Journal of
Cardiology & Cardiovascular Therapy
ISSN: 2474-7580

Juniper Publishers
key to the Researchers

Research Article
Volume 19 Issue 1 - December 2023
DOI: 10.19080/JOCCT.2023.19.556004

J Cardiol & Cardiovasc Ther

A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare

Mohan Raja Pulicharla*

MCA, 2004, Madras University, India

Submission: November 01, 2023; **Published:** December 20, 2023

***Corresponding author:** Mohan Raja Pulicharla, MCA, 2004, Madras University, India

Abstract

The heart, as the second most vital organ after the brain, is integral to maintaining bodily equilibrium, and disruptions to its function have profound health consequences. Heart disease, a leading global cause of mortality, often arises from cumulative daily physiological changes, emphasizing the importance of timely illness prediction. In healthcare, the fusion of data mining and machine learning, explored in this study using Support Vector Machine, Decision Tree, and Random Forest algorithms, addresses the challenges of diagnosing prevalent conditions like heart disease, particularly crucial in the field of cardiology.

Our proposed machine learning-based approach for diagnosing cardiac disease employs a range of classification algorithms and advanced feature selection techniques, demonstrating superior accuracy in detecting heart diseases from extensive datasets of unprocessed medical images. This technological advancement holds the potential to significantly enhance patient care in various healthcare settings, showcasing the promising impact of artificial intelligence tools on improving the quality of life for billions worldwide.

Keywords: Machine learning; Heart disease; Algorithms; Cardiovascular disease; Regression analysis

Abbreviations: ML: Machine Learning; LCS: Learning Classifier Systems; ILP: Logic Programming; ANN: Artificial Neural; SVMs: Network Support Vector; Machines; CVDs: Cardiovascular Diseases; TP: True Positive TN: True Negative; FN: False Negative; FP: False Positive LDL: Low-Density; Lipoprotein ECG: Electrocardiogram; CXR: Chest X-Ray; CVD: Cardiovascular Disease

Introduction

After the brain, the heart is regarded as the second-most significant organ. Every heart disruption causes the entire body to become upset. Heart disease is one of the top five killer diseases in the world. Disorders, including heart disease, are a result of the changes that occur to us daily. Consequently, it is crucial to predict a sickness at the appropriate time. Data mining is a fundamental and fundamental process for defining and discovering relevant data and uncovering hidden patterns in massive databases. By predicting and diagnosing various diseases, data mining, and machine learning techniques are used in the medical sciences to address genuine health-related challenges. This study compares the performance of three machine learning algorithms-support vector machine, decision tree, and random forest-for the prediction of heart disease.

Machine Learning-Based Approach for Diagnosing Cardiac Disease

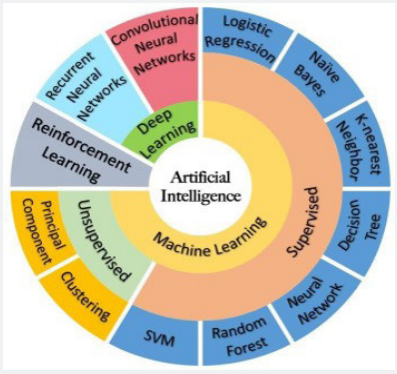

Figure 1: Various algorithms of artificial intelligence and Machine Learning applied in cardiovascular medicine.

The study emphasizes the critical need for swift and accurate heart disease identification, proposing a machine learning approach with Support Vector Machines, Logistic Regression, Artificial Neural Networks, K-Nearest Neighbors, Naive Bayes, and Decision Trees for classification. Efficiency is enhanced through feature selection algorithms and a conditional mutual information method, ensuring commendable accuracy, particularly with Support Vector Machines. This makes it a promising tool for rapid implementation in medical settings, crucial for early identification and interrupting cardiac disease progression. The analysis of diverse datasets identifies key features for heart disease prediction, utilizing seven machine learning methods. A hybrid dataset is created and analyzed with Python's Scikit-learn module using a univariate feature selection technique, offering a comprehensive approach to discern crucial factors in predicting and preventing heart disease.

Provides the Maximum Accuracy, Several Parameters Relating to Various Algorithms:

Datasets are split into training and testing using holdout and cross-validation techniques, adjusting algorithm parameters for maximum accuracy. Evaluation metrics, including a classification report and confusion matrix, gauge performance. Majority voting, combining logistic regression, SVM, and naive Bayes, achieves 88.89% accuracy on the first dataset, while the hybrid dataset lags individual ones. Project outcomes are compared with prior methodologies. Machine learning, an algorithmic system falling under AI, learns without explicit programming, relying on statistics and data for outcome prediction. Linked to data mining and Bayesian modeling, it operates by taking data as input and generating answers through algorithms, seen in applications like personalized recommendations, fraud detection, and predictive maintenanance [1].

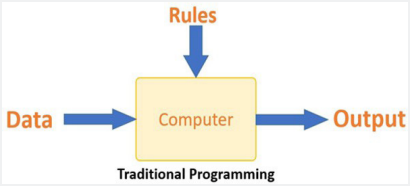

Figure 2: Traditional programming.

Machine Learning vs Traditional Programming

Traditional Programming: This problem is meant to be solved via machine learning. The computer creates a rule after learning how the input and output data are related. Every time there is fresh data, the programmers do not need to design new rules. The algorithms change because of fresh information and experiences, increasing their efficacy over time (Figure 2).

Machine Learning Approach

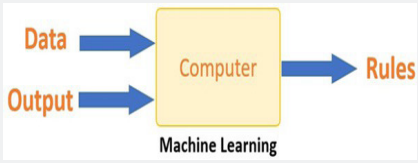

Figure 3: Machine Learning.

Machine learning mimics human learning through experience, succeeding in familiar scenarios with easier predictions. Like humans, machines train by observing examples for precise predictions but struggle with new instances. Central to machine learning is learning and inference, primarily achieved by identifying patterns in data. A crucial skill for data scientists is selecting data to create a feature vector, simplifying reality with sophisticated algorithms. This feature vector transforms the learning step into a condensed model that describes the data. (Figure 4).

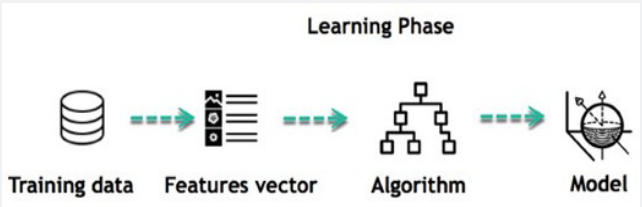

Figure 4: Learning phase.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

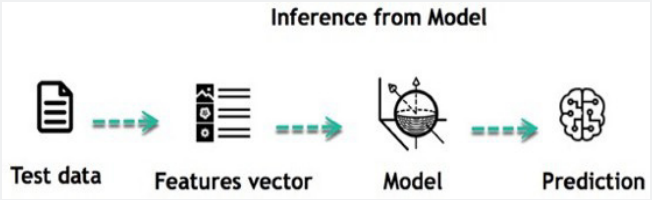

Figure 5: Inference model.

Key Points of Machine Learning Programs (Table 1)

Table 1:

The Life of Machine Learning Programs is Straightforward and Can Be Summarized in The Following Points:
1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

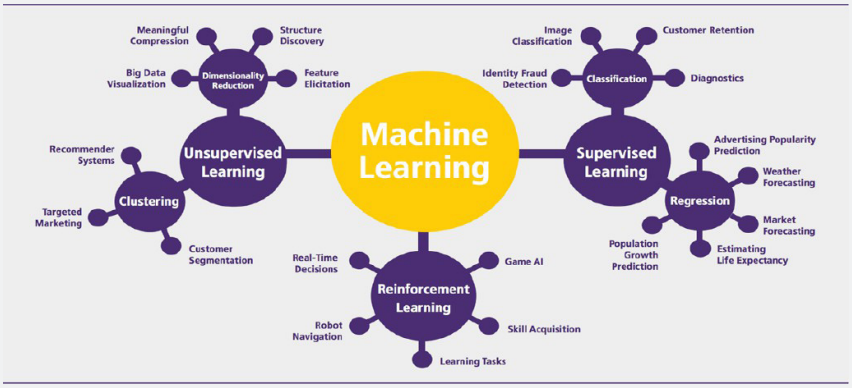

Figure 6: Machine Learning Algorithms.

Machine learning Algorithms

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms.

Supervised Learning: An algorithm learns the link between given inputs and a particular output using training data and feedback from humans. For instance, a practitioner can forecast sales using input data such as marketing expenses and weather predictions.

There are two categories of supervised learning:

- Classification task
- Regression task

Classification: To determine a customer's gender for a commercial, information is extracted from the database, including height, weight, occupation, salary, and purchase history. The classifier's goal is to assign a probability label (male or female) based on these features. Once the model learns to distinguish between genders, it can be used for predictions with new data. For example, if the classifier predicts a 70% probability of being male and 30% female, the algorithm confidently assigns the customer as male. Classifiers can have multiple classes for predicting items, like glass, table, shoes, each representing a different class [2,3].

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Types of Algorithms:(Table 2)

Algorithm Name	Description	Type
Linear regression	Finds a way to correlate each feature to the output to help predict future values.	Regression
Logistic regression	Extension of linear regression that's used for classification tasks. The output variable 3is binary (e.g., only black or white) rather the continuous (e.g., an infinite list of potential colors)	Classification
Decision tree	Highly interpretable classification or regression model that splits data feature values into branches at decision nodes until a final decision output is made	Regression Classification
Naive Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an even with the independent probability of each feature that can affect the even	Regression Classification
Support vector machine	Support Vector Machine, or SVM, is typically used for the classification task. SVM algorithm finds a hyper-plane that optimally divided the classes. It is best used with a non-linear solver.	Regression (non-very common Classification
Random forest	The algorithm, based on a decision tree, significantly enhances accuracy by employing a random forest. This approach generates numerous simple decision trees and utilizes the 'majority vote' method to deter-mine the final label for classification tasks. In regression tasks, the final prediction is the average of all the trees' predictions.	Regression Classification
AdaBoost	Classification or regression technique that uses models to come up with a decision but weighs them based on their accuracy i predicting the outcome	Regression Classification
Gradient-boosting trees	Gradient-boosting trees is a state-of-the-art classification/ regression technique. It is focusing on the error committed by the previous trees and tries to correct it.	Regression Classification

Unsupervised Learning: In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns) (Table 3)

Algorithm	Description	Type
K-means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as deter- mined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommender system	Help to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension

Machine Learning (ML) Algorithm and Practical Application

Numerous machine learning algorithms exist, chosen based on specific goals. In the following flower prediction example, ten algorithms predict flower types based on petal dimensions. The dataset is depicted in the top left image, divided into red, light blue, and dark blue groups. Classifications include the upper left of the second image in the red group, the middle exhibiting ambiguity and light blue, and the bottom in the dark category. Subsequent images illustrate various algorithms attempting to categorize the data. (Figure 7).

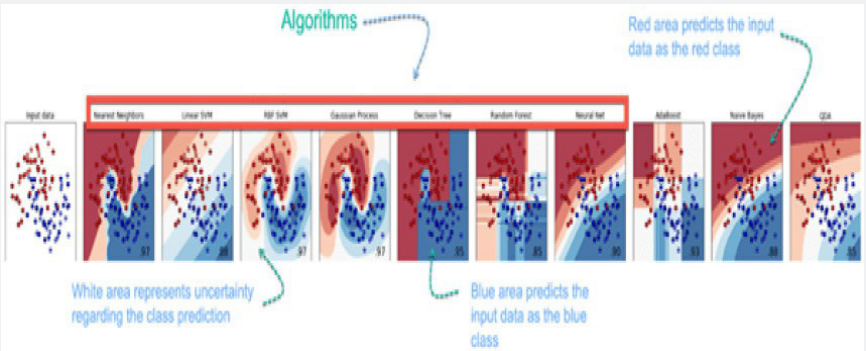

Figure 7: Algorithms categories in the data sets.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Machine learning as an AI subfield

Machine learning originated from AI research, initially exploring computers learning from data using symbolic techniques and "neural networks" like perceptron. These early efforts were later recognized as reimagined versions of generalized linear models in statistics. The divergence between AI and machine learning grew as AI favored a logical, knowledge-based approach, while machine learning focused on practical problem-solving using statistics and probability theory. In the 1990s, machine learning became a distinct field, moving away from the symbolic approaches of AI. The debate over whether all machine learning qualifies as AI persisted, with some arguing that only the "intelligent" subset does. Judea Pearl's distinction in The Book of Why emphasizes that AI involves active interaction for goal achievement, while machine learning relies on passive observations for learning and prediction.

Principal component analysis and cluster analysis

Feature learning is a vital preprocessing step for classification or predictions, utilizing algorithms to transform input data into a usable format without manual feature engineering. Supervised methods like supervised dictionary learning, multilayer perceptron, and artificial neural networks work with labeled data, while unsupervised techniques such as dictionary learning, independent component analysis, and autoencoders operate on unlabeled data. Manifold learning requires low-dimensional representations, enforcing sparsity through sparse coding techniques. Multilinear subspace learning directly extracts low-dimensional representations from tensor data. Deep learning hierarchically identifies abstract features. Intelligent machines separate variation sources through learned representations. Feature learning addresses the need for manageable data in machine learning tasks, with sparse dictionary learning (e.g., K-SVD) applied in various applications, linking new examples to classes based on the sparsest representation in the dictionary.

