Coronary Heart Disease Prediction Using Machine Learning Algorithms

Inooc Rubio Paucar[1], Cesar Yactayo-Arias[2], Laberiano Andrade-Arenas[3]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Departamento de Estudios Generales, Universidad Continental, Lima, Perú[2]
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú[3]

Abstract—**Cardiopathy is one of the most serious diseases worldwide with its high morbidity and mortality rates posing a latent risk over time. The objective of this research focuses on evaluating Machine Learning (ML) models such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Logistic Regression (LR) for the prediction of coronary heart disease (CHD), with the aim of identifying the most efficient model for this prediction. The model construction followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which comprises five stages: business understanding, data understanding, data preparation, modeling, and evaluation. The modeling results revealed the superior predictive capability of the XGBoost algorithm for detecting coronary heart disease, compared to Random Forest and Logistic Regression. The assessment of performance metrics (Accuracy, Precision, Sensitivity, and F1 Score) established XGBoost as the reference model, highlighting an F1 Score of approximately 90.8%. This superiority is attributed to its robustness in capturing nonlinear interactions among clinical variables. Consequently, the XGBoost model is selected as the optimal tool for integration into future medical decision support systems. In summary, this ML-based approach provides a highly predictive tool capable of identifying subtle risk patterns from real clinical data. The XGBoost model is a promising candidate for integration into decision support systems and for the optimization of primary prevention protocols for coronary heart disease.**

Keywords*—*Cardiovascular disease; machine learning; prediction; random forest; XGBoost

I. Introduction

CHD continues to be one of the leading causes of death worldwide. According to the World Health Organization (WHO), cardiovascular diseases account for approximately 17.9 million deaths annually, representing 32% of all global deaths, and within this group, coronary heart disease constitutes the largest proportion of cases. Similarly, the American Heart Association (AHA) estimates that one in five deaths in the United States is related to coronary events, reflecting the magnitude of the problem in both developed and developing countries [1]. Furthermore, the European Society of Cardiology (ESC) [2] has indicated that despite therapeutic advances, late detection and the lack of accurate predictive tools remain key limitations in reducing the global burden of the disease.

On the other hand, the problem is exacerbated in low- and middle-income regions, where access to early diagnostics and specialized follow-up systems is limited, contributing to a sustained increase in premature mortality and a significant burden of associated disability [3]. Additionally, deficits in hospital infrastructure, the shortage of specialized professionals, and economic barriers to access advanced technologies create a substantial gap compared to high-income countries. In Latin America, reports from the Pan American Health Organization (PAHO) [4] indicate that cardiovascular diseases are responsible for more than 30% of deaths, with a significant proportion attributed to CHD. Added to this is the high economic impact, both from direct healthcare costs and indirect productivity losses, further increasing pressure on health systems [5]. These data underscore the urgent need for innovative strategies and accurate predictive tools to improve early detection and diagnosis in heterogeneous clinical contexts, with the aim of reducing health inequities and optimizing medical decision-making.

In this scenario, ML has emerged as a promising approach in healthcare, offering the ability to analyze large volumes of clinical data and detect nonlinear patterns that traditional statistical methods do not always capture. Recent research has applied algorithms such as LR [6], RF [7], and XGBoost [8] for cardiovascular disease prediction, showing promising results in terms of accuracy, sensitivity, and specificity. However, most of these studies have been conducted on specific populations, with small sample sizes and without external validation, limiting the generalizability of their findings. Furthermore, there is a noticeable scarcity of studies focused on Latin American populations, creating a significant gap in the literature and highlighting the need for contextualized research.

In this context, the present study aims to evaluate and compare the performance of ML algorithms LR, RF, and XGBoost in the prediction of coronary heart disease, with the purpose of identifying the most efficient model and providing evidence that contributes to the development of tools supporting early diagnosis in clinical practice. In this way, the study seeks not only to strengthen the field of medical informatics but also to address the need for solutions applicable to local contexts, with potential impact on reducing cardiovascular mortality and optimizing healthcare resources.

II. Literature Review

This section is organized into two main parts. First, a review of related works is presented, analyzing previous research in the field of cardiovascular disease prediction. This analysis allows for contextualizing the object of study and provides a comprehensive view of the current state of knowledge. Second, the theoretical foundations are addressed, including the main ML algorithms and the clinical aspects associated with CHD, with the purpose of conceptually supporting the research. The following subsections elaborate on these aspects:

