2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN) | 979-8-3503-2443-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CICN59264.2023.10402340

Predictive Modeling for Heart Disease Detection with Machine Learning

Rushita Gandham
Department of CSE,
SRM University-AP, India
rushita_gandham@srmap.edu.in

Keerthi Reddy Manambakam
Department of CSE,
SRM University-AP, India
keerthi_manambakam@srmap.edu.in

Sai Venkat Naveen Madala
Department of CSE,
SRM University-AP, India
saivenkatnaveen_m@srmap.edu.in

Navya Sri Nannapaneni
Department of CSE,
SRM University-AP, India
navyasri_nannapaneni@srmap.edu.in

Srilatha Tokala
ACT Lab, Department of CSE,
SRM University-AP, India
srilatha_tokala@srmap.edu.in

Murali Krishna Enduri
ACT Lab, Department of CSE,
SRM University-AP, India
muralikrishna.e@srmap.edu.in

Abstract— **Debilitating health symptoms brought on by heart disease reduce people's quality of life and impose serious pain, discomfort, and restrictions on daily activities. It places a heavy burden on economies, healthcare systems, and society at large. Accurate cardiac disease prediction has the ability to significantly contribute to prevention, treatment, and essential assistance for healthcare personnel facing this ailment given its influence on public health. This study uses the most recent developments in machine learning techniques to build an accurate model for heart disease prediction. Heart disease prediction and the Cleveland datasets, which combine approximately 13 important patient history variables, are used to analyze data from people with and without heart disease. XGBoost, naive bayes, logistic regression, decision trees, support vector machines, random forests, and k-nearest neighbors are just a few of the machine learning techniques used in the model development for classification. We can increase the precision and effectiveness of identifying persons at risk of heart disease and enabling prompt therapies by applying these machine learning techniques. According to the findings of this study, XGBoost, decision trees, and random forests have consistently produced high accuracy predictions of heart disease.**

Keywords-heart disease; health care; machine learning; accuracy

I. NTRODUCTION

The global impact of cardiovascular disease is vast, posing a significant and extensive challenge to public health. With a considerable annual toll on lives, heart disease ranks among the primary causes of global mortality [1]. Its widespread prevalence underscores its significance as a critical health issue that transcends geographical borders, affecting individuals across diverse demographics. Machine learning algorithms can comprehend the links between input variables and the chance of developing heart disease by using medical data for training purposes. Machine learning algorithms analyze extensive sets of clinical data, encompassing patient medical histories, vital signs, laboratory findings, and lifestyle elements. Through this process, they can uncover hidden patterns and correlations that conventional methods might overlook [2]. By shedding light on risk factors associated with heart disease, targeted interventions, and preventative measures, we can enhance public health outcomes and alleviate the burden of cardiovascular disease. Detecting heart disease early and implementing lifestyle adjustments can lead to healthier individuals and reduce long-term healthcare expenses.

In recent years, cardiovascular diseases have become a leading cause of death worldwide. The World Health Organization (WHO) estimates that these illnesses cause 17.7 million fatalities per year, or roughly 31% of all deaths worldwide. This pattern is also seen in India, where heart-related illnesses are now the leading cause of death [1]. In fact, according to the startling 2016 Global Burden of Disease Report, which was issued on September 15, 2017, 1.7 million Indian lives were lost to cardiac ailments in 2016 alone. These heart-related problems increase healthcare costs while also lowering personal productivity. According to WHO estimates, the economic toll of cardiovascular illnesses in India between 2005 and 2015 could have been up to $237 billion [2]. Therefore, it is crucial to be able to anticipate heart-related illnesses with accuracy and viability.

II. RELATED WORK

Modern Machine Learning (ML) algorithms have opened a whole new realm of opportunities. They help in performing complex computations efficiently which helped researchers to tackle the problem of heart disease prediction. Some of the related works are:

Archana et al. [3] has performed algorithms named KNN, decision tree, SVM, linear regression on the UCI repository dataset and obtained the accuracies of 87%, 79%, 83%, and 78% respectively. Dimensionality reduction entails selecting a mathematical representation that captures a significant portion of the variance within a given dataset while excluding less impactful information. When dealing with data for a particular task, numerous attributes or dimensions

