

A HYBRID APPROACH TO DIAGNOSIS OF HEART DISEASE WITH MACHINE LEARNING TECHNIQUE

¹Vandana Agarawal, ²Dr. Virendra Kumar Swarnkar (Associate Professor)

¹Research Scholar, ²Supervisor

¹⁻² Department of Computer Science and Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh

:

Abstract : Due to the fact that heart disease continues to be one of the top causes of death around the globe, it is imperative that diagnostic tools that are both accurate and efficient be developed. Traditional diagnostic approaches, notwithstanding their efficacy, can call for a significant amount of medical experience and are moreover time-consuming. For this particular scenario, machine learning (ML) techniques present a viable answer since they make it possible to make decisions based on data and to spot problems at an earlier stage. In this work, a hybrid method to the detection of heart disease is proposed. This strategy involves the integration of different machine learning algorithms in order to improve the accuracy and reliability of prediction. The model makes use of numerical and categorical clinical data, such as age, blood pressure, cholesterol levels, and the kind of chest pain, in order to train and evaluate classifiers.

Keywords: Heart Disease Detection, Machine Learning (ML), Hybrid Model, Clinical Data Analysis, Predictive Modelling, Early Diagnosis.

1. Introduction

The machine learning models have showed effectiveness in cardiac disease prediction, most studies focus on the application of particular classifiers such as decision trees, logistic regression, or support vector machines (SVM). Although these models give useful insights, their forecast accuracy is generally hampered by dataset complexity, noise, and intrinsic algorithmic restrictions. Furthermore, a single classifier's performance might not translate well to a variety of patient populations. A hybrid machine learning strategy is therefore required in order to combine the advantages of several algorithms, minimize their drawbacks, and produce more consistent and trustworthy diagnostic results.

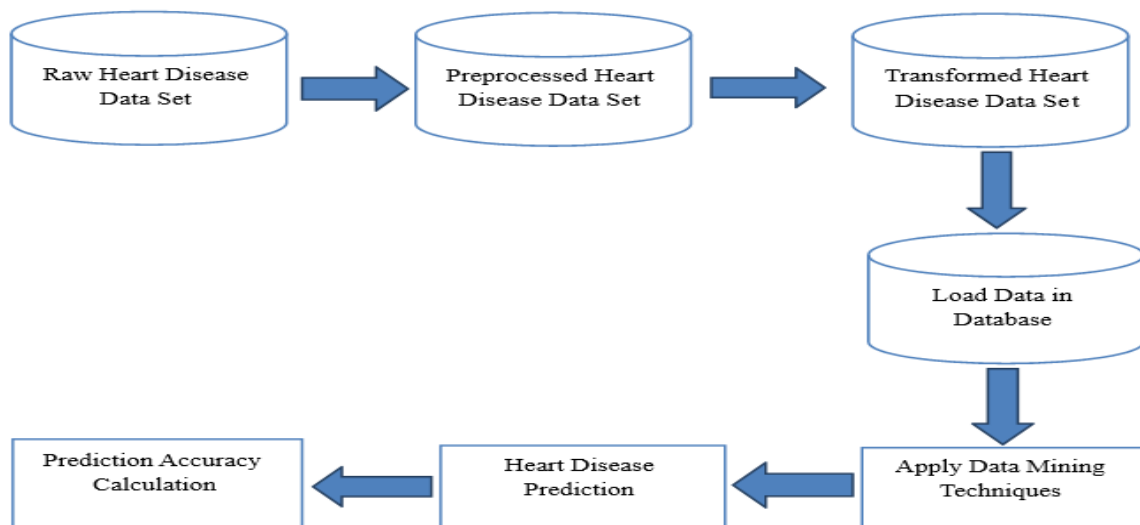


Figure 1. Prediction System

As of right now, there are no proven procedures that can reliably translate the raw data acquired by sensors into physiological signals that can be understood by humans. These signals might be used by medical professionals to identify cardiovascular disease symptoms. In order to address this problem, research is being carried out utilising data mining and statistical analysis techniques that are able to uncover connections from large datasets consisting of healthcare information. Although there is theoretical support for a great number of algorithms, they simply do not have the processing capability to be successfully implemented in real time. A glance at the architecture of the prediction system that developed exposes the essential components that make up the system (Figure 1.1).

1.1 Heart Diseases Prediction Using Artificial Neural Network

When it comes to difficulties that need broad spectrum categorisation, an Artificial Neural Network (ANN) is a primary instrument that is used. The Artificial Neural Network (ANN) will generate predictions if the output is continuous, and it will categorise the data if the output is discrete. According to Murthy and Meenakshi (2013), artificial neural networks (ANN) have the ability to predict and categorise due to their capacity to set the weights of neurones. Neural networks are constructed by neurones and the connections that exist between them. Using artificial neural networks (ANNs) that are focused on decision support may help to improve the general public's perception of the reliability of healthcare. Due to the fact that medical professionals do not have access to a data bank that can provide complete sickness prediction, even in a particular field, it may disclose unusual abnormal illnesses. In order to assist the implementation of the ANN model, which enables the accurate identification of physiological parameter variation during arrhythmia, the primary symptom in the diagnosis of cardiac illness, the availability of vast patient data in electronic form is a significant factor. Sayad and Halkarnikar (2014) provide the fundamental procedures that are used in the application of ANN for the diagnosis of heart illness in Figure 2.

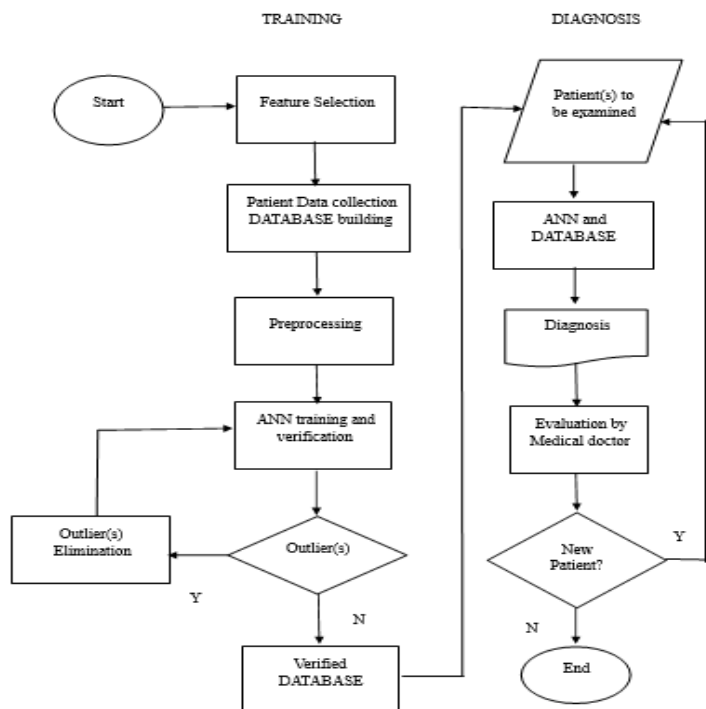


Figure 2. Essential Measures for Coronary Heart Disease

1.2 Data Classification for Heart Disease

When physicians and nurses are well-versed on the factors that lead to cardiovascular disease, they are able to swiftly identify the patients who are most at risk for developing the condition. Experts are essential in the process of analysing datasets in order to extract useful information, and they are also responsible for storing enormous volumes of data.

