International Journal of Advanced Research in Engineering Technology & Science
ISSN: 2349-2819
Email: editor@ijarets.org Volume-10 Issue-12 December-2023 www.ijarets.org

INVESTIGATION OF HEART DISEASE FORECASTING THROUGH DATA MINING AND MACHINE LEARNING METHODS

Ankush Goyal
Research Scholar, School of Technology and Computer Science
Glocal University, Mirzapur Pole Saharanpur (U. P.) India.

Dr. Manoj Kumar
Research Supervisor, School of Technology and Computer Science
Glocal University, Mirzapur Pole Saharanpur (U.P) India.

***Abstract*:**
The modern technology can help in early detection of such disease, reducing fatality risks. Modern technology like Machine Learning can be used for same, which uses other health data parameters to predict chances of any heart disease. Heart disease has been increasing, which can be caused due to modern lifestyle, shift in dietary patterns, and other factor as population aging. This study uses datasets readily available online to train and test machine learning algorithms, as how effective can it get with processing. It studies the difference in scores of algorithms before and after using feature selection.

Keywords: information gain, Anova, feature selection. machine learning, uci, preprocessing, normalization, xgb, regression imputer,

I. INTRODUCTION

Heart disease describes a range of conditions that affect the heart. Heart diseases include: Blood vessel disease, such as coronary artery disease, Irregular heartbeats (arrhythmias), Heart problems you're born with (congenital heart defects), Disease of the heart muscle, heart valve diseases. According to the World Health Organization (WHO), an estimated 17.9 million people died from cardiovascular diseases (CVDs) in 2019, representing 32% of all global deaths. The World Health Organization (WHO) reports that in 2020, approximately 19.1 million deaths were attributed to CVDs globally.In 2016, India reported 63% of total deaths due to NCDs, of which 27% were attributed to CVDs. CVDs also account for 45% of deaths in the 40–69 years age group. According to Statista, as per the results of a large-scale survey conducted across India, a majority of the people with heart problems in India in 2020 were aged between 45-54 years. Symptoms that may suggest a heart or blood vessel problem are shortness of breath, chest pain, chest pressure, heart palpitations, dizziness, sweating, numbness and weakness.Machine Learning algorithms play an important role for prediction of vast variety of things from historical data and its analysis. With upcoming developments in data analytics and machine learning techniques, the historical records of patients can be used to predict if person may or may not be suffering from heart condition.Machine learning techniques doesn't need any specific calculation unique for heart disease prediction. It comes with different computation algorithms which can effectively predict outcomes. Many individual algorithms may not effectively predict expected outcome. Accuracy can differ for various algorithm for different features. Features are columns in datasets, i.e., various factors that suggest presence of heart disease for an individual. To overcome such anomalies, Machine Learning has technique called ensemble. It, as name suggests, ensembles different algorithms in Machine Learning, and produce result, which may have accuracy better than individual algorithms.The heart disease dataset from the University of California, Irvine (UCI) Machine Learning Repository is a comprehensive dataset for heart disease prediction. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. These 11 common features are:

1) Age
2) Sex
3) Chest pain
4) Resting blood pressure

5) Serum cholesterol mg/dl
6) Fasting blood sugar > 120
7) Maximum heart rate achieved
8) Exercise induced angina
9) Resting ECG results
10) oldpeak = ST depression induced by exercise relative to rest
11) the slope of the peak exercise ST segment

The dataset mentioned above is most used dataset, which is freely available over Kaggle. It consists of Cleveland, Hungary, Switzerland, Long Beach VA and Statlog Dataset. Of these, Cleveland dataset is most widely used dataset.But, feeding dataset directly to Machine Learning Algorithms does not provide expected results. For achieving expected result with higher accuracy, the dataset should be processed before feeding them to some algorithm. It is called preprocessing in Machine Learning. Preprocessing data is a fundamental stage in data mining to improve data efficiency. The steps involved in preprocessing datasets are:Data cleaning: Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in data. This process involves several steps such as removing duplicates, filling in missing values, correcting errors, and dealing with outliers. By cleaning your data, you can ensure that your analysis is based on accurate and reliable information.Data transformation: This involves transforming the raw dataset into an understandable format. It involves cleaning, filtering, and organizing data so that it can be analyzed effectively.Data reduction: It is the process of reducing the size of a dataset by removing redundant features or instances. It involves removing data that is no longer useful or relevant to your analysis. There are several techniques for data reduction such as principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE). These techniques can help to reduce the dimensionality of a dataset while preserving the most important information and maintaining variance in data.Feature scaling: Feature scaling is a technique used to standardize the range of independent variables or features of data. It is a method used to normalize the range of independent variables or features of data. In layman's terms, feature scaling is like converting all the data into a common unit so that it can be compared on the same scale.

