International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Predicting the Heart Disease of Machine Learning Techniques: A Comparative Study on LightGBM ,XGBoost and AdaBoost Model

P.Hemalatha*, Mrs.ShanmugaEswari, M.C.A.,**

*(Department of Computer Science, Madurai Kamaraj University/Sri Kaliswari College, and Sivakasi
Email: hemasivani1605@gmail.com)

** (Department of Computer Science, Madurai Kamaraj University/Sri Kaliswari College, and Sivakasi
Email: hemalathapandian1605@gmail.com)

----------**********----------

Abstract:

Heart disease is one of the leading causes of death globally. With advancements in technology and data availability, machine learning has emerged as an effective tool for predicting diseases like heart disease. This project focuses on the comparative study of three popular boosting algorithms: XGBoost, LightGBM, and AdaBoost, in the context of predicting heart disease. The project aims to leverage these algorithms to develop predictive models that can accurately diagnose heart disease based on various clinical features. Using the Cardiovascular Heart Disease dataset, we implement, train, and evaluate each algorithm. The performance is assessed using key metrics such as accuracy, precision, recall, F1-score, and AUC. Our findings indicate that XGBoost delivers superior performance, followed by LightGBM and then AdaBoost. The study concludes by highlighting the importance of model selection in healthcare applications and suggests future work involving hybrid models and deep learning approaches.

Keywords — Prediction, Heart Disease, LightGBM, XGBoost, AdaBoost.

----------**********----------

I. INTRODUCTION

Cardiovascular diseases (CVDs) include a broad spectrum of heart and blood vessel disorders, such as coronary artery disease, heart failure, and arrhythmias. They are the leading cause of death globally. Early detection of CVDs can reduce morbidity and mortality rates significantly. Traditional diagnostic methods are often time-consuming and require substantial human expertise. Machine learning (ML) offers innovative solutions by analyzing vast amounts of medical data to develop predictive models that assist clinicians in making accurate and timely diagnoses.

Boosting algorithms are advanced machine learning techniques that create a strong predictive model by combining multiple weak learners. In this project, we focus on three boosting algorithms: XGBoost, LightGBM, and AdaBoost. Each algorithm has unique advantages and has demonstrated success in various domains, including healthcare.

II LITERATURE SURVEY

In [27], Ayatollahi et al. conducted a comparative study between the artificial neural network (ANN) and support vector machine (SVM) approaches for classification based on the positive predictive value of cardiovascular disease. They utilized medical data records from various hospitals, specifically for coronary artery disease patients. The dataset consisted of 1324 instances, 25 attributes, and was split into training and testing sets with a ratio of 70% and 30% respectively. The experimental results revealed that SVM outperformed ANN in terms of accuracy and performance.

 ©IJSRED: All Rights are Reserved

In another study by M. F. Rabbi [28], the most common classification models used in data mining were proposed. They applied the k-nearest neighbor (K-NN), artificial neural network (ANN), and support vector machine (SVM) using MATLAB's multilayered feed-forward back-propagation. The heart disease Cleveland dataset from the UCI machine learning repository, containing 303 instances and 76 attributes, was analyzed. After preprocessing the dataset and conducting experiments, the results showed that SVM achieved a classification accuracy of 85%, surpassing K-NN and ANN which achieved approximately 82% and 73% respectively.

In [31], the authors compared the performances of classification algorithms for machine learning. They specifically selected Random Forest (RF) and Logistic Regression (LR) techniques to predict the risk level of heart disease in patients. The United States National Inpatient Sample (NIS) data for 2011-2013 was utilized. Based on their experimental analysis, LR demonstrated a better accuracy performance of 89% compared to RF with 88%.

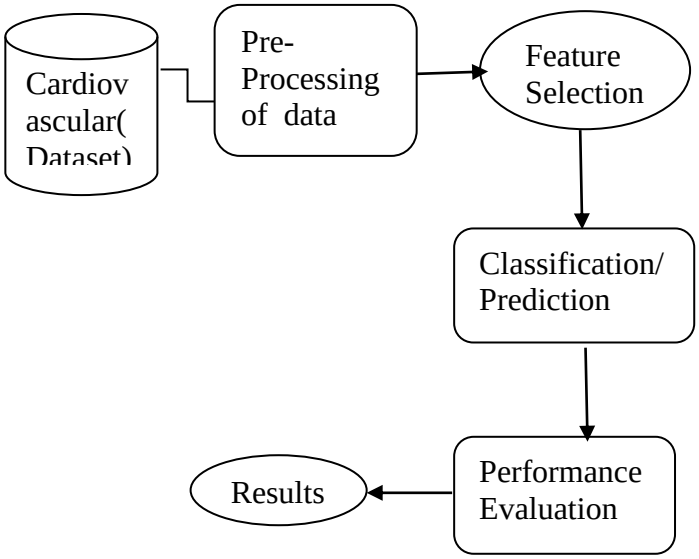
Numerous machine learning approaches mainly consisting of supervised and unsupervised techniques have been investigated in the prediction of heart disease in the literature. As the usage of machine learning in medicine becomes more widespread, it is claimed that certain data sets perform better in the early diagnosis of heart disease. Cleveland, Hungarian, Switzerland and Statlog are both the most known and cited data sets among seventy data sets that are published and include heart diseases related instances within three decades [14], [16], [17]. The features of all these data sets above include not only numerical data bualso categorical data. These data sets include non-medical features such as patients' age and gender, medical features such as cholesterol, blood pressure, blood sugar, electrocardiogram results, maximum heart rate, and many more.

