

Heart failure prediction: a comparative analysis of machine learning algorithms

Fida Husain Dahri^a, Asif Ali Laghari^{*b}, Dileep Kumar Sajnani^a, Asima Shazia^a, Teerath Kumar^c

^aSchool of Computer Science and Engineering, Southeast University, Nanjing 211100, Jiangsu, China; ^bSoftware College, Shenyang Normal University, Shenyang 110034, Liaoning, China;

^cNational University of Computer and Emerging Sciences, Islamabad, Pakistan

ABSTRACT

In recent, machine learning techniques have been employed to predict various diseases such as lung, blood, liver, heart, etc. As the heart is considered one of the human body's significant organs, this research proposes a comparative analysis of heart disease failure using various machine learning algorithms. Since the ECG and EEG are primary sources for analyzing heart performance, the lipid profile also provides information to determine good and bad cholesterol levels. Although this research proposes a comparative analysis of Heart Failure (HF) and other Cardiovascular diseases (CVDs), it assists in finding a better way to predict the root cause of HF and CVDs. The machine learning models are employed and trained, such as the Support Vector Classifier (SVC), Random Forest Classifier (RF), Logistic Regression (LR), Decision Tree Classifier (DT), and K-Nearest Neighbors Classifier (KNN) on the real-time dataset from Kaggle to predict and classify heart disease patients. The experimental set-up outcomes show that the Logistic Regression (LR) classifier has proven the best accuracy (88.00%) among all other machine learning classifiers. Our research results significantly contribute to predicting (HF) and nurturing advancements in AI-powered tools for improved heart failure (HF) prediction and patient care.

Keywords: Heart failure, machine learning, heart disease, classification, cardiovascular disease (CVD)

1. INTRODUCTION

In today's fast-moving world, people have become so busy with their daily work that they neglect their well-being. Therefore, the commonness of obesity, hypertension, anxiety, depression, and other forms of stress has become broadly observed as a consequence of such a distressing way of life. The factors mentioned above are the primary reasons for disorders in the body that ultimately result in the development of several kinds of cancer cells, diabetes, heart diseases, etc. Although numerous diseases cause fatalities annually, it is heart diseases or cardiovascular diseases (CVDs) that hold the biggest percentage of deaths in the healthcare sector worldwide¹. Around 17.9 million people died in 2019 due to (CVDs), as reported by (WHO), which is around 32% of the world's total death rate in 2019, of which 85% were only linked to heart attack and stroke, from the total number². CVDs are distinguished by higher rates of disability and death rates, making them a significant public health concern. Global Burden of Disease (GBD 2019) revealed that the weight of (CVD) diseases worldwide almost doubled between 1990 and 2019. From 271 million to 523 million people were affected by (CVDs) at this growth rate³. Further, the Global Burden of Disease (2023-GBD) revealed that from 12.4 million in 1990 to 19.8 million global deaths were caused by (CVDs), Primarily due to the aging and the growth of the worldwide population, together with the influence of preventable metabolic, environmental, and behavioral risk factors⁴.

The diagnosis of Heart Failure (CVD) is a time-consuming and complex procedure due to the requirement of high medical competence, which demands a wide medical test, medical experts competence, history of patients, and observing the patient's lifestyle⁵. In growing countries, there is a major shortage of competent medical experts, mainly in the cardiologist field. Due to this, few people get proper treatment and diagnoses of cardiac disease. In remote places, the treatment chances of a patient with (CVD) disease receiving from the appropriate healthcare expert to reduce the suffering is minimal⁶. So, this fast-growing rate of cardiovascular disease can be reduced by early detection, making an intelligent automatic system help to diagnose the patient in the early stage⁷. There are various primary methods to detect

*asiflaghari@synu.edu.cn

heart disease, such as echocardiogram (echo), electrocardiogram (ECG), cardiac MRI, computed tomography, blood tests, and more⁸.