Analyzing Anomalies

Anomaly detection, a data mining process for outlier discovery, identifies unusual occurrences indicating issues like fraud or structural flaws. In the context of misuse detection, periods of inactivity are more intriguing than rare items. Traditional statistical outlier definitions pose challenges for unsupervised algorithms, prompting an alternative approach using cluster analysis to identify micro-clusters. Anomaly detection methods categorize into three groups: unsupervised strategies identify instances deviating most from the dataset; supervised approaches train a classifier on labeled "normal" and "abnormal" data; and semi-supervised techniques model normal behavior before assessing a test instance's likelihood of conforming.

Robot Learning: In developmental robotics, robot learning algorithms generate their own sequences of learning experiences, also known as a curriculum, to cumulatively acquire new skills through self-guided exploration and social interaction with humans. These robots use guidance mechanisms such as active learning, maturation, motor synergies and imitation.

Rule of Association: Association rule learning, a rule-based machine learning technique, identifies meaningful connections between variables in large databases using a metric of "interestingness." Agrawal, Imieliski, and Swami applied association rules to supermarket transaction data, revealing patterns like the likelihood of a customer buying hamburger meat when purchasing potatoes and onions together. This knowledge informs marketing decisions and finds diverse applications in areas like Web usage mining and intrusion detection. On the other hand, Learning Classifier Systems (LCS) are rule-based algorithms that combine learning components with genetic algorithms for rule discovery, focusing on context-dependent rules. Inductive Logic Programming (ILP) employs logic programming to represent input instances and hypotheses, finding utility in natural language processing and bioinformatics. Established by Gordon Plotkin and Ehud Shapiro, the foundational framework for inductive machine learning was implemented in their 1981 Model Inference System using Prolog to infer logic programs from examples.

Models: Creating a model that can process more data to create predictions after being trained on a set of training data is called ma- chine learning. Machine learning systems have been studied and used with a variety of models.

An Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) emulates the interconnected structure of neurons in the human brain, with circular nodes representing artificial neurons and arrows indicating connections between them. Inspired by biological neural networks, ANNs, or connectionist systems, learn tasks through examples without specific rules. Each connection transmits a signal, and the network processes information through layers of artificial neurons, potentially with changing weights and thresholds. Originally designed to mimic the human brain, ANNs shifted focus to task execution, finding applications in computer vision, speech recognition, machine translation, social network filtering, games, and medical diagnosis. Deep learning, involving ANNs with multiple hidden layers, simulates how the brain processes vision and hearing, achieving success in computer vision and speech recognition.

A Decision Tree in ML

Decision tree learning uses a tree structure to move from item observations to predictions about the target value, applied in data mining and machine learning. It includes classification trees for discrete targets and regression trees for continuous values. Beyond data summarization, decision trees serve as formal tools for decision analysis, facilitating informed decision-making based on the resulting classification tree.

Support Vector Machines (SVMs) in ML

Support Vector Networks (SVMs), commonly known as Support Vector Machines (SVMs), are supervised learning techniques for regression and classification. The SVM training

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

algorithm creates a model for binary, non-probabilistic, and linear classification, predicting whether a new example belongs to one of two categories based on labeled training examples. Using the kernel trick, SVMs can also perform non-linear classification by implicitly mapping inputs into high-dimensional feature spaces.

Analysis of Regression

Regression analysis utilizes statistical techniques to discern relationships between input variables and features. Linear regression, a common method, establishes a line using criteria like ordinary least squares to best fit the data. Regularization, such as ridge regression, mitigates overfitting and bias. Logistic regression addresses non-linear challenges in statistical classification, while kernel regression introduces non-linearity via the kernel trick. Effective machine learning models require representative data for comprehensive training, spanning text corpora, image collections, and specific customer information. The risk of overfitting and biased predictions underscores the importance of meticulous data preparation to avoid algorithmic bias and skewed outcomes.

Research Design, Objectives and Methods

Introduction

The study outlines the research design and methods to evaluate performance metrics for heart disease identification using Machine Learning. Accurate predictions are crucial, as inaccurate ones can be lethal. The study employs various machine learning methods and data visualization techniques on a dataset with 14 key attributes for heart disease prediction. The research includes data pre-treatment and applies machine learning algorithms to datasets of different sizes, examining stability, accuracy, and precision. Heart disease, a leading cause of global deaths, is influenced by factors like high cholesterol, obesity, and hypertension. The American Heart Association provides warning signs, emphasizing the importance of prompt attention, especially for males who have a higher risk than females.

Background of the study

Millions of people worldwide suffer from heart disease, which is also the leading cause of death worldwide. To lower the true cost of di- agnostic tests, medical diagnosis should be quick, accurate, and supported by computer technology. Data mining is a software method that enables computers to construct and categorize different features. Several machine learning methods are used in this research to forecast and identify cardiac disease. And employed a variety of machine learning algorithms in this, along with the data cleaning procedures, evaluation, and description of the research dataset.

Research Methodology

This research employs machine learning techniques to simplify predicting cardiac disease, benefiting both patients and physicians. Multiple machine learning techniques are applied to the dataset to determine essential characteristics for improved precision. Identifying the most influential features helps streamline trials and potentially save costs by excluding less impactful patient characteristics.

Data Preparation

Large amounts of missing and noisy data are present in real-life information or data. This data has already been pre-processed to get around these problems and make confident forecasts. Cleaning up the data after collection, which can have missing values or be noisy. This data needs to be cleansed of noise and the missing values filled in in in order to produce an accurate and useful result. Transformation is the process of converting data from one format to another so that it is easier to understand.

Using Algorithm

This study explores machine learning methods (K Nearest Neighbors, logistic regression, and random forest classifiers) for accurate heart disease identification. The three-stage approach involves data collection, extraction of relevant values, and pre-processing. The model employs KNN, logistic regression, and random forest as classifiers, assessing accuracy with various indicators and utilizing 13 health-related variables for prediction. AI and machine learning, vital in scientific advancements, offer potential solutions in healthcare, particularly for cardiovascular diseases. The rising literature on AI applications in cardiovascular disorders emphasizes the technology's promising role. Addressing cardiovascular diseases is critical, constituting 31% of global fatalities, according to the World Health Organization. The increasing number of AI-related papers in cardiovascular literature underscores the technology's potential impact on healthcare. (Figure 8).

AI applications in healthcare, while advancing and expected to enhance expert judgment, face a "black-box dilemma." The model, without user involvement, accepts input and produces results. In critical applications, especially in medicine or when replacing human roles with machines, this poses trust issues and potential catastrophic outcomes.

Motivation for the study

Cardiovascular Diseases (CVDs), including coronary and cerebrovascular diseases, pose a global health threat with 18 million deaths, requiring swift symptom identification for prevention. AI systems utilize diverse algorithms for pattern recognition, emphasizing accuracy. Beyond metrics, trust is crucial, aligning AI with professional practices. Establishing a categorization system based on physical symptoms enhances user reliance. This initiative focuses on innovative heart disease prediction using a hybrid dataset, addressing CVD challenges through advanced AI applications. Specific goals include the identification of crucial characteristics and transforming healthcare practices through technology. The following are the work's specific goals:

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

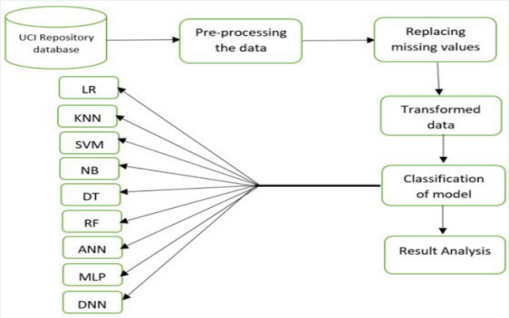

Figure 8: AI and ML applications in healthcare.

Objectives of the study

-To study the Machine Learning application in heart disease identification.

-To analyze the factors impacting on heart disease through using the Machine Learning application under Artificial Intelligence systems.

-To study the factors that affect heart disease risk the most across three separate datasets.

-To combine two separate datasets to create a hybrid dataset and to use three datasets.

-One hybrid dataset to apply several machine learning algorithms for the prediction of heart disease.

-To suggest efficient and effective measures to diagnose the heart disease identification through Machine Learning

Methodology

This project utilizes three datasets from Kaggle and the UCI machine learning repository, employing Python for data analysis through the Anaconda distribution. Various supervised machine learning algorithms, such as support vector machines, decision trees, k-nearest neighbors, naive Bayes, random forest, and logistic regression, are employed, including an ensemble technique, the majority voting classifier. The study explores different parameters and methods, such as adjusting C and gamma values for support vector machines and tuning k values for k-nearest neighbors. Feature selection is conducted using the univariate method, and a hybrid dataset is created from two distinct datasets, standardized, and evaluated. The research compares outcomes with literature, focusing on achieving the best accuracy in diagnosing heart disease.

Anaconda Distribution Package Configuration

Anaconda, a free and open-source distribution for R and Python, simplifies large-scale data processing and package management. This study leverages Anaconda's graphical user interface, Anaconda Navigator, for program launch and package, environment, and channel management. The navigator integrates various tools, such as RStudio, Spyder, Orange, Jupiter, and Jupiter Notebook. Specifically, Jupiter Notebook is used for running essential data analysis codes in an interactive, web-based computational environment.

Algorithm for Disease Prediction

Algorithm: Detection of heart disease using classifiers
Input: Heart disease dataset with several attributes
Output: Accuracy score/Confusion matrix/Classification report of predicted values. (Figure 9).

Process:

Step 1: Import libraries of sklearn, pandas, numpy

Step 2: Import the classifier functions

Step 3: Import *train_test_split* function

Or Import *cross_val_score* function

Step 4: i) Import *accuracy_score* function

ii) Import *confusion_matrix* function

iii) Import *classification_report* function

Step 5: Load the CSV file containing data using *read_csv*() function

Step 6: Separate the input and target attributes

Step 7: i) For holdout method, separate the train and test data using *train_test_split*() function

ii) Model the classifier using *model.fit*() function

iii) Predict the test data using *model.predict*() function

Or apply cross validation using *cross_val_score()* function

Step 8: i) Find accuracy using *accuracy_score*() function

ii) Find confusion matrix using *confusion_matrix*() function

iii) Find classification report using *classification_report*() function

Figure 9

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Performance Metrics

This study evaluates heart disease identification using various metrics, such as training accuracy, testing accuracy, precision, recall, and F1-measure, employing terms like True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). TP indicates correctly identified heart disease cases, TN represents correctly identified healthy cases, FN refers to missed heart disease cases, and FP signifies incorrectly identified cases. Accuracy, denoted by ac, measures the percentage of correctly identified vectors among all normal and abnormal samples. (Figure 10).

$$ac = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

Training accuracy and testing accuracy are defined as the accuracy obtained for training and testing samples, respectively. Precision, pr, can be mathematically written as follows.

$$pr = \frac{TP}{TP + FP} \quad (3.2)$$

The term recall, re, is also known as sensitivity given by

$$re = \frac{TP}{TP + FN} \quad (3.3)$$

The F1-Measure, f_1, is given by

$$f_1 = \frac{2 \times pr \times re}{pr + re} \quad (3.4)$$

Confusion matrix determines the performance of a classification algorithm by showing the predicted instances and the actual instances.

Figure 10:

Overview of Conceptual and Theoretical Aspects of Machine Learning Application in Heath Industry

Introduction

Machine learning in healthcare can enhance diagnostic tools for medical image analysis. ML algorithms applied to medical imaging, such as X-rays or MRI scans, utilize pattern recognition to identify specific conditions. These applications analyze vast datasets, aiding healthcare professionals in making more informed judgments.

Applications for machine learning (ML)

Machine Learning (ML) applications are pervasive and play a vital role in diverse practical domains, notably healthcare and patient data security. This study explores the utilization of ML to analyze medical records and predict diseases, addressing gaps in effective ML methods and applications within the healthcare industry.

Highlights of Machine Learning: (Table 6)

Table 6: Python
Python is an interpreted language, allowing runtime processing without the need for pre-compilation, like PHP and PERL.
Python is Interactive - To write programs, sit at a Python prompt and communicate with the interpreter directly.
Python's support for object -oriented programming, which encapsulates code within objects, makes it an object-oriented language.
Python is a fantastic language for beginning programmers and facilitates the creation of a wide range of programs, including simple text processing, web browsers, and games.

Machine Learning

Machine learning, a subset of artificial intelligence, creates predictive systems by learning from experiences and building models on datasets to uncover hidden patterns. In healthcare, machine learning applications optimize trial samples, increase data points, and play a pivotal role in early epidemic detection. The study emphasizes the transformative impact of machine learning on healthcare operations, allowing professionals to focus on patient care and addressing global healthcare challenges. (Figure 11).

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Figure 11: Smart features of machine learning for healthcare domain.

