OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12671276

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Heart-disease Prediction Using Ensemble-Learning Techniques

Sukruthi A[1], Sushma R[2], Vagdevi M N[3], Vaishnavi A*[4], Padmasree N[5]

[1,2,3,4]Student, Computer Science and Engineering, RVITM, Bengaluru, Karnataka, India
[5]Assistant Professor, Computer Science and Engineering, RVITM, Bengaluru, Karnataka, India

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Sukruthi, A., Sushma, R., Vagdevi, M. N., Vaishnavi, A., Padmasree, N. (2024). Heart-disease Prediction Using Ensemble-Learning Techniques. *Indiana Journal of Multidisciplinary Research*, *4*(3), 82-86.

Abstract: Heart-disease poses a growing threat to global health, driven by various lifestyle factors and demographic shifts. This study proposes ensemble model approach utilizing three Kaggle-sourced datasets to anticipate cardiac problems. Our methodology combines Random-Forest, XGBoost, and Logistic-Regression algorithms, alongside a Voting Classifier, to enhance predictive accuracy. Through 5-fold cross-validation, our ensemble model achieves a compelling training accuracy of 99.4% and testing accuracy of 91.7%. This amalgamation of datasets offers a thorough comprehension of multifaceted risk factors, including lifestyle behaviors, genetic predispositions, and clinical markers. Using machine-learning algorithms, our approach empowers healthcare practitioners with actionable insights for early detection and tailored intervention strategies. As heart-disease prevalence continues to rise, integrating advanced ensemble techniques holds promise in improving risk assessment, thereby mitigating its impact on public health.

Keywords: Heart-disease prediction, Ensemble model, Random-Forest, XGBoost, Logistic-Regression, Voting Classifier

INTRODUCTION

Heart-disease is the primary factor behind mortality worldwide, affecting the heart and coronary arteries in many different manners. It is a serious public health problem.. To improve patient outcomes and enable successful treatment, a timely and correct diagnosis is essential. Conventional diagnostic techniques, however, may have drawbacks. These techniques can be costly and time-consuming, and symptoms may not always be obvious, delaying detection.

Using machine learning (ML) to evaluate substantial volumes of healthcare information and identify complex patterns offers a viable way to tackle these issues. Through the integration of large-scale datasets that include demographics, clinical test results, and medical history, researchers may train machine learning algorithms to create models that predict a person's risk of acquiring heart-disease. This gives medical professionals the ability to proactively identify people who are at higher risk, allowing for earlier intervention and possibly saving lives.

This study examines the utilization of supervised-learning algorithms for predicting heart-disease. Training models using labeled data—where each data point is associated with a specific outcome or class label—is the process of supervised-learning. The model undergoes training with patient health-records, which have been categorized as either indicative of heart-disease or not, within the realm of heart-disease prediction. Upon training, the model is capable of forecasting the likelihood that a new patient will have the condition by analyzing previously unseen patient data.

In this investigation, we explored various machine-learning methods such as SVM, KNN, Random-Forest, and Logistic-Regression, to forecast heart-disease. Finding the best strategy for precisely estimating the likelihood of cardiovascular disease was our goal. After evaluating several methods, we found that the most successful results were attained by integrating them with an ensemble learning approach. We used a voting mechanism for predictions and combined the Random-Forest, XGBoost, and Logistic-Regression methods. This ensemble technique performed admirably for our problem description.

Real-world datasets frequently have erroneous characteristics, inconsistent values, and missing values. These issues are addressed by data pretreatment approaches, which also handle missing values, normalize features to guarantee scale consistency, and use feature selection techniques to identify the most informative features for the model. To enhance the precision and effectiveness of machine learning models in the prediction of heart-disease, effective data preparation is essential.

This project aims to contribute to the creation of a dependable in addition to reasonably priced method for identifying heart-disease at an early stage by utilizing the capabilities of machine learning and data science. A system like this could enable medical practitioners to make well-informed decisions about patient treatment, which may lead to improved patient outcomes and a

***Corresponding Author:** Vaishnavi A

decline in the occurrence of cardiovascular disease worldwide.

METHODOLOGY

This study's methodology entails a comprehensive assessment of supervised-learning algorithms. First, A wide range of patient data sets, including clinical test results and medical histories, are gathered and reprocessed to remove extraneous features and missing values. By determining the best algorithm for cardiac disease prediction, this scientific approach hopes to enhance outcomes for patients and healthcare decision-making.

