International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Cardio Vascular Disease Prediction Using Machine Learning

D. Nagendra Sai[1], Dr. N. Sri Hari[2], Dr. V. Ramachandran[3]

[1]PG Student, Department of Computer Science and Engineering, VVIT, Guntur, India
[2,3]Professor, Department of Computer Science and Engineering, VVIT, Guntur, India

Abstract:
Cardio vascular disease, commonly referred to as heart disease, is the major cause of death globally, responsible for an estimated 17.9 million deaths each year. This result highlights the critical importance of early detection and diagnosis to reduce mortality rates and improve patient outcomes. Traditional diagnostic approaches rely on clinical tests, physician assessments, and medical imaging, which, while effective, can be time-consuming and subject to human error or bias. In recent years, machine learning acts as a powerful tool to aid in predictive healthcare, enabling the development of models that can identify patterns in patient data and predict disease risk with high accuracy. These datasets are designed to predict the risk of CVD based on a symptoms, lifestyle factors, and medical history. Each row in the dataset represents a patient, with binary indicators for symptoms and risk factors, along with a computed risk label indicating whether the patient is at high or low risk of developing heart disease. The two algorithms' performances were revealed.

1. INTRODUCTION

Cardio vascular diseases continue to be a major cause of death globally, emphasizing the critical need of detection and accurate risk assessment. Traditional diagnostic methods, while valuable, often fall short in predicting the heart conditions, especially in asymptomatic individuals. Machine learning has introduced new methods for enhancing accuracy prediction, enabling healthcare professionals to identify the risk of patients more effectively and implement timely interventions.

Machine learning algorithms, Logistic regression and Extreme Gradient Boosting (XGBoost) have garnered significant attention for their efficacy in predicting heart disease. Logistic regression is renowned for its simplicity and interpretability, making it efficient in medical statistics for classification tasks which are in binary. Conversely, XGBoost, an advanced ensemble learning technique, excels in handling complex datasets and capturing nonlinear relationships, often outperforming traditional models in predictive tasks.

Here datasets used were Zheen hospital in Erbil, Iraq, from January 2019 to May 2019, EarlyMed dataset, developed by students of Vellore Institute of Technology (VIT-AP). and Heart Disease Data Set from UCI data repository.

2. LITERATURE SURVEY

Predicting heart disease is a critical area of research, as it can help healthcare professionals diagnose and treat patients more effectively. Over the years, various ML algorithms have been employed to improve

the efficiency and accuracy of heart disease prediction. In this literature survey, we explore ten key studies and their contributions in predicting heart diseases using ML algorithms.

1. Kaur and Singh (2017) - Logistic Regression for Heart Disease Prediction

In their study, Kaur and Singh (2017) applied **Logistic Regression** to the **Cleveland Heart Disease dataset** to predict the heart disease. They found that Logistic Regression achieved reasonable accuracy but struggled with the dataset's complexity, especially when the relationship between features was non-linear. The study highlighted that while Logistic Regression is interpretable and efficient, it may not always perform well on complex datasets where non-linear relationships between features exist.

2. Chaurasia and Pal (2018) - Decision Tree and Random Forest

Chaurasia and Pal (2018) explored the use of **Decision Tree** and **Random Forest** in prediction of heart diseases. They compared various classifiers, including Decision Tree and Random Forest, on the Cleveland Heart Disease dataset. They found that an ensemble learning method like Random Forest, outperformed Decision Tree due to its ability of overfitting reduction and handle complex interactions between features. The study concluded that Random Forest provided better accuracy and generalization than simpler methods like Decision Tree.

3. Sundararajan et al. (2020) - XGBoost for Heart Disease Prediction

Sundararajan et al. (2020) applied **XGBoost**, a popular gradient boosting technique to predict heart diseases. They evaluated the performance of XGBoost on the Cleveland Heart Disease dataset and achieved an **AUC of 0.87**. The study showed that XGBoost significantly outperformed Logistic Regression and Decision Trees due to its ability to handle imbalanced datasets, missing values, and complex relationships. The study also explored the interpretability of the ML model by analyzing feature importance.

