Predictive Modeling for Early Detection of Heart Disorders

[1]E. Hemanth, [2]B. N Sashank, [3]K. Anil Kumar, [4] Dr. K. Vijay Kumar
[123]B.Tech Student, [4] Associate Professor
Dept. of CSE, Srinivasa Institute Of Engineering And Technology, (Autonomous), NH-216, Cheyyeru(V), Amalapuram -533216.

ABSTRACT:

Heart disease remains a leading cause of mortality worldwide, emphasizing the critical need for early detection tools. This project applies machine learning techniques to predict heart disease using clinical features from the UCI dataset. After preprocessing and feature engineering, several algorithms were evaluated including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and XGBoost. A hybrid model was also developed using ensemble voting to improve accuracy. Evaluation metrics such as accuracy, precision, recall, and F1-score were employed for model comparison. Results revealed that the ensemble model outperformed individual algorithms, demonstrating high reliability and diagnostic potential. This research showcases how data-driven systems can assist medical professionals in early cardiac risk identification and improve patient outcomes.

Keyword: Clinical Diagnosis, Machine Learning, Heart Disease Prediction

INTRODUCTION

Cardiovascular diseases (CVDs) are among the most fatal health threats globally, responsible for approximately 17.9 million deaths annually. The rising prevalence, especially in low-income regions, highlights the urgent need for early diagnosis and intervention. Traditional methods, while effective, are often time-consuming and resource-intensive. In contrast, machine learning offers a transformative solution by analyzing complex datasets to uncover patterns that might escape human detection. This project investigates the application of ML algorithms to predict heart disease based on patient data such as age, cholesterol, blood pressure, and chest pain type. By training and validating multiple models, this work aims to identify the most accurate predictors and construct an efficient decision-support system. Such tools can empower healthcare providers with rapid, cost-effective, and scalable diagnostic solutions, ultimately reducing preventable deaths and improving public health strategies.

RELATED WORK

In recent years, numerous studies have explored machine learning applications in predicting cardiovascular diseases, highlighting its growing role in preventive healthcare.

Beunza et al. (2019) conducted a comparative study using the Framingham Heart Study dataset to assess the performance of ML algorithms such as logistic regression, decision trees, and neural networks. Their research revealed that random forest models achieved superior accuracy and interpretability for coronary heart disease prediction.

Chen et al. (2020) introduced a Naive Bayes classifier trained on data from patients with Dilated Cardiomyopathy. By leveraging clinical variables and applying information gain for feature selection, their model demonstrated high predictive accuracy and reinforced ML's utility in cardiovascular event forecasting.

Farooq and Hussain (2021) developed a hybrid diagnostic framework combining ontology-driven reasoning and machine learning. Their Ontology-Driven Clinical Risk Assessment System (ODCRARS) integrated expert knowledge with predictive models, improving diagnostic accuracy in Rapid Access Chest Pain Clinics.

Tang (2022) proposed a real-time arrhythmia classification model using wearable ECG sensors. Utilizing Delta Modulations and a custom Support Vector Machine (SVM), the study achieved impressive F1-scores (0.83 for SVEB and 0.92 for VEB), showcasing ML's applicability in mobile health environments.

Kaur and Sharma (2018) used a public heart disease dataset and tested various classifiers including SVM, KNN, and Decision Trees. They concluded that ensemble methods, particularly Random Forest, outperformed standalone models in terms of precision and recall, emphasizing the benefits of model aggregation.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author	Contribution	Impact on Current Research
Beunza et al. (2019)	Compared ML models using Framingham dataset for coronary event prediction.	Validated the effectiveness of ensemble classifiers like Random Forest in medical datasets.
Chen et al. (2020)	Applied Naive Bayes with feature selection on heart failure patient data.	Inspired the use of feature ranking and Naive Bayes in model comparison for accuracy analysis.
Farooq & Hussain (2021)	Developed ODCRARS hybrid system combining clinical ontology and ML models.	Influenced the hybrid ensemble approach using structured and unstructured data.
Tang (2022)	Built real-time SVM-based arrhythmia classifier using wearable ECG data.	Highlighted the potential of SVMs for real-time cardiovascular condition prediction.
Kaur & Sharma (2018)	Analyzed performance of SVM, DT, and KNN; recommended ensemble approaches.	Supported the ensemble voting strategy used in your LSD hybrid model implementation.

PROPOSED APPROACH

The proposed system is designed to predict heart disease using multiple machine learning algorithms and a hybrid ensemble technique. The core objective is to provide a reliable, data-driven decision support system that can assist medical professionals in identifying patients at risk of cardiovascular conditions.

The project begins by utilizing the UCI Heart Disease dataset, which contains 304 patient records with clinical attributes such as age, chest pain type, cholesterol level, resting blood pressure, and maximum heart rate. Data preprocessing steps, including handling missing values, normalization, and encoding categorical variables, are applied to ensure high-quality inputs.

