Global Scientific JOURNALS

GSJ: Volume 12, Issue 9, September 2024, Online: ISSN 2320-9186

www.globalscientificjournal.com

A Smart Heart Diagnostic Machine: Developing an Effective Predictive Classification Algorithm

Adekunle, J. D[1], Oyeniran, M. I[2], Dada, A. M[3], Robert, C. O[4], Sule H. S[5], Akinpelu T. T[6]

1 - 2. Department of Mathematics, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria

3. Information and Documentation department, National Horticultural Research Institute, Ibadan, Nigeria.

4. Department of Management Information Systems, Topdel Engineering Limited, Lagos, Nigeria.

5 – 6. Department of Statistics, Federal University of Agriculture Abeokuta

Corresponding Author: johdam01@gmail.com, Other Authors: oyeniranmatthew@gmail.com, dadaniyi@yahoo.com, rokechukwu123@gmail.com, harunasulesani@gmail.com, temitopeakinpelu98@gmail.com

ABSTRACT

Heart disease remains a leading cause of mortality worldwide, necessitating the development of accurate and reliable diagnostic models. This study aimed to optimize the application of machine learning in the medical field and better the process of decision-making in diagnoses process by analyzing the intricate relationship between various parameters and the precision of heart disease categorization by developing a heart disease detection system that utilises artificial neural networks and machine learning process. An artificial neural networks (ANN) with data scaling on the Cleveland Heart Disease dataset. The proposed Artificial Neural Network (ANN) model with data scaling achieved remarkable performance with an accuracy of 99.61%, surpassing various existing models. In addition to this thalassemia, Major Vessels, and oldpeak exhibited strong positive correlations with the target (presence of heart disease). Specifically, the number of major vessels showed the strongest positive correlation (0.58) with the presence of heart disease, followed by the thalassemia feature (0.52), and the ST depression induced by exercise (oldpeak) with a correlation of 0.44. Conversely, features such as thalassemia (maximum heart rate achieved) and slope (slope of the peak exercise ST segment) demonstrated negative correlations, suggesting their association with a reduced likelihood of heart disease. Furthermore, statistical comparisons between patients with and without heart disease revealed that individuals with heart disease tend to have lower maximum heart rates thalassemia and higher oldpeak values, further supporting the importance of these features in diagnosis. This results improve the precision of machine learning algorithms for diagnosing heart disease, but also provide valuable insights medical professionals. Ultimately, the outcomes contribute to improving understanding and decision-making in various domains related to human care, diagnoses, heart treatment and medical field at large.

Keywords: Heart Disease Prediction, Artificial Neural Network (ANN), Medical Diagnosis, Healthcare AI

1. Introduction

Problem solving is one of the unique ability of human life. The ability to handle complex and complicated issues which brings innovations. But as some point, some issues becomes too complicated, and almost impossible for human to handle. In the olden days, several problem were unsolved due to their complexity nature. The inventory of computer serves as a means of solving some whereas some remains unsolved. But in today's world, almost all human problem are been solved or there is a light to. Through data, things are possible. Several domain of knowledge such as computer science, mathematical science, biological science, medical and others depends greatly on information (data) to decide on critical issues. An availability of data serves as a stepping stone to decision making process - making it so easy for specialist to decide on what to do. In the realms of data, comes machine learning. According to Mahesh (2020), machine learning is the systematic examination of algorithms and statistical models that computer systems employ to execute a certain task without requiring explicit programming. It involve the process of training model which are mathematical in nature to study the behavior of an information (Sarker, 2021). In recent years, machine learning has performed beyond human imagination. And it has been used for several purposes. For example, in a study conducted by Rai et al. (2021), industry paradigm 4.0 was said to encourage the adoption of smart sensors devices, and machines to enhance and improve the rate of production in industries. Also, finance industry has applied machine learning algorithm into their systems for canonical modeling and frame work (Dixon *et al.,* 2020). Machine learning, which is a subset of artificial intelligent is known to be of categories. The simplest of it is supervised learning which allows the usability of labelled data for prediction purposes. The labeled data is utilised to train the model in order to provide predictions or classifications based on the input data it gets. These models initially does an analysis on the training data and creates a conditional function to map fresh examples. The highest setting likely enables the system to accurately assign class labels to the examples it covers. This requires the supervised learning algorithm to intelligently reduce the training data to relevant circumstances. Popular algorithms used in supervised learning encompass Decision Trees, Naive Bayes, and Support Vector Machines. Application of these learning are enormous. On the other hand, unsupervised machine learning are known to take unlabeled data. This type of learning approach aims to establish a learning framework solely for the sake of learning (Tyagi *et al*., 2022).

2. Application of Machine learning in Medicinal field

Within the medical field, several researcher has started applying machine learning for medical treatments, diagnosing health issues, and treatment of patients. A good example is a study conducted by Nashif et al. (2018) where a logistic regression was applied to assess the characteristics of several patients. Also, a random forest classification method is created to accurately detect cardiac disorder. It was reported that the model performed notable accuracy rate of roughly 83% over the training data. On the other hand, Li et al. (2020) built a system using various classification algorithms, including Support Vector Machine, Logistic Regression, Artificial Neural Network, K-Nearest Neighbour, Naïve Bayes, and Decision Tree. Additionally, standard feature selection algorithms such as Relief, Minimal Redundancy Maximal Relevance, Least Absolute Shrinkage Selection Operator, and Local Learning are employed to eliminate irrelevant and redundant features were applied. It was reported that, the experimental results demonstrate the feasibility of using the suggested feature selection method (FCMIM) in conjunction with the classifier support vector machine to develop a sophisticated intelligent system for heart disease identification. The FCMIM-SVM diagnosis system demonstrated superior accuracy in comparison to previously established approaches. This shows that machine learning presents a viable and advantageous approach in heart disease diagnoses and appraisal research, particularly in the context of medical diagnoses (Yadav *et al.,* 2020). Neural networks, a subset of artificial intelligence (AI) inspired by the human brain's neural structure, are computational models designed to recognize patterns and make decisions based on data. In recent years, their application in medical diagnostics has garnered significant attention due to their potential to revolutionize disease detection and patient care.

These networks consist of interconnected nodes (neurons) organized in layers, where each neuron processes information received from preceding layers and passes its output to subsequent layers, allowing for complex data processing and learning. Their ability to learn from large datasets and discern intricate patterns makes them particularly suited for tasks that involve image analysis, signal processing, and data interpretation—essential components of medical diagnostics. In medical diagnostics, neural networks hold promise across several fronts. In a study conducted by Sharma et al.(2018) a convolutional neural networks (CNNs) was used for image analysis. This network excel in analyzing medical images such as X-rays, MRI scans, and histopathological slides. They can identify abnormalities, classify diseases, and even predict outcomes with high accuracy, aiding radiologists and pathologists in making timely and precise diagnoses. Signal Processing: Recurrent neural networks (RNNs) and their variants are adept at processing sequential data, making them valuable for interpreting physiological signals like electrocardiograms (ECGs) or EEGs. They can detect subtle anomalies indicative of heart conditions, neurological disorders, and other diseases, facilitating early intervention and monitoring. NN serves as a decision support system and data integration to generate a comprehensive patient profiles. This holistic approach enables personalized diagnostics and treatment strategies tailored to individual genetic predispositions and health histories. Despite these advancements, challenges such as data privacy concerns, model interpretability, and integration into clinical workflows remain. Continued research and collaboration between clinicians, data scientists, and regulatory bodies are essential to harnessing neural networks' full potential in medical diagnostics while ensuring patient safety and ethical standards.

2. Heart diseases

Heart Disease (HD) has the greatest mortality rate worldwide when compared to other disorders (Rath et al., 2022). The mortality rate caused by heart dieases continues to increase, which is a persistent cause of worry among individuals. The researchers and clinicians are exerting significant efforts to preserve lives affected by HD. Existing literature indicates that numerous scholars are currently doing their studies in various parts of HD (Ahsan & Siddique, 2022; Shukur & Mijwil, 2023). Early detection of heart disease is crucial for optimal outcomes. An early diagnosis provides physicians with additional time to identify a treatment that can effectively manage the symptoms and mitigate the occurrence of potential health complications (Muhammad et al., 2020).