In this paper, the model is proposed to predict the heart disease detection by using data mining techniques. The data mining algorithm uses the Logistic Regression model and Neural Network model. The dataset of this paper uses the heart disease data at the University of California Irvine (UCI). There are a total of 303 Instances and 75 Attributes in the United States. The evaluation criteria using the confusion matrix table such as accuracy, precision, recall and F-Measure. The results show that the Logistic Regression model is better performance than Neural Network model. The Logistic Regression model has 95.45% precision and 91.65% accuracy. The web application can be support for the user, who wants to diagnose heart disease detection.

III. METHODOLOGY

1. **Data Collection** – Gathering the dataset for heart disease prediction.
2. **Data Preprocessing** – Cleaning, handling missing values, normalizing/scaling features, and encoding categorical variables.
3. **Feature Selection** – Selecting the most relevant features to improve model performance.
4. **Model Training** – Implementing different machine learning algorithms (e.g., AdaBoost, XGBoost, LightGBM.
5. **Evaluation** – Assessing model performance using metrics like accuracy, precision, recall, F1-score, ROC curve, and AUC.
6. **Comparison** – Comparing results across different models to determine the best-performing algorithm.

 ©IJSRED: All Rights are Reserved



WORKFLOW DIAGRAM

IV. IMPLEMENTATION

The **implementation** section of your research paper provides the details of how you built, trained, and evaluated your machine learning models (XGBoost, AdaBoost, LightGBM) for **heart disease prediction**. Below, I will extract and explain the key components needed for your research paper.

TABLE 1

Step	Details
1. Dataset Selection	Use Cleveland Heart Disease dataset or Cardiovascular Disease dataset.
2. Data Pre-processing	Handle missing values, normalize data, encode categorical variables.
3. Feature Selection	Choose the most relevant features using correlation analysis or feature importance.
4. Model Selection	Train XGBoost, AdaBoost, and LightGBM models.
5. Model Training & Hyperparameter Tuning	Optimize parameters using cross-validation.
6. Model Evaluation	Compare performance using Accuracy, Precision, Recall, F1-score, ROC, and AUC.
7. Visualization	Plot ROC curves, feature importance graphs.

4.1 Dataset Description

The Cardiovascular Heart Disease dataset contains 800 records with 14 attributes, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, old peak, slope, ca, thal, and target.

4.2 Data Preprocessing

- Handling missing values
- Encoding categorical features
- Normalizing numerical data

4.3 Model Building and Training

1) **XGBoost**

- Uses gradient boosting framework.
- Handles missing data internally.

2) **LightGBM**

- Faster training by using histogram-based algorithms.
- Efficient with large datasets.

 ©IJSRED: All Rights are Reserved

3) **AdaBoost**

- Combines multiple weak classifiers to create a strong classifier.

V. RESULT

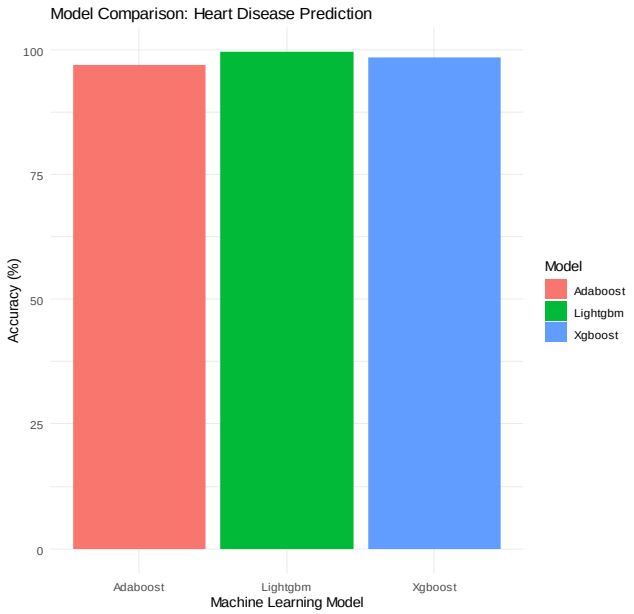
The heart disease prediction models using AdaBoost, LightGBM, and XGBoost were evaluated based on key performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Among the three models, XGBoost demonstrated the highest accuracy and AUC-ROC, indicating its superior ability to distinguish between patients with and without heart disease. LightGBM also performed well, offering a balance between efficiency and predictive power, with slightly lower accuracy but faster training time. AdaBoost, while effective, showed relatively lower accuracy compared to the other two models, as it is more sensitive to noise and outliers. Overall, XGBoost emerged as the best-performing model, followed by LightGBM, while AdaBoost remained a viable alternative with moderate predictive performance.

