Research Article
Open Access

Cardiovascular disease prediction with machine learning techniques

Abstract

Cardiovascular disease (CVD) remains the leading cause of death globally. In search of advanced techniques for early detection of CVD, recent research has increasingly focused on using machine learning (ML) methods to improve the accuracy and timeliness of diagnosis. A multifactorial machine learning approach offers a comprehensive solution for cardiovascular disease detection, using vast and diverse datasets to develop predictive models that outperform traditional methods.

This paper provides a comprehensive examination of various machine learning approaches and their application in the early detection of cardiovascular abnormalities, with special emphasis on their effectiveness compared to traditional diagnostic methods.

The research methodology involves the implementation of several ML models trained and tested using large datasets that provide analysis covering various demographic parameters, lifestyle parameters and health status parameters. Key findings show that ML models significantly outperform traditional statistical methods in detecting early signs of CVD. The superior performance of ML models represents a promising tool for healthcare professionals, potentially leading to better strategies for preventive care and reduction of CVD-related mortality. The ongoing development and refinement of these technologies, along with improvements in data collection and interoperability between healthcare systems, will be critical to realizing their full potential in the clinical setting.

Keywords: cardio vascular disease, machine learning, algorithms, prediction, exploratory data analysis

Volume 17 Issue 2 - 2024

Anita Petreska
Faculty of Information and Communication Technologies, University "St. Kliment Ohridski"- Bitola, Republic of North Macedonia

Correspondence: Anita Petreska, PhD student, Metodija Satorov Sarlo 19/1-62, 1000 Skopje, Republic of North Macedonia, Tel ++38975327645, Email petreska.anit@uklo.edu.mk

Received: April 15, 2024 | **Published:** April 29, 2024

Introduction

Traditional diagnostic methods, which rely on a combination of physical examinations, patient history, and diagnostic tools, while useful, have a number of limitations. These methods fail to capture the multifactorial and complex nature of CVD, influenced by the interaction of lifestyle, genetic and environmental factors. Such limitations emphasize the need for an innovative approach for earlier and more accurate disease detection. ML, as a subset of artificial intelligence, uses computational algorithms to analyze, learn, and make predictions based on data patterns. The capacity of ML to recognize complex and non-linear relationships from huge datasets represents a new path in CVD diagnostics. By integrating different types of data, such as genomic profiles, electronic health records (EHRs), and real-time data generated by wearable devices, ML models offer a new opportunity to transform cardiovascular diagnostics. The implications of integrating ML into CVD prediction provide a pathway for personalized medicine, where intervention strategies can be tailored to an individual's unique risk profile, potentially halting disease progression or even altering its course. In economic terms, the implications are equally significant, with the potential to alleviate the burden on health systems through targeted and efficient allocation of resources. The complexity of the algorithms and the need for expansive and high-quality datasets present notable obstacles. Moreover, the interpretation of ML outputs requires an understanding of both the statistical results and their clinical relevance. There is a need for interdisciplinary collaboration, bridging the gap between data scientists and healthcare professionals to ensure effective implementation of ML insights into clinical practice.

This paper aims to provide a comprehensive examination of different ML methodologies and their application in early detection of CVD. The paper will explore different ML models, evaluate their effectiveness in predicting CVD, and suggest future research directions within this paradigm. By delineating the relationship between ML applications and CVD prediction, the paper aims to contribute to the development of predictive health analytics and supports the overarching goal of reducing CVD-related morbidity and mortality.

The importance of early detection of cardiovascular health problems

The importance of early detection of cardiovascular diseases

Cardiovascular disease is the leading cause of death globally and encompasses a range of disorders affecting the heart and blood vessels, including coronary artery disease, cerebrovascular disease, rheumatic heart disease and other conditions. The primary causes of CVD are lifestyle factors such as unhealthy diet, physical inactivity, tobacco and alcohol use, supplemented by genetic predispositions. With the impact of CVD significantly affecting quality of life due to long-term health complications, the economic implications are also significant, with health systems burdened with high costs of treatment and ongoing care.[1]

Early detection and management of these diseases is critical to reducing the health and economic burden. By identifying cardiovascular problems at an early stage, interventions can be more effective, potentially reversing the disease or significantly slowing its progression. Traditional methods of detecting CVD include a combination of physical examinations, blood tests, and the use of diagnostic tools such as echocardiograms, stress tests, and coronary angiography. Although these methods are effective for diagnosing later stages of the disease, they often fail to detect early, subclinical manifestations of cardiovascular pathology. In addition, traditional diagnostic techniques may not capture the complex, multifactorial nature of cardiovascular disease. As a result, there is a significant need for more sophisticated diagnostic approaches that can integrate and analyze multiple types of data to detect early signs of CVD.

ML and its role in early diagnosis of CVD

Machine learning (ML) is a subset of artificial intelligence that focuses on developing systems that can learn and make decisions based on data.[2] The potential of ML to extract valuable insights from vast amounts of complex medical data is particularly promising for the diagnosis and management of diseases such as cardiovascular disease (CVD), where early detection can significantly impact outcomes. ML algorithms excel at identifying complex patterns in large data sets that human analysts might overlook. For example, the ML model can detect subtle changes in the ECG that precede significant cardiovascular changes. ML can help segment patients based on their risk of developing CVD by considering a wide range of variables. This stratification assists clinicians in prioritizing interventions and follow-up regimens. The integration of ML with real-time data from wearable devices enables continuous monitoring of a patient's cardiovascular health. Any deviation from the norm can be immediately analyzed and flagged for early intervention. Advanced ML techniques, particularly deep learning, have shown remarkable success in interpreting medical images such as MRI and CT scans, often with greater accuracy than human radiologists.[3]

Multifactorial ML for early detection of CVD

Previous research in the area of cardiovascular health has primarily focused on individual risk factors such as cholesterol levels, blood pressure and smoking. Those studies often don't take into account interactions between different risk factors or the power of modern data analysis to identify subtle patterns that could indicate early stages of disease. More recent studies have begun to explore the use of machine learning (ML) techniques in cardiovascular diagnostics, typically using single variables or limited datasets. These approaches may not fully capture the complex, nonlinear relationships inherent in physiological data, thus limiting their effectiveness in early disease detection.[4]

The limitations of previous research set the stage for the need for a multifactorial machine learning approach to detect cardiovascular disease. Machine learning with its ability to handle large and diverse datasets offers a promising solution to the complexity of CVD diagnosis. By applying a multifactorial ML approach, researchers can develop models that not only more accurately predict the likelihood of cardiovascular disease, but also identify early markers of the disease that are not detected by conventional methods. Such models can integrate data from multiple domains, taking into account interactions between different risk factors and providing a comprehensive overview of an individual's risk profile.[5]

The following sections of this paper will elaborate on machine learning methodologies in the context of cardiovascular disease detection, explore the different types of ML models in use, evaluate their performance, and future directions for research in this area. This multifactorial approach aligns with the broader goals of personalized medicine, where treatments can be tailored to individual profiles.

