IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

# PREDICTIVE MODELS IN MACHINE LEARNING FOR CARDIOVASCULAR DISEASE

**[1]M Prathibha A E, [2]M Sanjana S, [3]M Prajwal K, [4]M Rishikesh C**

[1]Undergraduate Student, CSE

[1]Mrs. Ashika S, Assistant Professor, CSE

[1]Sri Venkateshwara College of Engineering, Bengaluru, India

***ABSTRACT:*** Cardiovascular diseases (CVDs) continue to pose significant challenges in healthcare, being a leading cause of mortality worldwide. The complexity of predicting CVDs necessitates advanced expertise and tools due to the wealth of available data in healthcare systems. However, the current healthcare landscape often lacks the requisite analysis tools to uncover crucial relationships and patterns within this data. In response, this study investigates the capabilities of machine learning (ML) and deep learning (DL) techniques in predicting cardiovascular disease (CVD). Highlighting ML's capacity to unearth new genotypes, phenotypes, and risk factors, as well as its ability to model intricate relationships, this paper underscores its role in advancing CVD prediction. Additionally, it delves into the contributions of DL techniques, particularly convolutional neural networks (CNNs), in augmenting medical image recognition, diagnosis, prediction, and assessment. Moreover, the paper discusses the advantages of stacked fusion models, which amalgamate various models' strengths to achieve heightened performance levels.

Ultimately, this research suggests leveraging both ML and DL in conjunction to improve the precision of CVD prediction, advance preventive measures, and effectively identify individuals at high risk for cardiovascular diseases.

**KEYWORDS:** Cardiovascular Disease, Machine Learning, ML and DL, Data set, Data Mining, Algorithms, Random Forest.

## INTRODUCTION:

The paramount significance of health to people worldwide is undeniable, yet chronic ailments, notably cardiovascular diseases (CVDs), persist globally owing to factors like sedentary lifestyles, work-induced stress, environmental contaminants, and inadequate access to healthcare facilities. CVDs, which encompass various conditions affecting the heart and blood vessels, including hypertension, hyperlipidemia, thromboembolism, and coronary heart disease, impose a significant burden on public health systems internationally. Reports indicate that a majority of global deaths are attributed to complications related to CVDs, establishing it as a leading cause of mortality.

Despite advancements in medical science, early diagnosis and treatment of CVDs remain challenging, often leading to fatal outcomes or debilitating disabilities. Conventional risk-assessment models, while extensively employed, often oversimplify the complex interplay of diverse risk factors linked to CVD outcomes, presuming linear associations and neglecting nonlinear interactions among them. Hence, there is an urgent need to comprehensively incorporate multiple risk factors and discern subtle correlations between these factors and CVD outcomes.

The emergence of machine learning (ML) techniques offers promising avenues for improving CVD risk prediction accuracy and enhancing patient outcomes, particularly within primary care settings. ML algorithms such as knearest neighbors (KNN), logistic regression, and RF have demonstrated potential in analyzing extensive clinical data to identify novel risk predictors and unveil intricate relationships among them. By transcending the constraints of traditional regression models, ML approaches can better capture this complexity of CVD risk factors allows for more accurate prognostic evaluations.

Nevertheless, despite the potential advantages of ML in predicting CVD risk, there is a lack of extensive research in the literature concerning its widespread implementation and the comparison of various ML algorithms for this objective. This study aims to fill this gap by assessing the effectiveness of KNN, logistic regression, and random forest algorithms in enhancing cardiovascular risk prediction accuracy among primary care populations. Through a comprehensive review of existing literature and the analysis of routine clinical data, this research endeavors to identify the most effective ML algorithm for improving CVD risk assessment and guiding preventive interventions in primary care settings.

