Pegem Journal of Education and Instruction, Vol. 13, No. 3 , 2023 (pp. 542-556)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Machine Learning Framework for Cardiovascular Disease Prediction Using SVM

1. Tayyaba Tabassum, 2. Dr. Afroze Ansari, 3. Syeda Faqera Fatima, 4. Zameer Ahamad B

Assistant professor ,Department of Computer science and Engineering
Faculty of engineering and Technology ,Khaja Bandanawaz University
tayyaba@kbn.university
Assistant professor ,Department of Computer science and Engineering
Faculty of engineering and Technology ,Khaja Bandanawaz University
ansariafroze@kbn.university
Assistant professor ,Department of Computer science and Engineering
Faculty of engineering and Technology ,Khaja Bandanawaz University
Faqera@kbn.university
Assistant professor ,Department of Electronics and Communication Engineering
Faculty of engineering and Technology ,Khaja Bandanawaz University
Zameer@kbn.university

Abstract - Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, presenting a critical public health concern. Heart-related conditions, in particular, account for the majority of these deaths. Early and accurate prediction of CVD risk is therefore essential for timely intervention and effective management. This study focuses on assessing the risk of CVD occurrence among individuals aged over 50, across both genders, using supervised machine learning (ML) models as predictive tools. The primary objective is to evaluate and compare the performance of various ML classifiers—namely Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Random Forest—based on key metrics such as accuracy, sensitivity (recall), and Area Under the Curve (AUC) score. The experimental results demonstrate that Logistic Regression outperforms other models, achieving an accuracy of 68.4%, recall of 68.4%, and an AUC of 76.3%, thereby proving to be the most effective model for CVD risk prediction in this study.

Keywords: Cardiovascular disease, machine learning, risk prediction, logistic regression, support vector machine, AUC, recall, supervised learning

INTRODUCTION

The nomenclature "cardiovascular disease" encompasses a broad spectrum of disorders, encapsulating all pathological alterations affecting the heart and/or blood vessels. Within this category lie various conditions such as hypertension, coronary heart disease, heart failure, angina, myocardial infarction, and stroke, as elucidated by Kumar and Ramana in 2021. Over the past decade and half, cardiovascular diseases have consistently

How to cite this article: 1. Tayyaba Tabassum, 2. Dr. Afroze Ansari, 3. Syeda Faqera Fatima,4. Zameer Ahamad B. Machine Learning Framework for Cardiovascular Disease Prediction Using SVM, Vol. 13, No. 3, 2023, 542-556

Source of support: Nil
Conflicts of Interest:
None. **DOI:**
10.48047/pegegog.13.03.52

Received: 12.05.2023

Accepted: 12.06.2023 **Published:** 01.07.2023

ranked as the primary cause of mortality in developing nations, a trend anticipated to persist as projections indicate an annual death toll surpassing 20 million by 2030. A comprehensive classification of cardiovascular diseases is systematically presented in Table 1, offering a taxonomic framework for understanding their diverse manifestations and implications.

Heart diseases and stroke stand as formidable contributors to global morbidity and mortality, as underscored by Roth et al. in 2017, presenting a pressing public health challenge. Among the paramount behavioral risk factors associated with these cardiovascular afflictions, unhealthy diet, sedentary lifestyle, smoking, and excessive alcohol consumption prominently feature. The repercussions of such behavioral elements may manifest as elevated blood pressure, increased blood glucose levels, heightened blood lipids, and conditions of overweight and obesity. A detailed examination, as elucidated by Wilkins et al. in 2017 and Abdalrada et al. in 2022, outlines the principal risk factors instrumental in precipitating cardiovascular diseases.

• Elevated Body Mass Index (BMI): High BMI or obesity stands as an independent and significant risk factor for cardiovascular diseases (CVDs), necessitating heightened attention in risk assessment.

• Sedentary Lifestyle and Physical Inactivity: The adoption of a sedentary lifestyle significantly amplifies the susceptibility to CVDs, underscoring the critical role of physical activity in cardiovascular health.

• Alcohol Abuse: Excessive alcohol consumption contributes to an increased risk of cardiovascular diseases by elevating blood pressure levels and triglyceride concentrations, accentuating the need for moderation in alcohol intake.

• Nicotine Exposure and Smoking: The consumption of nicotine through smoking, as well as exposure to secondhand smoke, emerges as a pivotal factor in raising blood pressure, emphasizing the detrimental impact of tobacco use on cardiovascular health.

• Hyperlipidemia: Recognized as high cholesterol or hypercholesterolemia, hyperlipidemia denotes elevated levels of lipids in the blood, constituting a significant risk factor for cardiovascular diseases.

