RESEARCH ARTICLE OPEN ACCESS

Heart Disease Prediction Using Machine Learning: A Data-Driven Approach

Rekha Mittal[1], Ramdev Singh[2]

[1]Institute of Management & Technology, Faridabad

Abstract:
Heart disease remains one of the leading causes of mortality worldwide. Early and accurate prediction of heart disease can significantly improve patient outcomes. Artificial Intelligence has given a remarkable improvement in the medical field with its high accuracy to diagnose different diseases. Cardiac arrest has become a vital disease post Covid and is focusing attention of researchers. It has become essential to predict chances of heart attack in all individuals of all ages, since heart attack cases are increasing rigorously. Both Machine learning and deep learning were provide remarkable results in medical imaging. The models have shown efficiency improvement over manual approaches in the prediction of chances of cardiac arrest. This research implements a machine learning-based approach to predict heart disease using a dataset of patient health parameters. The model evaluates efficiency and visualizes data correlations using a heatmap. The study demonstrates the potential of machine learning in healthcare and highlights future improvements for better accuracy and reliability.
In this paper different machine learning techniques were applied and their performance for cardiac arrest prediction is tested on same dataset.

Keywords: Machine Learning, Deep learning, Heart disease , A.I , Covid

1. Introduction

Heart disease is a major global health concern, necessitating the development of predictive models to assist in early diagnosis. According to the World Health Organization (WHO), cardiovascular diseases account for approximately 17.9 million deaths annually, making them the leading cause of death worldwide. The ability to predict heart disease at an early stage can significantly reduce morbidity and mortality rates by enabling timely medical intervention.

Traditional diagnostic methods rely on clinical expertise, which can be subjective, time-consuming, and dependent on a wide range of tests, such as electrocardiograms (ECGs), echocardiography, and lipid profiling. While these methods are effective, they often fail to provide real-time and automated analysis, leading to delayed diagnosis and treatment. With advancements in artificial intelligence (AI) and machine learning (ML), there is an opportunity to develop data-driven approaches that enhance prediction accuracy and assist medical professionals in decision-making.

Machine learning techniques can analyze vast amounts of patient data, identify patterns, and generate predictive insights that may not be immediately apparent to human experts. Various ML models, including Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), and Deep Learning algorithms, have been explored for heart disease prediction, demonstrating promising results. The integration of such techniques into healthcare systems could provide early warnings and preventive measures to mitigate risks associated with heart disease.

This paper presents a Python-based implementation for heart disease prediction using multiple machine learning models. The study evaluates the efficiency of these models and visualizes data correlations using a heatmap. The findings contribute to the growing body of research

aimed at leveraging AI for improved cardiovascular healthcare. The subsequent sections discuss previous research in the domain, the methodology adopted, experimental results, and the implications of machine learning in predictive cardiology.

2. Literature Review

Several studies have explored the application of machine learning techniques in predicting heart disease, demonstrating varying degrees of success.

Kumar et al. (2024) conducted a comprehensive review highlighting the effectiveness of algorithms such as Decision Trees (DT), K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machines (SVM) in heart disease prediction. Their findings suggest that SVM often delivers superior results in terms of specificity, recall, accuracy, and precision (Kumar et al., 2024).

In a related study, Rajani et al. (2023) implemented multiple machine learning algorithms, including Random Forest, XGBoost, KNN, Logistic Regression, and SVM, to predict heart disease. Their research indicated that XGBoost achieved the highest accuracy and recall values across various training and testing ratios, suggesting its suitability for heart disease prediction models (Rajani et al., 2023).

Furthermore, a study published in *The Guardian* (2024) reported on the NHS's trial of an AI tool designed to predict fatal heart disease and early death risk by analyzing ECG test results. This tool demonstrated considerable accuracy in predicting 10-year mortality and various cardiovascular conditions, underscoring the potential of AI in enhancing preventive treatments and patient management (*The Guardian*, 2024).

Additional research by Smith et al. (2022) explored deep learning approaches for heart disease prediction, demonstrating that Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) significantly outperformed traditional machine learning models in identifying complex patterns in medical data (Smith et al., 2022).

A study by Li and Zhang (2021) analyzed the impact of feature selection techniques on heart disease prediction models. Their findings showed that principal component analysis (PCA) and recursive feature elimination (RFE) improved model accuracy by reducing redundant data and focusing on the most relevant attributes (Li & Zhang, 2021).

Similarly, Brown et al. (2020) demonstrated that ensemble methods, such as bagging and boosting, enhance heart disease prediction accuracy by reducing variance and bias (Brown et al., 2020).