II. LITERATURE REVIEW

Comprehensive Dataset with 1190 records and Cardiovascular dataset with 70,000 records. This study utilized Feature Selection (FS) and Feature Extraction (FE)techniques such as Chi Square, AnoVa and PCA (Principal Component Analysis) to eliminate redundant and irrelevant attributes to create subset of features.

This technique is called novel Quine McClusky Binary Classifier (QMBC). Chi square and Anova identify 10 features, and PCA then creates data subset with 9 topmost features. No cross validation was used, rather decide to use threThe preprocessing of dataset is very important for increasing accuracy of results from algorithm. Feature Selection is part of preprocessing dataset. Also known as attribute selection, feature selection helps select most prominent features for increasing accuracy and overall performance of algorithm. It also helps deal with overfitting issues. A. M. Quadri et al [1] proposed a novel Principal Component Heart Failure (PCHF) Feature Engineering technique to select the most prominent features to enhance performance. This technique was proposed after using nine algorithms for comparison. The dataset used had 14 features that are important for heart disease prediction. To get the maximum accuracy scores, the proposed PCHF process was improved by developing a new feature set as an innovation. The eight best-fit features serve as the foundation for the newly formed dataset. In contrast to the original features, the PCHF approach chooses the eight distinct dataset characteristics with the maximum variance. A linear transformation mechanism was employed in the suggested PCHF feature selection method. In this study, the newly developed characteristics using the PCHF technique were highly accurate in predicting heart failure. They also applied k-fold cross validation technique about 10 folds.

Abdallah et al [3] used Synthetic Minority Oversampling Technique abbreviated as SMOTE to handle imbalance distribution. They used Hyperparameter Optimization for purpose of finding best hyperparameter for ML classifier with SMOTE, for six different classifiers. Machine Learning algorithms used are SVM, SGD, k-NN, Extra Trees, XG Boost, and LR. They concluded that tree-based models were more effective in achieving quality results. Thus, SMOTE with Extra Trees optimized by Hyperband (HB). Ramdas Kapila et al [2] used ensemble technique for increasing performance of machine learning. It used seven standalone Machine Learning model, and voting classifier was used. This approach used three datasets including mentioned Cleveland Dataset, whole UCI comprehensive heart disease dataset and Cardiovascular dataset. Cleveland dataset with 303 records, e datasets only using train and test sets for evaluation. This is because the authors focus on using ensemble approach rather than standalone model and due to availability of enough data. It used python with Jupyter Notebook for implementation. Libraries used were Pandas, NumPy,

Matplotlib, Seaborn, Warnings, and Scikit-Learn.

Aishwarya et al [4] discusses the use of Explainable Artificial Intelligence (XAI) and Random Forest in interpreting cardiovascular disease, which is a novel approach in the medical field. The paper also highlights the potential societal, ethical, and safety implications of using AI in sensitive domains. Dataset used was from UCI Repository with 918 instances with 12 features. LIME and SHAP are feature-based model explainability techniques used in the context of cardiovascular disease evaluation in the paper.

LIME stands for Local Interpretable Model-agnostic Explanations, which is an explainable AI technique used to interpret the predictions made by a machine learning model at the local level. It provides an explanation of how a particular prediction is made by the model by highlighting the important features that contributed to the prediction.

SHAP stands for Shapley Additive Explanations, which is another feature-based model explainability technique that assigns each feature an importance value for a particular prediction. Both LIME and SHAP are used to increase transparency and trust in the machine learning model by providing interpretable explanations for its predictions.