• Dyslipidemia: An aberrant concentration of fats or cholesterol within the blood vessels characterizes dyslipidemia, representing a deviation from the norm and a key contributor to cardiovascular risk.

• Familial Predisposition, Psychosocial Stress, and Coexisting Chronic Conditions: Factors such as family history, psychosocial stressors, and the presence of other chronic ailments like Type 2 diabetes and arterial hypertension contribute synergistically to the overall risk profile for cardiovascular diseases.

Table 01: Taxonomy of CVDs

Heart Ailments	Explanations
Ailments in Coronary arteries	Ailments of blood in veins pumping blood to cardiac muscles
Cerebro-vascular	Ailments of blood veins sending blood to nervous systems
Ailments in arteries of the periphery	Ailments of blood veins sending blood to limbs

Rheumatism	Cardiac damages and valve damages by to bacteria (streptococcal)
Congenital diseases	Deformities of cardiac muscle from birth
Thrombosis of deep veins	Clotting of blood in veins of limbs that dislodges heart and limbs
Cardiac arrest	Extreme situations that happens due to blocking of blood flowing to brains and heart

The "2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk," authored by Yancy et al. in 2013, furnishes comprehensive and detailed recommendations for the precise estimation of cardiovascular disease risk within clinical settings. In the realm of clinical practice, this guideline meticulously accounts for a multitude of factors, encompassing age, gender, race, cholesterol and blood pressure levels, diabetes status, smoking habits, and the utilization of blood pressure-lowering medications. Within the European context, the estimation of the 10-year risk factor for fatal Cardiovascular Disease (CVD) relies on distinct charts established by the European Society of Cardiology. These charts serve as invaluable tools for assessing risk across both high-risk and low-risk populations throughout Europe. Importantly, these charts offer adaptability, enabling further customization to national or regional specificities through integration with published mortality data. This dynamic adaptability ensures a nuanced and region-specific approach to risk assessment, aligning with the diverse epidemiological characteristics prevalent across Europe.

Within the scientific literature, the prediction of Cardiovascular Diseases (CVDs) risk is methodically addressed through the utilization of either dedicated risk assessment tools or the integration of machine learning methodologies. In the work by Gale et al. in 2014, the Framingham cardiovascular disease risk score is applied to assess incident frailty, drawing insights from English cohort data, particularly focusing on aging participants. Additionally, the systematic coronary risk evaluation (SCORE) emerges as a valuable tool for predicting the 10-year risk of cardiovascular death within the European context. Alternatively, the QRISK model is recommended for predicting the composite outcome involving coronary heart disease and ischemic stroke. Other scholarly contributions, such as those presented by Mohan et al. in 2019 and Yang et al. in 2020, venture into the realm of machine learning techniques. These endeavors leverage advanced computational methodologies with the explicit aim of predicting potential risks associated with Cardiovascular Diseases (CVDs).

Machine Learning (ML), a subset of Artificial Intelligence (AI), proves itself as a potent asset within the medical domain, showcasing its prowess in disease prediction. The work by Dinesh et al. in 2018 delves into data-driven methodologies specifically tailored to predict both diabetes and cardiovascular diseases, with a particular emphasis on the latter within the cardiovascular medicine domain, as elucidated by Haq et al. in 2018. The primary objective is to discern predictive data patterns and identify high-risk Cardiovascular Disease (CVD) cohorts, particularly among the elderly demographic. This initiative extends to the development of personalized risk models, an integral component of the predictive AI tools slated for integration into the SmartWork system, outlined by Kocsis et

al. in 2019, and the GATEKEEPER systems. The method for CVD risk prediction is meticulously devised and independently validated, leveraging both publicly available datasets and pilot data from concurrent projects. The integration of ML models into the Long-term Risk Prediction tools within the SmartWork system aspires to architect a smart, age-friendly living and working environment tailored for office workers. On a parallel trajectory, the GATEKEEPER system aims to preserve the health of older individuals in their homes, preemptively addressing the onset of CVD, Type 2 Diabetes Mellitus (T2DM), high cholesterol, hypertension, and chronic conditions like Chronic Obstructive Pulmonary Disease (COPD) associated with Metabolic Syndrome (MetS), as expounded by Fazakis et al. in 2021, Dritsas et al. in 2021, and Hussain et al. in 2021, respectively.