Data Preprocessing

We imported the combined dataset, which contained crucial information for forecasting the likelihood of heart-disease, to begin our data preprocessing step. The dataset consists of 1632 cases, each of which is defined by 14 variables. These attributes encompass elements such as age, gender, kind of chest-pain, blood pressure, cholesterol, and other cardiovascular markers. To guarantee uniformity and compatibility throughout the dataset, we applied feature scaling using the Standard Scaler from the sklearn. preprocessing module. Through the procedure of scaling to unit variance and mean subtraction, the distribution of numerical features was normalized. After that, we divided the dataset into two categories: the goal variable (y), which represents the existence or non-existence of cardiovascular ailments, and the input attributes (X). We divided the data into training, validation, and testing groups using the train_test_split method from sklearn. model_ selection to enable robust model evaluation. We managed to train the model using a segment of the dataset, fine-tune its parameters using the validation set, and subsequently evaluate its performance based on hidden data from the trial set Furthermore, we used regularization methods in certain models of classification, such as Logistic-Regression, XG Boost, and Random-Forest to prevent overfitting and improve generalization. Finally, we saved both the scaler object and the trained ensemble model using joblib for future utilization and assessment.[2]

Dataset

The dataset comprises 1632 instances, each characterized by 14 attributes. Some of these attributes include age, sex, resting blood pressure, kind of chest pain, serum cholesterol levels, fasting blood sugar status, results of the resting electrocardiogram, maximum heart rate reached, presence of exercise-induced angina, ST depression caused by exercise relative to rest (old peak), slope of peak exercise ST segment, no. of major vessels coloured by fluoroscopy, and thalassemia classification. The dataset aims to predict the existence of cardiac-disease in the "target" field, with values of 0 signifying no disease and 1 indicating the existence of the disease. The ratio of testing to training has been 80:20 to enable robust model validation ensuring a comprehensive assessment of predictive performance during subsequent analysis.

Proposed model

In our heart-disease forecasting research, we used an ensemble modelling method to improve prediction performance. This approach involved integrating three robust machine learning algorithms: Random- Forest, XGBoost, and Logistic-Regression. Each algorithm was trained separately using pre-processed datasets sourced from the original dataset.

A. ***Random-Forest:***

Random-Forest is a well-known ensemble learning method that excels at managing data with a high number of dimensions while maintaining resilience. During training, it constructs multiple decision trees and returns the mean forecast for each tree or the class mode. Random-Forest enhances unpredictability by solely taking into consideration a random subset of qualities at each split point, allowing it to reduce overfitting and increase generalisation.[1]

B. ***XGBoost (Extreme Gradient Boosting):***

A sophisticated implementation of gradient-boosting machines intended to improve rapidity and efficacy is referred to by the name of XGBoost. The construction of decision-trees is methodical, with each subsequent tree rectifying faults resulting from the ones before it. XGBoost efficiently eliminates predetermined loss functions by leveraging an adaptive gradient descent optimisation technique, which makes it an excellent choice for large-scale datasets. Its capacity to identify complex patterns frequently results in better results in an assortment of regression and classification tasks.

C. ***Logistic-Regression:***

Logistic-Regression is a linear model utilized primarily for binary classification tasks. By fitting data to a logistic curve, an event's probability of happening is calculated. Despite its simplicity, Logistic-Regression is interpretable and serves as a baseline model for classification. Since it assumes a linear link between the log-odds of the outcome and the input data, it is especially useful in circumstances where there are few characteristics or linear correlations. By amalgamating these diverse algorithms into an ensemble model using the Voting Classifier from the scikit-learn library, we leveraged their collective decision-making to enhance overall predictive performance. Utilizing the unique benefits of every algorithm, this ensemble model produces a reliable and flexible forecasting approach for the categorization of cardiac disease. [1]

Five-fold cross-validation is a systematic evaluation strategy that we used to evaluate the performance of our ensemble model in our heart-disease prediction project. This method divides the dataset into five equalsized subsets. The ensemble model is then

***Corresponding Author:** Vaishnavi A

trained using four of these subsets, and performance is evaluated using the fifth subset. During each of the five iterations of this procedure, each subset is utilized as the validation set exactly once. We reduced the conceivable outcome of overfitting while ensuring a comprehensive validation procedure by carrying out several iterations. This comprehensive evaluation procedure yielded accurate perceptions of the ensemble model's performance as well as the capacity to generalize to new data.

