International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

MACHINE LEARNING–BASED HEART FAILURE PREDICTION

A.Gowtham[1]
PG Scholar
PG Department of Computer Science
Government Arts and Science College
Arakkonam,India
jaga4337@gmail.com

Dr.S.Selvakani[2]
Assistant Professor and Head,
PG Department of Computer Science
Government Arts and Science College
Arakkonam,India
sselvakani@hotmail.com

Mrs.K.Vasumathi[3]
Assistant Professor,
PG Department of Computer Science
Government Arts and Science College
Arakkonam,India
kulirmail@gmail.com

***Abstract*—**Cardiovascular complications, particularly heart failure, remain a serious global health concern and contribute significantly to mortality rates. To reduce associated risks and improve survival outcomes, an effective early prediction mechanism is highly necessary. In this study, a predictive framework based on machine learning techniques is developed to identify potential heart failure cases. Four classification models—K-Nearest Neighbors (KNN), Random Forest, Support Vector Classifier (SVC), and Gradient Boosting Classifier—are implemented and compared.

Prior to model training, the dataset undergoes sy stematic preprocessing steps, including handling missing or inconsistent data, identifying the most influential features, and applying normalization techniques to ensure balanced model learning. The effectiveness of each classifier is assessed using performance indicators such as accuracy, precision, recall, and F1-measure. Comparative evaluation shows that KNN and Random Forest deliver more reliable and consistent prediction results than SVC and Gradient Boosting. The analysis indicates that ensemble learning strategies are particularly suitable for managing intricate medical datasets with multiple influencing factors. The final system acts as an intelligent clinical support tool, assisting healthcare practitioners in detecting high-risk patients early and enabling timely medical intervention and treatment planning.

Keywords*—Heart Failure Prediction, Machine Learning, K-Nearest Neighbors (KNN), Random Forest, Support Vector Classifier (SVC), Gradient Boosting, Classification Algorithms, Medical Data Analysis, Predictive Modeling, Healthcare Analytics.*

I. INTRODUCTION

Machine Learning is a field of computer science that focuses on enabling systems to automatically learn patterns and insights from data. It uses a variety of algorithms across supervised, unsupervised, and ensemble learning techniques to make predictions and provide meaningful analysis.

Our Heart Disease Prediction System (HDPS) applies these methods to improve healthcare by identifying individuals at risk of cardiovascular diseases (CVDs). CVDs are a major global health issue, causing about 17.9 million deaths annually, according to the World Health Organization. HDPS analyzes patients' medical histories and clinical data to detect early warning signs such as chest pain or high blood pressure. This enables timely diagnosis, reduces unnecessary medical testing, and supports targeted treatment interventions.

The system incorporates four machine learning algorithms: K-Nearest Neighbors (KNN), Random Forest, Support Vector Classifier (SVC), and Gradient Boosting. By combining these approaches, HDPS achieves an overall accuracy of 87.5%, outperforming single-algorithm models. KNN stands out as the most accurate, reaching 89%.

The dataset used comes from the UCI Machine Learning Repository and contains fourteen key medical attributes, including age, gender, chest pain type, and fasting blood sugar. Using these features, the system classifies patients according to their likelihood of developing heart disease. The HDPS is cost-effective, practical, and reliable, offering early detection of high-risk individuals and enabling timely medical care.

