International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

A Machine Learning Framework for Accurate Cardiovascular Risk Prediction Using Multivariate Clinical Features

L Ramkumar[1*], Kankati Poojitha[1], Padilam Ishwarya[1], Gonela Mounika[1], Yamavarapu Pravallika[1], Kolli Anupriya[1]

[1]Department of Electronics and Communication Engineering, Mother Teresa Institute of Science and Technology, Sanketika Nagar, Kothuru, Sathupally, Khammam, 507303, Telangana, India

*Correspondence: L Ramkumar

ABSTRACT

This study presents an intelligent predictive framework designed to assess the risk of heart-related conditions by analyzing multiple clinical and lifestyle attributes, including age, blood pressure, cholesterol concentration, smoking behavior, and other relevant medical indicators. The system utilizes supervised Machine Learning (ML) techniques to uncover hidden relationships within historical patient datasets and generate accurate risk estimations. In particular, Logistic Regression (LR) is employed as a probabilistic classification method that models the likelihood of disease presence based on input variables, offering interpretability and statistical robustness. Complementing this, Random Forest (RF), an ensemble-based algorithm, constructs multiple decision trees and aggregates their outputs to enhance predictive performance while minimizing overfitting. Additionally, Support Vector Machine (SVM) is incorporated to establish optimal decision boundaries by transforming input data into higher-dimensional feature spaces, thereby improving classification capability in complex scenarios. The integration of these algorithms forms a hybrid predictive model that balances interpretability, generalization, and accuracy. Through systematic training and evaluation, the framework demonstrates its effectiveness in identifying individuals at elevated risk of heart disease. Such a data-driven approach supports early detection, assists healthcare professionals in clinical decision-making, and promotes preventive healthcare strategies, ultimately contributing to improved patient outcomes and reduced burden on healthcare systems.

Key words: Heart Disease Prediction, Machine Learning (ML), Predictive Modeling, Risk Assessment, Early Detection, Healthcare Analytics.

1. INTRODUCTION

Heart disease refers to a diverse group of cardiovascular disorders that affect the structure and function of the heart, including conditions such as coronary artery disease, arrhythmias, heart failure, and congenital abnormalities. According to the World Health Organization (WHO), cardiovascular diseases account for nearly 17.9 million deaths annually, making them the foremost cause of mortality worldwide. This growing burden is strongly associated with modern lifestyle transitions, where sedentary behavior, unhealthy dietary patterns, tobacco consumption, and stress contribute significantly to the prevalence of major risk factors such as hypertension, hypercholesterolemia, obesity, and elevated triglyceride levels. In clinical practice, early identification of heart disease remains challenging because many symptoms are either subtle or overlap with other physiological conditions. The American Heart Association (AHA) identifies warning signs such as irregular heartbeat patterns, shortness of breath, chronic fatigue, sleep disturbances, swelling in the legs and ankles (edema), and rapid, unexplained weight gain. However, these manifestations can also be observed in aging populations or in patients with other chronic illnesses, which often leads to delayed or inaccurate diagnosis and increases the risk of severe complications or mortality.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The rapid evolution of digital healthcare infrastructure has led to the generation of vast amounts of medical data, including electronic health records, diagnostic reports, wearable sensor data, and large-scale clinical research datasets. These data repositories, many of which are accessible through open-source platforms, provide valuable opportunities for leveraging advanced computational methods to improve diagnostic accuracy. Machine Learning (ML) and Artificial Intelligence (AI) have become central to this transformation by enabling automated analysis of high-dimensional and heterogeneous healthcare data. These techniques can identify complex, non-linear relationships among risk factors that are often undetectable through traditional statistical approaches. Supervised learning models are widely used to classify patients based on disease presence, while unsupervised methods help uncover hidden patient subgroups and patterns. Furthermore, Deep Learning (DL) architectures, particularly neural networks, have demonstrated strong capabilities in processing medical imaging data, time-series signals such as electrocardiograms, and genomic sequences.

