not a diabetic 1 = Diabetic
heartrate	Heart rate in beats/min	Continuous	37–220
cvd	Presence of cardiovascular disease	Nominal	0 = Free of disease 1 = Prevalent disease

3.1.2. UCI Heart Disease

The UCI Heart Disease is a publicly available dataset from the UCI Machine Learning Repository dating from 1981 to 1984 for the angiography-based diagnosis of coronary heart disease based on the narrowing of at least one of the coronary arteries of more than 50%. The dataset comprises data from four institutions: 1) Hungarian Institute of Cardiology; 2) University Hospital, Zurich, Switzerland; 3) University Hospital, Basel, Switzerland; and 4) V.A. Medical Centre, Long Beach, and Cleveland Clinic Foundation [46]. Although the full dataset consists of 1025 instances with 76 features, only 14 features were exclusively utilized by the related literature. The independent features include the results of the non-invasive diagnostic tests and other relevant patient information, as shown in Table 4. This study utilizes the UCI Heart Disease dataset as a benchmark for the proposed methods as it is among the well-established benchmark datasets in the related literature [29].

Table 4
Description of selected features of the UCI Heart Disease dataset.

Feature	Description	Data Type	Values or Range
age	Age at exam (years)	Discrete	29–77
sex	Gender of participants	Nominal	0 = Female 1 = Male
cp	Type of chest pain	Nominal	1 = Typical angina 2 = Atypical angina 3 = Non-anginal pain 4 = Asymptomatic
trestbps	Resting blood pressure (mmHg)	Continuous	94–200
chol	Serum cholesterol (mg/dl)	Continuous	126–564
fbs	Fasting blood sugar > 120 mg/dl	Nominal	0 = False 1 = True
restecg	Resting electrocardiographic results	Nominal	0 = Normal 1 = ST-T wave abnormality 2 = LV hypertrophy
thalach	Maximum achieved heart rate	Continuous	71–202
exang	Exercise-induced angina	Nominal	0 = False 1 = True
oldpeak	ST depression induced by exercise relative to rest	Continuous	0–62
slope	Slope of the peak exercise ST segment	Ordinal	1 = Up-sloping 2 = Flat 3 = Downsloping
ca	Number of major vessels coloured by fluoroscopy	Nominal	0–3
thal	Defect types	Nominal	3 = Normal 6 = Fixed defect 7 = Reversible defect

Feature	Description	Data Type	Values or Range
num	Diagnosis of heart disease	Nominal	0 = Free of disease 1 = Prevalent disease

3.1.3. UAE Retrospective Cohort Study

Although the Framingham Heart Study and UCI Heart Diseases were collected approximately four to five decades ago, some may consider these datasets old. However, according to [47], there is almost no change in the rank of modifiable cardiovascular disease risk factors. Nevertheless, this study utilizes a more recent publicly available cardiovascular dataset to validate the proposed methods on more recent data. The UAE dataset is a 9-year retrospective cohort study from 2008 to 2018 for the incidence of cardiovascular disease and its associated risk factors in the United Arab Emirates. The dataset consists of 977 observations and 19 features [19]. The target class is highly imbalanced, with 880 samples without CVDs and only 97 with CVDs. As a result, the UAE dataset is appropriate for validating the proposed methods. This study selected a subset of 17 features to validate the proposed methods, as shown in Table 5.

Table 5
Description of selected features of the UAE Retrospective Cohort Study dataset.

Feature	Description	Data Type	Values or Range
age	Age at baseline exam (years)	Discrete	19–81
gender	Sex of the participant	Nominal	0 = Female 1 = Male
dmeds	Use of diabetes medication at exam	Nominal	0 = Not currently used 1 = Current use
smoker	Participant cigarette smoking at exam	Nominal	0 = Not a smoker 1 = Smoker
bpmeds	Use of anti-hypertensive medication at exam	Nominal	0 = Not currently used 1 = Current use
dysmeds	Use of dyslipidaemia medication at exam	Nominal	0 = Not currently used 1 = Current use
totchol	Serum Total Cholesterol (mmol/L)	Continuous	1.7–11.3
hdlc	High density lipoprotein cholesterol (mmol/L)	Continuous	0.33-6.0
ldlc	Low density lipoprotein cholesterol (mmol/L)	Continuous	0.24–6.89
triglycerides	Triglycerides (mmol/L)	Continuous	0.18–11.56
cholhdlc	Total Cholesterol/HDL-C ratio	Continuous	0.6-10.19
hba1c	Glycosylated haemoglobin (%)	Continuous	3.5–18.1
egfr	Estimated glomerular filtration rate (mL/min.1.73 m ²)	Continuous	3.99-261.23
sysbp	Systolic blood pressure (mmHg)	Continuous	61.40–202.00
diabp	Diastolic blood pressure (mmHg)	Continuous	43.00-148.00
bmi	Body Mass Index (kg/m ²)	Continuous	12.07–60.11
cvd	Presence of cardiovascular disease	Nominal	0 = Free of disease 1 = Prevalent disease

3.2. Exploratory Data Analysis

This study performs exploratory data analysis (EDA) to understand the data in depth, learn the different characteristics, and uncover hidden patterns in the Framingham dataset via visualization. Figure 6 shows Pearson's correlation coefficient between the continuous variables in the dataset. We can observe a high positive correlation between systolic and diastolic blood pressure variables and between the total serum cholesterol and low-density lipoprotein cholesterol (HDL-C) variables. As a result, we may consider dropping highly correlated features as they do not bring additional information but increase the complexity of the developed models [48].

Figure 7 shows a dendrogram computed using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm to show the hierarchical clustering between variables. We can observe a high similarity between total cholesterol & LDL-C and smoker and the number of cigarettes smoked chunks. Surprisingly, the HDL-C is more similar to the gender chunk than the cholesterol chunks. The age chunk is somewhat related to CVD and death chunks. The systolic and diastolic blood pressure, hypertensive, and anti-hypertensive medication chunks are in a similar group as expected. Similarly, the glucose and diabetic chunks are similar to some degree.

Figure 8 shows a series of charts of gender (i.e., male and female), age groups (i.e., adults, middle-aged adults, and seniors), and diabetes, grouped by the presence or absence of CVD. In the following exploratory analysis, we found that the proportion of males with CVD was higher than females, even though the females have a higher frequency. Accordingly, it could be the case that males are more susceptible to CVDs than females. The age group with the highest ratio of presence/absence of CVD is the seniors (i.e., 65 and above), with almost half the seniors having CVDs. Although adults (i.e., 32 to 48) are as many as seniors, they have the least frequency of CVD. Hence, age can be considered a significant risk factor for CVDs, higher the age, the more susceptible to CVDs. Although the count of positive diabetes is relatively low, it is remarkable that more than half of the diabetic patients have CVD, meaning that diabetes might be an essential risk factor for CVD. Similarly, the glucose serum levels in diabetic patients are significantly higher compared to non-diabetic patients, as shown in Fig. 9. Smoking, glucose, total serum cholesterol, and BMI did not show a noticeable difference between patients with and without CVD. LDL-C and HDL-C levels among age groups had a notable difference between patients with and without CVDs, as shown in Fig. 10. Adult patients with CVD had significantly higher LDL-C levels and lower HDL-C levels compared to patients without CVDs.

3.3. Data Preprocessing

This study performs data preprocessing to improve the quality of data. Only the training data will be imputed for missing values, as the testing data needs to be unchanged to avoid bias in the results.

3.3.1. Missing Values and Outliers

Missing values in a dataset can negatively affect the performance of ML algorithms. Hence, missing values need to be handled with minimal data loss not to lose valuable information [44]. Table 6 shows the total number of missing values in the Framingham dataset. In this study, we primarily used group-

based imputation using the mean value. Group-based imputation uses a related attribute to the attribute in question to find the mean for each class and uses the corresponding class value as the imputation value.

Table 6
Total number of missing values by the features.

Attribute	Count of Missing Values
glucose	562
bpmeds	446
ldlc	237
hdlc	236
totchol	214
education	82
bmi	17
cigpday	14
heartrate	4

Outliers refer to values that fall outside the expected ranges or are unlike the other observations. However, not all identified outliers should be removed. Hence, we used box and scatter plots to examine possible outliers and reviewed whether the identified observations count as outliers based on existing literature.

3.3.2. Feature Selection

Feature selection is an essential component of ML, particularly with datasets with many variables. Feature selection helps improve the efficiency and accuracy of the machine learning algorithm where variables most relevant to the study are selected [44]. Initially, we chose 17 features from the Framingham dataset. However, not all variables may be relevant for our research and increase the computational time and complexity. In this study, we compare the performance of three feature selection methods: proposed domain knowledge-based, feature importance using the SelectKBest method, and LASSO regression. This study proposes feature selection based on the risk factors deemed as significant by the domain knowledge and compares the performance against the two mainly used feature selection methods in the related literature. The top 15 risk factors, as shown in Table 1, will be used as a reference when manually selecting features from the three datasets adopted in this study. Accordingly, only the risk factors present in each dataset will be used to create a subset of features for model development.

3.3.3. Data Imbalance

The target attribute in the Framingham dataset is imbalanced, with the positive class (i.e., has CVD) as a minority, as shown in the pie chart in Fig. 11a. The imbalance in the target class poses a challenge for accurate prediction as most ML algorithms are designed around the assumption that there is an equal distribution between the target class. As a result, the developed models have poor predicting performance, especially for the minority class. Thus, this study will evaluate the performance of several data sampling techniques applied only to the training data for handling the imbalance [21]. The pie chart in Fig. 11b shows the desired target class after sampling. This study will employ the following sampling techniques: under-sampling (random under-sampling, Tomek Links, edited nearest neighbour (ENN), repeated ENN), over-sampling (random over-sampling, SMOTE), and a combination of under- and over-sampling (SMOTE Tomek Links, SMOTE ENN and polynom-fit-SMOTE with Tomek Links).

Random under-sampling involves randomly selecting a certain number of instances from the majority class and removing them from the dataset until the desired class balance is achieved. Tomek Links is one of the modifications from the condensed nearest neighbours under-sampling method, where instead of randomly selecting examples, it uses more restrictive conditioning. A pair of observations from both classes, Tomek Links, can be used to locate all cross-class nearest neighbours where the majority examples are removed if they are found to be minority data. ENN is an under-sampling method to eliminate data points that do not align with the values of their neighbouring points. This can lead to removing points that are located close to the boundary between different classes and can also effectively eliminate noisy or incorrectly labelled training data. The repeated ENN approach involves repeatedly applying the ENN algorithm until no more data points can be removed. The Synthetic Minority Over-Sampling Technique (SMOTE) is used to increase the number of examples in a minority class. It creates new synthetic data points along the line segments connecting a given positive example with its k-nearest neighbours in the minority class. Random oversampling involves making copies of existing minority class examples to increase the number of instances in the minority class. SMOTE Tomek Links technique combines two different methods for handling imbalanced datasets. The SMOTE algorithm is used to create new synthetic instances of the minority class, while the Tomek Links algorithm is used to remove some examples of the majority class. The overall effect is to increase the number of minority class examples, while decreasing the number of majority class examples. SMOTE ENN method is a combination of two techniques, it uses SMOTE to increase the number of minority class samples, and then it applies the ENN method to clean the newly generated instances. It is a hybrid approach that combines over-sampling and under-sampling techniques.

Polynom-fit-SMOTE with Tomek Links presents a potent strategy for navigating class imbalances, merging both over- and under-sampling approaches. Polynom-fit-SMOTE, a variant of the SMOTE algorithm, synthesizes minority instances through polynomial fitting, potentially offering a more accurate portrayal of the minority class distribution [49]. This subsequently improves the predictive capability of the model [50]. Following the application of Polynom-fit-SMOTE, the Tomek Links technique is utilized to identify and eliminate instances from the majority class that are nearest neighbors to instances in the minority class, thus refining the decision boundary between the two classes. This combined approach, as previously tested in [51], enhances the model's performance by providing a more precise representation of

the minority class and minimizing the risk of overfitting to the majority class, positioning it as a promising solution for effectively addressing imbalanced datasets.

3.4. Classification Algorithms

Classification is a supervised learning procedure that predicts the outcome from existing data. In this study, we propose the prediction of CVDs using single classifiers and improve the performance of the single classifiers using an ensemble of classifiers, to find an optimal predictive model using the stacking technique. The working of the single classifiers is explained in the following sections.

3.4.1. Logistic Regression (LR)

LR is a statistical method used to predict the outcome of a categorical dependent variable based on a set of independent variables. It is a specific type of regression analysis, and the dependent variable in this type of analysis is always binary (i.e., it has two possible values). The primary purpose of logistic regression is to make predictions and to calculate the likelihood of a particular outcome occurring.

3.4.2. Naïve Bayes (NB)

NB is a probabilistic algorithm based on Bayes' theorem, which assumes the features in the data are independent of each other. As a result, changes in a feature do not affect another feature. NB starts by determining the frequency of different instances based on class labels. Then it calculates the probability of each one. The class label with the highest probability value is then chosen as the prediction. Compared to more complex algorithms, NB classifiers can be exceptionally fast, have few tuneable hyperparameters and are known to outperform complex algorithms. Consequently, it is often suitable for high-dimensional datasets.

3.4.3. Multilayer Perceptron (MLP)

MLP, a type of ANN modelled loosely after the human brain, is designed to recognize patterns. An MLP is a class of neural networks made up of at least three nodes that seek to learn the relationship between linear and non-linear data. A conventional MLP comprises one input & output layer and one or more hidden layers. In each MLP layer, neurons are linked and pass data through the network after computing values based on the inputs.

3.4.4. Decision Tree (DT)

DT is a supervised machine learning algorithm, where each path from the root represents a succession of data splitting based on a decision made (i.e., yes or no) at the leaf node by asking a series of questions, with the final nodes being the predicted value. The difficulty of this approach is identifying the root node. Selecting the root node with care decreases the algorithm's computational complexity and increases its efficiency. Furthermore, the decision tree is simple to comprehend, and the reader can deduce certain information from it.