## LITERATURE REVIEW:

Khandaker Mohammad Mohi Uddin study concentrates on utilizing machine learning techniques for diagnosing cardiovascular vascular disease, with a key goal of accurately predicting heart disease to prevent adverse outcomes. By amalgamating three datasets and leveraging various machine learning algorithms, the research achieved notable success, particularly with the algorithm known as Decision Tree attaining an impressive accuracy of 99.16%. Methodological steps included data interpretation, cleansing, and processing, alongside feature engineering, model selection, and deployment. Comparing with existing literature on heart disease prediction highlighted the effectiveness of the proposed approach. Furthermore, the study delved into statistical dataset analysis, feature prioritization, system specifications, and the practical implementation of the diagnostic system as a web application. These findings underscore the potential of machine learning in enhancing the precision of diagnosing cardiovascular disease has substantial implications for disease management and early intervention. The research team explored various supervised learning classifications for diagnosing heart issues, with Decision Tree emerging as particularly effective with its 99.16% accuracy rate [1].

Senthilkumar Mohan study delves into the utilization of Machine Learning (ML) algorithms, particularly focusing on Neural Networks, to predict heart disease. It highlights Carotid Artery Stenting (CAS) as a common procedure among elderly individuals with heart disease. The study showcases the efficacy of the Artificial Neural Network (ANN) model in forecasting heart disease outcomes. Additionally, Convolutional Neural Networks (CNN) are introduced without the need for segmentation., concentrating on heart cycles during testing. The research underscores the significance of cost-effective approaches that improve the accuracy of heart disease prediction, emphasizing the HRFLM method's accuracy in this regard. Recognizing the importance of early detection for preventing adverse health outcomes, the study proposes the hybrid HRFLM approach, which combines aspects of Random Forest (RF) and Logistic Model (LM) to enhance prediction accuracy. Overall, the research emphasizes the importance of integrating advanced ML and deep learning techniques to enhance heart disease prediction accuracy, with the ultimate goal of improving patient care and healthcare practices [2].

Rubini PE study focused on employing AI and Utilizing Machine Learning methodologies for predicting vulnerability to heart disease, emphasizing the critical role of early diagnosis in preventing cardiovascular diseases, a leading global cause of mortality. Utilizing the Random Forest Algorithm, researchers achieved high accuracy in predicting heart disease based on symptoms and risk factors. The dataset underwent thorough cleaning and preprocessing, with categorical data transformed into numerical values. Additionally, Kernel Functions and Logistic Regression were implemented, yielding promising results in heart disease prediction. The study highlighted the potential of AI in automating diagnosis and enhancing cardiovascular disease tracking, with future research aiming to further refine prediction methods. Overall, the study underscored the efficacy of Machine Learning algorithms in forecasting heart disease and emphasized the life-saving importance of early detection [3].

Madhumita Pal, Smita Parija, Ranjan K. Mohapatra, Kuldeep Dhama, and GP delve into the utilization of machine learning algorithms such as K-NN and MLP for predicting cardiovascular disease (CVD). They stress the importance of early CVD detection to prevent fatalities, noting that the MLP model outperforms K-NN in accuracy. Variables like chest pain type, fasting blood sugar levels, and exercise-induced angina play a significant role in predicting CVD. The study underscores the relevance of specific features like resting electrocardiographic results and old peak values in predicting CVD occurrence. Through data visualization plots, it's revealed that individuals aged 40 to 55 with heart rates surpassing 140 are at higher risk of CVD. The authors, including Madhumita Pal and others, actively contribute to the research and take responsibility for the manuscript's content. In summary, the research aims to enhance cardiovascular risk forecasting through the application of standard clinical data and machine learning methods.

Ahmed M. Alaa, Emanuele Di Angelantonio, James H. F. Rudd, and Mihaela van der Schaar aimed to improve cardiovascular disease (CVD) risk prediction using machine learning techniques, specifically the AutoPrognosis model. By integrating unconventional variables such as walking pace and self-reported health rating, the study demonstrated a significant enhancement in CVD risk prediction accuracy compared to traditional methods. The AutoPrognosis model displayed superior performance in forecasting CVD cases compared to the Framingham score, particularly among individuals with diabetes. The incorporation of a broader range of information in predictive models led to improved risk prediction outcomes. Additionally, the research underscored how the AutoPrognosis model effectively addresses patient subgroups typically neglected by current risk prediction models, suggesting its potential to address existing healthcare disparities. These findings underscore the importance of machine learning approaches in refining CVD risk prediction and ultimately advancing patient outcomes.

