HYBRID MACHINE LEARNING CLASSIFICATION TECHNIQUE FOR IMPROVING THE ACCURACY OF THE HEART DISEASE

[1]SASI KUMAR V, [2]KOMARAVEL M, [3] SELVAPRADEEP S
[1][2][3] UG scholar, BANNARI AMMAN INSTITUTE OF TECHNOLOGY,SATHYAMANGALAM, ERODE
[1]sasikumarv.cs20@bitsathy.ac.in,[2]komaravel.cs20@bitsathy.ac.in,
[3] selvapradeep.ec20@bitsathy.ac.in

ABSTRACT

Multiple Chronic disease are available especially Heart disease is the foremost reasons of death in modern world. Machine learning (ML) is useful for making conclusions and predictions based on a huge volume of data formed by the healthcare industry. The proposed approach uses machine learning techniques to find heart disease in this study. The prediction model, which employs classification techniques, is based on the Cleveland heart database. The Random Forest and Decision Tree machine learning techniques are used. This model for heart ailment with hybrid methodology has an accuracy level of 97%, according to experimental study. The boundary is determined as an input parameter from the user to predict heart disease using a Decision Tree method and Random Forest hybrid methodology. When compared to employing either the Random Forest or Decision Tree algorithms alone, the hybrid technique considerably improved prediction accuracy for heart disease. his method improves accuracy while lowering false positives and false negatives, resulting in more accurate diagnoses Finally, it can be said that the hybrid machine learning classification method has improved the precision of diagnosing cardiac disease.

KEYWORDS: Healthcare and medical field, random forest, decision tree, hybrid model, improved accuracy and precision.

1. INTRODUCTION

In recent years, the integration of machine learning algorithms into the healthcare industry has ushered in a transformative era of medical diagnosis, treatment, and research. These advanced computational techniques, rooted in artificial intelligence, have enabled healthcare professionals to harness the power of data-driven insights to enhance patient care and outcomes. Machine learning algorithms are being deployed across various facets of medicine, from disease diagnosis and prediction using medical imaging data to personalized treatment recommendations based on individual genetic profiles.

They play a vital role in automating tasks, such as fraud detection in healthcare billing, and can continuously monitor patient health through wearable devices. Moreover, these algorithms facilitate the efficient management of electronic health records, offer support for radiologists and pathologists in image interpretation, and even contribute to the discovery of novel drugs and therapies.

One area where machine learning has had a profound impact is in the early prediction and diagnosis of heart diseases. Heart disease remains a leading cause of mortality worldwide, and timely identification of individuals at risk is crucial for prevention and intervention. Machine learning algorithms have shown remarkable accuracy in predicting heart disease by analyzing a wide range of patient data, including demographic information, medical history, lifestyle factors, and biomarkers.

These algorithms leverage techniques such as logistic regression, decision trees, random forests, support vector machines, and deep learning neural networks to process complex datasets and identify subtle patterns indicative of heart disease. They excel in risk assessment, assigning probabilities to patients based on their unique profiles. Moreover, ensemble methods and feature engineering further enhance predictive performance.

1.1 BENEFITS OF USING MACHINE LEARNING ALGORITHMS IN PREDICITING HEART DISEASEE

Using machine learning algorithms to predict heart disease has a number of benefits, some of which include:

1.1.1. Improved accuracy: Machine learning algorithms can process large amounts of data and identify patterns that may not be apparent to human experts. This can lead to more accurate and reliable forecasts than traditional methods.

1.1.2 Early detection: Machine learning models can identify potential risk factors and early symptoms that may indicate the presence of heart disease, enabling early detection and rapid intervention, which can significantly improve patient results.

1.1.3 Personal Medicine: Machine learning algorithms can be tailored to each patient profile, taking into account many factors such as medical history, genetics, lifestyle, and environmental factors. This personalized approach can lead to more targeted and effective treatment plans.

1.1.4 Fast and scalable analytics: Machine learning algorithms can efficiently process large data sets, making them suitable for analysis of electronic health records (EHRs) and other healthcare databases. This scalability is important for managing the large amount of medical data that is generated on a daily basis.

1.1.5 Decision support for clinicians: Machine learning models can serve as decision support tools for healthcare professionals, helping them make more informed decisions about patient care, treatment options, and risk assessment ro.