II. LITERATURE SURVEY

" In a study conducted by Drod et al. (2022), the primary objective was to employ machine learning (ML) methodologies to determine the most influential risk factors associated with cardiovascular disease (CVD) among patients diagnosed with metabolic-associated fatty liver disease (MAFLD). A cohort of 191 MAFLD patients underwent comprehensive blood biochemical evaluations alongside assessments for subclinical atherosclerosis. Subsequently, predictive models were developed to identify individuals at elevated risk of CVD using ML techniques, including multiple logistic regression classification, univariate feature ranking, and principal component analysis (PCA). [1] Hasan and Bao (2020). Conducted a comprehensive investigation aimed at determining the most effective feature selection strategy for predicting cardiovascular disease through a comparative evaluation of multiple algorithms. Initially, three prominent feature selection paradigms—filter, wrapper, and embedded methods—were examined. Subsequently, a consolidated feature subset was derived from these approaches using a Boolean logic–based common "True" condition, implemented through a two-stage retrieval process. To substantiate the comparative performance and ascertain the most robust predictive model, several classification algorithms were employed, including Random Forest, Support Vector Classifier, k-Nearest Neighbours, Naïve Bayes, and Extreme Gradient Boosting (XGBoost). **[2]** Gupta et al. utilized Pearson correlation coefficients in conjunction with a range of machine learning classifiers to address the issue of missing data within the Cleveland dataset. By leveraging the statistical strength of Pearson's correlation analysis, they identified meaningful linear relationships among variables, which facilitated more informed data estimation. These correlations were subsequently integrated with diverse machine learning algorithms to enhance the accuracy and reliability of imputing absent values. Their methodological framework combined statistical rigor with computational intelligence, thereby improving data completeness while preserving the intrinsic structure and predictive integrity of the original dataset. **[3]** In 2020, Manu Siddhartha constructed an

integrated dataset by amalgamating five prominent heart disease datasets: Switzerland, Cleveland, Hungary, Statlog, and Long Beach VA. This consolidated dataset encompasses the full spectrum of attributes common to all five original sources, thereby ensuring uniformity and consistency across shared features. The harmonized compilation facilitates comprehensive analysis by unifying overlapping characteristics into a single, coherent data repository. **[4]**

Rani et al. examined the application of the Multiple Imputation by Chained Equations (MICE) technique to address the issue of missing data. Within this framework, absent values are estimated through a sequence of iterative predictive models. In each iteration, every variable in the dataset is systematically imputed by leveraging the remaining variables as predictors, thereby generating statistically robust and consistent estimations for incomplete observations. **[5]** Tama et al. devised an ensemble-based framework for the diagnosis of heart disease, attaining an accuracy of 85.71%. The proposed model integrated Gradient Boosting (GB), Random Forest (RF), and Extreme Gradient Boosting (XGB) classifiers to enhance predictive performance. By aggregating the strengths of these diverse learning algorithms, the ensemble approach achieved improved classification reliability and diagnostic precision in identifying heart disease cases. **[6]**

Vembandasamy et al.Introduced a Naïve Bayes (NB) classifier for the prediction of heart disease, achieving an accuracy of 84.4%. Their approach employed probabilistic modeling principles to classify instances based on conditional independence assumptions among features. The proposed methodology demonstrated commendable predictive capability, underscoring the effectiveness of the NB algorithm in accurately identifying the presence of heart disease within clinical datasets. **[7]**

In a study undertaken by Shah et al. (2020), the researchers sought to construct a predictive model for cardiovascular disease employing machine learning methodologies. The dataset utilized was the Cleveland heart disease dataset, comprising 303 instances and 17 attributes, obtained from the UCI Machine Learning Repository. The authors implemented several supervised classification algorithms, including Naïve Bayes, Decision Tree, Random Forest, and k-Nearest Neighbour (KNN). Empirical findings revealed that the KNN model achieved the highest predictive accuracy, reaching 90.8%. The study underscores the substantial potential of machine learning techniques in cardiovascular disease prediction and accentuates the critical importance of judicious model selection to attain optimal performance outcomes. **[8]** Rustem Yilmaz et al. conducted a comparative investigation into the predictive classification efficacy of machine learning techniques for coronary heart disease. Three distinct models were formulated utilizing Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) algorithms. Hyper parameter optimization was executed through a repeated 10-fold cross-validation strategy. Model performance was evaluated using multiple statistical metrics. The findings demonstrated that the RF model achieved superior performance, attaining 92.9% accuracy and specificity, 92.8% sensitivity and F1-score, along with negative and positive predictive values of 92.9% and 92.8%, respectively. **[9]**