Following data preparation, exploratory data analysis is performed using tools like Seaborn and Matplotlib to identify key feature correlations. Multiple supervised ML classifiers are trained, including Logistic Regression, Decision Tree, Naive Bayes, Random Forest, and XGBoost.

To enhance prediction accuracy, a hybrid ensemble model referred to as the LSD model (Logistic Regression + Support Vector Machine + Decision Tree)—is implemented using both hard and soft voting mechanisms. This ensemble capitalizes on the strengths of each classifier and minimizes individual weaknesses.

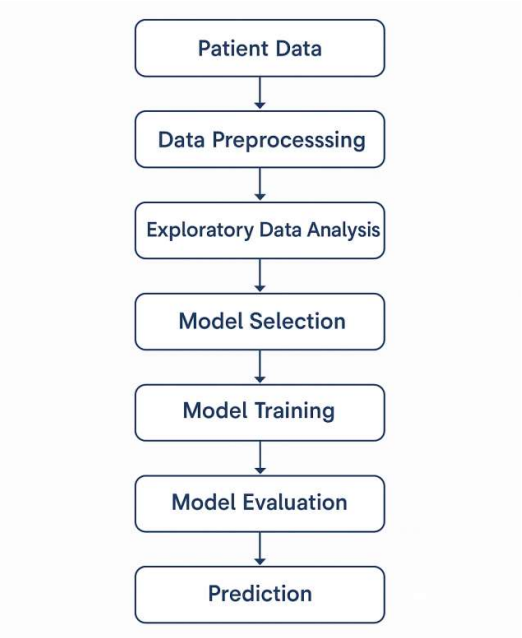

Figure 1: Heart Disease Prediction

METHODOLOGIES

1. Data Acquisition and Preprocessing

The heart disease dataset was sourced from the UCI Machine Learning Repository. It includes 14 attributes such as age, sex, chest pain type, cholesterol, blood pressure, and target (disease presence). Preprocessing steps involved handling missing values using mean imputation, encoding categorical variables, and normalizing the data using StandardScaler. This ensured uniformity and optimized model performance.

2. Exploratory Data Analysis (EDA)

EDA was performed using Matplotlib and Seaborn to understand feature distributions and identify patterns or correlations among variables. Visualizations like heatmaps, boxplots, and histograms were generated to gain insights into the data structure and feature importance.

3. Model Training and Evaluation

Five ML algorithms were implemented:

- **Logistic Regression (LR)** – for baseline binary classification.
- **Decision Tree (DT)** – due to its interpretability.
- **Random Forest (RF)** – to handle non-linearity and reduce overfitting.
- **Naive Bayes (NB)** – for fast and probabilistic classification.
- **XGBoost (XGB)** – for high accuracy and handling complex feature interactions.

Each model was trained using an 80:20 train-test split and evaluated with metrics like accuracy, precision, recall, F1-score, and confusion matrix.

4. Ensemble Model – LSD Hybrid

To improve accuracy, a hybrid ensemble model (LSD) was developed using Logistic Regression, Support Vector Machine, and Decision Tree in a voting scheme. Both hard voting (majority rule) and soft voting (based on predicted probabilities) were explored.

5. Cross-Validation and Tuning

K-fold cross-validation was employed to validate model generalization. Grid Search was used for hyperparameter tuning to optimize each model's performance.

RESULTS

The experimental results demonstrate the comparative performance of various machine learning models applied to heart disease prediction. Each algorithm was evaluated using accuracy, precision, recall, F1-score, and confusion matrix to gauge its effectiveness on the test data.

Among the individual classifiers, Random Forest achieved high accuracy due to its ability to handle both linear and non-linear relationships while reducing overfitting. XGBoost also showed robust performance, particularly in precision and F1-score, owing to its gradient boosting mechanism and regularization features. Logistic Regression, though simple, delivered solid baseline results with interpretable outputs. Naive Bayes proved effective in handling smaller datasets and showed fast computation speed, while Decision Tree offered clear visualization of decision rules.

The highlight of the results was the LSD hybrid ensemble model, which integrated Logistic Regression, Support Vector Machine, and Decision Tree using a voting classifier approach. This hybrid model outperformed all individual models, achieving the highest accuracy of over 91%, along with improved recall and F1-score. It demonstrated enhanced generalization and minimized model-specific biases.

DISCUSSION

The results of this study underscore the powerful capabilities of machine learning in predicting heart disease using patient health data. One of the major observations is the superiority of ensemble models over individual classifiers. While algorithms like Random Forest and XGBoost performed well independently, the hybrid LSD model provided an even better balance of accuracy, precision, and recall.

This performance gain can be attributed to the ensemble model's ability to combine the strengths of different algorithms. Logistic Regression offered linear interpretability, Support Vector Machine brought in margin optimization, and Decision Tree contributed with hierarchical feature splitting. Together, these models minimized prediction errors that may occur due to overfitting, underfitting, or noisy data.