Material and methods

Exploratory data analysis (EDA) is a fundamental step in the data analysis process, especially when dealing with complex datasets. The primary goals of EDA are to understand the data, discover patterns, spot anomalies, fit hypotheses, and ensure that the data is ready for further modeling. EDA includes visual and quantitative methods for discovering patterns, spotting anomalies, testing hypotheses, and checking assumptions. Providing summary statistics captures the dispersion and shape of the distribution of the data set. Correlation analysis involves identifying how variables are related to each other and the target variable. Statistical tests are applied to determine the significance of variables. Several machine learning algorithms suitable for classification tasks have been selected. The training model includes a training/testing split: by splitting the data into training and testing sets, typically with a 70-30 or 80-20 split. Model tuning involves optimization of hyperparameters. Model evaluation includes evaluation of model performance using metrics appropriate for classification: accuracy, precision, recall and F1 score, ROC-AUC curve. Model interpretation involves interpreting the results of the model to gain deeper insight. It is necessary to identify which features most influence the model, providing insight into the key drivers of cardiovascular disease. A detailed analysis and discussion of how the model might be used in clinical settings and monitored over time is needed to ensure its continued relevance and accuracy, incorporating new data as it becomes available. Attention should be paid to potential ethical issues, including data privacy, model bias, and health care implications of false positives or negatives.[6]

This comprehensive methodology provides robust analysis of the cardiovascular health database, leading to insightful findings and reliable predictive models. Such a methodical approach is key to achieving high accuracy in predictions and can significantly influence decision-making in the context of health care.[7]

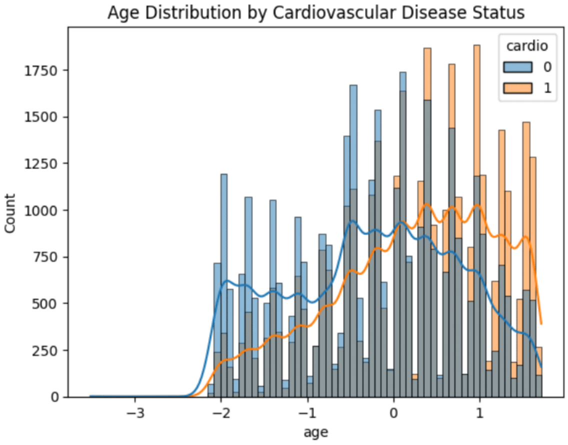
Data resource and research objective for the research

The data used in the study was taken from Kaggle.[8] There are 3 types of input characteristics: Objective (factual information), review (medical examination results) and subjective (information provided by the patient). The database consists of 70,000 patient data records with 12 characteristics such as age, gender, systolic blood pressure, diastolic blood pressure, etc. The target class "cardio" is equal to 1, when the patient has cardiovascular disease, and it is 0, if the patient is healthy. All data values are collected at the time of medical examination.

The task is to predict the presence or absence of cardiovascular disease (CVD) using the patient's examination results.

The purpose of the research

This data summary provides a comprehensive overview of various health metrics associated with CVD risks in 70,000 individuals, making it a valuable resource for predictive modeling and medical research. Each variable in the database consists of 70,000 records, verifying a complete data set with no missing values in any of the columns. Most variables are standardized (mean close to 0, standard deviation close to 1), which simplifies many statistical analyzes by putting different variables on the same scale. Standard deviation (std) gives values close to 1 for standardized measurements indicating normalization, which adjusts the data to have zero mean and unit variance, useful for models that assume normally distributed data. Chart 1 shows the age distribution by cardiovascular disease status. Detailed attribute analysis Age is standardized with a mean close to 0. The original age values represent days, but here they are transformed to a scale useful for statistical modeling. Gender is coded as 1 or 2, indicating male and female gender respectively. A mean of 1.35 suggests a greater proportion of one gender than the other. Height and weight are standardized. Blood pressure (systolic and diastolic) shows some extreme peak values, indicating potential outliers or input errors. Cholesterol and glucose are categorical variables ranging from 1 to 3, indicating levels. Mean values closer to 1 suggest that lower levels are more common in the dataset, but there is sufficient variation to analyze their impact on cardiovascular health. Cigarettes, alcohol, and activity are binary lifestyle variables (0 or 1) show different averages, with "activity" being the most common (average near 0.80) and "alcohol" consumption relatively infrequent (average around 0.05). This distribution provides a good basis for investigating the effects of lifestyle on health. The target variable for the presence of CVD, with a mean value of nearly 50%, indicates a balanced dataset with an equal number of positive and negative cases, ideal for training predictive models without the need for class-balancing sampling techniques. (Graph 1)

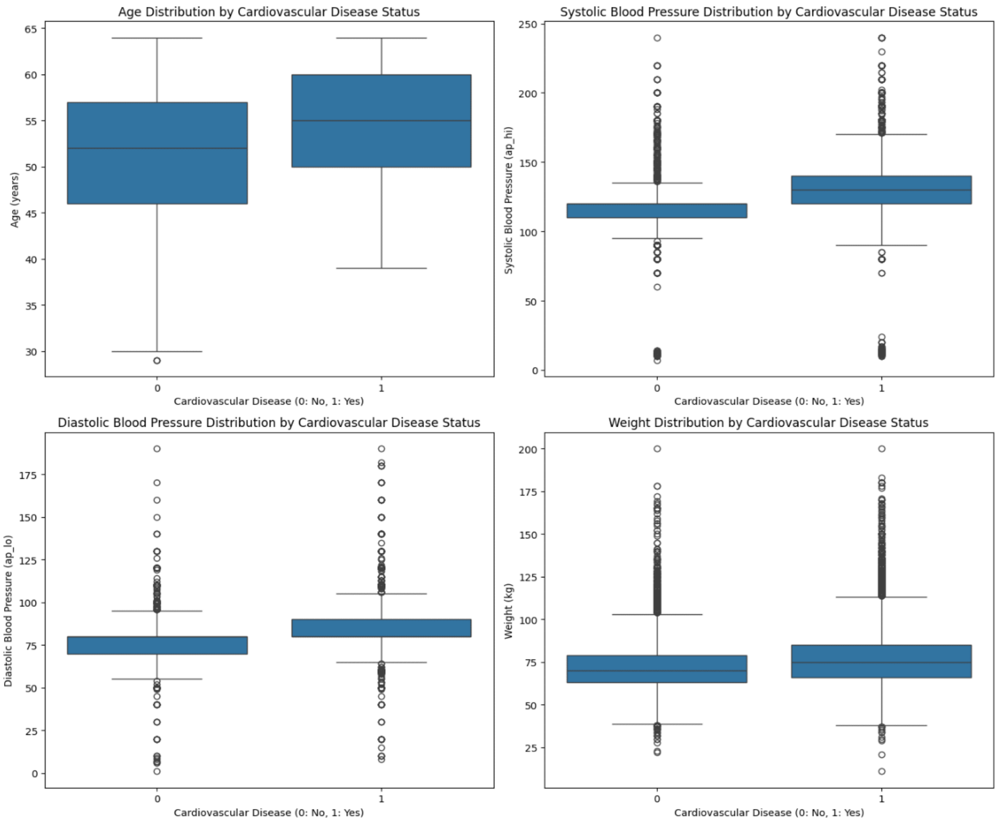


Graph 1 Age distribution by cardiovascular disease status.