3.4.5. Support Vector Machine (SVM)

SVM is a binary classification algorithm that classifies two classes by building a hyperplane or set of hyperplanes in two or more dimensions that separate those classes. The closest samples to the hyperplane are the support vectors, which affect the position and orientation of the hyperplane. The distance between a hyperplane and its support vectors is referred to as the functional margin. SVM can classify better if the functional margin is large. SVM performs well on a dataset with many features compared to the number of data points. Furthermore, SVM is robust to noisy data. Unfortunately, SVM does not perform well with large datasets, as training is time-consuming.

3.4.6. K-Nearest Neighbors (KNN)

KNN is a lazy supervised algorithm used for classification and regression. KNN is straightforward to understand and well-suited for datasets without prior knowledge of available data distribution. In other words, KNN is a classification algorithm based on the similarity of new and existing data [31]. Moreover, it is reasonably robust to noisy training data. However, KNN does not perform well with high-dimensional or large amounts of data that contain noisy and missing data. Furthermore, it is prone to overfitting [52].

3.5. Ensemble Algorithms

The ensemble is a technique used to improve the performance of single classifiers by combining predictions from two or more other models. There are different types of ensemble techniques. In this study, we will use different ensemble techniques to improve the accuracy of single classifiers. The adopted methods are boosting, bagging, majority voting and stacking. The objective of this study is to utilize stacking to produce an optimal predictive model. However, we also considered other ensemble techniques and single classifiers to compare against the proposed stacked ensemble classifier. In essence, weak learners do not perform well because they have a high bias or high variance. The concept of ensemble methods is to reduce the bias or variance of such weak learners by combining them to create a strong learner, achieving higher performance.

3.5.1. Boosting

giving importance to those observations that the preceding models did not correctly predict. In other words, the subsequent models focus on the misclassified observations by assigning higher weightage to obtain a strong learner with a lower bias. Being mainly focused on reducing bias, the base models often considered for boosting are models with a low variance but high bias. Due to the sequential nature of boosting, the algorithm cannot be trained in parallel. In this study, we adopted gradient boosting (GB), eXtreme gradient boosting (XGB), and adaptive boosting (AdaBoost) as the boosting ensemble classifiers. GB uses a boosting technique to repeatedly resample the data to form a weighted average data set. The boosting technique involves creating multiple decision trees that form a single predictive model. As a result, the successive samples of the data are adjusted to adapt to the inaccuracies found in the previous samples. On the other hand, XGB is an optimised GB algorithm designed to be highly efficient, flexible, and portable. XGB is a more regularised form of GB and uses advanced regularisation (i.e., L1 & L2), which improves model generalisation capabilities. AdaBoost is an algorithm that utilizes a step-by-step process to generate a more robust classifier by combining multiple basic classifiers. The

algorithm calls on a base learner during each iteration and assigns it a weight coefficient. The final prediction is based on the weighted vote of all the base classifiers. The weight assigned to each classifier is proportional to the accuracy of its predictions. The AdaBoost algorithm modifies the data distribution in each iteration by focusing on samples that were inaccurately classified in the previous iteration. These modified data and weights are passed on to the next base learner, and the final classifier is an amalgamation of all classifiers obtained during each iteration. We expect XGB to be the best-performing model in this study as it can be trained quickly, even on large datasets, and offer high predictive performance. Subsequently, XGB may be the meta-model used in the Stacking technique [27].

3.5.2. Bagging

Bagging is an ensemble that creates multiple bootstrap samples where the individual samples act as another independent dataset drawn from the true distribution. Then, a weak learner is fit to these bootstrapped samples and, in the end, aggregates the outputs from each of the models by voting or averaging to obtain an ensemble model with decreased variance. Being mainly focused on reducing variance, the weak learners considered for bagging are models with a high variance but low bias. We adopted Random Forest (RF) as the bagging ensemble classifier in this study. RF fits multiple trees on bootstrap samples and sampled features to build trees. Then, the votes of the trees are aggregated, and the class with the most votes is the predicted result. We anticipate RF to be among the base models as it is well-known to reduce overfitting and provide better predictive performance than other ML algorithms [12].

3.5.3. Majority Voting

The majority voting classifier is a meta classifier that combines any classifier through majority voting. The final class label would be the class label that most of the classifiers had predicted. We will explore two types of voting: hard (i.e., summing of votes) and soft (i.e., largest sum probability). We do not expect majority voting to perform well as it uses other models to calculate the majority vote. As a result, majority voting will not perform well unless the base learners have good predictive performance.

3.5.4. Proposed Stacking Ensemble

Stacking involves training a “meta-learner” through multiple base models. In other words, stacking utilises predictions from various models to produce an even better predictive model. The general framework of a stacked ensemble comprises two or more base models (i.e., Level-0 models) to fit the training data and predict the test data and a meta-model (i.e., Level-1 model) that fits the base models’ predictions and learns the best way to combine predictions. There is no established method for selecting the ideal meta-model. Hence, we will run a combination of the highest-performing models to see which provides the best predictive performance. As for the base models, we will try a combination of different models, such as GB, LR, and RF, to make the final model as robust as possible. The stacked ensemble aims to achieve the highest F1-score metric, as it assures us that our model is optimal in predicting the presence of CVDs, a minority class in our data.

3.6. Evaluation Metrics

In this study, we primarily adopt an 80:20 train-test split for our dataset, with 80% of the data serving as the training set and the remaining 20% used for testing. However, it's important to note that different data splitting techniques might be used on certain occasions. In such cases, the specific technique utilized will be mentioned explicitly; if not, it can be safely assumed that we are adhering to our default 80:20 split.

To gauge the performance of our models, we use a range of metrics associated with the confusion matrix - an essential tool that encapsulates the outcomes of our classifiers in a tabular format. This matrix includes key measures such as true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), which are all calculated based on the testing set.

Accuracy is one of the primary metrics applied to assess the performance of a classifier. It is computed as the percentage of the samples classified correctly to the total observations and can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

1

However, when the accuracy metric is applied to imbalanced data, such as in the case of this study, the obtained results may be misleading, as the metric assumes an equal distribution among the target classes. Therefore, the balanced accuracy metric is a more appropriate choice for imbalanced data, as it calculates the average accuracy of each class separately. The formula of the balanced accuracy metric is as follows:

$$Balancedaccuracy = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

2

Precision is the ratio of correctly predicted positive (CVD) observations to the total predicted positive observations, and the calculation is as follows:

$$Precision = \frac{TP}{TP + FP}$$

3

Recall is the ratio of correctly predicted positive observations to all the observations in the positive class and calculated as:

$$Recall = \frac{TP}{TP + FN}$$

4

F1-score is the weighted average of the precision and recall. F1-score is computed as follows:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5

The accuracy metric assumes the target attribute has an equal distribution of classes. However, prior to the data sampling, our dataset has an imbalance in the target class. Hence, we introduce precision, recall and F1-score to help in evaluating the minority class more accurately. F1-score is the deciding metric in this study, as it has been designed to work well on imbalanced data [53].

4. Results and Discussion

This study developed various single and ensemble classifiers to demonstrate the performance of the proposed domain knowledge-based feature selection and stacking ensemble. Firstly, we will discuss the results of the conducted feature selection methods: domain knowledge-based, k highest score (SelectKBest), and LASSO. This study selected the top ten subsets of features for a model performance comparison from the Framingham dataset based on the three feature selection methods, as shown in Table 7. Based on the results, we can observe that five variables are shared across the feature selection methods: hypertension, age, LDL-C, HDL-C, and systolic blood pressure. To note, hyperparameter tuning was performed using the grid search method on each classifier to maximise the performance of each developed model. Moreover, the ensemble classifiers were tested with different combinations of base learners to maximise the model's performance metrics. Also, the features mentioned in the feature selection have been used for model development. Lastly, the random state set during the train-test split has been ensured to achieve reproducible results across multiple executions.

Table 7

The selected top ten subsets of features based on three feature selection methods on the Framingham dataset.

Rank	Domain Knowledge	SelectKBest	LASSO
1	Hypertension	Systolic Blood Pressure	Gender
2	Age	Glucose	Age
3	Total Cholesterol	HDL-C	Systolic Blood Pressure
4	Diabetes	Age	Smoking
5	LDL-C	LDL-C	Diabetes
6	HDL-C	Diabetes	Hypertensive Meds
7	Smoking	Hypertensive Meds	Education
8	Gender	Diastolic Blood Pressure	HDL-C
9	BMI	Hypertension	LDL-C
10	Systolic Blood Pressure	Heart Rate	Hypertension

Table 8 shows the results of the trained ML models based on domain knowledge-based feature selection (i.e., the top ten risk factors from the domain knowledge available in the dataset). The proposed stacking ensemble has achieved the highest accuracy (80.25%), precision (78.04%), recall (80.25%), and F1-score (76.46%). However, all models' balanced accuracy is extremely poor, with the highest, Naïve Bayes, achieving a mere 63.31%. Consequently, the models trained are poor at predicting the minority (i.e., positive) class in the dataset due to the imbalance in the target class where the majority class heavily influences the model, rendering the model's usefulness [54]. To our surprise, there is no significant difference in performance between the single and ensemble classifiers.

Table 8
The results of the classifiers based on domain knowledge-based feature selection.

Classifier	Accuracy	Balanced Accuracy	Precision	Recall	F1-score
Single Classifiers					
SVM	78.56	53.76	75.86	78.56	71.60
K-Nearest Neighbour	79.02	56.25	76.12	79.02	73.62
Decision Tree	77.03	59.85	74.02	77.02	74.80
Logistic Regression	79.94	59.53	77.53	79.94	75.96
Multilayer Perceptron	79.48	59.96	76.65	79.48	76.00
Naive Bayes	77.49	63.31	75.68	77.49	76.33
Ensemble Classifiers					
Random Forest	78.41	57.08	74.66	78.41	73.98
Soft Voting	79.02	58.20	75.83	79.02	74.86
Hard Voting	79.33	59.38	76.39	79.33	75.63
Gradient Boosting	79.02	59.67	75.94	79.01	75.65
AdaBoost	79.02	60.15	76.01	79.02	75.88
XGB	77.80	62.30	75.48	77.80	76.16
Stacking Ensemble	80.25	60.21	78.04	80.25	76.46

Table 9 presents the results of the trained models based on the k -highest score (SelectKBest) feature selection technique. The proposed stacking ensemble has achieved the highest accuracy (78.25%), precision (75.46%), and recall (78.25%). On the other hand, logistic regression (LR) achieved the highest balanced accuracy (62.43%) and F1 score (75.74%). However, there is no significant difference in

performance between the classifiers. Moreover, the trained models have poor predictive power on the minority class, which can be observed in the balanced accuracy results.

Table 9
The results of the classifiers based on the k highest score feature selection.

Classifier	Accuracy	Balanced Accuracy	Precision	Recall	F1-score
Single Classifiers					
Multilayer Perceptron	74.27	53.80	68.56	74.27	69.63
SVM	77.48	55.24	74.51	77.48	71.32
K-Nearest Neighbour	76.87	56.40	72.89	76.87	72.05
Decision Tree	74.27	59.35	71.71	74.27	72.59
Naive Bayes	76.26	61.10	73.65	76.26	74.32
Logistic Regression	77.94	62.43	75.45	77.94	75.74
Ensemble Classifiers					
Gradient Boosting	75.65	56.26	71.20	75.65	71.61
AdaBoost	75.65	56.26	71.20	75.65	71.61
Hard Voting	76.56	57.31	72.64	76.56	72.54
Soft Voting	76.87	58.40	73.31	76.87	73.29
XGB	75.34	60.06	72.62	75.34	73.42
Random Forest	77.48	58.57	74.17	77.48	73.62
Stacking Ensemble	78.25	59.74	75.46	78.25	74.58

Table 10 presents the performance results for the classifiers trained on LASSO feature selection. Similar to the other feature selection methods, the stacking ensemble has achieved the highest accuracy (77.02%), precision (74.84%), and recall (77.02), while LR achieved the highest balanced accuracy (62.96%) and F1-score (73.98%). Nevertheless, the performance difference across the models is not significant.

Table 10
The results of the classifiers based on LASSO feature selection.

Classifier	Accuracy	Balanced Accuracy	Precision	Recall	F1-score
Single Classifiers					
SVM	75.34	54.43	73.93	75.34	67.88
K-Nearest Neighbour	75.49	55.10	73.75	75.49	68.56
Decision Tree	74.88	60.15	71.94	74.88	72.11
Logistic Regression	75.95	62.96	73.70	75.95	73.98
Multilayer Perceptron	75.34	58.39	72.17	75.34	71.22
Naive Bayes	72.74	60.78	70.70	72.74	71.40
Ensemble Classifiers					
Random Forest	75.49	59.25	72.48	75.49	71.82
Soft Voting	76.56	61.30	74.13	76.56	73.43
Hard Voting	75.80	58.14	73.01	75.80	71.17
Gradient Boosting	75.65	59.54	72.72	75.65	72.06
AdaBoost	74.42	59.09	71.17	74.42	71.34
XGB	74.11	61.71	71.89	74.11	72.50
Stacking Ensemble	77.02	61.42	74.84	77.02	73.70

Based on the pre-sampling results for the different feature selection techniques, the domain knowledge-based feature selection using ten features has achieved the highest performance metrics across the board. Although the results for the domain-based feature selection are not significant, the performance achieved is comparable to some of the frequently used feature selection techniques in the related literature. Hence, it is quite evident that a domain knowledge-based feature selection can be used during the feature selection process for CVDs to enhance the credibility of the models, resulting in more interpretable and trustable models.