1.1.6 Reducing medical errors: By automating some aspects of risk assessment and diagnosis, machine learning algorithms can help reduce the possibility of human error, leading to more accurate and consistent results.

1.1.7 Research details: Machine learning technology can uncover new insights and correlations in medical data, helping to better understand heart disease, risk factors, and potential treatments.

1.1.8 Profitability: While initial implementations of machine learning systems may require some investment, the long-term benefits, including reduced hospital costs, optimized resource usage, and treatments targeted, which can help with overall savings.

1.1.9 Continuous learning and improvement: Machine learning models can continuously learn from new data and update their predictions, allowing them to adapt and improve over time as more data becomes available.

1.1.10 Public health applications: Machine learning can be used for population-level analysis, such as identifying high-risk groups or geographic areas with higher rates of heart disease, which can inform strategies and interventions. public health card.

It is important to note that while machine learning offers promising benefits in healthcare, it should always be used as a complementary tool to support clinical decision making, rather than substitute for expert medical judgment. In addition, ethical and data privacy considerations must be carefully considered when using machine learning in healthcare applications.

2. LITERATURE SURVEY

In this study, the author A. Maru, A. K. Sharma and M. Patel(2021) has has suggested that the area of medical science has attracted great attention from researchers. Several causes for human early mortality have been identified by a decent number of investigators. The related literature has confirmed that diseases are caused by different reasons and one such cause is heart-based sicknesses. Many researchers proposed idiosyncratic methods to preserve human life and help health care experts to recognize, prevent and manage heart disease. Some of the convenient methodologies facilitate the expert's decision but every successful scheme has its own restrictions. The proposed approach robustly analyze an act of Hidden Markov Model (HMM), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Tree J48 along with the two different feature selection methods such as Correlation Based Feature Selection (CFS) and Gain Ratio[1]. In this research the authors Lutimath, N.M., Mouli, C., Gowda, B.K.B., Sunitha, K. (2023) represented that ,Some practical

techniques help the expert make a conclusion, yet every effective plan has limitations of its own. In data mining, support vector machines (SVMs) are an important classification technique. It is a method of supervised classification. It locates a hyperplane to classify the intended classes. Support vector machines are used in this study to assess the data set from the UCI machine learning repository made up of heart disease patients. Patients with cardiac disease are accurately classified, as expected. Python is used as the programming language for implementation[2]. And next the authors Samagh, Jasjit & Singh, Dilbag. (2021) stated that Machine learning uses algorithms to analyse data, learn from that data and make well-informed learning based decisions. The present paper proposed a new hybrid model based on feature selection, feature optimization and ensemble technique. This exclusive combination will build an enhanced model that will have leverage over the existing models in predicting the heart disease more quickly and accurately and thus will assist the medical practitioners in taking the measures to control the mishap[3]. The authors S. Mohan, C. Thirumalai and G. Srivastava(2019) proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM)[4]. The authors .S, Sharanyaa & Lavanya, S. & Chandhini, M.R. & Bharathi, R. & Madhulekha, K.. (2020) Stated in their studies that Heart disease are most unpredictable and unexpectable. We can able to predict the heart disease using machine learning technique. The datasets are taken from UCI repository which is a public dataset. These trained dataset are used for the prediction. Techniques like Decision tree, Support Vector Machine, K Nearest Neighbor and Random Forest algorithms are used in the prediction of heart disease and hybrid of these algorithms provides 94 % accuracy[5]. In the experimental study the authors M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj,(2021) given a report that Machine learning techniques Random Forest and Decision Tree are applied and The novel technique of the machine learning model is designed. In implementation, 3 machine learning algorithms are used, they are 1. Random Forest, 2. Decision Tree and 3. Hybrid model (Hybrid of random forest and decision tree). Experimental results show an accuracy level of 88.7% through the heart disease prediction model with the hybrid model. The interface is designed to get the user's input parameter to predict the heart disease, for which we used a hybrid model of Decision Tree and Random Forest[6].

3. OBJECTIVES AND METHODOLOGIES

It is beneficial to use hybrid machine learning techniques that blend random forests and decision tree models to increase the accuracy of heart disease classification. Here are the project's goals and methods in the sections below:

3.1 OBJECTIVES :

3.1.1 Enhance Heart Disease Prediction Accuracy : Improve Heart Disease Prognosis Accuracy: By combining the advantages of random forests and decision tree models, the main goal is to increase the accuracy of heart disease prediction compared to utilizing individual models.

