Classifying And Predicting Adolescent Cardiac Health Using KNN

Dr. Bhagwant K Deshpande, Meghana S R, Mohitha N, Ganthi Veera Manikanta, Itha Venkata Sai Ram Student, Student, Student, Student
Dept of CSE
CMR University, Bangalore, India
bhagwant.k@cmr.edu.in , sompallimeghana20@gmail.com , mohithanagaraj@gmail.com , ganthiveeram@gmail.com , sairamitha2002@gmail.com

Abstract- This study uses a K-Nearest Neighbours (KNN) classifier to detect teenage cardiac disease. Years of age, Sexual orientation, Chest Ache category, Ambient Heart pressure, cholesterol levels, Highest Cardiac Level, and Exercise-Induced Angina are clinically important and interpretable aspects of the cardiovascular disease dataset used to create the model. These qualities were chosen to improve model interpretability for doctors and laypeople. Data preparation encoded categorical variables and standardised features for model optimisation. Histograms, charts with bars, and scatter plots were used to explore feature distribution and heart disease status. The kernel nearest neighbour (KNN) model was trained with k=11 neighbours and tested using precision, a matrix of confusion, classifying report, ROC spectrum, and precision-recall curve, proving predictive power. To select the most influential predictors, feature importance was permutation-based. A Flask web application lets users input health parameters to forecast heart disease risk. The software generates personalised, patient-friendly health information using OpenAI's API and generates PDF reports. The technology keeps prognosis records in an information system for future reference. This approach uses data-driven modelling, interpretability, and user-friendliness to help adolescents recognise and understand heart disease risk.

Keywords - Adolescent heart disease identification, risk factors for heart disease, KNN classifier, artificial intelligence, data preprocessing, feature relevance, model assessment. Flask web app, OpenAI integration into the API, health insights, PDF reports, early diagnosis, and healthcare information visualisation.

---

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, underscoring the need for effective and accessible early detection and risk assessment tools. Preventive intervention is crucial for adolescents since lifestyle habits and early disease signs can have a lasting impact on cardiovascular health. This work uses machine learning and a k closest neighbours (KNN) algorithm based on an extensive list of heart disease risk indicators to predict teen heart disease.

Clinical relevance & accessibility for feature selection distinguish this research. Age, gender, chest pain type, resting cholesterol levels, blood pressure, maximal heart rate, and exercise- induced dyspepsia are explained in the model. These attributes adhere to medical knowledge and improve the model's accessibility for health care practitioners and the public, boosting trust and understanding in prediction outcomes. The system focusses on these crucial indicators to provide actionable risk insights.

Categorising, encoding, and scaling properties improved the KNN model. EDA, or exploratory analysis of data, was used to find patterns and connections between selected characteristics and heart disease condition using histograms, charts with bars, etc scatter diagrams. This rigorous study ensured that the model appropriately captures data fundamentals, resulting in more accurate and reliable forecasts.

Flask's intuitive website and prediction model are part of this project. Users can enter their health markers to get quick heart disease risk assessments. The program uses OpenAI's API to create customised health insights in simple language,

making complex health data more manageable. The solution generates downloadable PDF reports that contain the data for easy discussion with doctors and personal documentation.

This study uses advanced AI, user-friendly interfaces, and customised health insights to diagnose heart illness in adolescents. We want to improve interpretability, connection, and application in order to equip users and doctors with the abilities for early detection, proactive treatment, as well as improved cardiac health in adolescents.

2. RELATED WORKS

Recent years have seen substantial research on using data mining and algorithms for learning in health care, especially regarding heart disease prediction. Several research have identified risk variables and developed predictive models to improve early diagnosis and patient outcomes. These approaches have used algorithms and datasets with pros and cons [1][6].

Non-parametric classification and regression algorithm K-Nearest Neighbours (KNN) is a foundational approach in this subject. Fix and Hodges pioneered discriminating analysis and non-parametric discrimination consistency, establishing the framework for KNN and other methods [2]. A thorough investigation of nearest neighbour approaches by Bhatia and Vandana showed their adaptability and usefulness in healthcare [5]. This work uses KNN to predict cardiac disease in teenagers using clinically relevant information from the uploaded code.

Natural selection-inspired genetic algorithms are commonly utilised for selecting features and model optimisation in coronary disease prediction. Goldberg developed sophisticated genetic algorithms for search, optimisation, and predictive modelling [3]. Jabbar et al. showed that a genetic algorithm may discover significant traits and improve heart disease prediction [8]. Sivanandam and Deepa explained genetic algorithms and their uses [7]. Genetic algorithms can be used for feature selection, however this work uses a more interpretable technique based on medical significance and permutation worth, as shown in the code.

