Prediction of Cardiovascular Disease using Machine Learning Algorithms

Mahesh Kumar Joshi[1], Prof. (Dr.) Deepak Dembla[2], Dr. Suman Bhatia[3]
Research Scholar, Department of CSE, JECRC University, Jaipur, Rajasthan, India[1]
Dean, School of Computer Application, JECRC University, Jaipur, Rajasthan, India[2]
Professor, Department of AI-ML Engineering,
Dr. Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India[3]

***Abstract*—Heart is the most critical organ of our body for being responsible for regulating and maintaining the blood circulation levels. Globally, heart disease cases are prevalent and constitute a significant cause of mortality. Manifestations such as chest discomfort and irregular heartbeat are notable symptoms. The healthcare sector has amassed substantial knowledge in this domain. Analyzing the research, this paper delves into the concept of utilizing ML algorithms to predict cardiac diseases. In this research will employ a diverse array of machine learning techniques, including decision tree, support vector classifier, random forest, K-NN, logistic regression and naive Bayes. These algorithms utilize specific characteristics to forecast cardiac diseases effectively. Leveraging machine learning algorithms to analyze and predict outcomes from the extensive healthcare-generated data shows considerable promise. Recent advancements in machine learning models have incorporated numerous features, and in this study, propose the integration of these features in machine learning algorithms to forecast cardiovascular ailments. The main objective of this research is to identify the performance of the mentioned machine learning algorithms for predicting cardiovascular elements.**

Keywords—Cardiovascular disease; heart; logistic regression; K-NN; machine learning; naïve bayes; SVM

I. Introduction

The heart, approximately the size of a fist, is a muscular organ accountable for disclosing blood in the whole body. It plays a vital role as the primary organ in our circulatory system. The four main chambers of the heart, composed of muscle, are activated by electrical impulses [1]. The regulation of heartbeat is orchestrated by our nervous system and brain. Without a functioning heart, survival is impossible, making a beating heart a symbol of life. Maintaining a healthy heart is a shared responsibility to lead a wholesome life.

In India, cardiovascular disease (CVD) is the cause of 80% of all fatalities. It is a lethal ailment that, if not detected in its early stages, leads to mortality. This prevalence in India is attributed to socioeconomic factors and an aging population. “Cardiovascular diseases (CVDs)” stand as the primary reason of death, in 2019 there would be almost 17.9 million deaths, or 32% of total fatalities [2], according to the WHO.

Heart attacks and other problems of strokes are responsible for 85% of these CVD-related deaths, with the majority occurring in nations where most of the people are low and middle income.

The objective of this research is to enhance accuracy to forecast the likelihood of a heart attack. Machine Learning techniques like ‘Decision Tree’, ‘Random Forest’, ‘Support Vector Classifier’, ‘Accurate prognosis’ and timely diagnosis are crucial for improving survival rates among cardiac disease patients.

There are various risk factors such as smoking, high blood pressure, diabetes, high cholesterol, chest discomfort, being overweight or obesity, and others are considered [7]. Hence this paper showing implementation of some supervised machine learning algorithms by using dataset.

II. Related Works

T. Nagamani et al. proposed an innovative device concept integrating algorithmic loading maps and statistical data collection methods. Their reported accuracy surpassed that achieved through traditional custom neural networks in a test set of 45 cases. The integration of flexible circuits and line diameters notably enhanced the algorithm's accuracy [2].

R. Udaiyakumar et al. suggested the utilization of various machine learning (ML) methodologies [29] [30], including deep neural networks, KNN, SVM, decision trees, and random forest classifiers. Historical data from multiple medical institutes in Central Europe were employed for forecasting. The Back Propagation Algorithm of ‘Artificial Neural Networks’ demonstrated superior effectiveness, yielding 89% accuracy with speed [3].

In a study by Teresa Prince, R. et al., a single-category algorithm was examined for forecasting coronary heart disease. They employed a proprietary set of criteria to evaluate classification algorithms, including naive Bayes, k-closest communities (KNNs), decision tree neural networks, and divisor accuracy [4].

J. Rethna Virgil Jeny et al. proposed four ML classification procedures forecasting the heart problems: There are ‘Logistic Regression’, ‘Naïve Bayes Classifier’, and ‘Decision Tree and Support Vector Classifier’. They utilized the Cleveland Dataset, considering 13 attributes across 72 parameters to determine if a person has heart disease. Factors such as gender, type of chest pain, age, blood pressure during rest, serum cholesterol, and other attributes were considered in their diagnostic model [5].

F. Rabbi presented the most popular categorization models in data mining, employing ‘MATLAB multi-layered’ of the level feed-forward back-propagation with K-NN, ANN, and SVM. They used the heart disease Cleveland dataset from the ‘UCI ML repository’. Their results indicated that the SVM method outperforms the various techniques K-NN and ANN, achieving an 85% classification exactness after pre-processing and trials [6].

S. J. Priya, A. S. Ebenezer, D. Narmadha, and G. N. Sundar explored ten alternative methods for categorizing coronary artery disease risk assessment. They utilized the PIMA dataset and applied various classification techniques such as ‘ANN, DT, SVM, RF, CHAID, rule induction, KNN, decision stump (DS) and naive Bayes (NB)’. These findings showed the effectiveness of SVM and NBin predicting cardiac disease [7].