Heart disease, also known as cardiovascular disease (CVD), encompasses a range of conditions that affect the heart and blood vessels. It includes coronary artery disease (which can lead to heart attacks), cerebrovascular disease (which affects blood vessels supplying the brain), peripheral artery disease, rheumatic heart disease, congenital heart disease, and other conditions. Heart disease is the leading cause of death globally. According to the World Health Organization (WHO), an estimated 17.9 million people die each year from cardiovascular diseases, accounting for 31% of all global deaths. This staggering statistic highlights the profound impact of heart disease on public health worldwide. The prevalence of heart disease varies across regions and countries, influenced by factors such as socioeconomic status, lifestyle choices, healthcare infrastructure, and genetic predispositions. High-income countries tend to have higher rates of heart disease due to factors like sedentary lifestyles, unhealthy diets, and better diagnostic capabilities. However, low- and middle-income countries are increasingly affected as they undergo rapid urbanization, which often leads to lifestyle changes detrimental to cardiovascular health. The risk factors pertaining to CVD involves several modifiable and non-modifiable factors which contribute to the development of heart disease. According to Brown et al.(2024), factors such as age, sex and family history of cardiovascular disease are said to be Non-modifiable Risk Factors. Men are generally at higher risk until women catch up post-menopause. On the other hand, Some of the Modifiable Risk Factors identified by Ng et al.(2020) include unhealthy diet , physical inactivity, tobacco use, excessive alcohol consumption, obesity and overweight. Heart disease imposes a significant economic burden on individuals, families, and healthcare systems globally. The economic burden is particularly severe in low- and middle-income countries where resources for prevention, diagnosis, and treatment are limited.

3. Research Methodology

3.1. Data Collection

The Cleveland Heart Disease dataset is a well-known dataset in the field of medical research and machine learning. It is part of the larger collection of datasets created by the Hungarian Institute of Cardiology, the University Hospital in Zurich, and the V.A. Medical Center in Long Beach and Cleveland Clinic Foundation. Among these, the Cleveland dataset is the most commonly used for the development and evaluation of predictive models for heart disease. The dataset is available from the UCI Machine Learning Repository, a popular resource for machine learning datasets. It was collected from the Cleveland Clinic Foundation and includes detailed medical records of patients(Janosi et al., 1989).

Table 1: The Cleveland Heart Disease dataset Composition

Number of Instances: 303 patients.

Number of Attributes: 14 features, including the target variable.

Features	Description	Factors	Quantitative	Qualitative	Missing value
Age	Age of the patient in years		✓		0%
Sex	Gender of the patient	1:male 0:female		✓	0%
cp	Type of chest pain experienced by the patient	0:Typical angina 1:Atypical angina 2:Non-anginal pain 3:Asymptomatic		✓	0%
trestbps	Resting blood pressure (in mm Hg) on admission to the hospital			✓	0%
chol	Serum cholesterol in mg/dl			✓	0%
fbs	Fasting blood sugar > 120 mg/dl	1 = true, 0 = false		✓	0%
restecg	Resting Electrocardiographic Results	0: Normal 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria.		✓	0%
thalach	Maximum heart rate achieved during exercise		✓		0%
exang	Exercise-induced angina	(1 = yes, 0 = no).	✓		0%
Oldpeak	Measure of ST depression		✓		0%
Slope	Slope of the Peak Exercise ST Segment	0: Upsloping 1: Flat 2: Downsloping	✓		0%
Ca	Number of major vessels colored by fluoroscopy			✓	0%
Thal	Thalassemia	3: Normal 6: Fixed defect 7: Reversible defect	✓		0%

Target Variable	Diagnosis of heart disease (angiographic disease status):	0: < 50% diameter narrowing (No heart disease) 1: > 50% diameter narrowing (Heart disease)	✓		0%

Table 2: The Cleveland Heart Disease dataset detected outlier

Features	Outlier count	Outlier Ratio	Outliers Mean	With mean	Without mean
Age	0	0	-	54.4	54.4
Sex	0	0	-	0.696	0.696
Cp	0	0	-	0.942	0.942
trestbps	30	2.93	182	132	130
Chol	16	1.56	436	246	243
Fbs	153	14.9	1	0.149	0
restecg	0	0	-	0.530	0.530
thalach	4	0.390	71	149	149
exang	0	0	-	0.337	0.337
Oldpeak	7	0.683	5.86	1.07	1.04
Slope	0	0	-	1.39	1.39
Ca	87	8.49	3.21	0.754	0.527
Thal	7	0.683	0	2.32	2.34
Target Variable	0	0	-	0.513	0513

3.2. Data Preprocessing

Raw data undergoes preprocessing to ensure quality and consistency. This includes handling missing values (Kang, 2013), removing duplicates (Kwon, 2015), and outliers. Additionally, steps are taken to address potential biases in the data and ensure a balanced representation of heart disease(Figure 1). For training and evaluating of the models, the dataset was transformed to a binary with one and zero. It was ensured that the dataset is a representative of patient with heart disease and those without it as expected to encounter in the real-world scenario. The dataset was split into training and testing sets to assess generalization performance with the percentage, 70% and 30% (Toleva, 2021).

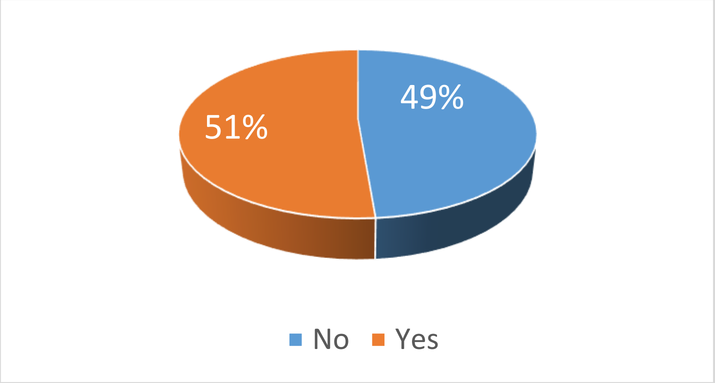


Figure 1: Distribution of heart disease and non-heart diseases

3.3. Tools, Equipment, software, and library or package.

For the implementation of this research project, the following tools and software was used. Programming language: The project was built primarily in R an efficient programming tool for machine learning implementation. Considering the preprocessing phase of the project before building the model, tidyverse which is a collection of R packages was used for data wrangling and data manipulation, numerical operation and array manipulation. Tidyverse, dplyr and the grammar of graphic (ggplot2) data visualization package for the statistical programming language R was used to handle the visualization part of the project. These tools offers and enhanced aesthetics and built-in function for complex, creative static, and interactive visualizations. For the prototyping aspect of the project, which has to do with the computing environments, an Rstudio development environment was used.

3.2 Neural Network Architecture

The input layer consists of 13 features. These features are represented as a vector x of samples of patients with coronary heart disease:

$$X = \ [age(x_1), sex(x_2), cp(x_3), \ldots thal(x_{13})]^T$$

Each input feature x_i is associated with a weights w_i and biases b_i. The weights and biases are initialized at random. For the first layer l (hidden layer), the weights $w^{(I)}$ and biases $b^{(I)}$. Each neuron in a hidden layer performs a weighted sum of its inputs, adds a bias, and applies an activation function. For the j-th neuron in the hidden layer and the k-th hidden layer with m, the computation is conducted (equation 1)

$$Z_j^{(k)} = \sum_{i=1}^{m} w_{ji}^{(k)} x_i + \ b_j^{(k)} ----(1)$$

K = 1, 2, 3, … 13

Where:

$w_{ji}^{(k)}$ is the weight from the i-th input to the j-th hidden neuron.
$b_j^{(k)}$ is the bias for the j-th hidden neuron.
$z_j^{(k)}$ is the pre-activation value for the j-th hidden neuron.
After applying an activation function g (•), the output of the j-th hidden neuron is:

$$a_j^{(k)} = g\left(z_j^{(k)}\right) ------(2)$$

This process repeats for each k-th hidden layers. Also, the binary classification, the output layer typically has a single neuron that produces a value $\hat{y}$ which represent the probability of heart diseases (one class). The pre-activation value for the output neuron is represented by the mathematical expression below

$$Z^{(2)} = \sum_{j=1}^{m} w_j^{(2)} a_j^{(1)} + \ b^2 ------- \ (3)$$

The output is obtained by applying the sigmoid activation function σ (•), which is commonly used in binary classification:

$$\hat{y} = \ \sigma\left(z^{(2)}\right) = \frac{1}{1 + \ e^{-z^{(2)}}} ------- \ (4)$$

For binary classification, we typically use the binary cross-entropy loss function:

$$\mathcal{L}(y, \hat{y}) = \ -(y \log \hat{y}) + (1-y)\log(1- \ \hat{y})) \ \text{L(y, yˆ) -------------- (5)}$$

Where:

y is the true label (0 or 1).
$\hat{y}$ is the predicted probability.