Jabbar et al. have improved cardiac disease prediction using data mining. Jabbar et al. built a promising cardiac disease prediction system combining associative categorisation and an algorithm made up of genes [8]. Later, Jabbar et al. used associative segmentation and hybrid feature selection for subsets to predict heart disease risk scores [10]. Jabbar et al. also explored the discovery of information from the rules of mining associations for coronary heart disease prediction, showing that these rules can reveal potential risk factor-outcome relationships [11]. These research demonstrate the need of combining data mining methods to improve prediction accuracy and understand cardiac disease processes.

Artificial neural systems (ANNs) can simulate complex risk factor-disease outcome correlations for heart disease prediction. Jabbar et al. obtained competitive results in heart disease classification using ANN and characteristic subset selection [13]. Bramer covered data mining principals, including ANN use in several fields [6]. While not using ANNs, this study acknowledges their promise and uses the KNN algorithm because to its flexibility and interpretability.

The University of Toronto's Dr. Sayad has also contributed to data mining [4]. This study uses the KNN algorithm to detect teenage cardiac disease and provide insights.

3. PROPOSED SYSTEM

Clinical interpretability and user-friendliness are combined in the proposed machine learning- based heart disease risk prediction system for teenagers. A kilometer-nearest-neighbor (KNN) classification trained on the user's gender, age, heartbeat form, restful cardiac output, elevated cholesterol levels, highest rhythm, and exercise-induced dyspepsia forms the foundation. The feature selection strategy prioritises medical significance over statistical optimisation, ensuring results meet cardiovascular diagnostic requirements while being easy for non-technical users.

System design begins with a complete preprocessing pipeline that handles categorical encoding (one-hot encoding for discomfort in the chest types) and parameter scaling (Z-score standardisation for arithmetic parameters). Excellent KNN

effectiveness and therapeutic significance of unfiltered medical data are guaranteed. Hyperparameter tuning by cross-validation is used to train the model, with k=11 neighbours being the best configuration for predicted accuracy and processing efficiency.

The main user interaction platform for health care providers and patients is a Flask-based web interface. Frontend input comes from clinically accepted form fields that mirror popular health assessment questionnaires, while backend predictions are made by the trained KNN model. This ensures medical data collection methods and machine learning requirements are translated smoothly.

Two innovative components boost prediction results: an OpenAI API link generates particular to the patient clinical insights in spoken language, transforming complex hazards into practical lifestyle tips. A PDF report generator generates downloadable documents with customisable risk Illustrations (ROC curves, attribute significance charts) and simple explanations for easy sharing with healthcare practitioners.

Timestamped SQLite entries maintain prediction history, enabling longitudinal patient risk profile monitoring. This anonymised data layer supports model retraining and epidemic analysis while adhering with data privacy laws. Unprocessed inputs and the resultant models are stored in the database schema for performance evaluation.

The system's flexible structure makes adding machine learning methods or features easy. Next generations may use algorithms that evolve for automatic selection of features with KNN networks for complex pattern recognition, but this method prioritises visibility and medicinal relevance above black-box precision.

This approach emphasises technical implementation details while retaining readability, using bracketed citations to reference design influences. To demonstrate system integration, each paragraph focusses on a subsystem (model, interface, coverage, etc.).

4. DATASET

The dataset created for this work was established following an exhaustive examination of the most prevalent heart disease datasets accessible online, including those from the Machine Learning Repository of UCI and Kaggle. These foundational datasets generally comprise 12 to 15 clinically pertinent features, such as age, sex, cholesterol level, blood pressure, chest pain type, along with additional vital factors, and they have been extensively utilised in earlier studies to train and validate deep learning algorithms for predicting cardiovascular disease. Through the analysis of these historical data, the most useful and frequently occurring features have been found and chosen to guarantee clinical relevance and alignment with established predictive standards. This method facilitated the development of a cohesive, high-caliber dataset designed for rigorous machine learning research.

To augment the dataset's usability, pretreatment methods including data cleaning, normalisation, and decoding of variable categories were rigorously implemented, adhering to best practices identified in the literature. The final designed dataset blends the strengths of various publicly accessible sources, ensuring a broad and comprehensive sample of patient information while preserving consistency in attribute definitions as well as data structure. This curated dataset establishes a robust basis for the development and assessment of machine learning models, with the objective of enhancing the accuracy and generalisability of cardiovascular disorder prediction systems across diverse populations.