The balanced accuracy is extremely poor in all the achieved results by the three feature selection methods, indicating that the models are poor at predicting the minority (i.e., positive) class in the dataset. Hence, we will evaluate a series of sampling techniques to balance the Framingham dataset and deduce which technique provides the best performance metrics. Table 11 presents the performance results and total observations of various data sampling techniques applied to the proposed stacking ensemble. This

study evaluated a series of under-sampling methods (edited nearest neighbours, repeated edited nearest neighbours, instance hardness threshold, random under-sampling, and Tomek Links), over-sampling methods (random over-sampling and SMOTE), and a combination of under- and over-sampling (SMOTE edited nearest neighbour and SMOTE Tomek Links). Regarding under-sampling methods, the general concept is to keep the minority class unchanged while decreasing the majority class. The most straightforward under-sampling technique, random under-sampling, randomly drops observations from the majority class to balance the class distribution. Consequently, random under-sampling has achieved worse performance than the original model, with a slightly higher balanced accuracy. On the other hand, the Tomek Links method has achieved the lowest performance metrics. The edited nearest neighbours (ENN) technique helps find ambiguous and noisy data points. As a result, ENN has achieved satisfactory performance metrics, with the F1 score reaching an 80% mark. Repeated ENN is essentially the ENN algorithm applied successively until no more data points are removed. We can observe better performance metrics achieved by the repeated ENN. The instance hardness threshold (IHT) has achieved the highest performance metrics in the under-sampling methods, with an F1-score of 90%. The IHT algorithm alleviates the class imbalance by removing hard samples. The hard samples are essentially the overlapping data points in the dataset, and the IHT removes data points classified with a low probability. Hence, the total number of observations is less than half of the original data. As for the over-sampling methods, we adopted random over-sampling and SMOTE, as they were the most frequently utilized sampling techniques in the related literature. Although the performance metrics achieved by random over-sampling are exceptionally high, the method randomly duplicates the observations from the minority class. As a result, the post-sampled data contained over 1,500 duplicated observations, indicating that the model has a high bias and is unreliable.

On the other hand, SMOTE generates synthetic examples of the minority class, thereby avoiding mere duplication as the generated instances are similar, but not identical. The SMOTE-based model yielded respectable results with an F1-score of 84.5%. Additionally, algorithms such as SMOTE Tomek Links and SMOTE ENN employ both under and over-sampling techniques. The philosophy behind using these methods is to oversample the minority class using SMOTE and perform cleaning (or under-sampling) using Tomek Links or ENN. While ENN and Tomek Links share similar objectives, Tomek Links employs more restrictive conditions, resulting in fewer deleted data points. While SMOTE ENN achieved the highest performance metrics across all the methods, the minority class became the majority following the sampling, suggesting that the model may have lost crucial information about the majority class. In contrast, SMOTE Tomek Links, with an equal class distribution, secured satisfactory results with an F1-score of 85%.

However, upon examining the results of the polynom-fit-SMOTE + Tomek Links combination, it's evident that it yields an even higher F1-score of 89.55%, alongside an accuracy of 89.66% and balanced accuracy of 89.27%. These results indicate that this method not only maintains a balanced representation but also enhances the model's performance. Therefore, given these promising results and the robustness of the combined approach in representing the minority class more accurately, this study will adopt the polynom-fit-SMOTE + Tomek Links as the sampling method.

Table 11

The results of the proposed stacking ensemble based on various sampling techniques.

Sampling Method	Total Observations	Accuracy	Balanced Accuracy	F1-score
Original	3263	80.25	60.21	76.46
Under-Sampling				
Random Under Sampling	1530	63.72	63.72	63.63
Tomek Links	3043	71.75	57.03	69.73
Edited Nearest Neighbors	2172	80.45	77.00	80.09
Repeated Edited Nearest Neighbors	1689	85.79	85.55	85.78
Instance Hardness Threshold	1587	90.09	90.01	90.07
Over-Sampling				
SMOTE	4996	84.50	84.50	84.49
Random Over Sampling	4996	96.30	96.30	96.29
Combination of Under and Over-Sampling				
SMOTE Tomek Links	4890	85.07	85.07	85.06
SMOTE ENN	3111	93.81	93.51	93.81
polynom-fit-SMOTE + Tomek Links	4775	89.66	89.27	89.55

Figure 12 visualizes the transformation of the dataset across different stages of data sampling. First, the Principal Component Analysis (PCA) is used to reduce the data dimensionality to two components. This enables the effective projection of high-dimensional data into a more digestible 2D space, after scaling all features to ensure a fair comparison across different stages. In the initial stage, the original dataset showcases a significant class imbalance, with instances of the majority class far outweighing those of the minority class. This imbalance could lead to a skewed learning process favoring the majority class, and thereby compromise the model's capability to accurately predict instances of the minority class. To address this issue, we apply the Polynom-fit-SMOTE technique, which generates synthetic instances by interpolating between minority class instances and their polynomial fitting. This process amplifies the representation of the minority class while preserving the underlying structure of its instances in the feature space. As a result, we can observe a clear increase in the density of minority class instances, especially in the center of the distribution. This indicates that critical regions of the problem space are now populated with more varied minority instances. In the final stage, we employ Tomek Links, an undersampling technique. Tomek Links identifies pairs of different class instances that are nearest

neighbors and removes the majority class instance from each pair. This action not only improves the balance between classes but also enhances the decision boundary, paving the way for improved classification performance. The visualization underscores the transformation of the dataset from an imbalanced state to a more balanced one, thus setting the stage for a more effective model training process.

This study performs validation on the proposed methods on two publicly available datasets: UCI Heart Disease [46] and UAE Retrospective Cohort Study [19]. Performing the domain knowledge-based feature selection on UCI Heart Disease yields the selection of the following five features: age, gender, resting blood pressure, total cholesterol, and fasting blood sugar. The dataset does not require sampling as the target class distribution is almost equal. Table 12 presents the performance results of the classifiers on the UCI Heart Disease dataset using five features from the domain. The proposed stacking ensemble achieved the highest performance metrics with an accuracy of 85.29%, balanced accuracy of 85.11%, precision of 85.32%, recall of 85.29%, and F1-score of 85.26%. The proposed stacking ensemble achieved satisfactory results, indicating the model can accurately predict both class instances given the five features from the domain knowledge-based feature selection. Apart from adaptive boosting, all ensemble classifiers achieved higher performance results than single classifiers.

Table 12
Validation results of the classifiers on the UCI Heart Disease dataset.

Classifier	Accuracy	Balanced Accuracy	Precision	Recall	F1-score
Single Classifiers					
Naive Bayes	71.00	70.38	71.38	71.00	70.62
Logistic Regression	71.42	70.93	71.59	71.42	71.18
SVM	72.68	72.37	72.69	72.68	72.58
Multilayer Perceptron	73.52	73.06	73.71	73.52	73.32
Decision Tree	74.78	75.04	75.30	74.78	74.78
K-Nearest Neighbour	78.99	79.21	79.41	78.99	79.00
Ensemble Classifiers					
AdaBoost	70.16	69.89	70.16	70.13	70.09
Hard Voting	82.35	82.53	82.66	82.35	82.36
XGB	83.19	83.13	83.19	83.19	83.19
Gradient Boosting	83.19	83.23	83.26	83.19	83.20
Soft Voting	84.03	84.07	84.10	84.03	84.04
Random Forest	84.87	84.87	84.89	84.87	84.88
Stacking Ensemble	85.29	85.11	85.32	85.29	85.26

The selected subset of features in the UAE dataset is age, systolic blood pressure, total cholesterol, HDL-C, smoking, gender, BMI, and triglycerides. There was a high positive correlation between LDL-C and total cholesterol. However, since correlated features do not generally add more information, we dropped the LDL-C variable from the feature subset to reduce the complexity of the models. There is a high imbalance in the target class with almost a 9:1 ratio. Hence, following the feature selection process, we applied the selected sampling method, polynom-fit-SMOTE + Tomek Links. Table 13 shows the results of the classifiers based on a domain knowledge-based feature selection and polynom-fit-SMOTE + Tomek Links sampling technique. The total observations increased from 977 to 1257 following the sampling process. The proposed stacking ensemble has achieved the highest performance metrics with 96.23% accuracy, 96.22% balanced accuracy, 96.27% precision, 96.23% recall and 96.23% F1-score. The achieved performance results indicate that the proposed model is good at predicting both target classes. Interestingly, all the ensemble classifiers, except for adaptive boosting, have achieved higher performance metrics than all the single classifiers, shining a light on the predictive power of ensemble models.

Nevertheless, MLP, KNN and SVM have achieved satisfactory results from the single classifiers. The achieved results confirm the validity of the proposed domain-based feature selection, as can be observed from the results on the validation dataset, where only the domain risk factors present in the dataset were selected as a feature subset.

Table 13
Validation results of the classifiers on the UAE Cardiovascular Disease dataset.

Classifier	Accuracy	Balanced Accuracy	Precision	Recall	F1-score
Single Classifiers					
Naive Bayes	72.46	72.51	76.20	72.46	71.46
Logistic Regression	79.13	79.15	79.81	79.13	79.01
Decision Tree	82.60	82.62	83.05	82.60	82.55
Multilayer Perceptron	88.98	88.99	88.32	88.98	88.96
K-Nearest Neighbour	89.27	89.28	89.50	89.27	89.26
SVM	89.85	89.86	90.24	89.85	89.83
Ensemble Classifiers					
AdaBoost	77.97	77.98	78.11	77.97	77.94
Hard Voting	89.85	89.87	90.34	89.85	89.82
Random Forest	90.72	90.73	90.92	90.72	90.71
Soft Voting	91.88	91.89	92.24	91.88	91.86
XGB	92.17	92.17	92.18	92.17	92.17
Gradient Boosting	95.07	95.07	95.08	95.07	95.07
Stacking Ensemble	96.23	96.22	96.27	96.23	96.23

We can observe from the results achieved on the Framingham, UCI Heart Disease, and UAE datasets that the goal to produce a stacked ensemble classifier that outperforms single classifiers and other ensemble techniques and achieve optimal prediction metrics has been successfully achieved. Figure 13 shows the base models and the meta-model of the best-performing stacked classifier after performing extensive combinations of different base models and meta-models. The base models selected for the stacking ensemble are KNN, SVM, RF, GB and XGB, and XGB as the meta-learner. An interesting observation in the models that performed well when stacked together is that they are heterogeneous. For example, KNN is a type of lazy algorithm which essentially does not perform training but stores the data during the training time. What makes KNN unique is that it does not assume the underlying data, which becomes ideal for

non-linear data. Based on the results, KNN has been the top-performing model on the single classifiers side. The potential reason for that could be explained by the low correlation between the selected subset of features in each dataset and the sampling method, Tomek Links, which performs data cleaning to get rid of noisy data. SVM was also one of the top performers in the single classifiers in the Framingham and UAE datasets. However, SVM performed poorly on the UCI Heart Disease dataset. One of the limitations of SVM is that it performs well when there is a clear margin of separation between classes, which the selected features in the UCI Heart Disease dataset may not have had. We did not perform sampling on the UCI Heart Disease, which may have caused noisy data to affect the performance of SVM. Discussing the tree-based models, RF is a bagging ensemble algorithm that uses several DTs, and generally, DTs tend to have high variance and low bias and, as a result, overfit easily. RF aims to solve the issues of DTs by sampling the data and the features, hoping to reduce the variance at a small cost of bias as much as possible. The mentioned phenomenon can be observed in the achieved results where RF has significantly outperformed DT in each dataset. GB and XGB are among the boosting ensemble classifiers that aim to boost the accuracy of the models at the expense of increased variance (i.e., can easily overfit the data). Since DTs are usually used as the base models in GB and XGB, they tend to have high accuracies but are also highly susceptible to overfitting, as outliers easily influence them. However, with the hyperparameter tuning offered by GB and XGB, it is possible to avoid overfitting, which can be observed in these models' precision and recall metrics. KNN and SVM have achieved the highest F1-scores for single classifiers in the Framingham and UAE datasets, and KNN on the UCI Heart Disease dataset, indicating that the classifiers are good at predicting positive instances and avoid misclassifying a negative instance as positive. Comparably, RF has achieved balanced results across all the metrics. Since GB is a boosting classifier, it is not surprising that it achieved the highest accuracy score compared to other classifiers in the Framingham and UAE datasets. XGB, a boosting algorithm, has not performed as well as other ensemble classifiers as expected. However, it performed the best as the meta-model in the stacked ensemble. A possible reason for such a case is that XGB learns the misclassifications and tries to correct these misclassifications in each iteration. Since the models KNN, SVM, RF, and GB provide independent predictions, these predictions are appended as features in the training data. Then, the transformed training data from the base models fit the meta learner, XGB, which makes the final predictions. Since XGB yielded knowledge from the base learners, it can generalize the predictions and resolve to boost the model's performance in the final prediction. Based on the results achieved, F1-score is the deciding measurement in this study. Indeed, the proposed stacked ensemble method has achieved the highest F1-score compared to all the ensemble and single classifiers used as a comparison. However, we must mention that the computation time taken to develop the stacked ensemble is slightly higher than other ensemble classifiers in the study.

This study performed a statistical hypothesis test using the 5×2 Cross-Validation Test, comparing the proposed ensemble for each dataset against developed classifiers to solidify the performance differences observed. The 5×2 Cross-Validation Test splits the dataset into two parts, 50% training, and 50% testing, and repeats the splitting process for five iterations. In each of the five iterations, both models are fit to the training split and use the testing split to evaluate the performance. Following the five iterations, the training and test split rotates, where the training becomes the test split and the test becomes the training

split, and the performance of both models is evaluated once more. The null hypothesis states that the stacked ensemble and the model in comparison have equal performance. If the p-value is smaller than α (i.e., 0.05), we reject the null hypothesis. Table 14 shows each dataset's results from the 5×2 Cross-Validation Test. Based on the hypothesis test, we can observe that the proposed stacked ensemble performs significantly better than all classifiers except KNN in the Framingham dataset. We think this could be due to KNN making almost no assumptions about the data. Nevertheless, the proposed stacked classifier outperforms the other ensemble classifiers in the Framingham dataset. However, the same cannot be said for the UCI Heart Disease and UAE datasets. The proposed stacking ensemble did not perform significantly better than GB, RF, ML and DT in the UCI Heart Disease dataset and KNN, soft voting, GB, and XGB in the UAE dataset. In many of the studies on cardiovascular disease prediction, authors who opt to use ensemble methods tend to use techniques such as majority voting, boosting, and bagging. On the other hand, stacking is among the least used ensemble methods for predicting cardiovascular diseases. In this study, we investigated the stacking ensemble and found that it outperforms other ensemble classifiers. As a result, the stacking ensemble has the potential to stretch to other health domains, such as specific cardiovascular diseases and diabetes.