Machine Learning Algorithms:

Supervised Learning Algorithms:
- Linear Regression
- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- Naive Bayes

Unsupervised Learning Algorithms:
- K-Means
- Hierarchical Clustering
- Principal Component Analysis (PCA
- Independent Component Analysis (ICA)

Semi-Supervised Learning Algorithms:
- Self-Training
- Multi-View Learning

Reinforcement Learning Algorithms:
- Q-Learning
- Deep Q Network (DQN)

Neural Network Architectures:
- Feedforward Neural Networks:
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
- Long Short-Term Memory (LSTM)
- Generative Adversarial Networks (GAN)

Ensemble Learning Algorithms:
- AdaBoost
- Gradient Boosting Machines (GBM)
- XGBoost

Clustering Algorithms:
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Mean Shift
- Affinity Propagation

Dimensionality Reduction Techniques:
- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- Autoencoders

Time Series Forecasting:
- ARIMA (AutoRegressive Integrated Moving Average)
- Prophet

Anomaly Detection:
- Isolation Forest
- One-Class SVM

Natural Language Processing (NLP) Algorithms:
- Word2Vec
- BERT (Bidirectional Encoder Representations from Transformers)
- Named Entity Recognition (NER)

Recommender Systems:
- Collaborative Filtering
- Content-Based Filtering
- Matrix Factorization

Transfer Learning:
- Fine-tuning
- Domain Adaptation

Hyperparameter Tuning Algorithms:
- Grid Search
- Random Search

Evolutionary Algorithms:
- Genetic Algorithms
- Particle Swarm Optimization (PSO)

Ensemble Learning Techniques:
- Stacking
- Bagging (Bootstrap Aggregating)

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

The potential of machine learning for healthcare

Rapidly advancing, machine learning (ML) holds intriguing promise for healthcare. Addressing complex challenges, from deciphering extensive patient data to elevating care quality and personalization, ML operates under the broader umbrella of artificial intelligence. Defined as a statistical method applying models to data and training AI, ML enables systems to discern patterns in vast datasets for future predictions. It involves developing algorithms and applications based on historical and real-time data, benefiting not only healthcare but also industries like banking, manufacturing, hospitality, agriculture, and even charitable efforts like humanitarian relief.

Machine learning trends in healthcare

Some of the most important machine learning developments in healthcare to be aware of are as follows:

1. Precision medicine, leveraging machine learning extensively, utilizes patient data to predict effective treatment procedures, facilitating highly customized and improved clinical outcomes.

2. Categorization applications – These procedures include determining whether or how likely it is that a patient will develop a specific

1. condition. Information can be used to guide policy, develop strong preventative measures, and assist providers in capacity planning.

3. Imaging analysis - Radiology and pathology images are already subject to machine learning analysis. It's also used to swiftly classify large numbers of photos. The application of machine learning to these procedures may advance in sophistication and accuracy during the ensuing years.

4. Machine learning streamlines claims and payment management by enhancing accuracy in claims data, reducing the risk of false claims, and simplifying the administration process for insurers, governments, and providers.

5. Machine learning supports various administrative tasks, including claims processing, clinical documentation, revenue cycle management, and medical data management. It extends to patient-facing solutions like telehealth chatbots and mental health assistance, optimizing interactions without direct doctor involvement.

6. Machine learning excels in predictive modeling and health policy. Population health models can pinpoint at-risk populations for specific health issues or readmissions. Moreover, it leverages social determinants of health data to inform policy decisions, enabling targeted interventions for preventable diseases like diabetes and heart disease.

7. Machine learning is instrumental in processing vast Electronic Health Records (EHR) data, especially unstructured free-form text submissions. Its capacity to rapidly interpret such data at scale for millions of patients enhances decision-making across the entire patient-care cycle.

8. Machine learning, notably through Clinical Decision Support tools (CDS), is increasingly employed for diagnosis and treatment recommendations. CDS systems analyze extensive data to enhance healthcare providers' decision-making, offering alerts for potential issues and facilitating preventive actions.

9. In drug development, machine learning aids researchers in forming trial cohorts, enhancing research efficiency, and expediting medication development by swiftly identifying patterns and trends for data-driven decisions.

The Importance of Machine Learning in Healthcare

In the healthcare shift to value-based care, machine learning facilitates cost reduction and improved clinical quality by providing end-to-end visibility into operations, data management, research, diagnosis, treatment, and regulatory procedures.

- Suggestions - Machine learning algorithms can extract and provide critical medical information without requiring you to look for it specifically.

- Classification - Aids in identifying and describing the type of medical case or condition a patient is experiencing.

- Prediction – Using past and current data coupled with common trends, smart algorithms can make an accurate prognosis on how future developments & events will unfold.

- Clustering – Can be leveraged to club similar medical cases together for analyzing patterns and conducting research.

- Ranking – Helps extract the most relevant information first, making the search fast and easy. • Detecting Anomalies – Enables easy identification of specimens that stand out from common patterns for timely intervention

- Automation – Can put standard, repetitive clinical operations such as appointment scheduling, inventory management, and data entry on the autopilot mode [4,5].

Data Analysis and Interpretation

Introduction

The research develops a cost-effective cardiovascular warning system using real-time data from the Clinical Foundation, achieving 87% accuracy with the Extreme Machine Learning technique. Outperforming ANN, this algorithm, with five outputs (0–4), analyzes new patients using past information. Future work targets missing characteristics and accuracy enhancement. The study establishes a cardiac disorder prediction model, comparing algorithms, and identifies the highest accuracy algorithm as the best for forecasting the illness.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Table 7:	
S. No	Methods & Description
1	**GET**: Sends data in unencrypted form to the server. Most common method.
2	**HEAD:** Same as GET, but without response body
3	**POST:** Used to send HTML form data to server. Data received by POST method is not cached by server.
4	**PUT:** Replaces all current representations of the target resource with the uploaded content.
5	**DELETE:** Removes all current representations of the target resource given by a URL

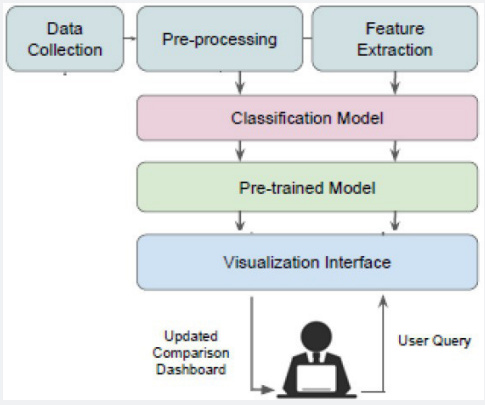

Figure 12: Diagram of data flow.

System architecture (Figure 12)

1. The Data Flow Diagram (DFD) is a bubble chart. It is a straightforward graphical formalism that can be used to represent a system in terms of the data that is fed into it, the different operations that are performed on it, and the data that is produced as a result of those operations.

2. One of the key modelling tools is the data flow diagram. The system's component models are created using it. These elements include. The system's operation, the data it uses, a third party that engages with it, and the way information moves through it.

3. DFD demonstrates the information's flow through the system and the various changes that affect it. It is a graphical method for rep- resenting information flow and the changes made to data as it travels from input to output.

4. Another name for DFD is bubble chart. Any degree of abstraction for a system can be represented by a DFD. DFD can be divided into

stages that correspond to escalating functional complexity and information flow. (Figure 13)

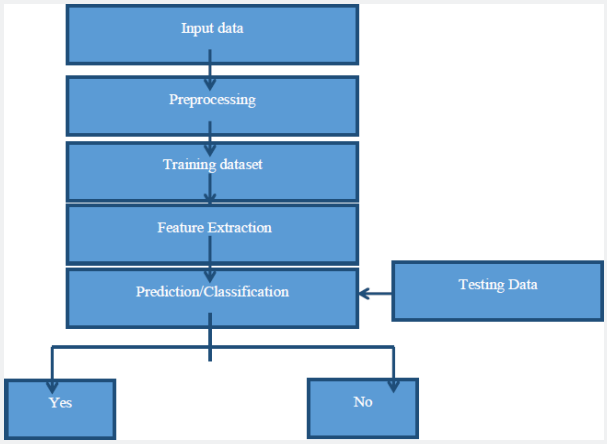

Figure 13: Information flow.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

System requirements:

Hardware requirements:	Software requirements:
System: Pentium IV 2.4 GHz. Hard Disk: 40 GB. Floppy Drive: 1.44 Mb. Monitor: 15 VGA Color. Mouse: Logitech. Ram: 512 Mb.	Operating system: Windows 7. Coding Language: Python Database: MYSQL

Feasibility study

In this stage, the project's viability is assessed, and a business proposal is presented with a very basic project plan and some cost projections. The proposed system's practicality must be investigated during system analysis. This will guarantee that the suggested solution won't burden the business. Understanding the main system requirements is crucial for the feasibility analysis. This will guarantee that the suggested solution won't burden the business. Understanding the main system requirements is crucial for the feasibility analysis. The feasibility analysis takes three important factors into account: (Table 8).

Table 8:
Economic feasibility
Technological potential
Social acceptability

Table 9:
Focus of functional testing is on the following areas:
Valid Input: Recognized valid input classes need to be accepted.
Invalid Input: Defined categories of invalid input need to be rejected.
Functions: It is necessary to use the listed functions.
Output: Certain classes of application outputs need to be tested.
Systems/Procedures: It is necessary to call interacting systems or processes.

Economic feasibility

This study assesses the financial impact of implementing the system within the company's limited budget for research and development. The system remained under budget by utilizing predominantly public domain technology, with expenses limited to acquiring specific, non-public domain goods.

Technical feasibility

This study evaluates the technical feasibility of the system, emphasizing the need for low demands on available technical resources to avoid burdening the client. The developed system is designed with modest requirements, minimizing the need for substantial changes during implementation.

Social feasibility

The study aims to gauge user acceptance of the system, emphasizing effective instructional design to ensure user-friendly operation without causing intimidation. Techniques for informing and familiarizing users significantly impact acceptance levels. Building user confidence is crucial for encouraging constructive criticism from the ultimate system user.

Process of Information

1. Designing the input involves transforming a user-centered input description into a computer-based system, crucial for error prevention in data entry and clear instructions for obtaining accurate information from the computerized system.

2. Input design simplifies and error-proofs data entry by creating user-friendly interfaces for handling large data volumes. The design ensures easy data manipulation and incorporates record viewing capabilities.

3. Input design ensures the verification of entered data, utilizing screens, and providing relevant messages to guide users and prevent confusion. The primary objective is to create a straightforward and easily understandable input layout.

Output Planning

Quality output, meeting user needs, is crucial for effective information transmission in any system. Output design determines how information is presented for immediate use and in hard copy. It serves as the primary source of information for users, aiding decision-making. A well-ordered and thoughtful design ensures efficient and intelligent interaction between the system and users. Creating correct and user-friendly output components is essential for successful system utilization. Analysis of computer-generated output focuses on pinpointing the specific output required to meet specifications.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Verify an activity (Figure 14-23)