979-8-3503-2443-3/23/$31.00 ©2023 IEEE
DOI: 10.1109/CICN.2023.55

might be present, but not all contribute equally to the desired output. A surplus of attributes can result in computational complexity and potential overfitting, leading to suboptimal outcomes. Therefore, dimensionality reduction assumes a critical role in model building. Typically accomplished through Feature Extraction and Feature Selection techniques, dimensionality reduction involves deriving a new feature set from the original one. Feature extraction entails transforming features, although this transformation may not be fully reversible, potentially leading to information loss. For feature extraction, Principal Component Analysis (PCA), as shown in [3] and [4], is frequently used. A popular linear transformation technique called PCA finds feature space directions that optimize variance and are orthogonal to one another. When used as a global algorithm, it produces the best reconstruction. While this was going on, Hodges et al. [13] published the K-Nearest Neighbour (KNN) rule, a nonparametric method for classifying patterns. KNN is a simple yet effective classification algorithm that makes few generalizations about the distribution of the data. When there is little prior data understanding, it is especially helpful. The method locates the k data points in the training set that are closest to an unlabeled data point and then uses their average value as a forecast.

According to [4], decision tree approaches were widely used to identify cardiac disease and had promising results. In order to predict cardiac disease, this work stresses the use of decision tree classifiers such the Logistic Model Tree method (j48) [4]. Comparatively speaking, J48 performed better in terms of sensitivity and accuracy, while LMT attained the best specificity. The study's findings concluded that j48 exhibited the highest accuracy. Hybrid algorithms combine the strengths of distinct individual algorithms to create a more robust and powerful solution. The study conducted by Mohan et al. [5] centers around predicting the presence of heart disease through a hybrid approach that amalgamates random forest and linear model techniques. This approach leverages the characteristics inherent to both random forest and linear model methodologies.

III. Methodology

Machine learning encompasses the automated generation of models, representing a type of data analysis. They can detect patterns and make predictions based on provided information with minimal intervention from humans. There are seven machine learning algorithms that are briefly discussed in this section.

A. *Logistic Regression*

This algorithm has the capability to predict whether an input is likely to be a part of either of two categories [6]. The model examines the connection between factors and a dependent variable. The formula

$$p(x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \qquad (1)$$

describes the model, which can be converted to

$$\frac{\ln(p(x))}{1 + \ln(p(x))} = b_0 + b_1 x \qquad (2)$$

In (1) and (2) $p(x)$, b_0, and b_1 represent the predicted output, bias term, and coefficient respectively. LR quantifies the likelihood of an individual having heart disease based on a combination of critical factors derived from medical observations and patient history. The aim is to minimize predicted vs. actual data differences via trained coefficients.

B. *Decision Tree*

It makes predictions by navigating a tree structure originating from input features [7, 8]. During prediction, input data traverses the tree, decisions are taken at each node based on feature values, and the process culminates at a leaf node providing the final majority-class prediction. The quality of splits is assessed using Gini impurity. It plays a critical role in ensuring that the tree structure maximizes information gain while minimizing impurity. This leads to the creation of effective decision boundaries that separate individuals at risk of heart disease from those who are not. It is calculated using the formula (3).

$$\text{G.I} = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad (3)$$

In (3) p_i is the proportion of class i in the set. Recursion halts when the maximum depth (set by the max_depth parameter) is reached, or when it's unable to find a split that decreases impurity.

C. *Naive Bayes*

This method relies on Bayes' theorem, denoted by formula (4), to compute the probability of an event given the likelihood of another event. It functions under the assumption of feature independence [8, 9], streamlining computations. If the posterior probability is above a certain threshold, the patient is classified as having heart disease; otherwise, they are classified as not having heart disease.

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)} \qquad (4)$$

D. *Random Forest*

It stands as a robust ensemble technique, amalgamating numerous individual decision trees to make predictions, thereby enhancing resilience and precision [10, 11]. Utilizing a combination of bagging and feature randomness, individual trees are trained on different data subsets and features. The collective outputs of these trees are fused, and predictions are established through a majority-based approach. This ensemble method effectively predicts heart disease by leveraging multiple decision trees for accuracy and reliability

E. K-Nearest Neighbors

In the context of heart disease prediction, k-Nearest Neighbors (KNN) is a learning method that classifies patients based on labeled examples. It accomplishes this by evaluating the similarity of individuals to their nearby counterparts [12, 13]. In this predictive process, various techniques assess the similarity of health data points, including measuring the difference between a patient's health profile and nearby individuals. The number of neighbors considered depends on the chosen 'k' value, ensuring personalized and accurate heart disease diagnosis.