Table 14
Result of the significance test using 5x2 Cross-Validation Test on three datasets.

Classifier	p (Framingham)	p (UCI Heart Disease)	p (UAE)
K-Nearest Neighbour	0.233	0.026*	0.064
Soft Voting	0.044*	0.028*	0.057
Gradient Boosting	0.008**	0.322	0.316
Hard Voting	0.006**	0.026*	0.005**
XGB	0.005**	0.042*	0.064
Random Forest	0.003**	0.200	0.003**
Multilayer Perceptron	0.003**	0.100	0.016*
Naive Bayes	< 0.001**	0.001*	< 0.001**
Decision Tree	< 0.001**	0.057	0.005**
Logistic Regression	< 0.001**	0.006**	< 0.001**
Support Vector Machine	< 0.001**	0.029*	0.005**
AdaBoost	< 0.001**	0.003**	< 0.001**

* $p < 0.05$; ** $p < 0.01$.

Table 15 presents a comparison of our proposed model with the best-performing models from the literature, all aiming to predict cardiovascular disease using the Framingham data. A variety of models have been implemented, including Decision Trees (DTs), Stacking, and Boosting Ensembles, tested under three different data split techniques: 70:30, 80:20, and 10-fold cross validation. In the 70:30 train-test

split, our proposed model using domain knowledge for feature selection and a stacked ensemble as a classifier scored fairly well, achieving 89.26% accuracy and an F1-score of 89.16%. This slightly trails the randomized decision tree model from [7], which garnered 91% in both accuracy and F1-score. When tested with the 80:20 train-test split, the proposed model demonstrated consistent performance with 89.66% accuracy and 89.55% F1-score. The Random Forest model in [42], using SelectKBest for feature selection, outperformed other models in terms of accuracy at 94.63%, although it didn't report an F1-score. The Decision Tree model from [40], applying covariance for feature selection, showcased a promising F1-score of 94.08%. During the 10-fold cross-validation test, our model presented an accuracy of 88.51% and an F1-score of 90.49%. The Boosting Ensemble from [44], which employed SelectKBest for feature selection, reached a high accuracy of 99.1%.

Although our model did not attain the highest F1-score or accuracy across all splits, it's important to highlight the balanced performance it displayed across different testing techniques. Our approach employs the polynom-fit-SMOTE + Tomek Links sampling technique, which ensures that valuable information from the minority class is preserved. This is crucial, particularly when dealing with an imbalanced target attribute. Additionally, our feature selection method, based on domain knowledge, performed competitively when compared with other literature methods such as k highest score and LASSO. The consistency in performance across different data splits and the robust handling of imbalances in the data make our proposed model a valuable contribution to this field of study.

Table 15
Comparison of proposed study with existing literature for Framingham dataset.

Author	Feature Selection	No. of Features	Classifier	Accuracy	F1-score
70:30 Train-Test Split					
[7]	-	15	Randomized Decision Tree	91.00	91.00
Proposed	Domain Knowledge	10	Stacked Ensemble	89.26	89.16
80:20 Train-Test Split					
[42]	SelectKBest	-	Random Forest	94.63	-
[41]	Multiple Logistic Regression	8	Boosted Decision Tree	85.00	10.86
[45]	-	15	Stacking Ensemble	90.24	90.00
[40]	Covariance	-	Decision Tree	92.59	94.08
Proposed	Domain Knowledge	10	Stacked Ensemble	89.66	89.55
10-Fold Cross Validation					
[43]	Random Forest	4	LightGBM	83.10	-
[44]	SelectKBest	5	Boosting Ensemble	99.10	-
Proposed	Domain Knowledge	10	Stacked Ensemble	88.51	90.49

Table 16 compares the proposed stacked ensemble against the related studies on the UCI Heart Disease dataset. Most studies adopted ensemble learning techniques, including stacking, boosting, and bagging. On the other hand, [32] adopted a deep learning technique. However, we cannot perform an appropriate comparison as we could not retrieve the F1-score for the proposed model. Studies by [33, 37, 45] proposed a stacked ensemble with two, five and seven base learners and SVM, LR, and nuSVM as the meta-learners, respectively. The highest F1-score, 97.15%, for stacked ensembles was achieved by two deep learning models as the base learner and SVM as the meta-learner. The study by [35] achieved a 100% for all the performance metrics using RF. However, the study used the UCI Cleveland and StatLog datasets containing duplicated observations. As a result, the developed model may potentially contain bias. Our proposed stacked ensemble did not achieve as high-performance metrics as other studies on the UCI Heart Disease dataset. A potential reason for such a case could be due to the subset of features selected or the possible existence of noise in the data, which affected some of the classifiers in the base learners. Nevertheless, the proposed stacked ensemble achieved satisfactory results, and the proposed domain knowledge-based feature selection could generalize to three utilized datasets. As a result, we

successfully adopted domain knowledge into the ML prediction process to promote the credibility of the models.

Table 16
Comparison of proposed study with existing literature for UCI Heart Disease dataset.

Author	Feature Selection	No. of Features	Classifier	Accuracy	F1-score
[32]	LASSO	9	Deep Learning	94.20	-
[37]	SHAP	10	Stacking Ensemble	89.86	91.36
[45]	-	13	Stacking Ensemble	95.83	96.00
[33]	Recursive Feature Elimination	6	Stacked SVM	97.17	97.15
[31]	-	-	XGB	99.03	99.00
[35]	LASSO	11	Random Forest	100.0	100.0
Proposed	Domain Knowledge	5	Stacked Ensemble	85.29	85.26

The adopted UAE dataset in this study was one of the first for the task of cardiovascular disease prediction. Hence, we could not find any other studies that used the dataset to perform a comparative study. Following the comparative analysis of our proposed methods against other relevant studies in the literature, we find it beneficial to delve into another critical facet of modern machine learning practices - explainable AI. While the primary objective of this research is to develop trustworthy models that are underpinned by well-established domain knowledge, another equally important aspect is the interpretability and transparency of these models. Indeed, as we venture into leveraging the predictive power of an ensemble of models through stacking, ensuring the interpretability of these models becomes vital. Explainable AI, sometimes referred to as XAI, is an area in machine learning that focuses on creating models whose actions can be easily understood by human beings. The importance of explainable AI in the field of medical research, and more specifically, in the prediction of cardiovascular diseases, cannot be overstated. It provides a way to understand how input features relate to the model's predictions, which in turn can help in gaining insights that can be useful for improving model performance, validating the model's decisions, and increasing human trust in AI.

LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are two such methodologies that can help shed light on the model's decision-making process. LIME is designed to explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model. SHAP, on the other hand, connects game theory with local explanations, uniting several previous methods and representing the only possible consistent and locally

accurate additive feature attribution method based on expectations. In this complex domain of cardiovascular disease prediction, where decisions could have significant impacts on patient care, explainable AI models such as those using LIME and SHAP can provide much-needed transparency. The use of these techniques allows us to not just develop models that perform well, but also understand why a prediction was made, increasing trust, and enabling further refinement of the models. In the following sections, we shall delve into the application and outcomes of using LIME and SHAP in the context of our study.

Figure 14 represents the LIME results for two different instances of the cardiovascular disease prediction model. The LIME output consists of three sections: prediction probabilities, feature contributions towards the classes, and table of features with their actual values. Figure 14 (a) showcases an instance predicted as 'No CVD', with the influential features contributing to the prediction: female gender about 20%, no daily cigarette consumption contributing 17% and HDL cholesterol level (hdlc) above 53, contributing about 16%. The age above 65 contributed towards the 'CVD' class, however, the aggregated contribution of the features was towards the 'No CVD' class. Figure 14 (b), on the other hand, represents an instance predicted as 'CVD', with male gender, age between 63 to 65 and daily cigarette consumption from 0 to 6 contributing the most towards the class prediction. These visual explanations provide insights into the model's decision-making process, shedding light on how each feature contributes to the final prediction. By doing so, we not only improve the transparency of the model but also enhance our understanding of the factors that influence the prediction of cardiovascular diseases. This would greatly help healthcare practitioners make more informed decisions when it comes to managing patient health.

While LIME offers valuable insights at the local level, understanding the global behaviour of our model is equally important. SHAP is an interpretability approach that not only offers local interpretability akin to LIME but also guarantees consistency and accuracy at a global scale. The SHAP Waterfall Plot, as displayed in Fig. 15, provides an interpretive snapshot of our model's decision-making process at an instance level for the same instances shown in LIME. Each bar corresponds to a feature's contribution towards the prediction, with the direction signifying the class it is influencing. In Fig. 15 (a), we see an instance from the No CVD class. The daily cigarette consumption (cigpday) of 0 is the most significant contributor, pushing the prediction towards No CVD. Following it, the female gender and high HDL cholesterol level (hdlc) further contribute negatively, reinforcing the No CVD prediction. However, it's worth noting the age of 74 and total cholesterol (totchol) of 183, which lean towards predicting CVD, suggesting these features could be risk factors. Conversely, in Fig. 15 (b), representing an instance from the CVD class, the contributions are quite balanced. Notably, a daily cigarette consumption (cigpday) of 6, age of 65, male gender, and systolic blood pressure (sysbp) of 144 all push the prediction towards the CVD class. Only the BMI and non-diabetic lean towards the No CVD class.

When comparing the interpretations of SHAP and LIME for the No CVD class, it becomes apparent that both methods emphasize the significance of the 'cigpday', 'gender' and 'hdlc' features in predicting the absence of cardiovascular disease. Lower daily cigarette consumption and female gender are seen in both LIME and SHAP as influential factors for the No CVD class. Considering the CVD class, the 'gender',

'age' and 'cigpday' features are identified as significant in both the LIME and SHAP explanations, contributing to the prediction of the CVD class.

Figure 16 features a SHAP Summary plot, a robust visualization tool that synergistically showcases feature importance alongside their respective effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The color represents the value of the feature from low to high. Starting with the 'cigpday' feature, it displays a range of contributions. Low number of cigarettes daily leans towards a positive influence, while a low-to-medium number leans towards a negative influence. This might be interpreted as a nuanced representation of how cigarette consumption influences cardiovascular risk – a low number of daily cigarettes might signal a healthier lifestyle or recent successful cessation attempts, thus reducing CVD risk. The 'age' feature shows that older age contributes positively towards the prediction of CVD, reflecting the well-established correlation between age and cardiovascular risk. The 'gender' feature shows that being male positively influences the model's prediction towards CVD. This corresponds with epidemiological data showing higher CVD risk in men. 'sysbp' and 'ldlc' display similar behavior, where high values push the prediction towards CVD, aligning with medical understanding that high systolic blood pressure and low-density lipoprotein cholesterol are risk factors for CVD. 'hdlc' shows a contrary effect, where low values tend to push the prediction towards CVD, corroborating the fact that low levels of high-density lipoprotein cholesterol (HDL-C, often called 'good cholesterol') are associated with increased CVD risk. The presence of diabetes also pushes the prediction towards CVD, reflecting the well-known association between diabetes and cardiovascular disease. Finally, 'totchol' shows a more subtle influence, where low values lean towards a negative prediction for CVD, while high values don't make much impact. This might be due to the complex relationship between total cholesterol and CVD, where both very low and very high levels can be associated with health risks.

Concluding the results and discussion section, this study has revealed several key findings that contribute to the field of cardiovascular disease prediction. Firstly, the utility of domain-based feature selection was demonstrated. By incorporating domain knowledge in feature selection, we achieved satisfactory results, reaffirming the value of expert knowledge in the ML process. This approach was compared with LASSO and SelectKBest chi2 feature selection methods, which identified the top ten features from the dataset, reinforcing the effectiveness of domain knowledge-based selection. Addressing the challenge of class imbalance in the dataset, the application of polynom-fit-SMOTE with Tomek Links was found to be the most effective technique in enhancing the prediction performance. This combination was preferred for its superior handling of minority class representation and reduction of noise or borderline majority instances. To validate our methodology, we performed experiments on two separate datasets – the UCI Heart Disease and UAE datasets. This demonstrated the generalizability and robustness of our model, revealing consistent results across different data. To further substantiate the performance improvement, hypothesis testing was conducted. The tests confirmed that the enhancements in our proposed methods were statistically significant, lending strong empirical support to our findings. A comparative study against other literature provided a benchmark, situating our results within the wider field of research. Finally, the use of LIME and SHAP for explainable AI introduced transparency into our model's

predictions. These techniques elucidated the contributions of individual features towards the prediction outcome, providing a valuable tool for interpretability and fostering trust in our model. As we move towards the conclusion, it is evident that the combination of domain knowledge-based feature selection, appropriate handling of class imbalance, and a commitment to model interpretability have paved the way for an effective and trustworthy model for predicting cardiovascular disease.

5. Conclusions

This study proposed the use of domain knowledge-based feature selection to enhance the credibility of ML models for predicting cardiovascular diseases (CVDs). This approach prioritizes a holistic understanding of the disease, combining this rich context with the predictive power of a stacking ensemble. The training and evaluation of these models primarily relied on the Framingham Heart Study dataset, while the UCI Heart Disease and UAE Retrospective Cohort Study datasets served to validate the proposed methods' generalizability. Exploratory data analysis revealed that gender, age, and diabetes are among the important risk factors for CVDs. This study compared the proposed domain knowledge-based feature selection against the two frequently used feature selection methods: k highest score and LASSO. Moreover, several under-sampling, over-sampling and combinations of under- and over-sampling techniques were evaluated on a highly imbalanced dataset. Several weak classifiers and a variety of ensemble classifiers were developed to serve as a comparison against the proposed stacked ensemble classifier.

