JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A PREDICTION MODEL FOR ADDRESSING THE CARDIAC HEALH ISSUE AMONG YOUNG PEOPLE USING BIG DATA ANALYTICS

[1]Siddesh.A.S, [2]Dr.K.T.Veeramanju
[1]Assistant Professor, [2]Research Supervisor
[1]Department of Computer Science & Engineering,
[1]Bapuji Institute of Engineering and Technology, Davangere, India

Abstract: This paper explores the changing landscape of cardiovascular diseases (CVDs), extending beyond aging populations to impact the youth due to modern lifestyle shifts. Leveraging big data analytics, the study delves into intricate factors influencing cardiac health among younger individuals, unravelling complexities related to sedentary lifestyles, dietary habits, and genetic predispositions. The evaluation of various machine learning classifiers for predicting heart disease reveals Logistic Regression and Random Forest as standout models, demonstrating high accuracy and balanced metrics. SVM and KNeighbors offer a well-balanced approach, while XGBoost showcases competitive precision. The proposed modified LightGBM presents a balanced alternative. Model selection depends on application priorities, emphasizing accurate predictions, effective capture of positive instances, or balanced metric performance. In summary this study contributes insights into constructing a robust prediction model tailored for addressing cardiac health issues in the younger demographic through machine learning methodologies. The findings, driven by big data analytics, offer transformative potential, reshaping our comprehension of cardiovascular health and guiding targeted strategies and preventative measures.

1. INTRODUCTION

Cardiovascular diseases (CVDs) pose a significant and pervasive global health challenge, traditionally recognized as ailments affecting the aging population. However, there is a growing and alarming trend of these diseases afflicting a younger demographic. This shift can be attributed to widespread lifestyle changes, sedentary behaviors, and dietary habits, contributing to a notable surge in cardiac risk factors among the youth. Addressing this concerning scenario requires innovative and pre-emptive approaches.

In response to the escalating health concerns among the younger population, the integration of big data analytics emerges as a transformative force. Big data analytics offers a unique capacity to process vast and diverse datasets, providing a valuable tool to unravel the intricate web of factors contributing to cardiac health issues in young individuals. In recent decades, a disconcerting trend has emerged, challenging the conventional association of cardiac health issues predominantly with aging populations. Instead, these issues are increasingly manifesting in individuals during their formative years.

This paradigm shift necessitates a more sophisticated and data-driven understanding of cardiovascular health, considering the complex interplay of sedentary lifestyles, dietary habits, and genetic predispositions. This paper aims to explore the transformative potential of big data analytics, equipped with its unparalleled ability to analyze extensive and varied datasets. By doing so, it seeks to decipher the complexities associated with cardiac health issues among the youth and pave the way for informed interventions and preventative measures. The integration of big data analytics holds promise as a powerful tool in reshaping our approach to cardiovascular health in the younger population, offering insights that can inform targeted strategies for healthier living.

2. Literature Overview

Cardiovascular diseases (CVD) have assumed a critical role in global health, directly claiming over 17.8 million lives annually. However, the transformative impact of the healthcare industry's extensive data repositories has yet to be fully realized [1]. Various risk factors contribute to cardiovascular diseases, including gender, smoking, age, family history, poor diet, physical inactivity, high blood pressure, weight gain, and alcohol consumption [2]. Cardiovascular issues traditionally rely on symptoms such as chest pain and fatigue, with nearly 50% of cases remaining undetected until after an adverse cardiac event [3].

Inheritance plays a significant role in predisposing individuals to cardiovascular disease, with high blood pressure and diabetes being notable examples. Additionally, several lifestyle factors contribute to an increased risk of developing cardiovascular issues. These include physical inactivity, excess weight, poor dietary habits, and the presence of symptoms such as back, neck, and shoulder pain, persistent fatigue, and a rapid heartbeat. Common indicators of potential heart problems encompass chest pain, shoulder pain, arm pain, shortness of breath, and a pervasive sense of weakness. Throughout history, chest pain has consistently stood out as the predominant and widely recognized sign indicative of insufficient blood supply to the heart [4].

For instance, Zhao et al. [5] employed a t-test-based AdaBoost approach to evaluate biological parameters related to coronary heart disease, focusing on Qi Deficiency Syndrome. The intersection of cardiology and big data analytics opens avenues for enhanced diagnostics and patient outcomes. Despite challenges such as noise and incompleteness in healthcare data, recent technological advancements, including big data analytics and machine learning (ML), play pivotal roles in reshaping cardiology [6]. Abdel-Motaleb and Akula [7] proposed a diagnosis method based on phonogram signals, utilizing Back Propagation and Radial Basis Function Artificial Neural Networks. Zhang et al. [8] introduced Support Vector Machine, achieving an 84.1% accuracy in heart disease prediction.

The Heart Disease Prediction System utilizes the Naive Bayesian Classification technique to facilitate decision-making. Through the analysis of an extensive database comprising past heart disease cases, this system reveals valuable insights. Its efficiency lies in the adept identification of patients at risk of heart disease. Notably, the model demonstrates prowess in responding to complex queries, underscoring its strengths in interpretability, access to comprehensive information, and accuracy [9]. In assessing the predictive capabilities of Support Vector Machine (SVM) and Naive Bayes algorithms regarding the occurrence of heart disease and patients' survival status, the author conducted a comprehensive investigation. These algorithms were applied to a dataset featuring sixteen attributes sourced from the University of California, Irvine’s Centre for Machine Learning and Intelligent Systems. The evaluation of model performance involved the use of a confusion matrix, providing visual insights into metrics such as accuracy, recall, precision, and error. Furthermore, a rigorous statistical analysis was conducted, utilizing the receiver operating characteristic (ROC) curve and calculating the area under the curve to effectively demonstrate the accuracy of the models [10].

While promising, studies like the one conducted by Setiawan et al. [11] underscore the need for rigorous evaluation methods such as k-fold cross-validation. In the realm of machine learning, ensemble learning, particularly the mixture of expert strategy, has demonstrated efficacy in learning intricate patterns, offering advantages in generalization and efficient processing of large datasets [12], [13].