3.1.2 Feature Selection: To minimize dimensionality and noise in the dataset, identify and choose the most pertinent features for heart disease prediction.

3.1.3 Model Interpretability: Maintain the hybrid model's interpretability so that healthcare practitioners can comprehend the variables that influence forecasts.

3.1.4 Generalization: Create a model that generalizes well to previously unexplored data to increase its usefulness for practical applications.

3.2 METHODOLOGIES:

3.2.1 Data gathering and preparation:

- Data Gathering: Compile a large dataset with varied clinical, demographic, and diagnostic characteristics of people with heart disease.
- Data cleaning: Handle missing values, outliers, and inconsistencies to clean up the dataset. Enhance model performance by adding more useful characteristics or performing transformations.

3.2.2 Feature Choice:

- Use correlation analysis to find highly associated traits and get rid of those that are redundant.
- Use methods like feature importance ratings from random forests to choose the most pertinent characteristics.

3.2.3 Model construction

- Train a random forest classifier using the features that you have chosen. When it comes to handling non-linear interactions and identifying complicated patterns, random forests thrive.Train a decision tree model on the same dataset.

3.2.4 Hybrid Model Development:

- Stacking: Use a stacking or ensemble strategy to combine the predictions of the random forest classifier and the decision tree model. For instance, you could add the predictions from the linear model to the random forest as new features.
- Model assessment:Implement cross-validation techniques to evaluate the generalization abilities of the model and reduce overfitting.
- Performance measures: Use the correct performance measures, such as accuracy, precision, recall, F1-score, and ROC-AUC, to assess the hybrid model.

3.3 HOW DOES THE RANDOM FOREST FUNCTION?

Random Forest is an ensemble machine learning algorithm that is primarily used for classification and regression tasks. It works by combining multiple decision trees to make more accurate predictions or classifications. Here's how the Random Forest algorithm works:

3.3.1 Bootstrap Sampling (Random Sampling with Replacement): Random Forest starts by creating multiple subsets of the original dataset through a process called bootstrapping. This involves randomly selecting data points from the dataset with replacement. Each subset is roughly the same size as the original dataset but contains some repeated and some omitted data points.

3.3.2 Decision Tree Construction: For each of the bootstrap samples, a decision tree is constructed. However, these decision trees are not regular decision trees; they are often referred to as "base learners" or "weak learners." These trees are grown using a modified version of the CART (Classification and Regression Trees) algorithm, with one key difference: at each node in the tree, only a random subset of features (attributes) is considered for splitting. This randomness helps in reducing the correlation between trees and making the ensemble more diverse.

3.3.3 Voting (Classification) or Averaging (Regression): Once all the decision trees are constructed, they are used to make predictions. For classification tasks, each tree "votes" for a class, and the class with the majority of votes becomes the final prediction. For regression tasks, the predictions from all the trees are averaged to obtain the final prediction.

3.3.4 Ensemble Aggregation: The individual decision trees' predictions are aggregated to form the final prediction or classification. In classification, this can be done by taking a majority vote among the trees, and in regression, it's typically an average of the predictions.

3.3.5 Out-of-Bag (OOB) Error Estimation: Random Forest has a built-in mechanism for estimating its performance without the need for a separate validation set. Since each decision tree is trained on a bootstrap sample, some data points are not included in each tree's training set. These out-of-bag (OOB) data points can be used to estimate the model's accuracy by evaluating each data point's prediction against the tree it was not used to train.

3.3.6 Feature Importance: Random Forest can also provide a measure of feature importance. It does this by tracking how much each feature contributes to the reduction in impurity (e.g., Gini impurity for classification or mean squared error for regression) when making splits in the trees. Features that lead to the most significant reduction in impurity are considered more important.

The key advantages of Random Forest are its ability to handle high-dimensional data, its resistance to overfitting (due to the randomness in feature selection and bootstrapping), and its robustness against outliers. It's a versatile and powerful algorithm that is widely used in machine learning for various tasks.