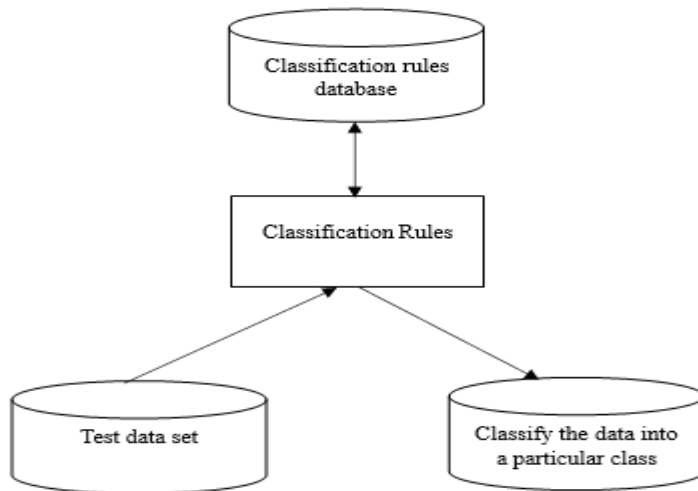


Figure 3. Rules uses for classifying unknown tuples

In order to make predictions about the labels of categorical classes, classification techniques are used. In order to classify data, a model is constructed by using a training set and values (class labels) in an attribute that is used for classification. After then, this model is used to classify newly collected data. The two components that comprise this process are the Model Construction and the Model Usage that are involved. Model construction is the process of giving an explanation for a preset group of classes. This approach is known as "model creation." A decision tree, clustering classification, and Naïve Bayes are the three classifiers that are used in the process of determining whether or not individuals are suffering from cardiac illness.

2. Lecture Survey

It has been determined by the World Health Organization that cardiovascular disease continues to be one of the leading causes of death on a global basis. A number of environmental and inherited factors, such as mental stress, a sedentary lifestyle, a heavy workload, and poor eating habits, are contributing factors in the increasing prevalence of heart disease. The majority of the time, diagnostic techniques that have been around for a long are dependent on the patient's medical history, the results of laboratory tests, and the responses that the physician provides to their clinical queries. However, as a result of the rapid development of computational technology, new ways have emerged for the early detection and diagnosis of cardiac diseases. These new methods are hybrid approaches that integrate data mining and machine learning techniques. When searching through huge medical datasets, hybrid models use a variety of algorithms to seek for patterns that would otherwise be missed by conventional diagnostic tools. This is done in order to identify these patterns. Because of the installation of such sophisticated technology, medical personnel are able to improve their decision-making abilities, which in turn helps them provide treatment that is both accurate and timely.

The fundamental purpose of Mohan et al. (2019) [2], is to find the unique attribute that, by using a machine learning technique, has the potential to increase the accuracy rate of heart disease forecasts. An extensive range of feature selection algorithms and feature extraction techniques are used in this process.

The fundamental objective of the research conducted by Reddy et al. (2020)[3], is to create a model that makes use of fuzzy logic in order to forecast the first cardiac stage. Another term for this model is the Adaptive Genetic algorithm, or AGAFL. The first step is to use rough set theory in order to isolate the feature that was harmed.

Utilising a genetic algorithm, Salem (2018)[4], conducted research on a wide variety of models for the purpose of predicting cardiac illness and selecting features. When the genetic algorithm is used in the most effective manner, the results of its performance are superior to those obtained using more traditional methods. They are put through real-time assessment using four different classifiers, namely Random Forest, Decision Tree, Navies Bayes, and Support Vector Machine, and then they are retested with a number of datasets that are associated with cardiac sickness. Based on the findings, it can be concluded that Navies Bayes is the most effective classifier for this particular dataset.

According to Gokulnath et al. (2019)[5], the major objective is to optimise the function via the significant feature selection process by using a mix of support vector machine (SVM) and genetic algorithm (GA). A number of

different approaches, including as CFS, chi-squared, Filtered subset, Info gain, and consistency subset, are evaluated in relation to the outcomes that are generated by the proposed model.

Choi et al. (2017)[6], have compared and contrasted new approaches with more traditional ones in order to determine the most effective way to diagnose heart failure at an early stage. In order to identify events that have a time stamp, the Recurrent Neural Network is used between the time periods of 12 and 18 months of observations.

The Differential Evolution method (DE) was proposed by Vivekanandan et al. (2017)[7], for the purpose of feature selection during the pre-processing stages. The seventh method, which is developed from the traditional DE technique and compared to other approaches, is used for the purpose of diagnosing cardiac illness.

Dwivedi (2018)[8], made use of the outcomes of six different machine-learning approaches in order to evaluate or investigate the performance metrics of the techniques that were supplied. For the purpose of determining whether approach is more successful in forecasting the incidence of heart disease, the purpose of this assessment is to do this. There are a great number of different classifiers available to select from when it comes to the prediction of cardiovascular disease. In order to anticipate the early stage of cardiac sickness, Nayak et al. (2019)[9], used a number of different classification algorithms. These approaches included Navies Bayes, Support Vector Machine, Decision tree classifier, and k-NN classifiers.

Karadeniz et al. (2021)[10], offered two different classifiers: the first was the Reference Vector classifier—RVC, which was used as a baseline estimate by analysing the distance randomisation using the medical Spectf dataset. The second classifier was the Random Forest classifier. In addition, a Shrunk Covariance Classifier (SCC) is provided for the statlog dataset, which is then delivered to the University of California, Irvine.

According to Maji et al. (2019)[11], a decision tree algorithm is a very effective classification technique that may be used for the purpose of predicting cardiovascular disorders. By combining decision trees with artificial neural networks, this technique employs a hybridisation strategy to improve the prediction of cardiac diseases using WEKA.

Patro et al. (2021)[12], used a variety of methods to enhance classification and feature selection in order to achieve their goal of improving the prediction method for heart illness. With the assistance of this classification technique, we were able to thoroughly investigate the identification model that was used for the purpose of data training and testing.

In the field of machine learning, classifiers are often used to identify a person's shortcomings. Initially, they were used to foresee cardiac issues by assessing a range of parameters, including haemoglobin A1c, cholesterol levels, ECG data, and heart rate. Kumar et al. (2021)[13] conducted a research in which they evaluated a number of different machine learning techniques and classifiers for the purpose of diagnosing cardiac diseases by making use of datasets for prediction. We rated the approaches based on their accuracy rate as well as the amount of time they took to compute. Classifiers such as Hoeffding trees, Random Forest, J48, Random Tree classifiers, and LMT-Logistic Model Trees are used in the process of identifying cardiac diseases.

Additionally, Sagir et al. (2017)[14], introduced a unique methodology for efficient classification that they referred to as the ANFIS—Adaptive Neuro Fuzzy Interference approach. This method was offered in addition to the Levenberg-Marquardt method. An integration of ANFIS-LSLM and ANFIS LSGD has been implemented in order to enhance the precision of the prediction of cardiovascular disease.

Data pre-processing techniques were also used by Dinesh et al. (2018)[15], in order to detect cardiovascular illnesses. These techniques included the elimination of noisy or missing data, the utilisation of classifiers, and the filling in of default values where deemed suitable. For the purpose of assessment, a variety of model-based performance measures are used.

Uddin et al. (2019)[16], used a supervised machine-learning approach in order to make a prediction for a specific sickness. The identification of cardiovascular diseases is accomplished by the use of two separate databases, namely Scopus and PubMed. Within the scope of this investigation

Using a grid search method, Javeed et al. (2019)[17], proposed a revolutionary diagnostic system that is capable of optimising itself. Two different sorts of prototypes were used by the writers in order to perfect the correctness of the prototype. The first strategy includes the creation of the random forest algorithm. In the second method, the random forest model and the random forest algorithm are combined in order to build upon the foundation that was previously established. With the use of the architecture proposed by Ganesan et al. (2019)[18], the UCI dataset and healthcare sensors have been utilised to make an accurate diagnosis of individuals suffering from cardiovascular ailment. Training and testing are the two unique processes that are included in the categorisation approach that is presently being used. During the training phase, the classifiers are provided with the dataset that is considered to be the benchmark.