In light of Metabolic Syndrome (MetS) serving as a convergence of risk factors propelling Cardiovascular Disease (CVD) and Type 2 Diabetes (T2DM) development, as outlined by Hoyas and Leon-Sanz in 2019, this study takes an initial stride by presenting a methodology aimed at accurately identifying individuals with a long-term risk of diagnosed CVD. The evaluation entails assessing the classification performance of diverse Machine Learning (ML) models on instances from a CVD dataset. Models exhibiting optimal recall, emphasizing heightened sensitivity, and Area Under Curve (AUC) signify adept prediction of the CVD class. The principal contribution lies in a comparative analysis of distinct ML models on a balanced dataset, culminating in the endorsement of a Logistic Regression model for long-term CVD risk prediction. Subsequent sections elucidate the procedural steps, commencing with the methods for long-term CVD risk prediction in Section 2, followed by an exploration of dataset features in Section 3, and a detailed exposition of pre-processing steps for training and testing dataset design and feature ranking in Section 4. Section 5 unfolds the experimental setup, presenting the classification performance of ML techniques. The paper concludes in Section 6, offering insights into future directions stemming from the current outcomes.

METHODOLOGIES OF MACHINE LEARNINGS

Within the medical domain, the pervasive utilization of data science, particularly machine learning, is notable for its extensive application in risk analysis pertaining to various chronic conditions. These models predominantly focus on discerning the most pertinent factors conducive to long-term risk prediction, with the overarching objective of averting severe health complications attributable to specific symptoms. The core purpose lies in fortifying healthcare management through strategic insights derived from the comprehensive analysis of risk factors, thereby optimizing proactive interventions and preventive measures. This proactive stance aligns with the broader paradigm of predictive and preventive healthcare, wherein the leveraging of advanced computational methodologies facilitates a nuanced understanding of risk dynamics. The emphasis on long-term risk prediction underscores the strategic importance of early identification and intervention, contributing to enhanced patient outcomes and streamlined healthcare resource allocation. In essence, the integration of data science, particularly machine learning, emerges as an indispensable tool in the medical arsenal, navigating the complexities of chronic condition risk analysis to elevate healthcare practices into

a realm of proactive, data-driven decision-making.

Within the confines of this investigation, the focus is directed towards unveiling the predictive efficacy of four distinct machine learning models. Specifically, the Naive Bayes, Support Vector Machine (SVM), Logistic Regression, and Random Forest models are harnessed to gauge and project the long-term risk associated with the diagnosis of cardiovascular disease in older individuals. The selection of these models is strategic, aligning with the pursuit of a comprehensive understanding of the forecasting landscape by encompassing diverse machine learning paradigms. Each model, with its unique algorithmic underpinnings, contributes distinct computational methodologies to the overarching goal of risk estimation. The Naive Bayes model, characterized by its probabilistic approach, operates in tandem with SVM, a model renowned for its efficacy in classification tasks through the identification of optimal hyperplanes. Simultaneously, Logistic Regression, a staple in statistical modeling, and the ensemble-based approach of Random Forest bring complementary perspectives to the task at hand. By synthesizing insights from these machine learning models, this study endeavors to provide a nuanced and multifaceted evaluation of their forecasting performance in the domain of long-term cardiovascular disease risk assessment for the elderly demographic.

The dataset undergoes a bifurcation, delineating a training set characterized by a magnitude denoted as M, and a test set with a size denoted as N. Within the dataset, a categorical variable denoted as c assumes the role of capturing the class label pertaining to a given instance i. In the specific context of this research endeavor, the investigative problem manifests with two potential classes, articulated as c = "CVD" or "Yes," and c = "Non-CVD" or "No." These classes encapsulate the binary nature of the problem, indicative of whether an instance belongs to the category of cardiovascular disease (CVD) or not. Furthermore, the features vector corresponding to an instance i finds representation through the vector $f_i = [f_{i1}, f_{i2}, f_{i3}, ... , f^{in}]^T$, with M significantly surpassing n in magnitude. This features vector encapsulates the multidimensional aspects of the instance, where each element f_{i1}, f_{i2}, f_{i3}, ..., fin corresponds to a specific feature contributing to the overall characterization of the instance. The transposition denoted by T underscores the vector's orientation, emphasizing the diverse array of features considered within the dataset, pivotal for subsequent analysis and model training.

The primary objective entails the attainment of elevated recall or sensitivity and a commendable Area Under Curve (AUC) utilizing the paradigm of supervised machine learning. This implies the accurate prediction of the cardiovascular disease (CVD) class. The methodological framework for CVD prediction encompasses a repertoire of models, each elucidated in subsequent sections for comprehensive understanding and clarity.

