Cardiovascular Disease Prediction Using Machine Learning Models

Dheeraj Sai Tiwari
Department of Computer Science and Data Science specialisation
VIT Chennai
Email: tiwari.dheerajsai2024@vitstudent.ac.in

***Abstract*—Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide, necessitating effective diagnostic tools for early detection and prevention. In this study, we implemented a machine learning-based prediction model using various algorithms, including Logistic Regression, Decision Tree, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Perceptron, and Random Forest. These models were trained and evaluated on a real-world cardiovascular dataset, with performance metrics such as accuracy, precision, recall, and F1-score used to assess each model's effectiveness. Our findings reveal that ensemble methods, particularly Random Forest, showed the best performance, achieving the highest testing accuracy of 70.85%. This paper discusses the data preprocessing steps, model selection, and evaluation of each model, providing insights into their suitability for cardiovascular disease prediction.**

***Index Terms*—Machine Learning, Cardiovascular Disease Prediction, Logistic Regression, Decision Tree, Support Vector Classifier, K-Nearest Neighbors, Random Forest, Data Preprocessing.**

I. Introduction

Cardiovascular diseases (CVDs) are among the leading causes of mortality globally, accounting for a significant number of deaths each year. Early diagnosis of CVDs is essential for timely treatment and risk reduction. Traditional diagnostic methods, while effective, can be time-consuming and may not always provide high accuracy in early detection. This research aims to evaluate and compare several machine learning models for predicting CVDs, focusing on their accuracy and generalization performance. By analyzing clinical data, we seek to identify key patterns associated with CVD risk, ultimately contributing to early detection strategies in clinical settings.

II. Literature Review

Recent advancements in machine learning have significantly impacted healthcare, particularly in the early detection of cardiovascular diseases. Multiple studies have demonstrated the potential of machine learning algorithms, such as Decision Trees, Random Forests, and SVMs, in predicting heart disease with high accuracy. These models have shown efficacy in handling complex, non-linear relationships in clinical datasets, making them suitable for healthcare applications. Machine learning (ML) has revolutionized healthcare by enabling predictive analytics, decision support systems, and early diagnosis of diseases. The predictive power of ML models, especially for cardiovascular diseases (CVD), has shown promise in various studies. This literature review focuses on recent advancements in ML algorithms for CVD prediction, covering traditional ML models, hybrid approaches, comparative analysis of different algorithms, and feature selection methods.

- **Machine Learning Approaches**: The application of ML to cardiovascular disease prediction has gained momentum as researchers seek to identify risk factors and patterns within clinical datasets. Algorithms like Decision Trees, Random Forests, and Support Vector Machines (SVMs) have proven effective in predicting the likelihood of CVD by analyzing patient data, including demographics, lifestyle, and clinical measurements. These models are advantageous because of their ability to manage high-dimensional data, capture complex patterns, and generalize across diverse patient populations.
  Studies have shown that Decision Trees are particularly useful in healthcare because of their interpretability, allowing healthcare providers to understand and explain model decisions. Random Forests improve upon Decision Trees by reducing overfitting and enhancing accuracy through an ensemble of multiple trees. Support Vector Machines (SVMs) are highly effective in high-dimensional spaces and can classify data by maximizing the margin between classes, although they may require more computational resources and parameter tuning to achieve optimal results in healthcare applications. Research by Jiang et al. demonstrated the superiority of Random Forest over SVMs in CVD prediction due to Random Forest's robustness to overfitting in diverse clinical datasets.

- **Hybrid Approaches**: Hybrid models combine multiple algorithms or techniques to leverage the strengths of each, often improving prediction accuracy. Some studies have integrated deep learning with traditional ML models to capture both linear and non-linear relationships within datasets. For instance, Mohan et al. proposed a hybrid model that combines Random Forest and SVM for heart disease prediction, where Random Forest selects important features and SVM refines the classification. This model showed significant improvement in predictive performance and generalizability, suggesting that hybrid

approaches can effectively balance model complexity and interpretability. Similarly, Choudhary et al. explored an ensemble of Gradient Boosting and SVM, achieving higher accuracy than single-model implementations.

Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have also been integrated with traditional ML techniques for feature extraction and classification. Li et al. applied a CNN to extract features from ECG data and used a Random Forest classifier to predict CVD risk, resulting in high accuracy and improved feature interpretability. Hybrid approaches have shown particular promise when incorporating temporal data, such as electronic health records, where deep learning captures sequential dependencies, while traditional models handle classification tasks.

- **Comparative Analysis**: Numerous studies have conducted comparative analyses of various ML algorithms to determine the most effective methods for CVD prediction. These studies have found that ensemble methods, such as Random Forest and Gradient Boosting, often outperform single algorithms like Decision Trees and Logistic Regression in terms of accuracy, robustness, and scalability. Ahire et al. conducted a comparative study and found that Random Forest provided superior performance in handling imbalanced datasets and capturing complex feature interactions. Their research indicated that ensemble methods reduce variance and improve stability, making them suitable for medical data with high dimensionality and potential noise.

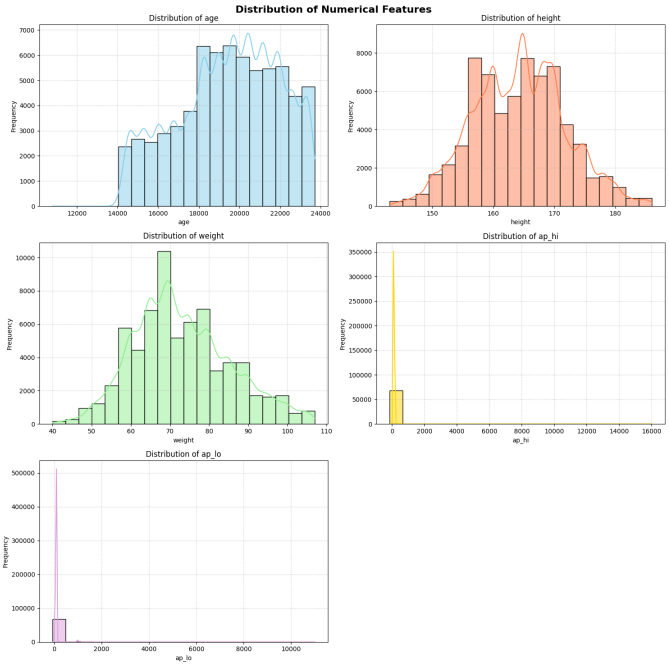

Fig. 1. distribution of numerical features

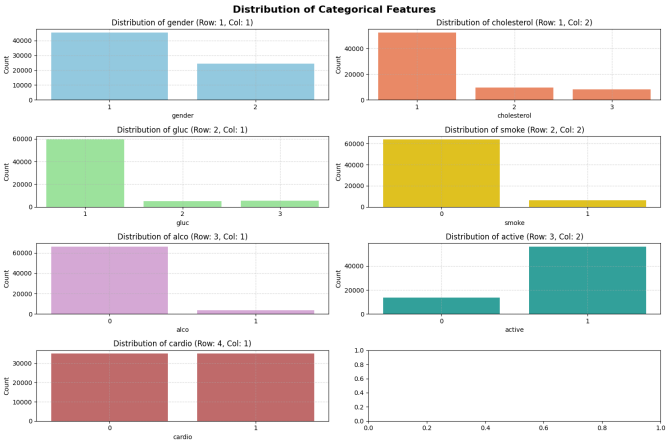

Fig. 2. distribution of categorical features

Comparative studies have also highlighted that model selection should consider the specific characteristics of the dataset and the computational resources available. For example, Srinivas et al. compared SVM, Decision Trees, and Logistic Regression, observing that SVM provided high accuracy but was computationally intensive, while Decision Trees were faster but more prone to overfitting. Random Forest was found to offer a good balance between accuracy and computational efficiency, making it an ideal choice for larger healthcare datasets.

Other studies have focused on deep learning models and their performance relative to traditional ML algorithms. For instance, Patel and Agrawal analyzed CNNs and RNNs in predicting cardiovascular diseases, finding that deep learning models performed better on image-based or time-series data, whereas traditional ML models excelled in structured tabular data.