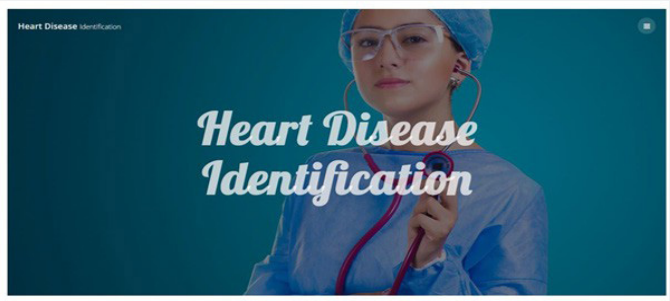

Figure 14

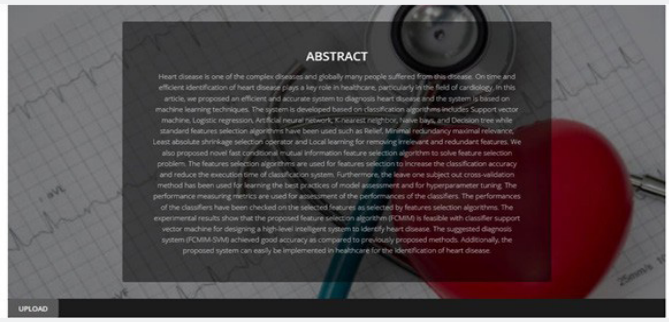

Figure 15

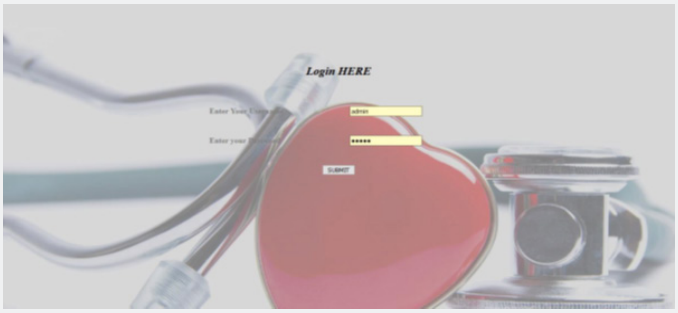

Figure 16

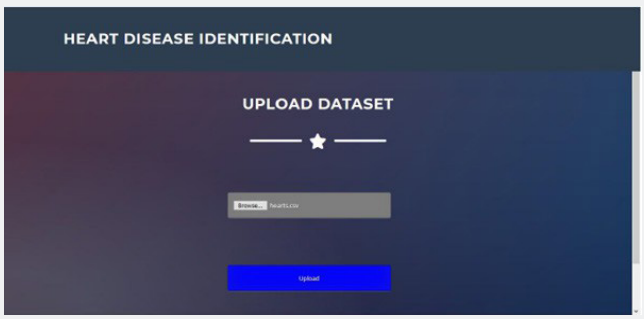

Figure 16

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

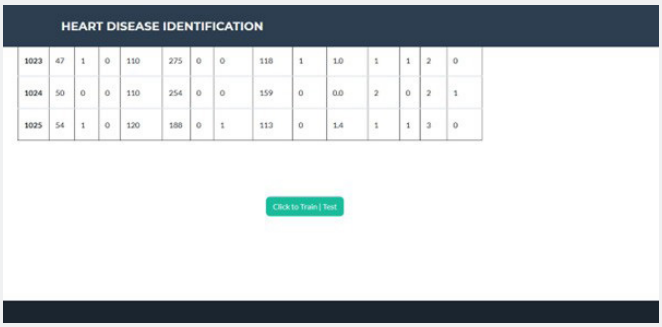

Figure 17

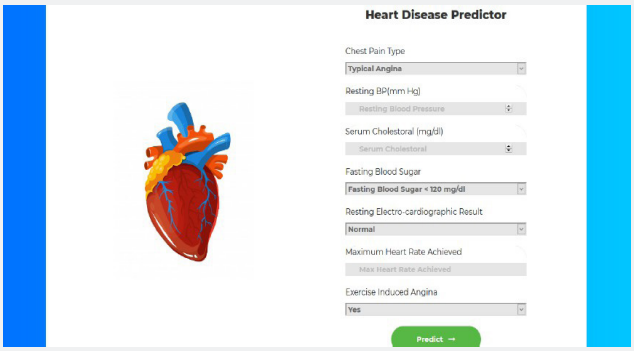

Figure 18

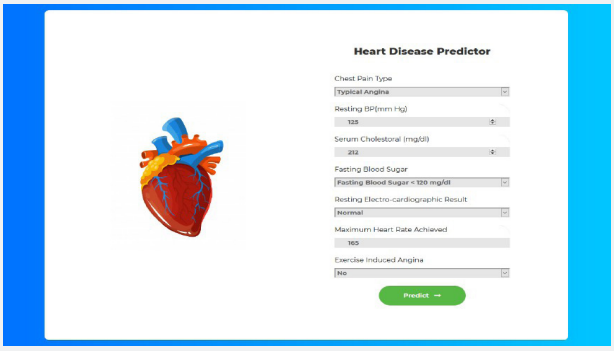

Figure 19

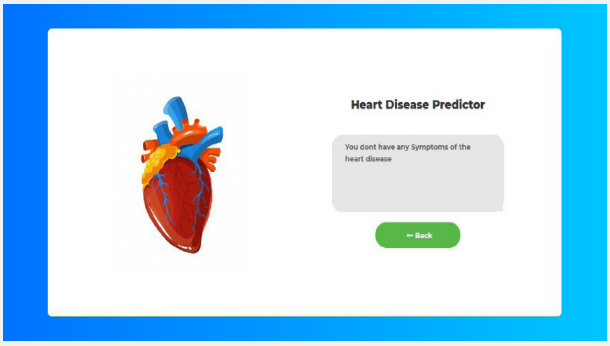

Figure 20

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

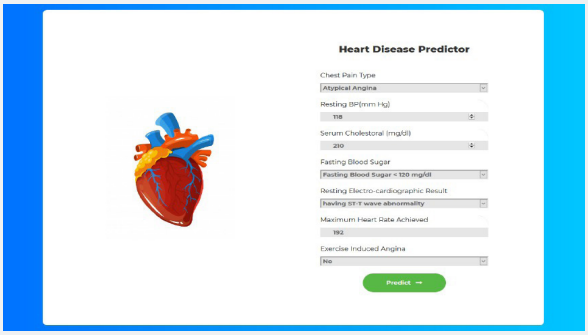

Figure 21

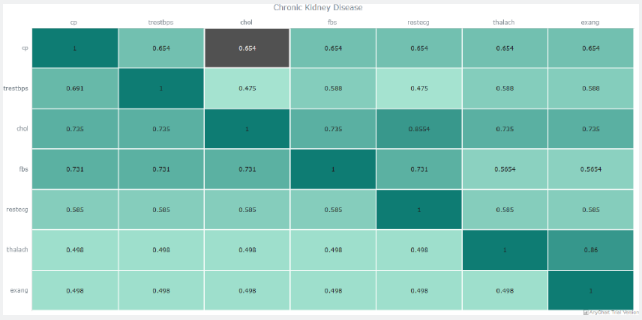

Figure 22

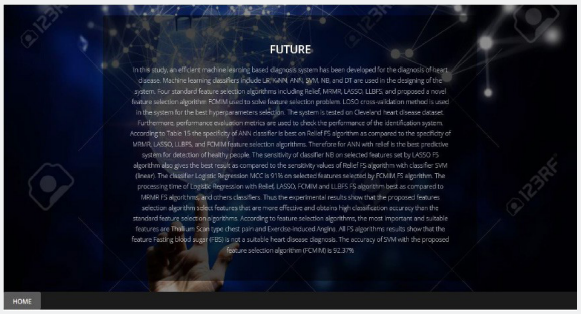

Figure 23

Model Selection: The Random Forests machine learning technique, absent in the original study but implemented here due to its 91.7% test set accuracy, is a favored method known for its strong predictive accuracy, minimal overfitting, and interpretability. Its simplicity allows for the determination of each variable's significance in tree decisions, contributing to overall interpretability. Utilizing a random forest for feature selection falls under embedded approaches, combining the benefits of filter and wrapper techniques for high accuracy, improved generalization, and simplicity.

Main Features from the Actual Dataset were Selected for Analysis and Prediction:

1. **Angina (Chest Pain):** Angina is a type of discomfort or pain in the chest that develops when the heart muscle isn't receiving enough oxygen-rich blood. Chest may experience pressure or squeezing. You may also experience pain in your back, neck, jaw, shoulders, or arms. Even the pain from angina can resemble dyspepsia.

2. **Resting Blood Pressure:** High blood pressure over a long period of time can harm the arteries that supply your heart. Your risk is even higher if you have high blood pressure together with another illness, such as diabetes, high cholesterol, or obesity.

3. **Serum Cholesterol:** The most likely cause of artery narrowing is a high level of Low-Density Lipoprotein (LDL) cholesterol, also known as "bad" cholesterol. Your risk of a heart attack is also increased by having high blood levels of triglycerides, a type of blood fat connected to your diet. Yet, having high levels of HDL cholesterol (the "good" cholesterol) reduces your risk of having a heart attack.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

4. Fasting Blood Sugar: When your body doesn't produce enough insulin or doesn't respond to it appropriately, your blood sugar levels rise, which raises your chance of having a heart attack.

5. Resting ECG: The USPSTF comes to the conclusion with a moderate degree of certainty that, for individuals at low risk for cardio- vascular disease, the risks of screening with resting or exercise ECGs are either equivalent to or greater than the potential benefits. There is currently inadequate information to determine the balance between screening's advantages and disadvantages for those at intermediate to high risk. The risk of a heart attack is reduced by (HDL) cholesterol, sometimes known as "good" cholesterol.

6. Maximum heart Rate Attained: The rise in cardiovascular risk brought on by the heart rate acceleration was equal to the rise in risk brought on by high blood pressure. The risk of cardiac death has been demonstrated to increase by at least 20% for every 10 beats per minute increase in heart rate, and this risk increase is comparable to that seen with an increase in systolic blood pressure of 10mmHg.

7. Exercise Induced Angina: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from moderate to severe. Angina is typically felt in the middle of the chest, but it can also affect one or both shoulders, as well as your back, neck, jaw, or arm. You can feel it in your hands as well. kinds of angina Angina Pectoris / Stable Angina / Unstable Angina Microvascular Angina c. Variant (Prinzmetal) Angina d.

On the test set, we achieved an accuracy of 91.7%.

Diagnosis of Heart Disease using Machine Learning

This project uses three datasets to analyze heart disease prediction, emphasizing its importance as a leading cause of death. It covers symptoms, types of heart conditions, and key risk factors like age, sex, blood pressure, cholesterol, diabetes, and lifestyle choices. The work introduces machine learning approaches, including data mining, as essential tools for identifying cardiac diseases.

Logistic Regression Models in Predicting Heart Disease

This study predicts the risk of heart disease among the elderly using logistic regression models and data mining technology to identify key pathogenic factors. Comparing the accuracy with other algorithms, logistic regression proves valuable for heart disease prediction. Cardiovascular diseases, responsible for a significant global mortality rate, necessitate cost-effective and reliable prediction methods. Logistic regression, by extracting disease risk factors and providing real-time incidence probability, addresses this challenge, offering an efficient approach for early detection and improved quality of life, especially in low-income or developing countries [6-8].

a)Research Aim and Scope: The aim of this research has developed an efficient way to predict the presence of the cardiovascular

disease. The steps as mentioned below.

i. The UCI dataset is used to predict the disease.

ii. The Features are selected based on high positive correlation values with the target and used random order of data.

iii. The performance of the model is evaluated by Five different training and testing ratio of dataset.

iv. To check the behavior of the model with low to high training and testing data.

This study assesses the logistic regression algorithm's effectiveness in predicting heart disease by comparing its accuracy with other algorithms like Naive Bayes, SVM, and Neural Network. It aims to identify crucial characteristics and incidence probabilities for heart disease prediction. (Figure 24).

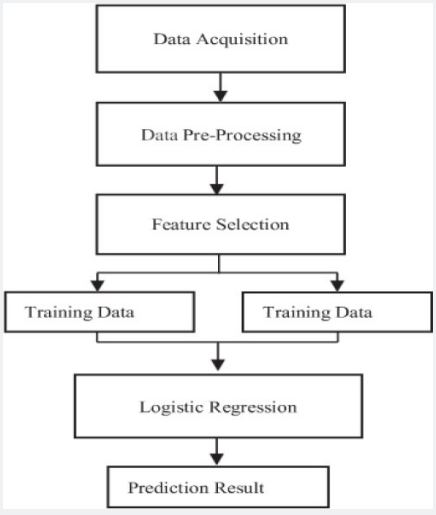

Figure 24: Flow chart of logistic regression model.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Algorithm: This paper focuses on the logistic regression model, widely used in machine learning for applications like data mining, disease diagnosis, and economic prediction. It specifically discusses its application in identifying risk factors for heart disease and predicting the probability of disease occurrence based on these factors. Logistic regression is commonly employed for classification tasks, especially those involving two categories, providing probabilities for each classification event.

Implementation: Logistic Regression Model: Dataset

The study collected data from the UCI machine learning repository, comprising 303 records with 14 attributes. Thirteen parameters served as eigenvalues for heart disease prediction, with one representing the output or forecast value for patients with heart disease ('num' for Numeric, 'nom' for Nominal). (Figure 25-36).

Attributes from UCI Dataset1

	Attribute	Description	Type
①	Age	Age in years	num
②	Sex	Sex (1 = male; 0 = female)	nom
③	Cp	chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic	nom
④	Trestbps	Resting blood pressure	num
⑤	Chol	Serum cholestoral in mg/dl	num
⑥	Fbs	Fasting blood sugar > 120 mg/dl	nom
⑦	Restecg	Resting electrocardiographicresults -- Value 0: normal -- Value 1: having ST-T waveabnormality -- Value 2: showing probable ordefinite leftventricular hypertrophy by Estes' criteria	nom
⑧	Thalach	Maximum heart rate achieved	num
⑨	Exang	Exercise induced angina	nom
⑩	**Oldpeak**	**ST depression induced by exercise relative to rest**	num

Figure 25:

Attributes from UCI Dataset2

11	Slope	The slope of the peak exerciseST segment Nominal -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping	nom
12	Ca	Number of major vessels (0–3) coloured by fluoroscopy	num
13	Thal	3 = normal; 6 = fixed defect; 7 =reversable defect	nom
14	Num	Diagnosis of heart disease(angiographic disease status) -- Value 0: no heart disease -- Value 1-4: presence of heartdisease	nom

Figure 26: UCI ML repository's Cleveland heart disease dataset—feature subset

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Attribute name	Attribute description
Age	Age in years
Sex	1 denotes male and 0 denotes female
CP	Chest pain type 1, typical angina; type 2, atypical angina; type 3, nonanginal pain; and type 4, asymptomatic
trestbps	Resting blood pressure (in mmHg at entry to the health center)
chol	Serum lipid level in mg/dL
fbs	1 denotes true, i.e., the fasting blood sugar level > 120 mg/dL; 0 denotes false
restecg	Resting ECG results: null, normal; 1, ST-T wave abnormality; and 2, probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; null = no)
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment (1, 2, and 3): 1, upsloping; 2, flat; and 3, downslopingca
	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia: 3 = normal, 6 = fixed defect, and 7 = reversible defect

Figure 27:

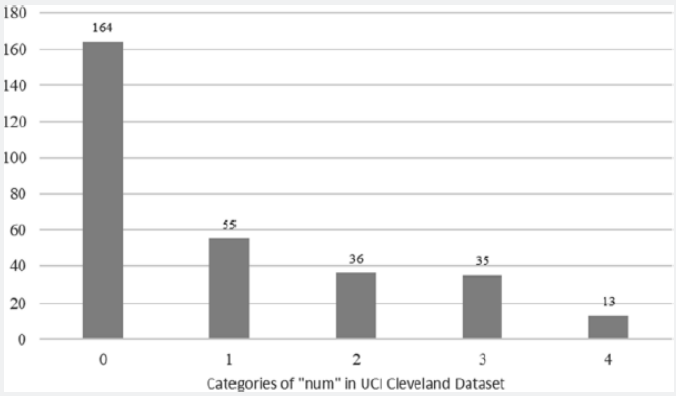

Figure 28: The proportion of 'num' in the data set.