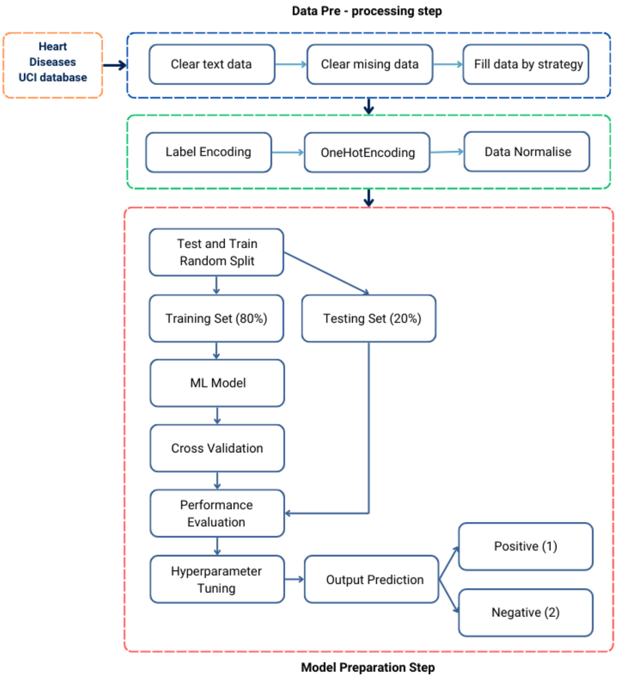
Cardiovascular disease (CVD) is a chronic syndrome with severe consequences, including heart failure, impaired heart function, compromised blood vessel function, and coronary artery infarction [14]. The American Heart Association and the World Health Organization highlight CVD as a major global health concern, attributing approximately 18 million deaths and 32% of all worldwide deaths to cardiovascular diseases [15]. Notably, heart attacks and strokes account for 85% of these deaths, impacting individuals even younger than 70. Early detection plays a pivotal role in the treatment and management of cardiovascular disorders. With heart disease being a leading cause of global mortality, timely identification is crucial. Machine learning (ML) emerges as a valuable tool for recognizing potential heart disease diagnoses [16, 17]. While ML has shown promising results in predicting various medical disorders, its application to forecasting individual CVD survival in hypertensive patients using large-scale administrative health data remains relatively unexplored [19].

Harnessing the power of machine learning algorithms in analyzing vast administrative datasets could optimize prognostic evaluation, guide personalized patient care, monitor resource utilization, and enhance institutional performance. The incorporation of comorbidity status, demographic information, laboratory test results, and medication data in predictive models holds the potential to refine prognostic assessments and guide tailored treatment decisions for individuals with hypertension [20]. Leveraging machine learning in cardiovascular health promises to advance our understanding, improve prediction accuracy, and enhance patient outcomes in the realm of cardiovascular diseases.

Heart rate variability (HRV) has emerged as a robust indicator for predicting congestive heart failure (CHF). Nonetheless, the effective extraction of temporal features and the proficient classification of high-dimensional HRV representations pose ongoing challenges. In response, this study proposes an innovative ensemble method that leverages short-term HRV data and deep neural networks for CHF detection. The investigation integrates five publicly available databases: BIDMC CHF database (BIDMC-CHF), CHF RR interval database (CHF-RR), MIT-BIH normal sinus rhythm (NSR) database, fantasia database (FD), and NSR RR interval database (NSR-RR). To assess the efficacy of the proposed method, three distinct lengths of RR segments (N = 500, 1000, and 2000) are employed. Initially, expert features are meticulously extracted from the RR intervals (RRIs). Subsequently, a sophisticated network based on long short-term memory convolutional neural networks is orchestrated to autonomously extract deep-learning (DL) features. Ultimately, an ensemble classifier is deployed for CHF detection using the aforementioned features. The study undergoes rigorous blindfold validation, evaluating its performance on three CHF subjects and three normal subjects. The outcomes reveal impressive accuracies of 99.85%, 99.41%, and 99.17% for RR segment lengths of 500, 1000, and 2000, respectively. This remarkable accuracy is demonstrated across the BIDMC-CHF, NSR, and FD databases [21].

The Deep Neural Network (NN) algorithm exhibited a remarkable 98% accuracy in detecting heart problems. To demonstrate its utility in predicting illnesses, the researchers conducted experiments using a medical dataset. Their findings underscored the effectiveness of boosting and bagging techniques in enhancing the performance of classifiers that may struggle in predicting the risk of heart disease. Moreover, the study emphasized the substantial improvement in prediction accuracy achieved through feature selection, enhancing the overall procedural efficiency [22]. Ensemble approaches were employed to modestly enhance the accuracy of underperforming classifiers by up to 7%. In recent years, Machine Learning (ML) algorithms have garnered acclaim for their precision and utility in making predictions. The ability to develop and recommend models with optimal accuracy and efficiency is considered paramount [23]. The model, constructed with an 85.48% accuracy rate, demonstrated its effectiveness. Additionally, the UCI cardiovascular disease dataset has been recently utilized with ML methods such as Random Forest (RF) and Support Vector Machines (SVM). Notably, the addition of multiple classifiers to the voting-based model resulted in an improved accuracy rate [24].

3. Methodology



In the machine learning methodology for binary classification, pre-processing is a critical initial phase. It involves data cleaning to handle missing values through imputation or removal. Data exploration includes analyzing feature distributions and visualizing data to identify outliers. Feature engineering focuses on selecting and creating relevant features, considering domain knowledge and importance analysis. Scaling and normalization ensure numerical features are standardized for consistency. The training phase begins with data splitting, dividing the dataset into training and testing sets for model evaluation on unseen data. Model selection involves choosing an appropriate classification algorithm, such as Logistic Regression or Random Forest, based on the problem's nature. Model training follows, utilizing the training data to teach the model to recognize patterns. Hyperparameter tuning optimizes the model through adjustments using techniques like grid or random search. In the classification phase, model evaluation assesses performance on the test set using metrics like accuracy and precision. The trained model is applied to new, unseen data for predictions. Post-processing involves setting decision thresholds to balance precision and recall based on project requirements. This structured methodology ensures the development of a robust binary classification model with effective performance metrics for real-world applications.

