Big Data analytics for Predicting Heart Disease: A Machine Learning Approach

Israel Morakinyo  
*University of Wolverhampton*   
I.Morakinyo@wlv.ac.uk

*Abstract*— Big Data in healthcare facilitates the use of advanced analytics to enhance patient outcomes and disease prediction. Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, making up for roughly 20.5 million deaths in 2021 (World Heart Federation, 2023). However, traditional diagnostic methods are time consuming and require professional assessment. I aim to apply machine learning techniques to predict heart disease risk based on patient health indicators. The dataset, sourced from Kaggle, includes medical attributes such as age, cholesterol levels, blood pressure, and lifestyle factors. Data preprocessing steps involve handling missing values, feature selection, and normalisation. A Logistic Regression model was trained and evaluated using accuracy, precision, and recall metrics. The results demonstrate that advanced models can provide reliable predictions, aiding healthcare professionals in early diagnosis and preventive care.

# Background of the study

Cardiovascular diseases (CVDs), including coronary heart disease (CAD), remain a leading cause of mortality worldwide, mostly because of atherosclerosis, where plaque buildup narrows arteries and restricts blood flow (Mayo Clinic, 2024). Among the main risk factors that can be changed are hypertension, which strains arteries, high cholesterol, which accelerates plaque formation, smoking, which damages blood vessels, and diabetes which increases vascular complications (CDC, 2024). Obesity, physical inactivity, unhealthy diets and alcohol abuse further heighten CVD risk (World Health Organization, 2021). Non-modifiable risk factors such as age, gender, family history and ethnicity also contribute to susceptibility (World Health Organization, 2021). Considering the impact of cardiovascular disease, early detection and management are essential through lifestyle modifications, routine screenings and pharmacological interventions for reducing mortality and improving patient outcomes (World Health Organization, 2021).

In the UK, heart disease is the second leading cause of death, accounting for approximately 170,000 deaths per year (British Heart Foundation, 2022). One person dies from heart or circulatory disease every three minutes. Despite medical advancements, numerous cases remain undiagnosed until late stages, largely due to dependence on traditional risk assessment methods. The elevated expenses associated with diagnostic procedures, including coronary angiography, result in accessibility challenges, which contribute to delays in detection and treatment.

A close-up of a poster

AI-generated content may be incorrect.

Figure 1.0 (World Heart Federation, 2021)

The aim of this study is to analyse a heart disease dataset to uncover key patterns, trends and risk factors associated with cardiovascular diseases. In this study, I will conduct exploratory data analysis (EDA) to identify relationships among variables including age, cholesterol levels, blood pressure, and lifestyle factors, evaluating their influence on heart disease risk. This study makes several contributions to big data analysis in healthcare by providing valuable insights into heart disease risk factors. Key contributions include

* Data Cleaning and Preprocessing – The dataset goes through data cleaning, handling missing values and normalising medical attributes to ensure reliable analysis.
* Exploratory Data Analysis (EDA) - This study examines relationships between key variables such as age, cholesterol levels, blood pressure etc, uncovering patterns that contribute to cardiovascular risk.
* Data Visualisation for Medical Insights – This study provides statistical visualisations (like histograms, box plots, correlation heatmaps) to help interpret trends and risk factors in a clear and informative manner.
* Healthcare Impact – The findings can assist healthcare professionals in identifying high risk individuals, enabling early intervention and preventive care strategies.
* Machine Learning for Risk Prediction – A simple classification model is explored to assess its potential in identifying patients at risk of heart disease.

This report presents a detailed analysis of heart disease through the application of big data techniques. Following this Background of the Study section, the Related Work section discusses existing research on heart disease prediction and data-driven healthcare analysis. The Methodology section details the data preprocessing steps, exploratory data analysis (EDA) techniques, and visualisation methods used in this study. The Results and Discussion section presents key findings, supported by statistical insights and graphical representations of the dataset. Finally, the Conclusion summarises the study’s key contributions, discusses the implications of the findings, and suggests potential areas for future research.

# Related Work

Many studies have examined the use of statistical and machine learning models in heart disease prediction. Statistical models such as Logistic Regression offers clear decision rules, and interpretability, making it particularly advantageous in clinical settings where understanding the rationale behind predictions is essential (Swetha et al., 2024). These techniques, however, frequently have trouble handling intricate, non-linear relationships in large datasets. For example, Decision Trees offer interpretability but are prone to overfitting (Chang, 2024).

