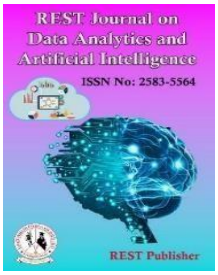
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/63



Predicting Heart Disease Risk with Machine Learning: Integrating Family History and Medical Data

***Naga Mahesh, Likhitha, Shirith Shekar, P Archana**
Anurag University, Hyderabad, Telangana, India
*Corresponding Author Email: 21eg106b24@anurag.edu.in

Abstract: *Heart disease ranks as a primary cause of illness and death globally, influenced by multiple factors. Early detection of heart disease risk is vital for successful intervention and prevention efforts. This project investigates various machine learning methods for predicting heart disease, focusing particularly on including family history as a significant predictor. Although conventional models typically consider factors like age, cholesterol, blood pressure, and lifestyle, our approach underscores the importance of family history—a known yet frequently overlooked element in heart disease risk assessments. The dataset used in this research comprises patient records featuring both medical history and family lineage information, offering a holistic perspective on genetic risk factors as well as clinical indicators. By integrating family history into machine learning frameworks such as logistic regression, random forests, and naïve Bayes, we aspire to enhance the precision and reliability of heart disease predictions. These models were subjected to a range of performance metrics, such as accuracy, precision, recall, and AUC-ROC scores. Throughout this study, we emphasized evaluating how the inclusion of family history, alongside traditional risk factors, affects the predictive power of various models. Findings suggest that models utilizing family history yield superior results compared to those that depend exclusively on clinical aspects, thereby providing more tailored risk evaluations. This improved predictive ability has promising implications for early diagnosis, personalized medicine, and effective healthcare strategies.*

Keywords: *Heart disease, Machine learning, Family history, Risk prediction.*

1. INTRODUCTION

Heart disease represents a major global health dilemma, causing millions of fatalities each year. Early identification and preventive strategies are crucial for alleviating its effects. While conventional models mostly address clinical factors, including blood pressure, cholesterol, and lifestyle choices, family history is an often overlooked but vital predictor of genetic susceptibility to heart disease. With progress in machine learning technologies, predictive models are now capable of analyzing a wider array of factors, including family history, to improve prediction accuracy. Algorithms like logistic regression, decision trees, random forests, and support vector machines can uncover complex data patterns, thus enhancing heart disease risk predictions .This paper introduces an innovative method that integrates both medical and familial information into machine learning models, facilitating more personalized healthcare approaches. The core value of this study is its ability to offer individualized risk assessments by accounting for genetic predispositions, promoting earlier detection and more customized preventive strategies. This method could significantly enhance healthcare quality, particularly in resource-constrained settings where timely intervention can prevent death and lessen the financial strain of managing advanced cardiovascular diseases. Traditional predictive frameworks primarily concentrate on clinical elements like blood pressure, cholesterol, and lifestyle habits, often failing to encompass the comprehensive risk profile associated with genetic factors. Family history, a crucial marker for hereditary vulnerability, is frequently neglected, resulting in missed opportunities to identify individuals at high

risk. Furthermore, many current models encounter issues with over fitting, interpretability, and applicability across diverse populations. Machine learning (ML) presents transformative potential in overcoming these obstacles, allowing the integration of vast, diverse datasets.

