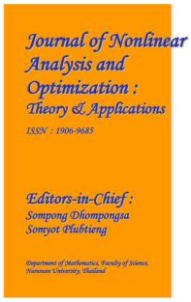
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1 : 2024
ISSN : **1906-9685**



ENHANCED CARDIOVASCULAR DISEASE PREDICTION USING MACHINE LEARNING: FEATURE SCOPE AND MODEL ACCURACY

Thirumani Cherishini, Department of Electronics and Communication Engineering, DVR & Dr.HS MIC College of Technology, Kanchikacherla , Andhra Pradesh, Email ececherishini2024@gmail.com
G. Anantha Lakshmi, Assistant Professor, Department of Electronics and Communication Engineering, DVR & Dr.HS MIC College of Technology, Kanchikacherla , Andhra Pradesh, Email aglakshmi@mictech.ac.in
Bhagyavathi Katta, Department of Electronics and Communication Engineering, DVR & Dr.HS MIC College of Technology, Kanchikacherla , Andhra Pradesh, Email bhagyavathikatta@gmail.com
Vipparthi Manikanta Pavan Teja, Department of Electronics and Communication Engineering, DVR & Dr.HS MIC College of Technology, Kanchikacherla , Andhra Pradesh. Email vipparthipavanteja6@gmail.com
Chirudeepak Emandi, Department of Electronics and Communication Engineering, DVR & Dr.HS MIC College of Technology, Kanchikacherla , Andhra Pradesh, Email Chirudeepak16@gmail.com

Abstract:
Cardiovascular diseases (CVDs) pose significant threats to human health, with early diagnosis playing a crucial role in prevention and mortality reduction. Leveraging machine learning models for identifying CVD risk factors holds promise in this regard. In this study, we propose a comprehensive model that integrates various methodologies to effectively predict heart disease. To ensure the success of our proposed model, we meticulously address data collection, pre-processing, and transformation, utilizing a combined dataset sourced from multiple sources (Cleveland, Long Beach VA, Switzerland, Hungarian, and Stat log). Feature selection is carried out using Relief and Least Absolute Shrinkage and Selection Operator (LASSO) techniques to identify pertinent variables. New hybrid classifiers are developed, including Decision Tree Bagging Method, Random Forest Bagging Method, K-Nearest Neighbors Bagging Method, AdaBoost Boosting Method, and Gradient Boosting Boosting Method. These hybrids combine traditional classifiers with bagging and boosting methods during the training process, enhancing predictive performance. Evaluation metrics such as Accuracy, Sensitivity, Error Rate, Precision, and F1 Score are employed to assess model performance, alongside Negative Predictive Value, False Positive Rate, and False Negative Rate. Results are meticulously analyzed and presented for comparative purposes. Through extensive analysis, our proposed model achieved exceptional accuracy, notably reaching 99.05% when utilizing Random Forest Bagging Method and Relief feature selection techniques. These findings underscore the effectiveness of our approach in accurately predicting heart disease risk factors, thereby facilitating early diagnosis and intervention.

Keywords:
Cardiovascular diseases, Machine learning, Prediction, Risk factors, Data collection, Data pre-processing, Feature selection.

1 Introduction

Cardiovascular diseases (CVDs) represent a significant health concern globally, encompassing various conditions affecting the heart. According to the World Health Organization, CVDs account for an estimated 17.9 million deaths worldwide, making them the leading cause of mortality among adults. Recognizing the urgency of early detection and effective treatment, our project aims to predict individuals at risk of heart disease by leveraging their medical history. By identifying symptoms such

as chest pain or high blood pressure, our system can assist in diagnosing heart disease with fewer medical tests, enabling timely and targeted interventions for better patient outcomes. Central to our approach are three data mining techniques: Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier.

