Impact of Data Pre-processing and Missing-Value Handling on Machine Learning-Based Heart Disease Prediction

Rebecca Nneka Oguzor[1], Dr. Chinagolum Ituma[2], Favour Ngozi Okoro[3]

[1, 2, 3] Department of Computer Science, Ebonyi State University, Abakaliki, Nigeria

ABSTRACT

The reliability of machine learning-based heart disease prediction depends not only on the choice of predictive models but also on the quality and preparation of the underlying clinical data. Missing values, inconsistent variable representations, and differences in feature scales can affect the reliability of predictive models and their ability to distinguish between patients with and without heart disease. This study investigated the impact of systematic data pre-processing and missing-value handling on machine learning-based heart disease prediction. The study employed a hybrid methodological framework combining the Cross-Industry Standard Process for Data Mining (CRISP-DM) and Object-Oriented Methodology (OOM). CRISP-DM guided the data-driven process from data understanding and preparation to modelling and evaluation, while OOM supported the modular design and implementation of the system components. The study utilized the Cleveland Heart Disease dataset obtained from the University of California Irvine (UCI) Machine Learning Repository. A structured pre-processing pipeline involving data quality assessment, missing-value treatment using K-Nearest Neighbours (KNN) imputation, categorical variable encoding, and feature normalization using Standard Scalar was applied to prepare the dataset for predictive modelling. Logistic Regression, Random Forest, and Boost models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The ROC-AUC results showed that Logistic Regression achieved the highest discriminative performance with an AUC of 0.591, followed by Boost with 0.566, while Random Forest recorded the lowest AUC of 0.514, indicating performance close to random classification. These findings demonstrate that classification accuracy alone may not adequately reflect a model's discriminative capability and highlight the importance of using multiple evaluation metrics in clinical prediction. The study provides a structured and reproducible pre-processing and modelling framework for preparing clinical heart disease data and supports more rigorous evaluation of machine learning approaches for heart disease prediction.

KEYWORDS

Heart disease prediction, machine learning, data pre-processing, missing-value handling, KNN imputation, CRISP-DM, Object-Oriented Methodology, ROC-AUC, clinical data

I. INTRODUCTION

A. Background of the Study

Heart disease remains a major health concern worldwide and continues to create a substantial burden on healthcare systems. Early identification of individuals who may be at risk is important because timely assessment and intervention can improve clinical

decision-making and patient management. The increasing availability of clinical datasets has encouraged the application of machine learning techniques to cardiovascular disease prediction. Machine learning provides methods for identifying patterns within clinical data and can support predictive and decision-support applications in healthcare.

However, the performance and reliability of a machine learning model depend considerably on the quality of the data used during model development. Clinical datasets may contain missing observations, categorical variables, inconsistent data representations, and numerical features measured on different scales. If these issues are not appropriately addressed, they may affect the learning process and consequently influence the predictive performance of the resulting models. Data pre-processing is therefore an important component of a reliable machine learning pipeline rather than merely a preliminary step before model training. Missing values are particularly common in clinical datasets. Patient information may be unavailable because a particular examination was not performed, a measurement was not recorded, or information was lost during data collection. Removing observations containing missing values can reduce the effective sample size and may also introduce undesirable bias. Consequently, imputation techniques have been developed to estimate missing observations while retaining otherwise useful records. K-nearest-neighbour (KNN) imputation is one approach that estimates missing values using information from observations with similar characteristics.

Appropriate representation of clinical variables is also important during pre-processing. Categorical attributes must generally be transformed into numerical representations before they can be used by machine learning algorithms. In addition, numerical variables can have substantially different ranges. Standardization can therefore be used to transform numerical features to a common scale, reducing the influence of differences in feature magnitude and producing a more consistent input representation for subsequent modelling. The evaluation of machine learning models also requires consideration beyond overall classification accuracy. Accuracy describes the proportion of correctly classified observations, but it may not provide sufficient information about the ability of a model to distinguish between different classes, particularly when class distributions are unequal. Measures such as precision, recall, F1-score, and receiver operating characteristic area under the curve (ROC-AUC) can provide complementary information about classifier performance. ROC analysis, in particular, provides a framework for examining the discriminatory behaviour of classification models across different decision thresholds.

In heart disease prediction, this distinction is particularly relevant because a model with comparatively high accuracy may not necessarily provide the strongest discrimination between individuals with and without the condition. Consequently, evaluating several complementary performance measures can provide a more comprehensive assessment of predictive models. This consideration is especially important when machine learning systems are intended to support healthcare-related decision making.

Although numerous studies have applied machine learning algorithms to heart disease prediction, considerable attention has traditionally been directed toward the selection and comparison of predictive algorithms. Less emphasis may be placed on examining the data

preparation process as an important determinant of the quality of the subsequent modelling process. In particular, missing-value handling, categorical encoding, feature standardization, and assessment of data completeness should be systematically considered when preparing clinical datasets for machine learning. To address this issue, this study investigates the impact of data pre-processing and missing-value handling on machine learning-based heart disease prediction. The study employs a hybrid methodological framework that combines the Cross-Industry Standard Process for Data Mining (CRISP-DM) with Object-Oriented Methodology (OOM). CRISP-DM provides a structured framework for progressing from data understanding and preparation to modelling and evaluation, while OOM supports the modular organization and implementation of the system components. The resulting pre-processing pipeline incorporates data quality assessment, KNN-based missing-value imputation, categorical variable encoding, and feature standardization.