F. Support Vector Machine

It finds extensive application in tasks related to classification [14, 15]. SVM delineates separate classes within a multi-dimensional space using a hyperplane. SVM creates a line (hyperplane) by adjusting it step by step to reduce errors. It strives to delineate an optimal line that effectively segregates distinct data groups, maximizing the separation space between them (referred to as the maximum marginal hyperplane) [16, 17]. For heart disease prediction, SVM's hyperplane aids in effectively distinguishing between different patient groups, facilitating accurate decisions.

G. XGBoost (Extreme Gradient Boosting)

It enhances accuracy by aggregating multiple decision tree's predictions. Employing a gradient boosting framework, it falls under the category of boosting algorithms that amplify predictive power by combining weak learners (decision trees) [12, 17]. Through gradient boosting, it optimizes the ensemble of decision trees, effectively mitigating errors and reducing overfitting. It aids in heart disease prediction via accurate feature ranking, class balancing, and interpretability.

The collected datasets were loaded into a Python environment using Google Collaboratory for further analysis and model development. To transform category variables into a numerical format, dummy variables are made. The characteristics are normalized using the Standard Scaler [13, 19]. The dataset is divided into training and testing valves in an 80:20 ratio. The algorithms LR, SVM, KNN, DT, RF, XGB, NB are fit to the training set. Grid search helped to fine-tune the models by finding optimal parameters [14, 19]. Measures like accuracy, F1-Score, precision, and recall are employed for model assessment. They are plotted for each algorithm using a bar graph to compare them.

Figure 1 illustrates the step-by-step implementation of the model, encompassing data preprocessing up to the stage of exporting the finalized model for implementation.

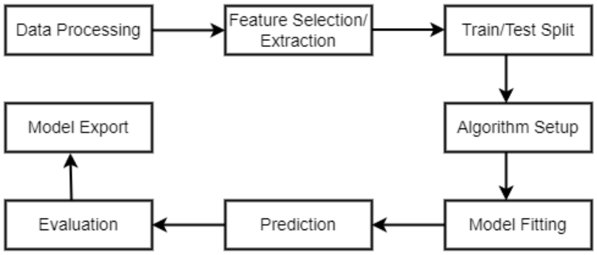

Fig. 1. Flowchart for the overall procedure

IV. Dataset Statistics

We have two datasets, Cleveland and Heart Disease Prediction, from the UCI Machine Learning Repository and Kaggle respectively. The UCI repository dataset encompasses 303 records, while the Kaggle dataset is more extensive, containing 1024 records. The datasets consist of the following 14 features:

Table I shows the features that are utilized for the prediction. These features encompass various aspects of the patient's health and diagnostic measurements.

Table I: Dataset Attributes

Continuous Numerical Features	**Categorical Numerical Features**	**Binary Features**	**Discrete Numerical Features**
Age depression thalach MaxHR trestbps oldpeak/ ST chol	thal restecg slope cp	sex fbs exang target	Number of vessels fluoro

A heatmap is employed to visualize data in a two-dimensional arrangement, utilizing colors to depict the intensity or value of data points [15, 20]. It is clear from Figure 6 that the relationship between cardiac disease and maximum heart rate is adverse. This suggests that people who have lower maximal heart rates are more susceptible to developing heart disease.

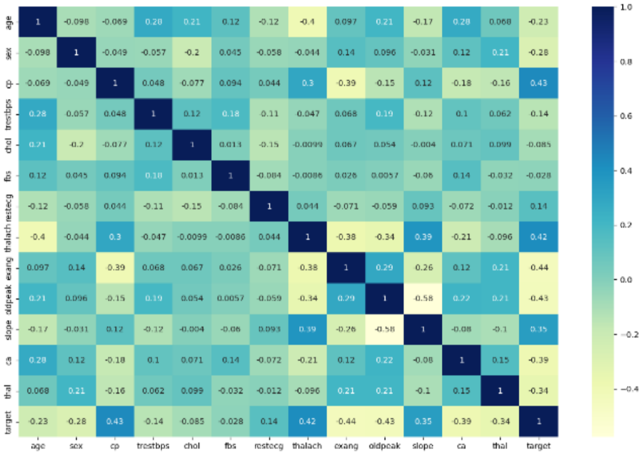

Fig.2. Heatmap for the Cleveland dataset

V. Results

A confusion matrix provides a structured representation that showcases how a model's classifications align with each class. This table helps to calculate important evaluation metrics [16, 21]. These metrics are used to evaluate and know how well the algorithms have performed [22].