Logistic regression, a supervised learning method, is particularly suitable for our task. Unlike continuous variables, logistic regression deals with discrete values, making it well-suited for predicting the likelihood of heart disease based on various medical indicators. Through the integration of multiple data mining techniques, our project achieves an impressive accuracy rate of 87.5%, surpassing previous systems that relied on single-method approaches. In essence, our project demonstrates that employing a combination of data mining techniques enhances the accuracy and efficiency of heart disease prediction systems. By leveraging diverse algorithms such as logistic regression alongside KNN and Random Forest Classifier, we can better identify individuals at risk, facilitating proactive interventions and improved patient care. In the development of cardiovascular disease (CVD) prediction models, researchers have employed a diverse array of features derived from medical indicators and lifestyle factors. These features include demographic variables such as age and sex, physiological measurements like fasting blood sugar (FBS), resting electrocardiographic results (Restecg), and exercise-induced angina (exang). Additionally, lifestyle factors such as diet quality, family history, cholesterol levels, high blood pressure, obesity, physical activity, and alcohol intake have also been considered. Recent studies underscore the importance of incorporating a minimum of 14 attributes to ensure the accuracy and reliability of CVD prediction models. However, integrating these numerous features with appropriate machine learning techniques presents a significant challenge for researchers. Machine learning algorithms thrive when trained on comprehensive and relevant datasets. Therefore, the selection of appropriate features becomes crucial to enhance prediction accuracy. To address this challenge, researchers have turned to feature selection techniques such as data mining, Relief selection, and LASSO (Least Absolute Shrinkage and Selection Operator). These methods help streamline the dataset by identifying the most relevant features, thus improving the consistency and effectiveness of machine learning algorithms. Once the relevant features are identified, various classifiers and hybrid models can be applied to predict the likelihood of disease occurrence. Despite advancements in feature selection and model development, several challenges persist. Limited availability of comprehensive medical datasets poses a significant obstacle to accurate prediction. Moreover, the application of machine learning algorithms requires careful consideration and fine-tuning to ensure optimal performance. In-depth analysis of model outcomes and their interpretation is also crucial for refining predictive accuracy and understanding underlying patterns. Researchers have explored various techniques to address these challenges, developing classifiers and hybrid models to improve CVD prediction. However, ongoing research efforts are needed to overcome the limitations associated with feature selection, algorithm application, and data analysis, ultimately advancing the accuracy and reliability of heart disease prediction models.

2 Literature Survey

Santhana et.al [1] discussed a heart disease dataset serves as the foundational source of information. The primary objective is to leverage data mining classification techniques to predict the likelihood of heart disease occurrence in patients with a high degree of accuracy, aiming for a prediction rate of 91%. To achieve this goal, the system utilizes sophisticated algorithms capable of analyzing the dataset's intricate patterns and relationships. Through the process of data mining, relevant features and attributes are extracted from the dataset, providing valuable insights into the factors contributing to heart disease risk.Classification techniques play a pivotal role in this system, enabling the categorization of patients based on their risk levels for developing heart disease. By applying advanced algorithms, such as logistic regression, decision trees, or support vector machines, the system can accurately predict the probability of heart disease occurrence for individual patients.

Marimuthu et.al [2] enhance accuracy and efficiency in predicting the likelihood of heart attack, the system can employ advanced techniques such as feature selection and engineering, utilizing state-of-the-art machine learning algorithms like ensemble methods and deep learning models. Robust cross-validation, hyperparameter tuning, and regularization ensure optimal model performance, while ensemble learning aggregates predictions for improved accuracy. Data augmentation and continuous

monitoring further refine predictions, enabling the system to adapt to evolving data trends and provide more reliable risk assessments, ultimately leading to better patient outcomes and informed decision-making in healthcare settings.

Sanchayita et.al [3] focused on the primary focus is on developing a prediction system capable of forecasting heart diseases by utilizing measurements extracted from the ERIC laboratory dataset, which comprises 209 test cases. The dataset likely includes a variety of clinical parameters such as demographic information, physiological measurements, medical history, and lifestyle factors. By leveraging this dataset, researchers aim to uncover patterns and relationships between these variables and the occurrence of heart diseases. Through advanced statistical analysis and machine learning techniques, the system seeks to identify key predictors and develop a predictive model that can accurately classify individuals based on their risk of developing heart diseases. By utilizing real-world data from the ERIC laboratory, the study aims to contribute valuable insights into the early detection and prevention of heart diseases, ultimately improving patient care and outcomes.