To update the weights and biases, a backpropagation was performed. This involves computing the gradients of the loss with respect to each weight and bias. At first, the output layer gradient of the loss with respect to the pre-activation value $z^{(2)}$ is:

$$\frac{\partial \mathcal{L}}{\partial z^{(2)}} = \hat{y} - y ---------- (6)$$

The gradients for the weights and bias of the output layer are:

$$\frac{\partial \mathcal{L}}{\partial w_j^{(2)}} = \frac{\partial \mathcal{L}}{\partial z^{(2)}} * a_j^{(1)} ---------- (7)$$

$$\frac{\partial \mathcal{L}}{\partial b^{(2)}} = \frac{\partial \mathcal{L}}{\partial z^{(2)}}$$

For the hidden layer, the gradient of the loss with respect to the pre-activation value $z_j^{(1)}$ is:

$$\frac{\partial \mathcal{L}}{\partial z_j^{(1)}} = \left(\frac{\partial \mathcal{L}}{\partial z^{(2)}} * w_j^{(2)}\right) * \acute{g}\left(z_j^{(1)}\right) -------- (8)$$

Where $\acute{g}\left(z_j^{(1)}\right)$ the derivative of the activation function used in the hidden layer. The gradients for the weights and biases of the hidden layer are:

$$\frac{\partial \mathcal{L}}{\partial w_{ji}^{(1)}} = \frac{\partial \mathcal{L}}{\partial w_j^{(1)}} * x_i -------- (9)$$

$$\frac{\partial \mathcal{L}}{\partial b_j^{(1)}} = \frac{\partial \mathcal{L}}{\partial z_j^{(1)}}$$

Using the gradients computed, the weights and biases were update using gradient descent for a learning rate η. This update is done similarly for the output layer weights and biases.

$$w_{ji}^{(1)} \leftarrow w_{ji}^{(1)} - \eta \frac{\partial \mathcal{L}}{\partial w_{ji}^{(1)}} ---------- (10)$$

$$w_j^{(1)} \leftarrow w_j^{(1)} - \eta \frac{\partial \mathcal{L}}{\partial b_j^{(1)}} ---------- (11)$$

During inference, the model predicts a patient has heart disease if $\hat{y} \geq 0.5$ and no heart diseases otherwise. This process iteratively improves the model through training, minimizing the loss function to achieve accurate binary classification.

$$\hat{y} = \sigma\left(W^{(2)} g\left(W^{(1)} x + b^{(1)}\right) + b^{(2)}\right) ------- (12)$$

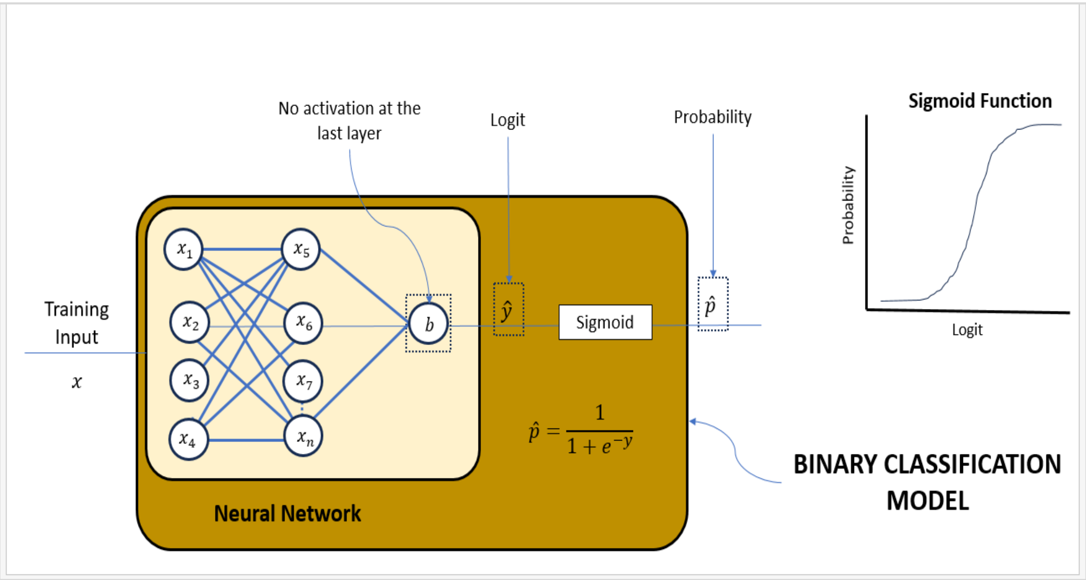


Figure 2: Binary Artificial neural network structure

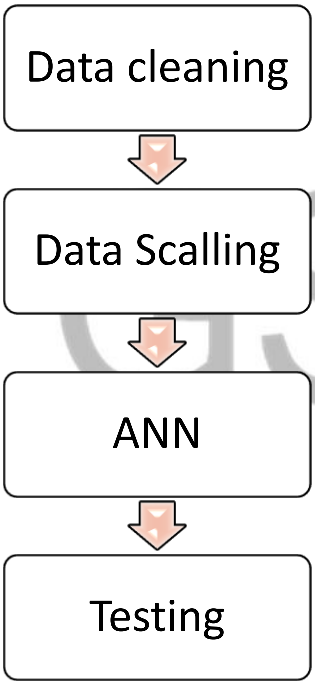
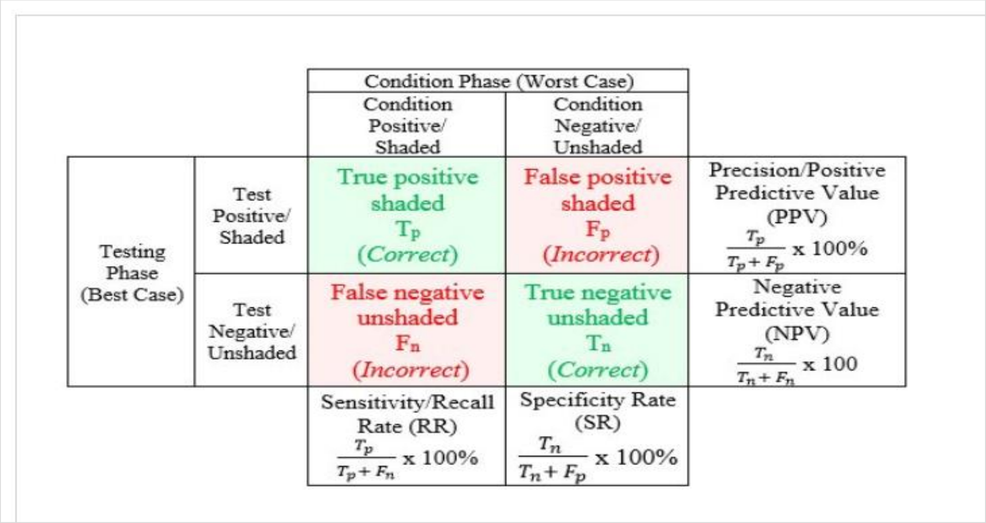


Figure 3: The model training flow

3. Performance Evaluation metrics

The confusion matrix(Figure 4) for our heart disease predictive classification algorithm can be described as follows:1)True Positives (TP): This represents the number of patients correctly identified as having heart disease by the neural network;2)True Negatives (TN): This is the number of patients correctly identified as not having heart disease;3)False Positives (FP): This represents the number of patients incorrectly identified as having heart disease when they do not actually have it;4)False Negatives (FN): This is the number of patients incorrectly identified as not having heart disease when they actually have it.