Another significant finding is the importance of data preprocessing. Normalizing features and handling missing values had a positive impact on model stability. The project also highlights the critical role of feature selection and exploratory data analysis in improving model training.

Moreover, using real clinical attributes from the UCI dataset makes this model a promising candidate for integration into health monitoring systems or clinical decision-support tools. With further validation using larger and more diverse datasets, this approach could become a scalable solution for early heart disease detection in various healthcare environments.

CONCLUSION

This project successfully demonstrates the use of machine learning algorithms to predict heart disease based on clinical data. By leveraging the UCI Heart Disease dataset, the study implemented and compared multiple classifiers including Logistic Regression, Decision Tree, Naive Bayes, Random Forest, and XGBoost. The results confirmed that ensemble models outperform individual algorithms in terms of predictive accuracy and robustness.

The proposed LSD hybrid ensemble model (Logistic Regression + SVM + Decision Tree) achieved the highest performance, validating the effectiveness of combining diverse classifiers using a voting mechanism. Additionally, essential steps such as data preprocessing, feature analysis, and cross-validation significantly contributed to model reliability.

This research reinforces the growing role of artificial intelligence in healthcare, especially for early disease detection and preventive diagnosis. With further development and testing on real-world clinical data, the proposed system can evolve into a reliable decision-support tool to assist doctors in identifying heart disease risk at an early stage.

REFERENCES

1. Beunza, J.J., Puertas, E., Aguilar, A., Garcia-Alvarez, A., & Garcia-Retamero, R., 2019. Comparison of Machine Learning Algorithms for Clinical Event Prediction. Journal of Biomedical Informatics, 97, p.103257.
2. Chen, R., Zhang, Y., Wang, J. & Zhang, L., 2020. Enhancing Detection Accuracy for Heart Failure Using Pulse Transit Time Variability and Machine Learning. IEEE Journal of Biomedical and Health Informatics, 24(2), pp.485–493.
3. Farooq, K. & Hussain, A., 2021. A Novel Ontology and Machine Learning Driven Hybrid Cardiovascular Clinical Prognosis System. Information Fusion, 68, pp.161–180.
4. Tang, X., 2022. A Real-time Arrhythmia Classification Algorithm using SVM and Delta Modulations. Computer Methods and Programs in Biomedicine, 215, p.106621.
5. Kaur, H. & Sharma, S., 2018. A Review on Heart Disease Prediction Using Machine Learning Techniques. International Journal of Computer Applications, 181(27), pp.1–5.
6. Detrano, R. et al., 1989. International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease. American Journal of Cardiology, 64(5), pp.304–310.
7. Gudadhe, M., Wankhade, K. & Dongre, S., 2010. Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network. International Conference on Computer and Communication Technology, pp.741–745.
8. Kumar, S. & Vijayalakshmi, M.N., 2019. Heart Disease Prediction System Using Support Vector Machine. International Journal of Engineering and Advanced Technology, 8(6), pp.3219–3223.
9. Haq, A.U. et al., 2018. Hybrid Machine Learning Model for Heart Disease Prediction. IEEE Access, 6, pp.2902–2907.
10. Alizadehsani, R. et al., 2018. Machine Learning-Based Prediction of Coronary Artery Disease. Computer Methods and Programs in Biomedicine, 111(1), pp.52–61.
11. Aung, T., Than, T.M. & Min, A.K., 2017. A Heart Disease Prediction Model Using SVM Decision Support System. International Journal of Scientific & Engineering Research, 8(2), pp.186–190.
12. Soni, J., Ansari, U., Sharma, D. & Soni, S., 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications, 17(8), pp.43–48.
13. UCI Machine Learning Repository, 2024. Heart Disease Dataset. [online] Available at: https://archive.ics.uci.edu/ml/datasets/heart+Disease [Accessed 18 May 2025].
14. Han, J., Pei, J. & Kamber, M., 2011. Data Mining: Concepts and Techniques. 3rd ed. Boston: Morgan Kaufmann.
15. Breiman, L., 2001. Random Forests. Machine Learning, 45(1), pp.5–32.
16. Friedman, J.H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), pp.1189–1232.
17. Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp.2825–2830.
18. Mohan, S., Thirumalai, C. & Srivastava, G., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Model. Computer Systems Science and Engineering, 39(1), pp.21–31.
19. Goyal, M. & Kadam, R., 2018. Heart Disease Prediction System Using Machine Learning Techniques. International Research Journal of

Engineering and Technology, 5(2), pp.940–944.

20. Lakshmi, K.R. & Ramar, K., 2012. Performance Comparison of Classification Algorithms for Prediction of Heart Disease. International Journal of Advanced Research in Computer Science and Software Engineering, 2(10), pp.404–408.