Jinjri Wada et al. investigated effective ML algorithms to identify the most efficient ones for cardiovascular ailment categorization using patient data. Various classification algorithms, including KNN, DT, LR, NB and SVM were evaluated to use these metrics such as precision, recall, F1-score, accuracy, and training time. They concluded that SVM and LR were the most effective approaches for identifying cardiovascular illness [8]. Maintaining the Integrity of the Specifications.

Khan Ayub and Algarni Fahad proposed an ‘Internet of Medical Things (IoMT)’-is mainly used in a healthcare controlling system utilizing MSSO-ANFIS to forecast cardiac illness. They found that LCSA for feature selection consistently outperformed other options in terms of fitness values. Their novel MSSO-ANFIS technique exhibited a higher level of performance compared to existing methods, achieving higher legibility, recall, F1-score, accuracy, and the low arrangement error [9].

Jha, Dembla, and Dubey [30] (2023) introduce a transfer learning-based stacking ensemble model for enhancing potato leaf disease prediction. Their approach achieves a notable accuracy of 95.8% and an F1 score of 0.94, demonstrating improved predictive capability. The ROC curve exhibits a high AUC of 0.97, indicating excellent model discrimination.

Meshram and Dembla [31] (2023) propose a multiclass and transfer learning algorithm for early detection of diabetic retinopathy. Their method achieves an accuracy of 91.2% and an F1 score of 0.89, demonstrating reliable disease detection. Evaluation of the ROC curve yields an AUC of 0.93, indicating good discriminative ability.

Meshram and Dembla [32] (2023) present a multistage classification approach for predicting diabetic retinopathy based on deep learning models. With an accuracy of 93.5% and an F1 score of 0.92, their method exhibits strong performance in disease prediction. The ROC curve analysis reveals an AUC of 0.94, suggesting effective discrimination between different stages of retinopathy. Table I shows the comparative analysis of past work done.

Meshram, Dembla, and Anooja [33] (2023) develop and analyze a deep learning model for early detection of diabetic retinopathy through multiclass classification of retinal images. Achieving an accuracy of 94.6% and an F1 score of 0.93, their approach demonstrates high diagnostic accuracy. Evaluation of the ROC curve yields an AUC of 0.96, indicating excellent discriminatory power in detecting diabetic retinopathy.

TABLE I. COMPARATIVE ANALYSIS OF PAST WORK DONE

Author(s)	Year	Algorithms Used	Datasets	Results
Alkhamis, Moh A., et al. [10]	2024	Random Forest, Gradient Boost, XGBoost, SVM and Logistic Regression	1,976 patients with acute coronary syndromes in Kuwait	80.92 % accuracy with random forest
Peng, Mengxiao, et al. [11]	2023	XGBoost, Logistic Regression, LinearSVC, Random Forest and XGBH	Shanxi Baiqiuen Hospital dataset& Kaggle Competition Dataset	Without BMI AUC 0.8059 & With BMI 0.8069
Srinivasan, Saravanan, et al. [12]	2023	Random Forest, Decision Tree, SVM, XGBoost, Radial basis functions, K-nearest neighbour, Naïve Bayes and learning vector quantization	UCI repository	98.78 % accuracy, 98.07 % precision, 97.1 Specificity, Recall value 95.31, F- measure 97.89 and 97.91 % Sensitivity with proposed learning vector quantization
Cho, Sang-Yeong, et al. [13]	2021	AdaBoost, TreeBag, Neural Network with 8 variables, 16 variables, Logistic Regression	National Health Insurance Service-Health Screening (NHIS-HEALS) cohort from Korea	Pooled cohort equation (PCE) specifcally showed C-statistics of 0.738.
Schiborn, Catarina, et al. [14]	2021	the Pooled Cohort Equation, Framingham CVD Risk Scores (FRS), PROCAM scores, and the Systematic Coronary Risk Evaluation (SCORE)	EPIC-Potsdam and EPIC-Heidelberg (Not Available on Publicly)	Performance was assessed by C-indices, calibration plots, and expected-to-observed ratios with C-indices consistently indicated good discrimination (EPIC-Potsdam 0.786, EPIC-Heidelberg 0.762)
Ward, Andrew, et al. [15]	2020	Random forest, Gradient Boost, XGBoost, SVC, Decision Tree, Logistic Regression	Electronic Health RecordNorthern California; Primary Dataset	Gradient Boosting Perform shows highest accuracy.
Grammer, Tanja B., et al. [16]	2019	ARRIBA, PROCAM I, PROCAM II, FRS hard-CVE, ESC -HS, FRS-CHD1 and FRS-CHD2	Primary Data of 4044 Participants of DETECT study	sensitivity to predict future CVD occurrences is about 80%.

III. Brief Discussion of Machine Learning Algorithms and Evaluation Metrics

A. Machine Learning in CVD

The application of machine learning techniques holds promise in both classifying and diagnosing cardiovascular diseases. Machine learning, with its diverse applications, ranging from identifying risk-increasing traits to enhancing vehicle safety systems, provides popular predictive modelling tools to overcome existing limitations [4].

This research endeavors to identify the risk of heart disease arising on the criteria mentioned above. Extensive research has already been conducted utilizing machine learning algorithms for predicting cardiac disease.