Nevertheless, these procedures might result in inaccurate results and be exceedingly time-consuming. Many researchers have successfully employed deep learning and machine learning techniques to analyze medical data, accurately predicting cardiovascular disease risk and symptoms^{9,10}. As demonstrated in recent research, ML techniques have received significant attention in the research community¹¹. ML approaches can achieve higher accuracy than previous data classification methods¹². Researchers now implement several popular machine learning algorithms, including Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine, to predict occurrences of heart disease^{13,14}. The following given below are the main contributions of our research work:

- First, we address the dataset issue and then refine and standardize the datasets to train and test classifiers.
- Second, our study compares and evaluates the effectiveness of machine learning classifiers (LR, SVC, DT, RF, and KNN) to find out which classifiers provide the most accurate results.
- Last, the proposed classifier (LR) gives state-of-the-art accuracy, which ultimately will ease the medical experts by finding Cardiovascular diseases (CVDs) and improve (HF) rapid treatment.

2. RELATED WORK

In developing countries, Heart disease (CVDs) remains a major cause of death globally, highlighting the importance of research that aims to uncover risk factors and early indications of the condition. In the following part, we present the related demand and the importance of our research through current studies. Bhatt et al.¹⁵ proposed a k-mode clustering technique in their research after testing different models, and they optimized the parameters using Grid SearchCV, a dataset from Kaggle, which consisted of 70000 instances, was used in the model. Cross-validated MLP performed well among all other algorithms with an accuracy of 87.28%. An accuracy of 87.10% was achieved by Ambrish et al.¹⁶ in their study using an LR classifier to classify cardiovascular disease. They used data from 303 records and 13 features from UCI Cleveland. Authors in¹⁷ achieved an accuracy of 85.25% by LR for early detection of CVD among 6 other models that were used: NB, SVM, DT, RF, & KNN.

Berke et al.¹⁸ used 10-fold cross-validation in their model, where the highest accuracy, 89%, is achieved by XGB for non-outlier & and 84.6% for outlier data, while KNN achieved 85.6% accuracy for non-outlier and 81%. KNN achieved an accuracy of 88.52% in the research conducted by Jindal et al.¹⁹; among KNN, two other classifiers, LR and RF, were used. The authors tested the dataset from the UCI repository with 14 attributes. Madhumita et al.²⁰ used their model on the UCI Cleveland dataset with 303 samples and 14 different attributes. Their model achieved an accuracy of 86.9% with a sensitivity rate of 90.6%. Besides this, an 82.7% specificity rate was achieved. Research conducted by Garg et al.²¹ used a dataset containing 303 samples with 14 attributes. After testing the model, the KNN method achieved 86.8% accuracy, while the RF algorithm achieved 81.9%. The authors of this work²² collected a dataset including 301 samples, each containing 12 clinical variables. The heart disease prediction involved logistic regression, decision tree, support vector machine (SVM), and Naive Bayes classification techniques. Significantly, logistic regression achieved an accuracy rate of 86.25%.

Amin et al. conducted a study evaluating different data mining approaches for predicting heart failure. They observed that the technique used with its highest classification accuracy was 87.4%²³. A dataset is based on a UCI machine learning repository comprising 303 records, including 14 distinct attributes. Seven data mining approaches were separately assessed on 8100 different feature combinations. Nabaouia et al.²⁴ utilized a range of machine learning techniques to identify cardiovascular disease (CVD), ultimately proposing the use of a Support Vector Machine (SVM) using a linear kernel approach. By utilizing 13 different features, they reached an accuracy rate of 86.8%. Mohan et al.²⁵ introduced a Hybrid Random Forest combined with a linear model, which demonstrated an accuracy of 88.7% in heart disease prediction; they used the UCI dataset for their study and applied binary classification and multi-class variable techniques for data pre-processing. Dwivedi et al. conducted a study where they evaluated 6 machine-learning methods for predicting heart diseases (CVD). The outcome showed that the LR had the highest accuracy of 85% on the dataset of 270 samples, each containing 13 features²⁶. The our research provide better results compare into terms of various measuring metrics compared to previous studies^{16,17,21-23}.