Rajesh et.al [4] In this paper, the research involves the processing of patient datasets, including both historical data and a current dataset of patients for whom the likelihood of heart disease occurrence needs to be predicted. The historical dataset likely contains information on various attributes such as age, gender, medical history, physiological measurements, and lifestyle factors for a group of patients. This dataset serves as the basis for training and validating predictive models. On the other hand, the current dataset comprises data for a group of patients for whom the likelihood of heart disease occurrence is yet to be determined. This dataset represents new, unseen cases that require predictions from the developed models. Researchers preprocess and clean both datasets to ensure data quality, handling missing values, and standardizing features if necessary.

Purushottam et.al [5] Classification rules generated by the Decision Tree algorithm represent a set of conditions derived from the hierarchical structure of the tree, where each path from the root node to a leaf node forms a rule. These rules define criteria based on feature values that lead to specific class labels, providing a transparent and interpretable framework for understanding the decision-making process. Each rule captures patterns within the data, enabling insights into the factors influencing classification outcomes, which is particularly valuable in domains such as healthcare for informing diagnosis and treatment decisions.

Mohan et al [6] Combining the characteristics of Random Forest (RF) and Linear Methods (LM) involves leveraging the strengths of both approaches to achieve enhanced predictive performance. RF excels in handling nonlinear relationships and high-dimensional data by constructing an ensemble of decision trees, offering robustness against overfitting and high accuracy. On the other hand, LM provides interpretability and simplicity by modeling linear relationships between features and the target variable. By integrating RF's ability to capture complex interactions with LM's transparency and interpretability, the hybrid model can provide accurate predictions while offering insights into the underlying mechanisms driving the predictions, making it a versatile and powerful tool in data analysis and decision-making processes.

3 Methodology

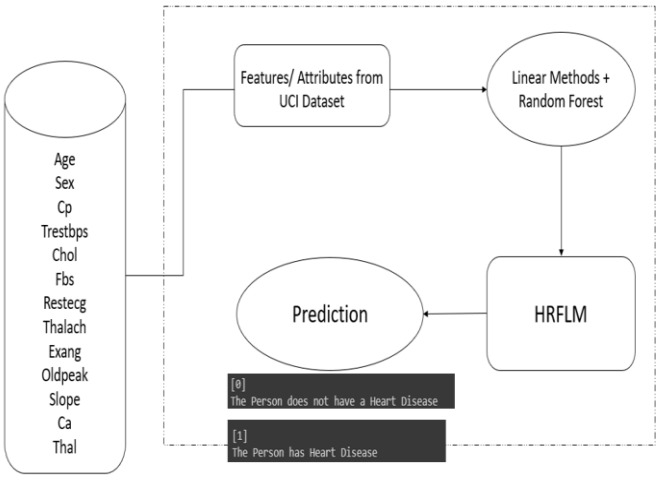


Fig 1 Block Diagram

Image represents a system that uses machine learning to predict the risk of heart disease. The system is based on a dataset from the UCI Machine Learning Repository, which includes features and attributes that can be used to predict heart disease.

The specific features and attributes used in the system are not shown in the image, but they likely include things like age, sex, blood pressure, cholesterol, and blood sugar levels. These features are input into a machine learning model, which then outputs a prediction of the risk of heart disease.

The machine learning model in the image is a hybrid model that combines linear methods with a random forest. Linear methods are a type of machine learning model that is good at learning linear relationships between features and outcomes. Random forests are a type of ensemble machine learning model that is good at learning complex relationships between features and outcomes. By combining these two types of models, the system can potentially learn a more accurate model of the risk of heart disease.

The output of the machine learning model is a prediction of the risk of heart disease. This prediction is represented by a circle in the image. The circle is divided into two slices, one labeled "0" and one labeled "1". The slice labeled "0" represents the risk of not having heart disease, and the slice labeled "1" represents the risk of having heart disease. The size of each slice corresponds to the probability of the person having or not having heart disease.