The results show that the proposed domain knowledge-based feature selection performed at a comparable level against other feature selection methods. Moreover, the proposed stacked ensemble classifier had the highest F1-score in each dataset, indicating it can accurately predict both cases of the target class. Regarding the sampling technique, we adopted the polynom-fit-SMOTE + Tomek Links sampling technique for its effective balance of information retention and performance. The significance test results indicate that our proposed model performs significantly better than all classifiers except KNN in the Framingham dataset. However, the proposed classifier took a slightly longer time to train as compared to other ensembles. The comparative analysis showed that the proposed ensemble classifier did not achieve the highest F1-score. However, the proposed method performed considerably well compared to the related literature. The incorporation of LIME and SHAP methods in this study ensured that our predictive models did not operate as "black boxes". These techniques offered a more nuanced understanding of our model's behavior, enabling us to interpret which features were driving predictions. For instance, using LIME and SHAP, we found features like 'cigpday', 'age', 'gender', 'sysbp', 'hdlc', 'diabetes', 'totchol', and 'ldlc' to be particularly influential. This transparency serves to improve trust in our model's predictions and provides insightful data for future research, clinical decisions, and health policymaking.

Future work could enhance the domain knowledge-based feature selection by integrating it with other feature selection methods, potentially yielding a more precise feature subset for cardiovascular disease prediction. Further, it would be interesting to test the efficacy of these models using real-world clinical

data to determine their applicability in everyday healthcare settings. We also suggest exploring graph-based methods like node embeddings to capture complex interrelations among cardiovascular disease risk factors. Finally, a broader exploration of explainable AI techniques could enhance the models' transparency, fostering trust among healthcare practitioners.

Declarations

Author Contributions: Conceptualization, Z.R., S.T. and J.R.; methodology, Z.R., J.R. and M.S.S.; software, Z.R.; validation, Z.R., J.R., V.B. and J.Y.T.; formal analysis, Z.R.; investigation, J.R., J.Y.T., M.S.S. and Z.R.; resources, J.R. and S.T.; data curation, Z.R.; writing—original draft preparation, Z.R., M.S.S. and J.Y.T.; writing—review and editing, Z.R., S.T., J.R. and N.Z.; visualization, Z.R.; supervision, N.Z., S.T. and V.B.; project administration, N.Z., S.T. and V.B.; funding acquisition, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the United Arab Emirates University Strategic Research Grant G00003676 (Fund No.: 12R136) through Big Data Analytics Centre.

Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article and there are no financial interests.

Ethical Approval: Not applicable.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable.

Data Availability Statement: The UCI Heart Disease dataset is available at <https://archive.ics.uci.edu/ml/datasets/heart+disease>, and the UAE Retrospective Cohort Study dataset available as the additional file in the study by [19].

Acknowledgments: The authors would like to thank the United Arab Emirates University Strategic Research Grant G00003676 (Fund No.: 12R136) through Big Data Analytics Centre.

References

1. World Health Organization Cardiovascular Diseases (CVDs) Available online: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
2. Doppala, B.P.; Bhattacharyya, D.; Janarthanan, M.; Baik, N. A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques. *J. Healthc. Eng.* **2022**, *2022*, doi:10.1155/2022/2585235.
3. Rustamov, Z. Clustering and Association Rule Mining of Cardiovascular Disease Risk Factors. In; 2022; pp. 389–396.

4. Hosni, M.; Carrillo de Gea, J.M.; Idri, A.; El Bajta, M.; Fernández Alemán, J.L.; García-Mateos, G.; Abnane, I. A Systematic Mapping Study for Ensemble Classification Methods in Cardiovascular Disease. *Artif. Intell. Rev.* **2021**, *54*, 2827–2861, doi:10.1007/s10462-020-09914-6.
5. Masih, N.; Ahuja, S. Prediction of Heart Diseases Using Data Mining Techniques. *Int. J. Big Data Anal. Healthc.* **2018**, *3*, 1–9, doi:10.4018/IJBDAH.2018070101.
6. Uddin, M.N.; Halder, R.K. An Ensemble Method Based Multilayer Dynamic System to Predict Cardiovascular Disease Using Machine Learning Approach. *Informatics Med. Unlocked* **2021**, *24*, 100584, doi:10.1016/j.imu.2021.100584.
7. Mienye, I.D.; Sun, Y.; Wang, Z. An Improved Ensemble Learning Approach for the Prediction of Heart Disease Risk. *Informatics Med. Unlocked* **2020**, *20*, 100402, doi:10.1016/j.imu.2020.100402.
8. Al'Aref, S.J.; Anchouche, K.; Singh, G.; Slomka, P.J.; Kolli, K.K.; Kumar, A.; Pandey, M.; Maliakal, G.; van Rosendaal, A.R.; Beecy, A.N.; et al. Clinical Applications of Machine Learning in Cardiovascular Disease and Its Relevance to Cardiac Imaging. *Eur. Heart J.* **2019**, *40*, 1975–1986, doi:10.1093/eurheartj/ehy404.
9. Wang, J.; Wang, H.; Oh, J.; Wiens, J. Learning Credible Models. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **2018**, 2417–2426, doi:10.1145/3219819.3220070.
10. Michos, E.D.; Khan, S.S. Further Understanding of Ideal Cardiovascular Health Score Metrics and Cardiovascular Disease. *Expert Rev. Cardiovasc. Ther.* **2021**, *19*, 607–617, doi:10.1080/14779072.2021.1937127.
11. PURISCH, S.E.; ALGODI, M.; TAUB, C.; GOFFMAN, D. Evaluation and Diagnostic Testing. *Clin. Obstet. Gynecol.* **2020**, *63*, 828–835, doi:10.1097/GRF.0000000000000566.
12. Zhenya, Q.; Zhang, Z. A Hybrid Cost-Sensitive Ensemble for Heart Disease Prediction. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 73, doi:10.1186/s12911-021-01436-7.
13. MB, A.A.; WA, K. Cardiovascular Diseases Risk Prediction Using the Framingham Risk Score. *Egypt. J. Occup. Med.* **2021**, *45*, 249–264, doi:10.21608/ejom.2021.193283.
14. Ofori, E.K.; Intiful, F.D.; Asante, M.; Asare, G.A.; Adjei, P.K.; Steele-Dadzie, R.K.; Amoako-Mensah, A.; Mensah, D.; Angmorterh, S.K. Prevalence of Cardiovascular Disease Risk Factors among Students of a Tertiary Institution in Ghana. *Food Sci. Nutr.* **2018**, *6*, 381–387, doi:10.1002/fsn3.565.
15. Kumma, W.P.; Lindtjørn, B.; Loha, E. Modifiable Cardiovascular Disease Risk Factors among Adults in Southern Ethiopia: A Community-Based Cross-Sectional Study. *BMJ Open* **2022**, *12*, e057930, doi:10.1136/bmjopen-2021-057930.
16. Roman, W.P.; Martin, H.D.; Sauli, E. Assessment of Risk Factors for Cardiovascular Diseases among Patients Attending Cardiac Clinic at a Referral Hospital in Tanzania. *J. Xiangya Med.* **2019**, *4*, 18–18, doi:10.21037/jxym.2019.03.05.
17. Lind, L.; Ingelsson, M.; Sundstrom, J.; Ärnlöv, J. Impact of Risk Factors for Major Cardiovascular Diseases: A Comparison of Life-Time Observational and Mendelian Randomisation Findings. *Open Hear.* **2021**, *8*, e001735, doi:10.1136/openhrt-2021-001735.

18. Avesta, L.; Rasoolzadeh, S.; Naeim, M.; Kamran, A. Prevalence of Cardiovascular Disease Risk Factors in the Women Population Covered by Health Centers in Ardabil. *Int. J. Hypertens.* **2022**, *2022*, 1–8, doi:10.1155/2022/2843249.
19. Al-Shamsi, S.; Regmi, D.; Govender, R.D. Incidence of Cardiovascular Disease and Its Associated Risk Factors in At-Risk Men and Women in the United Arab Emirates: A 9-Year Retrospective Cohort Study. *BMC Cardiovasc. Disord.* **2019**, *19*, 148, doi:10.1186/s12872-019-1131-2.
20. Khandia, R.; Puranik, N.; Lodhi, N.; Gautam, B.; Alqahtani, T.; Alqahtani, A.M.; Alamri, A.H.; Chidambaram, K. Comparing Heart Risk Scores to Identify the Most Important Risk Factors for Cardiovascular Diseases. *Eur. Rev. Med. Pharmacol. Sci.* **2021**, *25*, 7947–7963, doi:10.26355/eurrev_202112_27645.
21. Tran, D.-M.T.; Lekhak, N.; Gutierrez, K.; Moonie, S. Risk Factors Associated with Cardiovascular Disease among Adult Nevadans. *PLoS One* **2021**, *16*, e0247105, doi:10.1371/journal.pone.0247105.
22. Dahlöf, B. Cardiovascular Disease Risk Factors: Epidemiology and Risk Assessment. *Am. J. Cardiol.* **2010**, *105*, 3A-9A, doi:10.1016/j.amjcard.2009.10.007.
23. O'Donnell, C.J.; Elosua, R. Cardiovascular Risk Factors. Insights from Framingham Heart Study. *Rev. Esp. Cardiol.* **2008**, *61*, 299–310, doi:10.1157/13116658.
24. Schnohr, P. Coronary Heart Disease Risk Factors Ranked by Importance for the Individual and Community. A 21 Year Follow-up of 12000 Men and Women from The Copenhagen City Heart Study. *Eur. Heart J.* **2002**, *23*, 620–626, doi:10.1053/euhj.2001.2842.
25. Schnohr, P.; Marott, J.L.; Kristensen, T.S.; Gynzelberg, F.; Gronbaek, M.; Lange, P.; Jensen, M.T.; Jensen, G.B.; Prescott, E. Ranking of Psychosocial and Traditional Risk Factors by Importance for Coronary Heart Disease: The Copenhagen City Heart Study. *Eur. Heart J.* **2015**, *36*, 1385–1393, doi:10.1093/eurheartj/ehv027.
26. Wang, T.; Zhao, Z.; Yu, X.; Zeng, T.; Xu, M.; Xu, Y.; Hu, R.; Chen, G.; Su, Q.; Mu, Y.; et al. Age-Specific Modifiable Risk Factor Profiles for Cardiovascular Disease and All-Cause Mortality: A Nationwide, Population-Based, Prospective Cohort Study. *Lancet Reg. Heal. - West. Pacific* **2021**, *17*, 100277, doi:10.1016/j.lanwpc.2021.100277.
27. Scottish Intercollegiate Guidelines Network *Risk Estimation and the Prevention of Cardiovascular Disease: A National Clinical Guideline*, 2017; Vol. 1;.
28. Kaptoge, S.; Pennells, L.; De Bacquer, D.; Cooney, M.T.; Kavousi, M.; Stevens, G.; Riley, L.M.; Savin, S.; Khan, T.; Altay, S.; et al. World Health Organization Cardiovascular Disease Risk Charts: Revised Models to Estimate Risk in 21 Global Regions. *Lancet Glob. Heal.* **2019**, *7*, e1332–e1345, doi:10.1016/S2214-109X(19)30318-3.
29. Marateb, H.R.; Goudarzi, S. A Noninvasive Method for Coronary Artery Diseases Diagnosis Using a Clinically-Interpretable Fuzzy Rule-Based System. *J. Res. Med. Sci.* **2015**, *20*, 214–223.
30. Shah, S.M.S.; Shah, F.A.; Hussain, S.A.; Batool, S. Support Vector Machines-Based Heart Disease Diagnosis Using Feature Subset, Wrapping Selection and Extraction Methods. *Comput. Electr. Eng.* **2020**, *84*, 106628, doi:10.1016/j.compeleceng.2020.106628.

31. Ahmad, G.N.; Fatima, H.; Shafiullah; Salah Saidi, A.; Imdadullah Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and Without GridSearchCV. *IEEE Access* **2022**, *10*, 80151–80173, doi:10.1109/ACCESS.2022.3165792.
32. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. *Comput. Intell. Neurosci.* **2021**, *2021*, doi:10.1155/2021/8387680.
33. Almulihi, A.; Saleh, H.; Hussien, A.M.; Mostafa, S.; El-Sappagh, S.; Alnowaiser, K.; Ali, A.A.; Refaat Hassan, M. Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction. *Diagnostics* **2022**, *12*, 3215, doi:10.3390/diagnostics12123215.
34. Sarra, R.R.; Dinar, A.M.; Mohammed, M.A.; Abdulkareem, K.H. Enhanced Heart Disease Prediction Based on Machine Learning and X2 Statistical Optimal Feature Selection Model. *Designs* **2022**, *6*, 87, doi:10.3390/designs6050087.
35. Mohamed, S.; Malhat, M.; Elhady, G. Prediction of Cardiovascular Disease Using Machine Learning Techniques. *IJCI. Int. J. Comput. Inf.* **2022**, *17*, 0–0, doi:10.21608/ijci.2022.129472.1071.
36. Hastie, T.; Tibshirani, R.; Friedman, J. *Elements of Statistical Learning 2nd Ed.*; 2009; Vol. 27;.
37. Liu, J.; Dong, X.; Zhao, H.; Tian, Y. Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion. *Processes* **2022**, *10*, 749, doi:10.3390/pr10040749.
38. Saboor, A.; Usman, M.; Ali, S.; Samad, A.; Abrar, M.F.; Ullah, N. A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms. *Mob. Inf. Syst.* **2022**, *2022*, 1–9, doi:10.1155/2022/1410169.
39. Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. *Biomed Res. Int.* **2020**, *2020*, doi:10.1155/2020/9816142.
40. Patel, J.; Khaked, A.A.; Patel, J.; Patel, J. Heart Disease Prediction Using Machine Learning. In *Springer*, 2021; Vol. 203, pp. 653–665 ISBN 9789813340831.
41. Beunza, J.J.; Puertas, E.; García-Ovejero, E.; Villalba, G.; Condes, E.; Koleva, G.; Hurtado, C.; Landecho, M.F. Comparison of Machine Learning Algorithms for Clinical Event Prediction (Risk of Coronary Heart Disease). *J. Biomed. Inform.* **2019**, *97*, 103257, doi:10.1016/j.jbi.2019.103257.
42. Kwakye, K.; Dadzie, E. Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases. **2021**.
43. Guarneros-Nolasco, L.R.; Cruz-Ramos, N.A.; Alor-Hernández, G.; Rodríguez-Mazahua, L.; Sánchez-Cervantes, J.L. Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms. *Mathematics* **2021**, *9*, 2537, doi:10.3390/math9202537.
44. Rahim, A.; Rasheed, Y.; Azam, F.; Anwar, M.W.; Rahim, M.A.; Muzaffar, A.W. An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. *IEEE Access* **2021**, *9*, 106575–106588, doi:10.1109/ACCESS.2021.3098688.
45. Louridi, N.; Douzi, S.; El Ouahidi, B. Machine Learning-Based Identification of Patients with a Cardiovascular Defect. *J. Big Data* **2021**, *8*, 133, doi:10.1186/s40537-021-00524-9.