The algorithms evaluated are Random Forest, Decision Tree, Logistic Regression, Neighbors, and SVM. The results show that Random Forest has the highest accuracy. The study also includes visualizations of dataset features, scatter plots, and two-dimensional histograms.

Kuldeep V. et al [5] proposes a model for predicting and analyzing heart disease using machine learning and deep learning algorithms. The dataset used for model is UCI Cleveland dataset, with 303 records and 14 different attributes. Ten algorithms in all used in this study, namely Logistic Regression, NB, k-NN, SVM, MLP, ANN, DT, RF, XG Boost and Cat Boost.

Abdulwahab Ali et al [6] used in this paper a deep learning framework for heart disease prediction. The proposed methodology is divided into two stages: data acquisition and preprocessing, and classification models and proposed model. The proposed framework is based on Learning Vector Quantization (LVQ) algorithm for the prediction and analysis. Deep neural networks with numerous layers are used to obtain great analytical accuracy. The Dataset used is UCI repository dataset (comprehensive), i.e., one with 1190 instances. 10-fold cross validation is performed in this study. Final result is provided as average accuracy of all fold. Algorithms used in this study include k-Nearest Neighbors, Gaussian Process, Linear SVM, Decision Tree, Naive Bayes, QDA, AdaBoost, Bagging, Boosting, and Deep Neural Networks.

Learning Vector Quantization (LVQ) is a type of artificial neural network that is used for classification problems. It is a supervised learning algorithm that is used to classify data into different categories. The LVQ algorithm works by creating a set of prototypes, which are representative vectors for each class. These prototypes are then adjusted during the training process to better represent the data.

Aishwarya D. et al [7] proposed that using machine learning algorithms with ensemble learning, feature selection and bio-medical test values can help classify heart diseases more efficiently. Their study was mainly based on importance of preprocessing steps as Checking missing values, Label encoding and Data Scaling or Normalization. The dataset used for this purpose was heart disease dataset from UCI Repository. With and without preprocessing the Random Forest gave highest accuracy for dataset. After preprocessing the accuracy of Algorithms improved. The importance of preprocessing is truly displayed by SVM algorithm. The error rate for SVM without preprocessing was approximately half cent percent, which decreased to less than 15percent. Only algorithm that showed reversed result was Decision Tree, i.e., the accuracy actually decreased after preprocessing the dataset.

Year	Author	Objective	Technique	Dataset	Accuracy	Implementation Tool
2022	Abdallah et al [3]	Prediction of heart disease, and severity level	SMOTE + ET + HB	Cleavelan d and Statlog	99.20% & 98.52%	-
2023	Ramdas K. et al [2]	Prediction of Heart Disease	QMBC + Chi-square + PCA	CVD Dataset	99.92%	Python – Jupyter 6.3.0
2023	Aishwarya et al [4]	Prediction and Interpretation of CVD using RF and Explainable AI	LIME + SHAP + Random Forest	UCI Repositor y	87.5%	What-If (WIT)
2022	Kuldeep V. et al [5]	Heart Disease prediction using ML and DL Algorithms	RF / LR / XGBoost	Cleveland Dataset	88.52%	Python/R

2023	A. A. Almazroi et al [6]	Heart Disease prediction using Deep Learning	LVQ	UCI Repository	>80% ~83.03%	Python, keras, TensorFlow
2023	Aishwarya et al [4]	Analysis of CVD using ML algorithms.	-	UCI Repository	88.04%	Python

Table 1: Comparison of various Studies carried before and their accuracy scores.

III. DATA OVERVIEW AND PREPROCESSING

Datasets features and description.

Feature name	Description
Age	Age of the patient in years.
Sex	Gender of the patient (M = male; F = female)
ChestPainType	Type of chest pain experienced by the patient
RestingBP	Resting blood pressure of the patient in mm Hg
Cholesterol	Cholesterol levels of the patient in mg/dl.
FastingBS	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
RestingECG	Resting electrocardiographic results of the patient.
MaxHR	Maximum heart rate achieved by the patient.
ExerciseAngina	Exercise-induced angina
Oldpeak	ST depression induced by exercise relative to rest
ST_Slope	Slope of the peak exercise ST segment.
HeartDisease	Presence of heart disease (1 = present; 0 = absent)

Table 2 : Various features and their descriptions

A. Preprocessing

This [8] dataset has 1,759 null values for features trestbps, chol, thalch, restecg, fbs oldpeak, slope, thal, exang, ca; [59,30,55,2,90,62,309,486,55,611] respectively. Five of features having null value are of categorical type and other five having numerical data.