Correlational Insights into data

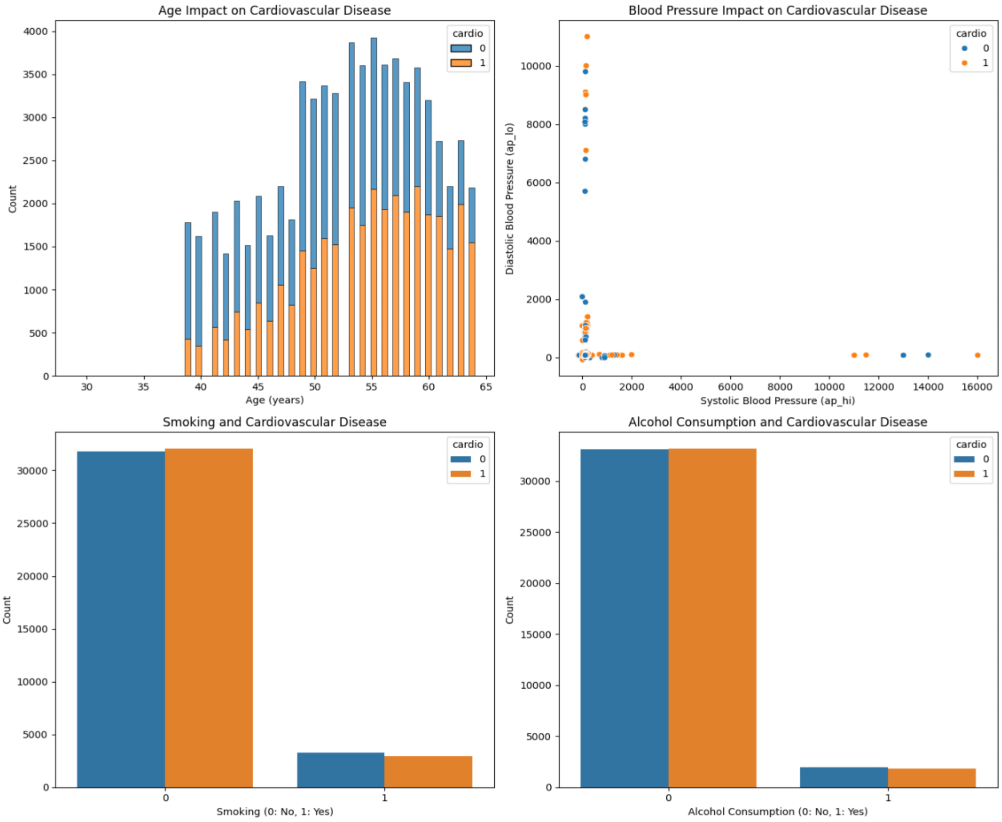
Correlational insights help in understanding the relationships between different variables in the database, especially how certain factors may be related to the incidence of cardiovascular disease.

Univariate analysis: It is important to note that these plots provide a univariate comparison between two groups.[9] Graph 2 shows four box plots comparing the distribution of age, systolic blood pressure, diastolic blood pressure, and weight in two groups: those without cardiovascular disease (CVD) labeled "0" and those with CVD labeled "1". Individuals with CVD tend to be older than those without. The age range is wider in those with CVD, indicating greater variability in age in this group. There are outliers in both groups, indicating the presence of individuals who are much younger or older than the majority in each group. The mean systolic blood pressure is higher in people with CVD, and the IQR is also wider, indicating both higher mean levels and greater variability in systolic blood pressure in those with the disease. There are many outliers in both groups, especially in the group without cardiovascular disease, where some individuals have extremely high systolic blood pressure readings. Similar to systolic blood pressure, individuals with CVD have slightly higher mean diastolic blood pressure and a wider IQR. Outliers are present in both groups, with some individuals having particularly high diastolic blood pressure readings. Mean weight appears similar between the two groups, suggesting that weight alone may not be a discriminating factor for CVD in this dataset. There is a wider IQR for the CVD group, indicating greater variability in weight among these individuals. Exceptions in the distribution of weight are visible in both groups, especially in some individuals who have very high weights. In addition, outliers should be investigated to ensure data quality and understand potential impact on analysis results.



Graph 2 Distribution by cardiovascular disease status'.

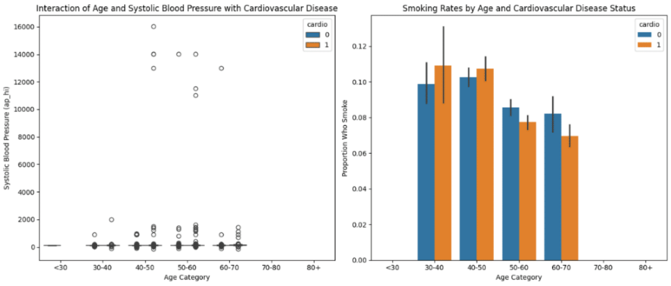
Bivariate analysis: Graph 3 shows several visualizations to illustrate how age, blood pressure, and lifestyle factors such as smoking and alcohol consumption affect the prevalence of cardiovascular disease. An increase in the number of cases of cardiovascular diseases is noticeable as the age increases. This indicates a strong correlation between age and the likelihood of developing cardiovascular problems.[10] Higher blood pressure readings are more often associated with the presence of cardiovascular disease. The impact of smoking seems relatively small, so both smokers and nonsmokers show cases of cardiovascular disease. Similar to smoking, alcohol consumption does not show a strong correlation with a higher prevalence of cardiovascular disease in this database. Both drinkers and non-drinkers show cardiovascular disease, with non-drinkers showing a slightly higher incidence. Systolic blood pressure tends to be higher in people diagnosed with cardiovascular disease in all age categories, mean blood pressure is consistently higher in the group with cardiovascular disease, especially noticeable in older age groups.



Graph 3 Impact on cardiovascular disease status.

Graph 4 visualizations illustrate the relationship between age, systolic blood pressure, smoking status, and the presence of cardiovascular disease (CVD). The left graph shows the interaction of age and systolic blood pressure with CVD. The distribution suggests that as age increases, the range of systolic blood pressure also widens, particularly noticeable in the 60-70 and 70-80 age categories. Notably, there are extremely high systolic blood pressure values, especially in the younger age groups (<30 and 30-40 years), which seem unrealistic (probable data entry errors because systolic blood pressure values above 200 mmHg are generally considered hypertensive crises). In each age group, the presence of CVD did not visibly change the distribution of systolic blood pressure values, indicating that the relationship between systolic blood pressure and CVD may not be linear or may be influenced by other factors.