Naive Bayes

A straightforward classifier known as Naive Bayes, as detailed by Dinesh et al. (2018), operates on the foundational principles of the Bayes theorem. This classifier makes a fundamental assumption of highly independent attributes, often referred to as predictors, with the overarching objective of maximizing probabilities. The crux of interest within this framework lies in the determination of posterior probabilities, a key facet in the

probabilistic model's decision-making process.

$$P(c|f_{i1}, \dots, f_{in}) = \frac{P(f_{i1}, \dots, f_{in}|c)P(c)}{P(f_{i1}, \dots, f_{in})} \quad (1)$$

In the above equation $P(f_{i1}, \dots, f_{in}|c) = \prod_{j=1}^{n} P(f_{ij}|c)$ represents the predictor probability class that is given and $P(f_{i1}, \dots, f_{in})$ is the probabilities of predictors (prior) and $P(c)$ is the class probability (prior). The information of testing has been classified on the basis of association probability.

$$\hat{c} = \arg\max P(c) \prod_{j=1}^{n} P(f_{ij}|c) \; where \; c \in \{CVD, Non - CVD\}$$

SVM

The Support Vector Machine (SVM), a potent machine learning algorithm renowned for its high classification efficacy, finds applications in the medical domain. Specifically, its utility extends to resolving binary classification predicaments, aligning with the current investigation's objectives. SVM adeptly handles both linear and non-linear classification challenges through the integration of Kernel functions, facilitating the mapping of non-linear data into a higher-dimensional feature space. In the linear scenario, instances undergo segregation by a hyperplane, denoted as the support vector, characterized by the equation wT f+b. Here, w represents an n-dimensional coefficient vector normal to the hyperplane, and b signifies an offset value from the origin. The derivation of w and b involves intricate calculations, culminating in the formulation of a linear discriminant function expressed as g(w) = sign(wT f+b).

Logical Regression

Prediction of Logitical Regressions predict label of classes of features of inputs on the basis of these values by the use of model of regression of binary nature. On assumption of $p = P(c = 'CVD), log_a\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{m} f_{ji}\beta_i, with \; (\beta_1, \beta_2, \dots, \beta_n)$ are the weight of regressions that are combines with row vector features $f_i, i = 1,2, \dots, M$. By the isolation of p, In case it is more than thresholds is slated to CVDs classes i.e. Yes; in other cases the setting is done to Non-CVDs classes. The coefficients of β value for algorithms of logistical regressions have been estimated from sets of data to train by utilization of estimators of Maximum Likelihood.

Randomized Forests

The Random Forest method, as elucidated by Yang et al. (2020), constitutes a supervised learning algorithm designed for categorical classification, distinguishing instances as either pertaining to CVD or Non-CVD categories. This algorithm operates as an ensemble of decision trees, each constructed based on distinct data samples. The strength of the resulting forest is directly proportional to the number of decision trees incorporated, contributing to heightened robustness. The synergy of multiple decision trees within the Random Forest algorithm converges in a comprehensive prediction mechanism. This amalgamation involves individual predictions from each tree, ultimately culminating in the determination of the most optimal outcome through a process of majority voting. This collective decision-making strategy enhances the reliability and accuracy of the overall classification process, solidifying the Random Forest method as a robust tool for discerning cardiovascular disease status.

Explanation of Datasets

The creation of the training and test datasets for the model predicting cardiovascular disease risk drew from the CVDs dataset, an open-source dataset accessible from Kaggle, encompassing 70,000 participants. This dataset boasts equilibrium between healthy and CVD-diagnosed participants, featuring 11 distinctive attributes—comprising 4 demographic, 4 examination, and 3 social history variables. The demographic variables encompass age (in years) and gender, while examination variables include weight (measured in kilograms) and height (in square meters), utilized to compute Body Mass Index ($BMI = \frac{weight}{(height)^2}$). Additional attributes encompass cholesterol and glucose levels categorized as normal, above normal, or well above normal, lifestyle factors such as physical activity, drinking, and smoking habits, each denoted with binary values (yes or no), and systolic and diastolic blood pressure readings (in millimeters of mercury). Notably, blood pressure measurements were taken during medical examinations. To identify linear correlations between these features and the target class, Pearson's correlation coefficient (CC), as defined by Mukaka (2012), was employed.

$$C = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(x_i - \bar{x})^2} \quad (2)$$

The outcomes displayed in Table 2 were computed with an alpha value set at 0.01, ensuring a 99% confidence interval. Examination of Table 2 reveals a substantial linear interdependence between BMI and weight attributes, with a coefficient of 0.7620. Additionally, a moderate correlation of 0.4520 is evident between glucose and cholesterol. A lesser correlation value of 0.3400 is identified in the relationship between smoking and alcohol consumption. Delving into the CVD class and features associated with age and cholesterol, the dataset portrays low correlation values, denoted by the interval $0.2 < C < 0.39$.