		Condition Phase (Worst Case)		
		Condition Positive/ Shaded	Condition Negative/ Unshaded	
Testing Phase (Best Case)	Test Positive/ Shaded	True positive shaded T_p *(Correct)*	False positive shaded F_p *(Incorrect)*	Precision/Positive Predictive Value (PPV) $\frac{T_p}{T_p+F_p}$ x 100%
	Test Negative/ Unshaded	False negative unshaded F_n *(Incorrect)*	True negative unshaded T_n *(Correct)*	Negative Predictive Value (NPV) $\frac{T_n}{T_n+F_n}$ x 100
		Sensitivity/Recall Rate (RR) $\frac{T_p}{T_p+F_n}$ x 100%	Specificity Rate (SR) $\frac{T_n}{T_n+F_p}$ x 100%	

Figure 4: Modified confusion matrix table for model accuracy (Opanin & Missah, 2017)

3. Result

This section presents the outcomes of model for heart disease detection - highlighting key findings from its application and it effectiveness in identifying patients with heart disease and provide insights into its overall diagnostic capabilities.

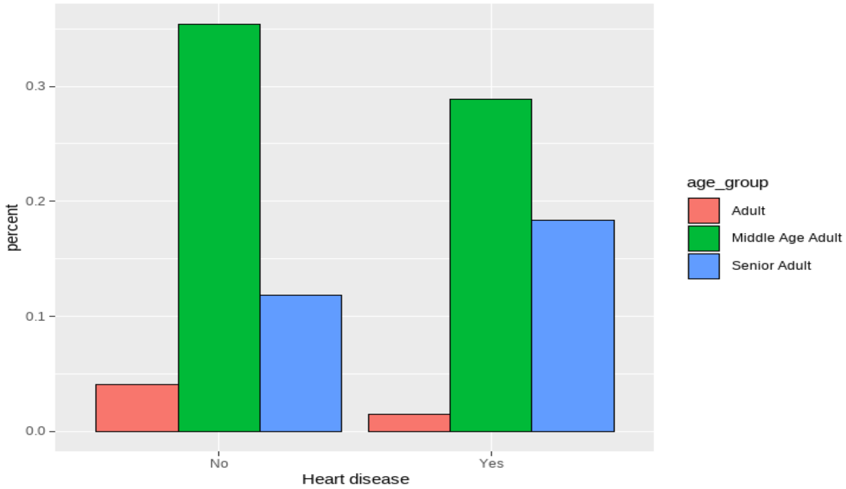
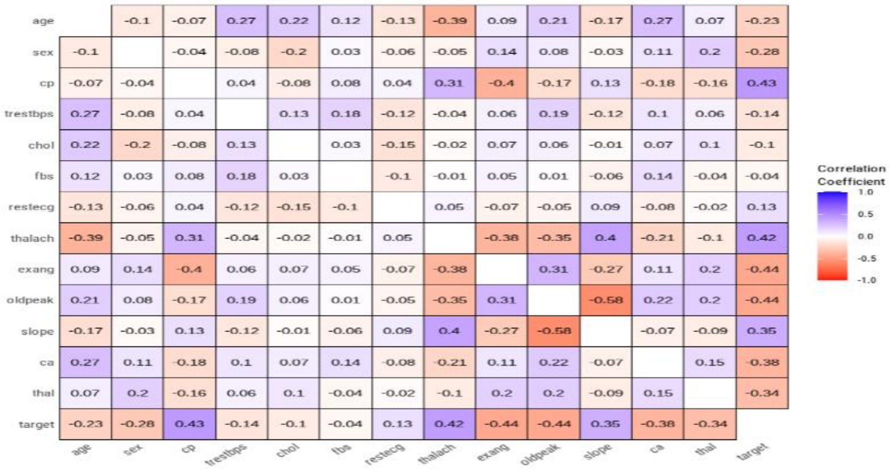


Figure 5: Distribution of heart disease by age group

Figure 5 shows that Middle-Age Adults have the highest percentage of individuals both with and without heart disease, indicating that this age group is most prevalent in the dataset. However, the proportion of Senior Adults (blue bars) with heart disease is notably higher compared to those without, suggesting an increased risk of heart disease as age progresses. In contrast, the Adult group shows a relatively low percentage for both categories, particularly for those with heart disease, implying that younger individuals are less affected by heart disease in.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age		-0.1	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
sex	-0.1		-0.04	-0.08	-0.2	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.2	-0.28
cp	-0.07	-0.04		0.04	-0.08	0.08	0.04	0.31	-0.4	-0.17	0.13	-0.18	-0.16	0.43
trestbps	0.27	-0.08	0.04		0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.1	0.06	-0.14
chol	0.22	-0.2	-0.08	0.13		0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.1	-0.1
fbs	0.12	0.03	0.08	0.18	0.03		-0.1	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
restecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.1		0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
thalach	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05		-0.38	-0.35	0.4	-0.21	-0.1	0.42
exang	0.09	0.14	-0.4	0.06	0.07	0.05	-0.07	-0.38		0.31	-0.27	0.11	0.2	-0.44
oldpeak	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31		-0.58	0.22	0.2	-0.44
slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.4	-0.27	-0.58		-0.07	-0.09	0.35
ca	0.27	0.11	-0.18	0.1	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07		0.15	-0.38
thal	0.07	0.2	-0.16	0.06	0.1	-0.04	-0.02	-0.1	0.2	0.2	-0.09	0.15		-0.34
target	-0.23	-0.28	0.43	-0.14	-0.1	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	

Figure 6: Relationship between the features

Figure 6 shows that notably, cp (chest pain type) and thalach (maximum heart rate achieved) show strong positive correlations with the target (0.43 and 0.42, respectively). Conversely, slope (slope of the peak exercise ST segment), exang (exercise-induced angina), and oldpeak (ST depression induced by exercise relative to rest) exhibit strong negative correlations with the target (-0.58, -0.44, and -0.44, respectively), indicating a potential inverse relationship. Additionally, ca (number of major vessels colored by fluoroscopy) shows a moderate negative correlation with the target (-0.38), while other features like chol (cholesterol level) and fbs (fasting blood sugar) display weak correlations, implying limited predictive value.

Table 3: Description of heart disease by different factors

	Heart disease	Average	Max	Min	sd	Range
Age	No	56.6	77	35	7.91	42
	Yes	52.4	76	29	9.63	47
Trestbps	Yes	134	200	100	18.6	100
	No	129	180	94	16.1	86
Chol	Yes	251	409	131	49.6	278
	No	241	564	126	53.0	438
Thalach	Yes	139	195	71	22.6	124
	No	159	202	96	19.1	106
Oldpeak	Yes	1.60	6.2	0	1.29	6.2
	No	0.57	4.2	0	0.771	4.2

On average, those with heart disease are younger (52.4 years) compared to those without (56.6 years) and exhibit higher resting blood pressure (134 mmHg vs. 129 mmHg) and cholesterol levels (251 mg/dL vs. 241 mg/dL). Additionally, individuals with heart disease have a lower maximum heart rate (139 bpm) compared to those without (159 bpm) and experience more pronounced ST depression (Oldpeak of 1.60 vs. 0.57).