The authors Venkatesh et al. (2019)[19], used a unique analytical model that depended on Big Data predictive analysis employing the Navies-Bayes Technique, which they referred to as BPA-NB. This model was utilised in order to anticipate the incidence of cardiovascular sickness. Categorisation is provided by the use of Navies Bayes



probability, which is particularly effective when applied to very large data sets. Our research is based on the machine learning dataset used by the University of California, Irvine (UCI). ‘

A technique for selecting feature wrappers was proposed by Al-Tashi et al. (2018)[20]. This approach combines support vector machines (SVM) with the grey wolf optimisation strategy. We are now able to choose valuable features from the Cleveland dataset by using this strategy, while simultaneously removing qualities that are either unneeded or repetitive. Through the use of this proposed strategy, an accuracy rate of around 91% has been accomplished. In the future, the author has suggested that there should be an enhancement that would make the process of selecting features more effective. Several other medical datasets have also used this grey wolf approach for the purpose of making fine-grained predictions about outcomes. In addition, Manikandan et al. (2019)[21] have proposed a model for the identification and prediction of diabetes by using the fuzzy rule concept and the grey wolf optimisation technique. There is a success rate of 71% for the ant colony optimisation approach, which is the foundation upon which this model is built. It is essential to maintain a high degree of precision in order to make progress in the future.

For the purpose of identifying and forecasting cardiac sickness, Uyar et al. (2017)[22] used a genetic algorithm that was based on RFNN, which stands for recurrent fuzzy neural networks. For this procedure, the dataset that has been used is the Cleveland dataset, which is derived from the data that is stored in the UCI repository

Jayaraman et al. (2019)[23], presented a unique method for the early diagnosis of heart illness. This method included integrating a neural network with a bee optimisation algorithm and an artificial gravitational cuckoo search algorithm in order to handle the features that are required for early stage prediction.

Using fuzzy analytic hierarchy processing (Fuzzy AHP), Samuel et al. (2017)[24], introduced a unique technique for acquiring information on the weights of components worldwide according to their individual contributions. This method is based on the concept of fuzzy AHP.

To improve the performance of the artificial neural network (ANN), Alsaeedi et al. (2020)[25] used the PSO-particle swarm optimisation approach. Artificial neural networks and the PSO method have been used in conjunction with one another to solve and evaluate non-linear models. This has been the case for the most part.

Diker et al. (2019)[26] provide a unique approach for the categorisation of electrocardiogram (ECG) signals. This method is referred to as the Physikalisch-Technische Bundesanstalt Diagnostic ECG Database (PTDB), and it is based on data that was retrieved from the database of Physionet. Methods such as pan-Tompkins and discrete wavelet may be used in order to extract critical points from the electrocardiogram (ECG).

A fresh concept that they referred to as Medical Data Mining-MDM was created by Hassoon et al. (2017)[27]. As part of its efforts to enhance diagnosis, it has used a variety of data mining methods, including clustering and classification, in order to determine the symptoms and classifications of various illnesses.

An innovative machine-learning approach was developed by Abdar et al. (2019)[28] with the purpose of accurately identifying coronary artery disease diagnoses. In this study, the major emphasis is on assessing the performance of ten classic machine-learning algorithm strategies by using three distinct types of support vector machines (SVMs). An innovative feature selection strategy was used in order to choose the most significant feature from the dataset that was supplied for the purpose of early identification of coronary artery diseases.

3. Research Methodology

The primary emphasis of our research agenda was on developing issues that are associated with existing technology. With the use of contemporary computer methods, the problem that was associated with heart diseases has been resolved. Through the use of the tactics and methods that were created by early mathematicians, these issues have been found and resolved. Through the use of analytical procedures, such as various machine learning techniques, it is possible to determine the numerical results and the domain in which they have significance. Through the use of computer software that plots the pertinent graphs and tables, the results and conclusions are presented to the interested parties. In the course of our research, we made use of Anaconda 3, which is based on the Python programming code.

3.1. Materials and Methods

The study resources and methods are projected and briefly explained in this part.

a) Dataset Description

Dataset-I: UCIKC HDDs: It is possible to access this HD dataset via the repository at the University of California, Irvine (UCI), which was provided by the Medical Centre and the Cleveland Clinic Foundation (Jindal et al., 2021).

This dataset has 14 features, 303 patients, and two classes.

In order to collect data for this research, cardiologists manually collected information from 333 patients, each of whom had 21 characteristics. After being precisely recorded in an Excel sheet, the name property was subsequently erased in the right manner, as stated by Hay et al. (2017).

The table 1, Combined UCIKCHSLBV HDDs: The Medical Centre, the Cleveland Clinical Foundation, the Hungarian Institute of Cardiology, Switzerland, and the Long Beach V Clinical Foundation are the organisations that make this hard disc drive (HDD) available to the public, as stated by AL Mustafa et al. (2020). This HDD contains 13 features, 1025 instances, and two classes. That information is kept in the repository at UCI.

Table 1: Description of the Datasets

	Dataset	Classes	Attributes	Instances
Dataset-I	UCIKC	2	14	303
Dataset-II	Erbil Heart Diseases	2	21	333
Dataset-III	UCIKCHSLBV	2	14	1025

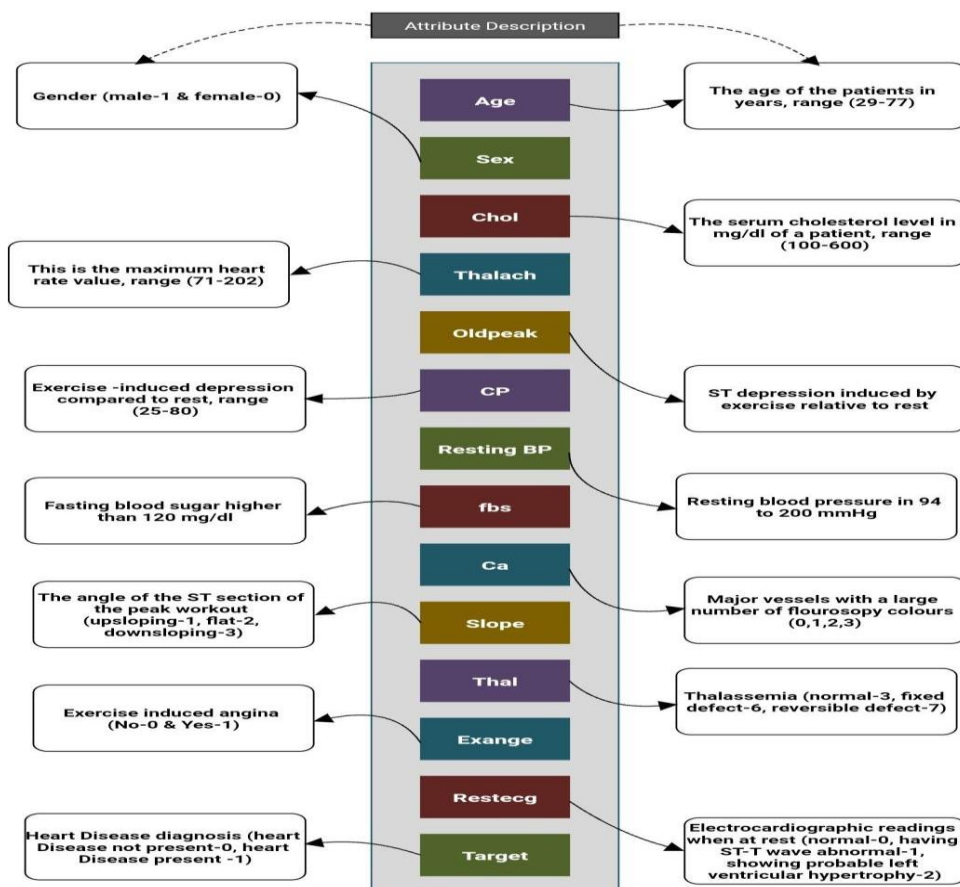


Figure 3. Parameters taken in dataset I and III

3.2. Machine Learning NLTF Preprocessing

In order to maintain the edge of smoothing and visible detail for N-dimensional signaling in computer graphics, image processing and computer vision applications, a high contrast image and mail-in trilateral filter has been introduced as a nonlinearly single-pass filter. In order to identify noisy pixels in random images affected by impulses noise the trilateral filter is used to insert the local image statistics into the BF. A NLTF value quantifies the difference in intensity between a center pixel and its closest comparable neighbors. For reducing Gaussian noise and uniform

noisy impulses and mixing, the trilateral filter is suitable.