4. Ravi et al. (2016) - Neural Networks for ECG-based Heart Disease Prediction

Ravi et al. (2016) focused on **Artificial Neural Networks (ANNs)** for prediction of heart disease based on **ECG data**. They explored the use of **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**, achieving high accuracy in classifying ECG signals for heart disease detection. The study concluded that deep learning models, such as CNNs and RNNs, could effectively learn complex patterns in time-series ECG data and significantly improve accuracy.

5. Shah et al. (2019) - Comparative Study of ML Algorithms

Shah et al. (2019) conducted a comparative study of several **machine learning algorithms** for heart disease prediction, including **SVM**, **Random Forests**, **Logistic Regression**, and **Neural Networks**. They used the Cleveland Heart Disease dataset and found that **Random Forests** and **SVM** provided the best performance, achieving high accuracy and F1-scores. The study explains the importance of model selection based on the nature of the data and highlighted that ensemble methods like Random Forests tend to outperform individual models.

6. Sultana et al. (2020) - Hybrid Machine Learning Approach

Sultana et al. (2020) proposed a **hybrid machine learning approach** combining **Random Forest** and **Logistic Regression** for heart disease prediction. The study showed that this hybrid model outperformed individual models by combining both the algorithms. The hybrid approach achieved an **accuracy of 90%**, significantly higher than either Random Forest or Logistic Regression alone.

7. Ashraf et al. (2021) - Feature Selection for Heart Disease Prediction Ashraf et al. (2021) focused on **feature selection techniques** to improve the performance and accuracy of heart disease prediction models. They tested various methods, including **filter**, **wrapper**, and **embedded** approaches, to find out

the most relevant features from the Cleveland dataset. The study found that incorporating feature selection techniques improved the accuracy of models like **Random Forests** and **SVM**, reducing overfitting and enhancing model interpretability.

3. METHODOLOGY

This section explains about the input datasets and models for heart disease prediction. Data collection datasets from Kaggle which are

Heart Attack Dataset, Zheen hospital in Erbil, Iraq

The attributes of this dataset are: age, gender, heart rate, systolic blood pressure, diastolic blood pressure, blood sugar, ck-mb and troponin with negative or positive output. According to the information provided, the dataset classifies either heart disease or none. The gender column in the dataset is normalized: the male is having value 1 and the female to 0. The glucose column is should be 1 if it is > 120; otherwise, 0. As for the output, positive is to be 1 and negative to 0.

Heart Disease Risk Prediction Dataset

The dataset contains 70,000 samples, making it suitable for training machine learning models for classification tasks. The agenda is to provide researchers, data scientists, and healthcare professionals with a clean and structured dataset to explore predictive modeling for cardiovascular health.

UCI heart disease data

This is a multivariate dataset which is having a variety of mathematical or statistical variables, multivariate numerical data analysis. It has 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia.

3.1. Models

Logistic Regression

A linear classifier used for classification problems. It calculates the probability of a class using the logistic function:

Logistic regression is a supervised machine learning algorithm used for classification tasks which is used to predict the probability that an instance belongs to a given class or not. Logistic regression is statistical algorithm which can analyze the relationship between two factors of data. The article explores the fundamentals of logistic regression, it's types and implementations.

XGBoost Classifier

XGBoost, a gradient boosting method that builds trees sequentially to correct previous errors. It supports regularization and is optimized for accuracy and speed.

XGBoost is an implementation of optimized Gradient Boosting and is a type of ensemble learning method. Ensemble learning combines multiple weak models to form a stronger model.

- XGBoost uses decision trees as its base learners combining them sequentially to improve the model's performance. Each new tree is trained to rectify errors made by the previous and this process is called boosting.
- It has built-in parallel processing to train models on large and complex datasets quickly. XGBoost also supports customizations allowing users to adjust model parameters to optimize performance based on the specific problem.

3.2.Data Preprocessing:

Scaling (StandardScaler): StandardScaler is used to standardize features by removing the mean and scaling to unit variance. This is important for many machine learning algorithms (like Logistic Regression, which is sensitive to feature scaling) as it ensures that no single feature dominates the learning process due to its scale.