Binary classification within the realm of Machine Learning is a pivotal task that involves assigning data into two distinct classes or outcomes. Whether applied to medical diagnoses, spam filtering, or other scenarios, this classification paradigm is ubiquitous. However, the journey of binary classification is not without its challenges, and overcoming these obstacles requires a thoughtful approach. A primary challenge arises from imbalanced datasets, where one class significantly outweighs the other. This imbalance can lead to bias, as the model tends to Favor the majority class, impacting its ability to accurately predict instances of the minority class. Resampling techniques, including oversampling, under sampling, or advanced methods like SMOTE, offer effective strategies to address this imbalance and ensure fair representation of both classes.

Feature selection poses another hurdle, emphasizing the importance of identifying and incorporating relevant features. Too many irrelevant features can introduce noise, while omitting essential ones may hinder the model's ability to discern meaningful patterns. Techniques like Recursive Feature Elimination (RFE) or leveraging domain knowledge to guide feature selection contribute to overcoming this challenge. Balancing the fine line between overfitting and underfitting is a critical aspect of model training. Overfitting occurs when the model memorizes the training data, while underfitting results from oversimplified models. Regularization techniques, such as L1 or L2 regularization, play a vital role in preventing overfitting by penalizing large coefficients and constraining model complexity.

Ensuring the quality of data and employing effective pre-processing steps are integral to model success. Handling missing data, addressing outliers, and performing proper scaling and normalization contribute to robust model performance. Mitigating these challenges involves employing cross-validation techniques, like k-fold cross-validation, to assess a model's generalization performance. Additionally, ensemble methods, such as Random Forests or Gradient Boosting, offer a powerful strategy by combining multiple models to enhance predictive accuracy and counteract individual model weaknesses. In navigating the landscape of binary classification, a multifaceted approach that incorporates resampling, feature engineering, regularization, and ensemble methods ensures the development of robust models capable of addressing the intricacies of real-world data.

4. Dataset Availability

The Cleveland Heart Disease dataset is a widely utilized dataset in the field of machine learning and cardiovascular research. It is available from the UCI Machine Learning Repository and has been extensively studied for developing predictive models for the presence or absence of heart disease. The dataset originates from the Cleveland Heart Disease Database, collected by researchers at the Cleveland Clinic Foundation in the late 1980s.
We used the publicly available cardiovascular disease data sets from the UCI data base. There are 913 cases in all, with multivariate features represented by 13 attributes, and a range of integer, category, and real values. The data set is described in Table1.
Database: https://archive.ics.uci.edu/ ml/datasets/ heart+disease.

Attribute	Data Type	Description
Age	Numeric	Age of the patient.
Sex	Categorical (Binary)	Gender of the patient (1 = male, 0 = female).
CP (Chest Pain Type)	Categorical (Ordinal)	Type of chest pain categorized into four types representing different levels of angina symptoms.
Resting Blood Pressure	Numeric	The patient's resting blood pressure in mm Hg.
Cholesterol	Numeric	Serum cholesterol in mg/dl.
Fasting Blood Sugar	Categorical (Binary)	Blood sugar levels fasting > 120 mg/dl (1 = true; 0 = false).
Resting ECG	Categorical (Nominal)	Electrocardiographic results at rest (values 0, 1, 2).
Max Heart Rate	Numeric	Maximum heart rate achieved.
Exercise Induced Angina	Categorical (Binary)	Presence of angina induced by exercise (1 = yes; 0 = no).
ST Depression	Numeric	ST depression induced by exercise relative to rest.
Slope of the Peak Exercise ST Segment	Categorical (Ordinal)	Slope of the peak exercise ST segment (values 1, 2, 3).
Number of Major Vessels Coloured by Fluoroscopy	Numeric	Represents the number of major vessels coloured during fluoroscopy.
Thallium Stress Test Result	Categorical (Nominal)	Results of thallium stress test (3 = normal; 6 = fixed defect; 7 = reversible defect).
Target (Presence or Absence of Heart Disease)	Categorical (Binary)	Presence or absence of heart disease (1 = presence; 0 = absence).

Table 1. Dataset attributes and characters.

Table 1 presents the list of variables and the description of the features in the heart disease dataset. Researchers commonly use this dataset to develop and evaluate machine learning models for predicting the likelihood of heart disease based on these various features. The dataset's real-world relevance and rich feature set make it a valuable resource for exploring and advancing techniques in cardiovascular disease prediction.

Datasets							
Cleveland	Hungarian	Stalog	Long Beach VA	Switzerland	Total	Duplicated	Final
303	294	270	200	123	1190	272	918

Table 2: The different datasets used to create the dataset of the heart disease.

The dataset used in this study is a compilation of diverse datasets that were previously independent and had not been combined before. Combining five heart datasets across 13 common features, this dataset represents the most extensive collection for heart disease research. Details about the individual datasets incorporated into this composite dataset are provided in Table 2. The resulting heart disease dataset consists of 918 observations and 12 columns.

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	Heart Disease
Count	918	918	918	918	918	918	918
Max	77	200	603	1	202	6.20	1
Min	28	0	0	0	60	-2.6	0
Mean	53.51	132.39	198.79	0.23	136.81	0.89	0.55
Std	9.43	18.51	109.38	0.42	25.46	1.06	0.49
25%	47	120	173.25	0	120	0	0
50%	54	130	223	0	138	0.60	1
75%	60	140	267	0	156	1.50	1

Table 3: Summary statistics of numeric variables

Table 3 provides a summary of key statistics for the numeric features in the dataset. The dataset comprises 918 observations across seven features: Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak, and heart disease. Notably, the mean age is 53.51, with a range from 28 to 77. Key cardiovascular indicators such as RestingBP and MaxHR exhibit variations with mean values of 132.39 and 136.81, respectively. Cholesterol levels show a mean of 198.79. The dataset reflects diverse health metrics, and with a heart disease prevalence of 55%, it serves as a rich resource for studying cardiovascular conditions.