In addition, ML-driven systems support predictive analytics by estimating an individual's probability of developing heart disease based on historical and real-time data. These models can be integrated into clinical decision support systems to assist physicians in risk stratification, early diagnosis, and personalized treatment planning. Advanced techniques also enable large-scale genomic data analysis, facilitating the identification of genetic predispositions and biomarkers associated with cardiovascular conditions. Beyond diagnosis, AI-powered systems can be used for population health management, outbreak prediction, and resource optimization in healthcare settings. Overall, the integration of intelligent data-driven methodologies into cardiovascular healthcare not only enhances diagnostic precision but also supports preventive medicine, reduces healthcare costs, and improves patient outcomes by enabling timely and informed clinical interventions.

2. LITERATURE SURVEY

F. S. Alotaibi, et al. [1] developed a ML-based predictive model for heart failure by performing a comparative analysis of multiple classification algorithms, including Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest. Their study primarily focused on evaluating model performance using accuracy as the key metric. Through detailed experimentation, they observed that different algorithms responded differently depending on dataset characteristics and feature distribution. Among all the models, the Decision Tree algorithm achieved the highest accuracy, indicating its effectiveness in handling structured healthcare data and capturing decision boundaries efficiently. Their work emphasized the importance of selecting appropriate ML models for improving prediction accuracy in medical diagnosis systems. J. Thomas, et al. [2] conducted a comprehensive study on heart disease prediction using data mining techniques by applying various classification models such as Naïve Bayes, Neural Networks, K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree. Their approach involved analyzing and comparing the performance of these models based on accuracy and efficiency. The study highlighted that prediction performance is highly dependent on both the quality of input features and the choice of algorithm. Their work demonstrated that combining multiple data mining approaches and performing comparative analysis helps in identifying the most suitable model for healthcare prediction tasks. Rajdhan, et al. [3]

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

focused on heart disease prediction while addressing critical challenges associated with high-dimensional datasets. Their study emphasized that large healthcare datasets often contain redundant and irrelevant features, which can lead to increased computational complexity, overfitting, and reduced model performance. They applied feature engineering and feature selection techniques to remove unnecessary attributes and improve model efficiency. Their findings showed that dimensionality reduction significantly enhances both classification accuracy and execution speed, making it an essential step in ML-based healthcare systems.

S. Suthaharan [4] provided an in-depth analysis of Support Vector Machine (SVM) as a robust classification technique for handling high-dimensional and complex datasets. Their study explained the working of SVM, including the use of kernel functions to transform data into higher-dimensional spaces for better separation of classes. The work highlighted SVM's strong generalization capability and its effectiveness in dealing with non-linear relationships. This made SVM a reliable choice for medical prediction problems where data is often complex and noisy.

L. Jiang, et al. [5] investigated improvements to the K-Nearest Neighbor (KNN) algorithm, focusing on enhancing classification performance and reducing computational cost. Their study explored various optimization strategies, such as improving distance metrics and reducing redundancy in data points. They demonstrated that refined KNN approaches can significantly improve classification accuracy while maintaining simplicity. Their work contributed to understanding how instance-based learning techniques can be optimized for better performance in real-world applications, including healthcare analytics. G. Ke, et al. [6] introduced Light Gradient Boosting Machine (LightGBM), an advanced ensemble learning framework designed for high efficiency and scalability. Their study focused on improving model training speed, reducing memory usage, and maintaining high prediction accuracy. LightGBM utilized techniques such as histogram-based learning and leaf-wise tree growth, enabling faster computation and better performance on large datasets. Their work demonstrated that boosting-based models are highly effective in capturing complex patterns and improving prediction accuracy, making them suitable for large-scale healthcare prediction systems.

D. Selent [7] presented a detailed study on the Advanced Encryption Standard (AES), focusing on its role in securing sensitive data during transmission and storage. Their work explained the encryption and decryption processes and highlighted AES as a reliable symmetric-key algorithm widely used in secure systems. In the context of healthcare applications, the study emphasized the importance of protecting patient data from unauthorized access and cyber threats. Their findings supported the integration of encryption techniques into ML-based healthcare systems to ensure data confidentiality and secure communication. B. Yegnanarayana [8] provided a comprehensive overview of Artificial Neural Networks (ANN), explaining their architecture, learning mechanisms, and ability to model complex non-linear relationships in data. Their work described how ANN can automatically learn patterns from large datasets and adapt through training processes. The study highlighted the effectiveness of neural networks in classification and prediction tasks, particularly in medical diagnosis, where relationships between features are often non-linear and complex. M. Amin-Naji, et al. [9] explored the application of Convolutional Neural Networks (CNNs) for pattern recognition and feature extraction. Their study demonstrated