Label Encoding: LabelEncoder transforms categorical labels into numerical values. This is necessary because most machine learning algorithms require numerical input. For example, if a column has categories 'A', 'B', 'C', LabelEncoder might assign them numerical values 0, 1, 2.

Handling Missing Data:

- Dropping Columns: Columns with a high percentage of missing values are dropped because imputing a large amount of missing data might introduce bias or noise into the model. The threshold of 50% is a common heuristic, but the optimal threshold depends on the specific dataset and problem.
- Imputation: Missing values are filled using the median for numerical columns and the mode for categorical columns. The median is less sensitive to outliers than the mean, making it a robust choice for numerical imputation. The mode represents the most frequent category, which is a reasonable imputation strategy for categorical data when no other information is available.

3.3.Data Integration (Merging):

Standardizing Column Names renaming columns to a consistent format (e.g., using lowercase and underscores) is essential for merging datasets that contain similar information but use different naming conventions.

Dropping Unique Features: To combine datasets, it's necessary to have a common set of features. Features present in only one or two of the datasets are dropped to ensure that the merged DataFrame has a consistent structure. This is a simplification for merging; more sophisticated techniques like feature engineering could be used in a real-world scenario to make use of unique features.

3.4.Model Training and Evaluation:

Splitting Data (train_test_split): The dataset is split into training and testing sets. The training set is used to train the machine learning models, and the testing set is used to evaluate their performance on unseen data. This helps to assess how well the models generalize to new examples and prevents overfitting (where the model performs well on the training data but poorly on the test data).

Logistic Regression: Logistic Regression is a linear model used for binary classification (predicting one of two classes). It models the probability of the target variable belonging to a specific class using a sigmoid function.

XGBoost (Extreme Gradient Boosting): XGBoost is a powerful gradient boosting algorithm that builds an ensemble of decision trees. It is known for its high performance and speed. Gradient boosting works by sequentially building trees, with each new tree trying to correct the errors of the previous ones.

Evaluation Metrics (Accuracy Score, Classification Report)

- Accuracy Score: Measures the proportion of correctly predicted instances. It is calculated as (Number of Correct Predictions) / (Total Number of Predictions).
- Classification Report: Provides a more detailed evaluation of the classification model, including precision, recall, F1-score, and support for each class.

- Precision: The proportion of positive predictions that were actually correct.
- Recall (Sensitivity): The proportion of actual positive instances that were correctly identified.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
- Support: The number of actual instances for each class in the test set.

4. RESULTS

The result in the context refer to the evaluation metrics of the trained machine learning models.

4.1.Existing approach

In the early stages of heart disease prediction, several traditional statistical and machine learning methods were used. These methods relied on simple algorithms for binary classification, though they had certain limitations when dealing with complex data. These traditional methods laid the groundwork for heart disease prediction. However, they struggle with non-linear relationships, high-dimensional data, and complex interactions, which is why more techniques and algorithms are now preferred for improved accuracy and efficiency.

4.2.Proposed approach

This proposal outlines to predict the likelihood of heart disease using the Heart Disease datasets through advanced machine learning algorithms, specifically XGBoost and Logistic Regression. By leveraging these models, the goal is to develop a predictive system that can identify at-risk individuals based on medical data, thus aiding healthcare professionals in making more informed decisions.

Tab1: Accuracy Score and the Classification Report for both Logistic Regression and XGBoost.

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0.0	0.93	0.93	0.93	7290
1.0	0.92	0.93	0.93	7102
2.0	0.33	0.10	0.15	21
3.0	0.00	0.00	0.00	27
4.0	0.00	0.00	0.00	8
accuracy			0.93	14448
macro avg	0.44	0.39	0.40	14448
weighted avg	0.92	0.93	0.92	14448

Fig1: Classification report for Logistic Regression

XGBoost Classification Report:

	precision	recall	f1-score	support
0.0	0.95	0.94	0.94	7290
1.0	0.94	0.95	0.94	7102
2.0	0.53	0.43	0.47	21
3.0	0.47	0.30	0.36	27
4.0	0.00	0.00	0.00	8
accuracy			0.94	14448
macro avg	0.58	0.52	0.54	14448
weighted avg	0.94	0.94	0.94	14448