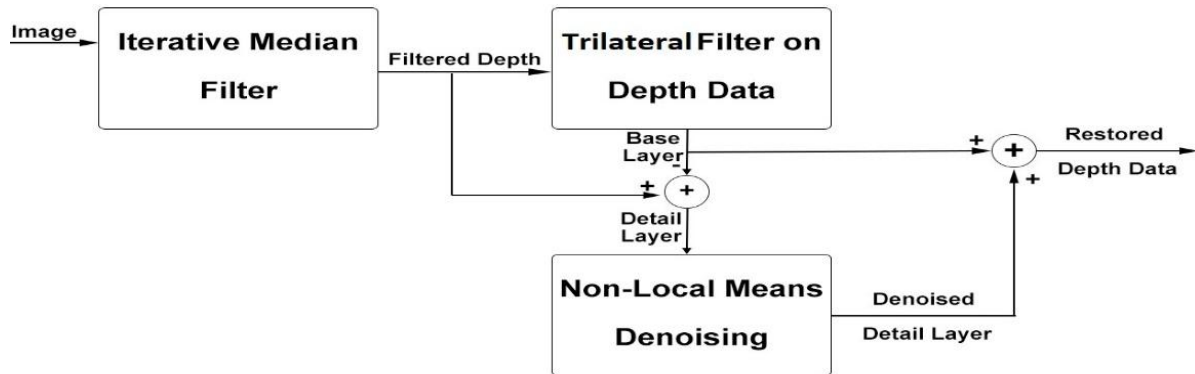


Figure 4. NLTF block diagram.

A pipeline architecture that aims to reduce the complexity problem of traditional trilateral filters with trilateral noise filters. Its very high temporal complexity is the major difficulty of the application of such a strong noise filter on real time imaging systems. The NLTF computation is bitwise performed and the exponential function evaluation is partly performed using linear approximations as shown in Figure 4.

The figure 5, an adaptive trilateral filter by employing NLTF and Rank Ordered Logarithmic Difference statistic was developed by using trilateral filters. The NLTF is already a decent statistic on pulse noise in imagery; however certain bruising pixels may possess intensity values near their surrounding pixels for randomly valued pulse noise. In such a scenario, the pixel NLTF cannot be large enough to differentiate between it and the free pixels of noise. The logarithmic function is called ROLD function, to enhance the input image.

Table 2: NLTF Model Information

Layer	Type	Input	Kernel	Output
1	Convolution	$28 \times 28 \times 1$	5×5	$24 \times 24 \times 32$
2	Max pooling	$24 \times 24 \times 32$	2×2	$12 \times 12 \times 64$
3	Convolution	$12 \times 12 \times 64$	5×5	$8 \times 8 \times 64$
4	Max pooling	$8 \times 8 \times 64$	2×2	$4 \times 4 \times 64$
5	Fully connected	$4 \times 4 \times 64$	4×4	512×1
6	Fully connected	512×1	1×1	2×1
7	Softmax	2×1	N/A	Result

The table 2, maximum pooling layer would then select the best deal from each kernel to reduce the number of features. In order to achieve the best results with the least quantity of computational work, multiple iterations of the convolution and max-pooling layers are used. Additionally, while mapping input to output features, a completely connected layer is used to maintain all neuron interconnections. Finally, the SoftMax classifier is always used to differentiate between the benign and malignant classes.

Segmentation methods for lung cancer typically involve analyzing medical images, such as computed tomography (CT) scans, to identify and delineate tumor regions within the lung. There are several segmentation techniques used in this context, including:

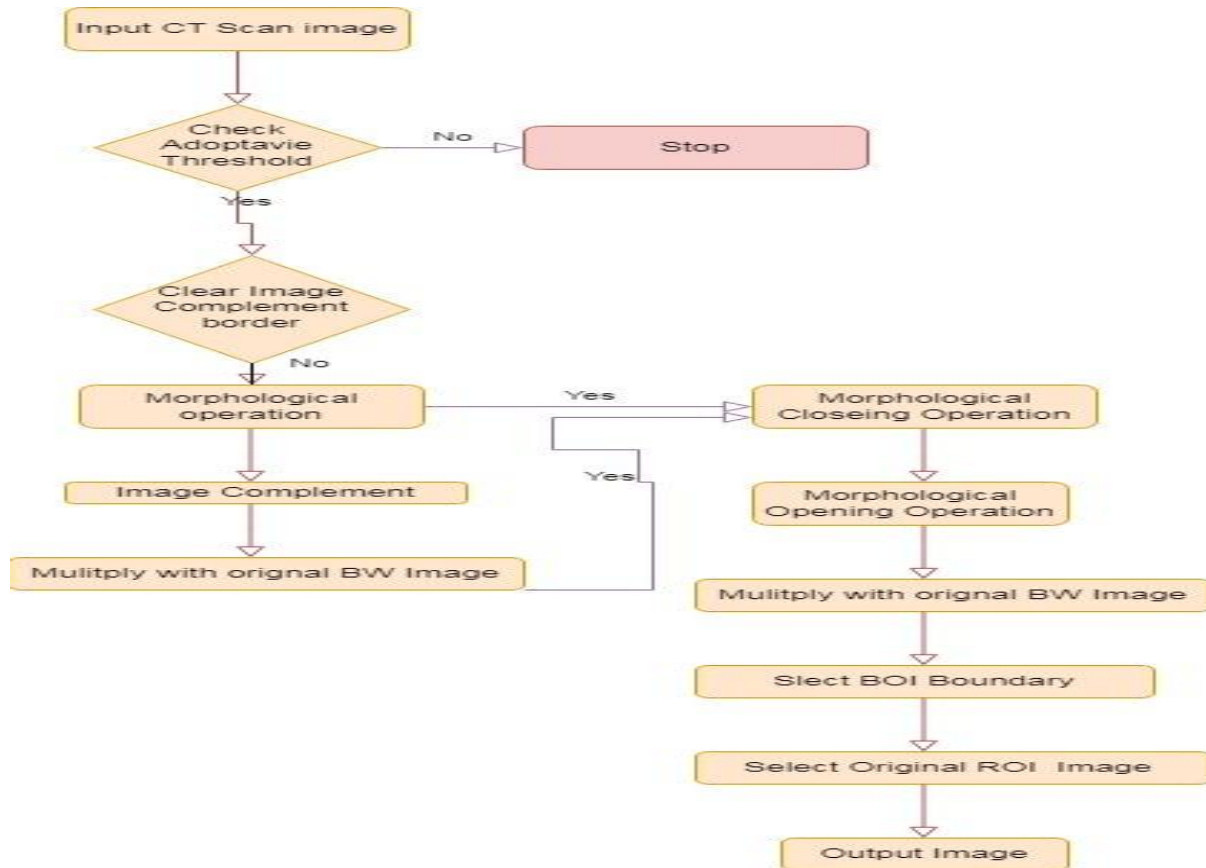


Figure 5: Flowchart for Segmentation Recommendation for NLTF.

Numerous medical computer vision applications used for classification of various types of lung cancer, make extensive use of the NLTF models. This method of comprehension has lately attracted a lot of attention due to its success. NLTF have the advantages of being much easier to learn and having fewer characteristics than fully connected networks.

3.3. HFM segmentation

The HFM (Hierarchical Feature Matching) algorithm, in mathematical terms, can be described using the following steps:

Feature Extraction: Let's denote the input image as I . Feature extraction involves identifying relevant characteristics or descriptors from the image. Let $F = \{f_1, f_2, \dots, f_n\}$ represent the set of extracted features from image I . Each feature f_i is a vector or descriptor that captures certain properties of the image region.