A. Related Work

In recent literature, various authors [9] have explored the potential of ML models in the detection and prediction of complications associated with cardiovascular diseases. In particular, [10] developed an XGBoost-based model aimed at anticipating unplanned one-year readmission in elderly patients with CHD. The study, conducted on a cohort of 2,137 individuals, reported a predictive performance with an Area Under the Receiver Operating Characteristic curve (AUROC) [11] of 0.704 and an AUPRC of 0.392 [12]. Moreover, the most clinically relevant variables were identified as length of hospital stay (LOS), age-adjusted Charlson comorbidity index (ACCI), monocyte count, blood glucose, and red blood cell (RBC) count, highlighting the utility of these parameters as significant predictors in assessing readmission risk. In contrast, another study aimed to develop an ML-based predictive model to estimate frailty risk in elderly patients with CHD [13],[14]. The main predictors identified were Activities of Daily Living (ADL) score [15], hemoglobin [16], low-density lipoprotein cholesterol (LDL) [17], total cholesterol [18], depression, and cardiac function classification [19], which were incorporated into an LR model and four ML algorithms (XGBoost, RF, LightGBM, and AdaBoost). Among them, AdaBoost showed the best performance, with an Area Under the Curve (AUC) of 0.803 in the validation set, demonstrating its robustness as a tool for clinical frailty risk assessment. Consistent with studies highlighting lipoprotein (a) as a cardiovascular risk marker, this study applied a decision tree ML model (Chi-square automatic interaction detection) to a multicenter registry of 2,301 patients with CHD or at high risk. Six clinical clusters were identified based on LDL cholesterol, CHD presence, family history, and age [20],[21]; these clusters were subsequently regrouped into three categories with progressive increases in Lp (a) and statistically significant differences $p < 0.001$. Consequently, the findings indicate that clinical clustering is an effective tool to stratify risk associated with elevated Lp (a) levels [22],[23]. Focusing on individuals aged 10 to 15 years, one study aimed to predict mortality from CHD using ML techniques [24]. The most relevant variables were selected, and based on them, various ML algorithms were trained, including LR [25], Support Vector Machines (SVM) [26], k-Nearest Neighbors (KNN) [27], Extra Trees Classifier (ETC) [28], and RF [29]. To ensure the validity of the results, 10-fold cross-validation was applied. The analysis showed that the KNN algorithm (k = 18) achieved the best performance, reaching an accuracy of 76.05%, an AUC of 79.41%, sensitivity of 81.08%, specificity of 71.28%, and precision of 73.48%.

Depression is also a key factor in the development of CHD, and in this context, ML emerges as an effective predictive tool [30],[31]. Various algorithms, including LR, RF, gradient boosting, SVM, XGBoost, decision trees, KNN, and neural networks, have been applied to improve risk estimation. Consequently, the RF model demonstrated the best performance, with an AUC of 0.987 in training and 0.996 in validation, as well as adequate calibration and an AUC of 0.928 after internal validation, confirming its high predictive capacity for assessing CHD risk in patients with depression. Another relevant aspect in the study of cardiopathy is exploring the feasibility of integrating ML algorithms for myocardial ischemia diagnosis. In a retrospective analysis of images from 206 patients, [32],[33] applied ML models such as XGBoost [34], RF [35], and logistic regression [36], all achieving accuracy levels above 0.80. Results showed that XGBoost performed best, with an accuracy of 0.903, an F1 score of 0.774, and an AUC of 0.931, demonstrating its high diagnostic potential for detecting myocardial ischemia.

A key element in cardiopathy diagnosis is the combination of artificial intelligence expert systems with neural networks, enabling early detection of at-risk patients and optimizing clinical decision-making. Data from the Framingham Heart Study, comprising 78,001 records, were used to construct a medical database. ML models such as RF, Decision Trees, and Neural Networks were implemented [37]. Consequently, the Decision Tree model achieved an accuracy of 90.08%, while the Neural Network reached 84.56%, demonstrating their capacity to predict the progression of coronary heart disease. Another study highlighted that cardiac disease classification improves the efficiency of medical decision support systems, while reducing associated diagnostic costs. Seven ML algorithms were evaluated, including LR, SVM, KNN, RF, decision trees, Naïve Bayes, and Gradient Boosting Classifier [38]. Hyperparameter optimization techniques such as Grid Search, Random Search, and Bayes Search were applied to maximize model performance [39]. Consequently, RF achieved the best performance, with an average accuracy of 92.85%, reaching up to 94.96% when ensemble methods were applied, evidencing its reliability and efficacy for cardiac disease classification in clinical settings.

