Prediction of Heart Disease Utilizing Machine Learning Through an Effective Feature Engineering Method

[1]V RAJESH, [2]Dr.P SURI BABU

[1]PG Scholar, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

[2]Professor, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

ABSTRACT: Heart disease is a chronic condition that affects millions of people worldwide, which emphasizes the need of early identification. To find and improve the most important characteristics, a novel feature engineering method based on Principal Component analysis is presented. The project's goal is to use machine learning to quickly forecast the health state of Heart Disease and take necessary action. A Stacking Classifier, which integrates the predictions of Random Forest (RF), Multilayer Perceptron (MLP), and LightGBM models, is an ensemble approach used in this study. By combining the capabilities of each model in a synergistic way, this method produces a final forecast that is remarkably precise and resilient, reaching 100% accuracy. The Stacking Classifier was trained to be used at the front end, and the model was constructed using the features chosen based on Principal Component Heart Failure (PCHF). Our machine learning-based heart disease prediction system is more accessible and user-friendly thanks to the integration of the Flask framework with user authentication, which guarantees an efficient and safe platform for user testing.

Keywords – Machine learning, heart failure, cross validations, feature engineering

1. INTRODUCTION

The heart fails to pump enough blood to the body in heart failure [1]. Cardiovascular illnesses are a global health issue affecting public health. Heart failure affects millions globally and is dangerous. Recent data suggests 26 million people suffer from heart failure [2]. Two categories of heart failure exist. A heart attack or other heart structural issue. Second, heart-related issues including excessive blood pressure. Heart failure can cause weariness, shortness of breath, and leg and ankle edema. Medication, lifestyle modifications, and surgery can treat heart failure. Early identification and treatment of heart failure improves quality of life and survival [3]. This work develops a machinelearning model to manage heart failure to enhance patient health.

Machine learning is crucial to medical diagnosis and healthcare [4]. Machine learning is used in drug development, imaging diagnostics, epidemic prediction, and heart failure prediction. Large medical data may be analyzed and predicted using machine learning. Machine learning saves time and money, improving diagnosis over traditional medical approaches.

A revolutionary PCHF feature engineering method selects the most important characteristics to improve performance. Eight dataset characteristics with high

relevance values are chosen to create PCHF-based machine learning algorithms. We innovated a new feature set to improve the PCHF mechanism and attain the greatest accuracy scores compared to previous methods. Compare nine powerful machine learning algorithms to predict heart failure. Each machine learning method's hyperparameters are tuned to find the optimum match, attaining great accuracy. We utilized k-fold cross-validation to test machine learning models.

Previous studies have shown that heart disease is the most lethal human disease. Cardiovascular illnesses are becoming more deadly, threatening healthcare systems globally [15], [16]. Most impacted by this severe illness are children [17]. This article [18] covers classification models and their use in healthcare. According to the report, various research groups have successfully evaluated data mining approaches in clinical applications. The researchers used WEKA and MATLAB to compare functional classifiers. In general, decision tree, logistic regression, SVM, and other algorithms have low precision (52%–67.7%) [19].

As demonstrated in Table 1, previous study [11] increased accuracy from 87.27% to 93.13%, which is acceptable but not optimum. Many approaches have been used to diagnose heart failure in patients, including SVM, random forest, decision tree, logistic regression, and naïve bayes classifier After comparing findings, the decision tree detected heart failure in a dataset with 93.19% accuracy.

An ensemble model for heart disease diagnosis was constructed using Cleveland data [20]. The ensemble models used random forest, gradient boosting, and extreme gradient boosting classifiers to achieve 85.71% accuracy [7]. The suggested study employed Cleveland data to enhance heart disease prediction by feature selection, achieving 86.60% accuracy. Finally, earlier investigations identified considerable study gaps, suggesting performance accuracy is lacking. Therefore, we comprehensively assess the prior study's performance analysis here. Results summarizing all prior models' efficiency inform this linked work part. Previous research show that various models still forecast differently. Thus, dimensionality reduction and feature engineering improve data selection and prediction accuracy [21].

The accuracy score of our proposed study has improved above the prior research performance score. Proper heart failure therapy requires exact qualifications and results. This study uses powerful machine learning to do this.