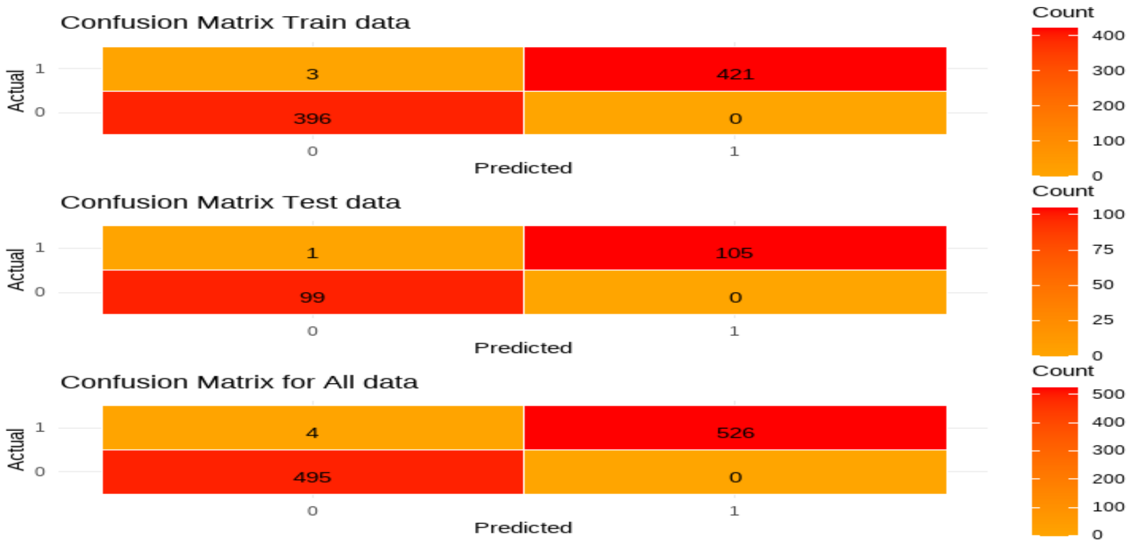


Figure 7: Confusion matrix of the model for the three situations

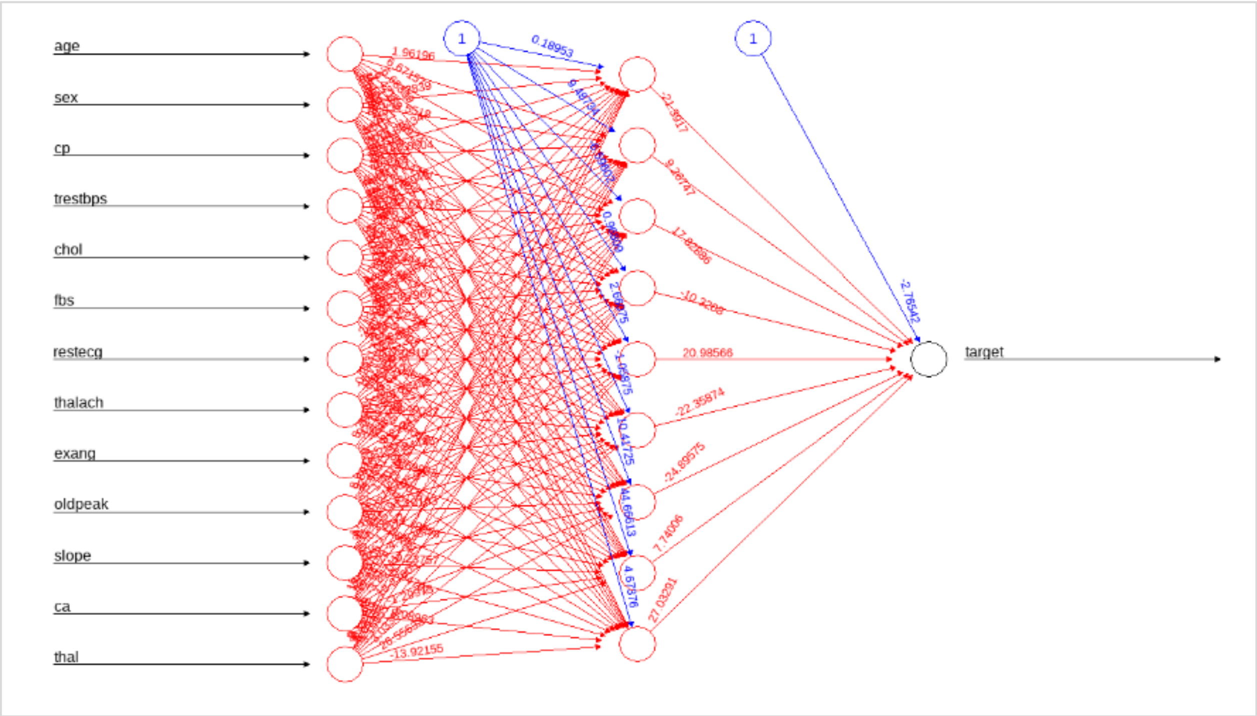
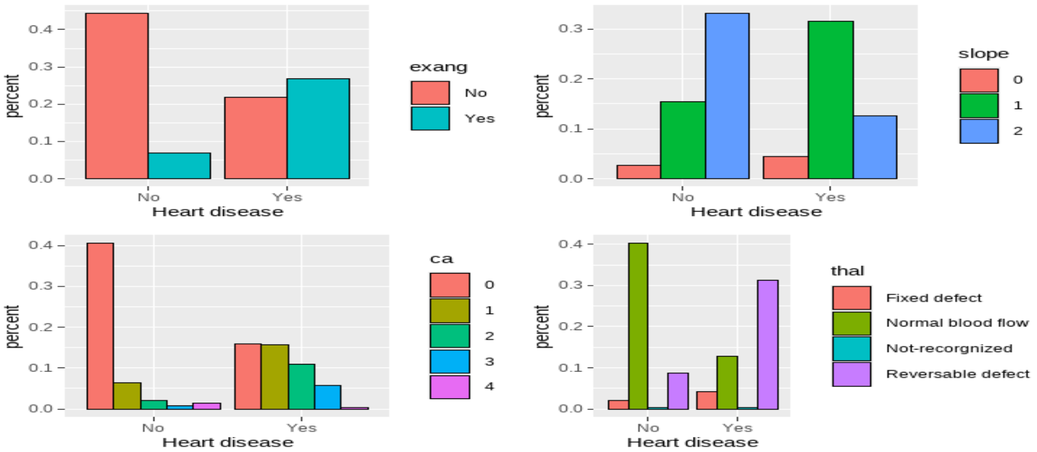


Figure 8: Feature interactions and their impact on the model's outcome.



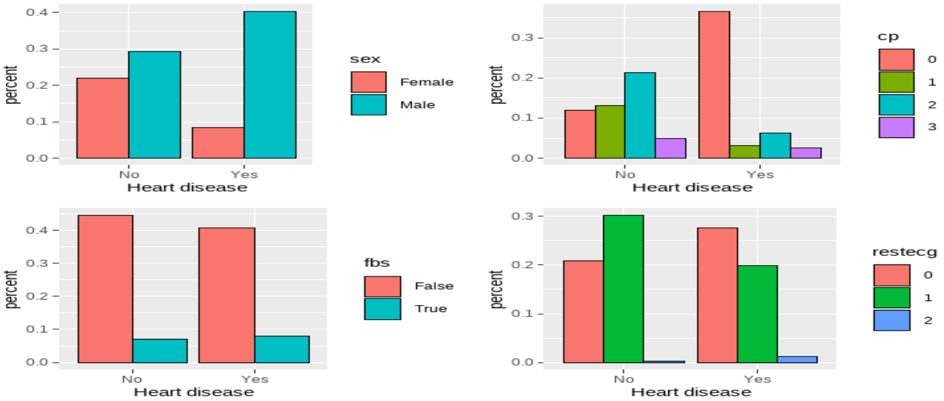


Figure 9: Description of patient physiology characteristics

Figure 9 shows that most patients without heart disease do not experience angina, while those with heart disease are more evenly split between having and not having angina. The slope of the peak exercise ST segment, revealing that slope 2 is common in both groups, but slope 1 is more prevalent among those with heart disease, suggesting a possible link. The number of major vessels colored by fluoroscopy (ca), indicating that patients without heart disease typically have no vessels colored, while those with heart disease show a more varied distribution, with higher counts of colored vessels possibly associated with the disease. The thalassemia (thal), where patients without heart disease predominantly exhibit normal blood flow, whereas those with heart disease have higher instances of fixed and reversible defects.

Table 4: Performance metrics of the model (A)

Metrics	Train_data	Test_data	All_data
Sensitivity	1	1	1
Specificity	0.9929	0.9906	0.9925
Pos Pred Value	0.9925	0.99	0.992
Neg Pred Value	1	1	1
Prevalence	0.4829	0.4829	0.4829
Detection Rate	0.4829	0.4829	0.4829
Detection Prevalence	0.4866	0.4878	0.4868
Balanced Accuracy	0.9965	0.9953	0.9962

The performance metrics for the model across the training, testing, and combined (training + testing) demonstrate consistently high accuracy (Table 4). The model achieves perfect Sensitivity (1.0) across all situations. Specificity and Predictive Value is slightly lower, around 0.99. The Negative Predictive Value (NPV) is perfect at 1.0. The Prevalence and Detection Rate are consistent at 0.4829. The Detection Prevalence is slightly higher, around 0.487. Finally, the Balanced Accuracy is nearly perfect at around 0.996, indicating the model's exceptional overall performance in distinguishing between positive and negative cases.