Feature Matching: The next step is to match the extracted features to find corresponding features across different image regions. Let $M = \{(f_i, g_j)\}$ represent the set of matched feature pairs, where f_i is a feature from image I and g_j is a feature from image J .

Hierarchical Grouping: The matched feature pairs are hierarchically grouped based on their similarities and spatial relationships. This grouping process can be represented using a hierarchical structure, such as a tree or a graph. Let $G = (V, E)$ represent the hierarchical grouping structure, where V is the set of nodes and E is the set of edges connecting the nodes.

Region Merging: The algorithm performs region merging iteratively to refine the segmentation and create coherent regions. Let $R = \{R_1, R_2, \dots, R_k\}$ represent the set of regions, where each region R_i is a collection of features. The merging process can be formulated using certain criteria or rules, such as similarity measures or clustering algorithms. The algorithm updates the set of regions R until convergence is achieved.

Region Classification: Once the merging process is complete, the resulting regions can be classified into different classes or labels. Let $C = \{C_1, C_2, \dots, C_l\}$ represent the set of class labels, where each region R_i is assigned to a specific class C_j based on certain classification criteria or machine learning models.

It's important to note that the mathematical representation of the HFM algorithm can vary depending on the specific

implementation and variations proposed by different research papers or publications. The above description provides a general framework for understanding the algorithm's mathematical terms and steps.

3.4 DLCNN Model Classification

The figure 6, DLCNN models are widely used in many medical image processing applications including multi class lung cancer classification. It is a kind of efficient identification approach that has lately gotten a lot of attention because of its effectiveness. The advantage of DLCNNs is that they are simpler to train and have a much smaller number of parameters than fully linked networks.

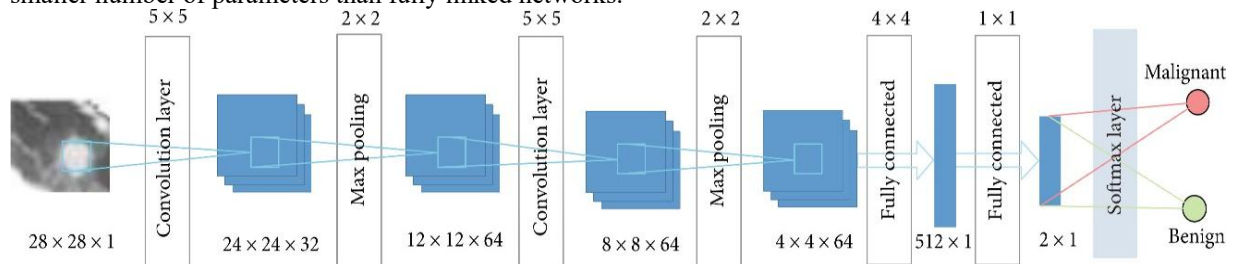


Figure 6. DLCNN classification

The steps involved in performing classification using a Deep Learning Convolutional Neural Network (DLCNN) typically include the following:

1. Data Preparation:
 - Collect and preprocess the dataset for training and testing.
 - Preprocessing steps may involve resizing images, normalizing pixel values, augmenting the dataset, and splitting it into training and testing sets.
2. Network Architecture Design:
 - Design the architecture of the DLCNN model.
 - Determine the number and arrangement of convolutional layers, pooling layers, and fully connected layers based on the problem at hand.
 - Consider the input dimensions, the number of classes, and any specific requirements or constraints.
3. Model Compilation:
 - Define the loss function to measure the difference between predicted and actual class labels.
 - Select an optimizer to adjust the model's weights during training.
 - Choose evaluation metrics, such as accuracy, to assess the performance of the model.
4. Training:
 - Train the DLCNN model using the prepared training dataset.
 - Feed the training data through the model, compute the loss, and backpropagate the gradients to update the model's weights.
 - Repeat this process for multiple epochs, adjusting the model's parameters to minimize the loss and improve accuracy.
5. Model Evaluation:
 - Evaluate the trained model's performance on the testing dataset.
 - Feed the testing data through the trained model and compute the predicted class labels.
 - Compare the predicted labels with the ground truth labels to calculate evaluation metrics such as accuracy, precision, recall, or F1-score.
6. Fine-tuning and Optimization:
 - If necessary, fine-tune the DLCNN model by adjusting hyperparameters or modifying the architecture.
 - Experiment with different optimization techniques, learning rates, regularization methods, or batch sizes to improve model performance.
7. Prediction:
 - Use the trained DLCNN model to make predictions on new, unseen data.
 - Provide the input data to the model and obtain the predicted class labels or probabilities.

4. Result Analysis

The analytical diagnosis of heart illness via the use of hybrid machine learning models is a novel approach that has the potential to be of immense benefit in terms of enhancing the effectiveness and precision of the identification of cardiovascular disease. In order to lessen the impact of heart disease, which continues to be one of the leading causes of mortality on a worldwide scale, early identification is very necessary.

4.1 Exploratory Data Analysis and Results Analysis

Cleveland dataset of figure 7, The identification and prediction of risk factors for binary heart disease is the fundamental objective of the study. The six different machine learning classification techniques are shown in Figure 4.1 below; these algorithms originate from a variety of machine learning applications. Procedure for the suggested method that is now being evaluated.

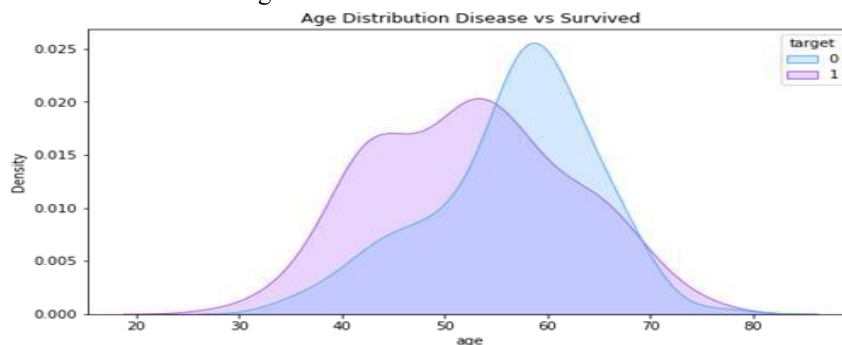


Figure 7: Age Wise Frequency Distribution of HD

The table 3, HDDs' average age is 54.37 years; the oldest is 77 years old and the youngest is 29 years. Cholesterol: The levels of cholesterol that have been recorded include a range that goes from 126 to 564 and an average of 246.5. Resting blood pressure: A resting blood pressure reading of 131.64 mm Hg is the norm. Levels range from 120 mmHg at the lowest to 200 mmHg at the highest. Maximal heart rate attained: With an average value of 149.74, the number of maxes_heart_rate_achieved values vary from 149 to 202. The standard deviation is 22.89. The usual value for st depression is 1.06. It ranges from 0 to 6.2 at its extremes. Major blood vessel count: It is observed that the number of major blood vessels varies from zero to four, with an average value of 0.734. The standard deviation is 1.024.