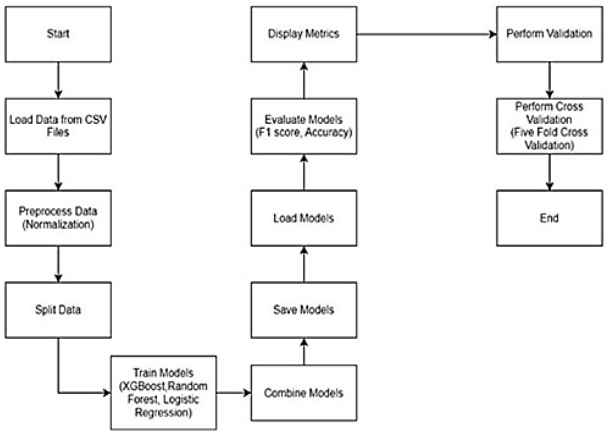

Figure 1: Flowchart

RESULTS

The primary aim of the study was to assess the likelihood of heart-disease by combining data from various sources using machine learning methods. With a 95.79% accuracy rate, the model was shown to have good predictive power. On average accuracy of 92.62% across five folds, cross-validation scores provided more evidence of the model's consistency. In each of the classes (0 and 1), the testing data's classification report showed balanced precision, recall, and F1-score values, demonstrating an efficient prediction of both positive and negative cases related to heart-disease. The model's capacity to correctly generalize to unknown inputs was demonstrated by the test results' 92% overall precision.

Evaluation Metrics:
Training Accuracy: 0.9942473633748802
Testing Accuracy: 0.9174311926605505
Precision: 0.920532199329345
Recall: 0.9174311926605505
F1-score: 0.9175316704798142
Validation Accuracy: 0.9578544061302682
Cross-Validation Scores: [0.92344498 0.90909091 0.94258373 0.93269231 0.92307692]
Mean CV Accuracy: 0.9261777695988224

Few misclassifications were seen in the confusion matrices of the training and testing data. Few occurrences within the dataset used for training were incorrectly identified, demonstrating the model's strong learning from training set. Similar to this, the confusion-matrix in the the testing data demonstrated a low amount of false-positives and false-negatives, highlighting the model's effectiveness in correctly categorizing cases of heart-disease. The findings suggest that the machine learning model developed in this research could potentially assist healthcare professionals in identifying and addressing cardiovascular disease prevention efforts early on. Improved patient outcomes and lower healthcare costs may arise from the model's increased validation and improvement in clinical settings.

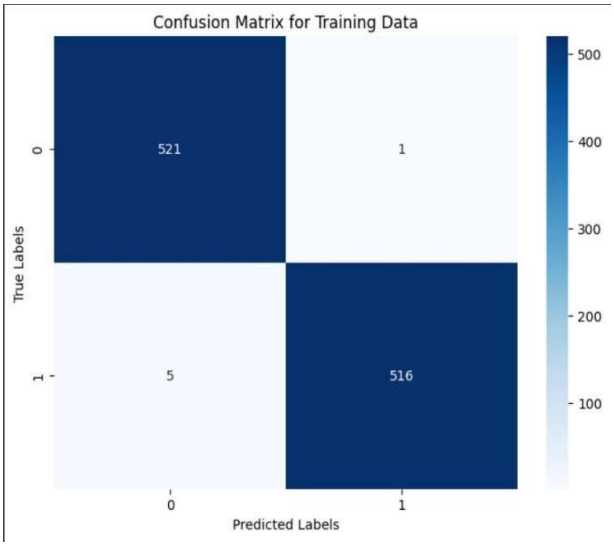

Figure 2: Confusion matrix - Training data

Figure 3: Confusion matrix - Testing data

DISCUSSION

Throughout our search for the best machine-learning model to make predictions, we made use of Generative-Adversarial-Networks (GANs). They consist of two neural networks: the generator and the discriminator. The generator produces fabricated data-samples. in an attempt to replicate the training-data distribution—such as text or images—from random noise. Concurrently, the discriminator learns to differentiate between genuine and fabricated samples. This adversarial training motivates the generator to

***Corresponding Author:** Vaishnavi A

produce more realistic samples, while the discriminator improves its discrimination capability. GANs find applications in diverse fields such as image generation, style transfer, and data augmentation. However, issues like mode collapse and training instability still exist, requiring ongoing research to get beyond these obstacles and improve GAN performance in generative modeling applications.