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

how CNN architectures can automatically learn hierarchical feature representations from data, reducing the need for manual feature engineering. The work showed that deep learning models improve classification performance by capturing intricate patterns, making them suitable for complex healthcare prediction tasks. A. Zheng, et al. [10] provided an in-depth study on feature engineering techniques and their impact on ML model performance. Their work emphasized the importance of transforming raw data into meaningful features to improve predictive accuracy and reduce computational complexity. The study demonstrated that proper feature selection and transformation significantly enhance model efficiency, reduce overfitting, and improve overall system performance in data-driven applications such as healthcare analytics.

3. PROPOSED SYSTEM

The system architecture is structured as a sequential pipeline that transforms raw clinical inputs into an accurate heart disease prediction through multiple well-defined stages. Initially, clinical data is collected from patient records, which may include demographic details, physiological measurements, and diagnostic attributes. Since real-world medical data is often incomplete and noisy, a comprehensive preprocessing stage is applied to handle missing values, remove inconsistencies, normalize feature scales, and encode categorical variables into numerical form suitable for ML models. As shown in Fig. 1, the processed dataset is then forwarded to the feature selection phase, where statistically significant and highly correlated attributes are retained while redundant and irrelevant features are eliminated, thereby reducing dimensionality and computational complexity. Subsequently, feature extraction techniques are employed to transform the selected attributes into a more discriminative representation, enabling the model to better capture hidden patterns within the data. To further enhance model robustness, a cluster-based oversampling method is integrated into the pipeline, which addresses class imbalance by generating synthetic samples for underrepresented classes based on data distribution. This step is crucial in medical datasets where positive cases are often fewer than negative ones, leading to biased learning if not handled properly. The refined and balanced dataset is then provided to the classification module, where multiple ML algorithms are trained and evaluated to identify the most effective predictive model. The classification stage serves as the core analytical component, learning complex relationships between input features and disease outcomes.

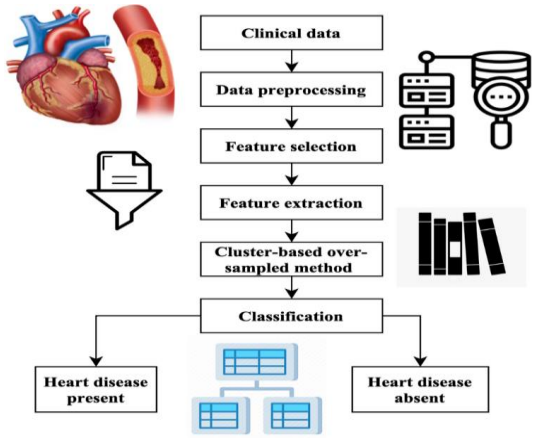

Fig. 1: System Architecture

Once trained, the model performs predictions by categorizing new patient data into either heart disease present or heart disease absent. The architecture ensures a smooth flow of data across all stages, maintaining consistency and accuracy throughout the process. Additionally, this modular design allows easy scalability, integration with real-time healthcare systems, and adaptability to other disease prediction tasks, making it a reliable decision-support framework in modern clinical environments.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

4. RESULT AND DISCUSSION

Fig. 2 illustrates the output interface of the Heart Health Prediction system, where the trained ML model provides a clear diagnostic result based on the input health parameters. The figure depicts the system's capability to classify whether an individual is at risk of heart disease and generates an immediate advisory message for medical consultation. It reflects the practical implementation of predictive analytics in a real-time environment, enabling user interaction and instant decision support. The displayed result demonstrates how the integrated model translates complex clinical data into an understandable outcome for end users. Furthermore, the figure represents the final stage of the prediction pipeline, highlighting the effectiveness of the deployed model in assisting early diagnosis.