On the other hand, machine learning algorithms like Support Vector Machines (SVM) and Random Forests have shown increased accuracy in disease prediction by identifying complex patterns in patient data. A study found that SVM achieved an accuracy of 83.81% in predicting heart disease, highlighting its effectiveness in pattern recognition (Chen, 2024). However, Machine learning models however do require high quality data and can be difficult to interpret which may hinder their practical application in clinical settings (Chang, 2024). Unlike previous studies that relies on traditional predictive models, this study focuses on an extensive exploratory data analysis (EDA) approach to uncover key trends and patterns before applying predictive analytics. This approach ensures a data driven understanding of heart disease risk factors which can provide deeper insights before employing machine learning models for classification.

# Methodology

In this section, I will outline the step-by-step process when analysing the heart dataset. The methodology is divided into five main phases which are data acquisition, data preprocessing, explanatory data analysis, data visualisation and predictive modelling.

## Data Acquisition

The dataset I used in this study was sourced from Kaggle which was loaded using the pandas library. The dataset includes patients characteristics like age, sex, chest pain type and so on and as well as a target variable showing the presence or absence of heart disease.

## Data Preprocessing

The data preprocessing involved several steps to guarantee the dataset’s quality and prepare it for analysis. I checked for missing values using .isnull().sum() and no missing data was found. I also checked for duplicate entries using .duplicated().sum() in order to maintain the integrity of the data.

Before model training, categorical features including sex, ChestPainType, RestingEcg, ExerciseAngina and ST\_Slope were encoded numerically using LabelEncoder. Additionally, the numerical features were standardised using StandardScaler to ensure all features contributed equally to model performance. These preprocessing steps were applied after the exploratory data analysis phase and immediately prior to building the model.

## Exploratory Data Analysis (EDA)

EDA was performed to gain insight into the characteristics of the data. Descriptive statistics were generated using .describe(include=”all”) to summarise the feature distributions and highlight any anomalies. A correlation heat map was produced to visualise the relationships between numerical variables helping to identify features that may influence heart disease risk. I also used boxplots to detect the presence of outliers in variables such as cholesterol. This phase established the groundwork for understanding underlying patterns before predictive modelling.

## D. Data Visualisation

I created various visualisations to explore distributions and relationships. Count plots and pie charts were used for categorical features such as sex, ChestPainType etc to show their distributions within the dataset. I used histograms and kde plots for numerical features like age, MaxHR etc to visualise their distribution. Bivariate analysis was conducted using count plots with a HeartDisease hue to observe how categorical variables vary between patients with and without heart disease. Boxplots and histograms comparing numerical features by disease status were also included to identify important trends.

## Predictive Modelling

The features of the dataset were split into independent variables, X, and the target variable, Y, which is HeartDisease. I trained a basic logistic regression model to evaluate the predictive power of the dataset, the data was split into training and test set in an 80 to 20 ratio using train\_test\_split(). After scaling and encoding, the logistic regression was trained on the training data and predictions were made on the testing data. I assessed the performance using the accuracy score, confusion matrix and classification report in order to provide detailed metrics on precision, recall and F1-score on each class thereby providing a thorough assessment of the model’s predictive capability.

# Results and Discussion

## Experimental Setup

The heart disease dataset obtained from Kaggle was used for the experimental analysis. Before the modelling, categorical features were label-encoded while numerical features were standardised using StandardScaler to ensure consistent scaling. The data was then split into training and test set in an 80 to 20 ratio using train\_test\_split(). I then trained a logistic regression model on the scaled training data with default hyperparameters from scikit-learn.

## Discussions and Analysis of Findings

I will be delving into each column of the dataset to understand the relationship between each feature and heart disease.

Starting off with the age column, Patients diagnosed with heart disease generally tend to be older compared to those without the disease. The distribution shows a higher concentration of cases above 50 years of age which is consistent with clinical knowledge that age is a major risk factor for cardiovascular disease.

A graph of age vs heart disease

AI-generated content may be incorrect.A blue square with black text

AI-generated content may be incorrect.

