

Developing a Novel Cardiac Disease Prediction Framework Utilizing Advanced Machine Learning Algorithms

Mukaila Olagunju
Department of Computer
Science,
Faculty of Science, Federal
University, Oye Ekiti, Nigeria.
mukaila.olagunju@fuoye.edu.ng

Abidemi Emmanuel Adeniyi
Department of Computer
Science, College of Computing
and Communication Studies,
Bowen University, Iwo, Nigeria.
abidemi.adeniyi@bowen.edu.ng

Odunayo Dauda Olanloye
Department of Computer
Science, College of Computing
and Communication Studies,
Bowen University, Iwo, Nigeria
odun.olanloye@bowen.edu.ng

Joseph Bamidele Awotunde
Department of Computer
Science, Faculty of
Communication and
Information Sciences,
University of Ilorin, Nigeria.
awotunde.jb@unilorin.edu.ng

ABSTRACT

Predicting and diagnosing cardiac disease has long been a crucial and difficult responsibility for medical professionals. Hospitals and other medical facilities provide pricey medicines and procedures to address cardiac ailments. Predicting cardiac disease in its early stages will thus be beneficial to the global population, allowing them to adopt preventative measures before the condition becomes serious. The study aims to revolutionize cardiac disease prediction and diagnosis through innovative machine learning methodologies. Addressing the challenge of early detection, which is crucial yet complex, the research seeks to implement a groundbreaking approach using advanced machine learning techniques. The novelty of this study lies in its use of two distinct machine learning algorithms - Logistic Regression and Random Forest - to analyze healthcare data. The obtained result shows that logistic regression model on the other hand had an accuracy of 80.48%, which is a fair performance, but still falls short of the random forest model's level of accuracy. This study will not only contribute to reducing mortality rates but also foster environments conducive to human development by enabling early intervention and effective treatment strategies. Data for this study is sourced from Kaggle, with Google Colab serving as the development platform, ensuring a robust and data-driven approach to cardiac healthcare.

Key words: Heart disease, Machine learning, Medical data, Algorithms, Diagnosis.

1. INTRODUCTION

Heart disease refers to a number of heart-related disorders, including infections, genetic anomalies, and blood-vessel ailments (Sobolewska-Nowak, 2023; Suleman et al. 2023). In this condition, the heart cannot pump the needed volume of blood to various regions of the body for normal functions, resulting in heart failure. It is crucial to diagnose cardiac illness as soon as possible so that counselling and medicine may be administered.

Machine Learning is becoming an important study subject in health care for providing prognoses and a better comprehension of medical data (Dash et al. 2019; Oladipo et al. 2023). In the past, several machine learning techniques were used to identify cardiac problems (Al Ahdal et al. 2023). For heart disease detection, a few common machine learning approaches are Random Forest and Logistic Regression. Diagnosis of cardiac diseases in an efficient and timely way is critical in healthcare, particularly in cardiology.

Cardiovascular disease is a huge concern in the modern day, with physical inactivity and bad food habits acting as its primary causes. Throughout the years, machine learning has demonstrated its usefulness in creating judgments and predictions based on the massive amount of data provided by health care organizations (Dash et al. 2019).

Heart disease, also called cardiovascular disease is a dangerous health condition, which if not detected early can lead to advanced health issues, and in very severe cases; death. With the increasing population of people with heart disease, a more effective method of assessing patient data has become vital. Machine learning plays a key role in achieving this, due to the fact that accessing medical services is becoming more of a challenge than it should be, machine learning presents a faster and more efficient way of handling patient data and producing highly accurate results.

In previous models developed by other researchers, machine learning models were developed but were occasionally inaccurate. This is because the data that was used in the library was compiled from big data sets and was mostly derived from large libraries, indicating that the data was prone to being erroneous or incomplete. Using approaches for dimensionality reduction, this may be avoided. Researchers often used complex combined algorithms (Adeniyi et al. 2022), which were not too accurate, due to conflicting results from similar datasets.

This study makes a meaningful contribution in the field of heart disease detection by addressing the limitations of existing machine-learning models. Previous models, while utilizing extensive datasets, often faced challenges with accuracy due to erroneous or incomplete data. This research seeks to enhance accuracy by employing dimensionality reduction techniques, avoiding the pitfalls of complex, combined algorithms that have produced conflicting results in the past. As heart disease prevalence rises, this study's approach offers a more effective and efficient method for patient data assessment, leveraging machine learning for faster, more accurate analysis, thereby improving early detection and treatment outcomes.