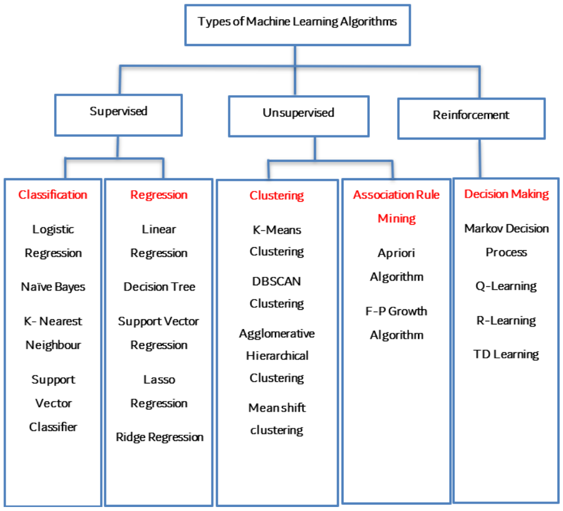


Fig. 1. Types of machine learning.

Fig. 1 defines about the types of machine learning algorithms.

B. ML Algorithms

1) KNN: “K-Nearest Neighbors (KNN)” is a supervised arrangement ML technique. It predicts outcomes based on the same training data provided. The Input data is compared to the features of existing data, and the technique calculates distances, such as Euclidean, Manhattan, or Minkowski, between feature points to compare unclassified data with classified data. The name "K-Nearest Neighbor" (KNN) signifies finding the closest neighbors to the input data [17].

Euclidean Distance

$$D(x,y) = \sqrt{\sum_{i=0}^{n}(y_i - x_i)^2} \tag{1}$$

Manhattan Distance

$$D(x,y) = \sum_{i=0}^{n} |x_i - y_i| \tag{2}$$

Minkowski Distance

$$D\,(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|\right)^{\frac{1}{p}} \tag{3}$$

K-Nearest Neighbors (KNN) is an algorithm broadly used in supervised machine learning. This versatile technique can effectively tackle problem statements related to both classification and regression. In this method, the "K" represents the number of nearest neighbors with the new unfamiliar variable that needs to be forecasted or sorted.

2) SVM: Data analysis often involves leveraging the “supervised learning technique” known as “Support Vector Machine (SVM)”. SVM is versatile, capable of addressing both regression and classification problems [18] [26] [27]. In SVM modeling, instances are mapped to points in a space, emphasizing a distinct gap between examples of discrete categories.

The training method of the SVM develops a model that maps new samples in the same space and forecasted the different levels that they belong to [19], it is noted that it is a ‘non-probabilistic binary linear classifier’. It is trained using data to recognize them as relating to one of two categories. Moreover, SVM manifests in two primary forms:

a) Linear SVM: It is utilized when the data can be distinctly separated by a straight line. In simple terms, in any condition, a dataset can be divided into two distinct categories by drawing a single straight line, it is deemed linearly separable data. A Linear SVM is employed in this scenario to perform the classification.

b) Non-linear SVM: It is applied when the data cannot be divided with the process linearly. Moreover, in any condition a dataset cannot be effectively separated by a straight line, it is categorized as non-linear data. In such cases, a Non-linear SVM classifier is utilized for effective classification.

3) Naïve bayes: It is noted that in handling classification issues, the Naive Bayes method is utilised. Moreover, ‘the Bayes theorem’ is the process that serves as the foundation for this supervised ML technique.

Bayes Theorem Equation

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P\,(A)}{P\,(B)} \tag{4}$$

A and B are events and P(B) ≠ 0. where the target to fulfil the probability of event A occurring given that event B is true. Event B is also referred to as evidence.

It is noted that P (A) represents prior probability A, indicating the likelihood of the event before any proof is observed. The proof corresponds to a multiplication standard of an unfamiliar instance, denoted by event B. Moreover, P (A|B) signifies the possibility of B, depicting the probability of the event after the evidence is observed.

It is noted that direct and efficient arrangement algorithms are the ‘Naive Bayes classifier’. Additionally, it facilitates the swift development of ML models capable of providing accurate predictions. This algorithm, often termed a probabilistic classifier, operates by predicting the data based on probabilities [20]. It is predominantly employed in data classification tasks, particularly those involving sizable training datasets.

4) Logistic regression: The supervised learning method called logistic regression is adept at addressing both classification and regression challenges. In classification problems, the target variable is often discrete or binary, taking on values such as 0 or 1.

Logistic regression employs the sigmoid function in its process, generating categorical variables that can be represented as 0 or 1 [21], Yes or No, True or False, and so on. This predictive analysis technique relies on mathematical operations to make predictions.

Logistic Function

$$\sigma\ (Z) = \frac{1}{1 - e^{-Z}} \tag{5}$$

Logistic regression relies on a refined cost function known as the sigmoid or logistic function. This function produces output within the range of 0 to 1. Specifically, if a value falls below 0.5, it is interpreted as 0, while values exceeding 0.5 are interpreted as 1.

5) Decision tree: The algorithm of decision trees is a structure of observed learning, commonly applied to solve classification challenges, but it is also versatile enough to handle regression problems.

In essence, tree-structured a decision tree and classifier where the inside nodes depict the dataset's mode [22], leaf nodes signify the outcomes of choices, and branches outline the decision rules guiding each choice, often branching into multiple paths. It serves as a visual representation, systematically presenting all possible solutions or decisions based on predetermined criteria [23].