Following pre-processing, Logistic Regression, Random Forest, and XG Boost models are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The experimental results provide an opportunity to examine whether systematic preparation of clinical data is associated with meaningful differences in predictive performance. In particular, the ROC-AUC results obtained in this study show that Logistic Regression achieved an AUC of 0.591, followed by XG Boost with 0.566, while Random Forest recorded 0.514. These results further demonstrate the importance of considering multiple evaluation measures rather than relying exclusively on accuracy when assessing machine learning models for heart disease prediction. The contribution of this study is therefore cantered on the development and evaluation of a structured and reproducible data pre-processing pipeline for clinical heart disease prediction. Rather than proposing a new classification algorithm, the study emphasizes the importance of systematic data quality assessment, missing-value treatment, categorical encoding, and feature standardization as integral components of a machine learning-based prediction process. The findings provide additional evidence for the need to consider data preparation and comprehensive model evaluation together when developing machine learning systems for healthcare applications.

B. Statement of the problem

The increasing application of machine learning to heart disease prediction has created opportunities for developing data-driven tools that can support early risk assessment and clinical decision-making. However, the reliability of these systems is strongly influenced by the quality of the clinical data used to develop them. Clinical datasets frequently contain missing values, categorical attributes, and numerical variables with different scales. If these characteristics are not appropriately addressed, the resulting dataset may not provide a consistent representation of the underlying clinical information, potentially affecting the performance and reliability of predictive models. Missing-value handling presents a particular challenge in clinical prediction. Eliminating observations with missing values may reduce the amount of information available for model development, especially when the dataset is relatively small. Conversely, inappropriate imputation may introduce distortions into the data and influence the relationships between clinical variables. Therefore, selecting an appropriate missing-value treatment method is important when preparing clinical datasets for machine learning.

Another problem concerns the tendency of some heart disease prediction studies to emphasize the selection of machine learning algorithms while providing comparatively limited attention to the data pre-processing procedures that precede model development. Although models such as Logistic Regression, Random Forest, and XG Boost can produce useful predictions, differences in pre-processing can influence the representation of the input data and consequently affect model performance. Furthermore, evaluating models primarily according to accuracy may provide an incomplete assessment of their discriminatory capabilities. Metrics such as precision, recall, F1-score, and ROC-AUC are necessary to provide a broader understanding of model behaviour. The problem addressed by this study is therefore the need for a systematic and reproducible approach to preparing clinical heart disease data before machine learning modelling. In particular, there is a need to appropriately assess data quality, handle missing observations, encode categorical variables, standardize numerical features, and subsequently determine how the prepared data perform when used with different machine learning algorithms. Addressing these issues can contribute to the development of more reliable and appropriately evaluated heart disease prediction systems.

C. Aim and objectives of the Study

The aim of this study is to investigate the impact of systematic data pre-processing and missing-value handling on the performance of machine learning models for heart disease prediction.

The specific objectives are to:

- assess the quality and completeness of the clinical heart disease dataset before and after pre-processing;
- identify and treat missing observations using K-nearest-neighbour (KNN) imputation;
- transform categorical variables into suitable numerical representations for machine learning;
- standardize numerical features to provide a consistent feature representation for predictive modelling;
- evaluate Logistic Regression, Random Forest, and XG Boost models using accuracy, precision, recall, F1-score, and ROC-AUC after data pre-processing; and
- Establish a structured and reproducible pre-processing framework for machine learning-based heart disease prediction.

D. Significance of the study

This study is significant because it addresses an important aspect of machine learning-based heart disease prediction that occurs before the actual classification process: the quality and preparation of clinical data. By systematically examining missing-value treatment, categorical encoding, feature standardization, and data completeness, the study provides a clearer understanding of how clinical datasets can be prepared for predictive modelling.

The study is also significant to healthcare and clinical decision-support research. Reliable prediction depends on appropriately prepared input data, and therefore systematic pre-processing can contribute to the development of more dependable machine learning applications for cardiovascular risk assessment. Although the models investigated in this study are not intended to replace professional clinical judgement, the findings can support research into data-driven tools for preliminary risk prediction. For machine learning researchers, the study provides a reproducible pre-processing workflow that can be applied when working with small clinical datasets containing missing and heterogeneous variables. The explicit documentation of the pre-processing stages also makes it easier for researchers to reproduce the experimental process and compare their findings with those of other studies.

The study is further significant for developers of clinical prediction systems because it demonstrates the importance of integrating data quality assessment and pre-processing into the overall system development process rather than treating them as isolated preliminary activities. The use of a hybrid CRISP-DM and Object-Oriented Methodology provides a structured approach that connects data preparation with subsequent modelling and evaluation. Finally, the study contributes to the broader body of knowledge on machine learning for cardiovascular disease prediction by demonstrating the importance of evaluating predictive models using multiple performance measures. The observed ROC-AUC results, in which Logistic Regression achieved 0.591, XG Boost achieved 0.566, and Random Forest achieved 0.514, illustrate why model assessment should extend beyond accuracy alone.

E. Contribution of the Study

The primary contribution of this study is the development of a structured and reproducible data pre-processing framework for machine learning-based heart disease prediction. The framework integrates data quality assessment, missing-value identification and treatment, categorical variable encoding, and numerical feature standardization into a unified pre-processing pipeline. A second contribution is the empirical assessment of KNN-based missing-value handling within a heart disease prediction context. Rather than simply removing incomplete observations, the study incorporates KNN imputation to preserve available records while producing a complete dataset suitable for subsequent analysis.

A third contribution is the systematic evaluation of multiple predictive models following the pre-processing process. Logistic Regression, Random Forest, and XG Boost are evaluated using several complementary metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. This provides a broader assessment of predictive behaviour than relying on a single performance measure. A fourth contribution is the integration of CRISP-DM and Object-Oriented Methodology into the study framework. CRISP-DM provides an organized data-mining process, while OOM supports modular system development. Their combination provides a structured approach for connecting data preparation, predictive modelling, and model evaluation.