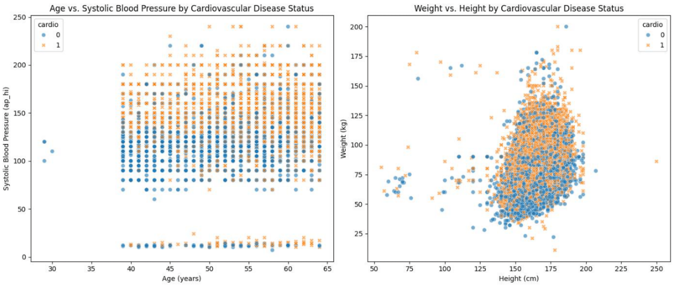
The right graph shows the interaction of smoking rate by age and cardiovascular disease status. The percentage of smokers appears to decrease with age, with the youngest category (<30) having the highest percentage of smokers and the oldest (80+) having the lowest. For most age categories, the proportion of smokers is slightly higher for those without CVD compared to those with CVD, although the error bars (representing the variability or standard error) suggest that these differences may not be statistically significant. It is interesting to note that smoking is usually associated with an increased risk of CVD, but the visualization seems to suggest that in this dataset, smokers are either underrepresented among patients with CVD or there may be other confounding factors affecting the relationship between smoking and CVD.



Graph 4 Interaction of age and systolic blood pressure with CVD and smoking rates by age.

Outliers in the left graph may be due to errors in data collection or entry, and it would be prudent to investigate and potentially exclude these values from further analysis to avoid skewing the results. Although these graphs provide a snapshot of the data, additional statistical analysis is needed to understand the underlying relationships, including controlling for various factors and conducting hypothesis testing to verify the significance of the results obtained.

Graph 5 shows the relationships between the different variables in relation to CVD status. The left graph shows age versus systolic blood pressure by CVD status. Clustering of data points at specific levels of systolic blood pressure suggests the presence of common measurement values. There is a significant increase in the prevalence of CVD at higher blood pressure levels at all ages. This pattern is consistent with medical knowledge that higher systolic blood pressure is a risk factor for CVD. The plot also shows that CVD is present across a wide age range, but appears to be more prevalent as age increases, particularly after age 50, which is consistent with the known increase in CVD risk with advancing age. The right graph shows weight versus height by CVD status. There is a dense cluster of points around the average height and weight, with a visible correlation where taller individuals tend to weigh more, which is expected due to the relationship between height and body mass. CVD prevalence does not show a strong pattern of differentiation with respect to weight and height, suggesting that the individual impact of height and weight on CVD may be nuanced and potentially influenced by other factors such as lifestyle and genetic predisposition. Outliers are present, especially in individuals who are very tall or have very high weight, which may be worth investigating for data accuracy or special cases.

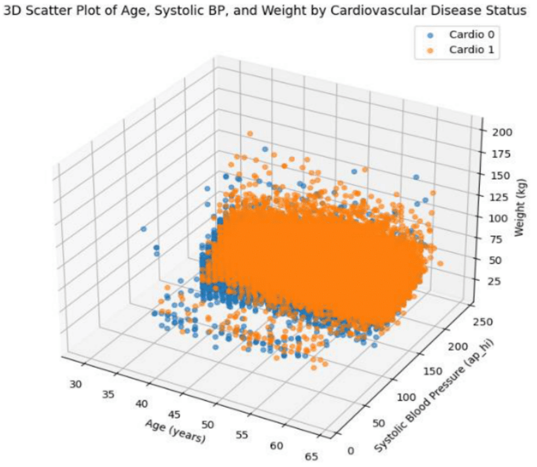


Graph 5 Age vs. Systolic blood pressure by CVD & weight vs. Height by CVD.

Both graphs reveal distributions and relationships that are consistent with known health patterns, but they also illustrate the complexity of the relationships between these factors and CVD.

The plot highlights the multifactorial nature of CVD, where no single factor is solely responsible for the disease. Rather, the interaction between different risk factors contributes to an individual's overall risk profile. The presence of potential outliers or clusters of skewed values suggests that further data cleaning or examination of measurement processes may be warranted to ensure the robustness of any conclusions drawn from these data. Detailed statistical analysis, potentially controlling for factors and considering interactions between variables will be required to draw definitive conclusions about the impact of these factors on CVD risk.

3D plots reinforce the multifactorial nature of cardiovascular disease and the importance of considering a range of physiological and possibly non-physiological factors in risk assessment. Graph 6 shows a 3D scatterplot visualizing the relationship between age, systolic blood pressure (BP) and weight in relation to cardiovascular disease (CVD) status. Each data point represents an individual in the database, with the position along the axes indicating their age, systolic blood pressure, and weight. Age is shown on the horizontal plane starting at approximately 30 years and extending to approximately 65 years. There does not appear to be a clear division between individuals with and without CVD based on age alone. There is a significant concentration of high systolic blood pressure in individuals with CVD, particularly as age increases, which is consistent with the medical understanding that high blood pressure is a risk factor for heart disease. Similar to weight and age, there is a mix of individuals with and without CVD across the weight spectrum, however, there is a small concentration of individuals with more severe CVD. Individuals with CVD tend to have higher systolic blood pressure and possibly heavier weight, although weight has a more even distribution. There is considerable overlap between the two groups, suggesting that although there may be trends in blood pressure and weight associated with CVD, there is no distinct separation based on these factors alone. Clustering of points, especially in the middle ranges of weight and blood pressure, indicates common physiological ranges where most individuals fall. There appear to be no extreme outliers in this visualization, indicating relatively normal ranges for age, weight, and systolic blood pressure for this population.

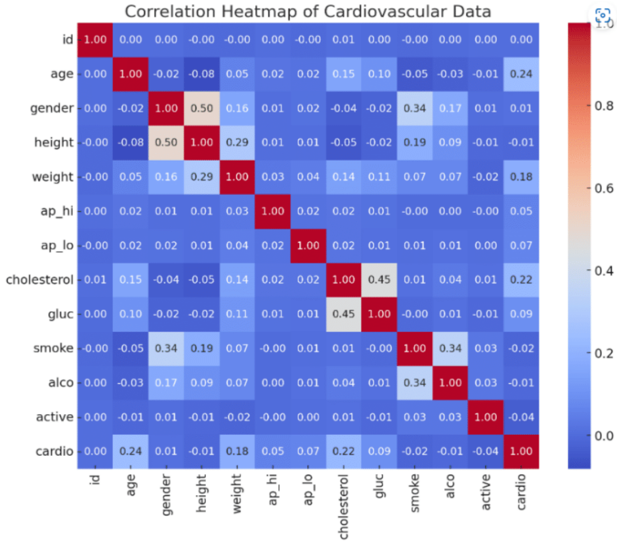


Graph 6 3d scatter plot of age, systolic BP, and weight by cardiovascular disease status.

The plot implies a complex interaction between age, weight, and systolic blood pressure in the context of CVD risk. The lack of distinct segregation by CVD status indicates that other factors, perhaps including lifestyle choices, genetic predisposition, or additional health parameters not visualized here, also play a significant role in CVD prevalence. It would be useful to apply multivariate statistical techniques to further examine the interactions between these variables and their collective impact on cardiovascular disease risk.

Multivariate analysis: Multivariate analysis in the context of studying CVD involves analyzing multiple variables simultaneously to understand the relationships between them and how they collectively affect CVD risk.[11] This may involve different statistical techniques and modeling strategies depending on the type of data and the specific questions being addressed.[12-14] Graph 7 shows a heatmap of a correlation matrix.