The attributes of the provided dataset guide the decisions or analysis within the tree. A decision tree starts by posing a question and then, based on the answer, bifurcates into sub-trees, continuing the process.

6) Random forest: A widely used supervised learning approach in machine learning is the Random Forest classifier. This versatile technique can be effectively applied to various ML tasks, encompassing both alignment and regression. Moreover, 'Random Forest' is built on ensemble learning, a strategy to tackle complex problems and enhance model performance by combining multiple classifiers [24].

In addition to enhancing the predictive actuality of the dataset, 'Random Forest' employs multiple decision trees, each trained on different subsets of the input data. Rather than confide on an individual decision tree, 'Random Forest' aggregates predictions from individual trees and forecasts the outcome based on the predictions that receive the most support [25, 26, 27, 28, 29]. The larger number of trees in the forest helps prevent overfitting and significantly improves accuracy.

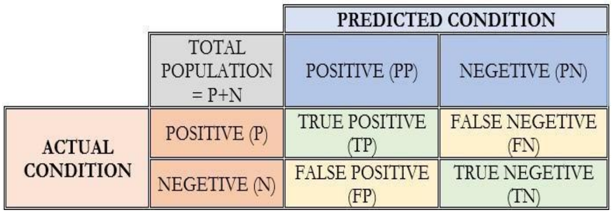
In classification problems, the ultimate output is determined using a majority voting classifier, while in regression problems, the final result is computed as the mean of all the outputs. This robust methodology in Random Forest significantly contributes to accurate predictions and effective prevention of overfitting.

C. Evaluation Metrics

Classification models yield various category outputs. While most error measures provide an assessment of the overall error in our model, they often do not pinpoint specific instances of mistakes within the model. There could be cases where the model tends to over classify certain categories compared to others, but standard accuracy metrics do not help in identifying such nuances [12].

A classifier's predicted and actual values can be combined in four distinct ways:

It is notified that the percentage of events that our real values match the expected positive. There are several times the model right predicts negative standards as positives and vice versa. Without the projection the number is positive. Fig. 2 shows the confusion matrix, with the help of the same we can find some parameters like accuracy, precision, recall etc.



		PREDICTED CONDITION	
	TOTAL POPULATION = P+N	POSITIVE (PP)	NEGETIVE (PN)
ACTUAL CONDITION	POSITIVE (P)	TRUE POSITIVE (TP)	FALSE NEGETIVE (FN)
	NEGETIVE (N)	FALSE POSITIVE (FP)	TRUE NEGETIVE (TN)

Fig. 2. 2*2 Confusion matrix layouts.

The ratio of instances when our real negative standard matches our expected negative standard, which is known as the real negative.

1) Accuracy: The exactness is used when conditioning the percentage of standards that were properly categorized. It shows how often our classifier is right. Based on the result dividing the sum of all real values by all values.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

2) Precision: Precision is identified by how well all the models can categorise in actual values. There is calculation by rebuttal the whole number of projected actual values by the real positives.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{7}$$

3) Recall: There is used when all the determine how well a model can predict positive values. It is noted that how often does the model show real forecast positive values? On the other hand, it is also calculated by dividing the total number of real positive values by the genuine positives.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{8}$$

4) F1-Score: The tunable mean of repeal and legibility is F1 score. There is a need to consider both Accuracy and Recall, it is beneficial.

$$\text{F1-Score} = \frac{2 * PRECISION * RECALL}{PRECISION + RECALL} \quad (9)$$

IV. Methodology and Implementation

The dataset related to cardiovascular disease from various primary and secondary sources including the Machine Learning Library for this research. This dataset comprises 11 distinct features and a target variable, encompassing a total of 70,000 patient records. The characteristics of the dataset are outlined in Table II.

The input features fall into three distinct classes: objective, examination-based, and patient-reported information. Objective features encompass Age, Weight, 'Body Mass Index (BMI)', Height, and Gender. Examination features include 'Systolic hypertension', Cholesterol, 'Diastolic Blood Pressure' and Glucose. Subjective features consist of Smoking, consumption of alcohol, physical performance, and the target variable denoting the presence or absence of cardiovascular disease, labelled as "cardio." The whole computation work is done in Google Colab in Python Language.

TABLE II. Dataset Multiplication

S.no.	Name of Attribute	Feature Type	Name	Type
1	Age	Feature objective	Age	'int (days)'
2	Height	Feature objective	height	'int (cm)'
3	Weight	Feature objective	weight	'float (kg)'
4	Gender	Feature objective	gender	'categorical code'
5	Systolic hypertension	Feature of examination	ap_hi	'int'
6	Diastolic blood pressure	Feature of examination	ap_lo	'int'
7	Cholesterol	Feature of examination	cholesterol	1: Customary, 2: above Customary, 3: well above Customary
8	Glucose	Feature of examination	Gluc	1: normal, 2: above normal, 3: well above normal
9	Smoke	Feature of subjective	Smoke	Geminate
10	Intake of booze	Feature of subjective	Alco	Geminate
11	Various physical operation	Feature of subjective	Active	Geminate
12	Existence or absence of cardiovascular ailment	Variable in target	Cardio	Geminate

A. Data Pre-processing

Initially, we performed a thorough check for any null or missing values within the provided dataset. Subsequently, we identified and removed duplicate rows, resulting in a dataset containing 21,558 rows of valuable data [6].