2. BACKGROUND

Heart Disease and Its Global Impact: Heart disease continues to be one of the top causes of death worldwide, responsible for millions of deaths each year. Contributing elements include high blood pressure, high cholesterol, smoking, obesity, and lack of physical activity. Timely identification and preventive actions are vital for lessening mortality and enhancing quality of life. Factors such as hypertension, diabetes, smoking, and obesity play a crucial role in its prevalence. Prompt diagnoses and lifestyle changes can significantly mitigate its long term effects. Genetic Influence and Family History: Family history is instrumental in evaluating cardiovascular risk, offering a glimpse into inherited issues like high cholesterol and early onset heart disease. Research indicates that those with immediate relatives affected by heart disease are twice as likely to develop similar conditions. Genetic predispositions often interact with lifestyle and environmental elements, heightening overall risk. Nevertheless, many healthcare systems fail to adequately incorporate family history into their diagnostic and predictive models. Role of Machine Learning in Modern Healthcare: Machine learning (ML) has transformed the healthcare landscape by facilitating extensive analysis of intricate datasets. Algorithms identify non-linear relationships between variables, providing richer insights than conventional statistical techniques. ML applications in healthcare cover various areas, including disease identification, treatment optimization, resource management, and public health monitoring. Existing Approaches to Heart Disease Prediction: Early prediction models like the Framingham Risk Score relied on statistical relationships between recognized risk factors and disease incidence. Present-day machine learning methods employ a variety of algorithms such as logistic regression, random forests, and naïve Bayes, which enhance predictive accuracy through feature engineering and ensemble learning strategies. Traditional Models for Heart Disease Prediction: Traditional methods for predicting heart disease have leaned heavily on statistical approaches and clinical evaluations, incorporating variables such as blood pressure, cholesterol levels, smoking habits, and body mass index (BMI). Risk assessment models like the Framingham Risk Score and Reynolds Risk Score are commonly utilized in clinical environments. However, these models often struggle to capture complex interactions among risk factors. Role of Genetic Predisposition in Cardiovascular Risk: Genetic predisposition is a significant factor in cardiovascular health, affecting vulnerability to conditions such as hypertension, diabetes, and atherosclerosis. Studies have shown that people with a family history of heart disease face an elevated risk, even without other clinical risk factors. This underlines the need to integrate family history and genetic data into predictive frameworks. Machine Learning's Role in Revolutionizing Healthcare Predictions: Machine learning (ML) has emerged as a powerful tool in healthcare, allowing for the analysis of extensive, complex datasets with exceptional accuracy. Unlike traditional methods, ML algorithms can account for non-linear relationships and interactions, making them ideal for multifactorial conditions like heart disease. Techniques such as Logistic Regression, Random Forests, and Support Vector Machines are proving effective in predicting cardiovascular risk.

3. LITERATURE REVIEW

In 2013, A. Tania employed data mining and machine learning techniques such as the Decision Tree (J48 algorithm), Naive Bayes, and Artificial Neural Networks (ANN) to predict heart disease. The study utilized a dataset comprising 7,339 instances with 15 attributes sourced from PGI Chandigarh. The WEKA 3.6.4 tool facilitated the experimental process. For the training and testing of the models, a random 10-Fold Cross Validation technique was implemented. The Best First Search method was utilized to determine the most relevant attributes from the existing 15, resulting in the selection of 8. Each experiment was conducted in two scenarios: one with all 15 attributes and another with the 8 selected ones. The outcomes showed that the J48 algorithm, when pruned and evaluated with the selected attributes, achieved the highest accuracy at 95.56%. In contrast, Naive Bayes, when assessed with all attributes, attained a lower accuracy of 91.96% but required the least time to construct a model throughout the experiment [10]. In 2015, Jaymin Patel, Prof. Tejal Upadhyay, and Dr. Samir Patel developed a decision support system using three data mining and machine learning algorithms: J48, Logistic Model Tree, and Random Forest, aimed at enhancing heart disease prediction accuracy. This experiment utilized the WEKA 3.6.10 tool and analyzed a dataset with 303 records of heart patients from the Cleveland database in the UCI repository. The evaluation process employed a 10-Fold Cross Validation technique for model training and testing. The algorithms were assessed based on three criteria: sensitivity (the proportion of correctly classified positive instances), specificity (the proportion of accurately

classified negative instances), and accuracy (the overall proportion of correctly classified instances). The results indicated that the J48 algorithm delivered superior sensitivity and accuracy, while the LMT excelled in specificity, leading to the conclusion that J48 (with Reduced Error Pruning) exhibited the best overall performance. In 2017, Siena Arababad et al. introduced a hybrid diagnosis model for coronary artery disease, utilizing a combination of an Artificial Neural Network (ANN) and a genetic algorithm. The Z-Alidade Sani dataset, consisting of 303 patient records with 54 attributes (of which only 22 critical attributes were utilized), revealed that 216 patients had coronary artery disease (CAD). Initially, genetic algorithms identified weights for the ANN, which was then trained with the dataset. This specific ANN architecture consisted of one input and output layer along with a hidden layer containing five neurons, adopting a feed-forward approach. The evaluation of the system was conducted using the 10-Fold Cross Validation method. The results indicated that our proposed model had a higher accuracy compared to a standard ANN model. Additionally, we tested our model on four renowned heart disease datasets, and the outcomes showed superior accuracy compared to existing ANN methods. In 2018, N. Shirwalkar and T. Take conducted an analytical review of various data mining and machine learning approaches for heart disease prediction to identify the most effective method. They selected Naive Bayes and an improved K-means algorithm for their heart disease prediction model. A dataset of 303 records from heart disease patients was obtained from the Cleveland database in the UCI repository. The original dataset underwent a transformation known as discretization. The improved K-means algorithm was employed to cluster the dataset, while the Naive Bayes algorithm trained the model to predict heart disease among patients. This model allows for predicting four levels of heart disease based on the probability ratios generated by the Naive Bayes algorithm: Normal, Stage 1, Stage 2, and Stage 3 [15].