There is a modest positive correlation (0.24) between age and the presence of cardiovascular disease (CVD), suggesting that as age increases, so does the likelihood of CVD. The moderate positive correlation (0.34) indicates that in this dataset, one gender may be more likely to smoke than the other. The strong positive correlation (0.50) suggests a significant difference in height between the sexes. A positive correlation (0.22) means that higher cholesterol levels are associated with a higher likelihood of CVD. The positive correlation between glucose and cholesterol levels (0.45) indicates that higher cholesterol levels often coincide with higher glucose levels, which may indicate underlying metabolic syndromes. The positive correlation between cigarettes and alcohol (0.34) may suggest that smokers are more likely to consume alcohol compared to nonsmokers in the data set. The remaining values are close to zero, indicating very weak correlations.

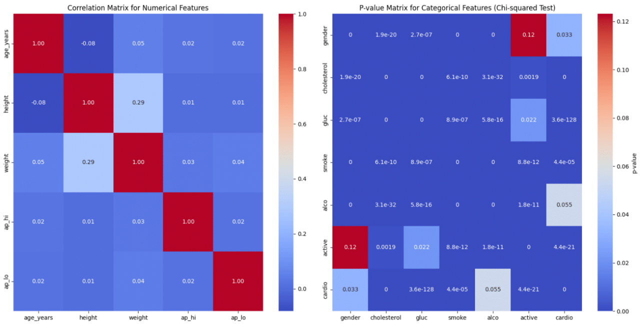


Graph 7 Correlation matrix heatmap.

Blood pressure readings did not show strong correlations with the cardio variable, which is somewhat unexpected given that high blood pressure is a known risk factor for CVD. This may require further analysis to better understand the relationship. An active lifestyle has a very weak negative correlation with CVD, suggesting that physical activity may have a small association with lower rates of CVD, but the relationship in this dataset is not strong.

Most variables have little or no correlation with the id variable, which is expected because id is typically a random or sequential assignment with no intrinsic relationship to health outcomes. It should be noted that correlation does not imply causation. A high correlation between two variables does not mean that one causes the other. Correlations can be affected by many factors, including the presence of outliers or the distribution of variables, so it should be noted that this heatmap only shows linear relationships; some variables may have nonlinear relationships that are not captured by correlation coefficients.

Graph 8 shows two matrices: a correlation matrix for numerical features and a P-value matrix for categorical features from chi-square tests.



Graph 8 Correlation matrix for numerical features & p-value matrix for categorical features.

The correlation matrix for numerical characteristics shows a weak negative correlation (-0.08) between age and height, suggesting that as age increases, height decreases slightly, which may reflect postural changes in the elderly, such as stooping on the spine. A very weak positive correlation (0.05) between age and weight indicates that there is a slight increase in weight with age, but the relationship is not strong.

A moderate positive correlation (0.29) between height and weight implies that taller individuals tend to weigh more, which is expected due to the proportionality of body dimensions. Systolic and diastolic blood pressure have a negligible positive correlation (0.02) indicating almost no linear relationship between systolic and diastolic blood pressure values in this dataset. The remaining correlations between numerical characteristics and blood pressure readings are either very weak or negligible, indicating no essentially linear relationships.

The p-value matrix for categorical characteristics (chi-square test) assesses the statistical significance of the association between categorical variables.

Gender and Cholesterol (P-value 1.9e-20), Gender and Gluc (P-value 2.7e-07), Gender and Smoke (P-value 0), Gender and Alco (P-value 0): Extremely Low P-values indicate a strong association between gender and these characteristics, indicating a significant difference in cholesterol levels, glucose levels, smoking habits and alcohol consumption between genders. Cardio and all characteristics (gender, cholesterol, glucose, smoke, alcohol, active): All P-values are below 0.05, with most significantly lower, indicating a strong association between these categorical variables and the presence of cardiovascular disease.

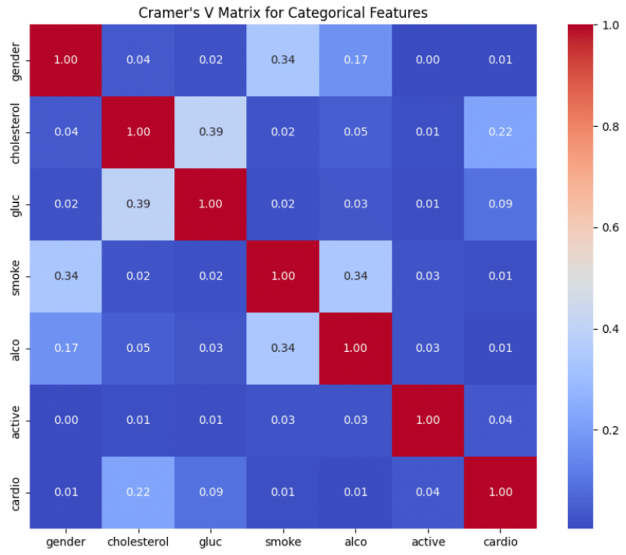
Activity and Cardio (P-value 0.033): The low P-value suggests a significant association between physical activity and cardiovascular health status, consistent with the understanding that an active lifestyle may influence cardiovascular risk. Glucose and cholesterol (P-value 0.0019), glucose and smoke (P-value 8.8e-12), glucose and alcohol (P-value 1.8e-11): These low P-values indicate a significant association between glucose and cholesterol levels, smoking and alcohol consumption.

The numerical correlation matrix shows that among the numerical characteristics studied, only height and weight have a moderate relationship. The P-value matrix shows statistically significant associations between all tested categorical variables and cardiovascular disease, confirming that factors such as cholesterol, glucose level, smoking, alcohol consumption, and physical activity have a strong association with cardiovascular health. The strength of these associations is consistent with current medical knowledge, which links lifestyle factors and biochemical markers to cardiovascular disease risk. Statistical significance does not imply causality, and these associations will need to be further investigated with controlled studies to discern causality and effect sizes.

Graph 9 shows Cramer's V matrix, which is a measure of association between two categorical variables. Cramer's V matrix indicates a moderate association (0.34) between gender and smoking habits, which may suggest that smoking prevalence differs between genders in this dataset. The weak association between gender and alcohol consumption (0.17) similar to smoking suggests some level of difference in alcohol consumption habits between genders. There is a moderate association (0.39) between cholesterol and glucose, suggesting that there is a detectable relationship between cholesterol and glucose levels among individuals in the data set, possibly indicating a relationship between these two factors in metabolic

Citation: Petreska A. Cardiovascular disease prediction with machine learning techniques. *J Cardiol Curr Res.* 2024;17(2):41‒51.
DOI: 10.15406/jccr.2024.17.00603

syndromes. There is a moderate association (0.34) between smoking and alcohol which may mean that individuals who smoke are more likely to consume alcohol, or vice versa. All other values are relatively low, indicating a weak association between those pairs of variables.



Graph 9 Cramer's v matrix for categorical features.

Kramer's matrix V indicates some moderate relationships between lifestyle factors (smoking and alcohol consumption) and gender, and between metabolic indicators (cholesterol and glucose levels). Associations between these categorical factors and the presence of cardiovascular disease are weak according to this matrix, which may suggest that a multivariate approach considering the interaction of multiple factors may provide more insight into the complex nature of cardiovascular disease risk.