A. ***Accuracy:*** It shows the proportion of patients who were accurately predicted to all of the patients, as indicated in (5)

$$\text{Accuracy} = (TN+TP) / (FN+TP+TN+FP) \quad (5)$$

B. ***Recall:*** It shows the proportion of patients with heart disease who are correctly diagnosed to all patients with heart disease, as illustrated in (6).

$$\text{Recall} = TP / (TP+FN) \quad (6)$$

C. ***Precision*****:** It is the ratio of accurately predicted patients with heart disease to the overall number of patients predicted as having heart disease as shown in (7).

$$\text{Precision} = TP / (FP+TP) \quad (7)$$

D. ***F1-Score:*** It is calculated using precision and recall. It represents the balance between them. Its formula is given by (8).

$$\text{F1-Score} = 2 \times ((\text{Precison} \times \text{Recall}) / ((\text{Precison} + \text{Recall})) \quad (8)$$

For each dataset, the metrics obtained for each algorithm are stored in a list and plotted with the algorithm names on x-axis.

Figure 3 presents the metric values achieved by each algorithm on the Heart Disease Prediction dataset. The outcomes are striking, random forest achieved a remarkable accuracy of 100%, closely followed by decision trees at 86%, and XGBoost at 82%.

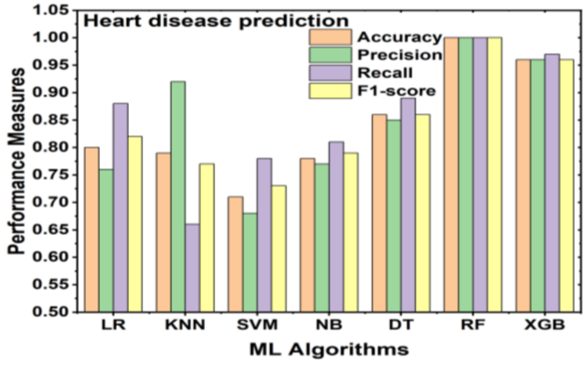

Fig. 3. Metrics Performance Comparison of Seven Algorithms for Heart disease prediction dataset.

Notably, these algorithms consistently maintained recall scores of 1.0, 0.89, and 0.85, respectively. This shows their reliability in identifying positive instances and their consistent effectiveness in recognizing relevant cases. Moreover, the precision scores were noteworthy, with each algorithm achieving a precision of 1.0, 0.85, and 0.81. This highlights their capability to minimize false positives. They also excelled in terms of the F1-score, achieving approximate values of 1.0, 0.86, and 0.83. This shows their adeptness at effectively balancing precision and recall.

We can see the metric values obtained by each algorithm for the Cleveland dataset in Figure 4. LR achieved the highest accuracy at 93%, followed closely by SVM at 92% and KNN at 90%. All three algorithms showed consistent recall and precision scores. Among the tested algorithms, logistic regression achieved the highest F1-score at approximately 0.94. Decision trees, random forest, and XGBoost exhibited accuracies near 0.89, with precision, recall, and F1-scores consistently exceeding 0.85.

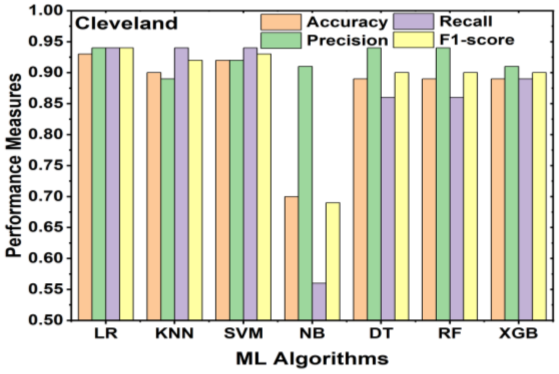

Fig. 4. Metrics Performance Comparison of Seven Algorithms for Cleveland dataset.

Employing multiple datasets enhances result reliability by assessing model consistency across diverse data. Though logistic regression, SVM, and KNN delivered impressive performance on the Heart Disease Prediction dataset, their effectiveness was lessened when applied to the Cleveland dataset. Naive Bayes demonstrated subpar performance on both datasets, yielding only 70% and 81% accuracies.