Data Analysis of heart diseases in the dataset

a) Data Preprocessing: To accommodate the logistic regression's dichotomous nature and the varying degrees of disease in the output values (0 to 4), data preprocessing was performed. Incomplete data were removed, and predictive attributes for heart disease were transformed from multi-category to binary values. Diagnostic values 2 and 4 were converted to 1, resulting in a final dataset with values of 0 (no heart disease) and 1 (potential heart disease). (Table 11-13).

Table 11: Logistic regression in determining Heart issues: Variable.

Variables in the equation							
		B	S.E.	Wald	Df	Sig.	Exp(B)
Step 1[a]	male (1)	-0.462	0.132	12.318	1	0	0.63
	age	0.063	0.008	60.897	1	0	1.065
	education			1.549	3	0.671	
	education (1)	-0.032	0.197	0.027	1	0.869	0.968
	education (2)	-0.57	0.21	0.073	1	0.787	0.945
	education (3)	-0.247	0.239	1.065	1	0.302	0.781
	Current Smoker (1)	-.0.046	0.187	0.06	1	0.806	0.955
	Cigs Per Day	0.018	0.007	6.029	1	0.014	1.018
	BPMeds (1)	-0.38	0.283	0.018	1	0.893	0.963
	prevalent Stroke (1)	-0.787	0.569	1.917	1	0.166	0.455

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

	Prevalent Hyp (1)	-0.256	0.166	2.374	1	0.123	0.774
	diabetes (1)	0.14	0.4	0.123	1	0.726	1.15
	tot Chol	0.002	0.001	1.374	1	0.241	1.002
	sysBP	0.016	0.004	12,417	1	0	1.016
	diaBP	-0.005	0.008	0.448	1	0.503	995
	BMI	0.011	0.015	0.505	1	0.477	1.011
	heart Rate	-0.004	0.005	0.622	1	0.43	0.996
	glucose	0.008	0.003	9.275	1	0.002	1.008
	Constant	-6.707	1.242	29.154	1	0	0.001

Note*: a. Variable(s) entered on step 1: male, age, education, current smoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, duiabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose

Table 12: Statistical outline of subset attributes.

Attributes	Age	Sex	CP	Trestbps	Chol	Fbs	Resteeg	Thalach	Exang	Oldpeak	Slope
mean	54.44	0.68	3.16	131.69	246.69	0.15	0.99	149.61	0.33	1.04	1.6
std	9.04	0.47	0.96	17.6	51.78	0.36	0.99	22.88	0.47	1.16	0.62
min	29	0	1	94	126	0	0	71	0	0	1
25%	48	0	3	120	211	0	0	133.5	0	0	1
50%	56	1	3	130	241	0	1	153	0	0.8	2
75%	61	1	4	140	275	0	2	166	1	1.6	2
max	77	1	4	200	564	1	2	202	1	6.2	3

Table 13: Feature selection using correlation.

Features	Correlation
EXang	0.436757
Cp	0.433798
Oldpeak	0.430696
Thalach	0.421741
Ca	0.391724
Slope	0.344029

Cardiovascular disease, a leading cause of global mortality, necessitates early diagnosis for timely intervention. This study employs Logistic Regression (LR) on the UCI dataset to classify cardiac disease. Enhancements include data pre-processing through cleaning, handling missing values, and feature selection based on positive correlations with the target value. The LR model achieves an accuracy of 87.10% with a dataset split ratio of 90:10 for training and testing. (Table 14).

Table 14: Split percentage of training and test set.

Serial Number	Training Set	Test Set
1	50%	50%
2	60%	40%
3	70%	30%
4	80%	20%
5	90%	10%

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Machine learning, a crucial field for rapidly extracting valuable information from vast datasets, employs powerful models to uncover hidden patterns and correlations. Logistic regression classifiers yield accurate results in this context. (Table 15).

Table 15: Training and Testing.

Training and Testing					
90:10	80:20	70:30		60:40	50:50
87.10%	85.25	83.52%		81.97%	81.58%

Various ML algorithms, including logistic regression, decision tree, and support vector machine, were tested on the UCI dataset. Logistic regression achieved an accuracy of 82.56%, while logistic regression support vector machine reached 84.85%. (Table 16).

Table 16: Classification report of logistic regression classifier.

Confusion MatriX			Classification Report			
	Pos	Neg	Precision	Recall	F1-Score	Accuracy
Pos	15	2	0.857	0.857	0.857	87.10
Neg	2	12				

Logistic regression was tested with the UCI dataset at various ratios, and the table below illustrates the accuracies. The model achieved 87.10% accuracy with a training-testing split ratio of 90:10, showcasing improved accuracy with increased training data. (Table 17).

Table 17: Comparative result of logistic regression classifier.

	Year	Author	Tool/ Techniques	Logistic Regression
UCI DATASET	2019	Z. Khan, D.K. Mishra, V. Sharma, A. Sharma,	Rapid Miner (Logistic Regression)	82.56%
	2019	Z. Khan, D.K. Mishra, V. Sharma, A. Sharma,	Rapid Miner (Logistic Regression Support Vector Machine)	84.85%
	2020	WHO	Python	85.04%
	2020	K Uday	Python	85.71%
	2021	This study	Python sklearn (90:10)	87.10%

In the medical field, predicting cardiovascular disease is crucial, utilizing patient data to discern its presence or absence. Logistic Regression, a supervised machine learning algorithm, is employed for classification in this study. Enhancing performance involves preprocessing steps like cleaning and handling missing values. Feature selection is a vital aspect, contributing to algorithm accuracy and behavior.

b) Feature Selection: Apart from predicting heart disease probability, the experiment aims to identify key factors contributing to heart diseases. To achieve this, feature extraction is conducted before data analysis, aiming to unveil the pertinent pathogenic factors influencing heart issues. The focus is on understanding the relationship between age, blood pressure, and various characteristic values, providing insights for preventive health measures. (Figure 29,38).

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

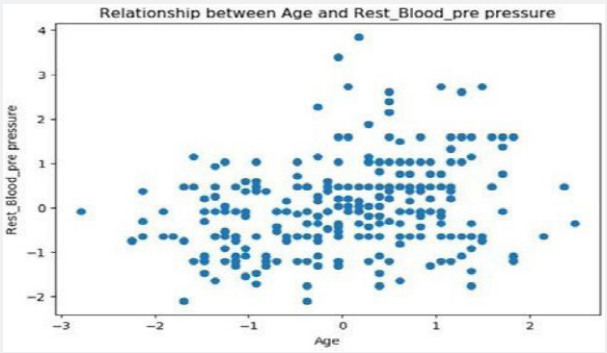

Figure 29: Relationship between age and blood pressure.

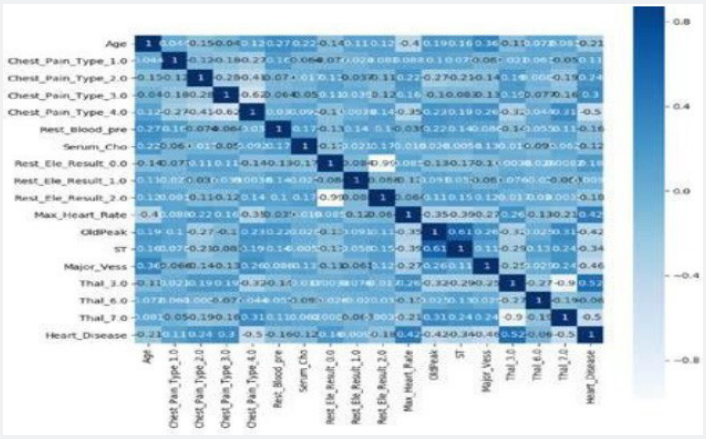

Figure 30: Correlation between each characteristic value.

According to the correlation between the eigenvalues, the combination with the most significant features was selected for data analysis, and different data mining techniques were used to test the selected combination.

Heart Disease Prediction using Exploratory Data Analysis

The healthcare industry handles vast volumes of big data, holding valuable insights for informed decision-making. Exploratory Data Analysis (EDA) in healthcare involves scrutinizing data for errors, identifying correlations among variables, and deriving patterns without statistical modeling. EDA is crucial for accurate predictions and preventive care, contributing to improved healthcare outcomes. This paper employs K-means clustering for predicting heart disease risk factors, analyzing a dataset with attributes like age, chest pain type, and blood pressure. The study emphasizes pre-processing, classifier performance, and evaluation metrics, demonstrating accurate predictions through visualized data.

Human-generated data has exceeded ten exabytes, highlighting the significance of EDA. The method uncovers hidden structures, identifies anomalies, and builds models to test assumptions. EDA is categorized into graphical/non-graphical and univariate/multivariate methods. Diagnostic methods in healthcare include invasive (incise procedures) and non-invasive approaches. Machine learning algorithms for non-invasive methods, such as SVM, K-means clustering, KNN, ANN, and Naive Bayes, play a vital role in disease diagnosis [9].

Heart Diseases

Heart disease, considered globally as the deadliest ailment, poses challenges in early diagnosis and treatment due to insufficient medical resources and personnel. Invasive techniques, relying on medical history and physical reports, are time-consuming and imprecise. Predicting heart disease based on symptoms like age and pulse rate using data analysis in healthcare is crucial for early detection and effective treatment. Technologies like ECG aid in screening irregular heartbeats. Risk factors include obesity, smoking, diabetes, blood pressure, and unhealthy diet. Unusual forms of heart disease, like acute spasms in coronary arteries, can lead to oxygen deprivation, affecting physiological systems beyond the heart. Early diagnosis is paramount for minimizing risks and enhancing the quality of life for heart patients.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Results and Discussion

This section presents the results of data analysis for predicting heart diseases, considering variables such as age, chest pain type, blood pressure, blood glucose level, ECG, heart rate, exercise angina, and four types of chest pain. The heart disease dataset undergoes effective preprocessing, eliminating unrelated records and handling missing values. The K-means algorithm is then applied to compose the preprocessed dataset, discussing four types of heart diseases: asymptomatic pain, atypical angina pain, non-anginal pain, and non-anginal pain. Histogram analysis shows a higher risk of heart disease in the age range of 50 to 55, where the development of coronary fatty streaks begins. (Figure 31).

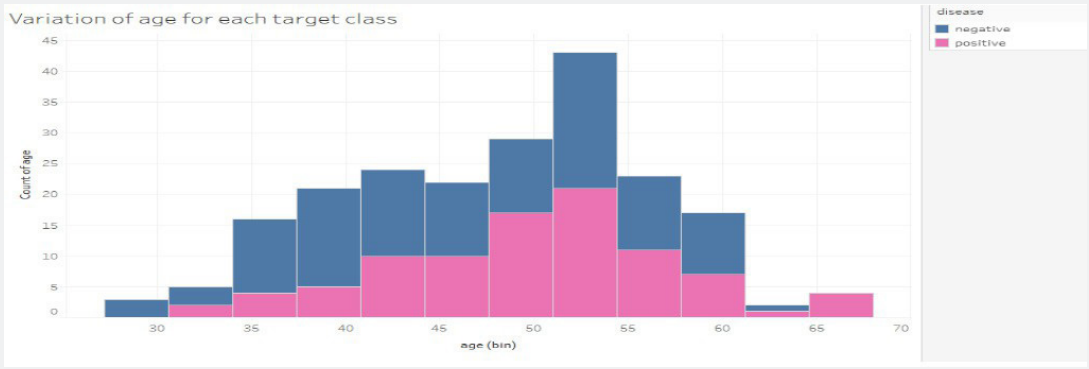

Figure 31: Histogram of variation of age for each target class.

Figure shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease (Figure 32).

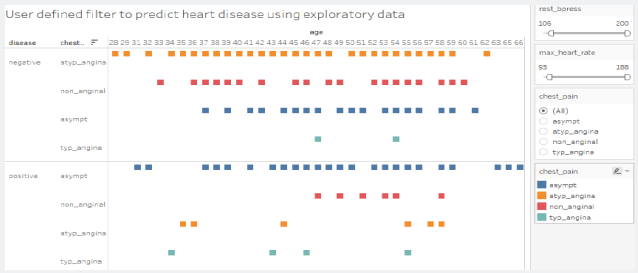

Figure 32: shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease.

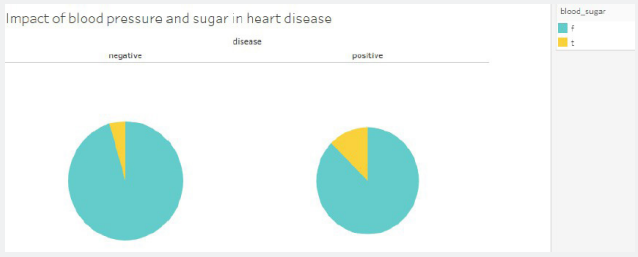

Figure 33: Impact of blood pressure and sugar in heart disease.

Figure shows the impact of blood pressure and sugar in heart disease. It is inferred that population with diabetics and high blood pressure is expected to get heart disease (Figure 33).

K-means Clustering: The K-means clustering algorithm is chosen for its efficiency, simplicity, capacity to generate even-sized clusters, and scalability in handling the dataset, ensuring accurate outputs with a minimum sum of squares. The dataset comprises 209 observations with 7 variables. (Figure 34)

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

$$\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

where S_k is the set of observations in the kth cluster and $\bar{x}_{kj}$ is the jth variable of the cluster center for the kth cluster.

Figure 34:

Chest Pain Type: Asymptomatic

The plot of Age vs. Max Heart Rate broken down by Disease. Colour shows details about disease.The screen shot of the clustering are described below (Table 19,20), (Figure 35)& (Table 21).