8. REVIEW OF RELATED WORKS

Table 1. Summary review of literature review

Author	Title	Method	Result	Limitation
Choudhary and Singh (2020)	Prediction of Heart Disease Using Machine Learning Algorithms	Decision Tree Classification Algorithm, Naïve Bayes Classifier	-	Research was limited to few algorithms

Ed-Daoudy and Maalmi (2019)	Real-time machine learning for early detection of heart disease using big data approach	Random Forest Algorithm, Real time data processing	-		The quality and amplitude of the signal in the dataset from which a perfect model is learned restrict a perfect model.
Martin-Isla et al. (2020)	Image-Based Cardiac Diagnosis With Machine Learning: A Review	Logistic Regression, Support Vector Machine (SVM), Random Forest		To demonstrate the validity of ML applied to cardiac imaging, the findings were studied from two perspectives: statistical validity and accuracy of the acquired statistical values.	Although intriguing, the applicability of such methods to cardiovascular applications remains challenging.
Nikhar and Karandikar (2016)	Prediction of Heart Disease Using Machine Learning Algorithms	Naive Bayes Classification Algorithm, Decision Tree		Although most studies are using various classifier approaches in the diagnosis of heart illness, using Naïve Bayes and Decision tree with information gain calculations produces superior results in the diagnosis of heart disease and greater accuracy as compared to other classifiers.	Due to the inherent linearity of the input set, standard medical scoring systems are unable to accurately describe the nonlinear complex interactions that occur in medical domains.
Bharti et al. (2021)	Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning	Machine Learning, Deep Learning		Deep learning a new field of artificial intelligence demonstrated promising outcomes in medical diagnosis with high accuracy	A typical issue is the high dimensionality of the data; the datasets utilized include enormous amounts of information, which cannot always be examined.
Sharma and Rizvi (2017)	Prediction of Heart Disease using Machine Learning Algorithms: A Survey	SUPPORT VECTOR MACHINE (SVM), K-NEAREST NEIGHBOUR ALGORITHM (KNN), DECISION TREE		It has been shown that deep learning, an emerging field of artificial intelligence, can accurately diagnose diseases.	The computer's processing capability and the study's time restriction were two of the tools employed in this research.

Khourdifi and Baha (2019)	Particle Swarm and Optimization of Ant Colonies are used to optimize algorithms for machine learning for coronary artery disease forecasting and categorization.	Particle Swarm Optimization, Ant Colony Optimization	In terms of coronary artery disease predictions and categorization, our optimized model beats previous models based on existing methodology and experiment results.	The instruments used in this examination, such as the computer's processing capabilities and the time limit for the study, are mentioned.
Learning (2017)	A review of using data mining and machine learning methods to diagnose and predict coronary artery disease.	Naive Bayes Algorithm, Decision Tree, K-Nearest Neighbour Algorithm	The findings demonstrate that the hybrid system of genetic algorithm and neural network performs significantly better than the performance of neural network alone	The lower values that result, which are near to zero, show the generalized format of the network, which is ready to solve the classification issue.
Arabasadi et al. (2017)	Computer-assisted making decisions for heart disease diagnosis utilizing a hybrid neural network-genetic approach.	Neural Network, Genetic Algorithm	The suggested technique has a significantly superior performance compares to Neural Network.	Probabilistic or fuzzy models are inefficient.
Marimuthu et al. (2018)	A Review of Cardiovascular Disease Prediction Through Learning Machines and Data Analytics.	Artificial Neural Network (ANN), Decision Tree, Fuzzy Logic, K-Nearest Neighbor (KNN), naive Bayes, and Support Vector Machine. (SVM).	This study's findings suggest that SVMs and neural networks are excellent tools for predicting cardiac disease.	Because of the fast development of digital technology, healthcare facilities now store enormous amounts of data that are very complicated and difficult to analyze.

9. MATERIAL AND METHODS

This study methodology involves creating a model capable of forecasting the occurrence of coronary artery disease using a set of features (risk factors) defining the condition. This research presents an overview of a variety of machine learning algorithms, which are Logistic Regression and Random Forest, which may assist practitioners or medical analysts in properly diagnosing heart problems. The suggested technique comprises the following steps:

Data Collection: The model's input is a dataset taken from the Kaggle repository. Collecting data for training the machine learning model is the initial stage in the machine learning process. Predictions made by machine learning systems are only as accurate as the datasets used to train them. The dataset contains the 14 attributes including age, sex, chest pain, cholesterol level, fasting blood sugar, electrocardiographic result, old peak, slope, heart rate and chest pain.