Seven categorical features are present in dataset. These are to be converted into numerical data. We used Label Encoding for this purpose. The label encoder comes from sklearn. Preprocessing package of sklearn library in Python.

Handling Null values: The null values in features of Categorical types were replaced with median value of entire feature, while for numerical we used Regression Imputer to fill null spaces.

Regression Imputer: It is a technique used in data preprocessing to fill in missing values in a dataset using regression analysis. A regression imputer estimates the missing values based on the relationships between the missing feature and other features in the dataset. sklearn.experimental and sklearn.impute libraries were used for this purpose.

Z-score: It is a statistical measure that helps you understand how far away a particular data point is from the average of a group of data points. It was used to handle outliers in UCI dataset. It indicates whether a value is typical or unusual compared to the other values in a dataset.

Normalization: It is process of adjusting values measured on different scales to a common scale, typically between 0 and 1. The goal of normalization is to bring all the features of a dataset to a similar range, making it easier to compare different variables and eliminating the effects of scale differences in the data. The features resting bp, cholesterol, age, maxhr, oldpeak,ca were normalized.

B. Data Splitting

In the context of machine learning, the integrity of data is critical, and the division of data into training and testing sets constitutes a crucial phase in model development. Mishandling the data division process can lead to issues such as underfitting or overfitting, potentially resulting in biased outcomes. This section underscores the significance of distinguishing the train and test sets without creating a separate validation set. Employing a standard approach, the dataset was divided into training and testing datasets using an

80:20 ratio, with 80% of the data allocated for training and the remaining 20% for testing. This enabled a dependable evaluation of the models.

The data splitting in python was done using train_test_split module of sklearn.model_selection library of python.

The training set functioned as the cornerstone for model training, facilitating the recognition of patterns and relationships within the analyzed data. In contrast, the test set served as an impartial metric for assessing the model's performance on unseen data, offering insights into its capacity to generalize and maintain robustness.

IV. MACHINE LEARNING TECHNIQUES

A. Random forest (RF)

Random Forest (RF) is a popular supervised machine learning technique used for solving classification and regression problems. This method is particularly effective because it leverages multiple decision trees to tackle complex issues and improve efficiency. In essence, an RF classifier takes many decision trees and combines their outputs based on different subsets of the dataset. This ensemble approach enhances predictive accuracy. The more decision trees, the better the accuracy, but there's a point where adding more doesn't significantly improve results. One of the key strengths of Random Forest is its ability to prevent overfitting, a common problem where a model fits the training data too closely and struggles to generalize to new data. Instead, Random Forest strikes a balance between fitting the data and maintaining high performance, making it a valuable tool in the world of machine learning.

$$\hat{y}_i = \frac{1}{M}\sum_{j=1}^{M} f_j(x_i)$$

Where $\hat{y}_i$ is predicted output for observation *i*, M is the number of trees, f_j is j^{th} decision tree, x_i is the input vector for observation *i*.

B. Extreme Gradient Boosting

(XGB) is a supervised machine learning technique used for tasks like classification and regression. It's part of the ensemble learning family, which involves combining multiple machine learning methods for better results. XGB is unique in that it combines many decision trees in a clever way. It doesn't create deep trees; instead, it trains them step by step. Each new tree is designed to correct the errors made by the previous one, which makes the model progressively better. The final prediction of XGB isn't just a simple average of the individual tree predictions. It's a weighted average, giving more importance to the trees that perform better. This ensemble approach not only improves accuracy but also helps prevent underfitting, where the model is too simplistic. In XGB, the model's performance is continuously refined using a technique called the gradient descent algorithm. This method adjusts the model's parameters to minimize errors, leading to a more accurate and efficient model. XGB is a valuable tool in machine learning, particularly when high performance and fine-tuning are essential.

$$\hat{y}_i = \sum k = 1^k f_k(x_i), f_k \in f$$

where $\hat{y}_i$ represents the predicted value for the *i*-th instance, f_k k-th represents the k-th weak learner added to the ensemble.