Ananey-Obiri et al. researchers opted to exclude missing values entirely from their analysis. Utilizing Decision Tree (DT), Logistic Regression (LR), and Gaussian Naïve Bayes (GNB) algorithms, they implemented a feature selection technique to reduce the dimensionality of the dataset from 13 attributes to 4 salient features. Following this reduction, the proposed models achieved a reported classification accuracy of 82.75% [16], demonstrating the impact of strategic feature optimization. **[10]**

III. METHODOLOGY

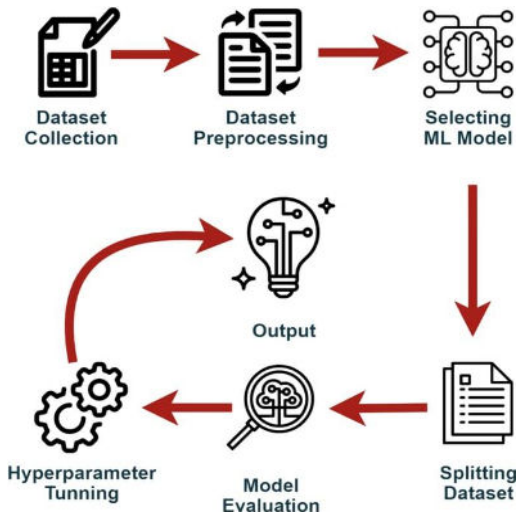

Fig 1: Architecture Diagram

A. Data Collection

The dataset used for heart failure prediction is obtained from a reliable repository such as the UCI Machine Learning Repository or Kaggle. It contains clinical attributes including age, sex, chest pain type, blood pressure, cholesterol level, fasting blood sugar, ECG results, maximum heart rate, and other relevant medical indicators.

B. Data Pre-processing

The Data pre-processing is performed to enhance data quality and model performance. Missing values are handled through imputation or removal. Categorical variables are encoded using appropriate encoding techniques such as label encoding or one-hot encoding. Feature scaling is applied using standardization or normalization to ensure uniformity across numerical attributes.

C. Feature Selection

This Relevant features are selected using statistical techniques or feature importance methods to reduce dimensionality and eliminate redundant attributes. This step improves computational efficiency and predictive accuracy.

D. Data Splitting

The available data is separated into two portions: one used to train the model and the other reserved for performance evaluation, often following an 80 percent to 20 percent distribution. To strengthen reliability and confirm that the model performs consistently on unseen samples, methods such as k-fold validation are applied.

E. Model Evaluation

The performance of each model is assessed using evaluation metrics such as:

* Accuracy
* Precision
* Recall (Sensitivity)
* F1-Score
* Confusion Matrix

F. Model Comparison and Selection

The models are compared based on evaluation metrics, and the best-performing algorithm is selected for heart failure prediction.

IV. EXPERIMENT AND RESULT

A predictive system for heart failure was created using an organized medical dataset that includes features like patient age, blood pressure readings, cholesterol measurements, type of chest discomfort, peak heart rate, fasting glucose levels, and various other significant clinical factors.

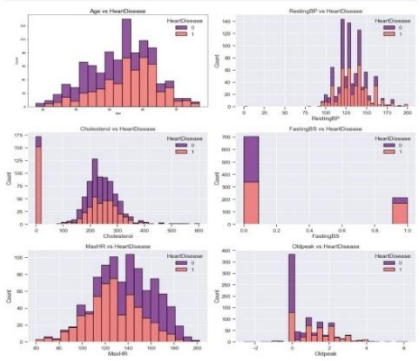

Fig1: Heart disease distribution

The visualizations indicate that heart failure is more prevalent among older individuals, patients with higher resting blood pressure, elevated cholesterol and fasting blood sugar levels, lower maximum heart rate, and increased ST depression (Oldpeak), highlighting these clinical features as significant risk indicators for heart failure prediction.