46. Andras, J.; William, S.; Matthias, P.; Robert, D. Heart Disease Data Set - UCI Machine Learning Repository Available online: <https://archive-beta.ics.uci.edu/ml/datasets/heart+disease> (accessed on 10 January 2023).
47. Roth, G.A.; Mensah, G.A.; Johnson, C.O.; Addolorato, G.; Ammirati, E.; Baddour, L.M.; Barengo, N.C.; Beaton, A.Z.; Benjamin, E.J.; Benziger, C.P.; et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *J. Am. Coll. Cardiol.* **2020**, *76*, 2982–3021, doi:10.1016/j.jacc.2020.11.010.
48. Ahsan, M.; Gomes, R.; Chowdhury, M.M.; Nygard, K.E. Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector. *J. Cybersecurity Priv.* **2021**, *1*, 199–218, doi:10.3390/jcp1010011.
49. Gazzah, S.; Amara, N.E. Ben New Oversampling Approaches Based on Polynomial Fitting for Imbalanced Data Sets. In Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems; IEEE, September 2008; pp. 677–684.
50. Kovács, G. Smote-Variants: A Python Implementation of 85 Minority Oversampling Techniques. *Neurocomputing* **2019**, *366*, 352–354, doi:10.1016/j.neucom.2019.06.100.
51. Rustamov, Z.; Rustamov, J.; Sultana, M.S.; Ywei, J.; Balakrishnan, V.; Zaki, N. Cardiovascular Disease Prediction Using Ensemble Learning Techniques: A Stacking Approach. In Proceedings of the 2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA); IEEE, March 3 2023; pp. 93–98.
52. Chang, A.C. *Intelligence-Based Medicine: Artificial Intelligence and Human Cognition in Clinical Medicine and Healthcare*; 2020;
53. Bej, S.; Davtyan, N.; Wolfien, M.; Nassar, M.; Wolkenhauer, O. LoRAS: An Oversampling Approach for Imbalanced Datasets. *Mach. Learn.* **2021**, *110*, 279–301, doi:10.1007/s10994-020-05913-4.
54. Leevy, J.L.; Khoshgoftaar, T.M.; Bauder, R.A.; Seliya, N. A Survey on Addressing High-Class Imbalance in Big Data. *J. Big Data* **2018**, *5*, 42, doi:10.1186/s40537-018-0151-6.

Figures



Figure 1

Study count distribution of the continents in the domain knowledge study.

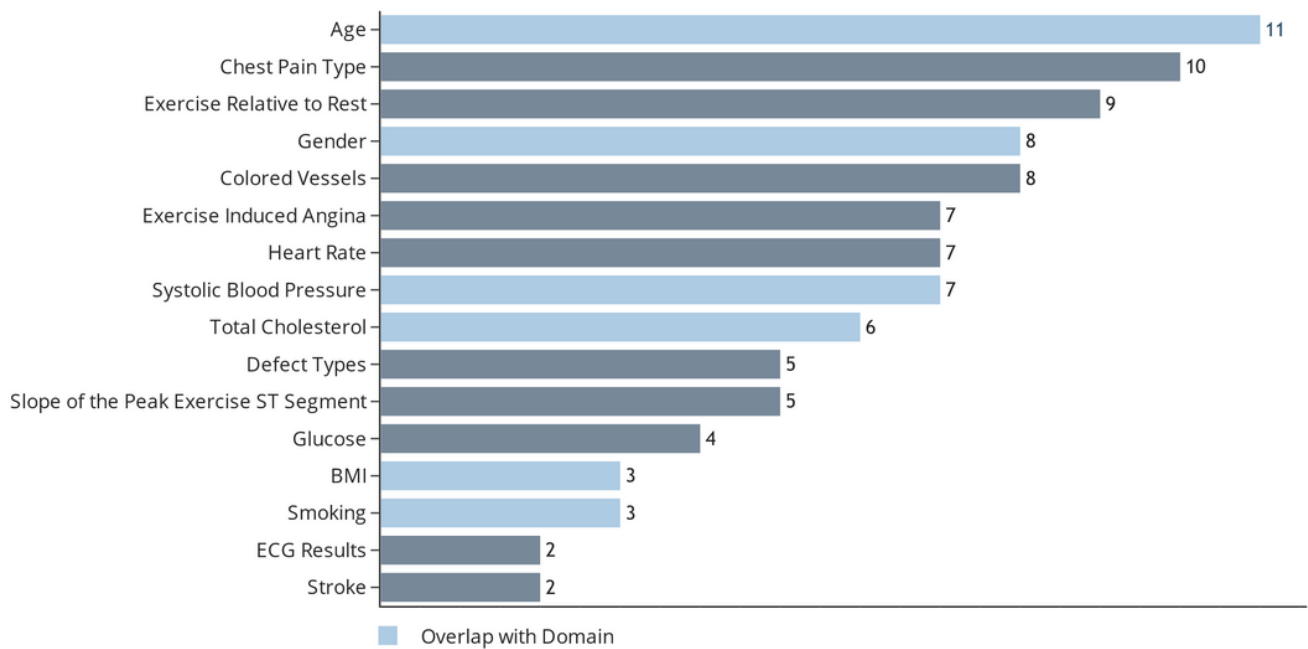


Figure 2

The total count of used features in the ML studies. The lighter shaded bars refer to the features that overlap with the risk factors analysed from the domain.

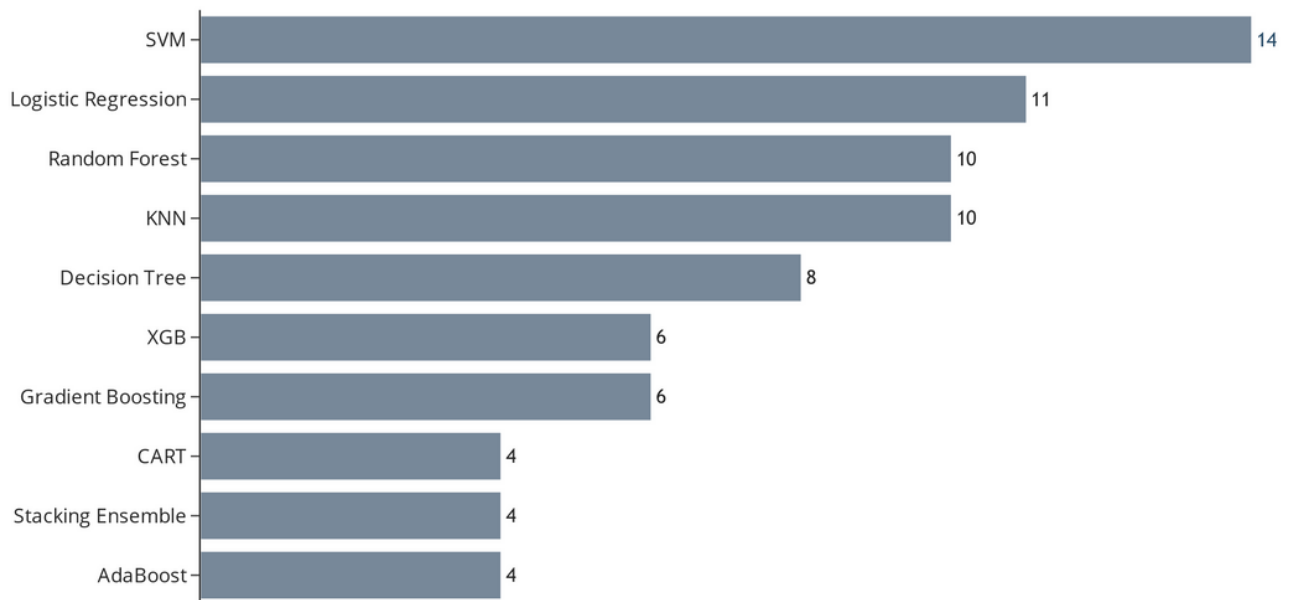


Figure 3

The total count of algorithms used in the ML studies.

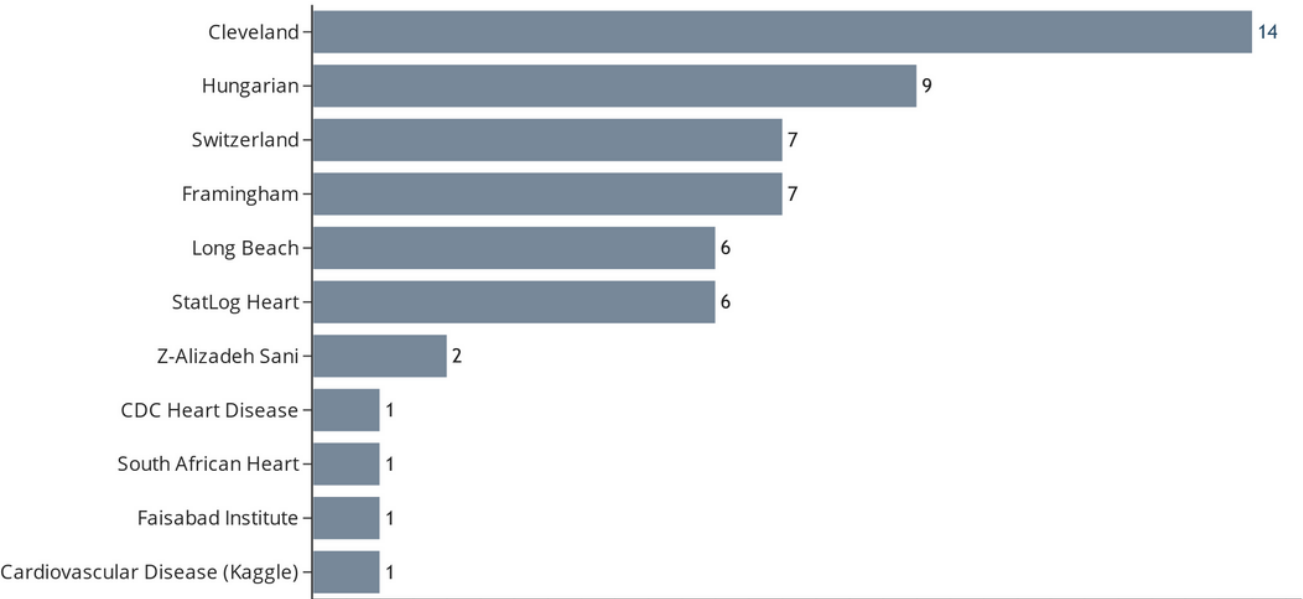


Figure 4

The total count of datasets used in the ML studies.

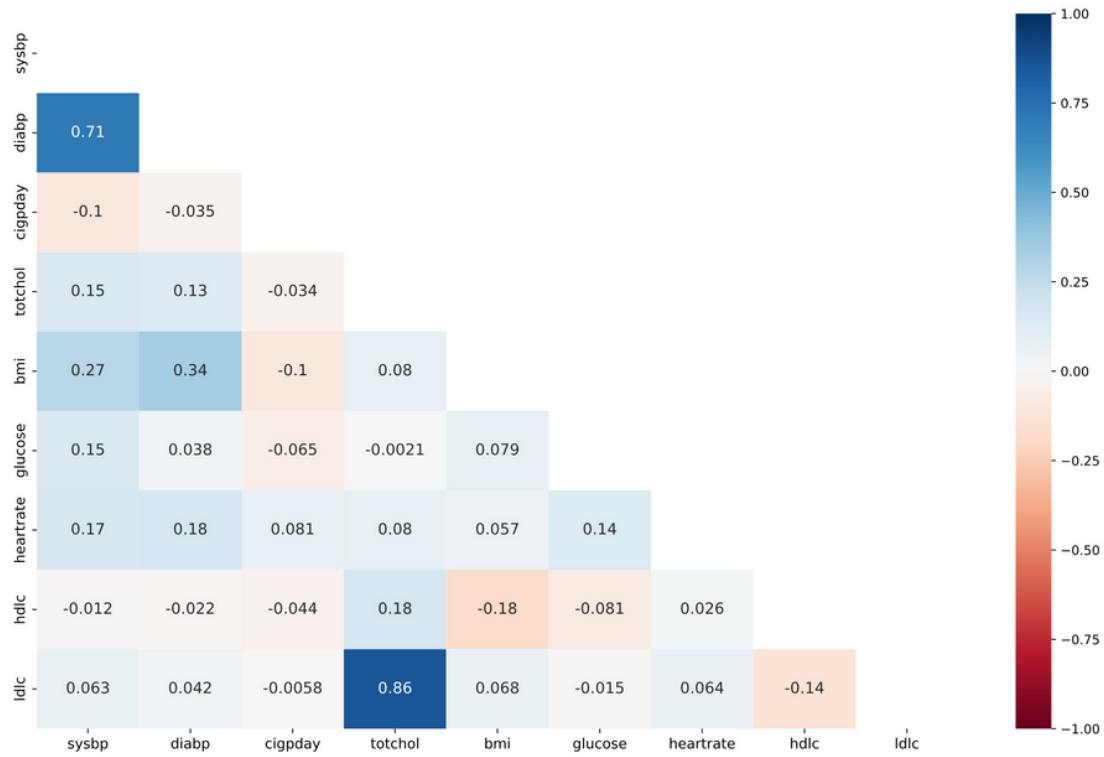


Figure 6

Pearson's correlation coefficient of continuous variables.