Subramani S, Varshney N, Anand MV, Soudagar MM, Al-keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K, and Rohini K. concentrate on the utilization of machine learning models and AI technologies for predicting cardiovascular disease (CVD) outcomes. They employ the SHAP methodology to isolate and evaluate individual characteristics and feature interactions, thereby enhancing outcome prediction accuracy. Feature selection utilizing the Tree SHAP approach aids in identifying significant features within the dataset, thereby improving model performance. The research incorporates Gradient Boosting Decision Trees (GBDT) employed as an integrated approach for feature selection, which demonstrates superior prediction performance compared to alternative methods. Furthermore, they employ stacking models with base learners like RF, LR, MLP, ET, and CatBoost to augment prediction accuracy. The study underscores the importance of standardizing medical data and utilizing AI to analyze large datasets for accurate disease prediction. Their research aims to contribute to the advancement of CVD prediction models by integrating deep learning techniques and IoT data, with the ultimate goal of enhancing patient outcomes and optimizing healthcare resources.

## OBJECTIVES:

1. User-Friendly Platform: This project aims to create an intuitive platform for patients to input medical information easily. By leveraging this data, the algorithm accurately detects heart disease presence and type with minimal human intervention. This enhances accuracy and speeds up identification of potential heart issues for both medical professionals and patients, even without direct supervision.

2. Automated Disease Detection: The algorithm autonomously analyzes input medical parameters like age and cholesterol levels to generate results swiftly and accurately. This reduces errors and allows for prompt action by patients and healthcare providers, expediting the healthcare process.

3. Tailored Recommendations: In addition to identifying heart disease types, the project provides personalized precautions to mitigate the impact of identified conditions. These tailored recommendations aid in treatment planning for both medical practitioners and patients, facilitating effective healthcare strategies.

4. Optimized Data Usage: Recognizing the importance of ample annotated data, the project emphasizes efficient utilization of available samples. Through robust data pre-processing techniques, it maximizes the use of existing data, even when dealing with limited sample sizes in medical datasets. This enhances data quality and prepares it for precise algorithmic analysis, improving overall model performance.

## PROPOSED METHODS:

Globally, cardiovascular diseases (CVD) present a substantial healthcare challenge, underscoring the need for precise prediction models to facilitate early intervention and preventive measures. Lately, algorithms from the field of machine learning have emerged as formidable tools for forecasting CVD risk, offering potential enhancements in both accuracy and efficiency when compared to traditional approaches. Among these algorithms, k-nearest neighbors (KNN), logistic regression, and random forest have attracted interest due to their effectiveness in predictive analytics. The objective of this proposed system is to harness the capabilities of these algorithms to construct a robust prediction model tailored for CVD.

### 1. K-NEAREST NEIGHBOR (KNN) ALGORITHM

The K-nearest neighbors (KNN) algorithm identifies similarities between predictors and dataset values without assuming a specific functional form. It's known as a non-parametric approach, offering flexibility in handling diverse data types. Known as a passive classifier, KNN memorizes training data instead of learning fixed weights, making most computations during classification. It assigns a new feature to the class closest to its location in the feature space. KNN is appreciated for its simplicity and versatility in handling both classification and regression tasks. However, it can become computationally intensive for large datasets due to the necessity of storing and comparing distances to all training examples during prediction. Moreover, KNN's performance may suffer with high-dimensional or sparse data. Nonetheless, it remains commonly utilized for its straightforward methodology and effectiveness across various applications in machine learning.

### 2. LOGISTIC REGRESSION

Logistic regression is often utilized as a statistical technique for binary classification purposes, particularly in predicting the presence or absence of cardiovascular disease (CVD). Unlike linear regression, which presupposes linearity, logistic regression utilizes a logistic function to estimate the probability of a binary outcome, mapping input features to a probability distribution. In the realm of CVD prediction, logistic regression scrutinizes patient data, integrating variables such as age, blood pressure, cholesterol levels, and smoking status to evaluate the probability of developing cardiovascular issues. This analysis provides comprehensible insights into the relationship between these variables and the probability of CVD occurrence.