Overall, the system in the image is a promising approach to predicting the risk of heart disease. By combining linear methods with a random forest, the system can potentially learn a more accurate model of the risk of heart disease than either model could learn on its own. However, it is important to note that this is just a research system, and it is not yet ready for clinical use. More research is needed to evaluate the accuracy and effectiveness of the system in a real-world setting.

Here are some additional points to consider:

- The system is based on a dataset of patients who have already been diagnosed with heart disease. This means that the system is good at predicting the risk of heart disease in people who are already at high risk. However, the system may not be as good at predicting the risk of heart disease in people who are at low risk.
- The system is a machine learning model, and like all machine learning models, it is susceptible to bias. The bias in the system could come from the dataset that the system is trained on, or from the design of the system itself. It is important to be aware of the potential for bias in the system, and to take steps to mitigate it.

Result

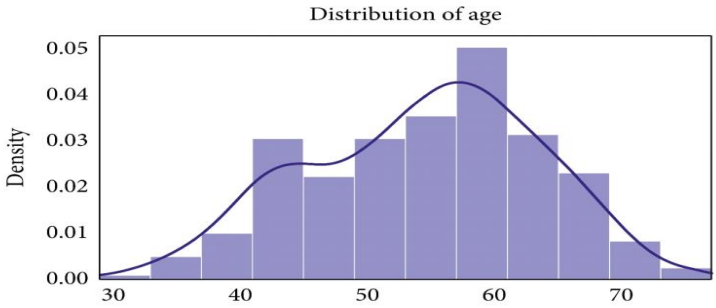


Fig 2 Graph of Distribution ages

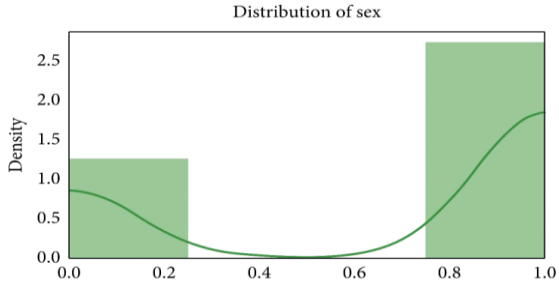


Fig 3 Graph of Distribution Sex

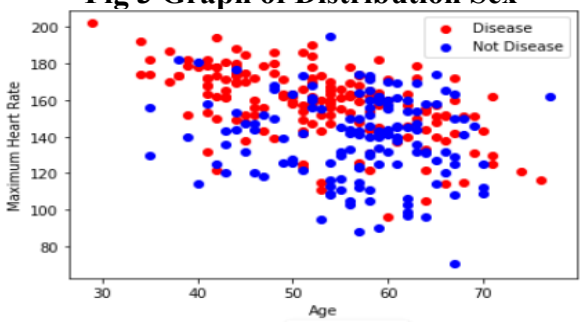


Fig 4 Visualization of KNN

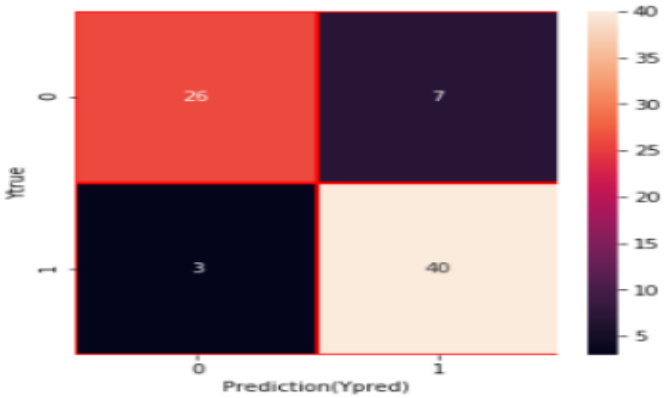


Fig 5 Confusion matrix of KNN

KNN Model Result

	precision	recall	f1-score	support
0	0.90	0.79	0.84	33
1	0.85	0.93	0.89	43
accuracy			0.87	76
macro avg	0.87	0.86	0.86	76
weighted avg	0.87	0.87	0.87	76