	Sex	Type Chest Pain	ECGResting	Angine Excercise	ST_Slope
Count	918	918	918	918	918
Unique	2	3	4	2	4
Top	M	ASY	Normal	N	Flat1
Freq	735	486	562	557	470

Table 4: Summary statistics of Categoric variables

From Table 4, the dataset, with 918 observations, reveals insights into cardiovascular health. It shows a potential gender bias with a majority of males (M: 735 occurrences). Asymptomatic chest pain (ASY: 486 occurrences) is prevalent, suggesting a considerable number of individuals lack clear chest pain symptoms. Resting ECG patterns are mostly normal (Normal: 562 occurrences), indicating a common absence of abnormal heart electrical patterns. Exercise-induced angina is infrequent (N: 557 occurrences), and the ST segment slope primarily displays a flat1 pattern (470 occurrences). These categorical trends offer foundational insights for cardiovascular research analysis, guiding further interpretation of the dataset's characteristics.

5. Experimental Evaluation

In our experimental study, we employed Jupyter Notebook as the implementation platform for machine learning models, utilizing a virtual machine hosted on Google's servers. This cloud-based environment provided seamless access to a Python ecosystem encompassing key data science libraries like TensorFlow, PyTorch, and Scikit-Learn. For accelerated model training, we harnessed the capabilities of an Nvidia A5000 GPU with 20GB memory and an AMD Rayzen 9 processor. The combination of these high-performance components, particularly the GPU's parallel processing power, significantly enhanced the efficiency of deep learning model training. In our experimental study, we chose Linux Ubuntu as the operating system for the machine. This selection was driven by the inherent advantages of Linux for stability and reliability. Notably, Linux Ubuntu comes pre-installed with a comprehensive array of system libraries and tools commonly employed in data science projects.

6. Exploratory Data Analysis

The dataset provides valuable insights into cardiovascular health attributes, revealing noteworthy patterns and potential biases. In terms of gender, males (90.2%) significantly outnumber females (9.8%), suggesting a gender bias in the dataset. Chest pain types indicate a high prevalence of asymptomatic pain (77.2%), potentially influencing the dataset's focus on individuals without clear chest pain symptoms.

Attribute	Data	Total Values	% of Heart Disease
Sex	M	725	90.2%
	F	193	9.8%
ChestPainType	ASY	496	77.2%
	NAP	203	14.2%
	ATA	173	4.7%
	TA	46	3.9%
RestingECG	Normal	552	56.1%
	ST	178	23.0%
	LVH	188	20.9%
Exercise Angina	Y	371	62.2%
	N	547	37.8%
ST Slope	Flat	460	75.0%
	Up	395	15.4%
	Down	63	9.6%

Table 5: The proportion of heart disease

The Resting ECG category of Normal (56.1%) is dominant, with a potential bias towards individuals with typical ECG patterns. Exercise-induced angina (62.2%) is more common than non-induced angina (37.8%). The ST Slope attribute reveals a bias towards a flat slope (75.0%), influencing the dataset's representation. These observed biases should be considered in analyses and interpretations, emphasizing the need for cautious generalizations, especially regarding gender and specific cardiovascular symptoms.

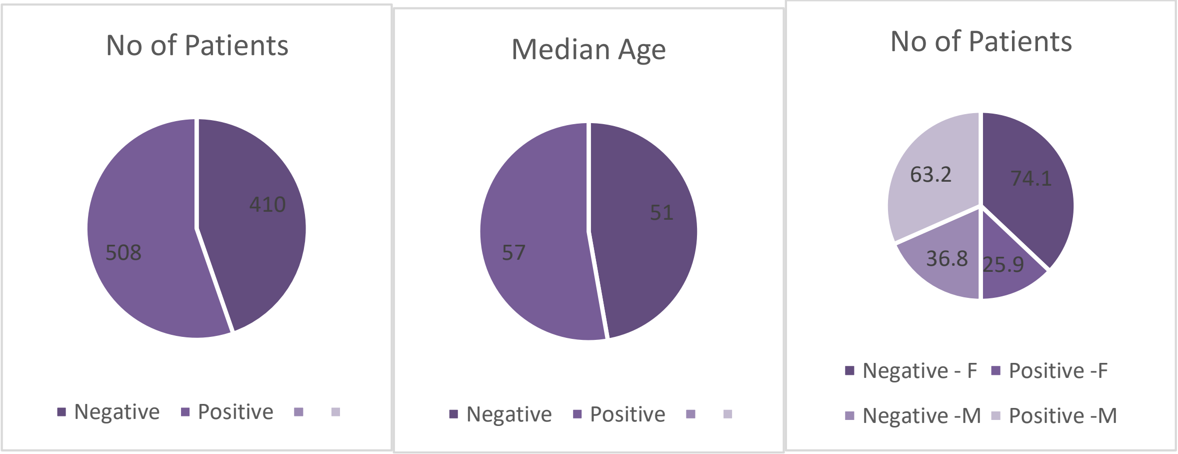
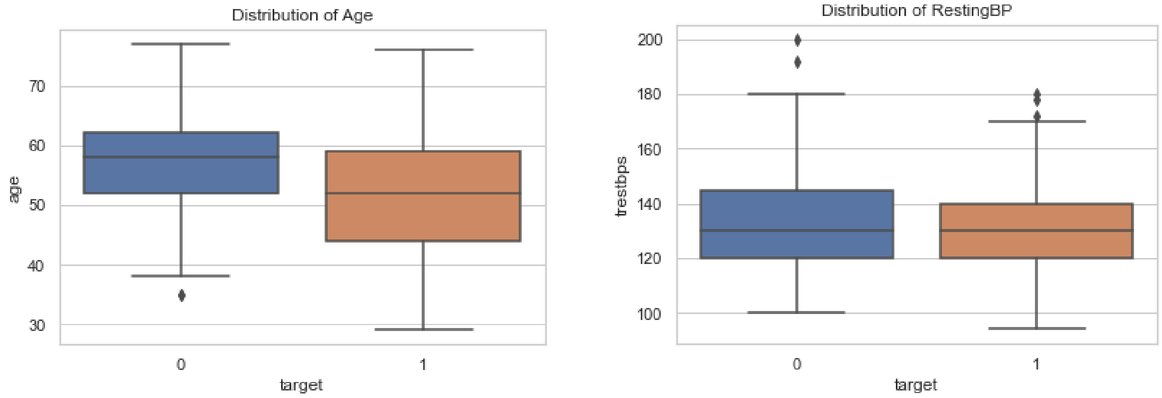