Table 19: The plot of Age vs. Max Heart Rate broken down by Disease factors.

Summary of Diagnostics	
No. of Clusters	2
No. of Points	102
Between-group Sum of Squares	20.285
Within-group Sum of Squares	9.5649
Total Sum of Squares	29.85

Table 20: Chest Pain Type: Asymptomatic.

No. of Clusters	Items	Ages (in Sum)	Sum Of Maximum Heart Rate	Disease
Cluster1	75	49.853	124.03	Positive
Cluster2	27	48.556	136.59	Negative

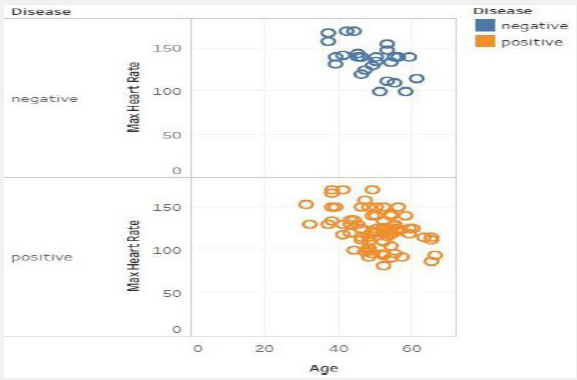

Figure 35: Age vs. Max Heart Rate broken down by Disease with asympt chest pain type.

Table 21: Chest Pain Type: Atypical Angina: Factors.

Summary of Diagnostics	
No. of Clusters	2
No. of Points	65
Between-group Sum of Squares	5.5109
Within-group Sum of Squares	8.3246

Table 22: Chest Pain Type: Atypical Angina.

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster1	59	45.492	147.47	Positive
Cluster2	6	47.5	139.5	Negative

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Table 24: Chest Pain Type: Non-Angina.

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster 1	15	39.533	162.8	Negative
Cluster 2	14	54.571	133.43	Negative
Cluster 3	7	52.857	140.29	Positive

Table 26: Chest Pain Type: Typical Anginal Pain.

No. of Clusters	Items	Ages (in Sum)	Sum of maximum heart rate	Disease
Cluster 1	2	40	177.5	Positive
Cluster 2	2	49	145.5	Positive
Cluster 3	2	50.5	143.5	Negative

Table 25: Chest Pain Type: Typical Anginal Pain: Factors.

No. of Clusters	3
No. of Points	6
Between-group Sum of Squares	2.3779
Within-group Sum of Squares	0.52542
Total Sum of Squares	2.9033

Table 23: Chest Pain Type: Non-Angina: Factors.

Chest Pain Type: Non-Angina	
Summary of Diagnostics	
No. of Clusters	3
No. of Points	36
Between-group Sum of Squares	8.89
Within-group Sum of Squares	2.251
Total Sum of Squares	11.141

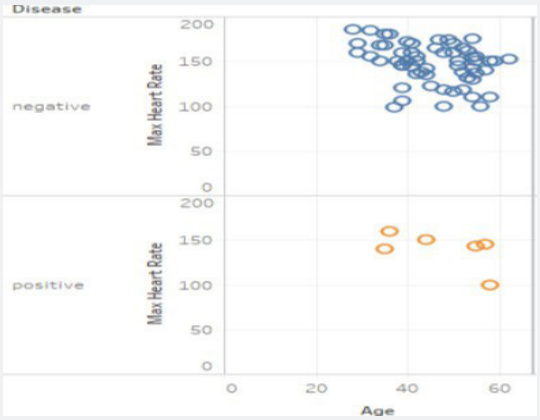

Figure 36: Age vs. Max Heart Rate broken down by Disease with atypical angina chest pain type.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

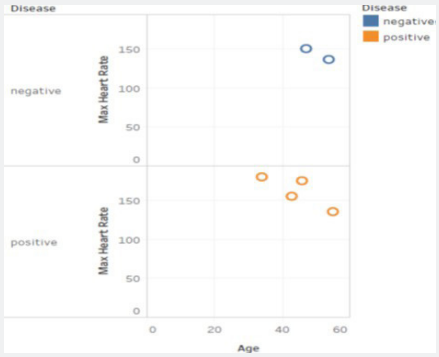

Figure 37: Age vs. Max Heart Rate broken down by Disease with non-angina chest pain type

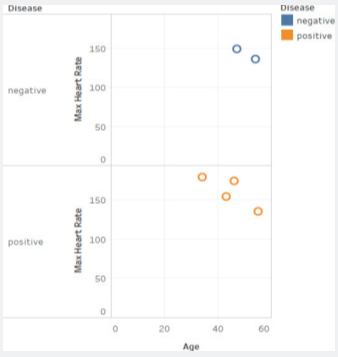

Figure 38: Age vs. Max Heart Rate broken down by Disease with typical angina chest pain type.

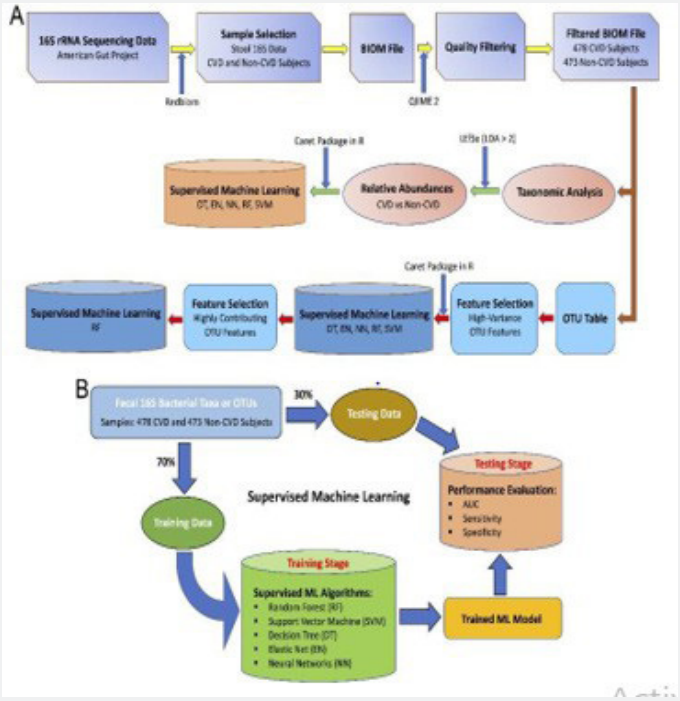

Figure 39: The study workflow.
a) Overall analysis.
b) Supervised machine learning.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Microbiome-Based Cardiovascular Disease Diagnostic Screening Using Machine Learning:

Cardiovascular disease (CVD), encompassing conditions like heart failure, hypertension, and atherosclerosis, poses a significant global health threat with an estimated death toll exceeding 23.6 million by 2030. To comprehensively assess cardiovascular health, various clinical procedures, including electrocardiogram (ECG), chest x-ray (CXR), and echocardiography, are often necessary. A rapid screening test could streamline this diagnostic process and facilitate timely and effective treatment. Utilizing QIIME 2, the BIOM file underwent further processing, excluding samples with a total frequency of less than 10,000. The resulting filtered BIOM file was employed to generate the Operational Taxonomic Units (OTUs) table for subsequent analysis, involving stool 16S data from 478 participants with CVD and 473 without CVD.

Statistical Analysis: ML Algorithms

The LDA score greater than 2.0 and the Kruskal-Walli's test determined the cutoff for discriminating features, utilizing 50 separate iterations for mean and standard deviation calculations of AUC, sensitivity, and specificity in ML modeling. Supervised ML models, enhanced with 39 taxonomic features, were employed to predict CVD and non-CVD classification. Testing five ML algorithms for CVD versus non-CVD classification revealed RF and NN outperforming others with an AUC of 0.58, while EN, SVM, and DT achieved 0.57, 0.55, and 0.51, respectively. RF and NN showed lower sensitivity but greater specificity compared to EN, DT, and SVM. (Table 27).

Table 27: Performance measures of supervised ML models

Features Algorithms	AUC	Sensitivity	Specificity
DT	0.51 ± 0.07	0.68 ± 0.18	0.41 ± 0.18
EN	0.57 ± 0.04	0.71 ± 0.17	0.37 ± 0.16
NN Bacterial Taxa	0.58 ± 0.04	0.59 ± 0.07	0.52 ± 0.06
RF	0.58 ± 0.04	0.59 ± 0.06	0.51 ± 0.04
SVM	0.55 ± 0.03	0.60 ± 0.08	0.49 ± 0.07
DT	0.52 ± 0.08	0.57 ± 0.10	0.53 ± 0.11
EN	0.56 ± 0.05	0.56 ± 0.09	0.55 ± 0.09
High-Variance NN			
OTUs	0.48 ± 0.04	0.59 ± 0.30	0.46 ± 0.28
RF	0.65 ± 0.03	0.70 ± 0.05	0.50 ± 0.04
SVM	0.57 ± 0.04	0.60 ± 0.07	0.52 ± 0.09

HCOF-Trained Supervised ML Models

Table 28: Performance measures of the RF model for classifying the CVD and non- CVD subjects using the highly contributing OTU features.

Top Features	AUC	Sensitivity	Specificity
Top 20	0.70 ± 0.03	0.69 ± 0.04	0.58 ± 0.05
Top 25	0.70 ± 0.03	0.70 ± 0.05	0.60 ± 0.05
Top 50	0.69 ± 0.03	0.69 ± 0.05	0.56 ± 0.06
Top 75	0.68 ± 0.03	0.71 ± 0.04	0.55 ± 0.06
Top 100	0.68 ± 0.03	0.70 ± 0.05	0.55 ± 0.06

For improved diagnostic classification and reduced feature space dimensionality, the top 500 high-variance OTU features were utilized to select High-Contributing OTU Features (HCOFs) in a Random Forest (RF) model. Training the RF model with the top 100 HCOFs, based on variable significance scores from 0 to 100, resulted in improved testing AUC (to 0.70). Re-implementing the RF method with the top 20, 25, 50, 75, and 100 HCOFs revealed outperformance by models trained with the top 20 and 25 HCOFs, demonstrating robust diagnostic classification for identifying individuals with CVD. (Table 28).

Discussion

This study explores the relationship between gut microbiota and cardiovascular health, aiming to identify cardiovascular disease (CVD) using gut microbiota information. It assesses supervised machine learning models' performance with top 500 high-variance Operational Taxonomic Unit (OTU) features, measuring AUC, sensitivity, and specificity. Seeking common early warning indicators for CVD, the analysis investigates if gut microbial composition changes can serve as markers. Successfully

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

using gut microbiome data for supervised machine learning, the study demonstrates promising predictive capabilities. The approach, novel for CVD, benefits from a large sample size (478 with CVD, 473 without), offering valuable insights without specific variable constraints.

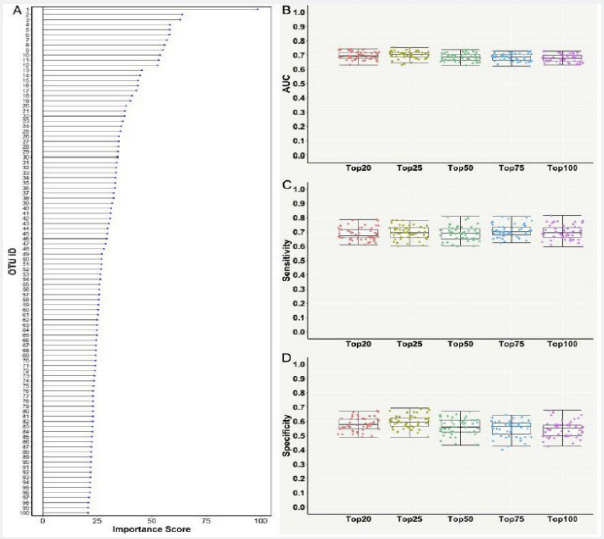

Figure 41: random forest (RF) model for categorizing people.

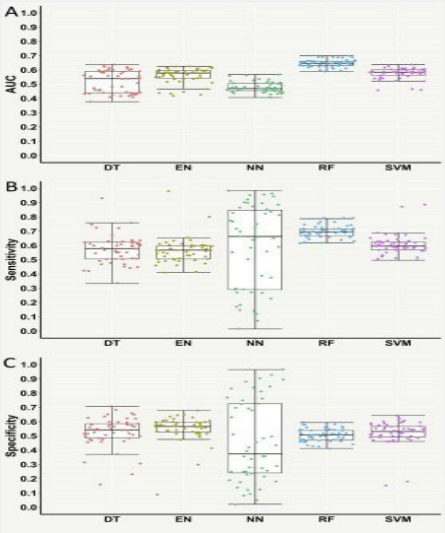

Figure 40: Performance metrics of supervised machine learning models.

The cohort represented various characteristics, allowing for a tolerant experimental design investigating ML model capability using gut microbiota for diagnostic classification of non-CVD vs. CVD. Acknowledging gut microbiota's susceptibility to additional factors, such as diet and medicine, the American Gut Project data lacked comprehensive assessment in the current ML analysis. Despite potential misreported or undiagnosed CVD cases, distinct gut microbial fingerprints were found between groups, indicating changed gut microbiota as a common factor in CVD manifestations. Initial ML modeling with differential bacterial taxa achieved a 0.58 AUC, prompting exploration of OTUs for increased predictive potential. Selecting top 500 high-variance OTUs, the RF model achieved an improved AUC of 0.65, with further enhancement to 0.70 using only 25 OTU characteristics. This study, unique in stool sample-based ML, outperforms previous approaches, showing promising potential for microbiome-based ML in predicting CVD, encouraging future calibration and improvement [10,11].