In Tables 4-8, the presented distributions delineate the consideration of diverse feature combinations and target classes (yes or no) among the selected participants. The gender-based analysis in Table 3 reveals that 32.31% of individuals afflicted with Cardiovascular Disease (CVD) are females, a proportion nearly double that of their male counterparts with CVD. Transitioning to Table 4, an observation emerges that approximately 70% of the overall participants fall within the elderly demographic (age ≥ 50), aligning with the focal demographic of this study. Among these older participants, about 40% are categorized under the CVD class.

Table 5 delineates a classification based on systolic and diastolic blood pressure values. The data indicates that 45% of participants have received a CVD diagnosis, specifically falling into the hypertensive categories (I or II). Moving to Table 6, the distribution of participants into Non-CVD and CVD classes is contingent on smoking and alcohol consumption features. It is noteworthy that a small percentage (1.18%) of individuals diagnosed with CVD concurrently engage in smoking and alcohol consumption.

The comprehensive examination of participant characteristics across these tables provides a nuanced understanding of the interplay between demographic factors and CVD prevalence. The observed patterns shed light on gender disparities, age-related correlations, and the influence of factors such as blood pressure, smoking, and alcohol consumption in the context of Cardiovascular Disease.

Table 3: Examining the dataset involves exploring how individuals, both healthy and diagnosed with Cardiovascular Disease (CVD), are distributed across gender groups.

	CardioVascular Disease		
Sex	Total	Yes	No
Male	65.49%	18.55%	35.04%
Female	34.51%	31.40%	15.01%
	100%	49.95%	50.05%

Table 4: Analyzing the dataset involves studying the distribution of individuals categorized as healthy and those diagnosed with heart disease across different age groups.

	CardioVascular Disease		
Age in years	Total	Yes	No
30-34	0.005%	0.005	0.000%
35-39	0.61%	0.12%	0.49%
40-44	14.78%	4.68%	10.10%
45-49	13.22%	5.22%	8.00%
50-54	28.43%	12.44%	15.99%

Table 5: The distribution of individuals labelled as healthy and those diagnosed with heart disease categorized by their respective blood pressure levels within the dataset.

	CardioVascular Disease		
Blood Pressures	Total	Yes	No
Slightly high	4.84%	2.04%	2.80%
High	59.82%	32.34%	27.48%
Very High	25.31%	10.20%	15.11%
Normal	13.67%	2.77%	9.90%
Net	100%	49.41%	50.59%

Table 6: The allocation of individuals categorized as healthy and those diagnosed with cardiovascular disease (CVD) concerning their smoking and alcohol-related characteristics within the dataset.

Smoking		CardioVascular Disease		
		Total	Yes	NO
Negative		90.95%	45.45%	45.5%
Alcoholic	Negative	89.14%	44.30%	44.84%
	Positive	2.75%	1.56%	1.19%
Positive		6.97%	4.08%	2.89%
Alcoholic	Negative	6.27%	3.17%	3.10%
	Positive	2.75%	1.26%	1.49%
total		100%	48.66%	51.34%

Table 7: The distribution of individuals classified as healthy and those diagnosed with Cardiovascular Disease (CVD) in relation to their glucose levels is being examined within the dataset.

	Cardiovascular Disease		
Sugar Levels	Total	Yes	No
Normal	74.64%	33.12%	41.52%
Just exceeding Normal	12.63%	8.22%	4.41%
Greater than Normal	11.91%	8.81%	3.10%
	100%	49.51%	50.49%

Table 8: In the dataset, the arrangement of individuals according to their health status (CVD), is examined with regard to the characteristic of cholesterol levels.

	CVD		
Level of Lipids in blood	Total	Yes	No
Just above Normal	13.57%	8.51%	5.06%
Normal	13.89%	8.91%	4.96%
Greater than Normal	11.31%	8.03%	3.28%
	100%	48.61%	51.39%

Tables 7 and 8 delineate the distribution of participants across Non-CVD and CVD categories based on glucose and cholesterol features. It is essential to acknowledge that a marginal proportion of participants (1.18%), diagnosed with CVD, concurrently engage in smoking and alcohol consumption. Additionally, there is an absence of information regarding the quantity of consumption, impeding a comprehensive understanding of the potential deleterious impact of participants' habits on their health.

Table 09 - Within the dataset, the categorization of individuals into distinct groups based on their health status, diagnosed with Cardiovascular Disease (CVD), is explored concerning their Physical Activity and BMI classifications.