3. METHODOLOGY

This research methodology discusses the various methodology sub-steps, as shown in Figure 1. The proposed method includes steps, data collection and preprocessing, feature selection, Heart Failure (HF) prediction modeling using the collected dataset, evaluation of models, and results.

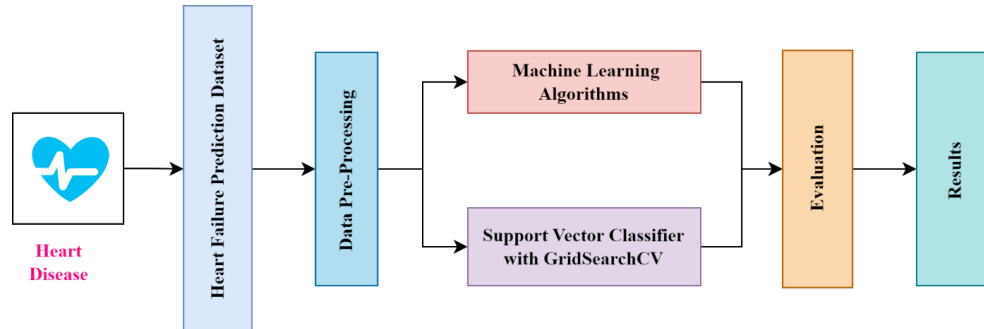


Figure 1. Flow diagram of proposed approach.

3.1 Data collection

The Heart Failure Prediction Dataset was gathered from Kaggle. The dataset contains 12 features with 1190 total observations, including different ages, where (44.70%) of observations have No Heart Diseases, and (55.30%) of observations have heart disease. This dataset combines five independent datasets to make it the largest heart disease dataset with 11 common features, and this dataset is available for research²⁷. The five individual datasets combined are 294 Hungarian observations dataset, 123 observations from the Switzerland dataset, 270 observations in the Stalog (Heart) Dataset, and 200 observations from the Long Beach VA dataset.

3.2 Data preprocessing

Data preparation and preprocessing are important for data mining or the machine learning approach to check how well the dataset is prepared, structured, and suitable for machine learning tasks because the efficiency of a machine learning approach depends on high-quality data input. This study applied multiple preprocessing steps to the selected dataset, including a replace missing values filter used to handle missing values, cleaning data and normalization depending on algorithms used, and standardization of numerical features using MinMaxScaler and StandardScaler, respectively. Then, the data is transformed into appropriate data types for more reliable and interpretable results and values. After that, the dataset was split into a training set (80%) and (20%) testing set. Multiple machine learning classifiers were used in this study to check the performance of various classifiers. However, the data type was converted into suitable formats for some attributes based on model specification. In the design of the experiment, mainly binary classification is used, which processes the categorizing of the dataset according to predefined classes²⁶.

3.3 Modelling

This study systematically investigates machine learning models for heart failure (HF) detection and classification. The dataset was collected and preprocessed for better performance of the classifiers. In this modeling section of the methodology, five machine learning classifiers, such as a Support Vector Classifier, Logistic Regression Classifier, K-nearest Neighbors Classifier, Decision Tree Classifier, and Random Forest Classifier, were applied to predict and classify heart failure (HF) as shown in Figure 2. All machine learning classifiers are trained on the training dataset, and the testing dataset is used to assess the classifiers' performances. The performance of each classifier is evaluated using the various evaluation metrics, including recall, accuracy, precision, F1 score, cross-validation score, and area under the receiver operating characteristic curve (AUC-ROC).