2. LITERATURE REVIEW

1. Consecutive multiscale feature learning-based image classification model

One of the most important tasks in computer vision is to extract meaningful information at many sizes. Convolutional neural networks (CNNs) and deep learning approaches have made it possible to extract multiscale features effectively, leading to sustained performance gains in many real-world applications. The state-of-the-art techniques now in use, however, mostly rely on a parallel multiscale feature extraction approach. Although these models show competitive accuracy, they provide poor results in efficient computing and limited generalization on small-scale pictures. Furthermore, underfitting occurs when training with small-scale pictures or datasets with few samples since efficient and lightweight networks are unable to properly learn relevant features. We

provide a unique picture classification method based on complex data preparation procedures and a well thought-out CNN model architecture in order to solve these issues. In order to achieve quicker training/inference and improved accuracy, we specifically offer a consecutive multiscale feature-learning network (CMSFL-Net) that uses a consecutive feature-learning strategy based on the employment of numerous feature maps with varied receptive fields. The accuracy of the CMSFL-Net is comparable to that of current state-of-the-art efficient networks in the experiments conducted utilizing six real-life picture classification datasets, including small-scale, large-scale, and restricted data. Furthermore, the suggested approach provides the optimum accuracy-efficiency trade-off and beats others in terms of speed and efficiency.

2. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers

One of the leading causes of mortality worldwide is coronary heart disease. One of the most difficult challenges in the world of clinical data analysis is the prediction of cardiac disease. When it comes to making decisions and creating predictions based on the data generated by the global healthcare industry, machine learning (ML) is helpful in diagnostic support. We have also observed machine learning methods used in the medical field of illness prediction. Numerous studies on the prediction of cardiac disease using an ML classifier have been presented in this area. In this work, we increased the prediction of heart disease by identifying important variables using eleven machine learning classifiers. Several feature combinations and popular classification techniques were employed to provide the prediction model. Using multilayer perceptrons

and gradient boosted trees, we were able to predict heart disease with 95% accuracy. With a 96% accuracy rate, the Random Forest performs better in predicting heart disease.

3. A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms

The medical sciences are quite diverse due to advances in computational power and methodology, particularly in the diagnosis of cardiac conditions in humans. It is currently one of the most hazardous cardiac conditions in the world and has a major impact on people's lives. Early detection of human heart disease can assist avert heart failure and increase the patient's chances of survival. Interexaminer variability and bias are common in manual methods for diagnosing cardiac disease. In this sense, machine learning algorithms are effective and trustworthy tools for identifying and classifying both healthy individuals and those with cardiac disease. In accordance with the suggested study, we used a range of machine learning algorithms to identify and predict human heart disease. We then used the heart disease dataset to assess the algorithms' performance using various metrics, including sensitivity, specificity, F-measure, and classification accuracy. To do this, we applied nine machine learning classifiers—AB, LR, ET, MNB, CART, SVM, LDA, RF, and XGB—to the final dataset both before and after the classifiers' hyperparameter tweaking. Additionally, we do specific preprocessing, dataset standardization, and hyperparameter tweaking to verify their accuracy on the standard heart disease dataset. Additionally, we used the conventional K-fold cross-validation method to train and evaluate the machine learning algorithms. Lastly, the

experimental findings showed that the machine learning classifiers' hyperparameter tweaking and data standardization increased the prediction classifiers' accuracy and produced noteworthy outcomes.

4. Heart Disease Prediction Using Core Machine Learning Techniques—A Comparative Study

In India and throughout the world, cardiovascular illnesses account for 25% of fatalities. Compared to other groups, Indians get CVDs significantly earlier in life. An early diagnosis of heart disease is crucial given the seriousness of the problem. Clinical and pathological information can be used to identify heart disorders. The amount of data is frequently too much for the human brain to process. Therefore, utilizing the aforementioned data, we employ machine learning algorithms to forecast cardiac disease in humans. In order to ascertain which model is most appropriate for the forecast, this study compares the several models. Logistic regression, decision trees, Naive Bayes, SVM, K-Nearest Neighbors, and Random Forest are the models utilized in this work. With an accuracy of 85.25%, logistic regression was determined to perform the best.