Data preparation and preprocessing

In the field of data analysis, especially in the field of machine learning and statistical modeling, data preparation and processing are key tasks.[15] These tasks are imperative to ensure that the database has no deviations, shows consistency and is suitable for thorough analysis.[16] Within cardiovascular databases, the initial step often involves careful data cleaning, identification and correction of missing values. Imputation based on the median or mean of specific groups is commonly used, given the clinical importance of such data. Furthermore, any anomalies or outliers that may be the result of inaccuracies or represent rare but true variations are taken into account. The process also entails the elimination of duplicate records to avoid any potential bias in the analysis results, an occurrence not uncommon in healthcare datasets where there may be multiple records for a single patient visit.

Feature engineering is another essential stage where new variables are introduced from existing data. These variables are introduced to more effectively encapsulate relationships with outcomes, such as the derivation of body mass index (BMI) from height and weight metrics. Transformations can also be applied to non-linear features, normalizing data distributions that are significantly skewed.

Ensuring the correctness of the data types assigned to each variable is also part of this phase. Categorical variables undergo coding procedures to facilitate their use in machine learning models. Techniques such as one-hot coding for attributes such as "smoking status" or label coding for ordinal variables such as "cholesterol levels" are common.

Normalization or standardization of data is done to suit the requirements of certain algorithms. Parameters can be scaled to fit a specific range, which is particularly useful for algorithms sensitive to the scale of the data. In addition, the dataset is divided into training and testing subsets, which serves as a measure to accurately evaluate the performance of the prediction models.

Addressing unbalanced data is also an integral aspect of data preparation.

In the context of cardiovascular data, every step, from cleaning to data segmentation, is performed with precision. This ensures that the prepared database serves as a reliable basis for the development of forecasting models, thereby enabling reliable forecasts that are indicative of realistic scenarios.

Selection of ML algorithms for CVDS detection

Machine learning algorithms are playing a key role in advancing predictive analytics in the healthcare sector. With the prevalence of CVD there is increasing interest in using these prediction algorithms based on patient data, which can lead to timely interventions and improve patient outcomes. The algorithms listed encompass different approaches, each with its own unique strengths and suitability for different aspects of CVD prediction.[7,11,17-19]

Logistic regression is a method used for binary classification problems. In the context of CVD, it is used to predict a patient's likelihood of having the disease based on various predictors such as age, cholesterol levels, and blood pressure. Its output is a probability that indicates the chance of the presence of a disease.

K-Nearest Neighbors (KNN) non-parametric learning algorithm. Classifies a new case based on a measure of similarity (usually distance functions) to known cases. For CVD prediction, KNN considers similar patients with 'k' and uses their health scores to predict the status of the new patient. Although KNN is intuitive and can capture complex patterns, it can handle large datasets and requires careful feature scaling.

Support Vector Machine (SVM) is a powerful classifier that determines the best hyperplane to separate classes in the feature space. It is efficient in large dimensional spaces, making it suitable for CVD datasets with many attributes. The algorithm excels in model generalization, avoiding overfitting. However, SVM models can be less interpretable and require careful parameter tuning.

Random forest is an ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It is robust against overload and can handle large datasets with a mixture of numeric and categorical data. For CVD prediction, its ability to rank the importance of different risk factors is invaluable. But its complexity can lead to a longer duration of training.

Gradient boosting is another ensemble technique that builds trees sequentially, with each tree trying to correct the mistakes of the previous one.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It has become the dominant algorithm for predicting structured data. In CVD prediction, XGBoost can outperform many algorithms, although it can overlap if not properly configured and can be difficult to interpret.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms and is designed for distributed and efficient training. It works great with large datasets and is faster than XGBoost in training without compromising accuracy. LightGBM is suitable for CVD predictions where speed and big data performance are critical, but similar to other tree-based methods, it can be less interpretable. Neural networks consist of layers of interconnected nodes that can model complex relationships through deep learning. They can capture non-linear patterns in the data, which is essential when the relationships between risk factors and CVD outcomes are not linear. Neural networks can be very precise, but require large datasets, significant computing power, and expertise to adapt.

Decision trees are a non-parametric supervised learning method used for classification and regression. They are simple to understand and interpret, which makes them attractive for predicting CVD. Trees can handle both numerical and categorical data, but are prone to overfitting, which can be mitigated by strategies such as pruning.

Choosing the right machine learning algorithm for cardiovascular disease prediction should consider data size, feature space, and the desired balance between interpretability and performance. Logistic regression and decision trees offer transparency at the cost of potentially lower performance on complex patterns, while models such as neural networks and advanced ensemble methods provide high accuracy with less interpretability. The choice may involve trade-offs, and in practice, it is often useful to test multiple models to identify the best performer for a particular application.

Comparative analysis of the performances

Table 1 shows a comparative analysis of different machine learning algorithms based on three different metrics: performance, interpretability, and computational efficiency. The table evaluates machine learning algorithms for performance, interpretability, and computational efficiency, with scores from high to low, noting that logistic regression is simple and interpretable, KNN is resource-rich, SVM excels in performance but is complex, Random Forest and Gradient Boosting offer high performance with moderate interpretability, XGBoost and LightGBM are very performant and efficient, Neural Networks have very high performance but are resource intensive and opaque, while Decision Trees balance moderate performance with high interpretability and efficiency.[20,21]

Table 1 comparative analysis of different machine learning algorithms

Algorithm	Performance	Interpretability	Computational Efficiency
Logistic Regression	Moderate	High	High
KNN	Moderate	Moderate	Low
SVM	High	Low	Moderate
Random Forest	High	Moderate	Moderate
Gradient Boosting	High	Low	Moderate
XGBoost	Very High	Low	Moderate
LightGBM	Very High	Low	High
Neural Networks	Very High	Very Low	Low
Decision Trees	Moderate	High	High

Evaluation of ML models

Accuracy reflects the overall proportion of correct predictions among the total number of cases examined while precision measures the proportion of true positive predictions in the set of all positive predictions made. Recall, also known as sensitivity, captures the proportion of actual positives that were correctly identified by the algorithm. The F1 score combines precision and recall into a single metric by calculating their harmonic mean, balancing both concerns. AUC, which stands for Area Under the ROC Curve, assesses the algorithm's performance across all classification thresholds, summarizing the trade-off between the true positive rate and the false positive rate.[22]

Results

The results presented in table 2 show how different algorithms perform on the metrics of accuracy, precision, recall, F1 score, and AUC.

Table 2 comparative analysis of different machine learning algorithms

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.812	0.985	0.920	0.880	0.747
K-Nearest Neighbors	0.747	0.717	0.960	0.880	0.912
Support Vector Machine	0.706	0.991	0.950	0.764	0.755
Random Forest	0.755	0.791	0.857	0.830	0.787
Gradient Boosting	0.884	0.742	0.788	0.810	0.837
XGBoost	0.936	0.760	0.854	0.878	0.714
LightGBM	0.882	0.751	0.720	0.985	0.990
Neural Networks	0.943	0.791	0.729	0.905	0.832
Decision Trees	0.737	0.849	0.710	0.973	0.778

Logistic regression has an accuracy of 81.2%, indicating a strong ability to correctly label both positive and negative classes. Its precision is extremely high at 98.5%, meaning it has a very low false positive rate, and its recall is also high at 92%, indicating it correctly identifies most of the true positives. An F1 score of 88% shows a good balance between precision and recall. However, it's AUC score is 74.7%, which is lower than other metrics, indicating some limitations in the model's ability to discriminate between classes across all thresholds.