Finally, the study contributes empirical evidence showing that the model with the highest accuracy does not necessarily have the strongest discriminative performance. In the present evaluation, Random Forest achieved the highest accuracy among the evaluated models, while Logistic Regression recorded the highest ROC-AUC. This finding reinforces the importance of using multiple evaluation criteria when assessing machine learning models for clinical prediction.

II. RELATED WORK

A. Machine Learning for Heart Disease Prediction

Machine learning has increasingly been applied to cardiovascular disease prediction because of its ability to learn patterns from clinical variables and generate predictions from combinations of patient characteristics. The Cleveland Heart Disease dataset from the UCI Machine Learning Repository has been particularly influential in this area. The dataset contains 303 patient records and a commonly used subset of 14 variables, consisting of 13 predictive attributes and a diagnostic target. The predictive variables include demographic, clinical, electrocardiographic, and exercise-related measurements.

Several studies have used the Cleveland dataset to compare different machine learning algorithms for heart disease prediction. Such studies have investigated conventional statistical models alongside tree-based and ensemble learning techniques. For example, a comparative study of coronary artery disease prediction examined multiple machine learning algorithms using the Cleveland data and reported that preprocessing decisions, including the treatment of missing observations, transformation of variables, and normalization, formed an important part of the modelling pipeline. Other studies have similarly demonstrated the application of machine learning classifiers to heart disease datasets. These investigations commonly include data cleaning, categorical-variable transformation, feature scaling, model training, and performance evaluation as sequential stages of the predictive process. However, differences in the preprocessing strategies adopted by individual studies make direct comparison of reported model performance difficult. This suggests that attention should be given not only to the predictive algorithm but also to how the input data are prepared before modelling.

B. Data Preprocessing in Clinical Machine Learning

Data preprocessing is a fundamental stage in data mining and machine learning because real-world datasets frequently require cleaning, transformation, and preparation before they can be used effectively for modelling. In clinical applications, preprocessing may involve missing-value treatment, data-type conversion, categorical encoding, feature scaling, outlier assessment, and data-quality verification. The importance of this stage has been demonstrated in medical data-mining research, where data preparation can represent a substantial portion of the overall analytical process.

The CRISP-DM framework provides a structured approach for organizing data-mining projects from business or problem understanding through data understanding, data

preparation, modelling, evaluation, and deployment. Its data-preparation stage is particularly relevant to clinical prediction because it provides a systematic process for transforming raw data into a form suitable for analytical modelling. The application of CRISP-DM to health-related data has also been reported as a means of organizing specialized data-preparation activities for medical and epidemiological datasets.

Recent heart disease studies continue to emphasize the importance of preprocessing. For example, recent work using the Cleveland dataset has incorporated missing-value treatment, categorical encoding, and feature normalization before predictive modelling. These studies reinforce the view that preprocessing is an essential component of the machine learning pipeline. Nevertheless, preprocessing is frequently presented as a supporting methodological step rather than being investigated as the primary subject of the study.

C. Missing-Value Handling in Heart Disease Datasets

Missing observations present a particular challenge when preparing clinical datasets for machine learning. The Cleveland heart disease dataset contains missing values in specific clinical variables, including the number of major vessels observed through fluoroscopy (ca) and the thal variable [6]. These missing observations require appropriate treatment because deleting incomplete records can reduce the available sample size, which is particularly important when working with relatively small clinical datasets.

Different studies have adopted different strategies for addressing missing values. One approach is complete-case analysis, in which records containing missing observations are removed before model training. A recent study using the Cleveland subset, for example, removed incomplete records and obtained 297 observations from the original 303 records. Another approach is imputation, in which missing values are estimated using information available from other observations.

KNN-based imputation represents one such approach. Rather than replacing a missing value using only a global statistic, KNN imputation uses neighboring observations that are similar according to the available feature information. Recent cardiovascular prediction studies have applied KNN-based imputation to missing clinical variables as part of their preprocessing pipelines. The choice between deletion and imputation can therefore affect both the size and composition of the dataset available for subsequent modelling.

The literature demonstrates that missing-value treatment is not simply a data-cleaning decision; it can influence the information available to the predictive model. However, there remains a need for empirical studies that explicitly examine missing-value handling as part of a reproducible preprocessing framework for small heart disease datasets.

D. Feature Encoding and Normalization

Clinical heart disease datasets contain both categorical and numerical variables. For example, the Cleveland dataset includes categorical variables such as sex, chest-pain type, fasting blood sugar status, resting electrocardiographic results, exercise-induced angina, slope, and thalassemia-related information, together with numerical variables such as age,

resting blood pressure, cholesterol, maximum heart rate, and ST depression [6]. Such mixed data types require appropriate transformation before they are supplied to machine learning algorithms.

Categorical encoding converts non-numerical or categorical information into numerical representations that can be processed by machine learning models. Feature scaling or normalization is similarly used to reduce the effect of differences in numerical magnitude among variables. Studies involving heart disease prediction have therefore incorporated categorical encoding and normalization as part of their preprocessing pipelines.

Although these procedures are routinely used, their importance can be underestimated when studies concentrate primarily on comparing classifier accuracy. A systematic preprocessing pipeline provides a consistent basis for subsequent model evaluation and improves the reproducibility of machine learning experiments.

E. Evaluation of Machine Learning Models

The evaluation of heart disease prediction models also requires careful consideration. Accuracy is commonly reported because it provides a straightforward measure of the proportion of correctly classified observations. However, accuracy alone may not adequately describe the behavior of a classifier, particularly when the distribution of classes is unequal or when the costs of different classification errors are not equivalent.

Consequently, several complementary metrics are commonly used, including precision, recall, F1-score, and ROC-AUC. ROC analysis provides a means of examining the trade-off between the true-positive and false-positive rates across different classification thresholds. The use of multiple metrics can therefore provide a more complete assessment of model discrimination than accuracy alone.