Data Pre-processing: The data pre-processing phase is the phase in which data is processed or encoded so that it may be swiftly parsed by a computer. Algorithms learn from their inputs, and their ability to solve a

problem relies heavily on the data they are fed, therefore pre-processing data is required before any machine learning gear can be used to tackle the issue at hand.

This is achieved by:

- i. Randomizing all accessible information. This guarantees that data is distributed evenly and that the order has no effect on the learning process.
- ii. Remove duplicate values, improper data types, and mistakes from data.
- iii. Visualizing data helps understand its structure and linkages.
- iv. The data has been divided into training and testing sets and is ready for analysis. It is the training set on which you will build your model. After training, you may use a testing set to see if your model is correct.

Training and Testing: Because it is already known if a patient has heart disease, we utilize current data to train our prediction algorithm. In other words, it's called monitoring and learning. Heart disease may be predicted using the trained model. The trained model is then employed to predict whether or not a user has heart disease. The test set is used by the ML model to predict outcomes.

Evaluating The Model: The model is evaluated using the datasets to determine the classification accuracy. This is accomplished by evaluating the model's performance using previously unknown data. The unseen data utilized is the testing set that you divided our data into before. A Logistic Regression Classifier and a Random Forest Classifier are then utilized to categorize the dataset's features.

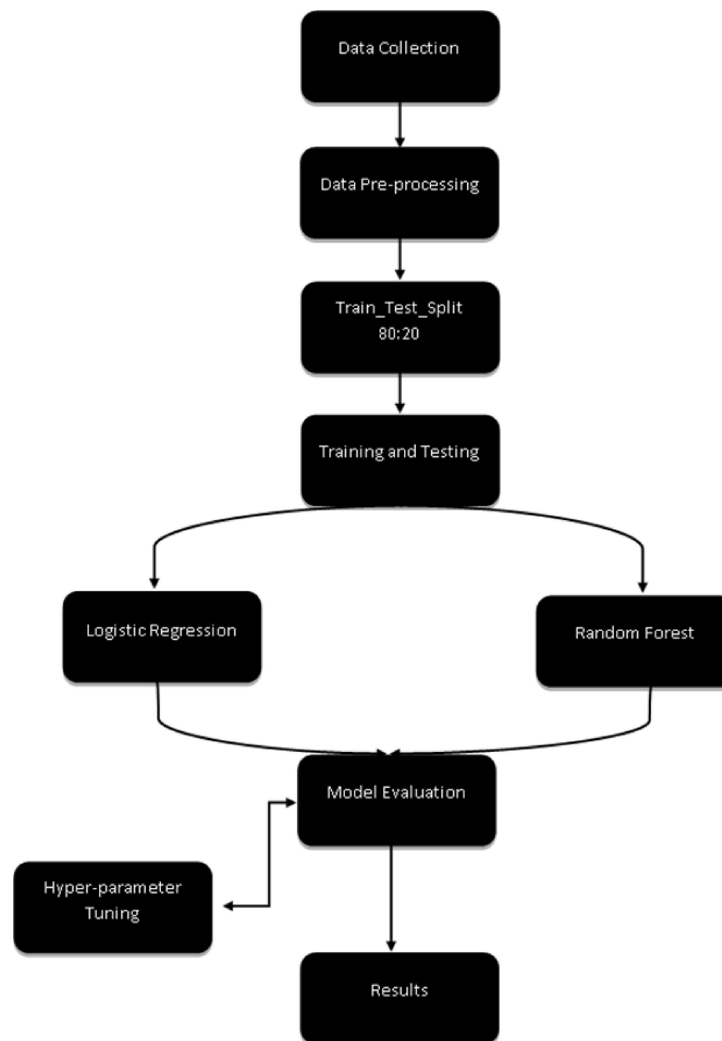
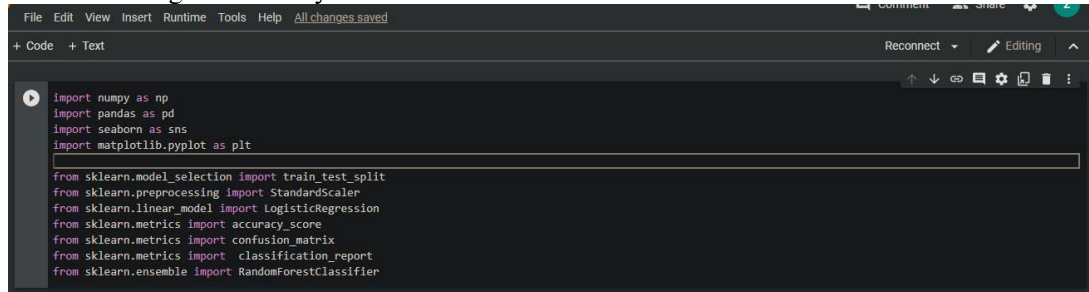