C. Logistic regression (LR)

Logistic regression (LR) is a widely applied supervised machine learning method capable of handling both regression and classification tasks. It's particularly well-suited for scenarios where we want to predict outcomes using probability. In LR, the focus is on binary classification, meaning we're interested in predicting one of two possible results. To do this, LR requires that the class variables are binary, mirroring the specific targets in the dataset you mentioned. For instance, if we take a dataset related to heart health, logistic regression can be used to predict the probability of heart failure. In this context, we often use a binary system: '0' might represent individuals with no risk of heart failure, while '1' signifies those at risk of heart issues in the dataset. Logistic regression is a valuable tool for making such binary predictions based on probability assessments.

$$p(y = 1|x) = \frac{1}{1 + e^{-z}}$$

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where $p(y = 1|x)$ represents the probability of a binary outcome variable y taking the value 1, given the predictor variable(s) x. The function e^{-z} is the logistic function, which maps any real value z to the range [0,1].

D. The Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a widely used supervised learning method, suitable for both classification and regression problems. Its primary goal is to establish optimal decision thresholds. SVM achieves this by creating a crucial decision boundary called a hyperplane, which effectively separates different classes within a multidimensional space. What makes this hyperplane particularly valuable is its ability to simplify the task of assigning new data points to their correct categories. SVM selects support vectors at the extreme edges of the data distribution to define this hyperplane. These support vectors are essentially the most critical data points for class separation, which is why the method is named a "support vector machine." This approach ensures that SVM can accurately categorize new data points by considering these pivotal support vectors that define the hyperplane.

$$\vec{w} \cdot \vec{x} + b = 0$$

where $\vec{w}$ is the weight vector, $\vec{x}$ is the input vector, and *b* is the bias term.

E. Decision Tree (DT)

Decision Trees are essentially tree-like structures at the core of the DT technique. These trees comprise multiple levels of nodes, with top-level nodes known as root or parent nodes, and the lower ones are termed child nodes. Decision Trees are often the preferred choice for managing extensive medical datasets due to their simplicity and ease of use. In this method, data is structured in a tree format, where internal nodes represent dataset attributes, branches dictate decision-making processes, and leaf nodes define the ultimate target outcomes. The Decision Tree algorithm divides the dataset into these nodes using criteria like the Gini index and entropy functions, helping determine the most effective decision criteria at each node.

M

$$f(x) = \sum_{m=1} c_m 1(x \in Rm)$$

where f(x) is the predicted output for input x, M is the number of leaf nodes in the tree, Rm is the region of input space corresponding to the m-th leaf node, c_m is the prediction value associated with the m-th leaf node.

F. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised machine learning technique used mainly for classifying and making predictions. What makes KNN unique is that it doesn't assume anything specific about the data—it's a non-parametric approach. In KNN, a new data point is assigned to a category by comparing its similarity to existing categories. Essentially, it assumes that the new data point is similar to those data points it's closest to. KNN typically measures this similarity using the Euclidean distance metric, which calculates how close or far apart data points are in a multidimensional space. In simple terms, KNN is a flexible method that makes decisions based on the closeness of data points. It offers a straightforward way to classify and predict outcomes without imposing rigid assumptions about the underlying data distribution.

$$y_q = y_{i1}, y_{i2}, \ldots\ldots\ldots, y_{ik}$$

where y_q represents the predicted label for a given query point, and *i1,i2,* , ik represent the indices of the k nearest neighbors of the query point.

V. FEATURE SELECTION (FS) STRATEGY

A. Chi-Square

Chi-Square is a statistical method employed in machine learning to identify the most pertinent attributes for a given task. It operates by calculating Chi-Square scores for each attribute, and then selecting the attributes with the highest scores. This is the fundamental process behind Chi-Square technique. In essence, if a feature is found to be independent of the target attribute, it is excluded from consideration. Conversely, if a feature demonstrates a relatively high Chi-Square (Chi2) score, it is regarded as more relevant to the target attribute. By focusing on the most informative attributes, this approach can substantially enhance performance while simplifying the dataset by reducing unnecessary complexity.