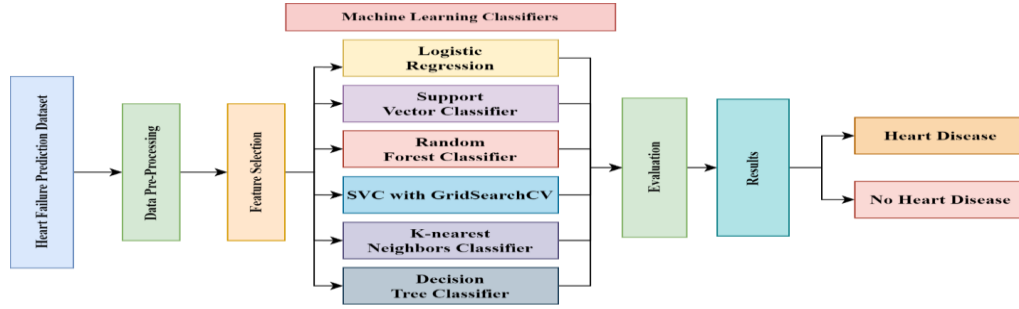


Figure 2. Model architecture diagram.

3.4 Performance evaluation metrics

This study evaluated the proposed approach's efficiency using state-of-the-art performance measures metrics, including recall, F-1 score, precision, accuracy, cross-validation score, and auc_roc score. The evaluation results of our research are shown in Table 1. The accuracy metric is the correctness or precision of the machine learning classifier. Mathematically, the representation of the accuracy metric is given by equation (1). The recall score analyzes the overall number of true or true positives that are artificial through the total figure of false negative cases. It is provided mathematically in equation (2). The precision metric predicts positive instances that are true positives. Mathematically shown in equation (3). An F-1 score metric is a harmonic mean of recall and precision. It takes the balance between recall and precision. It is given mathematically by equation (4). The AUC_ROC score displays the true positive rate (TPR) against the false positive rate (FPR). The confusion matrix is a 2×2 matrix used to measure the machine learning performance to evaluate the classification model's efficiency, in which predicted labels compare with the actual dataset labels. Possible four binary outcomes: TN, FP, FN, and (TP) mathematically, shown in equation (5).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$PRC = \frac{TP}{TP + FP} \quad (2)$$

$$RC = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \times \frac{PRC + RC}{PRC \times RC} \quad (4)$$

$$CM = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix} \quad (5)$$

Table 1. Evaluated results of classifiers.

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC score	Cross-validation score
Logistic regression	88.00%	0.88	0.87	0.87	87.43%	91.12%
Support vector classifier	87.50%	0.88	0.87	0.87	87.43%	90.53%
Random forest classifier	84.24%	0.85	0.84	0.84	84.06%	92.91%
K-nearest neighbors classifier	81.52%	0.82	0.81	0.81	81.36%	89.34%
Decision tree classifier	84.78%	0.85	0.85	0.85	84.62%	89.09%

3.5 Results & discussion

This research utilized the Heart Failure Prediction Dataset and the Kaggle platform for experiment set-up with a GPU P-100 computing processor with RAM 16 GB. The dataset was divided into two parts: 80% of training data was used to train the model, and 20% of test data was used to test the model. The algorithms used in this study were Logistic Regression, Random Forest Classifier, Support Vector Classifier, K-nearest Neighbors Classifier, and Decision Tree Classifier. This research included many performance metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve. The results of various machine learning classifiers are shown in Table 1. The results indicate that the Logistic Regression (LR) obtained the highest accuracy of 88.00%, along with high recall, precision, F1 score, AUC_ROC scores, and cross-validation scores of 87.00%, 88.00%, 87.00%, 87.43%, and 91.12%, respectively. All classifiers had an accuracy above 80.00%. The SVC accuracy is 87.00%, Precision score is 0.88, Recall score is 0.87, F1- Score is 87.43%, ROC-AUC SCORE is 87.43%, and Cross-Validation Score is 90.53%. The RF accuracy score is 84.00%, Precision 0.85, Recall 0.84, F1- Score 84.00%, ROC-AUC SCORE 84.06%, and Cross-Validation Score 92.91%. The KNN accuracy score is 81.52%, Precision 82.00, Recall 81.00, F1- Score 81.00%, ROC-AUC SCORE 81.36%, and Cross-Validation Score 89.34%. The DT accuracy score is 84.78%, Precision 85.00, Recall 85.00, F1- Score 85.00%, ROC-AUC SCORE 84.62%, and Cross-Validation Score 89.04%. We observe that Logistic Regression and SVC have predicted nearly the same accuracy. However, LR performed better than all the other classifiers in the accuracy metric. The confusion matrix results of each classifier are shown in Figures 3-7. All the classifiers were also compared using measuring metrics.