Body Activity		**CVD**		
		TOTAL	NO	YES
No		19.63%	9.11%	10.52%
Body Mass Index class	Healthy	7.54%	3.22%	4.32%
	Obese I	3.58%	1.32%	2.26%
	Obese II	1.73%	0.61%	1.12%
	Obese III	0.8%	0.58%	0.22%
	Overweight			
	Below-weight	0.21%	0.13%	0.08%
Yes		81.55%	40.03%	41.52%
Body Mass Index class	Healthy	30.01%	20.91%	9.20%
	Obese I	14.56%	10.28%	4.28%
	Obese II	5.13%	2.03%	3.10%
	Obese III	1.98%	0.91%	1.07%

	Overweight	30.19%	15.01%	15.90%
	Below-weight	0.17%	0.11%	0.06%
		100%	51.37%	48.63%

Table 9 documents the distribution of participants across Non-CVD and CVD categories based on their physical activity and BMI classifications. An observation indicates that 11.55% of participants, despite falling into the healthy BMI category and maintaining physical activity, have received a diagnosis of Cardiovascular Disease (CVD). Furthermore, within the obese and overweight categories, the count of participants diagnosed with CVD is roughly equivalent. In both instances, there exists a lack of information on exercise-related attributes (intensity, duration, frequency, type), hindering a comprehensive understanding of the extent to which physical activity may be beneficial for the health of individuals diagnosed with CVD.

PROCESSING OF DATA

The assessment of data preprocessing was conducted using the Stata V.14 toolkit, a versatile statistical software package developed by StataCorp for data manipulation, visualization, statistics, and automated reporting. The formulation of the training and testing dataset, derived from the CVDs dataset, involved a comprehensive examination of attributes correlated and potentially relevant to Cardiovascular Diseases identification. To ensure uniformity across participant attributes, harmonized variables were created by transforming all numeric values into nominal values according to predefined attribute rules. Additionally, the class, based on the CVD's dataset cardiovascular diseases feature, was established, resulting in a balanced distribution containing a total of 70,000 observations.

To assess the correlation of all attributes with the class, a feature selection method was employed, utilizing a ranking system based on the relevance of each attribute to the specific class. Specifically, a feature selection method, a variation of Random Forests (Genuer et al., 2010), was employed. According to this method, attributes are ranked using the Gini importance score of the model's trees. The Gini index (Sundhari, 2011) is calculated as follows.

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

In this context, the parameter c denotes the number of classes, and pi represents the relative frequency of class i within the dataset. Specifically, for our case, the parameter c equals 2, signifying the classes of CVDs and non-CVDs.

EVALUATIONS OF EXPERIMENTS

(a) SET-UP OF EXPERIMENTS

The experimental evaluations utilized the WEKA platform, a JAVA-based data mining toolkit acknowledged as free software under the GNU General Public License. WEKA offers an extensive library of methods and models encompassing classification, clustering, prediction, and feature selection. For this particular study, the dataset underwent a division into two segments, allocating 30% for testing and 70% for training. The initial step involved randomizing the dataset to establish a random permutation. Subsequently, the RemovePercentage method was applied with a 30% parameter, saving the resultant dataset as the training set. Simultaneously, the same filter, employing the

invertSelection option, was applied to extract the remaining 30%, constituting the testing dataset.

Table – 10: The efficacy of machine learning models in predicting the risk of cardiovascular diseases (CVDs) was assessed.

Algorithms	AUC	Accurateness	Recall
Logistical Regressions	75.11%	70.13%	75.34%
Randomized Forests	77.83%	69.33%	71.44%
Naïve Bayes	71.62%	60.21%	58.31%
Support Vector Machines	78.91%	69.87%	68.56%

F1-Score: represents weighted mean relating Recalls and Precision. The measure of F1-score have been used to evaluate the effective ness of different methods in detection of cardio-vascular dieases.

$$F1 - score = 2 \times \frac{(Recalls \times Precision)}{(Recalls + Precision)}$$

Table 1 Result enumerating F1-score classifier

Classifiers	F1- Scores
Logistical Regressions	71%
Randomized Forests	72%
Naïve Bayes	68%
Support Vector Machines	73%

(b) RESULTS OF EXPERIMENTS

In this particular experimental setup, the assessment of model performance focused on metrics such as accuracy, recall, and AUC. Additionally, the 10-fold cross-validation procedure was applied to appraise four distinct classifiers: Naive Bayes, SVMs, Logistic Regression, and Random Forests. The evaluation of these models is executed based on the confusion matrix, explicitly encompassing True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). To be more precise, the metrics utilized for computing accuracy and recall are explicitly defined as *Accuracy*.