Table 3: The Analytical Results of the MLA

Models	Accuracy(%)	Precision (%)	Recall(%)	F-1Score (%)	ROC_AUC(%)
DTC	82	82	82	82	85
AdaBoost	87	87	87	87	87
XGB	85	86	85	85	91
KNN	82	84	82	82	92
GBC	89	89	89	89	97
RFC	85	85	85	85	94

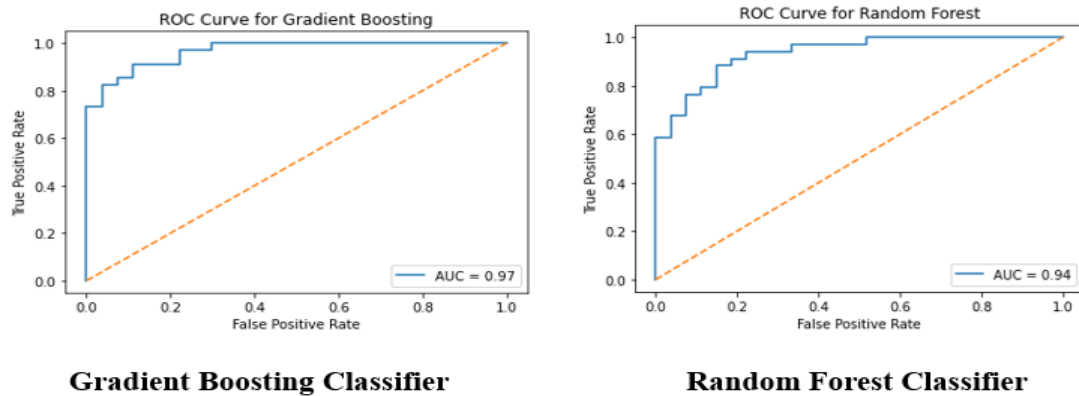


Figure 8: Demonstrates the Combination of ROC& AUC Curves

The figure 8, are a number of tried and tested models that may be used to estimate the risk of heart disease. The Gradient Boosting Classifier risk model approach has a ROC_AUC score of 97%, which places it in the top spot among all other strategies when it comes to determining if a person is healthy or sick. The dataset represent Table 4.

Table 4: Categorical Attributes of the Dataset-I

Attribute	Label	Count	Proportion	Missing	Distinct
Sex	0	96	31.68%	0	2
	1	207	68.32		
Cp	0	143	47.19%		
	1	50	16.50%	0	4
	2	87	28.71%		
	3	23	0.08%		
Fbs	0	258	85.18%	0	2
	1	45	14.82%		
Restecg	0	147	48.51%		
	1	152	50.16%	0	3
	2	4	1.33%		
Exang	0	204	67.33%	0	2
	1	99	32.67%		
Slp	0	21	6.90%		
	1	140	46.20%	0	3
	2	142	46.90%		
	0	175	57.75%		
Ca	1	65	21.45%		
	2	38	12.54%	0	5
	3	20	6.60%		
	4	5	1.66%		
Attribute	Label	Count	Proportion	Missing	Distinct
Thal	0	2	00.66%		
	1	18	5.94%	0	4
	2	166	54.78%		
	3	117	38.62%		
target	0	138	45.54%	0	2
	1	165	54.46%		

The figure 9, on the statistical characteristics of the attributes shown in Figure 4.10, which include the minimum (Min), mean (Mean), standard deviation (SD), missing values, and distinct values from Table 4.3, it was determined

that the numerical attributes of Dataset-I did not include any missing values.

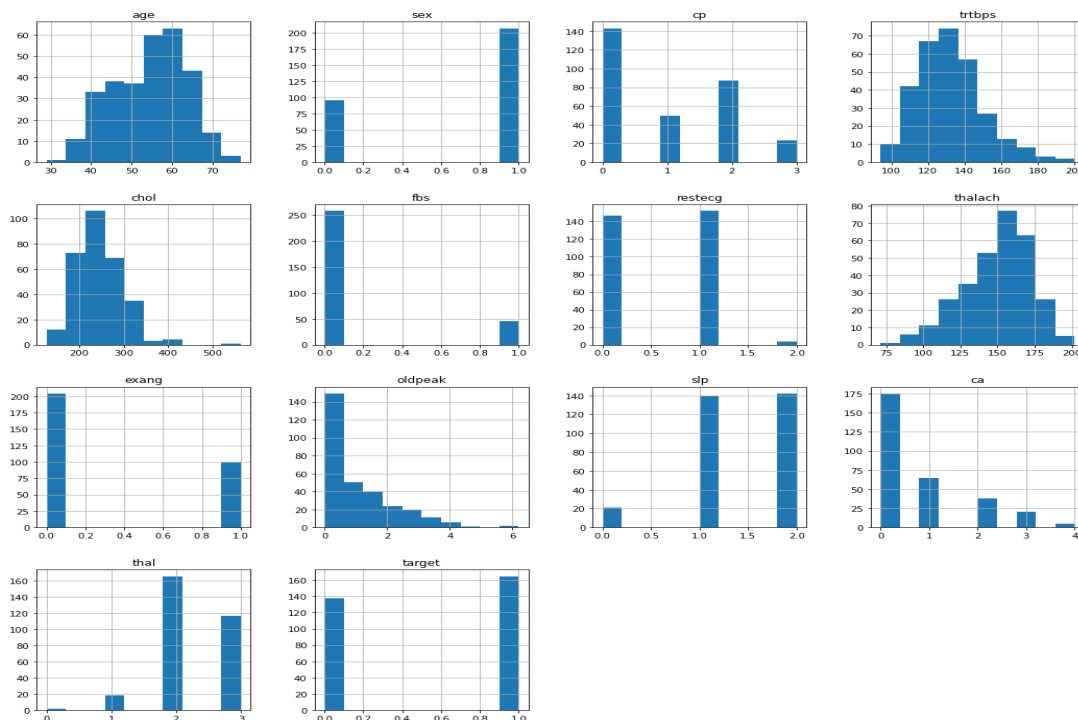


Figure 9: Numerical and Categorical Attributes Histogram in Dataset-I
4.2 Patient Classification according to Various Age Groups:

The figure 10, results of the pre-processing stage, which comprised the categorisation of our dataset into groups of persons with and without cardiac disease, are shown in Figure 4.11. This step was performed in order to ensure that the data was accurate. In contrast, just 45.44% of individuals do not have heart disease, whereas 54.56% of people have. As the bar diagram demonstrates, patients who suffered from heart illness saw a break-even time that was shorter than that of patients who did not suffer from heart disease

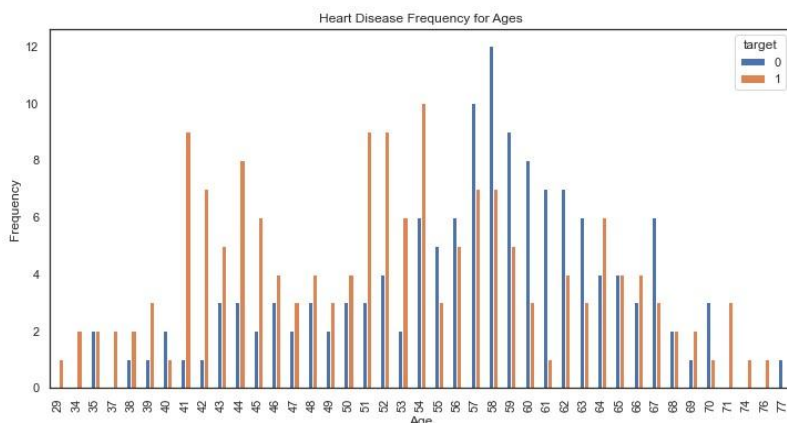


Figure 10: Classification of Individuals with and without heart disease, as well as Age-Specific Frequency in Dataset-I

Correlation is a statistical measure that shows the degree of relationship between two variables. There are mainly two types of correlations: positive and negative. The term "positive correlation" describes a relationship between two variables where the changes in one cause the other to expand. In this study piece, we will discuss correlation theory, introduce a matrix, and use Seaborn to create a heatmap. Data visualisation is made easy using Seaborn, a Python tool. If we are serious about honing our data visualisation abilities, we will keep you posted. We will discuss the use of Seaborn to create matrices and heatmaps once we have reviewed the basics of correlation.

Figure 11, illustrates the interrelationships that exist between the attributes, which are included in the dataset. 'Cp' (chest pain), 'thalach', and 'slp' slope features were found to have a significant connection to the objective elements, and the one-of-a-kind relationship that exists between these three characteristics was determined to be identified.