Fig 1: Methodology used for the process flow.

10. RESULTS AND DISCUSSION

Various python libraries were imported. They all have various functions, some of which are;

- i. Creating the machine learning model
- ii. Training the model
- iii. Visualization: this is the plotting of various data/information into visual graphs or maps
- iv. Getting the accuracy of the results



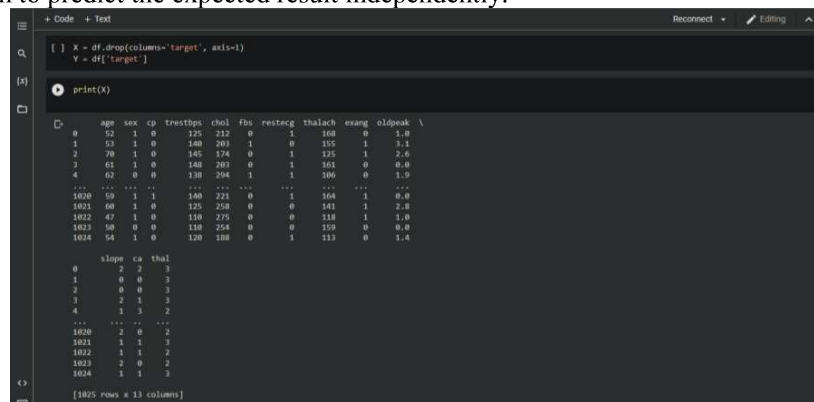
```
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.ensemble import RandomForestClassifier
```

Fig 2: Importing libraries

4.1 Defining the target

This implies removing a the predetermined target (result) from the dataset, thereby allowing the machine learning algorithm to predict the expected result independently.



```
+ Code + Text
[] X = df.drop(columns='target', axis=1)
Y = df['target']

print(X)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	\
0	54	1	0	125	212	0	1	168	0	1.6	
1	53	1	0	140	203	1	0	155	1	3.1	
2	70	1	0	145	178	0	1	125	1	2.6	
3	61	1	0	140	203	0	1	161	0	0.0	
4	62	0	0	138	294	1	1	106	0	1.9	
...
1020	59	1	1	140	221	0	1	164	1	0.0	
1021	60	1	0	125	258	0	0	141	1	2.8	
1022	42	1	0	110	226	0	0	118	1	1.0	
1023	50	0	0	110	254	0	0	159	0	0.0	
1024	54	1	0	120	188	0	1	113	0	1.4	
...
1020	2	0	2								
1021	1	1	3								
1022	1	1	2								
1023	2	0	2								
1024	1	1	3								