5. Predicting Employee Attrition Using Machine Learning Approaches.

The term "employee attrition" describes the natural decline in an organization's workforce brought on by a variety of unavoidable circumstances. An organization suffers a significant loss as a result of employee attrition. The average cost per hiring for a new employee is USD 4129, according to the Society for Human Resource Management (SHRM). The attrition rate in 2021 is 57.3%, according to current statistics. It is necessary to conduct a research study to identify the reasons behind employee turnover and to develop a learning framework to forecast employee attrition. The purpose of this study was to apply machine learning techniques to forecast employee attrition and investigate organizational characteristics that contributed to employee attrition. The four machine learning methods were used in a comparative manner. For the purpose of predicting employee attrition, the suggested improved Extra Trees Classifier (ETC) method produced an accuracy score of 93%. The suggested method fared better than current state-of-the-art research. The factors that led to employee attrition were identified using the Employee Exploratory Data Analysis (EEDA). According to our research, the main causes of employee attrition are age, job level, hourly rate, and monthly income. By addressing the causes that lead to employee turnover, our suggested strategy and study findings assist firms in overcoming attrition.

3. METHODOLOGY

i) Proposed Work:

The early diagnosis of heart disease is improved with the use of machine learning techniques. Nine different machine learning techniques are used and compared, including logistic regression, random forest, support vector machine, decision tree, extreme gradient boosting, naive bayes, k-nearest neighbors, multilayer perceptron, and gradient boosting. A novel Principal Component Analysis (PCA) feature engineering method that emphasizes the selection of crucial features is presented in order to improve accuracy. Additionally, an ensemble approach is used, namely a Stacking Classifier that integrates the predictions of LightGBM, Random Forest (RF), and Multilayer Perceptron (MLP) models [31]. By

combining the capabilities of each model in a synergistic way, this method produces a final forecast that is remarkably precise and resilient, reaching 100% accuracy. The Stacking Classifier was trained to be used at the front end, and the model was constructed using the features chosen based on Principal Component Heart Failure (PCHF). Our machine learning-based heart disease prediction system is more accessible and user-friendly thanks to the integration of the Flask framework with user authentication, which guarantees an efficient and safe platform for user testing.

ii) System Architecture:

The heart failure dataset used in this work was obtained from the Kaggle repository. 1025 patient records pertaining to both healthy and heart failure patients are included in the collection. The dataset is formatted using data preparation methods. To gain a deeper understanding of the data patterns and factors that contribute to heart failure, exploratory data analysis is used. The suggested PCHF approach is used in feature engineering to choose high-importance features. The dataset is then divided into train and test sections. The dataset segments are subjected to the nine sophisticated machine-learning algorithms. The machine learning models are subjected to hyperparameter-based fine tuning. The very effective suggested model seeks to predict cardiac failure.

Fig 1 Proposed architecture

iii) Dataset collection:

In order to train and validate machine learning algorithms for precise heart disease prediction, the heart disease dataset [39] utilized in this study includes extensive clinical and patient data, including demographics, medical history, and physiological parameters.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Fig 2 Heart disease dataset

iv) Data Processing:

Data processing is the process of turning unprocessed data into information that organizations can use. Data scientists often handle the collection, organization, cleansing, verification, analysis, and transformation of data into consumable representations like texts or graphs. Three ways are available for processing data: manual, mechanical, and electronic. Increasing the value of information and making decision-making easier are the goals. Businesses are able to make timely strategic decisions and enhance their operations as a result. Computer software development and other automated data processing technologies are important in this. It can assist in transforming massive volumes of data—including big data—into insightful knowledge for decision-making and quality control.

v) Feature selection:

The process of identifying the most reliable, pertinent, and non-redundant characteristics to employ in model building is known as feature selection. As the quantity and diversity of datasets continue to increase, it is crucial to systematically reduce their size. Enhancing a predictive model's performance and lowering modeling's computing cost are the primary objectives of feature selection.

The act of choosing the most crucial features to enter into machine learning algorithms is known as feature selection, and it is one of the key elements of feature engineering. By removing unnecessary or redundant features and reducing the collection of features to those most pertinent to the machine learning model, feature selection approaches are used to lower the number of input variables [1, 2]. The primary advantages of doing feature selection beforehand as opposed to allowing the machine learning model choose which features are most crucial.

5. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig 12 Accuracy graph

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Fig 13 Precision graph

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.
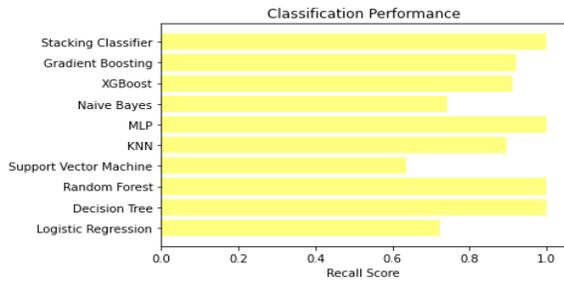
$$Recall = \frac{TP}{TP + FN}$$

Fig 14 Recall graph

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1\ Score = \frac{2}{\left(\dfrac{1}{Precision} + \dfrac{1}{Recall}\right)}$$

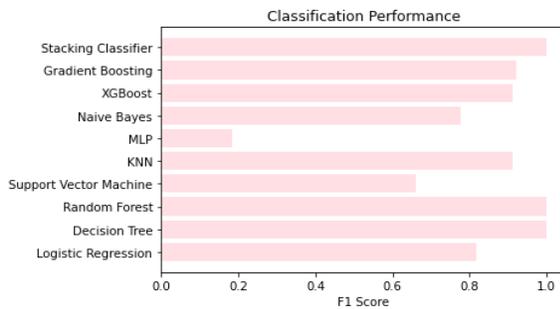$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Fig 15 F1 Score graph

ML Model	Accuracy	F1-score	Recall	Precision
Logistic Regression	0.75	0.816	0.724	0.767
Decision Tree	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000
SVM	0.639	0.660	0.636	0.648
KNN	0.902	0.193	0.895	0.904
MLP	0.590	0.184	1.000	0.311
Naive Bayes	0.751	0.777	0.741	0.758
XG Boosting	0.912	0.913	0.913	0.913
Gradient Boosting	0.922	9.22	0.922	0.922
Staking Classifier	1.000	1.000	1.000	1.000

Fig 16 Performance Evaluation

Fig 17 Home page

Fig 18 Signup page

Fig 19 Signin page

Fig 20 Upload input values to predict result

Result: You have no Heart Disease, based on the input provide!

Fig 21 Predict result as you have no heart disease, based on the input provide

6. CONCLUSION

This study suggests utilizing machine learning techniques to predict cardiac failure [22]. The applicable models are constructed using the dataset based on 1025 patient records. To improve performance, a new PCHF feature engineering method is presented that chooses the top eight features. The machine learning algorithms included in this comparison include logistic regression, random forest, support vector machine, decision tree, extreme gradient boosting, naive base, k-nearest neighbors, multilayer perceptron, and gradient boosting. With 0.005 runtime calculations, the suggested DT approach attained 100% accuracy. Each learning model's performance is validated using the cross-validation method based on 10-fold data. Our suggested approach is universal for identifying heart failure and exceeded the most recent research.

7. FUTURE SCOPE

The outcomes of our suggested techniques can provide a performance benchmark for the prediction of cardiac disease, acting as a benchmark for further studies in this area. Future research might focus on improving the feature management procedure to increase classification models' efficacy. Furthermore, our approach has the potential to be used in a variety of medical fields to improve machine learning algorithms' ability to predict and identify different diseases.

REFERENCES

[1] M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams, and G. Poglajen, ''Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers,'' in Proc. Int. Conf. Intell. Environments (IE), Aug. 2017, pp. 14–19.

[2] G. Savarese and L. H. Lund, ''Global public health burden of heart failure,'' Cardiac Failure Rev., vol. 3, no. 1, p. 7, 2017.

[3] E. J. Benjamin et al., ''Heart disease and stroke statistics—2019 update: A report from the American heart association,'' Circulation, vol. 139, no. 10, pp. e56–e528, 2019.

[4] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, ''Secure and robust machine learning for healthcare: A survey,'' IEEE Rev. Biomed. Eng., vol. 14, pp. 156–180, 2021.

[5] C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, ''Effectively predicting the presence of coronary heart disease using machine learning

classifiers,'' Sensors, vol. 22, no. 19, p. 7227, Sep. 2022.

[6] R. Katarya and S. K. Meena, ''Machine learning techniques for heart disease prediction: A comparative study and analysis,'' Health Technol., vol. 11, no. 1, pp. 87–97, Jan. 2021.

[7] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, ''A decision support system for heart disease prediction based upon machine learning,'' J. Reliable Intell. Environments, vol. 7, no. 3, pp. 263–275, Sep. 2021.