B. Information Gain

Information gain is a concept used in the field of machine learning and decision tree algorithms to measure the effectiveness of attributes in classifying data. It is particularly employed in the construction of decision trees to determine the sequence of attributes for splitting data. The concept of information gain is based on the principle of entropy from information theory. In the context of decision trees, information gain is used to quantify the reduction in uncertainty or randomness that results from partitioning a dataset based on a specific attribute. Higher information gain implies that a particular attribute is more effective at classifying the data. Mathematically, information gain is often calculated using the formula:

Information Gain = Entropy before split−Weighted average of entropies after split where entropy is a measure of the amount of uncertainty or randomness in the data. By selecting the attribute with the highest information gain at each step, decision tree algorithms can efficiently divide the dataset and create an optimal tree structure for classification tasks.

Mutual Information classifier calculates same value in python library sklearn.feature_selection. Scores of each feature in both datasets are used to consider important features. 11 best features in both cases were chosen.

Feature Names	Mutual info Values
ca	0.442096
cp	0.134812
oldpeak	0.123062
Sex	0.073648
Chol	0.070954
Age	0.066399
Thalch	0.063611
Rectecg	0.059522
exang	0.055868
Trestbps	0.048988
Thal	0.045427
slope	0.027746
fbs	0.021604

Table 3: Mutual Info Values of Features

C. ANOVA

ANOVA, which stands for Analysis of Variance, is a statistical technique employed to evaluate the significance of differences among groups or categories within a dataset. It's particularly useful when you want to understand if these differences are meaningful or merely the result of random variation. The F-test score, an integral part of ANOVA, is used to gauge the extent to which the variance (the spread or variability) in the target variable can be attributed to the variance in a specific feature or group. By calculating the F-test score, ANOVA enables you to make statistically informed conclusions about whether the differences you observe are likely to be meaningful or if they could have occurred by chance. It's a valuable tool for comparing multiple groups and determining if a particular factor or category has a substantial impact on the outcome studying.

After performing an analysis of variance (ANOVA), there are typically two main values that are:

1) *F-Statistic (F-Value):* The F-statistic, or F-value, is the test statistic used in ANOVA to assess whether the means of two or more groups are significantly different from each other. A larger F-value suggests that the variation between group means is greater than the variation within each group.
2) *p-Value:* The p-value is a measure of the probability that the observed data could have occurred by chance if the null hypothesis were true. In the context of ANOVA, a low p-value (usually less than a predetermined significance level, often 0.05) indicates that there is a significant difference between the means of at least two groups. This leads to the rejection of the null hypothesis, suggesting that there is evidence of a significant effect. On the other hand, a high p-value suggests that there is no significant difference between the means of the groups.

Features	P values	F values
Ca	144.775301	0.000000000000000
Exang	41.223358	0.000000000000000
Oldpeark	40.168852	0.000000000000000
Thalch	38.979224	0.000000000000000
Cp	36.263397	0.000000000000000
Sex	24.020727	0.000000000000000
Age	22.848036	0.000000000000000
Chol	17.579946	0.000000000000096
Slope	9.714975	0.000000120186475
Thal	7.227818	0.000010548397561
Trestbps	3.093567	0.015374976770680
fbs	2.564694	0.037228255311462
restecg	1.943089	0.101586052456979

Table 4: P values and F values of features

D. Entropy

Entropy is a fundamental concept used to measure the impurity or randomness in a dataset. It plays crucial role in various algorithms, especially in decision trees and information gain calculations. Entropy is used to quantify the uncertainty present in the dataset and is commonly employed in the context of classification problems. In the context of machine learning, entropy is often calculated using the formula:

$$\text{Entropy} = -\sum p_i(p_i)$$

Where p_i represents the probability of an item belonging to a particular class. Entropy reaches its maximum when the dataset is equally distributed among all classes, indicating high uncertainty, while it reaches its minimum when the dataset is perfectly classified into a single class, signifying low uncertainty or perfect order. The concept of entropy is closely related to the concept of information gain, which is used to measure the effectiveness of attributes in the context of decision trees and feature selection. By analyzing the entropy of the dataset, machine learning algorithms can make informed decisions about the most suitable attribute for partitioning the data and improving the overall predictive accuracy of the model.