```
[1025 rows x 13 columns]
```

Fig 3: Defining the target variable.

4.2 Data Visualization

Visualization of the data based on various metrics.

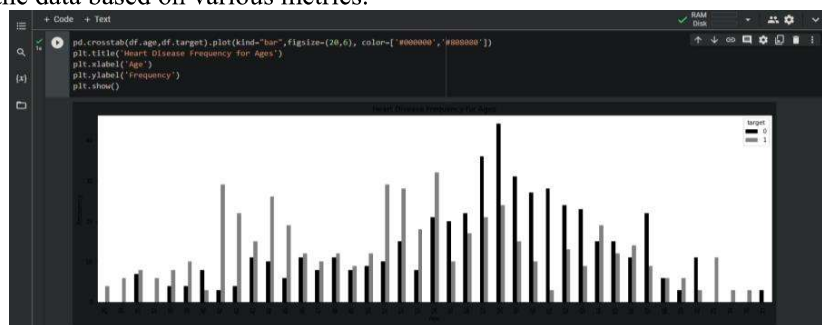


Fig 4: Heart Disease Frequency based on Age



Fig 5: Heart Disease frequency based on gender.

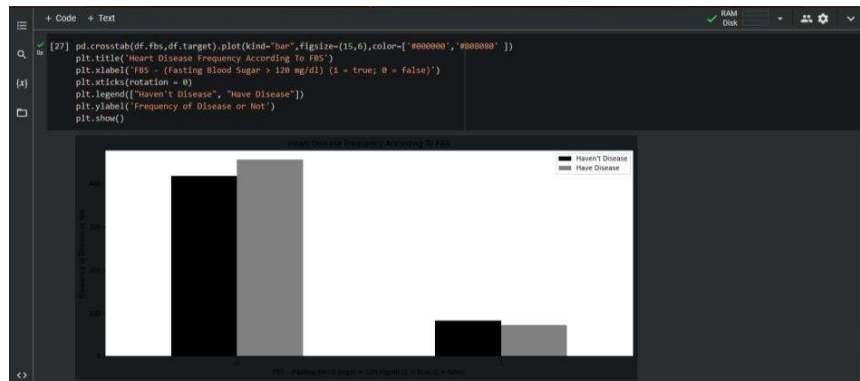


Fig 6: Heart Disease frequency based on Fasting Blood Sugar (FBS)



Fig 7: Heart Disease Frequency based on chest pain

4.3 Correlation and Correlation Heatmap

The correlation between two variables is a statistical measure of their relationship. It is most effective when applied to variables with a linear relationship. The data correlation is shown in Figure 8, and its visualization is depicted in Figure 9.

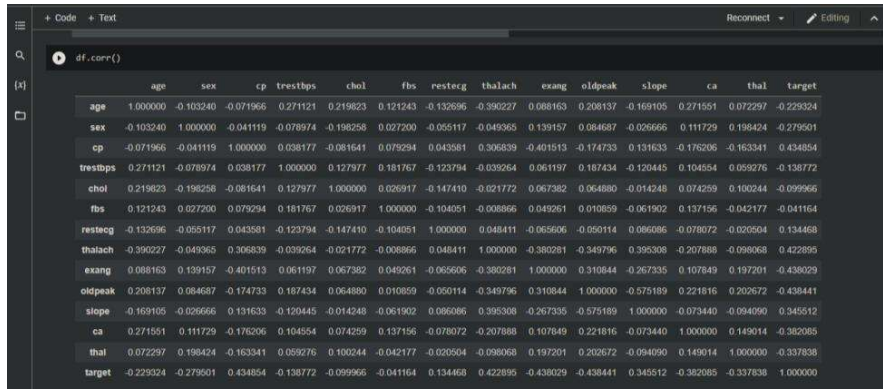


Fig 8: Dataset Correlation

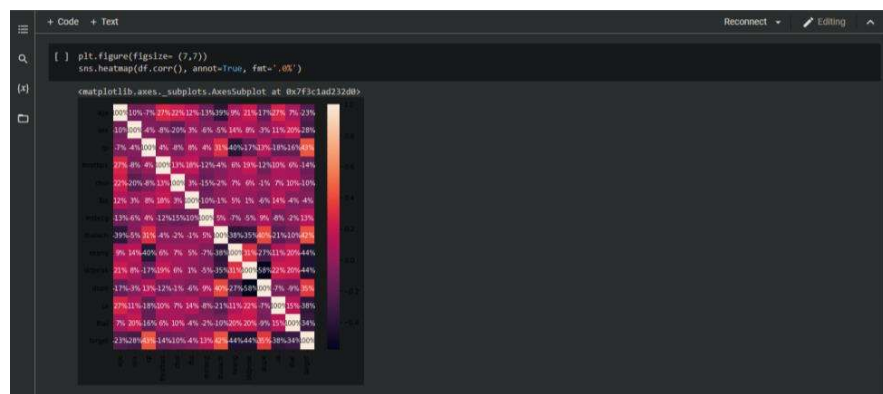


Fig 9: Correlation Heat Map

4.4 Logistic Regression Results

The data is separated into training and test data of 80% and 20% respectively. They are then tested for accuracy, and they give respective accuracies of 85.24% and 80.48%. Figure 10 shows how the assessment report (table 2) was used to assess the accuracy of the categorization technique's forecasts based on true positives, false positives, true negatives, and false negatives.

Table 2. Classification report of the algorithm

	Precision	Recall	F1-score	Support
0	0.72	0.86	0.7	84
1	0.89	0.77	0.82	121
Accuracy			0.80	205
Macro avg	0.80	0.81	0.80	205
Weighted avg	0.82	0.80	0.81	205

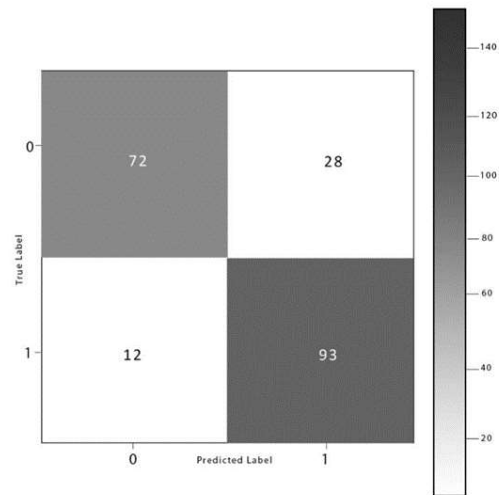


Fig 10: Logistic Regression Confusion Matrix

4.5 Random Forest Results

The data is separated into training and test data of 80% and 20% respectively. They are then tested for accuracy, and they give respective accuracies of 85.24% and 80.48%. Figure 11 shows how the classification report was used to assess the accuracy of a classification algorithm's predictions based on true positives, false positives, true negatives, and false negatives.

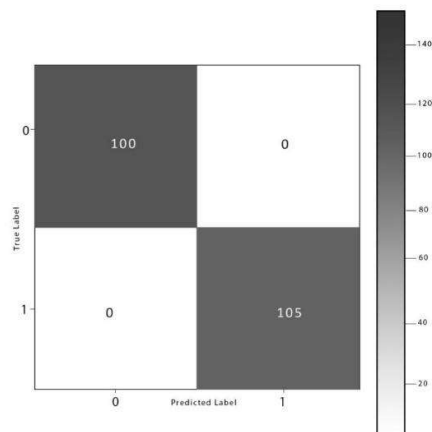


Fig 11: Random Forest Confusion Matrix

4.6 Evaluation Metrics

The efficiency of the model was evaluated using the metrics in the table below

Table 3. Matrices evaluation table

Performance Metrics	Logistic Regression(%) *100	Random Forest (%) *100	Formula
Accuracy	0.8048	1.00	$(TP+TN)/(TN+FP+FN+TN)$
Specificity	0.8571	1.00	$TN/(TN+FP)$
F1-Score	0.8416	1.00	$2TP/(2TP+FP+FN)$

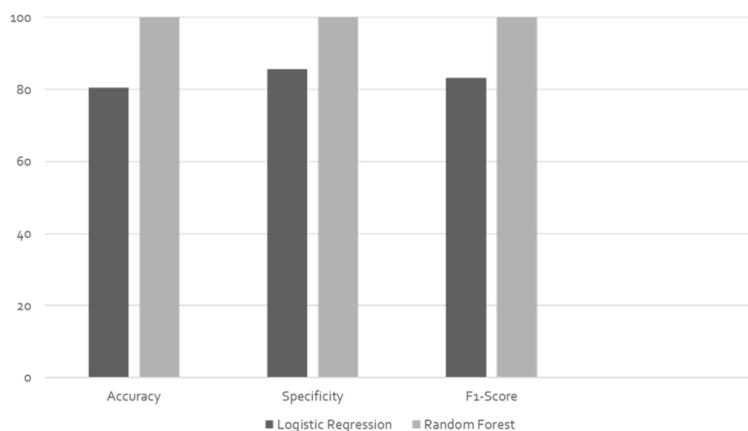


Fig 12: Comparison of the Results