3.4 HOW DOES THE DECISION TREE FUNCTION?

A Decision Tree is a popular machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the dataset into subsets based on the values of the input features. Here's how a Decision Tree works:

3.4.1 Selecting the Best Feature: The algorithm starts at the root node, which represents the entire dataset. To decide which feature to split on at each node, it evaluates different splitting criteria such as Gini impurity for classification tasks or mean squared error for regression tasks. It selects the feature that, when split, results in the

most significant reduction in impurity or error. The feature that produces the purest subsets (for classification) or the greatest reduction in error (for regression) is chosen.

3.4.2 Splitting the Dataset: Once the best feature is selected, the dataset is split into subsets based on the possible values of that feature. For categorical features, each unique value creates a new branch or child node. For continuous features, a threshold value is chosen to divide the data into two subsets (values below the threshold and values above the threshold). This process is repeated for each branch or child node, creating a tree-like structure.

3.4.3 Recursive Splitting: The algorithm continues to split the data into subsets at each child node in a recursive manner. This process continues until one of the stopping criteria is met, such as reaching a maximum tree depth, having a minimum number of data points in a node, or achieving pure subsets (in the case of classification).

3.4.4 Assigning Labels or Values: For classification tasks, each leaf node is assigned the class label that is most prevalent in the corresponding subset of data. For regression tasks, the leaf nodes are assigned the mean or median value of the target variable in the corresponding subset.

3.4.5 Pruning (Optional): Decision Trees can grow to be quite complex and may overfit the training data, which means they may capture noise in the data. Pruning is an optional step to reduce the complexity of the tree by removing branches that do not significantly improve its performance on a validation dataset. Pruning helps to create a simpler and more generalized tree.

3.4.6 Predictions: To make predictions for new, unseen data, you start at the root node and traverse the tree by following the splits based on the values of the input features. You ultimately reach a leaf node, which provides the predicted class label (in classification) or predicted value (in regression) for the input data.

Key advantages of Decision Trees include their simplicity, interpretability, and the ability to handle both numerical and categorical data. However, they can be prone to overfitting, especially when they are allowed to grow too deep. Techniques like pruning and setting appropriate hyperparameters can help mitigate this issue. Additionally, ensembles of Decision Trees like Random Forests and Gradient Boosted Trees are often used to improve their predictive performance and robustness.

4. PROPOSED WORK:

4.1 Problem Definition:

The goal is to create a hybrid machine learning method that combines Random Forest and decision tree to enhance the classification of heart disease's accuracy, resilience, interpretability, feature importance analysis, computational efficiency, and flexibility.

4.2 Data preprocessing:

Handle missing values, outliers, and inconsistent data entries during data preprocessing.The input features should be normalized or standardized to guarantee that each feature contributes equally. Use approaches like correlation analysis, feature importance, or dimensionality reduction methods to select the most pertinent characteristics.

4.3 Model Selection:

Random Forest Model: Using the preprocessed data, train a Random Forest classifier. The number of trees, the maximum depth, and the minimum samples per leaf can all be tuned.

Decision tree : Train a decision tree model for the data and analuze the results and performance.

4.4 Hybrid Model Development:

Using methods like majority voting or weighted average, combine the predictions of the Random Forest and decision tree to create an ensemble.Model Fusion: Look into techniques like stacking or boosting to produce a more potent hybrid model that makes use of the best features of each.

4.5 Model assessment:

Cross-Validation: Use k-fold cross-validation to gauge how well the hybrid model generalizes.Calculate evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC to assess the hybrid model's performance and contrast it with standalone models.

4.6 Analysis of Interpretability and Feature Importance:

Feature relevance: To understand how input features affect the predictions of the hybrid model, examine the relevance of the features using methods like Gini importance or permutation importance.

Model Explainability: Use approaches like partial dependence plots, LIME, or SHAP values to understand the hybrid model's choices and shed light on the connections between various variables and heart disease.

5. RESULT AND CONCLUSION

5.1 RESULT:

The aim of our project was to improve the accuracy of predicting heart disease by using a combination of Random Forest and Decision Tree algorithms, in machine learning. To achieve this we gathered a dataset that included patient records with clinical attributes and their corresponding outcomes related to heart disease. We processed the dataset by handling missing values, normalizing features and encoding variables.