### 3.RANDOM FOREST

The random forest algorithm is a technique in ensemble learning that merges numerous utilizing decision trees to enhance prediction accuracy and reliability. In the random forest methodology, each decision tree is trained on a portion of the dataset, generating independent predictions that are subsequently combined to generate a final prediction. In the realm of CVD prediction, random forest employs patient data to construct an ensemble of decision trees, with each tree analyzing specific subsets of features to uncover patterns related to cardiovascular risk. By combining the predictive capabilities of multiple trees, random forest becomes a powerful tool for assessing CVD risk, adept at capturing complex interactions among risk factors.

**METHODOLOGY:**

Random Forest (RF) ensemble learning methodology is commonly employed in many supervised procedures. It integrates numerous decision trees, each trained on a segment of the training data and input features. Despite using distinct characteristics and data sets, each decision tree within a random forest is constructed similarly to a standalone decision tree. This inherent unpredictability enhances the model's generalizability and mitigates overfitting. In regression scenarios, the random forest technique forecasts the average of all results, whereas in classification scenarios, it predicts results by a simple majority. Employing this ensemble technique can improve the model's accuracy and reliability, especially when confronted with noisy or incomplete data.

Random forest accepts both numeric and categorical input characteristics, making it adept at handling data with missing values. The feature importance ratings provided by random forest can help identify crucial features for predictions. Apart from accuracy, alternative metrics can be utilized to assess the efficacy of a random forest. Generally, it is preferred over single-decision trees because of its capability to reduce model variance, thus providing more consistent and reliable predictions. However, random forests can be computationally intensive and may not perform optimally on very large datasets.

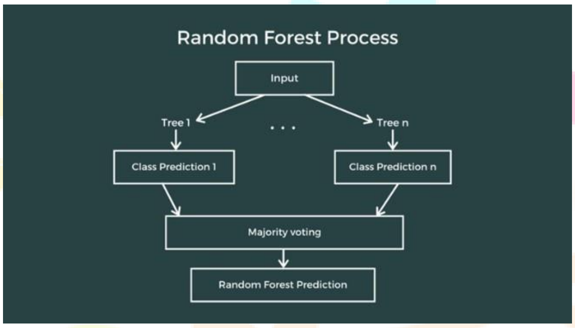

FIG (1)

**PROPOSED SYSTEMS** • Data Pre-processing:

Data pre-processing entails rectifying any inconsistencies or errors present in the dataset. This includes addressing outliers, managing missing values, and performing data normalization to ensure the dataset's appropriateness for model training.

- Feature Selection:

  Feature extraction seeks to identify the most pertinent features relevant to solving the given problem. This process involves eliminating redundant or irrelevant features, thereby reducing the model's complexity and enhancing its performance.

- Model Selection:

  Opting for the suitable method involves selecting the most effective algorithm for addressing the particular issue. This necessitates assessing the effectiveness of different algorithms on the dataset to determine the most appropriate one.

- Model Evaluation:

  The final stage involves evaluating the selected model using the test dataset. This entails comparing the anticipated outcomes with the actual results produced by the model. In our study, we conducted exploratory data analysis on the patient dataset attributes, including health metrics, to differentiate between individuals afflicted by disease and those who are healthy or unhealthy.

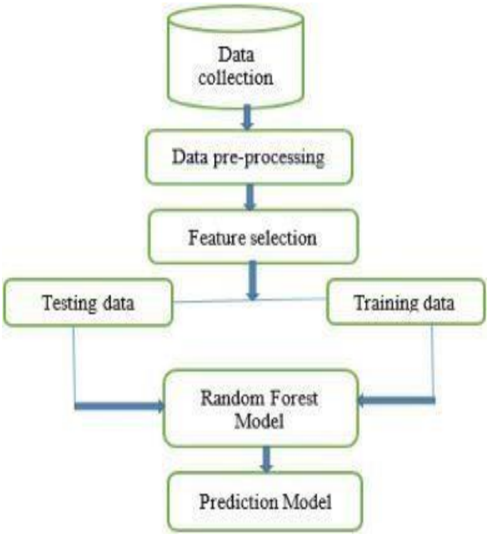

FIG (2)

**ALGORTTHM**

- Step 1: Each decision tree in the Random Forest model receives a set of features and data points. In simpler terms, n random records and m attributes are selected from a k-record data collection.
- Step 2: For every sample, a unique decision tree is constructed.
- Step 3: A result will be obtained from each decision tree.
- Step 4: Majority voting or averaging are used to assess the outcomes of the classification and regression.