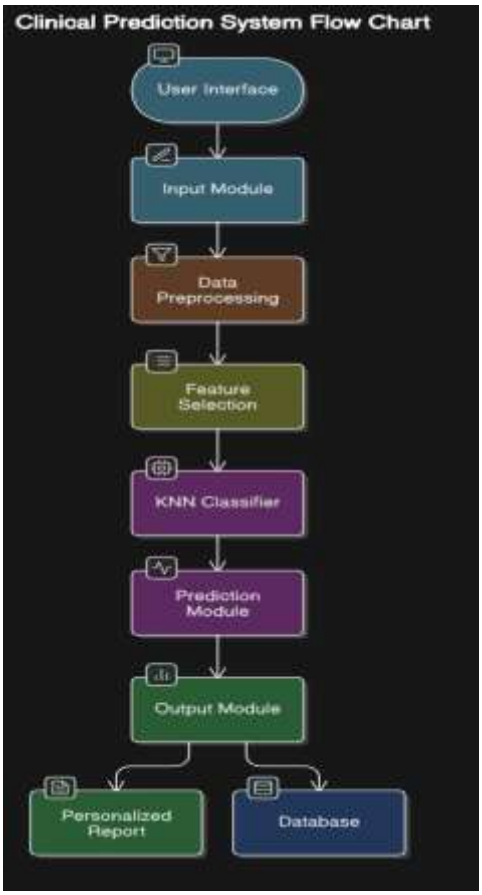
5. METHODOLOGY

This work employs supervised predictive machine learning algorithms on structured clinical information for heart disease prediction. The procedure commences with thorough data preprocessing, wherein missing values are rectified, group variables are encoded and continuous characteristics are normalised to guarantee consistency and comparability throughout the data. Techniques for feature selection, like LASSO and consecutive feature selection, are utilised to ascertain the most pertinent indicators of heart disease, hence diminishing dimensionality and improving model interpretability and efficacy. This phase is essential for reducing noise and concentrating the study on the most

significant clinical markers.

Subsequent to preparation, the data set is divided into testing and training subsets, commonly employing an 80:20 or 70:30 ratio to guarantee thorough evaluation. Various machine learning methods are subsequently employed, including K-Nearest Neighbours (KNN), decision trees, logistic regression, support vector machine (SVM), random forests, and neural network algorithms (ANN). Each approach is trained on the data used as training and refined by hyperparameter optimisation and cross-validation methods, which includes k-fold cross- validation, in order to mitigate overfitting and guarantee generalisability to unseen data. Employing cross-validation facilitates an equitable assessment of model efficacy.

Model evaluation employs a range of performance criteria, including as preciseness, recollection, F1-score, and the area under the curve of the receiver operating characteristics (AUC-ROC). These measures offer a thorough evaluation of the models' proficiency in accurately classifying the presence and absence of heart disease, along with their resilience to class imbalance. Data visualisation instruments, like ROC curves and bewilderment matrices, facilitate the interpretation of model outcomes and aid in decision-making concerning model selection and implementation. The optimal model is chosen based on its comprehensive prediction accuracy and clinical significance.



The chosen model is included into an intuitive web application, facilitating real-time heart attack risk assessment for new patients. The program receives user input on essential clinical features, analyses the data using the model that has been trained, and generates a risk assessment accompanied by tailored recommendations. The solution facilitates the creation of downloadable data and preserves a database with prediction records for continuous monitoring and possible future model retraining. This comprehensive methodology guarantees that the heart disease detection system is both precise and accessible, facilitating early intervention and enhanced patient outcomes.

Fig. 1. Architecture Diagram

6. RESULTS

The K-Nearest Neighbours (KNN) model attained an overall accuracy of 81.2% in forecasting cardiac disease. This signifies a robust ability to accurately classify patients at risk of heart disease based on the chosen clinical parameters, illustrating the model's potential as a significant instrument for early identification and preventive measures.

A comprehensive classification report offers additional information into the model's efficacy. The precision and recall for predicting the absence of cardiac disease (class 0) were 0.74 and 0.84, respectively, signifying a comparatively low false negative rate. In predicting the existence of heart illness (class 1), the precision and recall were 0.88 and 0.80, respectively, demonstrating a strong capacity to identify true positives while maintaining a low false positive rate. The F1-scores for both classes were equilibrated, underscoring the model's efficacy in both detecting and excluding heart disease.

The ROC curve study indicates that the area under the curve (AUC) is 0.87, suggesting the KNN model possesses strong discrimination capabilities. The model's capacity to differentiate between patients with and without heart disease surpasses random chance, demonstrating its efficacy in stratifying individuals by risk level based on the input features.