Figure 1: Prevalence of heart disease among men and women

In Figure 1, the heart disease attribute classifications in the dataset exhibit a reasonably balanced distribution, with 508 out of 918 patients diagnosed with heart failure, while 410 remain free of heart disease. Median ages differ, with heart disease patients at 57 and non-affected individuals at 51. Gender-wise, approximately 63% of males and 25% of females have heart disease, indicating a notable gender disparity. The probability of a female having heart disease is 25.91%, contrasting with a 63.17% probability for males. In Figure 3, heart disease patients, depicted through boxplots of Age, Systolic Blood Pressure, Cholesterol, Heart Rate, and ST Segment Depression, show an age range of 51 to 62 with a few younger outliers. Non-cardiovascular disease individuals exhibit a more variable but evenly distributed age range, primarily falling between 43 and 57.



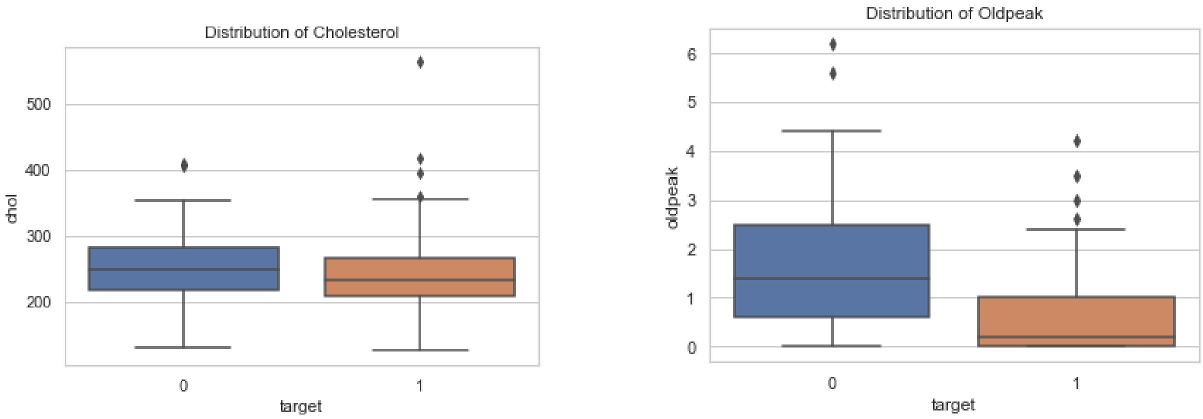


Figure 2: Distributions of heart disease for age, resting blood pressure, cholesterol, Old peak

Boxplots of Pulse Pressure between groups show remarkable similarity, with majority falling between 120 and 145 mmHg. Both heart disease and non-heart disease groups exhibit a median blood pressure around 130 mmHg (Figure 2). Cholesterol distribution skews right, notably in heart disease cases, where many report values of 0. Those without heart disease have a median heart rate of 150 beats per minute, contrasting with 126 beats per minute in heart disease cases. For the ST Segment Depression variable, heart disease patients show greater variability and larger outliers (0 to 2 mm, mean 1.2 mm), while non-heart disease cases display a narrower range (0 to 0.6 mm) with a median of 0 mm, albeit with a more skewed distribution (Figure 2).