**EXPERIMENTAL PROCEDURE AND DATA ANALYSIS**

Utilizing a dataset containing 14 health-related attributes for 301 patients of mixed gender (both male and female), various aspects of the patients' health data were extracted. Initially, the patients were categorized into two groups depending on their gender and health status: 206 males and 96 females. Among these, there were 92 male patients with ailments and 114 male patients in good health, while 24 females were healthy and 74 were ill. Figure 3 illustrates the relationship between the patients' health status and all variables.

Additionally, a correlation map was generated to compare all health metrics related to the patients, providing comprehensive understanding their health status concerning the likelihood of experiencing heart attacks.

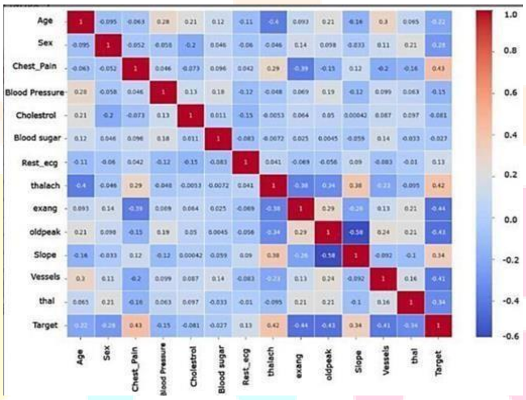

FIG (3)

**RESULTS:**

Comparing Model and Building Model - The research problem involves the utilization of seven distinct machine learning techniques. The Logistic Regression model demonstrated an accuracy of 85.25% on the trained dataset. Similarly, the Naive Bayes method achieved an accuracy score of 85.25%, mirroring that of the Logistic Regression model. The SVM simulation yielded a performance score of 81.97%. While other algorithms were evaluated, the model based on Decision Trees scored approximately 81.97%, K-nearest neighbor scored 67.21%, the model utilizing Random Forest scored 88.52%, XG-boost scored around 78.69%, and the neural network approach scored roughly 83.61%.

Upon analyzing all training algorithms, observations revealed that Random Forest technique produced the most likely outcome in 88.52% of all models. Consequently, the final predictive model for forecasting the risk of heart attacks in patients, based on their medical data characteristics, is built using this approach. Additionally, an interface was developed to incorporate all variables that could be utilized to determine a patient's high-risk status based on their health indicators. The designed Web interface for heart disease parameters is illustrated in Figure 4

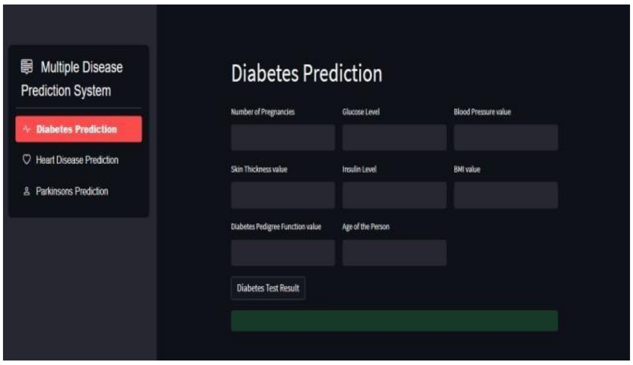

FIG (4)

## CONCLUSION:

Compared to other models such as K-nearest neighbor (KNN) and logistic regression, the Random Forest model demonstrated superior precision in forecasting the likelihood of heart attacks in patients using their medical data attributes. While KNN and logistic regression showed respectable accuracy scores, the Random Forest model consistently outperformed them, achieving an accuracy score of 88.52%. This indicates that Random Forest is an effective approach for CVD risk prediction, as it yielded the highest accuracy among the evaluated models. Thus, drawing from the comparative analysis, Random Forest emerges as the foremost effective model for accurately forecasting the likelihood of heart attacks in patients.