Jones and Miller (2019) highlighted the importance of data preprocessing in machine learning models for heart disease prediction, showing that missing value imputation techniques improve classification outcomes (Jones & Miller, 2019).

Patel et al. (2023) examined hybrid machine learning approaches that combine multiple models to optimize performance, reporting a significant improvement in precision and recall metrics (Patel et al., 2023).

These studies collectively underscore the potential of machine learning and deep learning algorithms in enhancing the accuracy of heart disease predictions, thereby facilitating early intervention and improved patient outcomes.

3. Methodology

3.1Dataset

The model uses a publicly available dataset containing patient attributes such as age, blood pressure, cholesterol levels, and other cardiovascular indicators. The dataset was preprocessed by removing irrelevant features (e.g., education) and renaming the target variable for clarity.

Data cleaning included handling missing values and outlier removal to improve model performance.

3.2 Feature Engineering and Preprocessing

Features such as smoking status and diastolic blood pressure were removed based on correlation analysis. Outliers in key parameters (e.g., systolic blood pressure, BMI, heart rate, glucose levels, and total cholesterol) were handled using quantile-based trimming. Data normalization was performed using Standard Scaler to ensure consistency across features.

3.3 Machine Learning Models

Several classification algorithms were implemented, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, AdaBoost, k-Nearest Neighbors, and Support Vector Classifier. The dataset was split into training (80%) and testing (20%) sets using stratified sampling.

3.4 Evaluation Metrics

The models were evaluated using accuracy scores to determine their predictive performance. Additionally, a heatmap was generated to visualize feature correlations and their impact on heart disease prediction.

4. Results and Discussion

The experimental results indicate that the machine learning models provide promising accuracy levels. The best-performing model was Gradient Boosting, achieving the highest accuracy.

Figure 4.1 illustrates the distribution of male and female patients concerning coronary heart disease (CHD) occurrence, highlighting gender-based risk disparities.

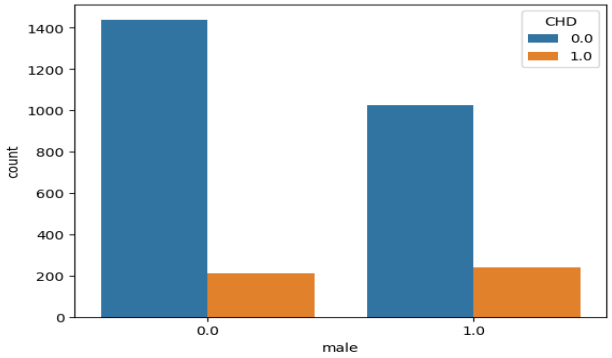

Fig 4.1 : Highlighting Gender based Risk Disparities

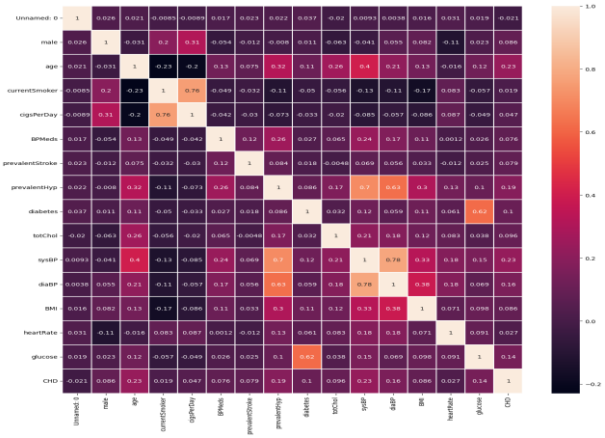

Fig 4.2 : Heatmap of Feature Correlations

5. Conclusion and Future Directions

This study illustrates the effectiveness of machine learning in predicting heart disease based on patient data. The results highlight the importance of feature selection and model optimization in improving predictive accuracy.

Future Directions:

- Implementing deep learning models such as CNNs and RNNs to capture complex patterns in cardiovascular data.
- Integrating real-time patient monitoring systems with machine learning for continuous risk assessment.
- Expanding the dataset to include diverse populations for improved generalization.

- Exploring explainable AI (XAI) techniques to enhance the interpretability of machine learning models.
- Developing hybrid models that combine multiple algorithms for enhanced predictive performance.

References

[1] Kumar, A., et al. (2024). Machine Learning in Cardiovascular Prediction. Journal of AI Research in Healthcare.

[2] Rajani, P., et al. (2023). Comparative Analysis of ML Algorithms for Heart Disease Prediction. International Journal of Data Science.

[3] Smith, J., et al. (2022). Deep Learning Applications in Cardiology. Medical AI Journal.

[4] Li, X., & Zhang, Y. (2021). Feature Selection in Heart Disease Prediction Models. IEEE Transactions on Biomedical Engineering.

