International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 03 | Mar 2025 www.irjet.net p-ISSN: 2395-0072

Cardiovascular Disease Prediction and Risk Assessment using Machine Learning Approaches

Manpreet Hire[1], Yasir Khan[2], Zuber Shaikh[3], Priyanshu Prajapati[4]

[1]*Asst. Professor, Department of Data Science, Thakur College of Science and Commerce, Mumbai, India*
[2]*Student, Department of Data Science, Thakur College of Science and Commerce, Mumbai, India*
[3]*Student, Department of Data Science, Thakur College of Science and Commerce, Mumbai, India*
[4]*Machine Learning Engineer Intern, BNK Infotech Pvt. Ltd., Delhi, India*

Abstract - *In this research, we develop a machine learning-based framework to predict patient-specific disease outcomes using a multifactorial healthcare dataset. The dataset encompasses a wide array of variables, including demographic details (such as age and sex), clinical indicators (general health condition, medical checkup frequency, physical activity), behavioral aspects (smoking history), and records of chronic illnesses like cardiovascular disease, skin cancer, diabetes, arthritis, and other cancers. Our exploratory data analysis revealed notable patterns in disease occurrence, highlighting how various health and lifestyle factors influence disease susceptibility. Furthermore, the dataset presented challenges in the form of incomplete data, underscoring the necessity for effective handling of missing values to ensure model robustness. The predictive models developed in this study offer valuable insights for identifying high-risk individuals and facilitating early interventions. The findings underline the potential of machine learning methodologies in enhancing clinical decision-making and advancing personalized healthcare.*

Key Words: Cardiovascular Disease Prediction, Machine Learning, Risk Assessment, Healthcare Data, Chronic Disease Forecasting, Clinical Analytics, Lifestyle Risk Factors, Predictive Modeling, Patient-Level Prognosis, Exploratory Data Analysis.

1. INTRODUCTION

The growing burden of cardiovascular and other chronic diseases represents a significant challenge for modern healthcare systems. These conditions are influenced by a complex combination of demographic, behavioral, and clinical factors, necessitating accurate and early risk prediction to improve patient outcomes and reduce healthcare costs. Conventional risk scoring systems often fail to capture the intricate relationships embedded within large-scale patient data, creating a demand for more advanced and flexible predictive tools. In this context, machine learning (ML) has emerged as a transformative approach in healthcare analytics, enabling the discovery of hidden patterns that traditional statistical models may overlook. This study focuses on leveraging ML, particularly gradient boosting algorithms such as XGBoost, to forecast disease prognosis using a multifaceted healthcare dataset that includes patient demographics, health behaviors, and clinical history.

The proposed model is developed and validated using a train-test split strategy to ensure robust and generalizable performance. A variety of evaluation metrics—including the confusion matrix, ROC curve, and Precision-Recall curve—are employed to provide a holistic assessment of the model's effectiveness, particularly in managing class imbalance common in healthcare data. By integrating these methods, the study demonstrates the capacity of ML to significantly enhance disease risk assessment and inform clinical decision-making processes. The findings underscore the potential of predictive modeling to support healthcare professionals in identifying high-risk patients and tailoring early intervention strategies. Additionally, this research paves the way for future improvements through advanced data balancing techniques and the inclusion of supplementary clinical features, ultimately contributing to the broader adoption of data-driven, personalized healthcare solutions.

1.1 BACKGROUND

Machine learning, particularly ensemble methods like gradient boosting, has proven to be highly effective in tackling these challenges. Algorithms such as XGBoost offer enhanced performance by iteratively reducing errors and handling imbalanced datasets commonly found in medical records. By integrating diverse patient-level data—ranging from demographic information to lifestyle factors—ML models can provide more precise and individualized risk assessments. This paradigm shift toward data-driven healthcare underscores the growing relevance of predictive analytics in clinical practice.

1.2 TOOLS & TECHNOLOGIES

This study primarily leverages Python, a versatile and widely adopted programming language in the field of data science and machine learning, due to its extensive libraries and ease of integration in healthcare analytics workflows. The core machine learning framework utilized in this

project is Scikit-learn (Sklearn), which offers a comprehensive suite of tools for data preprocessing, model training, hyperparameter tuning, and evaluation. Sklearn's modular and efficient implementation of algorithms such as gradient boosting, specifically XGBoost, allows for scalable model development and seamless handling of healthcare datasets characterized by class imbalance and missing values. Additionally, key Python libraries, including Pandas and NumPy, are employed for data manipulation, exploratory data analysis, and feature engineering, facilitating in-depth insights into patient-level records. Matplotlib and Seaborn are used to visualize patterns and relationships within the data, supporting both the analytical and presentation phases of the research. The combination of these tools enables the development of a robust and interpretable machine learning pipeline, tailored to address the challenges of disease prognosis prediction. This technology stack ensures reproducibility, scalability, and efficiency in building healthcare models capable of assisting clinical decision-making and enhancing patient risk assessment.