This issue is particularly relevant to the present study. The evaluated models produced different performance patterns across the selected metrics. Logistic Regression recorded an ROC-AUC of 0.591, XG Boost achieved 0.566, and Random Forest achieved 0.514. These results indicate that the model with the highest accuracy should not automatically be regarded as having the strongest discriminatory capability. A comprehensive evaluation is therefore necessary when assessing models for clinical prediction.

F. Research Gap

The reviewed literature demonstrates that machine learning has been extensively applied to heart disease prediction and that preprocessing operations such as missing-value treatment, categorical encoding, and feature normalization are commonly included in predictive workflows. However, many studies primarily emphasize classifier development and comparative predictive performance, while data preprocessing is often treated as a preliminary technical procedure rather than as an important analytical component.

Furthermore, studies using the Cleveland dataset have adopted different approaches to missing observations. Some have removed incomplete records, while others have applied

imputation techniques. This variation highlights the need for a more explicitly structured and reproducible approach to data preparation.

Therefore, the present study addresses this gap by focusing specifically on systematic data preprocessing and missing-value handling as integral components of machine learning-based heart disease prediction. The study combines data quality assessment, KNN-based imputation, categorical encoding, and feature standardization within a hybrid CRISP-DM and Object-Oriented Methodology framework. The resulting dataset is then evaluated using Logistic Regression, Random Forest, and XG Boost, with performance assessed using multiple complementary metrics. This approach shifts the emphasis from merely asking which algorithm performs best toward examining how systematically prepared clinical data can support more reliable and reproducible predictive modelling.

III. METHODOLOGY

A. Research Design

This study adopted a hybrid methodological framework that combines the Cross-Industry Standard Process for Data Mining (CRISP-DM) with Object-Oriented Methodology (OOM). The combination was selected to provide both a structured data-mining process and a modular approach to system development. CRISP-DM guided the analytical activities from understanding the heart disease dataset through data preparation, modelling, and evaluation, while OOM provided a structured basis for organizing the functional components of the proposed system. This combination supports a systematic and reproducible workflow in which data quality assessment and pre-processing are treated as integral stages of the prediction process.

The methodological workflow consisted of six major stages: problem understanding, data understanding, data preparation, predictive modelling, model evaluation, and system organization. Particular emphasis was placed on the data-preparation stage because the primary purpose of the study was to investigate the role of missing-value handling and other pre-processing operations in heart disease prediction.

B. Dataset Description

The study employed the Cleveland Heart Disease dataset obtained from the UCI Machine Learning Repository. The dataset contains 303 patient instances and 13 predictive attributes with one target variable [6]. The variables represent demographic, clinical, and diagnostic characteristics associated with heart disease. The dataset includes attributes such as age, sex, chest-pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia-related information.

The target variable represents the presence or absence of heart disease. The dataset was selected because it is a widely used benchmark dataset for evaluating machine learning approaches to cardiovascular disease prediction and contains a combination of numerical and categorical clinical variables.

C. Missing-Value Handling in Clinical Data

Missing values are a significant challenge in healthcare datasets because incomplete observations may reduce the information available to machine learning models. Removing records containing missing values can reduce the effective sample size, while inappropriate imputation can alter relationships among clinical variables. Recent research has therefore examined different approaches for handling missing clinical data.

Rios et al. investigated several missing-data strategies for machine learning-based cardiovascular event prediction using the REFINE SPECT registry [12]. Their results showed that models affected by missing values generally produced lower AUC values than the reference model without missing data, demonstrating that missing-value treatment can influence the discriminatory performance of cardiovascular prediction models. Similarly, Liu et al. reviewed deep-learning and other approaches for missing-value imputation in healthcare data and highlighted the importance of selecting appropriate imputation strategies for maintaining the quality and reliability of clinical datasets. Recent cardiovascular research has identified KNN, median, and other statistical or machine-learning-based approaches among the techniques used for missing-value treatment. However, the effectiveness of an imputation method depends on the characteristics of the dataset and the modelling task. Consequently, an imputation method should be evaluated within the specific analytical context rather than assumed to be universally optimal.

D. Data Pre-processing

The pre-processing stage consisted of four principal operations: missing-value identification and treatment, categorical-variable encoding, feature standardization, and final data-quality verification.

1. Missing-Value Identification and Treatment

Missing observations were first identified across the dataset. Instead of removing records containing incomplete values, the study employed K-nearest-neighbour (KNN) imputation. KNN imputation estimates a missing value using information obtained from observations that are similar to the incomplete observation.

The choice of KNN imputation was motivated by its ability to exploit relationships among observations rather than replacing missing values solely with a global mean or median. This approach allowed the study to retain the available patient records while generating complete input data for subsequent modelling.

The imputation process was performed before model training to ensure that the predictive algorithms received a consistent dataset without unresolved missing observations.

2. Categorical Variable Encoding

The dataset contains several categorical clinical variables. Because machine learning algorithms require appropriately represented numerical input, categorical attributes were transformed into numerical representations during pre-processing.

The encoding process ensured that categorical information was retained while producing a machine-readable representation of the clinical variables. This step was performed consistently across the dataset to avoid incompatible representations between training and testing data.

3. Feature Standardization

Following missing-value treatment and categorical encoding, numerical features were standardized using Standard Scaler. Standardization places numerical variables on a comparable scale and reduces the potential influence of differences in the original measurement ranges.

For a feature x, the standardized value z can be expressed as:

Where:

Feature value x → the raw observation.

Mean μ → the average of all observations for that feature.

Standard deviation σ → the measure of spread or variability.

Standardized value z → the transformed score showing how far x lies from the mean in units of standard deviation.

The transformation produces a standardized representation in which the feature distribution is centred on zero and scaled according to its standard deviation.