Regarding the next column, Sex, the dataset revealed a greater incidence of heart disease in male patients compared to female patients pointing to a possible risk factor associated with gender.

A graph with blue and orange bars

AI-generated content may be incorrect.

However, it is important to note that there are more male patients than female patients in the dataset.

A blue and orange pie chart

AI-generated content may be incorrect.

For the chest pain type column, there are 4 different types present in the dataset

A screenshot of a computer

AI-generated content may be incorrect.

ASY (Asymptomatic) is the most common type which shows that most patients have heart problems without obvious symptoms. NAP (Non-Anginal Pain) and ATA (Atypical Angina) are less common meaning that some patients have pain that is not associated with angina. TA (Typical Angina) is the least frequent but is indicative of more overt heart disease.

A graph of different colored squares

AI-generated content may be incorrect.

Analysing this count plot, the majority of ASY patients had heart disease despite the unclear symptoms.

Next the RestingBP (resting blood pressure), higher resting blood pressure values were observed among patients with heart disease. Elevated blood pressure is a well-established risk factor that can cause arterial damage leading to cardiovascular complications.

A graph of a graph of different colored lines

AI-generated content may be incorrect.A diagram with a blue rectangle

AI-generated content may be incorrect.

Even though most of the values are high, according to National Health Service (2023), normal blood pressure is considered to be between 90/60mmHg and 120/80mmHg, however from the box plot we can see there are a bunch of outliers, this might affect predictive modelling and even analysis. For example, a value of 0mmHg can’t be possible because that means there is no blood pressure, so that must be dropped.

When analysing the cholesterol feature, a significant issue comes up, it is impossible to have a cholesterol level of zero but when looking at the cholesterol distribution, a large number of patients had a cholesterol level of zero.

A graph of different colored lines

AI-generated content may be incorrect.A diagram of a number of objects

AI-generated content may be incorrect.

There are also a lot of outliers that can be seen in the box plot, this also will impact the data quality and predictive modelling. To be exact, there are 171 rows that have a cholesterol value of 0 suggesting a potential data entry issue, I can’t just drop these rows so further strategies are needed.

I identified cholesterol values of zero and replaced them with NaN and then imputed them using the median value of non-zero cholesterol values to maintain data quality without significantly reducing dataset size. The new distribution and box plots are shown below:

A graph of a normal heart disease

AI-generated content may be incorrect.A diagram of a number of cholesterol levels

AI-generated content may be incorrect.

The next feature is Fasting Blood Sugar (FastingBS) which indicates whether a patient’s FastingBS exceeds 120 mg/dl. In this dataset, only 23.3% of patients had elevated FastingBS, yet among these individuals, a noticeably higher proportion were diagnosed with heart disease.

A graph of different colored bars and a pie chart

AI-generated content may be incorrect.

This suggests that while abnormal FastingBS is less prevalent, it is strongly associated with heart disease presence. It is also consistent with existing clinical knowledge that diabetes and elevated blood sugar are associated with a higher risk of cardiovascular disease.

The RestingECG feature distribution is as follows

A graph of different colored bars

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

The next feature is maximum heart rate achieved (MaxHR), patients with heart disease tended to have lower MaxHR during exercise compared to healthy individuals. This suggests reduced cardiovascular fitness and exercise tolerance among those affected.

A graph of different colored lines

AI-generated content may be incorrect.

If a patient had ExerciseAngina (Y), it indicates a higher risk of heart disease while those with no ExerciseAngina (N) have a lower risk.

A graph of different colored bars

AI-generated content may be incorrect.

The next feature, Oldpeak represents ST depression during exercise relative to rest. When analysing this particular feature, I realised it had negative values which is not possible physiologically. This means there was a data entry issue or it was mislabelled.

A graph of different colored lines

AI-generated content may be incorrect.

So, to fix this issue, I treated the negative values as missing and imputed them using the median of valid values to ensure data quality. Oldpeak new distribution now looks like

A graph of different colored lines and numbers

AI-generated content may be incorrect.

The last feature is ST\_Slope which is the slope of the peak exercise ST segment, it can either be up, flat or down. Majority of the values were flat and were also diagnosed with heart disease. Up which indicates normal blood flow is the opposite, having less proportion of patients with heart disease.