2. LITERATURE REVIEW

Sr. No.	Title	Objectives	Findings
1.	**A Machine Learning-Based Approach for the Prediction of Cardiovascular Diseases**	To develop a machine learning framework for early detection of cardiovascular diseases using supervised learning algorithms such as SVM, KNN, XGBoost, Random Forest, LightBoost, and SGD.	The study achieved a 92.76% detection rate using the SGD classifier on an 80:20 training/testing split, indicating the effectiveness of machine learning models in predicting cardiovascular diseases.
2.	**A Novel Approach for Cardiovascular Disease Prediction Using Machine Learning Algorithms**	To enhance the prediction accuracy of heart diseases by employing ensemble machine learning models, including k-Nearest Neighbor, XGBoost, AdaBoost, and Random Subspace classifiers, and to identify influential features using Linear Support Vector Feature Measure.	The proposed ensemble model achieved a prediction accuracy of 96%, outperforming existing approaches. Key performance metrics included 97% precision, 95% sensitivity, and 95% F-measure, demonstrating the model's robustness in heart disease prediction.
3.	**Cardiovascular Disease Prediction Using Various Machine Learning Algorithms**	To develop an efficient model for predicting the probability of cardiovascular diseases by analyzing critical factors and employing algorithms such as Logistic Regression, SVM, MLP with PCA, Deep Neural Network, and Bootstrap Aggregation using Random Forests.	The study utilized the UCI repository dataset and found that the proposed model could effectively predict heart diseases. The integration of various machine learning algorithms provided a robust framework for cardiovascular disease prediction.
4.	**NHS in England to Trial AI Tool to Predict Risk of Fatal Heart Disease**	To evaluate the effectiveness of an AI tool, AI-ECG risk estimation (Aire), in predicting fatal heart disease and early death by analyzing ECG test results.	The AI tool, Aire, demonstrated the ability to predict 10-year mortality, future heart failure, serious heart rhythm issues, and atherosclerotic cardiovascular disease with considerable accuracy, highlighting its potential to enhance preventive treatments and patient management.
5.	**Algorithm Could Help Prevent Thousands of Strokes in UK Each Year**	To develop a machine learning algorithm capable of identifying undiagnosed atrial fibrillation (AF) to prevent strokes by analyzing factors such as age, sex, ethnicity, and existing medical conditions.	The algorithm, developed using anonymized data from millions of GP records, successfully identified high-risk patients for AF. Early detection through this tool could lead to timely treatment, significantly reducing stroke incidence and associated healthcare costs.
6.	**Predicting Cardiovascu**	To compare various machine	The study concluded that

	lar Disease Using Machine Learning Algorithms	learning algorithms, including Decision Trees, Random Forest, and Gradient Boosting, in predicting cardiovascular diseases using patient data.	ensemble methods like Gradient Boosting provided higher accuracy in predicting cardiovascular diseases compared to individual classifiers, emphasizing the importance of using advanced machine learning techniques for effective prediction.

3. ALGORITHMS/TECHNIQUES

3.1. LOGISTIC REGRESSION

A logistic model (or logit model) is a statistical model that models the log-odds of an event as a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) estimates the parameters of a logistic model (the coefficients in the linear or non linear combinations). The parameters of a logistic regression are most commonly estimated by maximum-likelihood estimation (MLE).[1]

Owing to its simplicity and interpretability, Logistic Regression is used to establish a linear decision boundary between healthy and at-risk patients. Despite its linear nature, it is highly effective when applied to healthcare datasets with well-separated classes and assists in providing an initial risk stratification of patients.