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \quad (4)$$

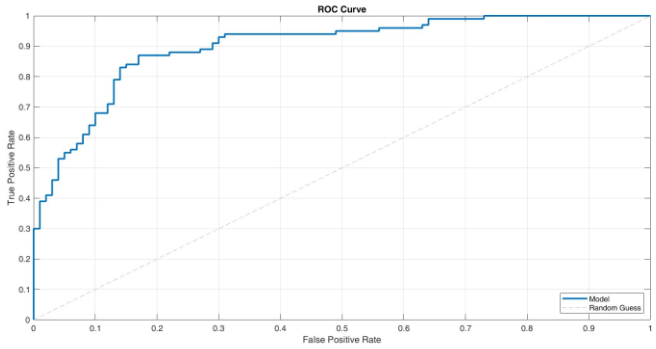
$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Table 10 illustrates the performance outcomes of the machine learning algorithms concerning the utilized metrics.

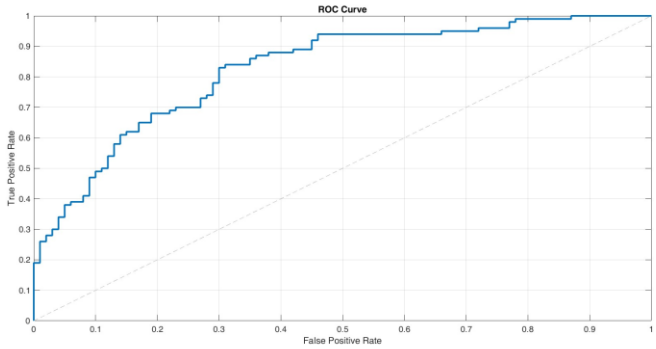
The Logistic Regression model exhibits promising predictive capabilities in the context of cardiovascular diseases, showcasing superiority in terms of accuracy, recall, and AUC compared to other machine learning models. The heightened AUC signifies an enhanced ability of the model to effectively discriminate between CVD and Non-CVD classes. Specifically, the AUC metric indicates a 78.4% probability that the Logistic Regression model will adeptly distinguish between these two classes.

Table 2 Comparative values of ROC-AUC values for the different Classification methods considered for comparison

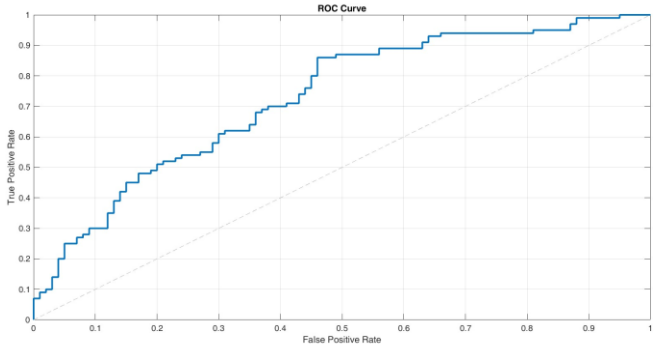
Classifiers	ROC-AUC Scores
Logistical Regressions	0.7511
Randomized Forests	0.7783
Naïve Bayes	0.7162
Support Vector Machines	0.7891



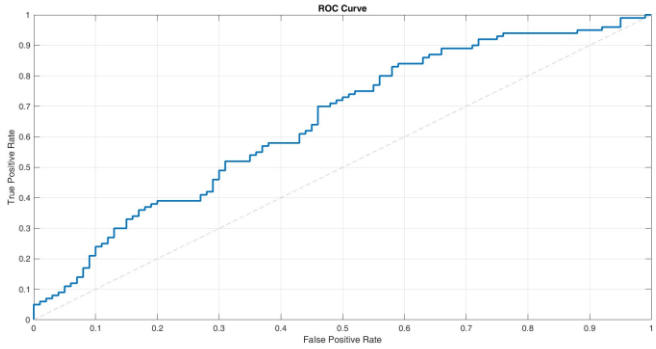
(a) AUC=0.7891 using proposed SVM technique



(b) AUC=0.7783 for Randomized Forest technique



(c) AUC=0.7511 for Logistical Regressions



(d) AUC=0.7162 for Naïve Bayes technique

Figure 1 AOC-ROC curves for techniques mentioned in Table 2

CONCLUSIONS

Within the scope of this investigative inquiry, a meticulously structured supervised machine learning approach was applied to evaluate the prolonged likelihood of cardiovascular disease (CVD) occurrence. The study's findings carry

valuable insights from a clinical perspective, potentially aiding healthcare practitioners in deciphering data and implementing optimal algorithms tailored to the dataset (Al'Aref et al., 2019).