Table 5: The performance metrics of the model (B)

	Accuracy	95% CI	No Information Rate	P-Value [Acc > NIR]	Kappa	Mcnemar’s Test P-Value
Train_data	0.9963	(0.9893, 0.9992)	0.5171	<2e-16	0.9927	0.2482
Test_data	0.9951	(0.9731, 0.9999)	0.5171	<2e-16	0.9902	1
All_data	0.9961	(0.99, 0.9989)	0.5171	<2e-16	0.9922	0.1336

Error	4.284659

The model demonstrates exceptional performance across all situations (training, testing, and training+testing), with accuracy rates of 99.6% for the training and training+testing, and 99.5% for the testing (Table 5). The 95% confidence intervals for these accuracy rates are narrow, underscoring the reliability of the model's predictions, though the test data shows a slightly wider range due to its smaller size. The model significantly outperforms random chance (NIR=51.71%, p < 0.01). The Kappa statistic, around 0.99 across all situations, reflects near-perfect agreement between the model's predictions and actual outcomes. Similarly, there is no significant difference in the model's error rates between the two classes, implying balanced performance (McNemar's Test p>0.01). The low overall error rate further confirm the accuracy and effectiveness in classifying cases.

Table 6: Comparative Performance of Various Machine Learning Methods for Heart Disease Prediction

Author	Method	No Attr	Acc
Perumal et al.(2020)	LR with PCA	7	87
Latha et al.(2019)	Majority vote with NB, BN, RF,and MP	9	85.48
Vishnu et al.(2021)	Chi-Square + SMO	11	86.468
Saqlain et al.(2018)	Forward feature selection with Radia Basis Function SVM	7	81.19
Tama et al.(2020)	Two-tier ensembe PSO	7	85.6
Gazeloglu et al.(2020)	Fuzzy Rough set and Chi-Square with Radial bias Fuction Network	7	81.188
Gazeloglu et al.(2020)	Correlation-based feature selection with NB	6	84.818
Pavithra et al.(2021)	HRFLC (RF + AdaBoost + Pearson Coefficient)	11	79
Tougui et al.(2020)	ANN	14	85.86
Kodati et al.(2018)	SMO	14	84
Sultana et al.(2016)	SMO	14	84.07
Gupta et al.(2019)	FAMD + RF	28	93.44
Kumar et al.(2020)	Random Forest	10	85.71
Ananey-Obiri et al.(2020)	LR and GNB with Single valuedecomposition	4	82.75
Naresha (2023).	Extreme Gradient Boost	14	90.16393
Syed (2024).	Decision Tree	14	98.5366
Hardik (2020).	Logistic model	14	86.885
Propose Method	**ANN+Data Scaling(DS)**	**14**	**99.61**

4. Discussion

The result of this study underscore the significant physiological differences associated with heart disease, particularly in blood pressure, cholesterol, heart rate, and exercise-induced ST depression. The overall trend highlights that heart disease prevalence increases with age, particularly in the Senior Adult group(Table 3). Also, there is a greater variability in these metrics among individuals with heart disease, indicating a more diverse set of risk factors and health outcomes within this group. There is a clear distinction between patients with and without heart disease (Table 3). On average, patients with heart disease tend to be younger (52.4 years) compared to those without the condition (56.6 years)(Mbakwem et al., 2023; Ambroziak et al., 2020). Blood pressure (trestbps) and cholesterol (chol) levels are notably higher in those with heart disease, with average values of 134 mmHg and 251 mg/dL, respectively, compared to 129 mmHg and 241 mg/dL in those without heart disease (Cleveland Clinic, 2022; Satoh et al., 2021). Maximum heart rate achieved is also lower in those with heart disease (139 bpm) compared to those without (159 bpm). Finally, oldpeak, which measures ST depression during exercise, is significantly higher in patients with heart disease (average 1.60) compared to those without (0.57), indicating greater cardiovascular stress in the former group (Lanza et al., 2004; Shahjehan, 2023).

It was found that chest pain type has a strong positive correlation with heart disease (0.43), indicating that patients experiencing certain types of chest pain are more likely to have heart disease, making this a crucial factor in diagnosis(Figure 6). Similarly, maximum heart rate achieved also has a positive correlation (0.42) with heart disease, suggesting that higher heart rates during stress tests could signal a greater risk for cardiac issues. Interestingly, the number of major vessels affected (ca) shows a strong negative correlation (-0.59) with heart disease, implying that fewer affected vessels may be linked to a higher risk of heart disease, or that specific diagnostic approaches using this variable are more indicative of certain forms of the disease. Additionally, ST depression during exercise (oldpeak) has a negative correlation (-0.44), reflecting that greater ST depression often points to increased likelihood of heart disease, as it typically indicates stress or ischemia in the heart during exertion(Hickam, 1990).

Exercise-induced angina (exang) has a negative correlation with both maximum heart rate (-0.38) and heart disease (-0.44), suggesting that those experiencing angina during exercise are more likely to have heart disease and may have reduced exercise capacity. This reinforces the role of exercise testing in identifying heart conditions. On the other hand, the weak negative correlation between sex and heart disease (-0.28) suggests that men may be slightly more prone to heart disease, though the association is not particularly strong. Other variables, like fasting blood sugar (fbs) and resting electrocardiographic results (restecg), show negligible correlations, suggesting they may be less relevant in predicting heart disease (Bekkouche et al., 2013; Hickam, 1990).

The performance metrics of the model which was evaluated on three instances (training, testing, and combined) reveal a crucial information in clinical settings where missing a positive case could have serious consequences (Table4 – Table 5). The sensitivity (or True Positive Rate) is perfect across all instances (1.0). This result was further enforced by the specificity which is also high but slightly lower than sensitivity, with values of 0.9929 for training, 0.9906 for testing, and 0.9925 for the combined. This means that in all instances, the model accurately identifies all cases of heart disease and without heart disease patients. Other metrics such as Prevalence, Detection Rate, and Detection Prevalence show consistent results across the datasets, indicating that the model maintains stability regardless of the sample size or type of data (train/test).

The model's high sensitivity and specificity make it a valuable tool in clinical settings, particularly in screening for heart disease. This support the result of a study conducted by Yan et al.(2019) which found the significant of integrating artificial intelligent into existing workflows by using it as a decision-support tool during patient assessments while it assist healthcare providers in determining the necessity for further diagnostic tests (such as ECG or stress tests) based on initial screening results. According to Setyati et al.(2024), patients benefit from early detection and accurate diagnosis, which are critical for managing heart disease. Moreover, the model high Negative Predictive Value(Table 4) ensures that clinicians can confidently rule out heart disease in patients who test negative, potentially reducing unnecessary follow-up tests and associated healthcare costs.

One limitation of the current model is that it was developed using a dataset with a specific set of clinical variables. It may not generalize well to other populations or regions where risk factors for heart disease might differ. To improve the model, incorporating additional relevant features—such as lifestyle factors, family history, and more granular clinical data—could enhance its predictive power. Additionally, retraining the model using a more diverse and larger dataset may increase its generalizability. Further research could explore the model's performance across different populations to evaluate its robustness. Finally, research into developing user-friendly interfaces for healthcare professionals could facilitate the model's broader adoption in clinical environments, ensuring that its benefits are realized in practice.

Conclusion

The key findings of this study demonstrate that the smart heart diagnostic machine offers highly accurate predictive capabilities for diagnosing heart disease, with perfect Sensitivity and high Specificity across the three instances (training, testing, and combined datasets). This shows that smart heart diagnostic machine has the potential to revolutionize healthcare by improving the accuracy and speed of heart disease diagnosis. Its integration into clinical workflows can lead to better patient outcomes through early detection, timely treatment, and more personalized care. For healthcare providers, the model can optimize resource allocation, streamline decision-making, and reduce the burden on medical staff by automating the initial stages of diagnosis. In conclusion, the smart heart diagnostic machine represents a significant advancement in the application of machine learning in healthcare. Its ability to deliver highly reliable predictions, combined with its ease of integration into clinical settings, makes it a powerful tool for improving cardiovascular health outcomes. We encourage healthcare institutions to adopt and further develop this technology to enhance patient care and ultimately reduce the global burden of heart disease.

8. References

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 128, 102289.

Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine LearningAlgorithms. Int. J. Comput. Appl. 2020, 176, 17–21

Choy, Lennon H. T., and Winky K. O. Ho. (2023). "The Use of Machine Learning in Real Estate Research" Land 12, no. 4: 740. https://doi.org/10.3390/land12040740

Dixon, M. F., Halperin, I., & Bilokon, P. (2020). Machine learning in finance (Vol. 1170). New York, NY, USA*: Springer International Publishing*.

Gazelo ˘glu, C. Prediction of heart disease by classifying with feature selection and machine learning methods. Prog. Nutr. 2020,22, 660–670

Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access2019, 8, 14659–14674.

Hardik, D. (2020). Heart Disease UCI-Diagnosis & Prediction, Medium, accessed on 6/22/2024, from: https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7

Kodati, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange and Weka Sri Satya Sai UniversityAnalysis of Heart Disease using in Data Mining Tools Orange and Weka. Glob. J. Comput. Sci. Technol. 2018, 18

Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine LearningClassifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems(ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21.

Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques.Inform. Med. Unlocked 2019, 16, 100203

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, 8, 107562-107582.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. 9(1), 381-386.

Muhammad, Y., Tahir, M., Hayat, M., & Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. *Scientific reports*, 10(1), 19747.

Nashif, S., Raihan, M. R., Islam, M. R., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, 6(4), 854-873.

Naresha, B. (2023). Heart Attack Prediction Using Different ML Models, Kaggle, accessed on 6/22/2024, from: https://www.kaggle.com/code/nareshbhat/heart-attack-prediction-using-different-ml-models Kaggle

Opanin Gyamfi, Enoch & Missah, Yaw. (2017). Pixel-Based Unsupervised Classification

Pavithra, V.; Jayalakshmi, V. Hybrid feature selection technique for prediction of cardiovascular diseases. Mater. Today Proc. 2021,22, 660–670

Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. Int. J. Adv.Sci. Technol. 2020, 29, 4225–4234

Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications. *International Journal of Production Research*, 59(16), 4773-4778.

Rath, A., Mishra, D., Panda, G., & Satapathy, S. C. (2022). An exhaustive review of machine and deep learning based diagnosis of heart diseases. *Multimedia Tools and Applications*, 81(25), 36069-36127.

Saqlain, S.M.; Sher, M.; Shah, F.A.; Khan, I.; Ashraf, M.U.; Awais, M.; Ghani, A. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowl. Inf. Syst. 2018, 58, 139–167

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3), 160.

Shukur, B. S., & Mijwil, M. M. (2023). Involving machine learning techniques in heart disease diagnosis: a performance analysis. *International Journal of Electrical and Computer Engineering*, 13(2), 2177.

Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of the 20163rd International Conference on Electrical Engineering and Information and Communication Technology, iCEEiCT 2016, Dhaka,Bangladesh, 22–24 September 2016; pp. 1–5.

Syed, A. H. (2024). *Heart Disease Detection*, Kaggle, accessed on 6/22/2024, from: https://www.kaggle.com/code/syedali110/heart-disease-detection

Tama, B.A.; Im, S.; Lee, S. Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier ClassifierEnsemble. BioMed Res. Int. 2020, 2020.

Tougui, I.; Jilbab, A.; El Mhamdi, J. Heart disease classification using data mining tools and machine learning techniques. HealthTechnol. 2020, 10, 1137–1144.

Tripathy JP. (2013). Secondary Data Analysis: Ethical Issues and Challenges. *Iran J Public Health*. 42(12):1478-9. *PMID:* 26060652; PMCID: PMC4441947.

Vishnu Vardhana Reddy, Karna & Elamvazuthi, Irraivan & Aziz, Azrina & Paramasivam, Sivajothi & Chua, Hui Na & Pranavanand, S.. (2021). Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Applied Sciences. 11. 8352. 10.3390/app11188352.

Yadav, S. S., Jadhav, S. M., Nagrale, S., & Patil, N. (2020, March). Application of machine learning for the detection of heart disease. In 2020 2nd *international conference on innovative mechanisms for industry applications* (ICIMIA) (pp. 165-172). *IEEE*.

Yazdani, A., Varathan, K.D., Chiam, Y.K., Malik, A.W., Wan Ahmad, W.A. (2021). A novel approach for heart disease prediction using strength scores with significant predictors. BMC Med Inform Decis Mak. 21(1):194. doi: 10.1186/s12911-021-01527-5.

Tyagi, K., Rane, C., Sriram, R., Manry, M. (2022). Chapter 3 - Unsupervised learning. In R. Pandey, S. K. Khatri, N. K. Singh, & P. Verma (Eds.), Artificial Intelligence and Machine Learning for EDGE Computing (pp. 33-52). *Academic Press*. ISBN 9780128240540. https://doi.org/10.1016/B978-0-12-824054-0.00012-5.

Mbakwem AC, Amadi CE, Ajuluchukwu JN, Kushimo OA. Trends and outcomes of cardiovascular disease admissions in Lagos, Nigeria: a 16-year review. Cardiovasc J Afr. 2023 Jul-Aug 23;34(3):140-148. doi: 10.5830/CVJA-2022-037. Epub 2022 Aug 30. PMID: 36044243; PMCID: PMC10658729.

Ambroziak M, Niewczas-Wieprzowska K, Maicka A, Budaj A. Younger age of patients with myocardial infarction is associated with a higher number of relatives with a history of premature atherosclerosis. BMC Cardiovasc Disord. 2020 Sep 11;20(1):410. doi: 10.1186/s12872-020-01677-w. PMID: 32912162; PMCID: PMC7488448.

Cleveland Clinic. (2022, May 24). High cholesterol diseases. https://my.clevelandclinic.org/health/articles/11918-cholesterol-high-cholesterol-diseases

Satoh M, Ohkubo T, Asayama K, Murakami Y, Sugiyama D, Waki T, Tanaka-Mizuno S, Yamada M, Saitoh S, Sakata K, Irie F, Sairenchi T, Ishikawa S, Kiyama M, Okayama A, Miura K, Imai Y, Ueshima H, Okamura T; Evidence for Cardiovascular Prevention from Observational Cohorts in Japan (EPOCH–JAPAN) Research Group. A Combination of Blood Pressure and Total Cholesterol Increases the Lifetime Risk of Coronary Heart Disease Mortality: EPOCH-JAPAN. J Atheroscler Thromb. 2021 Jan 1;28(1):6-24. doi: 10.5551/jat.52613. Epub 2020 Apr 8. PMID: 32269207; PMCID: PMC7875142.

Lanza GA, Mustilli M, Sestito A, Infusino F, Sgueglia GA, Crea F. Diagnostic and prognostic value of ST segment depression limited to the recovery phase of exercise stress test. Heart. 2004 Dec;90(12):1417-21. doi: 10.1136/hrt.2003.031260. PMID: 15547017; PMCID: PMC1768611.

Hickam DH. Chest Pain or Discomfort. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths; 1990. Chapter 9. Available from: https://www.ncbi.nlm.nih.gov/books/NBK416/

Shahjehan RD, Bhutta BS. Coronary Artery Disease. [Updated 2023 Aug 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK564304/

Bekkouche NS, Wawrzyniak AJ, Whittaker KS, Ketterer MW, Krantz DS. Psychological and physiological predictors of angina during exercise-induced ischemia in patients with coronary artery disease. Psychosom Med. 2013 May;75(4):413-21. doi: 10.1097/PSY.0b013e31828c4cb4. Epub 2013 Apr 10. PMID: 23576766; PMCID: PMC3646947.

Yan Y, Zhang JW, Zang GY, Pu J. The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine? J Geriatr Cardiol. 2019 Aug;16(8):585-591. doi: 10.11909/j.issn.1671-5411.2019.08.010. PMID: 31555325; PMCID: PMC6748906.

Setyati, Rina & Astuti, Aldiana & Utami, Tyas & Adiwjaya, Saputra & Hasyim, Dadang. (2024). The Importance of Early Detection in Disease Management. Journal of World Future Medicine Health and Nursing. 2. 51-63. 10.55849/health.v2i1.692.