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

Figure 3.1: Logistic Regression
Where:

- P(Y=1|X) = Probability of positive class given feature vector X
- β0 = Intercept term (bias)
- β1,β2,...,βn = Coefficients for each feature
- x1,x2,...,xn = Input features (e.g., age, sex, smoking history, etc.)
- e = Euler's number (approx. 2.718)

3.2. RANDOM FOREST CLASSIFIER

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that works by creating a multitude of decision trees during training. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the output is the average of the predictions of the trees.[2]

In the context of cardiovascular disease prediction, Random Forest is particularly useful for handling high-dimensional feature spaces and extracting critical health and lifestyle factors that contribute significantly to disease risk

$$\hat{Y} = majority\_vote(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M)$$

Figure 3.2: Random Forest Classifier

Where:

- $\hat{Y}$ = Final prediction of the Random Forest
- $\hat{Y}_i$ = Prediction of the ith decision tree
- M = Total number of decision trees in the forest

3.3. ADABOOST CLASSIFIER

AdaBoost (short for Adaptive Boosting) is a statistical classification meta-algorithm. It can be used in conjunction with many types of learning algorithms to improve performance. The output of multiple weak learners is combined into a weighted sum that represents the final output of the boosted classifier. Usually, AdaBoost is presented for binary classification, although it can be generalized to multiple classes or bounded intervals of real values.[3]

Within our framework, AdaBoost complements other models by increasing the sensitivity to minority or borderline disease cases, making it highly valuable in medical risk assessment tasks where false negatives could have serious clinical consequences.

$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Figure 3.3(A): AdaBoost Classifier

Where:

- H(x) = Final strong classifier
- ht(x) = Weak classifier (e.g., a decision stump) at iteration t
- α_t = Weight assigned to ht(x) based on its accuracy
- T = Total number of weak classifiers

The weights (αt) are calculated as:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$$

Figure 3.3(B): AdaBoost Classifier

Where:

- ϵ_t = Error rate of ht(x)

3.4. HARD VOTING CLASSIFIER

A Voting Classifier is an ensemble learning technique that aggregates the predictions of multiple base models (e.g., Logistic Regression, Random Forest, Support Vector Machines) to make a final prediction. The core idea is that a group of diverse models will collectively perform better than any single model alone. The Voting Classifier is especially useful when the base models capture different perspectives or patterns in the data, making the final ensemble more robust.[4]

The final predictive model is an ensemble Voting Classifier employing a hard voting mechanism, which synthesizes the outputs of Logistic Regression, Random Forest, and AdaBoost classifiers. This ensemble model adopts a majority-vote strategy, where each individual model contributes a single vote towards the final class prediction for each patient.

$$\hat{Y} = \arg\max_{c \in \{0,1\}} \sum_{m=1}^{M} I(\hat{y}_m = c)$$

Figure 3.4: Hard Voting Classifier

Where:

- $\hat{Y}$ = Final ensemble prediction (class label)
- c = Class label (0 for no disease, 1 for disease)
- M = Number of base models (in this case, 3)
- I(·) = Indicator function returns 1 if yhat_m = c else 0
- $\hat{y}_m$ = Prediction from the mth base classifier

4. DESIGN AND METHODOLOGY

The architecture of this project is built upon a straightforward yet effective machine learning pipeline, aimed at predicting the presence or absence of heart disease based on patient-level health data.

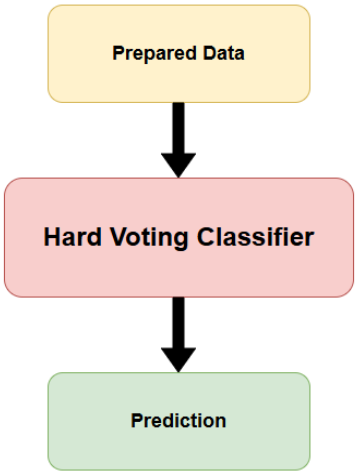

Fig -4.1: High Level Flow of the Overall System

The process begins with data preparation, where the healthcare dataset, comprising various health and lifestyle indicators such as age, gender, general health status, smoking habits, exercise frequency, and other disease histories, is cleaned and preprocessed. Missing values are handled using imputation techniques to ensure data completeness. Additionally, categorical features are encoded, and numerical features are normalized where necessary to enhance model interpretability and performance.

Following data preparation, the refined dataset is fed into a Hard Voting Classifier. This ensemble model integrates the predictive capabilities of three distinct machine learning algorithms: Logistic Regression, Random Forest, and AdaBoost Classifier. Each of these base learners captures different patterns and relationships within the data, contributing to a more robust decision-making process. Logistic Regression models the linear relationship between input features and disease occurrence, while Random Forest captures non-linear patterns through multiple decision trees. AdaBoost focuses on reducing bias and variance by iteratively improving weak learners.