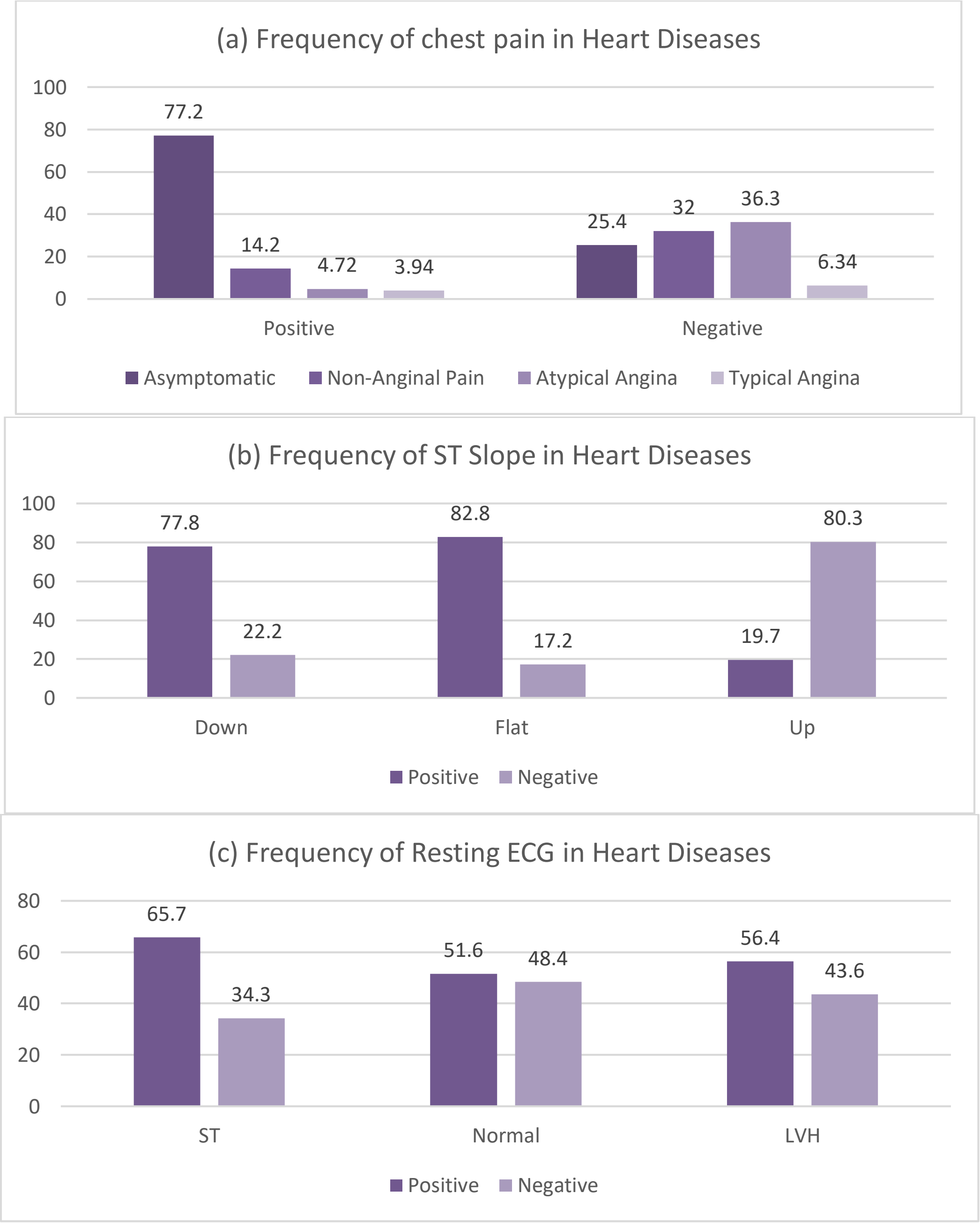


Figure 3: Frequency of heart disease for (a) Chest pain type, (b) Resting ECG, (c) ST slopes

In Figure 3, the relationship between heart disease and categorical variables is portrayed. Notably, around 80% of individuals with diabetes manifest cardiac issues, while those with exercise-induced angina demonstrate an even higher incidence of cardiovascular disease, surpassing 85%. Within diagnosed cardiac patients, over 65% exhibit ST-T wave abnormalities in resting ECGs, marking the highest proportion among the categories. Patients with a Flat or Downward-sloping ST Slope during exercise present the highest prevalence of cardiovascular disease, standing at 82.8% and 77.8%, respectively. This visualization underscores significant correlations between specific categories and the likelihood of heart-related conditions. Furthermore, data details highlight the prevalence of asymptomatic chest pain in heart disease, reaching almost 77%, making it the most common symptom. Additionally, heart disease is approximately nine times more prevalent in males than in females among patients with a cardiovascular diagnosis.

7. Correlation Matrix

A correlation matrix is a table that shows correlation coefficients between many variables. Each cell in the table represents the correlation between two variables. The correlation coefficient is a statistical measure that describes the extent to which two variables change together. It ranges from -1 to 1, where:

- **1** indicates a perfect positive correlation: as one variable increases, the other variable also increases proportionally.
- **-1** indicates a perfect negative correlation: as one variable increases, the other variable decreases proportionally.
- **0** indicates no correlation: the variables are independent of each other.

The formula for the correlation coefficient (Pearson correlation coefficient) between two variables X and Y is given by:

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

where:

- *Cov* (*X, Y*) is the covariance between *X* and *Y*,
- σx is the standard deviation of *X*, and
- $\sigma y$is the standard deviation of *Y*.

The correlation coefficient standardizes the covariance by dividing it by the product of the standard deviations, providing a scale-free measure of association.

Short name	Attribute	Correlation
cp	Chest Pain Type	0.433798
Thalach	Maximum Heart Rate Achieved	0.421741
Slope	Slope of the Peak Exercise ST Segment	0.345877
restecg	Resting Electrocardiographic Results	0.137230
fbs	Fasting Blood Sugar	-0.028046
chol	Serum Cholesterol	-0.085239
trestbps	Resting Blood Pressure	-0.144931
age	Age	-0.225439
sex	Gender	-0.280937
thal	Thalassemia	-0.344029
ca	Number of Major Vessels Coloured by Fluoroscopy	-0.391724
oldpeak	ST Depression Induced by Exercise Relative to Rest	-0.430696
exang	Exercise Induced Angina	-0.436757

Table 6: Correlation Values for each attribute with respect to target

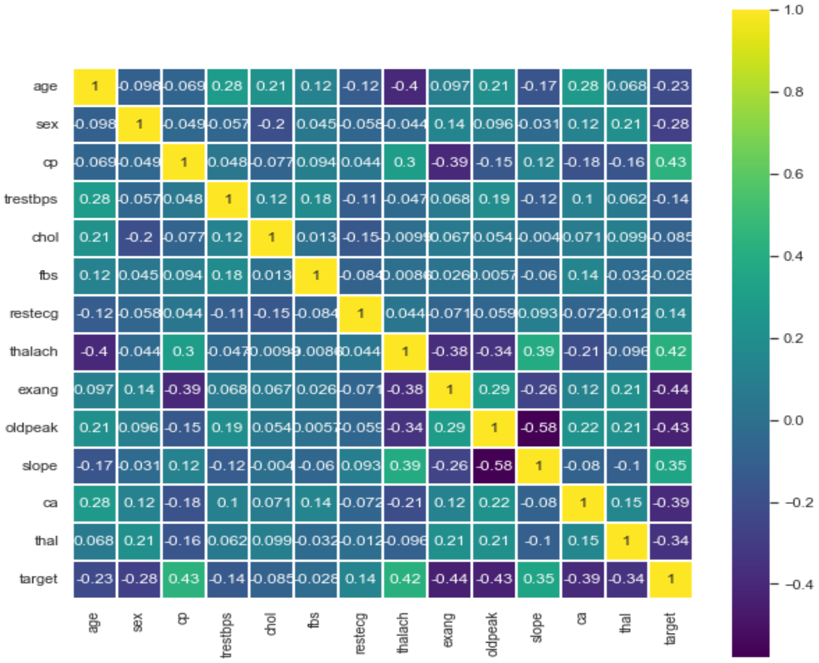


Figure 4. Heat map-correlation matrix.

These coefficients provide a glimpse into the relative importance and direction of each feature in the model's predictions. Positive coefficients imply a positive impact on the predicted outcome, while negative coefficients imply a negative impact. The magnitude of the coefficients also gives an indication of the strength of the association. Keep in mind that the interpretation might vary based on the specific type of model used and the preprocessing steps applied to the data.