Fig2: Classification report for XGBoost

CONCLUSION:

In conclusion, the proposed models offers more accuracy in predicting heart disease using more complex datasets. Machine learning algorithms, Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost) have garnered significant attention for their efficacy in heart disease prediction. Logistic regression is renowned for its simplicity and interpretability, making it a staple in medical statistics for

binary classification tasks. Conversely, XGBoost, an advanced ensemble learning technique, excels in handling complex datasets and capturing nonlinear relationships, often outperforming traditional models in predictive tasks. In the analysis XGBoost achieved 94% and Logistic regression attained 92% accuracy. These findings underscore the potential of integrating advanced machine learning algorithms into clinical practice, paving the way for more personalized and proactive healthcare solutions.

FUTURE ENHANCEMENTS

The current study demonstrated that both **XGBoost** and **Logistic Regression** have their strengths and weaknesses when applied. While **XGBoost** provides superior accuracy and performance, **Logistic Regression** offers interpretability and simplicity. However, there is significant potential for enhancing the performance, usability, and applicability of heart disease prediction models. There is substantial room for enhancement. Incorporating more data sources, optimizing hyperparameters, and focusing on model interpretability can significantly improve prediction performance and make the models more usable in real-world clinical settings.

Further research should focus on expanding feature sets to include genetic data and real-time patient monitoring, improving model explainability with SHAP and LIME, and developing deep learning models capable of handling larger, more complex datasets. By addressing challenges such as class imbalance and providing real-time predictions, the future of heart disease prediction models will be even more impactful in saving lives and improving healthcare outcomes.

REFERENCES

1. Liu J, Dong X, Zhao H, Tian Y. Predictive classifier for cardiovascular disease based on stacking model fusion. Processes 2022;10(4). https://doi.org/10.3390/ pr10040749
2. Mohan N, Jain V, Agrawal G. Heart disease prediction using supervised machine learning algorithms. 2021 5th Int Conf Inf Syst Computer Networks, ISCON 2021 2021;136(May). https://doi.org/10.1109/ISCON52037.2021.9702314.
3. Gasparrini A. Climate change and cardiovascular disease : implications for global health. June. https://doi.org/10.1038/s41569-022-00720-x; 2022.
4. Shorewala V. Informatics in Medicine Unlocked Early detection of coronary heart disease using ensemble techniques. Inform Med Unlocked 2021;26(June):100655. https://doi.org/10.1016/j.imu.2021.100655.
5. Wang X, Jiang Y, Bai Y, Pan C, Wang R, He M, Zhu J. Association between air temperature and the incidence of acute coronary heart disease in northeast China. Clin Interv Aging 2020;15:47–52. https://doi.org/10.2147/CIA.S235941.
6. Zeng J, Zhang X, Yang J, Bao J, Xiang H, Dear K, Liu Q, Lin S, Lawrence WR, Lin A, Huang C. Humidity may modify the relationship between temperature and cardiovascular mortality in Zhejiang province, China. Int J Environ Res Publ Health 2017;14(11). https://doi.org/10.3390/ijerph14111383.
7. Mnyawami YN, Maziku HH, Mushi JC. Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: a case of Tanzanian's secondary schools. Appl ArtifIntell2022;36(1). https://doi.org/10.1080/08839514.2022.2071406.
8. Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. Liashchynskyi, P. arXiv preprint arXiv:1912.06059 2019:1–11. http://arxiv.org/abs/1912.06059.

9. V.S.K. Reddy, P. Meghana, N.S. Reddy, B.A. Rao, Prediction on cardiovascular disease using decision tree and naïve Bayes classifiers, J Phys Conf Ser 2161 (1) (2022) 012015, https://doi.org/10.1088/1742-6596/2161/1/012015.
10. P. Gupta, D.D. Seth, Early detection of heart disease using multilayer perceptron, in: Micro-Electronics and Telecommunication Engineering Proceedings of 6th ICMETE 2022, 2023, pp. 309–315, https://doi.org/10.1007/978-981-19-3615-1_36.