Brown JC, Gerhardt TE, Kwon E. Risk Factors for Coronary Artery Disease. [Updated 2023 Jan 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554410/

Ng R, Sutradhar R, Yao Z, Wodchis WP, Rosella LC. Smoking, drinking, diet and physical activity-modifiable lifestyle risk factors and their associations with age to first chronic disease. Int J Epidemiol. 2020 Feb 1;49(1):113-130. doi: 10.1093/ije/dyz078. PMID: 31329872; PMCID: PMC7124486.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Sharma, Neha & Jain, Vibhor & Mishra, Anju. (2018). An Analysis Of Convolutional Neural Networks For Image Classification. Procedia Computer Science. 132. 377-384. 10.1016/j.procs.2018.05.198.

8. Supplement

```
library(caTools)
library(neuralnet)
library(tidyverse)
heart <- read_csv("C:\\Users\\DELL\\Desktop\\2024_Projects\\Project work from Deji\\Smart heart diagnostic
        machine dataset\\heart.csv")
head(heart,5)
glimpse(heart)
library(dlookr)
diagnose_outlier(heart)
plot_na_pareto(heart)

library(ggcorrplot)
plot_correlate(heart)

set.seed(123)

Age <- heart %>% group_by(target) %>% summarize(ave =mean(age), max =max(age), min=min(age), sd=sd(age),
range=max(age)-min(age))

trestbps <- heart %>% group_by(target) %>% summarize(ave =mean(trestbps), max =max(trestbps),
                              min=min(trestbps), sd=sd(trestbps), range=max(trestbps)-min(trestbps))

chol <- heart %>% group_by(target) %>% summarize(ave =mean(chol), max =max(chol),
                              min=min(chol), sd=sd(chol), range=max(chol)-min(chol))

thalach <- heart %>% group_by(target) %>% summarize(ave =mean(thalach), max =max(thalach),
                                min=min(thalach), sd=sd(thalach), range=max(thalach)-min(thalach))

oldpeak <- heart %>% group_by(target) %>% summarize(ave =mean(oldpeak), max =max(oldpeak),
                                min=min(oldpeak), sd=sd(oldpeak), range=max(oldpeak)-min(oldpeak))

rbind(Age,trestbps,chol,thalach,oldpeak)
Heart <- heart %>%  mutate(age_group = case_when( age >= 13 & age <= 19 ~ "Teen",
 age >= 20 & age <= 39 ~ "Adult",
 age >= 40 & age <= 59 ~ "Middle Age Adult",
 age >= 60 ~ "Senior Adult",
 TRUE ~ "Other"
) , sex =ifelse(sex==1,"Male", "Female"), fbs = ifelse(fbs ==1, "True", "False"),
thal = case_when(thal==0~"Not-recorgnized",thal== 1~"Fixed defect",
          thal==2~"Normal blood flow",thal==3~"Reversable defect",TRUE ~ "Other"),
exang=ifelse(exang==1,"Yes","No"), `Heart disease` = ifelse(target==1, "No", "Yes"))

agre <- Heart %>% group_by( `Heart disease`, age_group) %>% summarize(percent = n()/nrow(heart))

ggplot(agre, aes(x=  `Heart disease`, y=percent, fill =age_group))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")


Heart$cp <- as.factor(Heart$cp)

sex <- Heart %>% group_by( `Heart disease`, sex) %>% summarize(percent = n()/nrow(heart))

k1<-ggplot(sex, aes(x=  `Heart disease`, y=percent, fill =sex))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")

cp <- Heart %>% group_by( `Heart disease`, cp) %>% summarize(percent = n()/nrow(heart))
k2<-ggplot(cp, aes(x= `Heart disease`, y=percent, fill =cp))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")

fbs <- Heart %>% group_by( `Heart disease`, fbs) %>% summarize(percent = n()/nrow(heart))
k3<-ggplot(fbs, aes(x=  `Heart disease`, y=percent, fill =fbs))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")
Heart$restecg <- as.factor(Heart$restecg)
rest <- Heart %>% group_by( `Heart disease`, restecg) %>% summarize(percent = n()/nrow(heart))
k4 <-ggplot(rest, aes(x=  `Heart disease`, y=percent, fill =restecg))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")

exang <- Heart %>% group_by( `Heart disease`, exang) %>% summarize(percent = n()/nrow(heart))
k5<-ggplot(exang, aes(x= `Heart disease`, y=percent, fill = exang))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")
Heart$slope <- as.factor(Heart$slope)
slope <- Heart %>% group_by( `Heart disease`, slope) %>% summarize(percent = n()/nrow(heart))
k6<-ggplot(slope, aes(x=  `Heart disease`, y=percent, fill = slope))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")

Heart$ca <- as.factor(Heart$ca)
ca <- Heart %>% group_by( `Heart disease`, ca) %>% summarize(percent = n()/nrow(heart))
k7<-ggplot(ca, aes(x=  `Heart disease`, y=percent, fill = ca))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")

thal <- Heart %>% group_by( `Heart disease`, thal) %>% summarize(percent = n()/nrow(heart))
k8<-ggplot(thal, aes(x=  `Heart disease`, y=percent, fill = thal))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")

library(gridExtra)
grid.arrange(k1,k2,k3,k4)
grid.arrange(k5,k6,k7,k8)

Har <- heart %>% group_by(target)%>% summarize(percent = n()/ length(heart$target))
# Create a pie chart
ggplot(Har, aes(x = factor(target), y = percent)) +
  geom_bar(stat = "identity", fill = "red")

# Feature Engineering
sapply(heart, class)
heart$age <- scale(heart$age)
heart$trestbps <- scale(heart$trestbps)
heart$chol <- scale(heart$chol)
```

```
heart$thalach <- scale(heart$thalach)
heart$oldpeak <- scale(heart$oldpeak)
#heart$sex <- as.factor(heart$sex)
#heart$cp <- as.factor(heart$cp )
#heart$fbs <- as.factor(heart$fbs)
#heart$restecg <- as.factor(heart$restecg )
#heart$exang <- as.factor(heart$exang)
#heart$slope <- as.factor(heart$slope)
#heart$ca <- as.factor(heart$ca)
#heart$thal <- as.factor(heart$thal)


sample <- sample.split(heart$target, SplitRatio = 0.8)
Train <- subset(heart, sample==T)
Test <- subset(heart, sample==F)

length(Train$target)
length(Test$target)
length(heart$target)

ann1 <- neuralnet(target~., data = Train, hidden = c(9,6,3,2), err.fct = "ce", linear.output = F
                , act.fct = "logistic")

summary(ann1)


plot(ann1, col.entry = "red", col.hidden = "red", col.hidden.synapse = "red")


library(caret)
pr<-predict(ann1, Train%>%select(-target))
prT <- predict(ann1, Test%>%select(-target))
PTT <- predict(ann1, heart%>%select(-target))
p <- ifelse(pr>0.5,1,0)
pT <- ifelse(prT>0.5,1,0)
PTT <- ifelse(PTT>0.5,1,0)
k<-confusionMatrix(as.factor(Train$target), as.factor(p))
a<-confusionMatrix(as.factor(Test$target), as.factor(pT))
b<-confusionMatrix(as.factor(heart$target), as.factor(PTT))
k
a
b
l<-k$table
r<-a$table
d<-b$table
cm_df <- as.data.frame(l)
colnames(cm_df) <- c("Predicted", "Actual", "Count")

# Plot confusion matrix heatmap
g1<-ggplot(cm_df, aes(x = Predicted, y = Actual, fill = Count)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "orange", high = "red") +
  geom_text(aes(label = Count), vjust = 1) +
  labs(title = "Confusion Matrix Train data", x = "Predicted", y = "Actual") +
  theme_minimal()

cm_g <- as.data.frame(r)
colnames(cm_g) <- c("Predicted", "Actual", "Count")

# Plot confusion matrix heatmap
g2<-ggplot(cm_g, aes(x = Predicted, y = Actual, fill = Count)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "orange", high = "red") +
  geom_text(aes(label = Count), vjust = 1) +
  labs(title = "Confusion Matrix Test data", x = "Predicted", y = "Actual") +
  theme_minimal()

cm_d <- as.data.frame(d)
colnames(cm_d) <- c("Predicted", "Actual", "Count")

# Plot confusion matrix heatmap
g3<-ggplot(cm_d, aes(x = Predicted, y = Actual, fill = Count)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "orange", high = "red") +
  geom_text(aes(label = Count), vjust = 1) +
  labs(title = "Confusion Matrix for All data", x = "Predicted", y = "Actual") +
  theme_minimal()
library(gridExtra)
gridExtra::grid.arrange(g1,g2,g3)
```