Initially, we explored GAN technology because of its ability of data augmentation and overcoming data scarcity issues, particularly given our dataset's limited size. Notably, when applying GANs to a subset of our dataset containing 300 to 388 rows, we achieved a promising accuracy of 76.92%. However, as we expanded our investigation to a larger subset comprising 300 to 865 rows, the accuracy decreased to 68.21%. Despite our initial optimism, the synthesized data generated by the GAN failed to produce precise forecasts for heart-disease. This result is likely attributable to issues with data fidelity, class imbalance, and model convergence issues. Consequently, we shifted our focus towards combining three different datasets into one, pre-processing it to remove redundancy and handling various data inconsistencies. We then employed a variety of machine-learning models such as Support Vector Machines, the Random-Forest model, K-Nearest Neighbor, Logistic-Regression, and XG Boost techniques. Then we chose to combine various models using Ensemble Learning techniques, combining Random-Forest, SVC, and Gradient- Boosting and combining Random-Forest, XG Boost, and KNN produced accuracy scores that were encouraging. Upon further exploration, we combined Random-Forest, XG Boost, and Logistic-Regression algorithms, produced forecasts that were more accurate. This ongoing process underscores the importance to assess new techniques rigorously and choose strategies that successfully handle the unique problems our dataset presents.

Sl. No.	ML Algorithms applied	Training accuracy	Testing Accuracy	Precision	Recall	F1 Score
1	Random-Forest	1.0	0.93272	0.93399	0.93272	0.9328
2	SVC	0.74616	0.75229	0.75226	0.75229	0.75142
3	K-Nearest Neighbor	0.93865	0.84098	0.84234	0.84098	0.84116
4	Logistic-Regression	0.74923	0.76146	0.76127	0.76146	0.76090
5	XG Boost	1.0	0.93272	0.93463	0.93272	0.93280
6	Ensemble (Random-Forest, SVC, and Gradient-Boosting)	0.87577	0.85933	0.87007	0.85933	0.85922
7	Ensemble (Random-Forest, XGBoost and KNN)	0.99521	0.91437	0.91704	0.91437	0.91448
8	Ensemble (Random-Forest, XGBoost, and Logistic-Regression)	0.99424	0.91743	0.92053	0.91743	0.91753

CONCLUSION

Our study concludes by demonstrating the successful application of ensemble machine-learning techniques for highly accurate prediction of susceptibility to cardiovascular disease. Through the integration of diverse datasets and the utilization of advanced algorithms such as Random-Forest, XGBoost, and Logistic- Regression within a Voting Classifier framework, we achieved notable advancements in prediction performance. The obtained train-accuracy of 99.4% and test-accuracy of 91.7% underscore the resilience and practicality of our model.

By conducting careful feature engineering and selection, we identified crucial variables contributing to the risk of heart-disease, which can aid in early-intervention-strategies and enhance understanding. Our study highlights the importance of machine-learning in proactive risk-assessment and personalized patient treatment within the medical industry. Additionally, the prototype can be further refined by exploring additional data sources and incorporating cutting-edge methods to improve prediction capabilities.

Ultimately, our work paves the way for the development of reliable, data-driven tools for treating cardiovascular diseases, ultimately leading to improved patient outcomes and reduced medical expenses.

CONFLICT OF INTEREST

A conflict of interest does not exist.

ACKNOWLEDGEMENTS

We have the utmost gratitude for the guidance and support provided by the CSE faculty members, the head of CSE at RVITM, and the principal of RVITM in completion of this project.

REFERENCES

1. Hossain, M. I., Maruf, M. H., Khan, M. A. R., Prity, F. S., Fatema, S., Ejaz, M. S., & Khan, M. A.

***Corresponding Author:** Vaishnavi A

2. S. (2023). Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison. *Iran Journal of Computer Science*, *6*(4), 397-417.
3. Swathy, M., & Saruladha, K. (2022). A comparative study of classification and prediction of Cardio-Vascular Diseases (CVD) using Machine Learning and Deep Learning techniques. *ICT Express*, *8*(1), 109-116.
4. Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.
5. Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.
6. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 345.
7. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.
8. Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672.
9. Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962-969.
10. Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology*, *11*(1), 87-97.
11. Sharma, V., Yadav, S., & Gupta, M. (2020, December). Heart disease prediction using machine learning techniques. In *2020 2nd international conference on advances in computing, communication control and networking (ICACCCN)* (pp. 177-181). IEEE.
12. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684- 687.
13. Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In *2020 international conference on electrical and electronics engineering (ICE3)* (pp. 452-457). IEEE.
14. Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.
15. Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88.
16. Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In *2016 international conference on circuit, power and computing technologies (ICCPCT)* (pp. 1-5). IEEE.
17. Taneja, A. (2013). Heart disease prediction system using data mining techniques. *Oriental Journal of Computer science and technology*, *6*(4), 457-466.
18. Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. *Computer Science & Information Technology Journal*, *24*, 53-59.
19. Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263-275.

***Corresponding Author:** Vaishnavi A