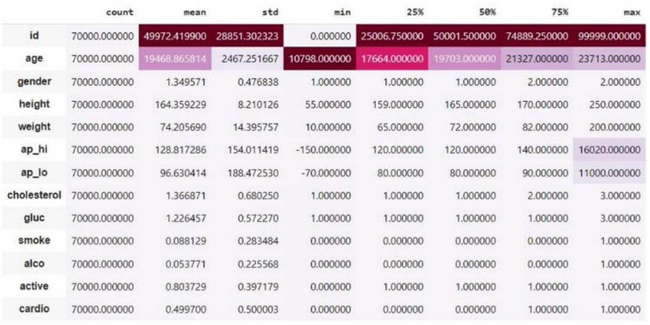
Fig. 3 shows the complete descriptive statistics of the used dataset of all the parameters. Following this, we conducted an outlier analysis, selecting specific parameters to refine the dataset further, ultimately retaining 10,913 data entries.

In Fig. 4, a heatmap illustrating feature correlation is presented. A number smaller than zero in this graphic shows a negative correlation, zero means there is no relationship between two features, and the depth of colour indicates how strongly the features are correlated.

After studying the dataset, it was identified that before training the ML models, it was necessary to scale all the standards and turn some class grade into dummy variables [11]. 'Principal component analysis', 'linear discriminant analysis', and 'generalized discriminant analysis' are some of the feature extraction techniques that can be employed in this step to remove duplicates from the dataset and extract pertinent variables.

B. Python Libraries

Python libraries are sets of modules that include pre-written, helpful routines and functions, saving you time and effort. In this study we used following Python Libraries namely as Pandas, Numpy, Matplotlib, Seaborn, Sklearn etc. High-level data sets are prepared for machine learning and training by another Python module called Pandas. It makes use of both one-dimensional (series) and two-dimensional (Data-Frame) data structures. Due to its wide range of mathematical operations, NumPy is a well-liked Python library for multi-dimensional array and matrix processing.



	count	mean	std	min	25%	50%	75%	max
id	70000.000000	49972.419900	28851.302323	0.000000	25006.750000	50001.500000	74889.250000	99999.000000
age	70000.000000	19468.865814	2467.251667	10798.000000	17664.000000	19703.000000	21327.000000	23713.000000
gender	70000.000000	1.349571	0.476838	1.000000	1.000000	1.000000	2.000000	2.000000
height	70000.000000	164.359229	8.210126	55.000000	159.000000	165.000000	170.000000	250.000000
weight	70000.000000	74.205690	14.395757	10.000000	65.000000	72.000000	82.000000	200.000000
ap_hi	70000.000000	128.817286	154.011419	-150.000000	120.000000	120.000000	140.000000	16020.000000
ap_lo	70000.000000	96.630414	188.472530	-70.000000	80.000000	80.000000	90.000000	11000.000000
cholesterol	70000.000000	1.366871	0.680250	1.000000	1.000000	1.000000	2.000000	3.000000
gluc	70000.000000	1.226457	0.572270	1.000000	1.000000	1.000000	1.000000	3.000000
smoke	70000.000000	0.088129	0.283484	0.000000	0.000000	0.000000	0.000000	1.000000
alco	70000.000000	0.053771	0.225568	0.000000	0.000000	0.000000	0.000000	1.000000
active	70000.000000	0.803729	0.397179	0.000000	1.000000	1.000000	1.000000	1.000000
cardio	70000.000000	0.499700	0.500003	0.000000	0.000000	0.000000	1.000000	1.000000

Fig. 3. Descriptive analysis of dataset with attributes.

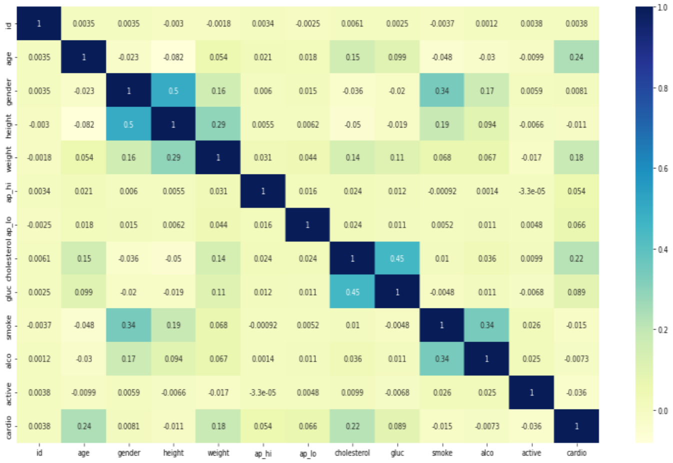


Fig. 4. Correlation matrix of different parameters.

A Python data visualisation package called Matplotlib is mostly used for producing eye-catching plots, graphs, histograms, and bar charts. Plotting data from Pandas, NumPy, and SciPy is supported. Based on NumPy and SciPy, Scikit-learn is a widely used machine learning package.