Once trained, each base model independently makes a prediction on whether a patient has heart disease. The Hard Voting mechanism then aggregates these predictions through a majority voting rule, where the class (presence or absence of heart disease) with the highest number of votes from the base models is selected as the final output.

5. RESULTS

The performance evaluation on the heart disease dataset revealed that the Hard Voting Classifier achieved the best generalization performance, securing the highest test accuracy of 73.16%. This demonstrates the effectiveness of aggregating Logistic Regression, Random Forest, and

AdaBoost models to deliver stable and reliable predictions.

Interestingly, while individual classifiers like Decision Tree Classifier and Random Forest reported extremely high training accuracies of 99.98%, their test accuracies dropped to 64.00% and 71.56%, respectively, indicating overfitting. On the other hand, the Soft Voting Classifier recorded a training accuracy of 83.19%, but it still failed to outperform the Hard Voting model on unseen data.

Additionally, ensemble models such as XGBoost and LGBMClassifier performed competitively with test accuracies of 73.09% and 73.00%, yet the Hard Voting Classifier retained the edge by providing a better balance between training and testing performance.

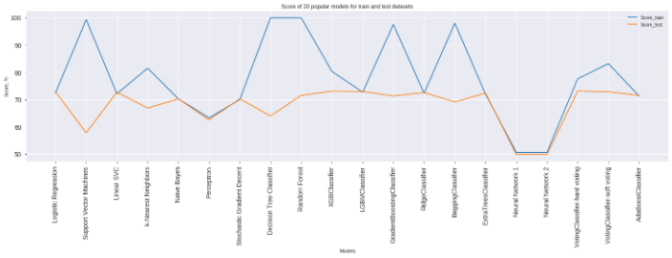

Fig -5.1: Results

Sr. No	Algorithm	Training Accuracy	Test Accuracy
1.	Voting Classifier - Hard	77.62%	73.16%
2.	XGBClassifier	80.32%	73.09%
3.	LGBMClassifier	72.71%	73.00%
4.	VotingClassifier - Soft	83.19%	72.87%
5.	Random Forest	99.98%	71.56%
6.	AdaBoostClassifier	71.28%	71.53%
7.	GradientBoostingClassifier	97.52%	71.34%
8.	BaggingClassifier	98.02%	69.11%
9.	Decision Tree Classifier	99.98%	64.00%

6. CONCLUSIONS

In this study, the Hard Voting Classifier emerged as the most effective approach for predicting the presence of heart disease. This result underscores the power of model diversity in ensemble learning, where the collective decision-making process helps reduce bias and variance, leading to better generalization on unseen data. Unlike models prone to overfitting, such as the Decision Tree and Random Forest, the Hard Voting approach maintained a healthy balance between training and test accuracies, making it a reliable solution for healthcare prediction tasks. The success of this ensemble strategy highlights its potential for practical implementation in clinical decision support systems, where robust and consistent performance is crucial for patient risk assessment and early diagnosis.

REFERENCES

[1] Logistic Regression on Wikipedia. https://en.wikipedia.org/wiki/Logistic_regression

[2] Random Forest on Wikipedia. https://en.wikipedia.org/wiki/Random_forest

[3] AdaBoost on Wikipedia. https://en.wikipedia.org/wiki/AdaBoost

[4] Sharma, G. (n.d.). Voting Classifier for Binary Classification Problems on Medium. https://medium.com/@iGirijesh/voting-classifier-for-binary-classification-problems-e5fef5e6c54d

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[6] Chollet, F. (2017). Deep learning with python. Manning Publications.

[7] Ahmad AA, Polat H. Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm. Diagnostics (Basel). 2023

[8] Damen, Johanna AAG, et al. "Prediction models for cardiovascular disease risk in the general population: systematic review." bmj 353 (2016).

[9] Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. International Journal of Computer Applications, 181(18), 20-25.

[10] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[11] Murray, Thomas H. "Ethics, genetic prediction, and heart disease." The American journal of cardiology 72.10 (1993): D80-D84.

[12] Alwakid, Ghadah, et al. "Optimized machine learning framework for cardiovascular disease diagnosis: a

novel ethical perspective." BMC Cardiovascular Disorders 25.1 (2025): 123.

[13] Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media, Inc.

[14] Miao, Kathleen H., Julia H. Miao, and George J. Miao. "Diagnosing coronary heart disease using ensemble machine learning." International Journal of Advanced Computer Science and Applications 7.10 (2016).

[15] Bajaj, Madhvan, et al. "Heart Disease Prediction using Ensemble ML." 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2023.