	Before FS	After FS
Training Time	4.648164	8.583903
No of Features	13	11
Best Accuracy	74.208145	76.018100

Table 5: Entropy values before feature selection and feature selection

VI. CONCLUSION

To ensure data robustness, preprocessing techniques such as label encoding and meticulous null value handling were employed. Specifically, for managing missing values, a regression imputer was utilized for numerical features, while categorical features were imputed using the mode values of respective features. This research proposes a model that aids healthcare professionals in accurately predicting and detecting heart diseases at an early stage through the utilization of an optimized set of features. The model not only reduces the time required to obtain accurate outputs but also enhances the overall quality of healthcare services. Additionally, it effectively minimizes the costs associated with the life- saving treatment of individuals. Furthermore, the study delved into the impact of feature selection on the model's accuracies, revealing notable improvements in predictive performance. These comprehensive data preprocessing strategies contributed significantly to the reliability and efficacy of the proposed model, promising a valuable and reliable tool for early detection and diagnosis of heart diseases.

REFERENCES

[1] M. Qadri, A. Raza, K. Munir, and M. Almutairi, “Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning,” vol. 11, p. 56214, Jan. 2023, DoI: 10.1109/access.2023.3281484.
[2] R. Kapila, T. Ragunathan, S. Saleti, T. J. Lakshmi, and M. W. Ahmad, “Heart Disease Prediction using Novel Quine McCluskey Binary Classifier (QMBC),” vol. 11, p. 64324, Jan. 2023, DoI: 10.1109/access.2023.3289584.
[3] Abdellatif, Abdallah, et al. "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods." ieee access 10 (2022), DoI: 10.1109/ACCESS.2022.3191669.
[4] Aishwarya D., Pratiksha K. et al “Interpreting Cardiovascular Disease using Random Forest and Explainable AI” IJRASET Volume 11, Issue V, May 2023 DoI: 10.22214/ijraset.2023.52922.
[5] K. Vayadande et al., “Heart Disease Prediction using Machine Learning and Deep Learning Algorithms,” May 2022, DoI: 10.1109/cises54857.2022.9844406.
[6] A. A. Almazroi, E. Aldhahri, S. Bashir, and S. Ashfaq, “A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning,” vol. 11, p. 61646, Jan. 2023, DoI: 10.1109/access.2023.3285247.
[7] Aishwarya Dabir et al “Analysis of Cardiovascular Disease using Machine Learning Techniques”, Volume 11 Issue V May 2023, DoI: 10.22214/ijraset.2023.52789.
[8] https://archive.ics.uci.edu/dataset/45/heart+disease.zip
[9] https://raw.githubusercontent.com/asdeokar/Dataset/main/heart.csv
[10] S. Ouyang, “Research of Heart Disease Prediction Based on Machine Learning,” Apr. 2022, doi: 10.1109/aemcse55572.2022.00071.
[11] R. G. Franklin and B. Muthukumar, “Survey of Heart Disease Prediction and Identification using Machine Learning Approaches,” Dec. 2020, doi: 10.1109/iciss49785.2020.9316119.
[12] A. U. Haq, M. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, “Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection,” Mar. 2019, doi: 10.1109/i2ct45611.2019.9033683.
[13] A. Newaz and S. Muhtadi, “Performance Improvement of Heart Disease Prediction by Identifying Optimal Feature Sets Using Feature Selection Technique,” Jul. 2021, doi: 10.1109/icit52682.2021.9491739.
[14] A. Jain, K. Kumar, R. G. Tiwari, N. B. Jain, V. Gautam, and N. K. Trivedi, “Machine Learning-Based Detection of Cardiovascular Disease using Classification and Feature Selection,” Apr. 2023, doi: 10.1109/csnt57126.2023.10134672.