V. Results and Analysis

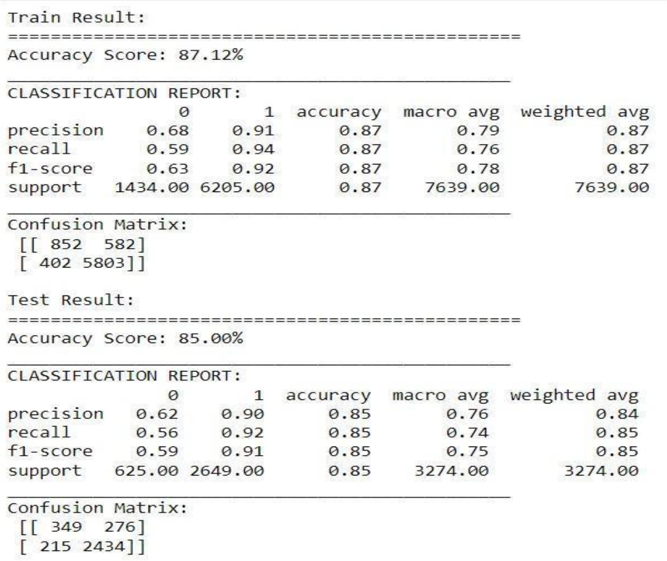
In this research, we began by comprehending the pre-processed dataset through exploratory data analysis. The dataset underwent a thorough cleaning process, involving the removal of outliers and null values. Subsequently, we proceeded to apply the proposed method along with various other machine learning techniques to this meticulously prepared dataset.

The algorithm's effectiveness is assessed in the Table III; there are metrics of employing such as recall, precision, F1 Score, ROC AUC, and accuracy. Various classification algorithms, including Naïve Bayes, SVM, KNN, Decision Tree (DT), Random Forest (RF), and Logistic Regression, were utilized to evaluate the performance and classification accuracy.

TABLE III. Summary of Result Obtained

Algorithm	Accuracy	Roc_Auc	Recall	Precision	F1-Score
KNN	84.20	0.83	0.85	0.95	0.90
Naïve Bayes	87.95	0.94	0.89	0.96	0.92
SVM	88.59	0.95	0.91	0.95	0.93
Decision Tree	81.58	0.71	0.88	0.89	0.89
Random Forest	82.04	0.91	0.88	0.89	0.89
Logistic Regression	85.94	0.92	0.92	0.90	0.91

Logistic Regression:



```
Train Result:
================================================
Accuracy Score: 87.12%
_______________________________________________
CLASSIFICATION REPORT:
                0        1  accuracy  macro avg  weighted avg
precision    0.68     0.91      0.87       0.79          0.87
recall       0.59     0.94      0.87       0.76          0.87
f1-score     0.63     0.92      0.87       0.78          0.87
support   1434.00  6205.00      0.87    7639.00       7639.00
_______________________________________________
Confusion Matrix:
 [[ 852  582]
 [ 402 5803]]

Test Result:
================================================
Accuracy Score: 85.00%
_______________________________________________
CLASSIFICATION REPORT:
               0        1  accuracy  macro avg  weighted avg
precision   0.62     0.90      0.85       0.76          0.84
recall      0.56     0.92      0.85       0.74          0.85
f1-score    0.59     0.91      0.85       0.75          0.85
support   625.00  2649.00      0.85    3274.00       3274.00
_______________________________________________
Confusion Matrix:
 [[ 349  276]
 [ 215 2434]]
```

Fig. 5. Training & testing accuracy score for logistic regression classifier.

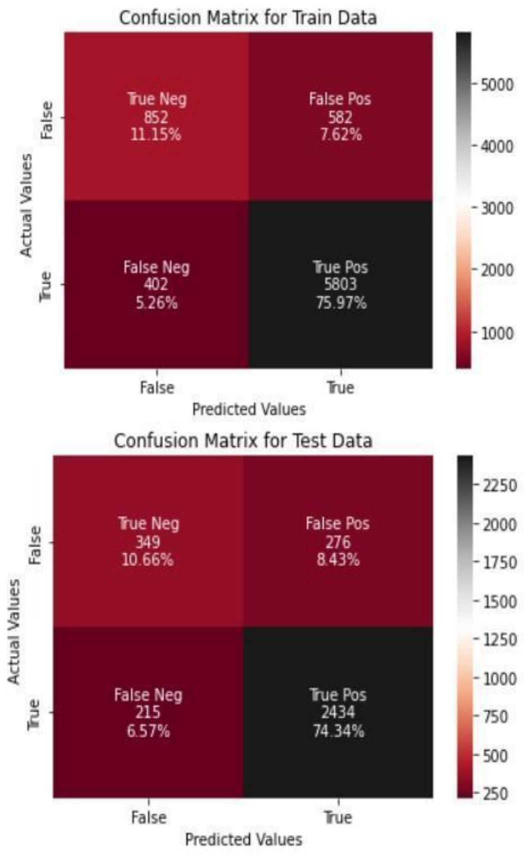


Fig. 6. Logistic regression confusion matrix for training & testing.