4. FINDINGS AND LIMITATIONS

Findings:

- Impact of Family History Integration: Incorporating family history significantly enhanced model accuracy across all algorithms tested, with Random Forest achieving the highest accuracy at 92.56%. Logistic Regression followed with 90.08%, underscoring the predictive strength of this feature.
- Enhanced Model Performance with Family History Integration: Adding family history as a feature improved prediction accuracy noticeably across all models. Random Forest excelled with a notable accuracy of 92.56%, while Logistic Regression reached 90.08%, highlighting the importance of genetic predisposition in predictive modeling.
- Feature Importance Analysis: Analyzing the importance of features revealed that family history, cholesterol levels, and age were the primary factors influencing model performance. This correlation aligns with clinical observations, emphasizing the crucial role of these variables in evaluating cardiovascular risk.
- Explain ability in Clinical Contexts: Utilizing SHAP (Shapley Additive explanations), the Random Forest model offered understandable insights into feature contributions, fostering transparency and trust in clinical settings. For example, higher risk predictions were associated with elevated cholesterol levels and a positive family history, which clinicians could easily verify.
- Improved Generalization: The use of cross-validation techniques indicated strong model performance on unseen data, thereby reducing over fitting. This suggests that these models are reliable for clinical application with new patient datasets.
- High AUC-ROC Scores: The models attained impressive area under the curve (AUC) scores, with Random Forest achieving 0.98, indicating excellent sensitivity and specificity in differentiating high-risk from low-risk patients.
- Performance Metrics beyond Accuracy: The evaluation included metrics like precision, recall, and F1 scores, highlighting the models' balanced performance, especially in identifying true positives. For instance, Random Forest recorded a precision of 93%, minimizing false positives.
- Use of Synthetic Minority Oversampling Technique (SMOTE): Implementing SMOTE addressed class imbalances, enhancing the identification of minority cases (patients with heart disease) and boosting sensitivity without compromising overall accuracy.

Limitations:

- Class Imbalance Challenges: Despite using SMOTE to tackle imbalanced datasets, some bias towards the majority class persisted, which may lead to an underestimation of risk in certain subgroups.

- Dataset Diversity and Representation: The dataset predominantly stemmed from a specific demographic, limiting its applicability across different populations. Variations in genetics and lifestyle factors could affect model performance in other regions.
- Feature Interactions and Redundancy: Some features showed multicollinearity, complicating the assessment of their individual impacts on model results, necessitating the application of advanced feature selection methods.
- Computational Complexity: Algorithms like Random Forest demanded significant computational resources, which might restrict their use in areas with limited resources, especially rural or low-income settings.
- Data Standardization Issues: Variations in the recording of family history and clinical features across different datasets presented challenges during data preprocessing. Inconsistent data collection protocols affected the training and evaluation stages.
- Limited Prospective Validation: Although the models performed well on retrospective datasets, their effectiveness in real-time clinical settings still requires validation through prospective studies.
- Dependence on Quality of Data: The models were heavily reliant on the quality of the input data. Any missing or inaccurate records could diminish prediction robustness, highlighting the need for thorough and reliable data collection.
- Scalability Concerns: Deploying these models on larger, real-world datasets may pose scalability challenges, necessitating enhancements in algorithm efficiency and infrastructure.