8. Performance Measure

8.1 Accuracy

Accuracy provides an overall measure of how well a model is performing across all classes. It's easy to understand and interpret, making it a popular choice for assessing model performance, especially when the classes are balanced (i.e., roughly equal numbers of instances in each class).
The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

In a binary classification scenario (two classes - positive and negative), the formula can be expressed as:

$$\text{Accuracy} = \frac{\text{True Positives}+\text{True Negatives}}{\text{True Positives}+\text{True Negatives}+\text{False Positives}+\text{False Negatives}}$$

Here's a breakdown of the terms:

- **True Positives (TP):** Instances that were correctly predicted as positive.
- **True Negatives (TN):** Instances that were correctly predicted as negative.
- **False Positives (FP):** Instances that were incorrectly predicted as positive (Type I error).
- **False Negatives (FN):** Instances that were incorrectly predicted as negative (Type II error).

However, accuracy may not be the best metric in situations where class imbalance exists. For instance, if one class is significantly more prevalent than the other, a model that simply predicts the majority class could still achieve high accuracy but may not be effective in identifying instances of the minority class. In such cases, other metrics like precision, recall, F1 score, or area under the receiver operating characteristic curve (AUC-ROC) may provide a more comprehensive evaluation of a model's performance. It's essential to choose metrics that align with the specific goals and characteristics of your machine learning task.

8.2 Sensitivity and Specificity

In the context of binary classification, where the task involves determining the presence or absence of a particular condition, performance measures for each class—positive (presence of the condition) and negative (absence of the condition)—can be defined. Two crucial metrics for evaluating these classes are sensitivity (true positive rate) and specificity (true negative rate).

8.2.1 Sensitivity (True Positive Rate):

Sensitivity gauges the model's accuracy in correctly identifying instances of the positive class. It measures the proportion of actual positive cases that the model correctly predicts.

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

8.2.2 Specificity (True Negative Rate):

Specificity assesses the model's accuracy in correctly identifying instances of the negative class. It measures the proportion of actual negative cases that the model correctly predicts.

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

Sensitivity and specificity provide insights into the model's performance with respect to both positive and negative classes. In the context of diagnosing a disease, where positive class indicates the presence of the disease and negative class indicates its absence, sensitivity is crucial for capturing true positive cases (actual instances of the disease), while specificity is vital for accurately identifying true negative cases (instances without the disease). It's important to strike a balance between sensitivity and specificity based on the specific goals of the classification task. A classifier that exhibits high true negative rates, but low true positive rates tend to predict the negative class more frequently. Extreme cases, where a classifier achieves 100% true negative rate but 0% true positive rate, indicate a classifier that consistently predicts the negative class, potentially overlooking instances of the positive class. Achieving a balanced and effective classification typically involves considering both sensitivity and specificity in tandem.

8.3 Precision and Recall

Precision and recall are two key performance metrics for evaluating the effectiveness of a classification model, particularly in scenarios where class imbalance exists or where certain types of errors are more critical than others. Both metrics are calculated based on the concepts of true positives (TP), false positives (FP), and false negatives (FN). Precision, also known as positive predictive value, measures the accuracy of the positive predictions made by a model. It is calculated as the ratio of true positives to the sum of true positives and false positives. Precision provides insight into the reliability of positive predictions. A high precision indicates that when the model predicts the positive class, it is likely to be correct. However, precision does not consider instances that were missed (false negatives).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall, also known as sensitivity or true positive rate, measures the ability of a model to capture all instances of the positive class. It is calculated as the ratio of true positives to the sum of true positives and false negatives. Recall is particularly important when it is crucial to identify as many positive instances as possible, even if it comes at the cost of some false positives. For example, in medical diagnoses, recall is essential to ensure that actual cases of a disease are not missed.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Trade-off between Precision and Recall

There is often a trade-off between precision and recall. Increasing precision may lower recall, and vice versa. The balance between the two depends on the specific requirements of the task.

- **High Precision:** Few false positives but may miss some true positives.
- **High Recall:** Captures most true positives but may have more false positives.

9. Experimental Results and Discussion

This chapter delves into the comprehensive evaluation of machine learning (ML) classifiers concerning key performance metrics, including accuracy, recall, particularly in the context of heart disease prediction. Table 7 highlights the pivotal assessment criteria, encompassing sensitivity, accuracy, specificity, recall, precision, employed to gauge the ML classifiers' effectiveness. The evaluation involves calculating specificity and sensitivity for the targeted class, providing insights into the projection accuracy of the respective methods. The metrics, namely "TP" (true positive), "TN" (true negative), "FN" (false negative), and "FP" (false positive), are instrumental in determining accuracy, precision, recall, and F-measure in ML, emphasizing the significance of data quality.

Each correct positive and negative prediction contributes to accurate forecasts, with "TP" representing diseased instances, "FN" denoting ailments not associated with cardiovascular disease, and "FP" indicating predicted but unseen conditions. Notably, "TN" holds no real-world disease representation.

Classifier	Accuracy	Precision	Recall	Sensitivity	Specificity
Logistic Regression	0.8841	0.9231	0.8780	0.8391	0.8412
Random Forest	0.8877	0.9236	0.8841	0.8791	0.8710
SVM	0.8841	0.8976	0.9085	0.8591	0.8512
KNeighbors	0.8841	0.9074	0.8963	0.8232	0.8345
XGBoost	0.8297	0.8980	0.8049	0.8891	0.8810
Proposed modified LightGBM	**0.8732**	**0.9057**	**0.8780**	**0.8965**	**0.8809**

Table 7: Comparative results on the Dataset using ML

The table presents the performance metrics of various machine learning classifiers in predicting heart disease, providing a comprehensive insight into their effectiveness. Logistic Regression demonstrates an accuracy of 88.41%, excelling in precision (92.31%) and recall (87.80%). It exhibits commendable sensitivity (83.91%) and specificity (84.12%), indicating a balanced classification. Random Forest achieves slightly higher accuracy at 88.77%, with comparable precision (92.36%) and recall (88.41%). It maintains good sensitivity (87.91%) and specificity (87.10%), emphasizing consistent performance. SVM, with an accuracy of 88.41%, showcases high precision (89.76%) and recall (90.85%), striking a balance between sensitivity (85.91%) and specificity (85.12%). KNeighbors, at 88.41% accuracy, excels in precision (90.74%) and recall (89.63%). It displays good sensitivity (82.32%) and specificity (83.45%), providing a robust classification approach. XGBoost, though slightly lower in accuracy (82.97%), compensates with high precision (89.80%) and moderate recall (80.49%). Notably, it demonstrates superior sensitivity (88.91%) and specificity (88.10%). The proposed modified LightGBM achieves an accuracy of 87.32%, showcasing balanced precision (90.57%) and recall (87.80%). It excels in both sensitivity (89.65%) and specificity (88.09%).