A graph of different colored bars

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

Finally, the target column, HeartDisease, the distribution is below

A graph with numbers and a pie chart

AI-generated content may be incorrect.

Out of the total datset, 507 individuals (55.3%) were diagnosed with heart disease while 410 individuals (44.7%) were not. This slight class imbalance indicates that more than half of the dataset have heart disease which justifies the importance of predictive modelling for early detection. The distribution is relatively balanced which is good for classification models as it reduces the risk of bias towards the majority class.

Having explored key trends in the data and observed meaningful relationships between variables and heart disease, the next phase involved developing a predictive model. Logistic regression was chosen to assess how well the available features could classify heart disease. The results below reflect the model’s performance on unseen test data:

The logistic regression model achieved an accuracy of 84% on the test set which shows a strong performance for a baseline model. I generated a confusion matrix to show the number of correctly and incorrectly classified instances.

A diagram of a confusion matrix

AI-generated content may be incorrect.

The matrix shows that 61 true negatives (patients without heart disease) and 94 true positives (patients with heart disease) were correctly identified. However, the model misclassified 11 patients without heart disease (false positives) and 18 patients with heart disease (false negatives).

The classification report elaborated on the model’s performance metrics.

A number of numbers in a row

AI-generated content may be incorrect.

The precision for class 1 (patients with heart disease) was 0.90 while the recall was 0.84, producing an F1-score of 0.87. These results show that the model is good at correctly identifying individuals with heart disease although a moderate number of cases were missed. The high precision signifies that when the model predicts heart disease, it is very probable to be accurate, hence reducing false alarms.

# Conclusion

This report applied data analytics and machine learning techniques to explore and predict heart disease risk using a publicly available dataset. After thorough preprocessing and exploratory analysis, I identified key risk factors such as age, cholesterol and exercise-induced angina. A logistic regression model trained on this data achieved 84% accuracy showing that even simple models can effectively support early detection when combined with proper data handling. These results highlight the value of data driven approaches in healthcare.

For future work, more advanced models and richer feature engineering could further boost predictive performance. This report demonstrates how big data workflows, from cleaning through visualisation to modelling can yield meaningful insights in healthcare and support early intervention methods for cardiovascular disease.

# References

British Heart Foundation (2022). *Facts and Figures*. [online] Bhf.org.uk. Available at: <https://www.bhf.org.uk/what-we-do/news-from-the-bhf/contact-the-press-office/facts-and-figures>.

CDC (2024). *Heart Disease Risk Factors*. [online] Heart Disease. Available at: <https://www.cdc.gov/heart-disease/risk-factors/index.html>.

Chang, X. (2024). Comparative Analysis of Machine Learning, Decision Trees, and K-Nearest Neighbors for Heart Disease Prediction. *Applied and Computational Engineering*, 82(1), pp.188–192. doi:https://doi.org/10.54254/2755-2721/82/20241186.

Chen, Y. (2024). Predicting Heart Disease Using Machine Learning: Analysis and New Insights. *Science and Technology of Engineering Chemistry and Environmental Protection*, 1(10). doi:https://doi.org/10.61173/ygj64y98.

Mayo Clinic (2024). *Heart Disease*. [online] Mayo Clinic. Available at: <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>.

National Health Service (2023). *Blood Pressure Test*. [online] NHS. Available at: <https://www.nhs.uk/conditions/blood-pressure-test/>.

Swetha R, Joy, H.K., Yassin, W., Sridevi. R, N, D.M. and Pohrmen, F.H. (2024). Predicting Heart Disease with Machine Learning: A Comparative Study. [online] pp.815–822. doi:https://doi.org/10.1109/icicnis64247.2024.10823237.

World Health Organization (2021). *Cardiovascular Diseases (CVDs)*. [online] World Health Organization. Available at: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>.

World Heart Federation (2021). *Cardiovascular Disease Infographic*. [online] World Heart Federation. Available at: https://world-heart-federation.org/resource/cardiovascular-disease-infographic/.

World Heart Federation (2023). *Deaths from cardiovascular disease surged 60% globally over the last 30 years: Report*. [online] World Heart Federation. Available at: https://world-heart-federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/.