As part of an ongoing research initiative, the initial phase involves the development of several traditional models to scrutinize data quality and pinpoint the model demonstrating the most superior predictive performance. The present outcomes will serve as a foundation for shaping the subsequent stages of the research, identifying optimal models, and refining performance metrics. The evaluation results showcased comparable accuracy and recall, a consequence of the balanced distribution of participants across two classes. An avenue showing promise for elevating the achieved outcomes (accuracy, recall, AUC) involves incorporating deep learning models and techniques (Swathy and Saruladha, 2021), leveraging their ability to delineate intricate decision boundaries for a more nuanced fit to the training data. Ultimately, the research aims to concentrate its analysis on i) anomaly detection techniques to pinpoint instances with inaccurate values and ii) dimensionality reduction techniques to optimize the efficacy of machine learning models.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovasculardiseases-(cvd)
[2] Kelly, B. B., & Fuster, V. (Eds.). (2010). Promoting cardiovascular health in the developing world: a critical challenge to achieve global health. National Academies Press.
[3] Poirier, Paul, et al. "Obesity and cardiovascular disease: pathophysiology, evaluation, and effect of weight loss: an update of the 1997 American Heart Association Scientific Statement on Obesity and Heart Disease from the Obesity Committee of the Council on Nutrition, Physical Activity, and Metabolism." Circulation 113.6 (2006): 898-918.
[4] Bhatnagar, Prachi, et al. "Trends in the epidemiology of cardiovascular disease in the UK." Heart 102.24 (2016): 1945-1952.
[5] Beunza, Juan-Jose, et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)." Journal of biomedical informatics 97 (2019): 103257.
[6] Zhao, Lina, et al. "Enhancing Detection Accuracy for Clinical Heart Failure Utilizing Pulse Transit Time Variability and Machine Learning." IEEE Access 7 (2019): 17716-17724.
[7] Borkar, Sneha, and M. N. Annadate. "Supervised Machine Learning Algorithm for Detection of Cardiac Disorders." 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 2018.
[8] Omar Boursalie, Reza Samavi, Thomas E. Doyle. "M4CVD: Mobile Machine Learning Model for Monitoring Cardiovascular Disease." Procedia Computer Science, 63 (2015): 384-391.
[9] Chen, Rui, et al. "Using Machine Learning to Predict One-year Cardiovascular Events in Patients with Severe Dilated Cardiomyopathy." European Journal of Radiology (2019).
[10] Dhar, Sanchayita, et al. "A Hybrid Machine Learning Approach for Prediction of Heart Diseases." 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018.
[11] Dinesh, Kumar G., et al. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms." 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE,
2018.
[12] Mezzatesta, Sabrina, et al. "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis." Computer Methods and Programs in Biomedicine 177 (2019):

9-15.
[13] Terrada, Oumaima, et al. "Classification and Prediction of atherosclerosis diseases using machine learning algorithms." 2019 5th International Conference on Optimization and Applications (ICOA). IEEE, 2019.
[14] Alić, Berina, Lejla Gurbeta, and Almir Badnjević. "Machine learning techniques for classification of diabetes and cardiovascular diseases." 2017 6th Mediterranean Conference on Embedded Computing (MECO).
IEEE, 2017.
[15] Awan, Shahid Mehmood, Muhammad Usama Riaz, and Abdul Ghaffar Khan. "Prediction of heart disease using artificial neural network." VFAST Transactions on Software Engineering 13.3 (2018): 102-112.
[16] Fathalla, Karma M., et al. "Cardiovascular risk prediction based on Retinal Vessel Analysis using machine learning." 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2016.
[17] Gjoreski, Martin, et al. "Chronic Heart Failure Detection from Heart Sounds Using a Stack of Machine-Learning Classifiers." 2017 International Conference on Intelligent Environments (IE). IEEE, 2017.
[18] Metsker, Oleg, et al. "Dynamic mortality prediction using machine learning techniques for acute cardiovascular cases." Procedia Computer Science 136 (2018): 351-358.
[19] Balasubramanian, Vineeth Nallure, et al. "Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure." 2009 36th Annual Computers in Cardiology Conference (CinC). IEEE, 2009.
[20] Groselj, C., et al. "Machine learning improves the accuracy of coronary artery disease diagnostic methods." Computers in Cardiology 1997. IEEE, 1997.
[21] Zhou, Yijiang, et al. "Machine Learning-Based Cardiovascular Event Prediction For Percutaneous Coronary Intervention." Journal of the American College of Cardiology 73.9 Supplement 1 (2019): 127.
[22] Singh, Manpreet, et al. "Building a cardiovascular disease predictive model using structural equation model & fuzzy cognitive map." 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2016.
[23] Alaa, Ahmed M., et al. "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants." PloS one 14.5 (2019): e0213653.