Upon thorough analysis, it becomes apparent that decision trees (DT), random forest (RF), and XGBoost consistently demonstrated robust performance. This makes them promising candidates for future datasets. Both random forest and XGBoost are ensemble algorithms, granting an advantage by constructing multiple trees and leveraging majority voting for decisions. The reliability and consistent accuracy of DT, RF, and XGBoost in predicting heart disease are evident, as shown in Figure 5, with an average accuracy of nearly 90% across the two datasets.

Apart from these metrics, we can consider other metrics for the evaluation of heart disease prediction such as the ROC-AUC curve, specificity, NPV (Negative Predictive Value), PPV (Positive Predictive Value), and F2 score which can provide valuable insights into model behavior.

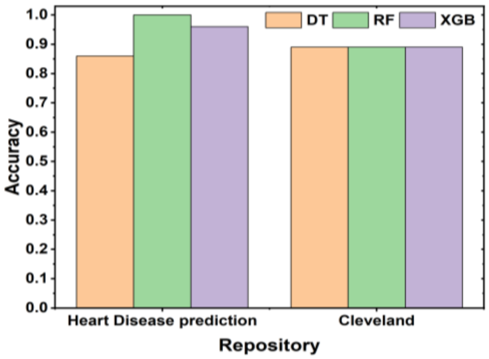

Fig. 5. Accuracies obtained for Decision Trees, Random Forest, XGBoost on the two datasets.

VI. Conclusion and Future Work

Predicting a heart disease holds a lot of importance in the medical industry due to its high impact on public well-being. Through the analysis of datasets housing crucial medical history attributes, data mining can unearth concealed patterns that contribute to heart conditions. The preprocessing and normalization of data are pivotal in heightening accuracy, maintaining data uniformity, and mitigating biases. ML techniques, such as DT, RF, and XGB, hold great promise in improving prediction accuracy and supporting healthcare professionals in early identification and intervention.

In the future, we can improve the prediction of heart disease by collecting larger datasets that encompass a more extensive and diverse patient population. This can help to reduce potential biases and ensure the predictive models are general and applicable to a broader range of individuals. We can further explore the integration of advanced methods like convolutional neural networks to find detailed data patterns, potentially improving prediction accuracy.

References

[1] 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.

[2] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4): e0174944.

[3] 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).

[4] Patel, J., Upadhyay, T., & Patel, S. (2016). Heart Disease prediction using Machine learning and Data Mining Technique. doi:10.090592/IJCSC.2016.018

[5] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[6] 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.

[7] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

[8] Liu, Kailong, et al. "Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification." IEEE/ASME Transactions on Mechatronics (2021).

[9] Bouali H, Akaichi J. "Comparative study of different classification techniques: heart disease use case", 2014 13th International conference on machine learning and applications. IEEE. p. 482–86.

[10] Mahmoud, S., Hussein, M., & Keshk, A. "Predicting Future Products Rate using Machine Learning Algorithms", *International Journal of Intelligent Systems & Applications*, *12*(5), 2020.

[11] Meyer, D., & Wien, F. T. "Support vector machines", *The Interface to libsvm in package e1071*, *28*(20), 597, 2015.

[12] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques", *IEEE Access*, *9*, 19304-19326, 2020.

[13] Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019). Heart disease prediction using machine learning. *International Journal of Research in Engineering, Science and Management*, *2*(2), 352-355.

[14] Prabu, S., Thiyaneswaran, B., Sujatha, M., Nalini, C., & Rajkumar, S. (2022). Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters. *Computer Systems Science & Engineering*, *43*(2).

[15] Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, *63*(2), 179-184.

[16] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, *77*, 5198-5219.

[17] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, 1-6.

[18] 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.

[19] 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.

[20] Rubini, P. E., Subasini, C. A., Katharine, A. V., Kumaresan, V., Kumar, S. G., & Nithya, T. M. (2021). A cardiovascular disease prediction using machine learning algorithms. *Annals of the Romanian Society for Cell Biology*, 904-912.

[21] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple Disease prediction using Machine learning algorithms. *Materials Today: Proceedings*, *80*, 3682-3685.

[22] Vayadande, K., Golawar, R., Khairnar, S., Dhiwar, A., Wakchoure, S., Bhoite, S., & Khadke, D. (2022, May). Heart Disease Prediction using Machine Learning and Deep Learning Algorithms. 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (pp. 393-401). IEEE.