4. Final Data-Quality Verification

After the pre-processing operations were completed, the resulting dataset was examined again to confirm that missing values had been appropriately addressed and that the transformed variables were suitable for machine learning. The final completeness level was compared with the initial completeness level.

This verification stage was important because it provided evidence that the pre-processing pipeline had successfully transformed the original dataset into a suitable input for predictive modelling.

5. Predictive Modelling

Following pre-processing, three machine learning algorithms were used to evaluate the predictive utility of the prepared dataset:

- Logistic Regression (LR);
- Random Forest (RF); and
- Extreme Gradient Boosting (XG Boost).

These models were selected because they represent different modelling approaches. Logistic Regression provides a conventional statistical classification approach, Random Forest is an ensemble of decision trees, while XG Boost uses gradient-boosting principles to construct an ensemble of predictive learners.

The models were trained using the pre-processed heart disease dataset. The same prepared data were supplied to the evaluated models to provide a consistent basis for comparison. This ensured that differences in observed predictive performance could be examined within the same pre-processing framework.

6. Model Evaluation

The predictive models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The use of multiple evaluation measures was intended to provide a broader assessment of model performance rather than relying exclusively on accuracy.

Accuracy measures the proportion of observations correctly classified by the model. Precision measures the proportion of predicted positive cases that were actually positive, while recall measures the proportion of actual positive cases correctly identified by the model. The F1-score provides a combined measure of precision and recall.

ROC-AUC was also used to assess the discriminative ability of the models. A higher ROC-AUC indicates stronger ability to distinguish between the evaluated classes across classification thresholds [5]. The ROC-AUC results obtained in the study were shown in table 1.

Table 1: ROC-AUC Performance of the Machine Learning Models

Model	**ROC-AUC**
Logistic Regression	0.591
XG Boost	0.566
Random Forest	0.514

The ROC-AUC results indicate differences in the ability of the three machine-learning models to discriminate between the outcome classes. Logistic Regression achieved the highest ROC-AUC value of 0.591, indicating that it provided the best overall class-discrimination performance among the three models based on this metric. XG Boost recorded a ROC-AUC of 0.566, placing it second, while Random Forest obtained the lowest value of 0.514. Therefore, based on ROC-AUC alone, Logistic Regression demonstrated the strongest discriminative performance, followed by XG Boost and Random Forest.

However, the ROC-AUC values are all relatively close to 0.5, which represents chance-level discrimination. Thus, although Logistic Regression performed better than the other two models in terms of ROC-AUC, the results indicate limited overall discriminatory ability of the models on the evaluated dataset. This finding should be considered alongside the other performance measures, particularly recall, precision, F1-score, and accuracy, rather than using ROC-AUC alone to determine the best model.

7. Integration of CRISP-DM and Object-Oriented Methodology

The CRISP-DM framework provided the overall analytical structure for the study. Its stages were adapted to the research process as follows:

Table 2: Application of the CRISP-DM Framework to the Heart Disease Prediction Study

CRISP-DM Stage	Application in the Study
Business/Problem Understanding	Definition of the heart disease prediction problem
Data Understanding	Examination of the Cleveland dataset and its characteristics
Data Preparation	Missing-value treatment, encoding, standardization, and completeness assessment
Modelling	Training of LR, RF, and XG Boost
Evaluation	Assessment using accuracy, precision, recall, F1-score, and ROC-AUC
Deployment/System Integration	Organization of the resulting predictive components

OOM complemented CRISP-DM by providing a modular perspective for organizing the system components. The major functional components were conceptualized as separate modules for data acquisition, pre-processing, predictive modelling, and model evaluation. This modular organization facilitates maintenance, modification, and potential extension of the prediction system.

8. Experimental Workflow

The complete experimental workflow can be summarized as:

Cleveland Heart Disease Dataset
↓
Data Quality Assessment
↓
Missing-Value Identification
↓
KNN Imputation
↓
Categorical Variable Encoding
↓
Feature Standardization
↓
Data Completeness Verification
↓
Logistic Regression / Random Forest / XG Boost
↓
Performance Evaluation

↓
Accuracy, Precision, Recall, F1-Score and ROC-AUC

9. Comparative Analysis

This workflow ensures that the three predictive models were evaluated using systematically prepared clinical data and that their performance could be examined using several complementary measures.

10. Reproducibility Considerations

To promote reproducibility, the pre-processing operations were defined as sequential and consistent transformations rather than as independent manual modifications. The same pre-processing principles were applied before predictive modelling, and the model evaluation was based on predefined performance measures. This approach reduces ambiguity in the preparation of the dataset and provides a basis for repeating the experiment with the same dataset or evaluating the workflow on related clinical datasets.

The methodology therefore treats data preparation as an integral part of the predictive system, rather than merely a preliminary cleaning activity. This distinction is central to the study because the principal research interest is to examine how systematic handling of missing values and other data-preparation requirements supports machine learning-based Heart disease prediction.

IV. RESULTS

A. Dataset Pre-processing Results

The heart disease dataset was examined to determine its completeness and suitability for machine learning analysis. The pre-processing assessment identified missing observations within the dataset, which were subsequently addressed using K-nearest-neighbour (KNN) imputation. Rather than deleting incomplete records, the imputation procedure retained the available observations for subsequent model development.

Following missing-value treatment, the categorical attributes were transformed into numerical representations, while the numerical features were standardized. These pre-processing steps produced a consistent dataset for training and evaluating the Logistic Regression, Random Forest, and XG Boost models

The pre-processing assessment identified missing observations within the dataset, which were subsequently addressed using K-nearest-neighbour (KNN) imputation. Rather than deleting incomplete records, the imputation procedure retained the available observations for subsequent model development. Following missing-value treatment, the categorical attributes were transformed into numerical representations, while the numerical features were standardized. These pre-processing steps produced a consistent dataset for training and evaluating the Logistic Regression, Random Forest, and XG Boost models. The overall

data pre-processing and missing-value handling workflow applied in the study is illustrated in Fig. 1.