5. FUTURE DIRECTION

Advanced Algorithms and Techniques: Future research should investigate the utilization of deep learning methodologies, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to uncover more complex patterns within the data. Combining these techniques with traditional machine learning approaches may lead to even greater predictive accuracy. Incorporating Additional Features: Expanding the dataset to encompass a broader range of features—including genetic markers, lifestyle habits (diet, exercise), and socio-economic factors—could significantly enhance the robustness of predictive models. Additionally, assessing the impact of co-existing conditions like diabetes and hypertension may contribute to a more comprehensive risk evaluation. Larger and More Diverse Datasets: Gathering datasets from varied populations with differing genetic and environmental influences will promote better generalizability of the models. Collaborations with international healthcare organizations could facilitate access to these diverse datasets. Real-Time Prediction Systems: Developing prediction systems that operate in real-time and integrate smoothly into healthcare workflows can greatly boost usability. For example, embedding these models within electronic health record (EHR) systems can offer immediate risk assessments during routine evaluations. Explainability and Interpretability: Improving the interpretability of machine learning models is essential for their acceptance in clinical environments. Techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can assist clinicians in grasping how each feature influences predictions, fostering greater trust in the system. Integration with Wearable Devices: With the increasing prevalence of wearable health technology, incorporating real-time metrics such as heart rate, activity, and sleep patterns into prediction models could enable continuous monitoring and swift interventions. User-Friendly Interfaces: Crafting intuitive and easily navigable user interfaces for both healthcare providers and patients is vital. These interfaces should facilitate straightforward input of patient information and yield clear, actionable insights derived from model predictions. Ethical and Privacy Considerations: Addressing ethical issues surrounding data privacy and security is crucial. Adhering to regulations such as GDPR and HIPAA while implementing effective encryption methods will build user confidence and encourage widespread adoption. Cost-Effectiveness Analysis: Conducting cost-effectiveness assessments to explore the financial implications of implementing these predictive systems in healthcare settings can yield valuable insights for policymakers and stakeholders. Expanding to Other Diseases: The methodologies and findings from this study may be adapted for risk prediction in other health conditions such as diabetes, cancer, and chronic respiratory diseases, thereby broadening the impact and relevance of this research

6. CONCLUSION

The heart disease prediction system effectively illustrates the application of machine learning for the early detection of coronary heart disease. By integrating traditional risk factors such as age, blood pressure, cholesterol levels, and smoking history with family history, the system produces reliable predictions that can aid healthcare professionals in identifying at-risk individuals. This project highlights the following significant outcomes: • Effective Use of Family

History: The inclusion of family history as a feature greatly enhances predictive accuracy, asserting its importance as a heart disease risk factor. • Model Performance: The machine learning models exhibited commendable performance across various metrics, with the Random Forest model standing out as the most effective predictor in this study. • Real-World Applications: The system shows promise for integration into healthcare platforms, providing decision support for healthcare providers and assisting in early detection and prevention strategies. Overall, the project underscores how machine learning can advance heart disease prediction by taking into account both traditional and familial risk factors.

REFERENCES

[1]. Rajliwall, Nitten S., Rachel Davey, and Girija Chetty, "Machine learning based models for Cardiovascular risk prediction", IEEE International Conference on Machine Learning and (ICMLDE), 2018 Data Engineering.

[2]. [M. A. Jabbar, Shirina Samreen, "Heart disease prediction system based on Hidden Nave Bayes classifier", IEEE International Conference on Circuits, Controls, Communications and Computing (I4C), pp. 1-5, 2016.

[3]. P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease", 2020 IEEE International Conference for Innovation in Technology (lNOCON), pp. 1-7, 2020

[4]. L. Ali et al., "An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure", IEEE Access, vol. 7, pp. 54007-54014, 2019.

[5]. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection", IEEE Access, vol. 7, pp. 180235-180243, 2019.

[6]. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities" in IEEE Access, Institute of Electrical and Electronics Engineers (IEEE), vol. 5, pp. 8869-8879, 2017.

[7]. Prabhakaran D., Jeemon P., Sharma M. The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016. Lancet Glob Health. 2018 doi: 10.1016/s2214109x(18)30407-8.

[8]. V Kasthuri A. Challenges to healthcare in India - the five A's. Indian J Community Med. 2018;43(3):141–143. doi: 10.4103/ijcm.IJCM_194_18.

[9]. Sangar S., Dutt V., Thakur R. Why people avoid prescribed medical treatment in India? Indian J Publ Health. 2019;63:151–153. doi: 10.4103/ijph.IJPH_218_18.

[10]. V an der Heijden A.A., Abramoff M.D., Verbraak F., van Hecke M.V., Liem A., Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmol. 2017;96(1):63–68. doi: 10.1111/aos.13613.

[11]. Alexander C.A., Wang L. Big data analytics in heart attack prediction. J Nurs Care. 2017;6(2) doi: 10.4172/21671168.1000393.

[12]. Shafenoor Amin M., Kia Chiam Y., Dewi Varathan K. Identification of significant features and data mining techniques in predicting heart disease. Telematics Inf. 2018 doi: 10.1016/j.tele.2018.11.007.

[13]. Kausar N., Abdullah A., Samir B., Palaniappan S., AlGhamdi B., Dey N. Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. J Med Imag Health Inform. 2016;6(1):78–87.