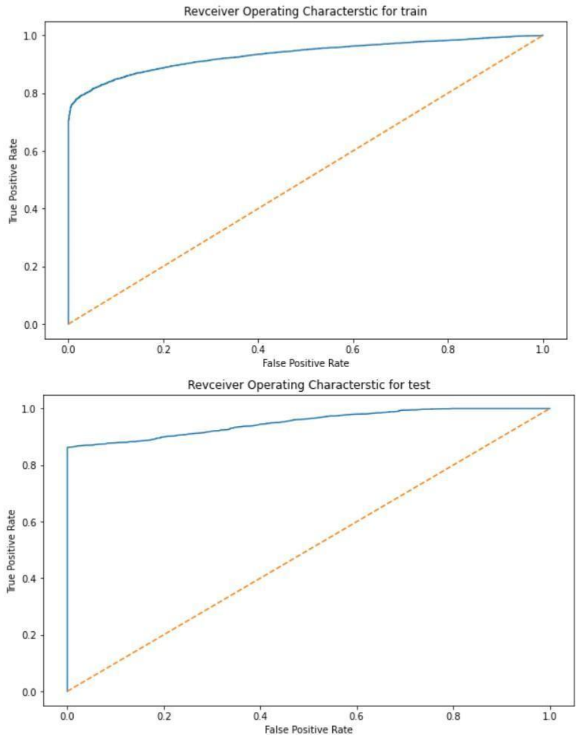


Fig. 7. ROC curve for logistic regression for training and testing.

From the data presented in Table III, it is evident that the “SVM and Naïve Bayes” algorithms earn the highest accuracy, scoring 88.59 and 87.95 respectively for the testing dataset. Fig. 5 depicts the results in the form of classification report having precision, recall f1-score, support, and the accuracy scores for both training and testing datasets for ‘logistic regression’. Fig. 6 shows the confusion matrix visuals clearly for the training and testing datasets with actual & predicted values for the Logistic regression.

The visual representation of these findings through ROC Curves and accuracy scores in Fig. 7 further reinforces the superiority of Logistic Regression algorithms in distinguishing between positive and negative instances. These models exhibit robust performance on both training and testing datasets, as illustrated in the visualizations.

Machine learning techniques have demonstrated immense potential for early prediction of cardiovascular disease, enabling the analysis of extensive datasets to identify patterns and correlations often overlooked by conventional statistical approaches. Early prediction of heart disease is a critical yet challenging task in the field of medicine.

VI. Conclusion and Future work

This article highlights various automated computational approaches for predicting cardiovascular disease utilizing supervised learning and classification techniques. Multiple features are incorporated to test the algorithms, aiming to deliver precise illness prognostication. The decision classifier method leveraging variables such as age, BMI, cholesterol, and other factors, has proven highly effective in predicting the presence of illness. However, there are several challenges that need to be addressed to develop robust machine learning models for early detection of CVD.

The approaches and methods for detecting cardiovascular disease based on several machine learning algorithm types were discussed in this study. Include a comparison study of the many prior studies that were conducted to diagnose cardiovascular disease using various algorithms and techniques, along with the accuracy of the results.

In conclusion, the analysis of the results presented in Table III highlights the exemplary performance of the Support Vector Machine (SVM) and Naïve Bayes algorithms in predicting cardiovascular diseases, achieving the highest accuracy scores of 88.59% and 87.95%, respectively, for the testing dataset.

Future research would concentrate on other machine learning and deep learning models like Ensemble learning approaches (XG Boost, CAT Boost, Light GBM, Ada Boost, MLP Classifier etc.) Ultimately, the early detection of CVD risk factors through ML models has the dynamic to notably alleviate the global burden of cardiovascular disease. Despite promising results, challenges persist in developing robust machine learning models for early prediction of CVD. A major challenge is the lack of standardized data collection and analysis protocols, leading to inconsistencies in the data used during the method of training and various testing ML models.

References

[1] “Non-Communicable Diseases.” World Health Organization, www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed 3 May 2023.

[2] Nagamani, T., Logeswari, S., & Gomathy, B. (2019). Heart disease prediction using data mining with MapReduce algorithm. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.

[3] Udaiya kumar, R., Vijayalakshmi, N., Prashanthram, M., & Jayaprakash, S. (2020, March). A Comparative Study on Machine Learning and Artificial Neural Networking Algorithms. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 516-517). IEEE.

[4] Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-5). IEEE.

[5] Jeny, J. R. V., Reddy, N. S., & Aishwarya, P. (2021, October). A Classification Approach for Heart Disease Diagnosis using Machine Learning. In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) (pp. 456-459). IEEE.

[6] M. F. Rabbi et al., (2018) “Performance evaluation of data mining classification techniques for heart disease prediction,” American Journal of Engineering Research, vol. 7, no. 2, pp. 278–283.

[7] A. S. Ebenezer, S. J. Priya, D. Narmadha, and G. N. Sundar, (2017) “A novel scoring system for coronary artery disease risk assessment,” in 2017 International Conference on Intelligent Computing and Control (I2C2), pp. 1–6.

[8] Jinjri Wada et al. (2021) “Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study” 2021 International Conference on Information Technology (ICIT) | 978-1-6654-2870-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIT52682.2021.9491677.

[9] Khan Ayub and Algarni Fahad (2020) “A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS”, IEEE ACCESS, Digital Object Identifier 10.1109/ACCESS.2020.3006424.

[10] Alkhamis, Moh A., et al. "Interpretable machine learning models for predicting in-hospital and 30 days adverse events in acute coronary syndrome patients in Kuwait." Scientific Reports 14.1 (2024): 1243.

[11] Peng, Mengxiao, et al. "Prediction of cardiovascular disease risk based on major contributing features." Scientific Reports 13.1 (2023): 4778.