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Perspectives

The AI-driven ML modeling offers promising potential for practical diagnostic screening of CVD using gut microbiota composition. This supervised ML approach could serve as an initial tool for routine cardiovascular health monitoring before resorting to various clinical tests. The presented ML-based feature selection method highlights significantly contributing OTUs, showcasing that a small number of informative OTUs not only reduce computational complexity but also enhance diagnostic classification abilities. With hypertension being a major CVD risk factor, this technique could be applied for regular monitoring of cardiovascular damage caused by hypertension. The study identifies gut microbiome traits associated with cardiovascular health and disease, emphasizing the strong segregation of gut microbiota signatures between non-CVD and CVD subjects. This study marks the successful application of AI-driven gut microbiome-based ML modeling for potential CVD diagnostic screening.

CVD diagnostic screening.

This study paves the way for further exploration of machine learning's potential for precise cardiovascular disease diagnosis. Acknowledging the complexity of illness states and considering confounding factors in cardiovascular disorder diagnosis through innovative machine learning approaches opens avenues for future research. Further studies can delve into taxonomic traits to discern if the high abundance of certain microorganisms is a cause or effect of cardiovascular disorders. Exploring multiple metabolic pathways can unveil connections between microbe presence, metabolites, and cardiovascular disease. Additionally, the non-invasive nature of this research prompts investigation into its application for diagnosing cardiovascular disease from patient facial samples in a clinical setting.

Exploratory Analysis of ML Application in Diagnoses of Heart Problems

The initial exploration of this dataset involves checking the proportion of patients in each class. Among the 296 individuals, 160 belong to class "1," signifying good health, while 136 are in a different class, indicating a risk of developing heart disease. The dataset appears balanced in accurately representing both classes. (Figure 42).

Figure 42: Distribution of Target classes; 1: Healthy, 0: Not healthy.

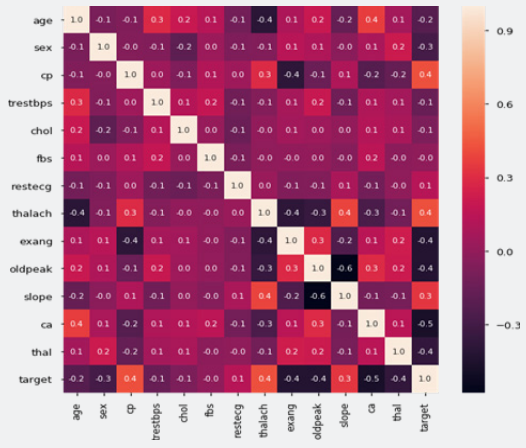

Figure 43: Correlation matrix between the variables.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

The correlation between variables is examined, and the presence of high correlation implies redundancy among them (Figure 43). However, in the correlation matrix shown, there is no evident association between any variables. Consequently, none of the variables can be dismissed as they may all contribute to the classification of the condition. Subsequent figures will explore the connections between both target classes and the various variables. (Figure 44).

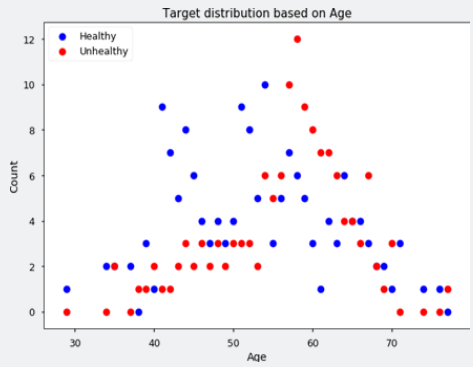

Figure 44: The relation between the variable Age and the classes of the target.

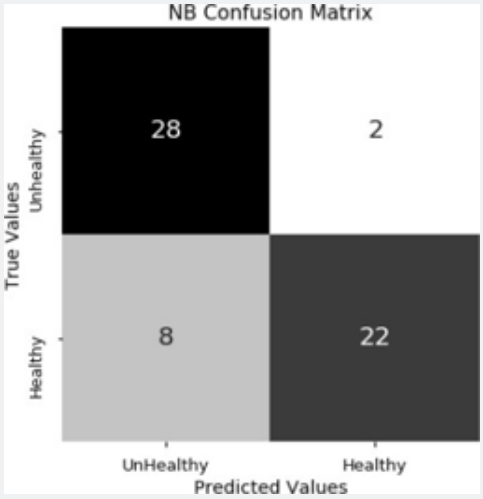

Figure 45: Mixed Naive Bayes confusion matrix.

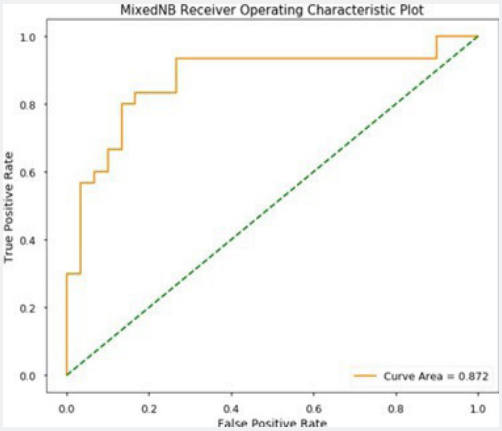

Figure 46: AUC of ROC curve for the mixed naive Bayes (MixedNB) mode Per class.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

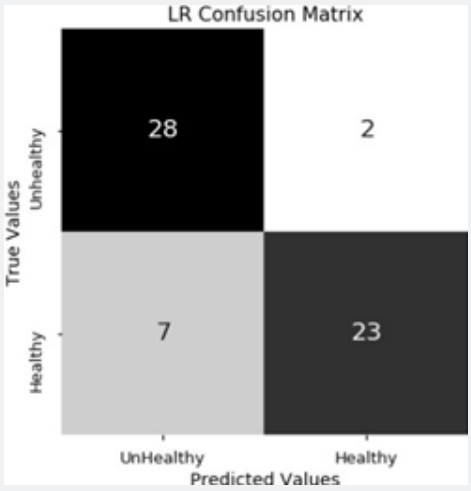

Figure 47: Logistic Regression Confusion Matrix.

Decision Trees

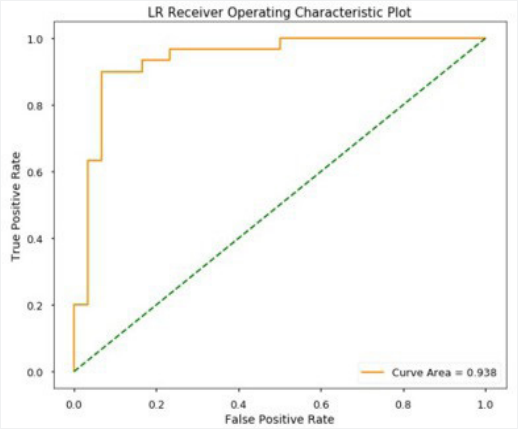

Figure 48: Logistic Regression AUC.

Weight	Feature
0.0767 ± 0.0806	ca
0.0333 ± 0.0365	thalach
0.0267 ± 0.0267	cp_2
0.0267 ± 0.0618	oldpeak
0.0200 ± 0.0249	cp_3
0.0100 ± 0.0163	trestbps
0.0067 ± 0.0163	fbs
0.0067 ± 0.0340	slope_2
0.0033 ± 0.0133	slope_1
0 ± 0.0000	restecg_2
-0.0000 ± 0.0211	exang
-0.0033 ± 0.0533	thal_3
-0.0067 ± 0.0163	cp_1
-0.0067 ± 0.0267	chol
-0.0067 ± 0.0267	age
-0.0100 ± 0.0400	sex
-0.0100 ± 0.0340	restecg_1
-0.0167 ± 0.0211	thal_2

Figure 49: Mixed Naive Bayes confusion matrix.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Although achieving an accuracy value of 0.84 on the training dataset by utilizing a 10-fold cross validation and fine-tuning the various parameters, the accuracy of the decision tree suffered from overfitting and fell to 0.75 on the test dataset. Figures 5-5 and 5-6 display the AUC curve and confusion matrix, respectively (Figure 57

The decision tree specifically struggled with sensitivity for the group of healthy individuals, where it received a recall of 0.6. The mod- el's f-1 score was 0.78 for class 0 and 0.71 for class 1, respectively. The same applications were run on SVM and ANN, and table 5-1 displays a summary of all outcomes (Table 29).

Table 29: Summary of the measure for all the used models.

		Precision		Recall		F1-Score		
Model	Accuracy	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	AUC
NB	0.83	0.78	0.92	0.93	0.73	0.85	0.81	0.87
LR	0.85	0.8	0.92	0.93	0.77	0.86	0.84	0.94
DT	0.75	0.69	0.86	0.9	0.6	0.78	0.71	0.81
SVM	0.83	0.78	0.92	0.93	0.73	0.85	0.81	0.93
ANN	0.85	0.8	0.92	0.93	0.77	0.86	0.84	0.85

Table 29 summarizes results from various strategies, with logistic regression standing out as the top-performing model across metrics, achieving 85% accuracy and a 94% AUC. Feature importance is assessed using permutation significance, revealing the LR model's crucial feature, "ca," impacting accuracy. Positive values indicate significant features, while negative numbers suggest improved accuracy after rearranging certain features.

Descriptive Statistics and Exploratory Data Analysis:

Table 30: Descriptive statistics of heart disease dataset.

	Age	Trestbps	Chol	Thalach
Mean	54.43	131.68	246.69	149.60
Std	9.03	17.59	51.77	22.87
Min	29.00	94.00	126.00	71.00
Median	56.00	130.00	241.00	153.00
Mode	58.00	120.00	197.00	162.00
Max	77.00	200.00	564.00	202.00

This section encompasses both descriptive statistics and exploratory data analysis. Out of the 14 data properties, only five are numeric, and the table includes information on four of them. The "Old peak" attribute is omitted from the table due to its conclusions necessitating a comprehensive understanding of heart physiology. (Table 30).

Table 31: Correlation matrix between numeric attributes

	Age	Trestbps	Chol	Thalach
Age	1.00	0.28	0.20	-0.39
Trestbps	0.28	1.00	0.13	-0.04
Chol	0.20	0.13	1.00	-0.003
Thalach	-0.39	-0.04	-0.003	1.00

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

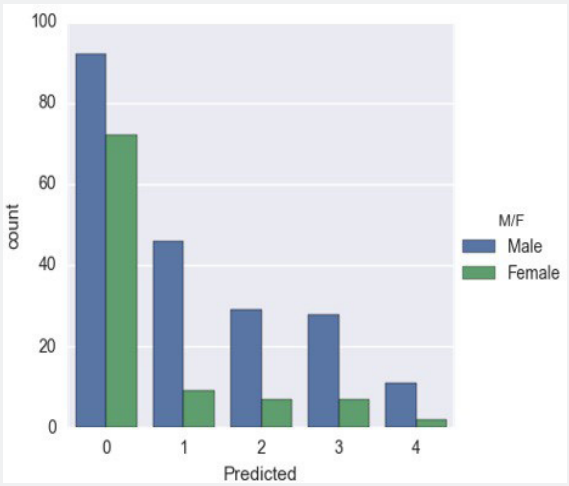

Figure 50: Disease rate in the male and the female.

Figure 11 shows the heart health statuses (0, 1, 2, 3, and 4), which range from extremely healthy to very ill. Male and female populations are shown by blue and green bars, respectively. It is clear from this data set that men are more likely than women to have cardiac disease. The Chi-square test is used to confirm the null hypothesis, H0, which is proposed. The null hypothesis asserts that there is no association between disease and population gender; if this hypothesis is disproven by the chi-square test, the hypothesis that there is some correlation between disease and population gender is established (Table 32).

Table 32: Number of diseased patients based on sex.

	0	1	2	3	4	Total
Male	92	46	29	28	11	206
Female	72	9	7	7	2	97
Total	164	55	36	35	13	303

0,1,2,3,4 are the columns spanning from no heart disease to highly unhealthy heart disease, and their total values are 164, 55, 36, 35, and 13 accordingly. In Table 32, there are 206 men and 97 women in total (Table 33).

Table 33: Expected number of diseased patients based on sex.

	0	1	2	3	4
Male	111.4	37.39	24.47	23.79	8.83
Female	52.50	17.60	11.52	11.20	4.16

The anticipated number of ill patients is determined in Table 33. The two rows of men and women with the five columns (0, 1, 2, 3, 4). (R-1) * (C-1) are the degrees of freedom for this table, where R and C are the corresponding rows and columns. The degree of freedom is therefore 4.

Experiments

The dataset underwent testing with various machine learning algorithms. Initially, the study modeled the dataset without feature selection and obtained results. In the second phase, modeling was done exclusively using features from SBS. Methodologies such

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

as parameter tweaking and k-fold cross-validation were applied across all studies. K-fold cross-validation aids in determining suitable parameters for the model, while parameter tweaking helps prevent over- and under-fitting in the dataset.