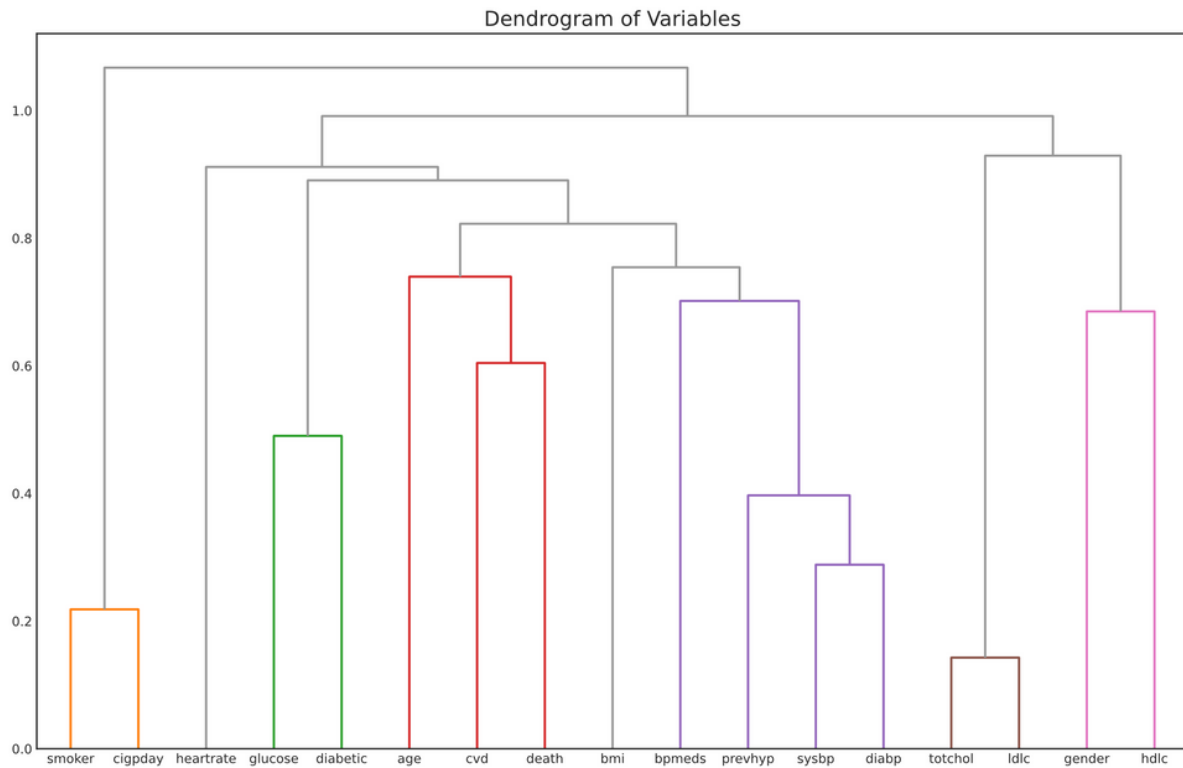


Figure 7

Dendrogram of the variables.

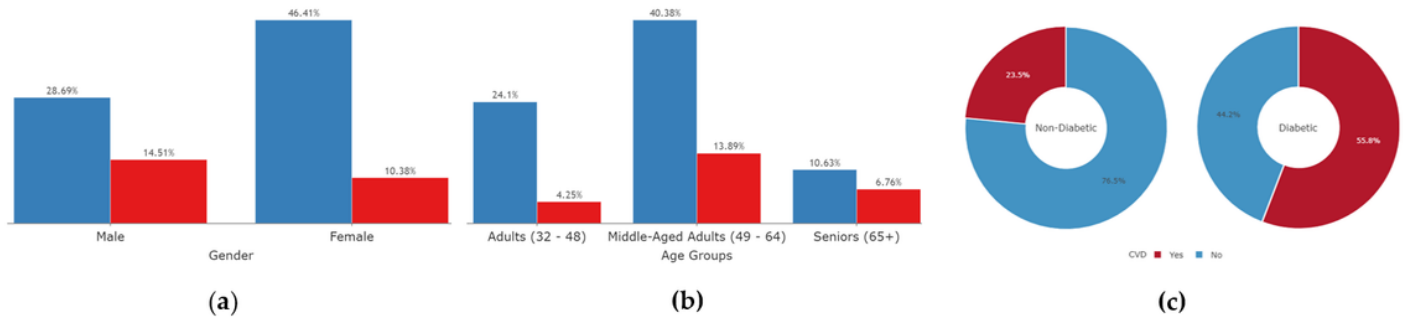


Figure 8

Series of charts for exploratory data analysis on the Framingham dataset: (a) Bar chart of gender and CVD; (b) Bar chart of three age groups and CVD; (c) Pie chart of diabetes and CVD.

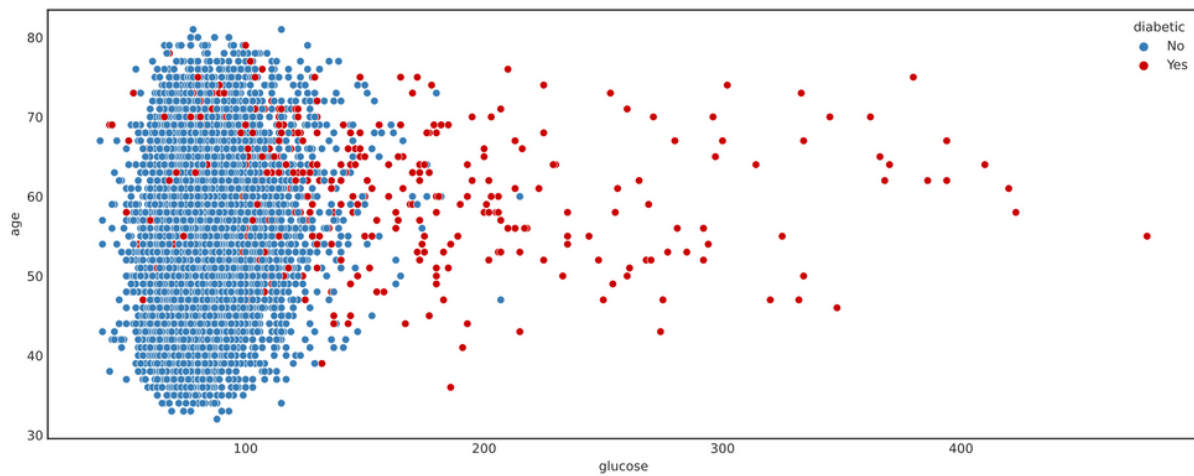


Figure 9

Scatter plot of casual serum glucose against age.

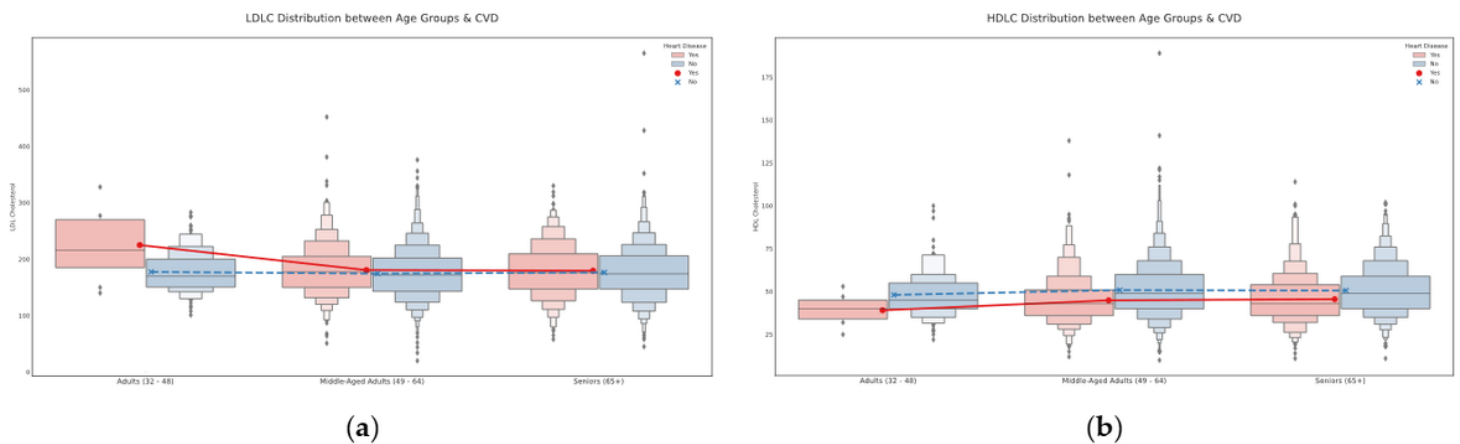
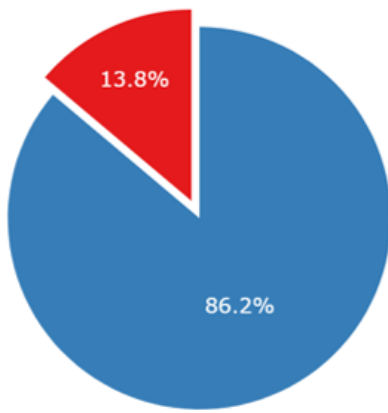
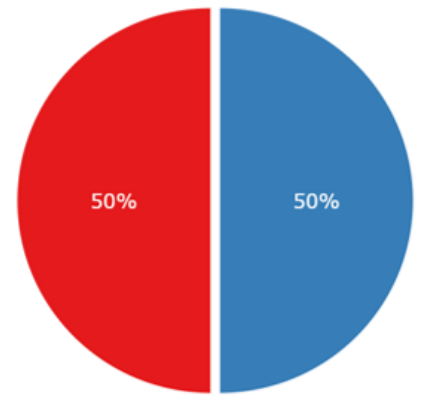


Figure 10

Series of Letter-value Plots of: **(a)** Low-Density Lipoprotein Cholesterol against Age Groups and CVD; **(b)** High-Density Lipoprotein Cholesterol against Age Groups and CVD.



(a)



(b)

Figure 11

Pie chart of the target cardiovascular disease class: **(a)** Original target class distribution in the dataset; **(b)** Desired target class distribution for training ML algorithms.

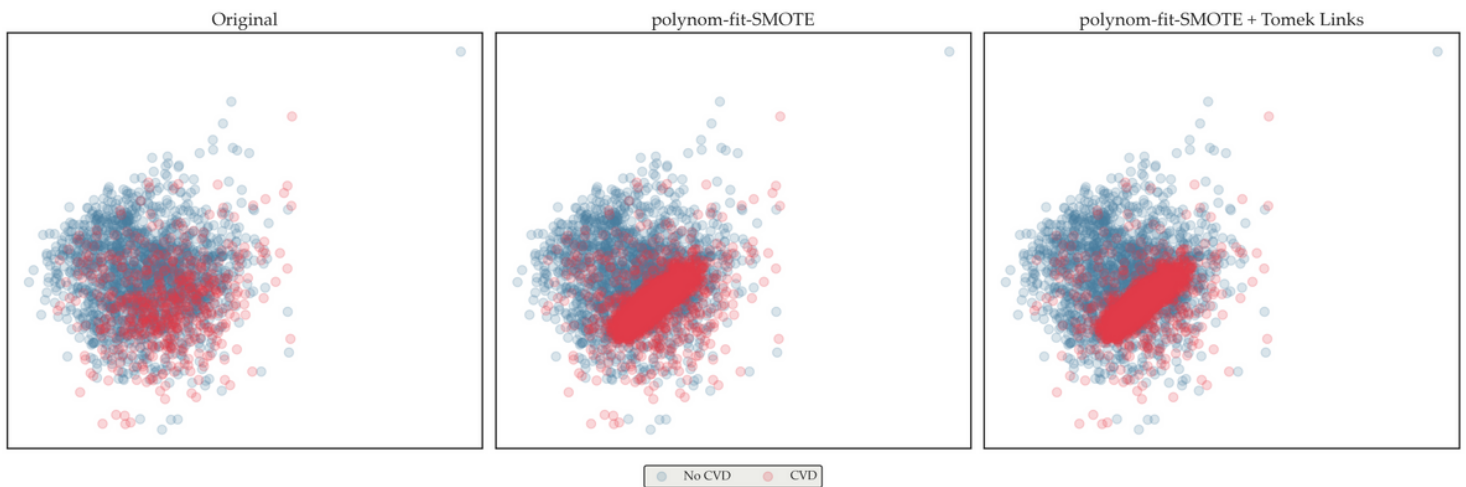


Figure 12

Stages of Dataset Transformation: From original class imbalance to balanced distribution through the application of Polynom-fit-SMOTE and Tomek Links.