## REFERENCES:

1. Khandaker Mohammad Mohi Uddin, Rokaiya Ripa, Nilufar Yeasmin , Nitish Biswas , Samrat Kumar Dey , "Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset", Elsevier science direct, 2017.

2. SENTHILKUMAR MOHAN , CHANDRASEGAR THIRUMALAI1 , AND GAUTAM SRIVASTAVA, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE EXPLORER, date of publication June 19, 2019

3. Rubini PE1, Dr.C.A.Subasini2, Dr.A.Vanitha Katharine3, V.Kumaresan4,S.GowdhamKumar5, T.M. Nithya. "A Cardiovascular Disease Prediction usingMachine Learning Algorithms". Annals of R.S.C.B, February 2021.

4. World Health Organization. Global Status Report on Noncommunicable Diseases Geneva, Switzerland: World Health Organization, 2014. [Google Scholar]

5. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA Guideline regarding the Evaluation of Cardiovascular Risk: A Report from the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2013; 135(11): 1–50. [PubMed] [Google Scholar]

6. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Forecasting cardiovascular risk in England and Wales: prospective development and validation of QRISK2. BMJ 2008; 336(7659): 1475–82. 10.1136/bmj.39609.449676.25 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

7. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. Comprehensive Cardiovascular Risk Profile for Primary Care Implementation: The Framingham Heart Study. Circulation 2008; 117(6): 743–53. 10.1161/CIRCULATIONAHA.107.699579 [PubMed] [CrossRef] [Google Scholar]

8. Ridker P, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The reynolds risk score. JAMA 2007; 297(6): 611–9. 10.1001/jama.297.6.611 [PubMed] [CrossRef] [Google Schol

9. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AM, Kastelein JJP, et al. Rosuvastatin to Prevent Vascular Events in Men and Women with Elevated C-Reactive Protein. New England

Journal of Medicine 2008; 359(21): 2195–207. 10.1056/NEJMoa0807646 [PubMed] [CrossRef] [Google Scholar]

10. K. Akyol, E. Çalik, Ş. Bayir, B. Şen, A. Çavuşoɫlu  Analysis of demographic characteristics creating coronary artery disease susceptibility using random forests classifier
Procedia Comput Sci, 62 (Scse) (2015), pp. 39-46, 10.1016/j.procs.2015.08.407

11. M. Pal, S. Parija
Prediction of heart diseases using random forest
J. Phys. Conf. Ser., 1817 (1) (2021), 10.1088/1742-6596/1817/1/012009

12. F.S. Alotaibi
Implementation utilization of a machine learning model for heart failure disease prediction Int J Adv Comput Sci Appl, 10 (6) (2019), pp. 261-268, 10.14569/ijacsa.2019.0100637

13. X. Su, et al.
Cardiovascular disease prediction using laboratory data: An examination of the random forest model J Clin Lab Anal, 34 (9) (2020), pp. 1-10, 10.1002/jcla.23421
P.A. Juma et al. authored a study on the non-communicable disease prevention policy process in five African countries, which was published in BMC Public Health in 2018.

14. The World Health Organization published the Global Status Report on Noncommunicable Diseases in 2014.

15. The WHO released the Regional Action Plan for Communicable Diseases in the Western Pacific in 2014.
16. Q. Wang et al. investigated out-of-pocket expenditure on chronic non-communicable diseases in  Sub-Saharan Africa, focusing on rural Malawi. Their findings were published in PLoS One in 2015.

17. R. Katarya and P. Srinivas conducted a survey on predicting heart disease at early stages using machine learning, presented at the International Conference on Electronic Sustainable Communication Systems (ICESC) in 2020.

18. M.R. Nahimana et al. provided a population-based national estimate of the prevalence and factors linked to hypertension in Rwanda, aiming to inform prevention and control efforts. Their research was published in BMC Public Health in 2017.