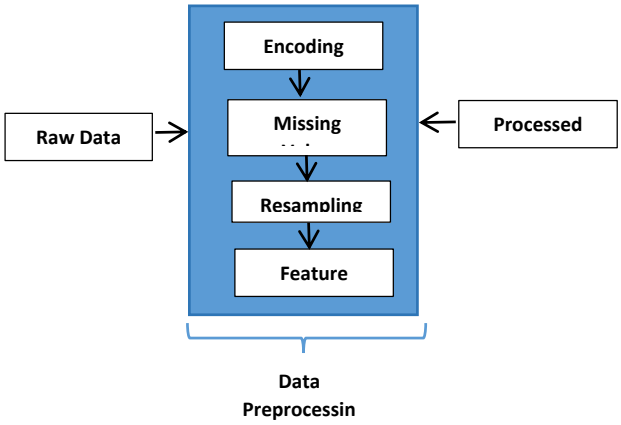

Figure: 1. Distribution of missing observations in the heart disease dataset before pre-processing

The figure1 presents the structured pipeline employed to transform raw data into a consistent format suitable for machine learning analysis. The workflow begins with encoding categorical data, ensuring that non-numeric attributes are converted into numerical representations. This is followed by missing value imputation, where incomplete records are addressed using KNN-based methods to preserve valuable observations. Subsequently, resampling is applied to mitigate class imbalance, thereby enhancing the fairness of model training. Finally, feature scaling standardizes numerical features, aligning them to comparable ranges for improved algorithmic performance. Collectively, these steps form the Data Pre-processing stage, producing a processed dataset that supports robust training and evaluation of models such as Logistic Regression, Random Forest, and XG Boost. The figure thus encapsulates the methodological foundation of the study, offering a clear visualization of the pre-processing strategy that underpins subsequent analysis.

B. Predictive Performance of the Models

The three machine learning models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The results are presented in Table 3.

Table 3. Predictive Performance of the Machine Learning Models

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	67.21	0.3158	0.4615	0.3750	0.5913
Random Forest	80.33	0.6000	0.2308	0.3333	0.5144
XG Boost	73.77	0.3333	0.2308	0.2727	0.5657

The table 3, reports the evaluation metrics for three machine learning models applied to the heart disease dataset. Logistic Regression achieved an accuracy of 67.21%, with

moderate recall (0.4615) but relatively low precision (0.3158), resulting in an F1-score of 0.3750 and a ROC-AUC of 0.5913. Random Forest produced the highest accuracy at 80.33% and the strongest precision (0.6000), though its recall was limited (0.2308), yielding an F1-score of 0.3333 and a ROC-AUC of 0.5144. XG Boost attained an intermediate accuracy of 73.77%, with balanced though modest precision (0.3333) and recall (0.2308), leading to an F1-score of 0.2727 and a ROC-AUC of 0.5657. Overall, the results highlight trade-offs across models: Random Forest demonstrated superior accuracy and precision, Logistic Regression offered better recall and ROC-AUC, while XG Boost provided a middle ground in terms of accuracy and discrimination ability. This comparative analysis underscores the importance of considering multiple evaluation metrics rather than relying solely on accuracy when assessing predictive performance in medical datasets.

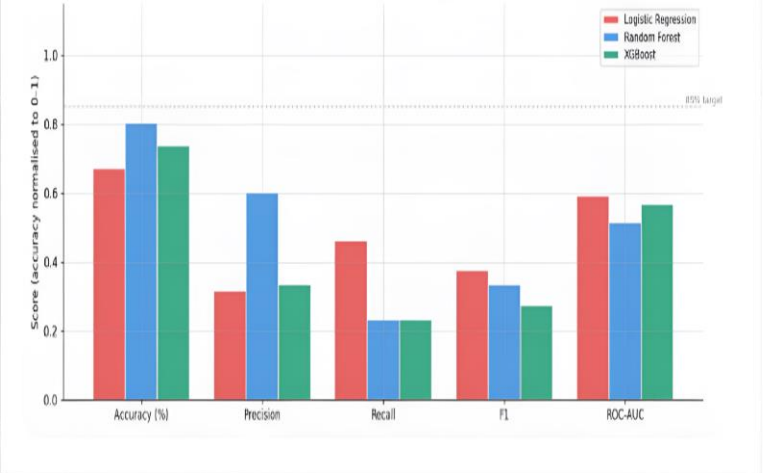

Figure. 2. Comparative performance of the Logistic Regression, Random Forest, and XG Boost models after data pre-processing.

The figure 2 above provides a visual comparison of three machine learning models: Logistic Regression, Random Forest, and XG Boost evaluated on the heart disease dataset. Performance is illustrated across five key metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Logistic Regression demonstrates relatively stronger recall and F1-score, alongside the highest ROC-AUC, suggesting better discrimination ability. Random Forest achieves the highest accuracy and precision, though its recall remains limited, reflecting a trade-off between predictive confidence and sensitivity. XG Boost shows intermediate performance, balancing accuracy with moderate values across the other metrics. By presenting these results in bar chart form, the figure complements the tabular data, enabling readers to quickly discern the relative strengths and weaknesses of each algorithm. This visualization underscores the importance of multi-metric evaluation in medical prediction tasks, where reliance on a single measure such as accuracy may obscure critical differences in model behaviour.

C. ROC-AUC Analysis

ROC-AUC was used to further assess the discriminatory ability of the three models. The results are presented in Table 4.