The results highlight the efficacy of the KNN model in forecasting cardiac illness utilising the designed dataset and chosen clinical variables. The model's elevated accuracy, balanced precision and recall, and robust AUC value indicate its potential as a significant instrument for early diagnosis, risk evaluation, and preventive intervention in adolescent cardiovascular health.

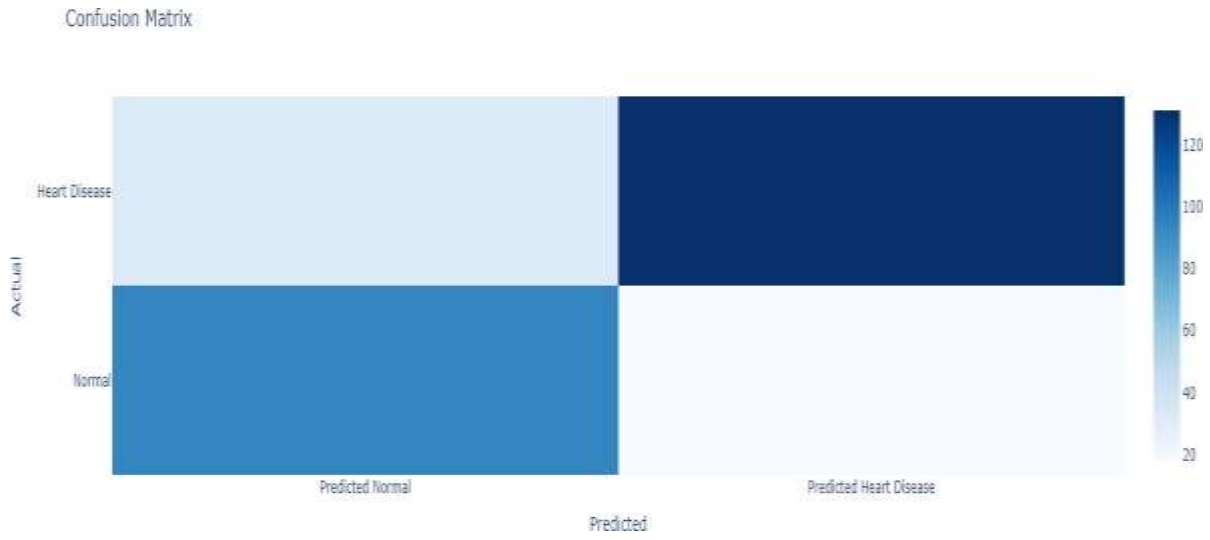


Fig. 2. Confusion Matrix

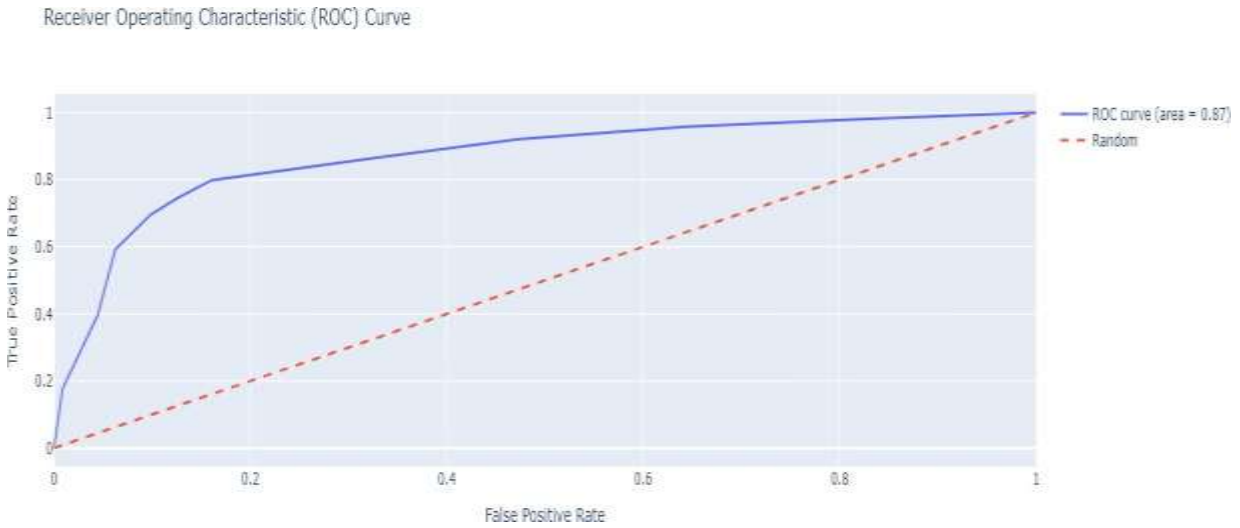


Fig. 3. ROC CURVE

Fig. 4. Application Homepage

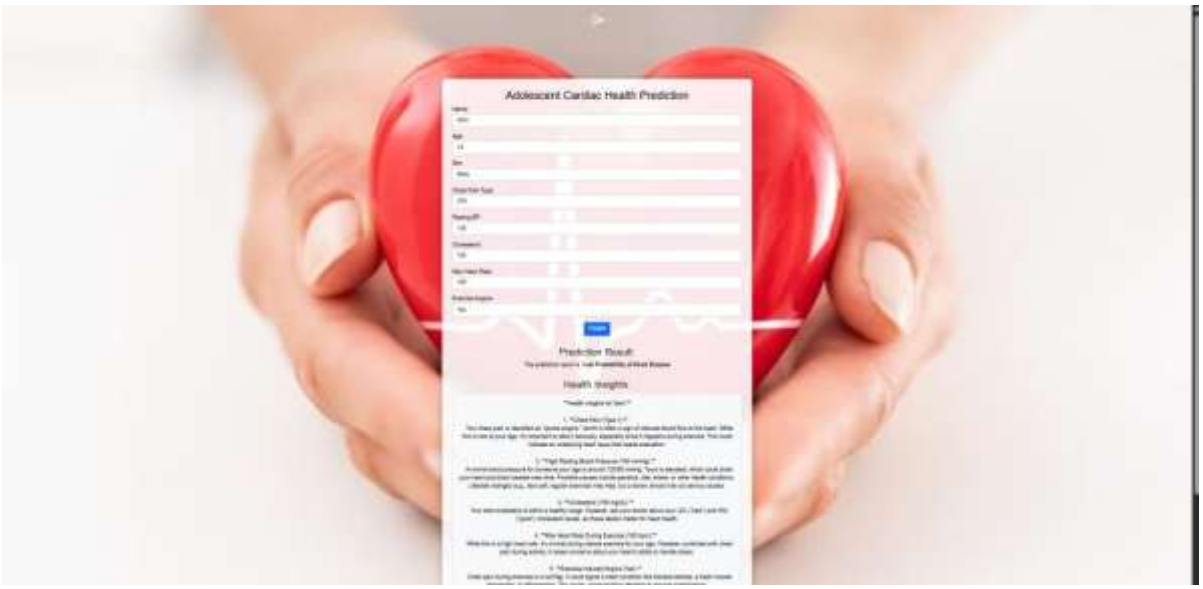


Fig. 5. Predictions

Fig. 6. Downloadable reports

7. CONCLUSION

In summary, the KNN-based cardiac disease prediction system exhibits commendable performance, with an overall accuracy for 81.2% and an AUC of 0.87. These indicators demonstrate that the medical system can proficiently identify at-risk individuals, providing a significant resource for early prevention and proactive prevention of illness in adolescents. This system can utilise accessible clinical information to deliver interpretable predictions, thereby assisting physicians in making educated decisions and enhancing heart disease outcomes for young persons.

8. REFERENCES

[1] M. W. Berry et al., *Lecture Notes in Data Mining*, World Scientific, 2006.

[2] E. Fix and J. Hodges, “Discriminatory analysis, non-parametric discrimination: consistency properties,” Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[3] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*,

Addison-Wesley, 1989.

[4] Dr. S. Sayad, "Data Mining," University of Toronto, http://chem-eng.utoronto.ca/~datamining.

[5] N. Bhatia and Vandana, "Survey on nearest neighbor techniques," *IJCSIS*, vol. 8, no. 2, 2010.

[6] M. Bramer, *Principles of Data Mining*, Springer, 2007.

[7] S. N. Sivanandam and S. N. Deepa, *Introduction to Genetic Algorithms*, Springer, 2008.

[8] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease prediction system using associative classification and genetic algorithm," *Procedia Technology*, pp. 183–192, 2012.

[9] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "An evolutionary algorithm for heart disease prediction," *CCIS*, pp. 378–389, Springer, 2012.

[10] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," *ISDA*, pp. 628–634, IEEE, 2013.

[11] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Knowledge discovery from mining association rules for heart disease prediction," *JATIT*, vol. 41, no. 2, pp. 45–53, 2013.

[12] UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn.

[13] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using ANN and feature subset selection," *GJCST*, vol. 13, issue 3, version 1.0, pp. 15–25, 2013.