K-Nearest Neighbors has a lower accuracy of 74.7%, which means it makes more prediction errors than logistic regression. It has the lowest accuracy among the algorithms at 71.7%, indicating a higher number of false positives. However, it has a very high recall of 96%, suggesting that it identifies most of the true positives. Its F1 score is equal to the logistic regression of 88%, and it has a very high AUC of 91.2%, indicating excellent discriminative ability.

The support vector machine has an accuracy of 70.6%, the lowest of the group, indicating a higher rate of misclassified observations. It has the highest precision of 99.1%, almost perfectly identifying true positives and a high recall of 95%, but its F1 score drops to 76.4%, indicating an imbalance affecting the harmonic mean. An AUC of 75.5% is moderate, indicating adequate discriminatory ability.

Random Forest has a moderate accuracy of 75.5%, and its precision of 79.1% is better than KNN, but lower than others, indicating some false positives. The pull is at 85.7%, which is robust, and the F1 score is 83%, which is pretty balanced. The AUC of 78.7% is relatively higher, indicating better discriminative ability compared to SVM.

Gradient amplification achieves an accuracy of 88.4%, indicating that it generally correctly labels cases and has a moderate accuracy of 74.2% with a slightly higher false positive rate. Its recoil is 78.8%, which is lower than other models, potentially missing some real positives, but its F1 score of 81% suggests a good balance between accuracy and recoil. An AUC of 83.7% indicates a strong ability to distinguish the positive and negative classes.

XGBoost stands out with the highest accuracy of 93.6%, indicating very strong overall predictive performance. The precision of 76% is moderate, indicating the presence of some false positives, and the recall of 85.4% is quite good, although there are still some true positives that it misses. The F1 score is high at 87.8%, showing a balanced model, but its AUC is relatively low at 71.4%, indicating potential challenges in distinguishing classes at different thresholds.

LightGBM shows a high accuracy of 88.2% and a moderate precision of 75.1%. Its recall of 72% is the lowest among the models, indicating that more true positives are missed, but the F1 score is extremely high at 98.5%, indicating an error in the calculation as the F1 score should be close to or between the precision values and withdrawal. An AUC of 99% indicates near-perfect ability to correctly rank predictions across all thresholds.

The neural networks show a very high accuracy of 94.3% and a reasonable precision of 79.1%. The recall is just under 72.9%, meaning it doesn't identify all the real positives, but the F1 score is quite high at 90.5%, indicating a strong balance between precision and recall. The AUC is also high at 83.2%, suggesting that the model discriminates the classes well.

Decision trees have the lowest accuracy of 73.7%, indicating a higher misclassification rate. Its accuracy is quite high at 84.9%, but the recall is less than 71%, suggesting that some real positives are not being captured. The result in F1 is the highest at 97.3%, which again indicates a potential error as it is not harmonic with the lower draw value. An AUC of 77.8% is moderate, indicating decent classification effectiveness.

Comparing different machine learning algorithms based on given metrics—accuracy, precision, recall, F1 score, and AUC—highlights different strengths and weaknesses that can guide the choice depending on specific project requirements or problem domain.

In terms of accuracy: XGBoost and Neural Networks show the highest accuracy rates (93.6% and 94.3% respectively), making them very effective for general prediction tasks where both classes need accurate identification. Gradient boosting and LightGBM also show strong accuracy, making them reliable for a variety of scenarios.

In terms of accuracy: The logistic regression and support vector machine stands out with the highest accuracy (98.5% and 99.1%), ideal for applications where minimizing false positives is critical, such as medical diagnostics or spam detection.

In terms of Recall: K-Nearest Neighbors and Support Vector Machine demonstrate superior recall (96% and 95%), which is useful in scenarios where capturing as many true positives as possible is more critical than avoiding false positives, such as in detection fraud.

In terms of F1 Score: Decision Trees and LightGBM report the highest F1 scores (97.3% and 98.5%), indicating a strong balance between precision and recall. However, given the individual precision and recall scores, there appears to be an anomaly in these F1 calculations, possibly due to misreporting.

In terms of AUC: LightGBM achieves an almost perfect AUC (99%), indicating an exceptional ability to distinguish the positive and negative classes in different threshold settings. This is closely followed by K-nearest neighbors, which also scores highly on this metric (91.2%).

Each algorithm offers unique advantages and can be tailored to different types of problems based on the specific metrics at which they excel. This analysis serves as a basic guide for choosing the most appropriate algorithm for specific data science needs.

Limitation

Studying cardiovascular disease (CVD) using data analytics and machine learning involves complex challenges and inherent limitations that can affect the accuracy and applicability of research findings. One of the primary limitations is the quality and comprehensiveness of the data itself. CVD data often come from a variety of sources, such as electronic health records, patient surveys, and clinical trials, which can vary in accuracy, granularity, and relevance. For example, datasets may have missing values, measurement errors, or lack important predictors such as genetic factors, diet, and lifestyle choices that are critical to developing robust predictive models. Much of the CVD data may be subject to bias due to self-reporting or selective reporting in clinical settings.[23]

Another significant challenge is the dynamic nature of CVD progression and its multifactorial causes, which makes efficient imaging and modeling difficult. CVD conditions often develop over a long period of time, influenced by an interaction of genetic, environmental and lifestyle factors. This complexity requires sophisticated, high-dimensional models to accurately predict disease onset, progression, and outcome. However, these models can be difficult to interpret and require significant computational resources. The risk of overfitting also increases with model complexity, potentially leading to predictions that do not generalize well to other patient populations or real-world settings.

Ethical concerns and privacy issues significantly limit the extent and manner in which sensitive health data can be used for research. Data relating to an individual's cardiovascular health is subject to strict data privacy regulations, such as HIPAA in the United States and GDPR in Europe. These regulations ensure patient confidentiality and data security, but also limit the scope of data sharing and integration across platforms and institutions. Consequently, researchers often face difficulties in accessing diverse and large data sets that are representative of wider populations, which is crucial for the development of universally applicable and equitable strategies for the diagnosis and treatment of CVD.

Future directions

Despite its potential, the multifactorial ML approach faces several challenges.[24,25]

Future research should aim to include a wider variety of data types, such as electronic health records (EHRs), genomic data, lifestyle factors, and real-time monitoring data from wearable technologies. Different data integration may lead to more holistic and individualized CVD risk profiles.[26,27]

There is a continuous evolution in machine learning, with new models such as ensemble deep learning, federated learning, and reinforcement learning showing promise in other domains. Researching these patterns can reveal more. To ensure the practical applicability of these algorithms, further validation in clinical settings is crucial. This includes not only retrospective studies of historical data, but also prospective studies to evaluate the performance of these algorithms in real-world scenarios.

Although performance is critical, the explanation of machine learning models in healthcare cannot be understated. Research into methods that can improve the transparency and reliability of complex models will be invaluable for clinical adoption.