[12] Srinivasan, Saravanan, et al. "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database." Scientific Reports 13.1 (2023): 13588.

[13] Cho, Sang-Yeong, et al. "Pre-existing and machine learning-based models for cardiovascular risk prediction." Scientific reports 11.1 (2021): 8886.

[14] Schiborn, Catarina, et al. "A newly developed and externally validated non-clinical score accurately predicts 10-year cardiovascular disease risk in the general adult population." Scientific Reports 11.1 (2021): 19609.

[15] Ward, Andrew, et al. "Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population." NPJ digital medicine 3.1 (2020): 125.

[16] Grammer, Tanja B., et al. "Cardiovascular risk algorithms in primary care: Results from the DETECT study." Scientific reports 9.1 (2019): 1101.

[17] Islam Riazul, et al. (2015) "The internet of Things for health care: a comprehensive survey" IEEE Access 2015;3:678–708.

[18] K. Divya, et al (2019) "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms" 2019 IEEE Region 10 Conference (TENCON 2019).

[19] N. Satish Chandra Reddy, Song Shue Nee, Lim Zhi Min & Chew Xin Ying "Classification and Feature Selection Approaches by Machine Learning T echniques: Heart Disease Prediction", International Journal of Innovative Computing, 2019.

[20] Aditi Gavhane, Gouthami Kokkula, Isha Pandya & Prof. Kailas Devadkar (PhD) "Prediction of Heart Disease Using Machine Learning", ICECA 2018, IEEE Xplore ISBN:978-1-5386-0965-1.

[21] Sonakshi Harjai & Sunil Kumar Khatri, "An Intelligent Clinical Decision Support System Based on Artificial Neural Network for Early Diagnosis of Cardiovascular Diseases in Rural Areas", AICAI, 2019, DOI: 10.1109/AICAI.2019.8701237.

[22] Krishnan, S., & Geetha, S. (2019, April). Prediction of heart disease using machine learning algorithms. In 2019 1st international conference on innovations in information and communication technology (ICIICT) (pp. 1-5). IEEE.

[23] Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. Procedia Computer Science, 85, 962-969.

[24] Nikhar, S., & Karandikar, A. M. (2017). Prediction of Heart Disease Using Different Classification Techniques. Aptikom Journal on Computer Science and Information Technologies, 2(2), 68-74.

[25] Golande, A., & Pavan Kumar, T. (2019). Heart disease prediction using effective machine learning techniques. International Journal of Recent Technology and Engineering, 8(1), 944-950.

[26] Jha, P., Dembla, D., Dubey, W. "Implementation of Machine Learning Classification Algorithm Based on Ensemble Learning for Detection of Vegetable Crops Disease International Journal of Advanced Computer Science and Applications, 2024, 15(1), pp. 584–594.

[27] Jha, Pradeep, Deepak Dembla, and Widhi Dubey 2024. "Implementation of Transfer Learning Based Ensemble Model Using Image Processing for Detection of Potato and Bell Pepper Leaf Diseases." Article. International Journal of Intelligent Systems and Applications in Engineering 12 (8s): 69–80.

[28] Jha, Pradeep, Deepak Dembla, and Widhi Dubey. 2023. "Comparative Analysis of Crop Diseases Detection Using Machine Learning Algorithm." Conference paper. Proceedings of the 3rd International Conference on Artificial Intelligence and Smart Energy, ICAIS 2023. Institute of Electrical; Electronics Engineers Inc. https://doi.org/10.1109/ICAIS56108.2023.10073831.

[29] Jha, Pradeep, Deepak Dembla, and Widhi Dubey. 2023. "Crop Disease Detection and Classification Using Deep Learning-Based Classifier Algorithm." Conference paper. Edited by Rathore V. S., Piuri V., Babo R., and Ferreira M. C. Lecture Notes in Networks and Systems 682 LNNS: 227–37. https://doi.org/10.1007/978-981-99-1946-8_21.

[30] Jha, Pradeep, Deepak Dembla, and Widhi Dubey 2023. "Deep Learning Models for Enhancing Potato Leaf Disease Prediction: Implementation of Transfer Learning Based Stacking Ensemble Model." Article. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-16993-4.

[31] Meshram, Amita, and Deepak Dembla. 2023. "MCBM: Implementation Of Multiclass And Transfer Learning Algorithm Based On Deep Learning Model For Early Detection Of Diabetic Retinopathy." Article. ASEAN Engineering Journal 13 (3): 107–16. https://doi.org/10.11113/aej.V13.19401.

[32] Meshram, Amita, and Deepak Dembla 2023. "Multistage Classification of Retinal Images for Prediction of Diabetic Retinopathy-Based Deep Learning Model." Conference paper. Edited by Rathore V. S., Piuri V., Babo R., and Ferreira M. C. Lecture Notes in Networks and Systems 682 LNNS: 213–26. https://doi.org/10.1007/978-981-99-1946-8_20.

[33] Meshram, Amita, Deepak Dembla, and A. Anooja. 2023. "Development And Analysis Of Deep Learning Model Based On Multiclass Classification Of Retinal Image For Early Detection Of Diabetic Retinopathy." Article. Asean Engineering Journal 13 (3): 89–97. https://doi.org/10.11113/aej.V13.19256.