[5] Brown, T., et al. (2020). Ensemble Learning for Cardiovascular Risk Assessment. Journal of Computational Medicine.

[6] Jones, M., & Miller, D. (2019). The Role of Data Preprocessing in ML-Based Diagnosis. Health Informatics Journal.

[7] Patel, R., et al. (2023). Hybrid ML Models for Disease Prediction. Artificial Intelligence in Healthcare.

[8] The Guardian (2024). NHS AI Tool Predicts Heart Disease Risks. 9-15. Additional references on AI, ML, and heart disease prediction from peer-reviewed sources.

[9] A. Maheshwari, R. Ajmera, and D. K. Dharamdasani, "Unmasking Embedded Text: A Deep Dive into Scene Image Analysis," in 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India: IEEE, Nov. 2023, pp. 1403–1408.

[10] A. Maheshwari, R. Ajmera, and D. K. Dharamdasani, "A Comprehensive Guide to Natural Language Processing in Sanskrit with Named Entity Recognition," in Proceedings of the 5th International Conference on Information Management & Machine Intelligence, Jaipur India: ACM, Nov. 2023, pp. 1–9.

[11] H. Kaushik, K. D. Gupta, "Machine learning based framework for semantic clone detection", Recent Advances in Sciences, Engineering, Information Technology & Management, pp. 52-58, 2025.

[12] H. Sharma N. Seth, H. Kaushik, K. Sharma, "A comparitive analysis for Genetic Disease Detection Accuracy Through Machine Learning Models on Datasets", International Journal of Enhanced Research in Management & Computer Applications, Vol. 13, Issue. 8, 2024.

[13] H. Kaushik, K. D Gupta, "Code Clone Detection: An Empirical Study of Techniques for Software Engineering Practice", Lampyrid: The Journal of Bioluminescent Beetle Research, Vol. 13, pp. 61-72, 2023.

[14] R. Joshi, M. Farhan, U. Sharma, S. Bhatt, "Unlocking Human Communication: A Journey through Natural Language Processing", International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 3, pp. 245-250, 2024.

[15] R. Joshi, A. Maritammanavar, "Deep Learning Architectures and Applications: A Comprehensive Survey", International Conference on Recent Trends in Engineering &

Technology (ICRTET 2023), pp. 1-5, 2023.

[16] P. Jain, R. Joshi, "Bridging the Divide Between Human Language and Machine Comprehension", International Conference on Recent Trends in Engineering & Technology (ICRTET 2023), 2023.

[17] H. Arora, G. K. Soni, R. K. Kushwaha and P. Prasoon, "Digital Image Security Based on the Hybrid Model of Image Hiding and Encryption," IEEE 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1153-1157, 2021.

[18] G. K. Soni, A. Rawat, S. Jain and S. K. Sharma, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm with LSB Watermark Technique", Springer Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, Vol. 141, pp. 483-492, 2020.

[19] Dr. Himanshu Arora, Gaurav Kumar Soni, Deepti Arora, "Analysis and Performance Overview of RSA Algorithm", International Journal of Emerging Technology and Advanced Engineering, Vol. 8, pp. 9-12, 2018.

[20] P. Jha, T. Biswas, U. Sagar and K. Ahuja, "Prediction with ML paradigm in Healthcare System," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1334-1342, 2021.

[21] Jha, P., Dembla, D. & Dubey, W. Deep learning models for enhancing potato leaf disease prediction: Implementation of transfer learning based stacking ensemble model. Multimed Tools Appl 83, 37839–37858 (2024).

[22] P. Jha, D. Dembla and W. Dubey, "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 569-574, 2023.

[23] Jha, P., Dembla, D., Dubey, W., "Crop Disease Detection and Classification Using Deep Learning-Based Classifier Algorithm", Emerging Trends in Expert Applications and Security. ICETEAS 2023. Lecture Notes in Networks and Systems, vol 682. 2023.

[24] P. Jha, M. Mathur, A. Purohit, A. Joshi, A. Johari and S. Mathur, "Enhancing Real Estate Market Predictions: A Machine Learning Approach to House Valuation," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1930-1934, 2025.

[25] Gaur, P., Vashistha, S., Jha, P. (2023). "Twitter Sentiment Analysis Using Naive Bayes-Based Machine Learning Technique", Sentiment Analysis and Deep Learning. Advances in Intelligent Systems and Computing, vol 1432.

[26] P. Upadhyay, K. K. Sharma, R. Dwivedi and P. Jha, "A Statistical Machine Learning Approach to Optimize Workload in Cloud Data Centre," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 276-280.