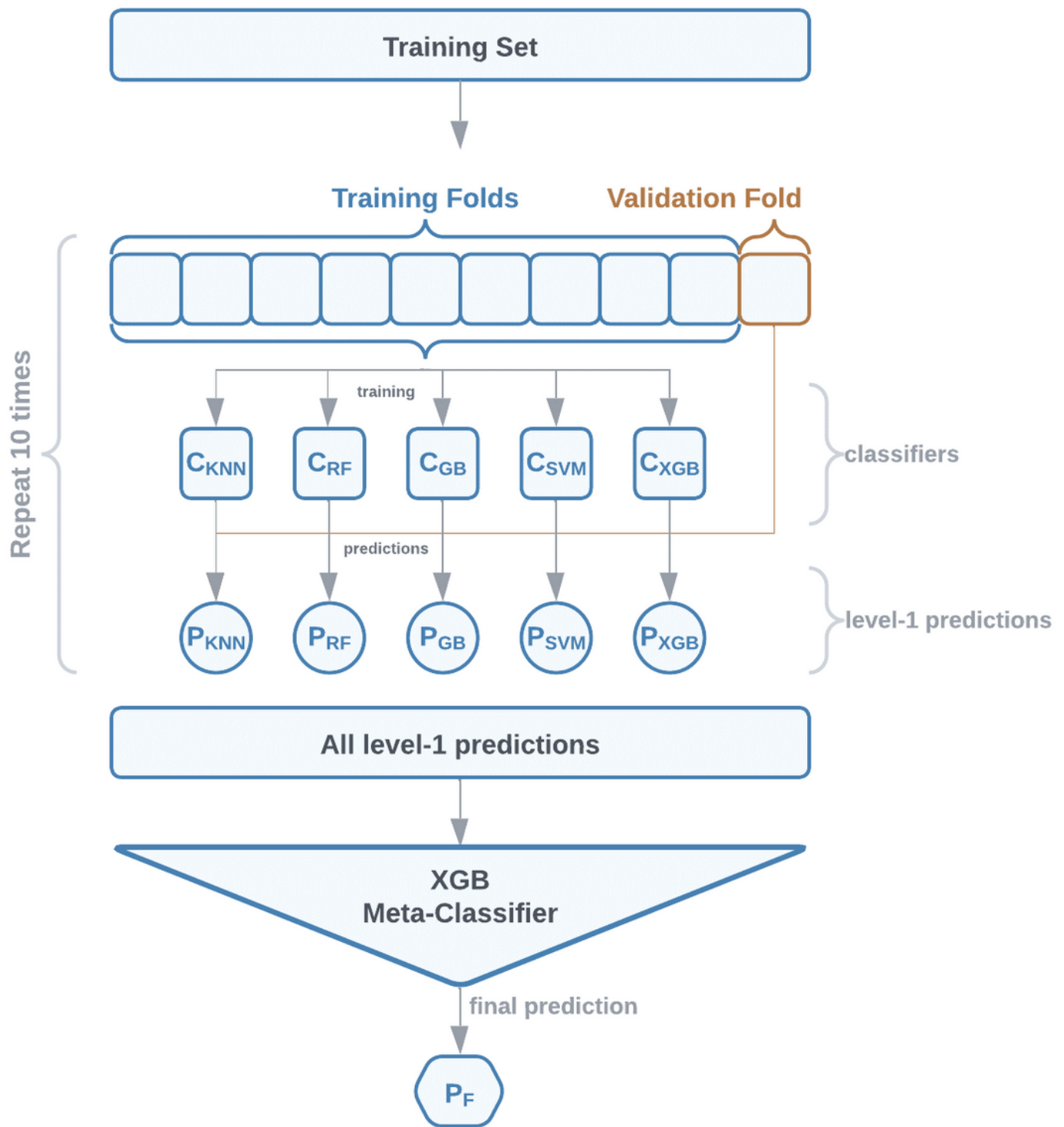


Figure 13

The process of the proposed stacked ensemble base and meta models.

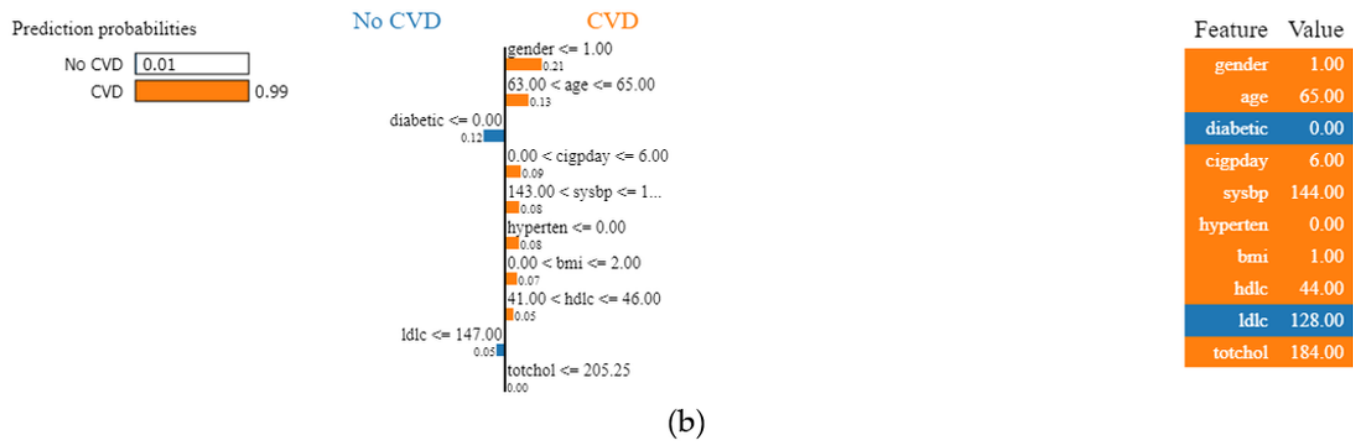
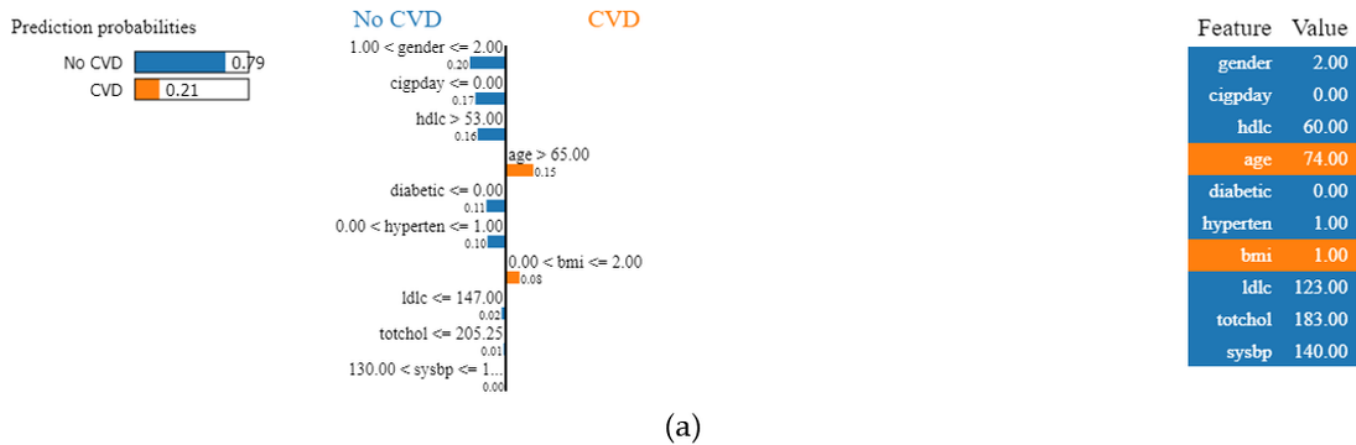
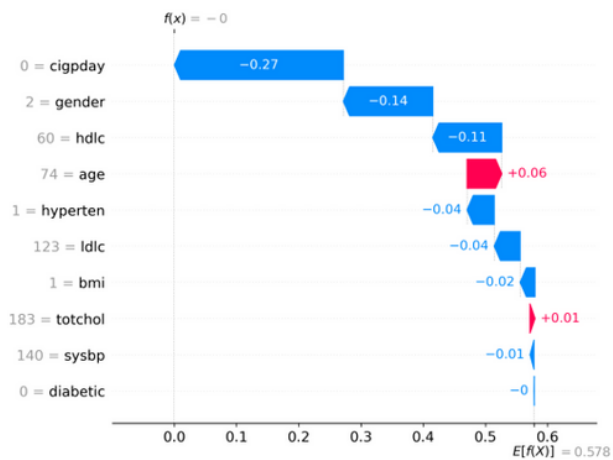
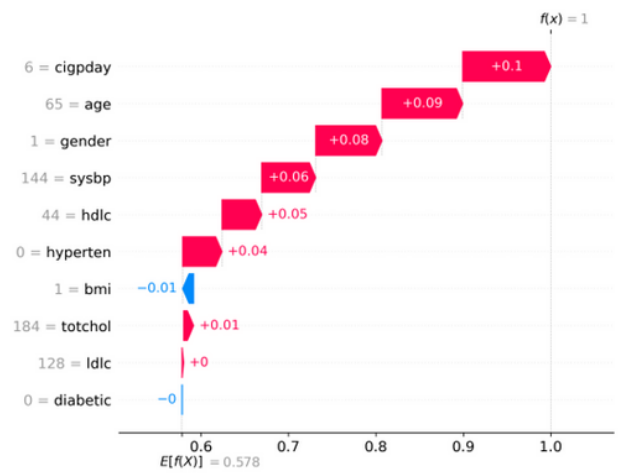


Figure 14

LIME-based explanation for two instances from the cardiovascular disease prediction model. (a) represents an instance classified as 'No CVD', displaying the top features and their corresponding contributions that led to this prediction. The key influencers here include gender, number of cigarettes smoked per day, HDL cholesterol level, and diabetic. (b) delineates an instance predicted as 'CVD'. Influential features for this prediction include gender, age, number of cigarettes smoked per day and systolic blood pressure. Each bar represents the contribution of a feature towards the prediction, promoting a comprehensive understanding of the model's decision-making process for individual instances.



(a)



(b)

Figure 15

SHAP Waterfall Plot displaying the contribution of each feature towards the prediction for two instances, one from each class (No CVD and CVD). (a) represents an instance from the No CVD class, illustrating the descending contribution of features towards the No CVD prediction. (b) pertains to an instance from the CVD class, depicting the contributing features towards the CVD prediction.

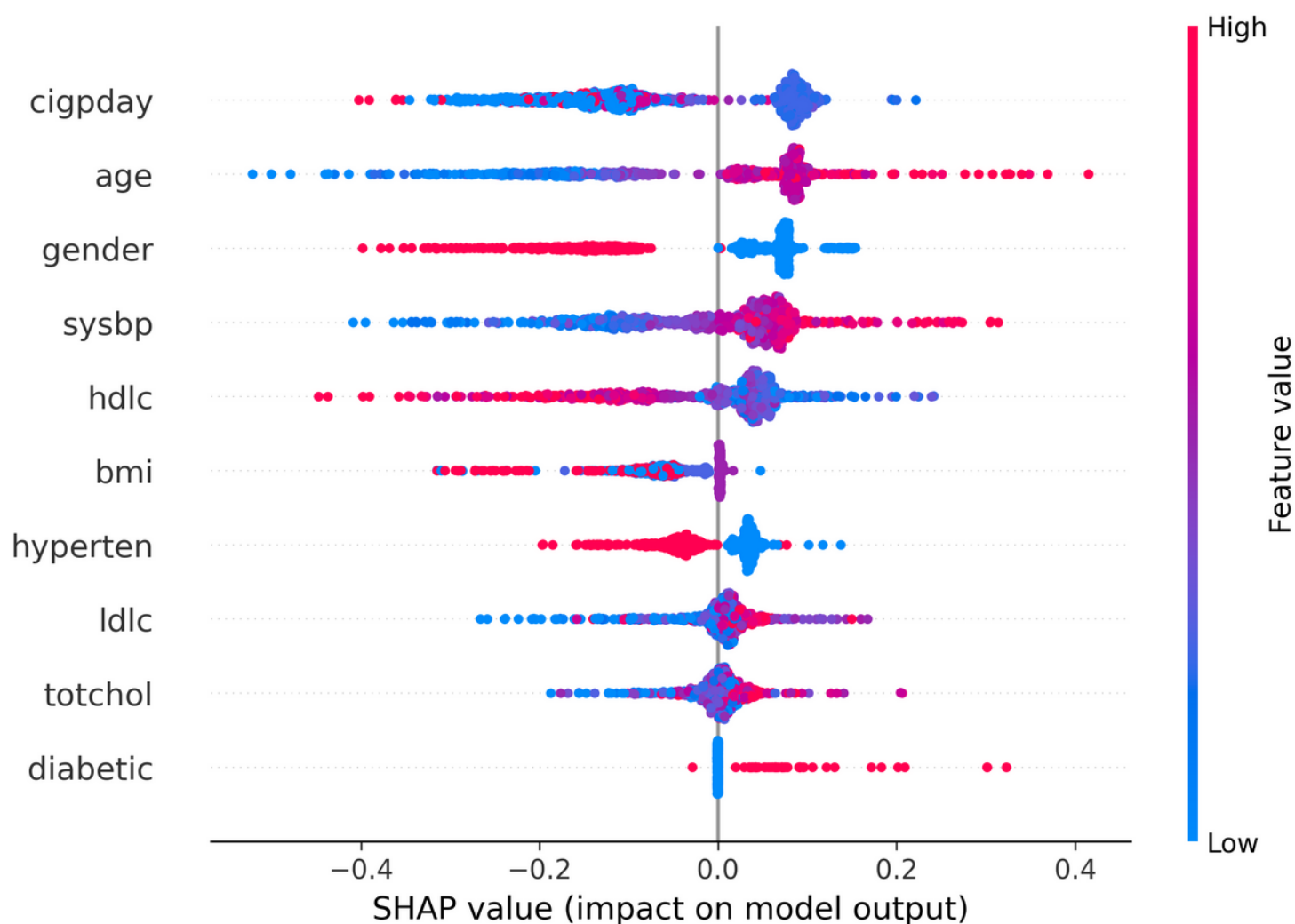


Figure 16

SHAP Summary Plot visualizing the impact of each feature on the model output. The plot demonstrates how individual features contribute towards the positive or negative prediction of CVD, capturing a range of influences across different feature values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFile1.docx](#)