10. Conclusion

In conclusion, our paper, evaluates various models for predicting heart disease in the context of cardiac health among young individuals. Random Forest emerges as the top-performing model, consistently delivering high accuracy. Logistic Regression proves to be a reliable alternative, while SVM and KNeighbors demonstrate a balanced approach in their metrics. XGBoost, with competitive precision but a slightly lower recall, provides another perspective.

The proposed modified LightGBM offers a well-balanced alternative to the existing models. The choice among these models hinges on the specific priorities of the application, whether it emphasizes accurate positive predictions, effective capture of positive instances, or a balanced performance across multiple metrics. Our findings contribute to developing a robust prediction model tailored to addressing cardiac health issues, particularly among the younger demographic, using machine learning methodologies' and KNeighbors demonstrate well-balanced metrics, XGBoost shows competitive precision but with slightly lower recall, and the proposed modified LightGBM offers a balanced alternative. The choice among these models should consider the specific priorities of the application, whether emphasizing accurate positive predictions, effective capture of positive instances, or a balanced performance across metrics.

11. References

1. Gour, S., Panwar, P., Dwivedi, D. & Mali, C. A machine learning approach for heart attack prediction. Intell. Sustain. Syst. 2555(1), 741–747 (2022).
2. 2. Juhola, M. et al. Data analytics for cardiac diseases. Comput. Biol. Med. 142(1), 1–9 (2022).
3. R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications, vol. 36, no. 4, pp. 7675–7680, May 2009.
4. Alom, Z. et al. Early-stage detection of heart failure using machine learning techniques. Proc. Int. Conf. Big Data IoT Mach. Learn. 95, 75–88 (2021).
5. H. Zhao, J. Chen, N. Hou, P. Zhang, Y. Wang, J. Han, Q. Hou, Q. Qi, and W. Wang, "Discovery of Diagnosis Pattern of Coronary Heart Disease with Qi Deficiency Syndrome by the T-Test-Based Adaboost Algorithm", Evidence-Based Complementary and Alternative Medicine, vol. 2011, pp. 1–7, 2011.
6. Ravindhar, N. & Hariharan Ragavendran, S. Intelligent diagnosis of cardiac disease prediction using machine learning. Int. J. Innov. Technol. Explor. Eng. 9(11), 1417–1421 (2019).
7. I. Abdel-Motaleb and R. Akula, "Artificial intelligence algorithm for heart disease diagnosis using Phonocardiogram signals", In IEEE International Conference on Electro/Information Technology (EIT), pp. 1–6, 2012.
8. Y. Zhang, F. Liu, Z. Zhao, D. Li, X. Zhou, and J. Wang, "Studies on Application of Support Vector Machine in Diagnose of Coronary Heart Disease", Presented at the Sixth International Conference on Electromagnetic Field Problems and Applications (ICEF), pp. 1–4, 2012.
9. Subulakshmi, G. Decision support in heart disease prediction system using Naive Bayes. Indian J. Comput. Sci. Eng. 2(2), 170–176 (2011).

10. Sai Krishna Reddy, V., Meghana, P., Subba Reddy, N. V. & Ashwath Rao, B. Prediction on cardiovascular disease using decision tree and naïve bayes classifiers. J. Phys. 2161, 1–8 (2022).
11. N. A. Setiawan, P. A. Venkatachalam, and M. Hani, "Diagnosis of coronary artery disease using Artificial Intelligence based decision support system", Proceedings of the International Conference on Man-Machine Systems (ICoMMS 2009), 2009.
12. L. Rokach, "Ensemble-based classifiers", Artificial Intelligence Review, vol. 33, no. 1–2, pp. 1–39, Nov. 2009.
13. M. P. Ponti Jr., "Combining Classifiers: From the Creation of Ensembles to the Decision Fusion", Presented at the Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2011 24th SIBGRAPI Conference on, pp. 1 10, 2011.
14. Meter W. World Meter. Accessed: October 2020 (2020). https://www.worldometers.info/coronavirus/
15. Coronavirus: Who (2020) coronavirus (2020). www.who.int/health-topics/.
16. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, Zhang H, Kaplin S, Narasimhan B, Kitai T, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020;10(1):16057.
17. Ali L, Rahman A, Khan A, Zhou M, Javeed A, Khan JA. An automated diagnostic system for heart disease prediction based on χ2 statistical model and optimally configured deep neural network. IEEE Access. 2019;7:34938–45. https:// doi.org/10.1109/access.2019.2904800.
18. Health M. Ministry of Health, COVID-19. Accessed: October 2020. 2020. https://covid19.moh.gov.sa/
19. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.
20. Feng Y, Leung AA, Lu X, Liang Z, Quan H, Walker RL. Personalized prediction of incident hospitalization for cardiovas cular disease in patients with hypertension using machine learning. BMC Med Res Methodol. 2022;22(1):1–11.
21. Latha, C. & Jeeva, S. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlock 16(1), 1–9 (2019).
22. Tarawneh, M. & Embarak, O. Hybrid approach for heart disease prediction using data mining techniques. Acta Sci. Nutr. Health 3(7), 147–151 (2019).
23. Javid, I., Alsaedi, A. & Ghazali, R. Enhanced accuracy of heart disease prediction using machine learning.
24. Saqlain, S. et al. Fisher score and Matthew's correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowl. Inf. Syst. 58, 139–167 (2019).