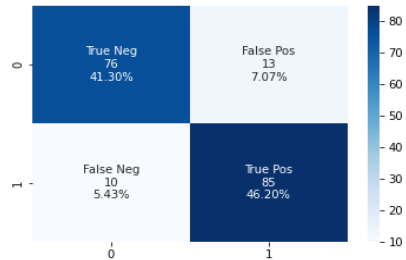


Figure 3. LR confusion matrix.

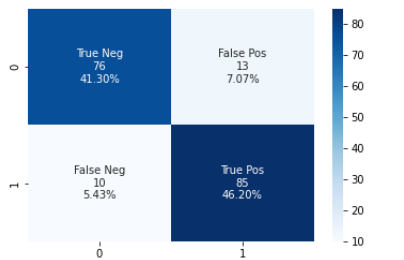


Figure 4. SVC confusion matrix.

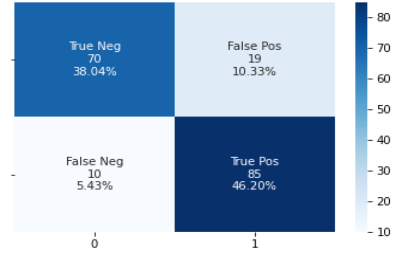


Figure 5. RF confusion matrix.

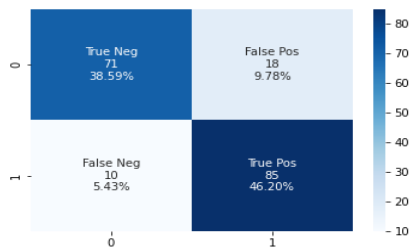


Figure 6. KNN confusion matrix.

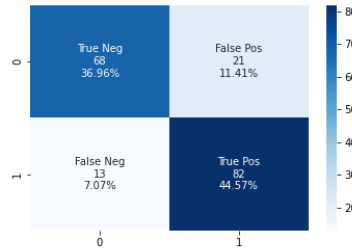


Figure 7. KNN confusion matrix.

4. CONCLUSION

The objective of this research study was to utilize various supervised machine learning algorithms to predict the CVDs and possibility of heart disease patients and classify them based on real-time data. The machine learning algorithms used to predict and classify heart disease were the Logistic Regression, Support Vector Classifier, Random Forest, Decision Tree Classifier, and K-nearest Neighbors Classifier. This study has shown promising results in terms of various measuring metrics compared to previous studies. The findings demonstrated the potential use and efficacy of the approach for physicians and cardiac surgeons in promptly assessing the chance of a heart attack in a patient. Logistic Regression (LR) has achieved the best accuracy score of 88.00% in predicting heart disease patients.

On the other hand, the SVC classifier has achieved an accuracy score of 87.50%, which is near to LR in predicting heart disease patients. Therefore, the LR classifier is more effective in managing medical datasets. This intelligent (HF) and (CVDs) prediction system using a machine learning classification algorithm may be used to predict or detect several

more diseases in the future. The study has the potential to be advanced or enhanced to automate the examination of cardiac disease, including several different machine-learning methods.