The purpose of this study is to explore how heart diseases can be detected and prevented. The goal of machine learning is to discover patterns in a dataset and then make predictions based on often complex patterns to answer questions and further solve complex problems. Due to the fact that an attempt is being made to tackle the issue of heart disease, this article corresponds well with the machine's purpose. This study's dataset was collected from the Kaggle repository. Heart disease is the most prevalent cause of death in the globe. Early identification and treatment alleviate suffering and prevent complications such as heart failure and stroke.

11. CONCLUSION

The machine learning algorithms have been trained and tested on a heart disease dataset, and proved to have a high prediction accuracy. Both algorithms performed relatively well, but the random forest model significantly outperformed the logistic regression model. The Random Forest model when used with this data set was essentially perfect, as it had an accuracy of 100%, which also means that it scored a 100% as well across all other evaluation platforms. The logistic regression model on the other hand had an accuracy of 80.48%, which is a fair performance, but still falls short of the random forest model's level of accuracy. Improvements can be made on this model, since when the random forest model was evaluated on a different dataset (which was smaller than the original), it gave an accuracy of 90%. To enhance this study, it could be suggested that future research accommodate more datasets for a higher chance of an increase in the accuracy of the model.

12. REFERENCES

- Sobolewska-Nowak, J., Wachowska, K., Nowak, A., Orzechowska, A., Szulc, A., Plaza, O., & Gałeczki, P. (2023). Exploring the heart–mind connection: unraveling the shared pathways between depression and cardiovascular diseases. *Biomedicines*, *11*(7), 1903.
- Suleman, M., Crovella, S., Hussain, T., Khan, M. U., Hassan, S. S., Majid, M., ... & Ahmad, Z. (2023). Cardiovascular challenges in the era of antiretroviral therapy for AIDS/HIV: A comprehensive review of research advancements, pathophysiological insights, and future directions. *Current Problems in Cardiology*, *10*(23), 102353.
- Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of big data*, *6*(1), 1-25.
- Oladipo, I. D., AbdulRaheem, M., Awotunde, J. B., Bhoi, A. K., Adeniyi, E. A., & Abiodun, M. K. (2021). Machine learning and deep learning algorithms for smart cities: a start-of-the-art review. *IoT and IoE driven smart cities*, 143-162.
- Al Ahdal, A., Rakhra, M., Rajendran, R. R., Arslan, F., Khder, M. A., Patel, B., ... & Jain, R. (2023). Monitoring Cardiovascular Problems in Heart Patients Using Machine Learning. *Journal of Healthcare Engineering*, *2023*.
- Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of big data*, *6*(1), 1-25.
- Adeniyi, J. K., Adeniyi, A. E., Oguns, Y. J., Egbedokun, G. O., Ajagbe, K. D., Obuzor, P. C., & Ajagbe, S. A. (2022). Comparison of the performance of machine learning techniques in the prediction of employee. *ParadigmPlus*, *3*(3), 1-15.
- Choudhary, G., & Singh, S. N. (2020, October). Prediction of heart disease using machine learning algorithms. In *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)* (pp. 197-202). IEEE.
- Ed-Daoudy, A., & Maalmi, K. (2019, April). Real-time machine learning for early detection of heart disease using big data approach. In *2019 international conference on wireless technologies, embedded and intelligent systems (WITS)* (pp. 1-5). IEEE.
- Martin-Isla, C., Campello, V. M., Izquierdo, C., Raisi-Estabragh, Z., Baeßler, B., Petersen, S. E., & Lekadir, K. (2020). Image-based cardiac diagnosis with machine learning: a review. *Frontiers in cardiovascular medicine*, *7*, 1.
- Nikhar, S., & Karandikar, A. M. (2016). Prediction of heart disease using machine learning algorithms. *International Journal of Advanced Engineering, Management and Science*, *2*(6), 239484.
- Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Computational intelligence and neuroscience*, *2021*.
- Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), 99-104.
- Khourdifi, Y., & Baha, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International journal of Intelligent engineering & systems*, *12*(1).
- Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. *Advances in Computational Sciences and Technology*, *10*(7), 2137-2159.

Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Computer methods and programs in biomedicine*, 141, 19-26.

Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications*, 181(18), 20-25.

Author's Brief Profile



Abidemi Emmanuel Adeniyi received the B.Sc. and M.Sc. degrees in computer science from Ladoke Akintola University of Technology, Ogbomoso, and Landmark University, Omu-Aran, Nigeria, in 2019 and 2021, respectively, and he is currently a research student at the department of Computer Science, University of Ilorin and a lecturer in the department of Computer Science, Bowen University, Iwo, Nigeria. He has authored or coauthored more than 50 journal and conference papers, and 50 book chapters with Elsevier and Springer. His research interests include the Internet of Things Security, Information security, applications of artificial intelligence, application of Machine Learning for intrusion detection, Internet of Medical Things, and Cryptographic algorithms. He can be contacted at email: abidemi.adeniyi@bowen.edu.ng