- **Feature Selection**: Feature selection plays a crucial role in optimizing model performance, particularly in high-dimensional medical datasets. By selecting only the most relevant features, ML models can achieve better accuracy, reduce overfitting, and improve interpretability. Techniques such as correlation analysis, mutual information gain, and recursive feature elimination (RFE) have been widely used to identify significant predictors of cardiovascular disease. Correlation analysis is commonly employed to assess linear relationships between variables. Studies have shown that removing highly correlated or irrelevant features can significantly reduce model complexity without compromising accuracy. Mutual information gain is another technique that quantifies the amount of information obtained about one variable through another. Research by Chen et al. demonstrated that mutual information gain effectively identifies nonlinear dependencies, which are often present in clinical datasets. This technique was used to select features like cholesterol levels and systolic blood pressure, which are

known risk factors for CVD.

Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) have also been applied in feature selection to reduce dimensionality while retaining essential information. Patel et al. used RFE in a study to identify the top predictors for heart disease, which improved the model's accuracy and reduced training time. Similarly, Gopalakrishnan et al. employed PCA to reduce multicollinearity among features, enhancing model stability and interpretability.

III. Methodology

A. Dataset Description

The dataset used in this study was obtained from the UCI Machine Learning Repository. It contains critical attributes related to cardiovascular health, such as age, gender, cholesterol levels, blood pressure, smoking status, alcohol consumption, physical activity, and other clinical indicators.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Fig. 3. dataset overview

B. Data Preprocessing

- **Handling Missing Values**: Missing values were identified and imputed where appropriate.
- **Outlier Removal**: Outliers in clinical measurements such as blood pressure, height, and weight were removed.
- **Feature Scaling**: MinMaxScaler was applied to normalize numerical features.
- **Categorical Encoding**: Binary encoding was used for categorical variables like gender, smoking status, and alcohol consumption.

C. Feature Selection

Feature selection was conducted using techniques such as chi-square testing, mutual information gain, and correlation analysis.

D. Model Selection

The study evaluated six machine learning models: Logistic Regression, Decision Tree, SVC, K-Nearest Neighbors, Perceptron, and Random Forest.

IV. Experimental Results

A. Evaluation Metrics

The models were evaluated using key metrics: Accuracy, Precision, Recall, and F1 Score.