Table 4: ROC-AUC Performance of the Models

Model	AUC Score	Interpretation
Logistic Regression	0.591	Highest discriminative ability among the evaluated models
XG Boost	0.566	Moderate but limited discriminatory capability
Random Forest	0.514	Performance close to random classification

The table 4, presents the Area under the Curve (AUC) scores for three machine learning models applied to the heart disease dataset, along with their interpretations. Logistic Regression achieved the highest AUC score (0.591), indicating the strongest discriminative ability among the evaluated models. XG Boost recorded a moderate AUC of 0.566, reflecting limited but acceptable discriminatory capability. In contrast, Random Forest produced the lowest AUC score (0.514), suggesting performance close to random classification.

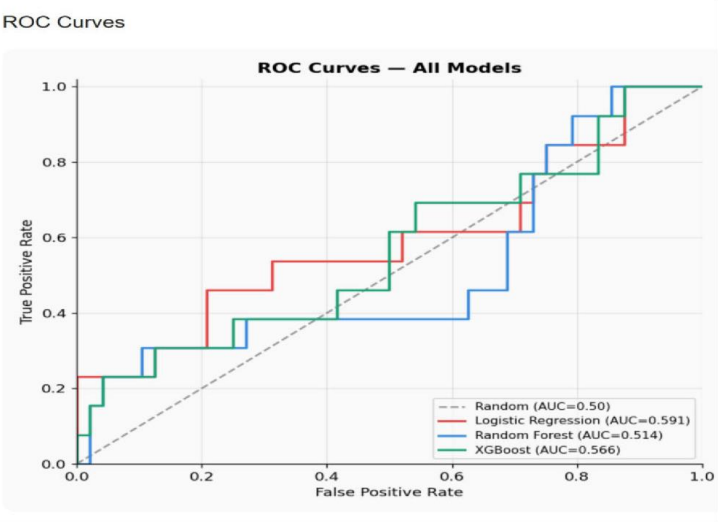

Figure. 3. ROC curves and AUC values of the Logistic Regression, Random Forest, and XG Boost models

The figure 3 compares the Receiver Operating Characteristic (ROC) curves of three machine learning models: Logistic Regression, Random Forest, and XG Boost against a random baseline classifier. The diagonal dashed line represents random performance (AUC = 0.50), serving as a reference point. Logistic Regression achieved the highest AUC (0.591), indicating superior discriminative ability relative to the other models. XG Boost followed with an AUC of 0.566, reflecting moderate discriminatory capability, while Random Forest produced the lowest AUC (0.514), suggesting performance close to random classification. This visualization reinforces the tabular results by illustrating the trade-offs in model performance. It highlights that, despite Random Forest's higher accuracy in earlier evaluations, Logistic Regression provided stronger overall class separation when assessed

through ROC-AUC. The figure underscores the importance of ROC analysis in medical prediction tasks, where sensitivity and specificity must be balanced to ensure reliable diagnostic support.

D. Comparative Analysis of Model Performance

The results demonstrate that no single model achieved the best performance across all evaluation measures. Random Forest achieved the highest accuracy and precision, whereas Logistic Regression produced the highest recall, F1-score, and ROC-AUC.

The relatively higher recall obtained by Logistic Regression is particularly relevant to heart disease prediction because recall reflects the proportion of positive cases correctly identified by the model. However, the ROC-AUC values of all three models were relatively modest, indicating limited discriminatory performance under the experimental conditions.

E. Discussion of Results

The comparative evaluation of Logistic Regression, Random Forest, and XG Boost models on the heart disease dataset revealed distinct strengths and limitations across performance metrics. Random Forest achieved the highest overall accuracy (80.33%) and precision (0.6000), indicating strong predictive confidence. However, its ROC-AUC score (0.5144) was close to random classification, suggesting limited discriminative ability despite its accuracy advantage. In contrast, Logistic Regression demonstrated the highest recall (0.4615), F1-score (0.3750), and ROC-AUC (0.5913). These results highlight its superior ability to correctly identify positive cases and provide better class separation, making it more reliable in medical prediction contexts where sensitivity is critical. XG Boost offered intermediate performance, with an accuracy of 73.77% and an ROC-AUC of 0.5657. While not outperforming the other models in any single metric, it provided a balanced compromise between accuracy and discriminative power. The ROC curve analysis further reinforced these findings: Logistic Regression consistently outperformed the other models in terms of class separation, while Random Forest's curve remained close to the random baseline. This divergence between accuracy and ROC-AUC rankings underscores the importance of employing multiple evaluation measures. Relying solely on accuracy can be misleading, as it may mask deficiencies in a model's ability to distinguish between classes a critical requirement in medical diagnosis.

Overall, the results suggest that while Random Forest excels in accuracy and precision, Logistic Regression provides more robust discriminative performance, and XG Boost offers a middle ground. These findings emphasize the necessity of multi-metric evaluation in healthcare-related machine learning applications, ensuring that models are assessed not only for predictive correctness but also for their diagnostic reliability.

V. CONCLUSION

This study investigated the impact of data pre-processing and missing-value handling on machine learning-based heart disease prediction. The pre-processing pipeline incorporated missing-value identification and KNN imputation, categorical-variable encoding, and numerical-feature standardization before the development of the predictive models. The results demonstrate that systematic preparation of clinical data is an important component

of the machine learning pipeline because it provides a consistent dataset for model development and comparison.

The comparative results showed that Random Forest achieved the highest accuracy (80.33%) and precision (0.6000), while Logistic Regression achieved the highest recall (0.4615), F1-score (0.3750), and ROC-AUC (0.5913). XG Boost achieved an accuracy of 73.77% and an ROC-AUC of 0.5657, while Random Forest recorded an ROC-AUC of 0.5144. The results therefore demonstrate that the model with the highest accuracy was not necessarily the model with the strongest discriminatory performance.