Results

Table 35: Evaluation of algorithms in test set without parameter tuning.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
KNN	0.82	0.83	0.81	0.81	0.81	0.44
SVC	0.79	0.78	0.78	0.78	0.78	0.45
Random Forest	0.69	0.68	0.68	0.68	0.69	11
Naïve Bayes	0.81	0.81	0.8	0.8	0.8	1.1

Table 34: Evaluation of algorithms in training set using k-fold.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
KNN	0.84	0.89	0.73	0.83	0.90	0.42
SVC	0.82	0.81	0.79	0.82	0.89	0.41
Random Forest	0.67	0.64	0.63	0.66	0.85	11.30

Table 36 reports a Nil entry for Naive Bayes, as it requires no parameter tweaking. Therefore, the performance result from Table 8 was used for comparison. KNN performance decreases after parameter adjustment. Post-adjustment, SVC outperforms Naive Bayes, Random Forest, and other models. It's worth noting that log loss, being negative and potentially exceeding one, is unique among performance measures. (Figure 51).

Table 36: Evaluation of algorithms in test set using parameter tuning.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
KNN	0.78	0.83	0.78	0.77	0.77	0.54
SVC	0.8	0.82	0.8	0.8	0.79	0.43
Random Forest	0.73	0.74	0.74	0.73	0.73	0.49
Naïve Bayes	Nil	Nil	Nil	Nil	Nil	Nil

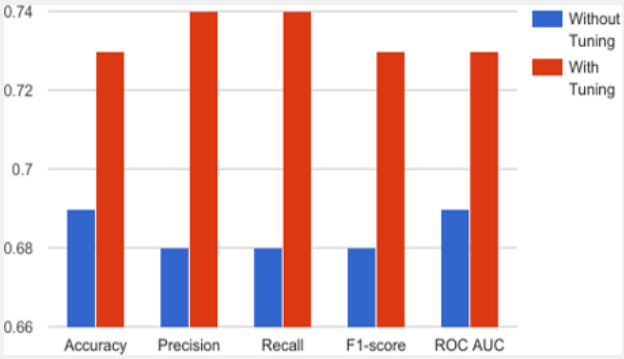

Figure 51: Random Forest with and without parameter tuning.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Table 38: Evaluation of tuned and calibrated algorithms with feature selection.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
KNN	0.78	0.78	0.78	0.78	0.77	0.42
SVC	0.85	0.86	0.86	0.86	0.85	0.37
Random Forest	0.78	0.78	0.78	0.78	0.77	0.45
Naïve Bayes	0.79	0.80	0.79	0.79	0.78	0.92

Table 37: Evaluation of algorithms in test set, parameter tuned using calibration.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
KNN	0.78	0.78	0.78	0.78	0.77	0.46
SVC	0.79	0.8	0.79	0.79	0.78	0.43
Random Forest	0.75	0.73	0.73	0.72	0.72	0.5
Naïve Bayes	0.8	0.81	0.8	0.8	0.79	0.48

Table 40: Evaluation of boosted, tuned and calibrated algorithms with feature selection.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
SVC	0.82	0.83	0.82	0.82	0.82	0.38
Random Forest	0.80	0.80	0.80	0.80	0.80	0.5
Naïve Bayes	0.83	0.84	0.84	0.83	0.83	0.43

Table 39: Evaluation of boosted algorithms in test set after without feature selection.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
SVC	0.78	0.79	0.78	0.78	0.77	0.43
Random Forest	0.72	0.73	0.73	0.72	0.72	0.5
Naïve Bayes	0.72	0.73	0.73	0.72	0.72	0.49

Table 40: Evaluation of boosted, tuned and calibrated algorithms with feature selection.

Algorithms used	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
SVC	0.82	0.83	0.82	0.82	0.82	0.38
Random Forest	0.80	0.80	0.80	0.80	0.80	0.5
Naïve Bayes	0.83	0.84	0.84	0.83	0.83	0.43

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

Table 41: Evaluation of Artificial Neural Network.

Algorithms	Accuracy	Precision	Recall	F1-score	ROC AUC	Log loss
ANN	0.91	0.92	0.89	0.90	0.87	0.23

Table 41 ANN employs the "adam" algorithm as an optimizer and three hidden layers. The Keras machine learning package was used to parameterize the ANN.

Observations and Findings:

Algorithm performance varies based on factors like cross-validation, grid search, calibration, and feature selection. Each method excels under different circumstances. Random Forest performs better with numerous datasets, while Support Vector Machine excels with a limited number of datasets. Interestingly, algorithm performance improves without boosting in non-feature-selected data but deteriorates with feature-selected data. However, after boosting in feature-selected data, performance improves, emphasizing the importance of feature selection before boosting. Performance measurements are commonly employed for dataset comparison after feature selection, parameter adjustment, and calibration.

Chapter-VI Findings, Recommendations and Conclusions

Myocardial infarction, commonly known as a heart attack, is a leading global cause of mortality and disability. Coronary artery disease is its primary cause. Early diagnosis is crucial for patient care, necessitating risk assessment based on various factors. Machine learning (ML) is explored for disease prediction, enhancing medical data analysis for improved patient treatment and community services. This study focuses on identifying key characteristics associated with heart disease by examining multiple datasets and employing various ML algorithms. The main contribution is a comparison of ML algorithms for early cardiovascular disease (CVD) prediction, achieving high accuracy. Future research could explore deep learning and fuzzy logic for more precise knowledge patterns. ML applications in healthcare, particularly for cardiovascular conditions, show promise in accurate diagnosis and prognosis, aiding in appropriate patient care. The study emphasizes the need for a reliable system to forecast heart diseases, offering potential for reduced fatality rates. The study's findings highlight the effectiveness of the Random Forest algorithm in predicting heart disease, emphasizing its potential for practical applications in healthcare.

Cardiovascular diseases (CVD) account for a significant portion of global fatalities, emphasizing the need for early diagnosis. This study focuses on ML algorithms for predicting heart disease, utilizing a diverse dataset and various algorithms. Results indicate the Random Forest algorithm's superior accuracy of over 90%. The study envisions the development of an online application using a larger dataset for improved accuracy and efficiency in disease forecasting. CVD prediction is essential to decrease mortality rates, and ML algorithms, particularly Random Forest, offer promising outcomes for effective diagnosis and prognosis.

Heart-related illnesses contribute significantly to global mortality, necessitating effective diagnostic tools. This study explores machine learning algorithms to predict heart disease, comparing their accuracy and investigating algorithmic variations. Gaussian Naive Bayes and Random Forest demonstrate the highest accuracy at 91.21%, while Decision Tree lags at 84.62%. The study emphasizes the importance of selecting the appropriate algorithm based on specific instances and datasets. It envisions potential applications, such as devices monitoring heart activities and aiding in disease diagnosis, especially in areas lacking heart disease experts.

Machine learning (ML) presents a viable approach for diagnosing heart disease, offering various benefits. This study comprehensively analyzes deep learning, ensemble learning, and ML-based cardiac prediction algorithms. It underscores the frequent use of a small heart disease dataset with limited features, indicating challenges in generalizing high accuracy results. The study highlights the need for more diverse datasets to achieve generalized classification and prediction accuracy. The research aims to develop an effective predictive framework that addresses current flaws, emphasizing real-time data analysis and clinical validation.

In conclusion, the exploration of machine learning algorithms for heart disease prediction reveals significant potential for early diagnosis and improved patient outcomes. The identified algorithms, such as Random Forest, showcase promising accuracy levels, paving the way for practical applications in healthcare. The studies underscore the importance of diverse datasets, algorithm selection, and ongoing research to enhance diagnostic frameworks and reduce the global burden of cardiovascular diseases.

Gather the Patient's History with Precision:

ML aids doctors in accurately gathering a patient's medical history by suggesting relevant questions based on various parameters, streamlining the process for both the patient and healthcare professionals. Additionally, ML applications offer valuable insights and predictions, improving patient care by providing intelligent reminders, scheduling assistance, accident

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

prevention through obstacle detection, optimizing routes, and facilitating prompt assistance for individuals with limited mobility.

Enhance Patient Interaction with Healthcare Services:

The main objective of ML-assisted platforms is to enhance the experience of healthcare services for the largest possible population.

- In established businesses, maximizing profit is the ultimate goal of the systems already in place. Strong ML technologies for hospital

operations management must set themselves apart from conventional systems by fusing compassion with a goal of making money.

- The pharmaceutical industry is increasingly relying on ML methods to address the exasperatingly difficult problem of effective drug discovery.

- Numerous start-up businesses also use ML systems to analyze multi-channel data using the most recent NLP and reinforcement learning methodologies.

Robot-Assisted Procedures and Other Image-Guided Treatments

ML tools significantly enhance surgical displays by providing surgeons with crucial information, such as the location of malignancies during robotic surgeries and image-guided therapies. These tools are invaluable to radiologists as healthcare organizations digitize exabytes of medical data. To support robotic procedures, robust AI-enabled solutions are necessary to connect to patient databases and analyze diverse data types. The developed systems should reveal hidden patterns, allowing healthcare professionals to work confidently and transparently, transforming findings into human-readable formats. ML offers automated messaging warnings and personalized content, promoting actions at critical moments. Additionally, ML applications, such as a bot system, reduce treatment time, facilitate neural network formation, and detect harmful cells, demonstrating precision in oncology comparable to skilled doctors. A voice-controlled virtual nurse acts as a healthcare assistant, providing information on illnesses, health conditions, and medications.

Take prompt Action

In healthcare, making timely decisions is crucial, and ML in bioinformatics plays a key role by processing vast amounts of data to provide valuable insights. This aids healthcare professionals in planning treatments, lowering risks, and making quick judgments. ML enhances patient assessments by improving access to medical histories and estimating outcomes based on treatment and lifestyle. The potential of ML extends to anticipating future patients' disease risks using data from routine exams and early screening tests, addressing information scarcity in healthcare.

By automating tasks and improving human decision-making, ML can revolutionize medical practice, particularly in addressing the challenge of reducing hospital readmissions, which can be costly due to Medicare payments being based on readmission rates. Identify a Health Issue

Examining Patient Data

Machine Learning (ML) utilizes patient data analysis to detect hard-to-find disorders, particularly effective in processing large volumes of radiology and pathology data for quicker decision-making in medical imaging. ML aids in tumor identification, offering precise visual images to guide medical professionals. Additionally, ML models assist researchers in analyzing data, selecting optimal paths for improving test results, and contribute to personal health monitoring through wearable smart devices.

Boost the Precision of the Outcomes:

Electronic medical gadgets are increasingly prevalent, benefiting from technological advancements and the enhanced accuracy provided by ML and AI applications. These gadgets, including diagnostic tools like computerized tomography scanners, ventilators, pacemakers, heart-lung machines, diabetes monitoring equipment, and infant incubators, play a critical role in providing accurate information about the patient's condition, necessitating precise operation and measurement by medical personnel.

Machine Learning (ML) accelerates the pace of human evolution, impacting healthcare by treating cancer and preventing pandemics through user-friendly smartphone apps. ML's substantial advancements benefit various industries, including healthcare, where it enhances patient care from drug discovery to diagnostics. ML resolves essential healthcare challenges by processing massive medical data, offering insights for quick decision-making. The technology's impact includes automating diagnostic recommendations, reducing healthcare expenses, and improving hospital administration. Challenges in ML adoption in healthcare include obtaining quality patient data, addressing privacy regulations, and preparing data for analysis. Despite obstacles, ML contributes to accurate cardiac disease prediction, supporting early detection and personalized treatment plans. The study utilizes ML classification techniques, achieving high accuracy for cardiovascular disease prediction, and suggests future exploration of deep learning and additional datasets. Heart disease remains a global health concern, and ML applications play a vital role in early detection and prediction, improving patient outcomes and reducing the economic burden.

References

1. P Resources and S Writers (2022) Python Resources for Programmers. Pp: 1-16.

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004

2. M Gudadhe, K Wankhade, S Dongre (2010) Decision support system for heart disease based on support vector machine and artificial neural network. 2021 Int Conf Compute Common Technol ICCCT-2010 Pp: 741-745.

3. K Thenmozhi, P Deepika (2014) Heart Disease Prediction Using Classification with Different Decision Tree Techniques. Int J Eng Res Gen 2(6): 6-11.

4. PPR Patil, PSA Kinariwala (2017) Automated Diagnosis of Heart Disease using Data Mining Techniques. Int J Adv Res Ideas Innov 3(2): 560-567.

5. SK Mohan, C Thirumalai, G Srivastava (2019) Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 7.

6. SP Bingulac (1994) On the Compatibility of Adaptive Controllers. Proc Fourth Ann Allerton Conf Circuits and Systems Theory Pp: 8-16.

7. S Nikhar, AM Karandikar (2016) Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Advanced Engineering, Management and Science (IJAEMS) 2(6): 617-627.

8. IS Jacobs, CP Bean (1963) Fine particles, thin films and exchange anisotropy. In Magnetism III, GT Rado, H Suhl Eds. New York: Academic Pp: 271-350.

9. Aditi G, Gouthami K, Isha P, Kailas D (2018) Prediction of Heart Disease Using Machine Learning. Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

10. Abhay K, Ajay K, Karan S, Maninder P, Yogita H (2018) Heart Attack Prediction Using Deep Learning. International Research Journal of Engineering and Technology (IRJET) 5(4): 4420-4423.

11. A Lakshmana Rao, Y Swathi, PSS Sundareswar (2019) Machine Learning Techniques for Heart Disease Prediction. International Journal of Scientific & Technology Research 8(11): 374-377.

DOI: 10.19080/JOCCT.2023.19.556004

How to cite this article: Mohan Raja Pulicharla. A Study On a Machine Learning Based Classification Approach in Identifying Heart Disease Within E-Healthcare. J Cardiol & Cardiovasc Ther. 2023; 19(1): 556004 DOI: 10.19080/JOCCT.2023.19.556004