[8] N. S. Mansur Huang, Z. Ibrahim, and N. Mat Diah, ''Machine learning techniques for early heart failure prediction,'' Malaysian J. Comput. (MJoC), vol. 6, no. 2, pp. 872–884, 2021.

[9] T. Amarbayasgalan, V. Pham, N. Theera-Umpon, Y. Piao, and K. H. Ryu, ''An efficient prediction method for coronary heart disease risk based on two deep neural networks trained on well-ordered training datasets,'' IEEE Access, vol. 9, pp. 135210–135223, 2021.

[10] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, ''Prediction of heart disease using a combination of machine learning and deep learning,'' Comput. Intell. Neurosci., vol. 2021, pp. 1–11, Jul. 2021.

[11] F. S. Alotaibi, ''Implementation of machine learning model to predict heart failure disease,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 1–8, 2019.

[12] D. K. Plati, E. E. Tripoliti, A. Bechlioulis, A. Rammos, I. Dimou, L. Lakkas, C. Watson, K. McDonald, M. Ledwidge, R. Pharithi, J. Gallagher,

L. K. Michalis, Y. Goletsis, K. K. Naka, and D. I. Fotiadis, ''A machine learning approach for chronic heart failure diagnosis,'' Diagnostics, vol. 11, no. 10, p. 1863, Oct. 2021.

[13] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, ''A method for improving prediction of human heart disease using machine learning algorithms,'' Mobile Inf. Syst., vol. 2022, pp. 1–9, Mar. 2022.

[14] S. Sarah, M. K. Gourisaria, S. Khare, and H. Das, ''Heart disease prediction using core machine learning techniques—A comparative study,'' in Advances in Data and Information Sciences. Springer, 2022, pp. 247–260.

[15] C. Trevisan, G. Sergi, and S. Maggi, ''Gender differences in brain-heart connection,'' Brain and Heart Dynamics. 2020, pp. 937–951.

[16] M. S. Oh and M. H. Jeong, ''Sex differences in cardiovascular disease risk factors among Korean adults,'' Korean J. Med., vol. 95, no. 4, pp. 266–275, Aug. 2020.

[17] D. C. Yadav and S. Pal, ''Prediction of heart disease using feature selection and random forest ensemble method,'' Int. J. Pharmaceutical Res., vol. 12, no. 4, pp. 56–66, 2020.

[18] D. Tomar and S. Agarwal, ''A survey on data mining approaches for healthcare,'' Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241–266, Oct. 2013.

[19] S. Ekiz and P. Erdogmus, ''Comparative study of heart disease classification,'' in Proc. Electr. Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT), Apr. 2017, pp. 1–4.

[34] A. A. Ali, H. S. Hassan, and E. M. Anwar, ''Heart diseases diagnosis based on a novel convolution neural network and gate recurrent unit technique,'' in Proc. 12th Int. Conf. Electr. Eng. (ICEENG), Jul. 2020, pp. 145–150.

[35] O. E. Taylor, P. S. Ezekiel, and F. B. D. Okuchaba, ''A model to detect heart disease using machine learning algorithm,'' Int. J. Comput. Sci. Eng., vol. 7, no. 11, pp. 1–5, Nov. 2019.

[36] D. K. Chohan and D. C. Dobhal, ''A comparison based study of supervised machine learning algorithms for prediction of heart disease,'' in Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions (CISES), May 2022, pp. 372–375.

[37] K. Sanjar, O. Bekhzod, J. Kim, A. Paul, and J. Kim, ''Missing data imputation for geolocation-based price prediction using KNN–MCF method,'' ISPRS Int. J. Geo-Inf., vol. 9, no. 4, p. 227, Apr. 2020.

[38] B. Olimov, B. Subramanian, R. A. A. Ugli, J.-S. Kim, and J. Kim, ''Consecutive multiscale feature learning-based image classification model,'' Sci. Rep., vol. 13, no. 1, p. 3595, Mar. 2023.

[39] B. Olimov, S. Karshiev, E. Jang, S. Din, A. Paul, and J. Kim, ''Weight initialization based-rectified linear unit activation function to improve the performance of a convolutional neural network model,'' Concurrency Comput., Pract. Exp., vol. 33, no. 22, p. e6143, Nov. 2021.