REFERENCES

- [1] Gaidai, O., Cao, Y. and Loginov, S., "Global cardiovascular diseases death rate prediction," *Curr. Probl. Cardiol.* 48, 101622 (2023).
- [2] Hussain, M. M., Rafi, U., Imran, A., Rehman, M. U. and Abbas, S. K., "Risk factors associated with cardiovascular disorders: Risk factors associated with cardiovascular disorders," *Pakistan BioMedical Journal* 3-10 (2024).
- [3] Zhang, B., et al., "Global burden of cardiovascular disease from 1990 to 2019 attributable to dietary factors," *J. Nutr.* 153, 1730-1741 (2023).
- [4] Mensah, G. A., Fuster, V. and Roth, G. A., "A heart-healthy and stroke-free world," *J. Am. Coll. Cardiol.* 82, 2343-2349 (2023).
- [5] Ahsan, M. M. and Siddique, Z., "Machine learning-based heart disease diagnosis: A systematic literature review," *Artif. Intell. Med.* 128, 102289 (2022).
- [6] Shah, S. M. S., Shah, F. A., Hussain, S. A. and Batool, S., "Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods," *Computers & Electrical Engineering* 84, 106628 (2020).
- [7] Cohn, J. N., et al., "Screening for early detection of cardiovascular disease in asymptomatic individuals," *Am. Heart J.* 146, 679-685 (2003).
- [8] Abubaker, M. B. and Babayiğit, B., "Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods," *IEEE Transactions on Artificial Intelligence* 4, 373-382 (2022).
- [9] Swathy, M. and Saruladha, K., "A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using machine learning and deep learning techniques," *ICT Express* 8, 109-116 (2022).
- [10] Dahri, F. H., Dahri, N. A. and Soomro, M. A., "Image caption generator using convolutional recurrent neural network feature fusion," *Journal of Xi'an Shiyou University, Natural Science Edition* 9, 1088-1095 (2023).
- [11] Wong, K. K. L., Fortino, G. and Abbott, D., "Deep learning-based cardiovascular image diagnosis: a promising challenge," *Future Generation Computer Systems* 110, 802-811 (2020).
- [12] Moushi, O. M., Ara, N., Helaluddin, M. and Mondal, H. S., "Enhancing the accuracy and explainability of heart disease prediction models through interpretable machine learning techniques," In *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 1-6 (2023).
- [13] Ghosh, P., et al., "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access* 9, 19304-19326 (2021).
- [14] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R. and Jeganathan, S., "Prediction of cardiac disease using supervised machine learning algorithms," In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 570-575 (2020).
- [15] Bhatt, C. M., Patel, P., Ghetia, T. and Mazzeo, P. L., "Effective heart disease prediction using machine learning techniques," *Algorithms* 16, 88 (2023).
- [16] Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C. and Mensinkal, K., "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proceedings* 3, 127-130 (2022).
- [17] Sarah, S., Gourisaria, M. K., Khare, S. and Das, H., "Heart disease prediction using core machine learning techniques—a comparative study," In *Advances in Data and Information Sciences: Proceedings of ICDIS*, 247-260 (2022).
- [18] Akkaya, B., Sener, E. and Gursu, C., "A comparative study of heart disease prediction using machine learning techniques," In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1-8 (2022).
- [19] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., "Heart disease prediction using machine learning algorithms," In *IOP Conference Series: Materials Science and Engineering*, 012072 (2021).
- [20] Pal, M., and Parija, S., "Prediction of heart diseases using random forest," In *Journal of Physics: Conference Series* 012009 (2021).
- [21] Garg, A., Sharma, B. and Khan, R., "Heart disease prediction using machine learning techniques," In *IOP Conference Series: Materials Science and Engineering*, 012046 (2021).

- [22] Islam, S., Jahan, N. and Khatun, M. E., "Cardiovascular disease forecast using machine learning paradigms," In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 487-490 (2020).
- [23] Amin, M. S., Chiam, Y. K. and Varathan, K. D., "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics 36, 82-93 (2019).
- [24] Louridi, N., Amar, M. and El Ouahidi, B., "Identification of cardiovascular diseases using machine learning," In 2019 7th Mediterranean Congress of Telecommunications (CMT), 1-6 (2019).
- [25] Mohan, S., Thirumalai, C. and Srivastava, G., "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access 7, 81542-81554 (2019).
- [26] Dwivedi, A. K., "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Comput. Appl. 29, 685-693 (2018).
- [27] Soriano, F., "Heart failure prediction dataset," Kaggle, 2020, <<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>>