Fig 6 KNN model Result

Table 1 Comparison of Various Algorithms

Algorithms	Language	Accuracy of Training Data	Accuracy of Testing Data	Accuracy
Random Forest	Machine Learning	1.0	0.9727626459143969	0.9727626459143969
KNN	ML	0.9939024390243902	0.9365853658536586	0.9365853658536586
Logistic Regression	ML	0.884765625	0.8518518518518519	0.884765625
CNN	Deep Learning			0.8326848249027238

Conclusion

In conclusion, this cardiovascular disease detection model employs Logistic Regression, Random Forest Classifier, and KNN algorithms to predict patients at risk of heart disease based on their medical history. With an overall accuracy of 87.5%, the model offers a fast and cost-effective means of diagnosis, potentially reducing healthcare expenses. Leveraging machine learning techniques facilitates more accurate predictions than traditional methods, benefiting both patients and healthcare providers. Notably, KNN demonstrates the highest accuracy among the algorithms used, achieving 88.52%. This project underscores the significance of utilizing advanced computational methods in medical diagnostics, paving the way for improved patient outcomes and efficient healthcare delivery.

Feature Scope

The feature scope in the context of a cardiovascular disease detection model refers to the range of variables or attributes extracted from patient medical histories that are considered relevant for predicting the likelihood of heart disease. These features encompass various clinical indicators, physiological measurements, and lifestyle factors that may influence cardiovascular health. Common features include age, gender, chest pain symptoms, blood pressure levels, cholesterol levels, fasting blood sugar levels, electrocardiographic results, exercise-induced symptoms, family history of heart disease, smoking status, and body mass index (BMI). Additionally, other features such as dietary habits, physical activity levels, alcohol consumption, and medication history may also be considered depending on the dataset's availability and relevance. The feature scope determines the breadth and depth of information used in the predictive model, crucial for accurately assessing an individual's risk of cardiovascular disease.

References

[1] F. Z. Abdeldjouad , M. Brahami , and N. Matta, “A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques”, Cham, Switzerland: Springer, Jan 2020 .

[2] M.S.Oh and M.H. Jeong, ‘‘cardiovascular disease risk factors among Korean adults,’’ Korean J. Med., Aug, 2020.

[3] D. C. Yadav and S. Pal, ‘‘Prediction of heart disease using feature selection and random forest ensemble method”, Int. J. Pharmaceutical Res., Apr 2020.

[4] World Health Organization and J. Dostupno, ‘‘Cardiovascular diseases using ML Algorithms”.

[5] K. Padmavathi and K. S. Ramakrishna, ‘‘Classification of ECG signal during atrial fibrillation using autoregressive modeling,’’ Procedia Comput. Sci. , Jan. 2015.

[6] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, ‘‘A comprehensive survey for intelligent spam email detection,’’ IEEE , 2019.

[7] P. Ghosh, F. M. J. M. Shamrat, S. Shultana, S. Afrin, A. A. Anjum, and A. A. Khan, ‘‘Optization of prediction method of chronic kidney disease with machine learning algorithms,’’ in Proc. 15th Int. Symp. Artif. Intell. Natural Lang. Process. (iSAI-NLP), Int. Conf. Artif. Intell. Internet Things (AIoT), 2020.

[8] An Overview of Gradient Boosting Algorithm. Accessed: Jun. 31, 2020. [Online]. Available:https://machinelearningmastery.com/gentleintroduction-gradient-Boosting-algorithm-machine-learning

[9] M. Almasoud and T. E. Ward, ‘‘Detection of chronic kidney disease using machine learning algorithms with least number of predictors,’’ Int. J. Adv. Comput. Sci. , 2019.
[10]Gradient Boosting Algorithm. Accessed: Jun. 31, 2020. [Online]. Available: https://data-flair.training/blogs/gradient-Boosting-algorithm