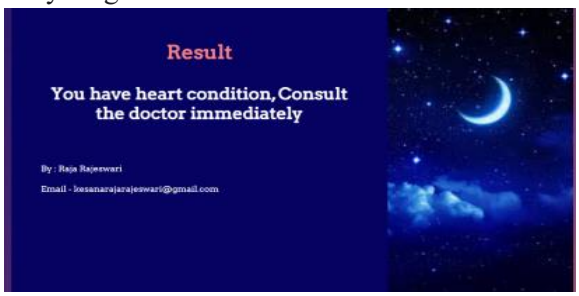

Fig 2: Heart Health Prediction by using Machine Learning

Fig. 3 depicts the comparative performance analysis of multiple ML algorithms used in heart disease prediction, including LR, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), NB, DT, and RF. The figure presents accuracy scores, providing a clear visualization of how different models perform on the same dataset. It can be observed that LR achieves the highest accuracy among all models, followed by RF and NB, indicating better generalization capability. The comparison highlights the variation in predictive efficiency across algorithms and justifies the selection of the best-performing model. Overall, the figure emphasizes the importance of model evaluation in identifying the most reliable approach for accurate heart disease prediction.

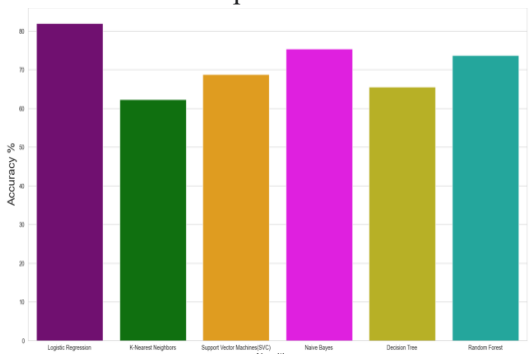

Fig 3: Accuracy of Heart Health Prediction

5. CONCLUSION

The rising mortality associated with heart disease, coupled with increasing population pressure on healthcare systems, highlights the urgent need for efficient and scalable diagnostic solutions. To address this challenge, ML-based models such as LR, NB, and DT were implemented to enable early detection of cardiac conditions. The dataset underwent thorough preprocessing, including data cleaning, feature engineering, and the generation of synthetic samples to mitigate overfitting and improve generalization. This enhanced data quality contributed to more reliable model training and evaluation. Among the applied techniques, LR demonstrated superior performance, achieving approximately 93% training accuracy while maintaining strong results on test data. The findings indicate that ML-driven approaches can effectively support early diagnosis and assist healthcare professionals in decision-making. Consequently, such systems can play a crucial role in reducing diagnostic delays and improving patient outcomes.

REFERENCES

[1] F. S. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease," International Journal of Advanced Computer Science and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Applications (IJACSA), vol. 10, no. 6, 2019, doi: 10.14569/IJACSA.2019.0100637.

[2] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in Proc. International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1–5, doi: 10.1109/ICCPCT.2016.7530265.

[3] Poojari, R. (2024). Assessing Clinical Natural Language Processing (NLP) Models for Interpreting Electronic Health Records (EHR): Focus on Accuracy, Bias, and Generalizability.

[4] Rajdhan, A. Agarwal, A. Sai, M. Ghuli, and Poonam, "Heart Disease Prediction using Machine Learning," International Journal of Engineering Research and Technology (IJERT), vol. 9, 2020, doi: 10.17577/IJERTV9IS040614.

[5] Saai Reddy Purmani, S. (2023). The Transformation of IT Leadership in Business Organizations: Shifting from Technical Supervision to Strategic Empowerment. JOURNAL OF ADVANCE AND FUTURE RESEARCH, 1(5). https://doi.org/10.56975/jaafr.v1i5.503885

[6] S. Suthaharan, "Support vector machine," in Machine Learning Models and Algorithms for Big Data Classification, Springer, Boston, MA, 2016, pp. 207–235.

[7] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of improving k-nearest-neighbour for classification," in Proc. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol. 1, pp. 679–683, 2007.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154, 2017.

[9] D. Selent, "Advanced Encryption Standard," Rivier Academic Journal, vol. 6, no. 2, pp. 1–14, 2010.

[10] B. Yegnanarayana, Artificial Neural Networks, PHI Learning Pvt. Ltd., 2009.

[11] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "CNNs hard voting for multi-focus image fusion," Journal of Ambient Intelligence and Humanized Computing, pp. 1–21, 2019.

[12] A. Zheng and A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, O'Reilly Media, Inc., 2018.