The higher recall obtained by Logistic Regression is particularly noteworthy for heart disease prediction because it indicates that the model identified a larger proportion of positive cases than the other evaluated models. However, the ROC-AUC of 0.5913 remains relatively modest, indicating that the model's discriminatory ability was limited. Consequently, the findings should be interpreted as a comparative assessment of the evaluated models rather than evidence of a clinically ready prediction system.

Overall, the study demonstrates that data pre-processing and missing-value handling are important components of machine learning-based heart disease prediction, but pre-processing alone does not guarantee strong predictive performance. The observed differences among the models further demonstrate the importance of evaluating predictive systems using multiple performance measures rather than relying exclusively on accuracy.

A. Direction for Future Work

Future research can extend this study in several directions. First, different missing-value strategies, including Multiple Imputation, Iterative Imputation, and other model-based imputation techniques, can be compared with KNN imputation to determine their influence on predictive performance.

Second, future studies should evaluate the proposed pre-processing approach using larger, more diverse, and independent clinical datasets. This would help determine whether the observed findings generalize beyond the dataset used in the present study.

Third, future work should investigate the effect of feature selection and feature engineering on model performance. Identifying the most informative clinical predictors may reduce unnecessary variables and potentially improve model discrimination.

Fourth, researchers could investigate hyper parameter optimization, ensemble learning, and alternative machine learning or deep-learning algorithms to determine whether improved predictive performance can be obtained after appropriate pre-processing.

Finally, future research should consider a direct before-and-after pre-processing experiment. Models could be trained and evaluated using the original dataset and then using the fully pre-processed dataset under the same experimental conditions. Comparing the two results would provide stronger empirical evidence of the actual impact of missing-value handling, encoding, and standardization on model performance. This would be particularly valuable because the present study primarily evaluates model performance after pre-processing, rather than experimentally isolating the contribution of each pre-processing operation.

RECOMMENDATIONS

Based on the findings of the study, the following recommendations are proposed:

- Systematic pre-processing should be incorporated into heart disease prediction workflows. Missing-value assessment, appropriate imputation, categorical encoding, and feature standardization should be performed before model training to improve the consistency and quality of clinical data.
- Accuracy should not be used as the sole model-selection criterion. Precision, recall, F1-score, and ROC-AUC should also be considered, particularly because heart disease prediction involves the identification of positive cases where false-negative predictions may be important.
- Missing-value treatment should be selected based on the characteristics of the dataset. KNN imputation was used in this study, but researchers should evaluate alternative imputation techniques to determine whether they provide better predictive outcomes.
- Models should be evaluated using clinically relevant performance measures. The relatively higher recall achieved by Logistic Regression suggests that recall deserves particular attention when the primary concern is identifying patients who may have heart disease.
- External validation should be considered before clinical application. The models evaluated in this study should not be considered clinically deployable based solely on the results obtained from the dataset used in the experiment.

REFERENCES

[1] "An explainable artificial intelligence framework for ischemic heart disease prediction using enhanced squirrel search feature selection," Scientific Reports, 2026. The study specifically discusses missing-value handling, KNN imputation, categorical encoding, and feature normalization in a UCI heart disease prediction pipeline.

[2] "Artificial intelligence for heart disease prediction and imputation of missing data in cardiovascular datasets," Cogent Business & Management, 2024, doi: 10.1080/23311916.2024.2325635. The study investigates heart disease prediction under missing-data conditions and compares machine learning approaches and imputation strategies.

[3] "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," Computers in Biology and Medicine, vol. 176, 108432, 2024, doi: 10.1016/j.compbiomed.2024.108432. The study emphasizes pre-processing, missing/noisy data, feature selection, and precision-oriented evaluation in heart disease prediction.

[4] "Deep learning for cardiovascular disease: A comprehensive review of detection and risk forecasting," Frontiers in Artificial Intelligence, 2026. The review identifies KNN, median, regression, and other approaches for missing-data treatment and emphasizes standardized pre-processing pipelines for reproducible cardiovascular disease modelling.

[5] "Artificial Intelligence-Based Medical Data Mining," Journal of Personalized Medicine, vol. 12, no. 9, 1359, 2022. The study discusses CRISP-DM for medical data mining and describes data preparation as including selection, cleaning, imputation, transformation, integration, and formatting.

[6] "Symptoms and causes of obstacles in AI-based tele monitoring: Strategies for system development through a CRISP-DM lens," Frontiers in Artificial Intelligence, 2026. The study applies a CRISP-DM perspective to healthcare and emphasizes data preparation activities including cleaning missing values, handling outliers, checking data quality, and formatting data for modelling.

[7] UCI Machine Learning Repository, "Heart Disease," Dataset 45, Cleveland Heart Disease Database. Note: This is the dataset source rather than a research-literature reference, so I recommend retaining it even though the dataset originates before 2020. The repository identifies the Cleveland dataset and its characteristics.

[8] "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," Applied Sciences, vol. 11, no. 18, p. 8352, and 2021.

[9] M. A. Islam, M. Z. H. Majumder, M. S. Miah, et al., "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," Computers in Biology and Medicine, vol. 176, p. 108432, 2024, doi: 10.1016/j.compbiomed.2024.108432.

[10] "Optimization of multidimensional feature engineering and data partitioning strategies in heart disease prediction models," Alexandria Engineering Journal, vol. 107, pp. 932–949, 2024, doi: 10.1016/j.aej.2024.09.037.

[11] "Deep learning for cardiovascular disease: A comprehensive review of detection and risk forecasting," Frontiers in Artificial Intelligence, 2026.

[12] R. Rios et al., "Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: Insights from REFINE SPECT registry," Computers in Biology and Medicine, vol. 145, p. 105449, 2022, doi: 10.1016/j.compbiomed.2022.105449.

[13] M. Liu et al., "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," Artificial Intelligence in Medicine, vol. 142, p. 102587, 2023, doi: 10.1016/j.artmed.2023.102587.