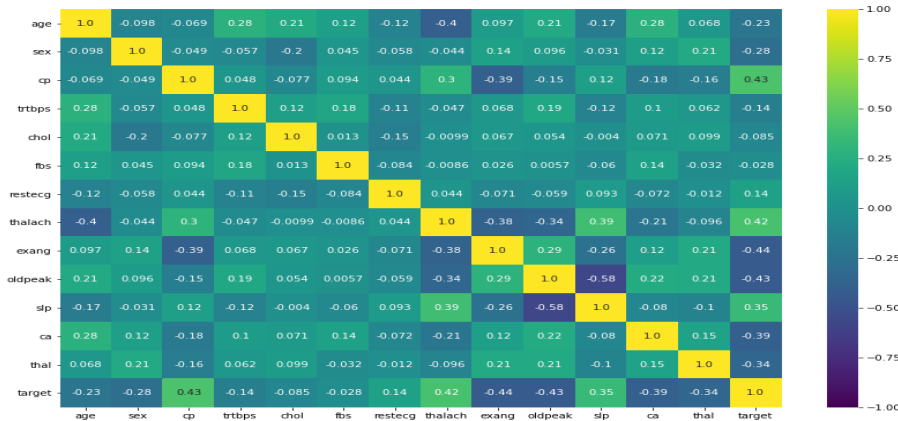


Figure 11: Correlation Heat Map to Measure Accuracy, Precision, Recall, f-1 Score in Dataset-I

There is a total of 215 people who do not exhibit any signs of heart illness, which accounts for 64.56 percent of the population. On the other hand, there are 118 patients who do have heart disease, which accounts for 35.44 percent of the population. On the basis of the information, the numerical characteristics of Dataset-II do not include any values that are missing. The histogram that was produced. Minimum (Min), mean (Mean), standard deviation (MSD), missing values, and distinct values are the statistical features of the numerical value attributes. Other qualities include unique values and missing values represent table 5.

Table 5: Statistical Numeric Attributes of the Dataset-II

Attributes	Min.	Max.	Mean	SD	Missing	Distinct	Proportion
Age	20	90	55.17	14.15	0	59	17.70%
Years	0	50	4.7	11.24	0	33	9.90%
IDL	26	260	112.92	37.97	0	154	46.20%
Hight	128	192	162.10	11.29	0	50	15.00%
Weight	41	134	82.15	15.39	0	92	27.60%
Hr	40	140	83.88	14.63	0	68	20.40%
bpsys	80	220	123.61	21.34	0	21	6.3%
bpdias	46	140	74.87	12.68	0	77	5.1%

The findings of figure 12, of the preprocessing stage, which categorized our dataset according to the presence or absence of cardiac illness in patients, are shown in Figure. This step involves sorting patients into distinct age groups from the beginning of the procedure. A total of 35.44 percent of persons are affected by heart disease, whereas 64.56 percent are not affected by it.

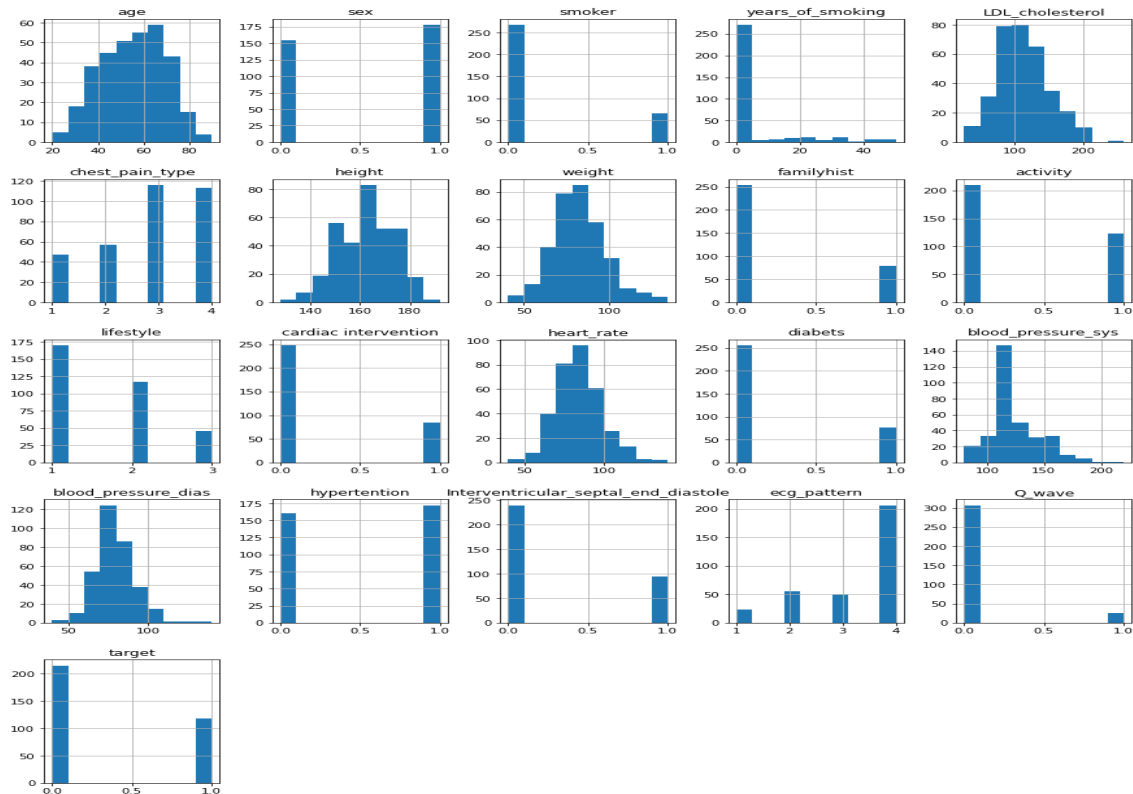


Figure 12: Statistical Numerical and Categorical Histogram Dataset-II

As the bar diagram demonstrates, patients who suffered from heart illness saw a break-even time that was shorter than that of patients who did not suffer from heart disease. Within the age range of 23 to 68 years, the proportion of patients who do not have heart disease is higher than the proportion of patients who do have heart disease. Beginning at the age of 67 and continuing until the age of 82, the proportion of people who have heart disease in comparison to those who do not have heart disease progressively rises across all age groups.

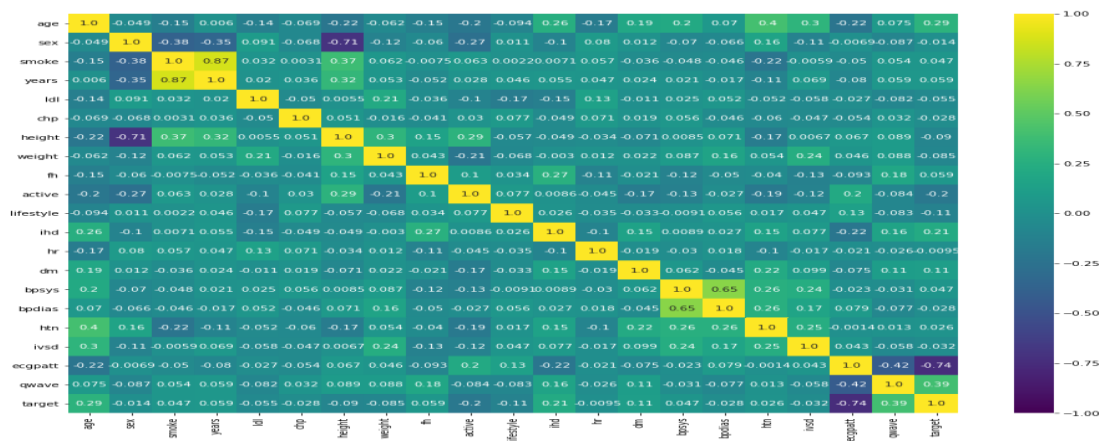


Figure 13: Correlation Heat Map to Measure Accuracy, Precision, Recall, f-1 Score. Dataset-II

The following relationships between the characteristics of the dataset are shown in Figure 13. The following relationships were discovered: (smoke, years) = 0.86; (bpsys, bpdias) = 0.65; (Q-wave, target) = 0.39; (smoke, height) = 0.37; and (years, height) = 0.37. All of these correlations were investigated. It was evident that the target qualities were connected to the smoke, bpsys, Q-wave, height, and years characteristics, and the distinct connections between them were also crystal plain.