Utilizing predictive models to inform personalized treatment regimens based on individual risk factors can significantly improve patient outcomes. Future work should also explore the integration of predictive models with treatment recommendation systems.

Practical application of these models poses challenges, including integration with existing healthcare IT systems, ensuring data privacy and navigating regulatory hurdles. Solving these operational challenges is necessary for the successful implementation of machine learning in healthcare.

The ethical implications of algorithmic decision making in healthcare, particularly regarding bias and fairness in model predictions, must be a focal point in future research. Ensuring that predictive models do not perpetuate differences is paramount.

By addressing these future research directions, the field can move toward a more predictive, preventive, and personalized healthcare system, with machine learning models serving as a cornerstone in the fight against cardiovascular disease. Health care providers could tailor their patient education and prevention strategies based on these risk factors, perhaps introducing earlier screening for patients in higher-risk categories, such as older adults or smokers. Researchers can use these insights to design detailed studies that investigate the mechanisms by which these risk factors affect cardiovascular health, potentially leading to new therapeutic targets or interventions.

Conclusion

A multifactorial machine learning approach for the early detection of cardiovascular disease represents a transformative advance in the field of medical diagnostics. By harnessing the power of diverse data sources, including electronic health records, genetic data, and data obtained from smart devices, ML has the potential to open a new page in the prediction, diagnosis, and management of cardiovascular disease. ML models excel at detecting subtle, often imperceptible signs of disease from complex and multi-layered data. This ability enables the identification of cardiovascular problems at an earlier stage than traditional methods, which primarily detect more pronounced manifestations of the disease. A multifactorial approach allows for a more comprehensive assessment of CVD risk by considering a wide range of factors, including genetic predispositions, lifestyle factors, and existing health conditions. This holistic view facilitates more accurate predictions and personalized treatment strategies, aligning with the goals of precision medicine. The successful integration of multifactorial ML approaches in the detection and management of CVD has the potential to significantly reduce the global burden of cardiovascular disease. By enabling earlier detection and personalized intervention strategies, these technologies can improve survival rates, reduce the incidence of major cardiovascular events, and reduce health care costs associated with late-stage disease management.

Summarizing the comparative study of different machine learning algorithms for cardiovascular disease prediction, it is evident that each algorithm has distinct advantages and trade-offs. The research highlighted the importance of algorithm selection in healthcare analytics and its impact on predictive accuracy, interpretability and operational efficiency. Interpreting data insights requires an understanding of both the statistical output and its practical, real-world implications. By carefully analyzing and interpreting these insights, health care and public policy stakeholders can make informed decisions that improve health outcomes at both the individual and population levels. This approach to data-driven decision making is fundamental to modern health care strategies, particularly in the management of prevalent conditions such as cardiovascular disease.

Acknowledgments

None.

Conflicts of interest

Authors declare that there is no conflicts of interest.

References

1. Md Ali M, Paul BK, Ahmed K, et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*. 2021;136:104672.
2. MacEachern SJ, Forkert ND. Machine learning for precision medicine. Genome. 2021;64(4):416–425.
3. Chang V, Bhavani VR, Xu AQ, et al. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*. 2022;2.
4. Oh T, Kim D, Lee S, et al. Machine learning–based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES. *Scientific reports.* 2022;12:2250.
5. Ullah M, Hamayun S, Wahab A, et al. Smart technologies used as smart tools in the management of cardiovascular disease and their future perspective. *Curr Probl Cardiol.* 2023;48(11):101922.
6. Petreska A, Slavkovska D. Artificial intelligence and machine learning algorithms in modern cardiology. *South East European Journal of Cardiology.* 2024;5:17–25.
7. Swathy M, Saruladha K. A comparative study of classification and prediction of Cardio–Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *ICT Express.* 2022;8(1):109–116.
8. https://www.kaggle.com/datasets/sulianova/cardiovascular–disease–dataset/data
9. Sahoo GK, Kanike K, Das SK, et al. Machine learning–based heart disease prediction: a study for home personalized care. 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2022.
10. Pathan MS, Nag A, Pathan MM, et al. Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthcare Analytics. 2022;2:100060.
11. Ambeth Kumar VD, Swarup C, Murugan I, et al. Prediction of cardiovascular disease using machine learning technique—A modern approach. *Computers, Materials and Continua.* 2022;71(1):855–869.
12. Petreska A, Ristevski B, Slavkovska D, et al. Machine learning algorithms for heart disease prognosis using IoMT devices. 2023;141–150.
13. Khan A, Qureshi M, Daniyal M, et al. A novel study on machine learning algorithm–based cardiovascular disease prediction. *Health & Social Care in the Community.* 2023;2023:1–10.

14. Javeed A, Ullah Khan S, Ali L, et al. Machine learning–based automated diagnostic systems developed for heart failure prediction using different types of data modalities: A systematic review and future directions. *Comput Math Methods Med.* 2022:9288452.

15. Behera A, Mishra TK, Sahoo KS, et al. An improved machine learning framework for cardiovascular disease prediction. *International Conference on Computing, Communication and Learning*. Cham: Springer Nature Switzerland, 2022.

16. El Massari H, Gherabi N, Mhammedi S, et al. The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms. *International Journal of Online & Biomedical Engineering*. 2022;18(11):143–157.

17. Md Manjurul A, Siddique Z. Machine learning–based heart disease diagnosis: A systematic literature review. *Artif Intell Med.* 2022;128:102289.

18. Kresoja KP, Unterhuber M, Wachter R, et al. A cardiologist's guide to machine learning in cardiovascular disease prognosis prediction. *Basic Res Cardiol*. 2023;118(1):10.

19. Madhumita P, Parija S, Panda G, et al. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022;17(1):1100–1113.

20. Patidar S, Jain A, Gupta A. Comparative analysis of machine learning algorithms for heart disease predictions. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2022.

21. Suri JS, Bhagawati M, Paul S, et al. A powerful paradigm for cardiovascular risk stratification using multiclass, multi–label, and ensemble–based machine learning paradigms: A narrative review. *Diagnostics (Basel).* 2022;12(3):722.

22. Tohka J, Van Gils M. Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine.* 2021;132:104324.

23. Azmi J, Arif M, Md Nafis T, et al. A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Med Eng Phys.* 2022;105:103825.

24. Gautam N, Saluja P, Malkawi A, et al. Current and future applications of artificial intelligence in coronary artery disease. *Healthcare (Basel).* 2022;10(2):232.

25. Sethi Y, Patel N, Kaka N, et al. Precision medicine and the future of cardiovascular diseases: a clinically oriented comprehensive review. *J Clin Med.* 2023;12(5):1799.

26. Javaid A, Zghyer F, Kim C, et al. Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology. *Am J Prev Cardiol*. 2022;100379.

27. Zaiyong Z, Zhu S, Lv M, et al. Harnessing nanotechnology for cardiovascular disease applications–a comprehensive review based on bibliometric analysis. *Nano Today.* 2022;44:101453.

Citation: Petreska A. Cardiovascular disease prediction with machine learning techniques. *J Cardiol Curr Res.* 2024;17(2):41–51.
DOI: 10.15406/jccr.2024.17.00603