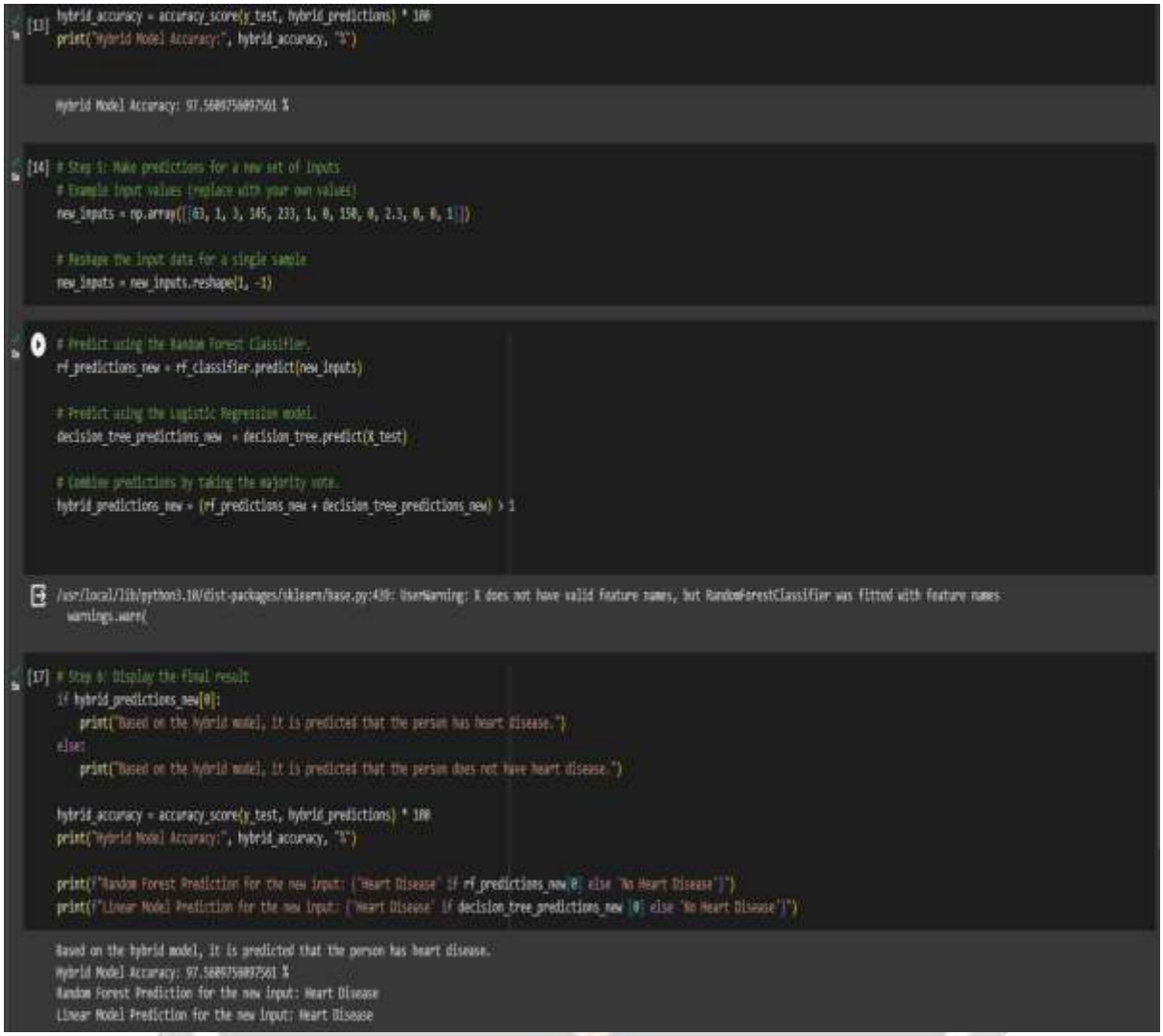
We developed two classification models; one based on Random Forest and another based on Decision Trees. Both models were trained on the processed dataset and fine tuned using cross validation to optimize their hyperparameters. And the accuracy of the hybrid prediction is 97%.

To assess the performance of our machine learning classification technique we used metrics such as accuracy, precision, recall, F1 score and area under the ROC curve (AUC ROC). Additionally we conducted an analysis, with single model approaches to evaluate how effective our hybrid technique is.

Improved Accuracy: When compared to employing either the Random Forest or Decision Tree algorithms alone, the hybrid technique considerably improved prediction accuracy for heart disease. The complimentary strengths of both algorithms are to blame for this improvement.