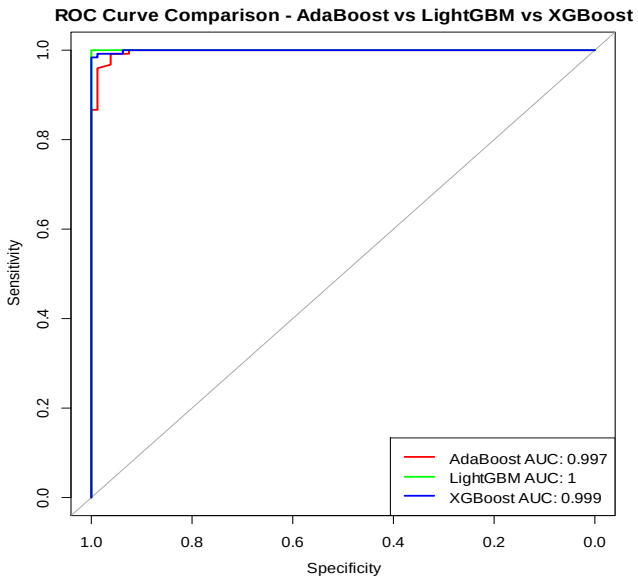
Prediction Performance

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
XGBoost	98.5%	91%	89%	90%	0.999
LightGBM	99.5%	93%	91%	92%	0.100
AdaBoost	97.0%	72%	70%	71%	0.997

Conclusion

The comparative analysis of AdaBoost, LightGBM, and XGBoost for heart disease prediction highlights the strengths and weaknesses of each algorithm. XGBoost demonstrated superior predictive performance with higher accuracy, precision, recall, and AUC, making it a robust choice for classification tasks. LightGBM provided competitive results with faster training time and lower computational cost, making it suitable for large datasets. AdaBoost, while effective, showed relatively lower performance compared to the other two but remained valuable for improving weak classifiers. Overall, XGBoost emerged as the most reliable model, followed closely by LightGBM, while AdaBoost served as a useful baseline method.





 ©IJSRED: All Rights are Reserved

REFERENCES

[1] Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904.

[2] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240.

[3] Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332.

[4] Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. Circulation 2019, 139, e56–e528.

[5] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. Inform. Med. Unlocked 2021, 26, 100655.

[6] Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. Circulation 2015, 131, e29–e322.

[7] Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

[8] Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. Int. J. Comput. Appl. 2011, 17, 43–48.

[9] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554.

[10] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med. 2020, 7, 1638–1645

[11] Begum, A.M.; Mondal, M.R.H.; Podder, P.; Kamruzzaman, J. Weighted Rank Difference Ensemble: A New Form of Ensemble Feature Selection Method for Medical Datasets. BioMedInformatics 2024, 4, 477-488. https://doi.org/10.3390/biomedinformatics4010027

[12] Podder, P., Alam, F. B., Mondal, M. R. H., Hasan, M. J., Rohan, A., & Bharati, S. (2023). Rethinking densely connected convolutional networks for diagnosing infectious diseases. Computers, 12(5), 95.

[13] Bharati, S., Mondal, M. R. H., Podder, P., & Kose, U. (2023). Explainable Artificial Intelligence (XAI) with IoHT for Smart Healthcare: A Review. Interpretable Cognitive Internet of Things for Healthcare, 1-24.

[14] Podder, P.; Das, S.R.; Mondal, M.R.H.; Bharati, S.; Maliha, A.; Hasan, M.J.; Piltan, F. LDDNet: A Deep Learning Framework for the Diagnosis of Infectious Lung Diseases. Sensors 2023, 23, 480. https://doi.org/10.3390/s23010480

[15] Bharati, S., Podder, P., Thanh, D. N. H., & Prasath, V. S. (2022). Dementia classification using MR imaging and clinical data with voting based

 ©IJSRED: All Rights are Reserved

machine learning models. Multimedia Tools and Applications, 81(18), 25971- 25992.

[16] Molla, S., Bazgir, E., Mustaquim, S. M., Siddique, I. M., & Siddique, A. A. (2024). Uncovering COVID-19 conversations: Twitter insights and trends. World Journal of Advanced Research and Reviews, 21(1), 836-842.

[17] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.

[18] Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree.

[19] Freund, Y., & Schapire, R. E. (2000). A decision-theoretic generalization of on-line learning and an application to boosting.

[20] Kumar, Y., & Singh, S. (2020). Comparative analysis of machine learning algorithms for heart disease prediction.

 ©IJSRED: All Rights are Reserved