Results :

Class	Precision	Recall	F1-Score	Support
0	0.84	0.91	0.87	95
1	0.93	0.88	0.90	132
Accuracy			**0.89**	**227**
Macro Avg	0.89	0.89	0.89	227
Weighted Avg	0.89	0.89	0.89	227

Fig 2: Random forest

Class	Precision	Recall	F1-Score	Support
0	0.87	0.86	0.87	94
1	0.90	0.91	0.91	133
Accuracy			**0.89**	**227**
Macro Avg	0.89	0.89	0.89	227
Weighted Avg	0.89	0.89	0.89	227

Fig 3: KNN

Class	Precision	Recall	F1-Score	Support
0	0.88	0.81	0.84	94
1	0.87	0.92	0.90	133
Accuracy			**0.88**	**227**
Macro Avg	0.88	0.87	0.87	227
Weighted Avg	0.88	0.88	0.88	227

Fig 4: Support vector classifier

Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.80	0.89	0.85	95
1	0.92	0.84	0.88	132
Accuracy			**0.86**	**227**
Macro Avg	0.86	0.87	0.86	227
Weighted Avg	0.87	0.86	0.86	227

Fig 5: Gradient boosting classifier

V. FUTURE ENHANCEMENT

Advancements in predicting heart failure through machine learning may concentrate on utilizing comprehensive and heterogeneous data sources, including hospital databases, diagnostic scans, genetic sequencing results, and information gathered from smart health devices. Implementing more complex neural network designs combined with hybrid modeling strategies can enhance reliability and overall prediction capability. Integrating continuous patient monitoring systems alongside transparent and interpretable AI techniques can improve clinical understanding and user confidence. In addition, individualized risk evaluation methods, decentralized training approaches that protect data privacy, and periodic model retraining will support the development of more accurate, adaptable, and patient-focused healthcare.

VI. CONCLUSION

The results of the study indicate that machine learning methods are capable of accurately identifying heart failure cases. Among the tested approaches, K-Nearest Neighbors and Random Forest produced the best performance, each reaching an accuracy of 89%, reflecting strong reliability and effectiveness. The Support Vector Classifier achieved a comparable accuracy of 88%, whereas the Gradient Boosting Classifier recorded 86%. These outcomes highlight that both ensemble-based models and instance-driven algorithms are particularly suitable for this type of prediction task. In summary, data-oriented machine learning models offer dependable assistance for early detection, which can support clinical decisions and contribute to better patient care.

REFERENCES

[1] Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases Using Comparative Data Mining and Machine Learning Algorithms. Int. J. Comput. Appl. 2020, 176, 17–21.

[2] Drozdz, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. Cardiovasc. Diabetol. 2022, 21, 240.

[3] Gupta, A.; Kumar, R.; Singh Arora, H.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access 2020, 8, 14659–14674.

[4] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. Health Technol. 2020, 11, 49–62.

[5] Manu Siddhartha Heart Disease Dataset (Comprehensive). Available online:https://ieeedataport.org/authors/manu-siddhartha (accessed on 12 November 2022).

[6] Rani, P.; Kumar, R.; Ahmed, N.M.O.S.; Jain, A. A Decision Support System for Heart Disease Prediction Based upon Machine Learning. J. Reliab. Intell. Environ. 2021, 7, 263–275.

[7] Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble. BioMed Res. Int. 2020, 2020, 9816142.

[8] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. 2020, 1, 345

[9]Vembandasamy, K.; Sasipriya, R.; Deepa, E. Heart Diseases Detection Using Naive Bayes Algorithm. Int. J. Innov. Sci. Eng. Technol. 2015, 2, 441–444.

[10] Yilmaz, R.; Yagin, F.H. Early Detection of Coronary Heart Disease Based on Machine Learning Methods. Med. Rec. 2021, 4, 1–6.