5. Conclusion

This work conclude that is being done nowadays requires the automated detection of risks and statistical numerical

aspects. Some examples of these are age, cholesterol, resting blood pressure, ST-depression, and the number of main blood arteries. To satisfy this requirement, a number of different machine learning approaches that are based on DT, AdaBoost, XGB, KNN, GBC, and RFC are currently being researched. A GBC approach was presented in this research as a means of delving further into the subject of how to improve the accuracy of predictions. This approach is based on the Cleveland datasets, which have been used in a way that has been both vast and successful in previous research. The GBC that was proposed is used in order to put into effect the tactics that shown the highest level of success. ROC-AUC accuracy of 97% and accuracy of 89% are both achieved by the suggested technique, which is superior to the study that was done before. After being put into action, the strategy has shown that it is highly advantageous. In preparation for the subsequent classification run, our objective is to develop a GBC model that is more precise.

In the field of medicine, machine learning algorithms (MLAs) are an essential component because they have the potential to forecast human cardiac disease via the analysis of medical data. Cardiovascular diseases (CVDs) are a significant public health problem that need the efforts of academics and medical professionals to ameliorate. This research takes into consideration twelve various methodologies in order to enable it to carry out comparative analysis and get satisfactory outcomes. In comparison to learning algorithms that are directed by humans, the results of this research showed that machine learning algorithms performed much better. The comparison techniques that are used include, but are not limited to, accuracy, recall, precision, F1-score, and ROC-AUC. On the basis of the 14-parameter dataset, the performance of the random forest classifier in the machine learning approach was increased by the use of data pre-processing. Using MLAs, a dataset of clinical data pertaining to cardiovascular disease (CVD) is analysed in order to figure out the primary risk factors that influence the progression of CVD. To find a solution to the issue, this is done. Additionally, the decision tree and random forest are used in order to improve the training and testing procedures for the algorithm.

REFERENCES

1. Triyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An effective heart disease prediction model for a clinical decision support system. *IEEE Access*, 8, 220990–221008. <https://doi.org/10.1109/ACCESS.2020.3041307>
2. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *Procedia Computer Science*, 152, 59–65. <https://doi.org/10.1016/j.procs.2019.05.016>.
3. Salem, [Initials]. (2018). *Study and analysis of prediction model for heart disease: Feature selection using genetic algorithm*. **International Journal of Pure and Applied Mathematics**, 119(16), 5323–5336.
4. Salem, H. (2018). Study and analysis of prediction model for heart disease: Feature selection using genetic algorithm. *International Journal of Pure and Applied Mathematics*, 119(16), 5323–5336.
5. Gokulnath, C. B., & Shantharajah, S. P. (2019). *An optimized feature selection based on genetic approach and support vector machine for heart disease*. *Cluster Computing*, 22(S6), 14,777–14,787. <https://doi.org/10.1007/s10586-018-2416-4>
6. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2), 361–370. <https://doi.org/10.1093/jamia/ocw112>.
7. Vivekanandan, T., & Iyengar, N. C. S. N. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Computers in Biology and Medicine*, 90, 125–136. <https://doi.org/10.1016/j.combiomed.2017.09.011>.
8. Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural Computing and Applications*, 29(10), 685–693. <https://doi.org/10.1007/s00521-016-2604-1>.
9. Nayak, S., etc. (2019). *Early-stage cardiac disease detection using classification algorithms: Naïve Bayes, SVM, Decision Tree, and k-NN with ensemble enhancements in R.*
10. Karadeniz, T., Tokdemir, G., & Maraş, H. H. (2021). Ensemble methods for heart disease prediction. *New Generation Computing*, 39, 569–581. <https://doi.org/10.1007/s00354-021-00124-4>.
11. Maji, S., Arora, A., & Acharjya, D. P. (2019). Decision tree algorithms for prediction of heart disease. *Proceedings of the International Conference on Computational Intelligence in Data Science (ICCIDS 2019)*, 1–5. IEEE. <https://doi.org/10.1109/ICCIDS.2019.8862046>.
12. Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*, 26, 100696. <https://doi.org/10.1016/j.imu.2021.100696>.
13. Kumar, A., (2021). Machine learning classifiers for cardiac disease prediction using medical datasets.

14. Sagir, M., [Other Authors, if any]. (2017). Adaptive Neuro-Fuzzy Inference System (ANFIS) with Levenberg–Marquardt learning for cardiovascular disease prediction..
15. Dinesh, (2018). Data pre-processing and classification techniques for cardiovascular disease detection.
16. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, 19(1), 1–16. <https://doi.org/10.1186/s12911-019-1004-8>.
17. Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., Nour, R., & Khelifi, A. (2019). An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. *IEEE Access*, 7, 180235–180243. <https://doi.org/10.1109/ACCESS.2019.2952107>.
18. Ganesan, M., Ramesh, S., & Kannan, R. (2019). An intelligent healthcare framework for cardiovascular disease prediction using UCI dataset and sensor data. *International Journal of Advanced Computer Science and Applications*, 10(6), 432–439. <https://doi.org/10.14569/IJACSA.2019.0100654>.
19. Venkatesh, S., Kumar, R., & Prasad, P. (2019). BPA-NB: A big data predictive analytics model using Naïve Bayes for cardiovascular disease prediction with Hadoop Spark. *International Journal of Recent Technology and Engineering*, 8(3), 1024–1031. <https://doi.org/10.35940/ijrte.C4243.098319>.
20. Al-Tashi, Q., Kadir, S. J. A., Razzaque, M. A., Mirjalili, S., & Alhussian, H. (2018). Feature selection based on grey wolf optimizer for cardiovascular disease classification. *IEEE Access*, 7, 70657–70666. <https://doi.org/10.1109/ACCESS.2019.2918660>.
21. Manikandan, T., & Bharathi, N. (2019). Diabetes prediction model based on fuzzy rules and grey wolf optimization. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), 2192–2197. <https://doi.org/10.35940/ijitee.L3273.1081219>.
22. Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based recurrent fuzzy neural networks. *Procedia Computer Science*, 120, 588–593. <https://doi.org/10.1016/j.procs.2017.11.280>.
23. Jayaraman, S., Anandakumar, H., & Kannan, A. (2019). Early diagnosis of heart disease using a hybrid neural network with bee optimization and artificial gravitational cuckoo search. *International Journal of Innovative Technology and Exploring Engineering*, 8(9), 2504–2510. <https://doi.org/10.35940/ijitee.I8653.078919>.
24. Samuel, O. W., Asogbon, M. G., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and fuzzy AHP for heart failure risk prediction. *Expert Systems with Applications*, 68, 163–172. <https://doi.org/10.1016/j.eswa.2016.10.020>.
25. Alsaeedi, A., Abushariah, M. A., & others. (2020). *Optimizing artificial neural network performance using particle swarm optimisation (PSO): A comparative study with backpropagation*.
26. Diker, A., [Co-authors]. (2019). *ECG signal classification using extreme learning machine and genetic algorithm-based wavelet optimization..*
27. Hassoon, M., [Co-authors]. (2017). *Medical Data Mining (MDM): Optimizing boosted C5.0 classifier rules with genetic algorithm for improved disease diagnosis.*
28. Abdar, M., [Co-authors]. (2019). *A novel machine learning framework for coronary artery disease diagnosis using hybrid optimisation and support vector machines..*