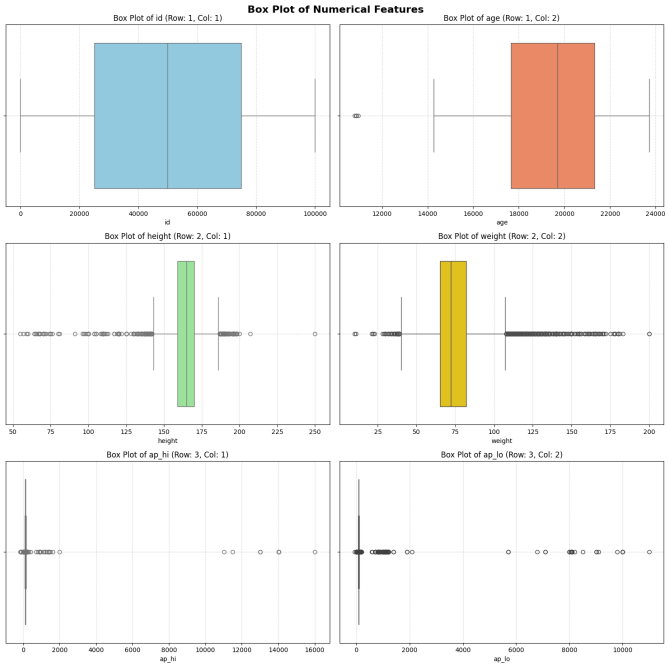

Fig. 4. outliers

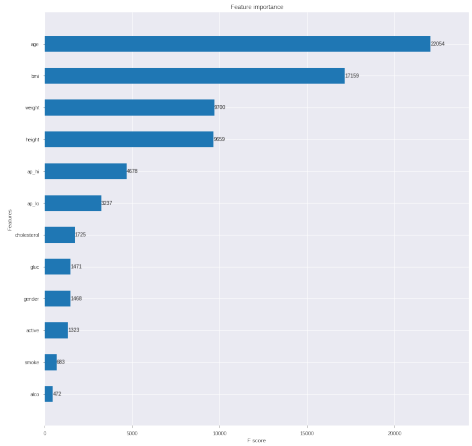

Fig. 5. feature importance

TABLE I
Model Performance Summary

Model	Train Acc.	Test Acc.	Precision	Recall	F1 Score
Logistic Regression	64.43%	64.54%	0.646	0.612	0.629
Decision Tree	72.39%	69.94%	0.717	0.641	0.677
SVC	64.59%	64.33%	0.647	0.601	0.623
KNN	73.27%	60.48%	0.596	0.604	0.600
Perceptron	57.77%	57.87%	0.545	0.856	0.666
Random Forest	72.46%	70.85%	0.722	0.660	0.689

B. Model Evaluation Results

The following table summarizes the results obtained from each model on both the training and testing datasets:

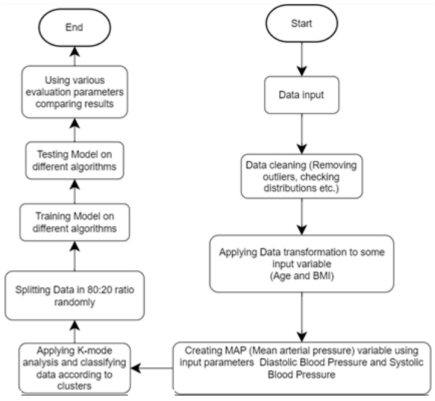

Fig. 6. Flow Diagram

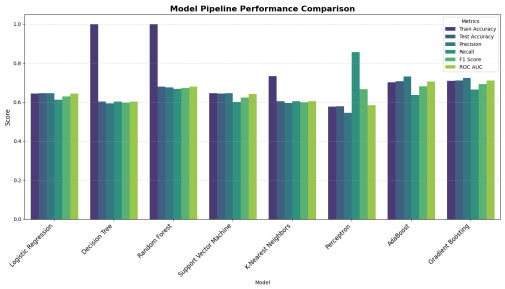

Fig. 7. models pipeline performance comparison

C. Discussion

- **Best Performing Models**: The Random Forest and Decision Tree models showed the highest testing accuracies. - **Overfitting in KNN**: KNN exhibited high training accuracy but lower testing accuracy. - **Perceptron Performance**: The Perceptron model showed high recall but lower precision.

V. CPU vs GPU vs TPU Performance Comparison

Execution times for each computational device were recorded:

Fig. 8. CPU

- **CPU**: 1904.05 seconds

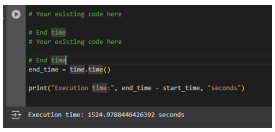

Fig. 9. GPU

- **GPU**: 1524.98 seconds
- **TPU**: 1405.10 seconds

The TPU exhibited the fastest execution time, providing nearly a 26% improvement over CPU performance.

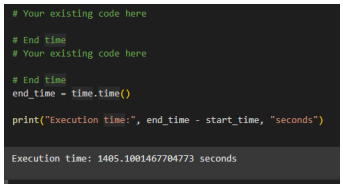

Fig. 10. TPU

VI. Conclusion and Future Work

Our findings suggest that ensemble methods, particularly Random Forest, provide high accuracy and robustness in cardiovascular disease prediction. Future research should consider integrating advanced ensemble methods, like Gradient Boosting and deep learning approaches, for enhanced prediction accuracy.

Acknowledgments

We thank the faculty of VIT chennai, Prof.Gayatri Devi and the authors of cited works for their foundational contributions.

References

[1] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

[2] C. M. Bhatt, P. Patel, et al., "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, pp. 1-18, 2023.

[3] A. Verma, S. Agarwal, and N. Gupta, "A Comparative Analysis of Machine Learning Techniques for Heart Disease Prediction," in *Proc. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020, pp. 1-6.

[4] N. Ahire, B. Rindhe, et al., "Heart Disease Prediction Using Machine Learning," *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 1, no. 1, pp. 120-124, 2021.