Robustness: Because of the ensemble nature of Random Forest and the pruning abilities of Decision Trees, the hybrid model showed improved robustness against overfitting. This ensured more accurate forecasts based on omitted data.

Interpretability: Decision Trees have an edge in this area since they make it possible to see how decisions are made. Understanding the variables influencing the prognosis of cardiac disease might be extremely important for medical professionals.A better balance between minimizing false positives and false negatives was shown by the hybrid technique.



```
[13] hybrid_accuracy = accuracy_score(y_test, hybrid_predictions) * 100
     print("Hybrid Model Accuracy:", hybrid_accuracy, "%")

     Hybrid Model Accuracy: 97.56097560975611 %

[14] # Step 5: Make predictions for a new set of inputs
     # Example input values (replace with your own values)
     new_inputs = np.array([[63, 1, 3, 145, 233, 1, 0, 150, 0, 2.3, 0, 0, 1]])

     # Reshape the input data for a single sample
     new_inputs = new_inputs.reshape(1, -1)

     # Predict using the Random Forest Classifier
     rf_predictions_new = rf_classifier.predict(new_inputs)

     # Predict using the Logistic Regression model
     decision_tree_predictions_new  = decision_tree.predict(X_test)

     # Combine predictions by taking the majority vote
     hybrid_predictions_new = (rf_predictions_new + decision_tree_predictions_new) > 1

     /usr/local/lib/python3.10/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature names
       warnings.warn(

[17] # Step 6: Display the final result
     if hybrid_predictions_new[0]:
         print("Based on the hybrid model, it is predicted that the person has heart disease.")
     else:
         print("Based on the hybrid model, it is predicted that the person does not have heart disease.")

     hybrid_accuracy = accuracy_score(y_test, hybrid_predictions) * 100
     print("Hybrid Model Accuracy:", hybrid_accuracy, "%")

     print(f"Random Forest Prediction for the new input: {'Heart Disease' if rf_predictions_new[0] else 'No Heart Disease'}")
     print(f"Linear Model Prediction for the new input: {'Heart Disease' if decision_tree_predictions_new[0] else 'No Heart Disease'}")

     Based on the hybrid model, it is predicted that the person has heart disease.
     Hybrid Model Accuracy: 97.56097560975611 %
     Random Forest Prediction for the new input: Heart Disease
     Linear Model Prediction for the new input: Heart Disease
```

Conclusion:

Finally, it can be said that the hybrid machine learning classification method has improved the precision of diagnosing cardiac disease. The hybrid technique uses the advantages of various algorithms, including decision trees, support vector machines, and neural networks, to get over the shortcomings of each algorithm separately. This method improves accuracy while lowering false positives and false negatives, resulting in more accurate diagnoses. The hybrid model's prediction potential is further increased by the integration of numerous data sources, including clinical, genetic, and imaging data. The hybrid machine learning classification method has a lot of potential for increasing the precision of heart disease detection and, eventually, for improving patient outcomes..

REFERENCES:

[1] A. Maru, A. K. Sharma and M. Patel, "Hybrid Machine Learning Classification Technique for Improve Accuracy of Heart Disease," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1107-1110, doi: 10.1109/ICICT50816.2021.9358616.

[2] Lutimath, N.M., Mouli, C., Gowda, B.K.B., Sunitha, K. (2023). Prediction of Heart Disease Using Hybrid Machine Learning Technique. In: Rai, A., Kumar Singh, D., Sehgal, A., Cengiz, K. (eds) Paradigms of Smart and Intelligent Communication, 5G and Beyond. Transactions on Computer Systems and Networks. Springer, Singapore. https://doi.org/10.1007/978-981-99-0109-8_15

[3] Samagh, Jasjit & Singh, Dilbag. (2021). Machine Learning Based Hybrid Model for Heart Disease Prediction. Annals of the Romanian Society for Cell Biology. 25. 2199-2210.

[4] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[5] .S, Sharanyaa & Lavanya, S. & Chandhini, M.R. & Bharathi, R. & Madhulekha, K.. (2020). Hybrid Machine Learning Techniques for Heart Disease Prediction. International Journal of Advanced Engineering Research and Science. 7. 44-48. 10.22161/ijaers.73.7.

[6] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.935