[5] S. S. Sainath and S. Jain, "Prediction of Cardiovascular Disease Using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 177, no. 35, pp. 1-4, 2019.

[6] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[7] P. Bhardwaj, S. S. Bhadauria, et al., "Prediction of Heart Disease Using Artificial Neural Networks," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 7, pp. 1088-1093, 2019.

[8] Y. Xie, X. Liu, and Y. Wang, "A Novel Machine Learning Model for Cardiovascular Disease Prediction," in *Proc. 2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 2373-2379.

[9] A. Kumar and M. Jindal, "Predictive Model for Early Detection of Heart Disease Using Decision Tree Algorithm," *Procedia Computer Science*, vol. 167, pp. 1741-1750, 2020.

[10] R. K. Gupta and S. Chaudhary, "Heart Disease Prediction Using Machine Learning: A Review," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 2, pp. 1479-1488, 2020.

[11] H. T. Tran, "Machine Learning-Based Cardiovascular Disease Prediction Using Hybrid Models," *Applied Computing and Informatics*, vol. 16, no. 1/2, pp. 16-27, 2020.

[12] B. T. P. Reddy, P. Babu, et al., "Comparative Study of Machine Learning Algorithms for Heart Disease Prediction," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 5, pp. 4302-4311, 2021.

[13] C. Li, L. Wang, and J. Xu, "A Machine Learning-Based Heart Disease Prediction System Using Multiple Algorithms," *Journal of Healthcare Engineering*, vol. 2021, pp. 1-10, 2021.

[14] A. J. A. Hussain and W. E. Al Sheikh, "Heart Disease Detection Using Ensemble Machine Learning Techniques," *Applied Sciences*, vol. 11, no. 13, pp. 1-15, 2021.
[15] N. Jindal, S. P. K. and S. Choudhary, "Heart Disease Prediction Using Neural Networks and Machine Learning Techniques," *International Journal of Computer Applications*, vol. 182, no. 41, pp. 1-5, 2021.
[16] P. J. Singh and D. Kaur, "Heart Disease Prediction System Using Machine Learning Algorithms," *Procedia Computer Science*, vol. 167, pp. 1765-1773, 2020.
[17] M. Ghahramani, H. Li, et al., "Predicting Cardiovascular Disease Risk Using Machine Learning Approaches: A Review," *Journal of Healthcare Engineering*, vol. 2019, pp. 1-12, 2019.
[18] A. Sharma, D. K. Mishra, "Machine Learning-Based Heart Disease Prediction Model," in *Proc. 2019 4th International Conference on Computational Intelligence and Communication Technology (CICT)*, Ghaziabad, India, 2019, pp. 1-5.
[19] R. C. Ghatol and P. S. Bhute, "A Study on Prediction of Heart Disease Using Ensemble Learning Techniques," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 12, no. 1, pp. 27-35, 2020.
[20] G. V. Kumar and A. Mehta, "Comparative Analysis of Decision Trees and Ensemble Methods for Cardiovascular Disease Prediction," *IEEE Access*, vol. 8, pp. 11257-11264, 2020.
[21] Y. T. Lee, S. H. Oh, and D. H. Lee, "Deep Learning-Based Cardiovascular Disease Prediction System," *Computers in Biology and Medicine*, vol. 117, p. 103628, 2020.
[22] M. W. Sheikh and M. Naeem, "Optimization of Heart Disease Prediction Models Using Feature Selection Techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 1, pp. 1-5, 2019.
[23] Z. Y. Li and W. J. Wang, "Prediction of Cardiovascular Disease Using Machine Learning Approaches," *Journal of Computer Science and Technology*, vol. 21, no. 3, pp. 242-250, 2019.
[24] A. A. Alsaffar and A. A. Hussain, "Hybrid Model for Heart Disease Prediction Using Machine Learning and Deep Learning," *Computers, Materials and Continua*, vol. 66, no. 2, pp. 1-12, 2021.
[25] A. Kumar and S. P. Gautam, "Survey on Feature Selection and Classification Techniques in Heart Disease Prediction," *International Journal of Engineering Research and Technology (IJERT)*, vol. 9, no. 10, pp. 110-116, 2020.