Authors in studies such as [40],[41] developed ML models for predicting coronary disease in individuals with periodontitis. This study analyzed data from 3,245 patients, randomly dividing the sample into training and validation sets and considering variables such as age, race, history of myocardial infarction, chest pain, use of lipid-lowering medication, and serum levels of uric acid and creatinine. Five ML algorithms were trained: LR, Gradient Boosting Machine (GBM), SVM, KNN, and Classification and Regression Tree (CART), among which the KNN model showed the best performance, achieving an AUC of 0.977. Calibration metrics, Brier score, and decision curve analysis (DCA) confirmed the model's accuracy and clinical applicability, demonstrating the effectiveness of ML in predicting coronary disease in patients with periodontitis. Early detection of CHD has also been addressed using deep learning and ML techniques [42], employing seven classifiers, including KNN, SVM, LR, Convolutional Neural Networks, Gradient Boost, XGBoost, and RF. Class imbalance handling strategies were applied, with particular optimization of the XGBoost model [43], which achieved outstanding performance: 98.50% accuracy, 99.14% precision, 98.29% sensitivity, and 98.71% F1-score, demonstrating high efficacy for early cardiovascular disease identification. Continuing these investigations, the identification of congenital heart disease patients in large hospital databases involved ML analysis of 19,187 patients, of whom 3,784 were confirmed as CHD [44],[45]. Various algorithms, including Gradient Boosting Decision Tree, SVM, and Decision Tree, were evaluated and compared with regularized LR, using area under the precision-recall curve (AUPRC) as the main metric [46]. External validation with data updated until 2010 showed that the Gradient Boosting Decision Tree model performed best, with an AUPRC of 99.3%, sensitivity of 98.0%, and specificity of 99.7%, demonstrating the ability of ML methods to efficiently automate the identification of patients with complex diseases like CHD. In summary, cardiac disease enables the identification of high-risk patients and guides more effective prevention strategies. Likewise, cardiac disease prediction has been explored using RF-based classifiers to identify the most relevant features associated with the disease [47],[48]. A recent study employed an RF classifier optimized via hyperparameter tuning using a grid search approach. This model was evaluated in terms of accuracy [49], error rate, and recall [50], and compared with a traditional system. Results showed that while the traditional system achieved accuracies between 81.97% and 90.16%, the hyperparameter-tuned model reached higher accuracies, ranging from 84.22% to 96.53%, demonstrating that ML-based methodology allows for more precise prediction of cardiac disease by optimally identifying the most relevant features.

In summary, the authors proposed an innovative approach for cardiac disease prediction using the Grey Wolf Optimization (GWO) algorithm combined with stacked ensemble techniques to enhance diagnostic accuracy [51]. A patient database was used to evaluate model performance compared to traditional methods. The model demonstrated significantly superior performance, achieving 93% accuracy, with strong metrics in precision (91%), sensitivity (95.3%), F1-score (92.9%), a Matthews correlation coefficient of 0.83, and reduced Log Loss of 2.87, confirming it as a more reliable and effective alternative for early cardiac condition diagnosis. Additionally, the study incorporated the Multi-Criteria Decision-Making (MCDM) approach using the Combined Compromise Solution (CoCoSo) technique to comprehensively evaluate different ML models applied to cardiac disease prediction [52]. This procedure identified LR and SVM as the most consistent and effective algorithms across three clinical datasets, demonstrating superior performance compared to other alternatives [53]. Evaluation was based on six performance metrics, confirming that both models offer greater robustness and reliability for constructing predictive systems aimed at early cardiac disease diagnosis. Similarly, identifying the most suitable model for addressing cardiovascular diseases is a significant challenge in healthcare, motivating the application of ensemble methods and deep learning frameworks on a test dataset obtained from the Kaggle platform [54]. Key predictive variables such as age, gender, cholesterol levels, blood pressure, and lifestyle-related factors were considered. The evaluation showed that both ensemble methods and deep learning outperformed traditional models, achieving higher levels of accuracy, sensitivity, and area under the ROC curve (AUC-ROC), highlighting their potential as more effective tools for early cardiovascular disease diagnosis. Despite these advances, most of the reviewed studies present limitations such as small sample sizes, lack of external validation, and limited application in Latin American populations, restricting the generalizability of the models.

B. Theoretical Bases

1) ML algorithms: ML has established itself as a rapidly expanding scientific discipline, driven by the design of advanced algorithms and its ability to uncover hidden relationships within complex datasets. Through the use of mathematical models and statistical approaches, this field enables computational systems to autonomously adjust their performance, emulating human cognitive processes [55],[56]. Its growing importance lies in its potential to contribute significantly to the generation of new knowledge and the strengthening of decision-making processes across multiple application domains [57]. However, building ML models requires the implementation of concepts derived from specialized programming languages, among which Python [58] stands out for its versatility and widespread adoption in the scientific community. Table I presents the most representative algorithms in the field of ML, highlighting their main characteristics according to different analytical criteria. Consequently, its relevance is based on the ability to handle processes and meet needs autonomously, providing value without requiring direct human intervention.

2) Coronary heart disease: CHD is defined as a cardiovascular disease caused, in most cases, by atherosclerosis of the coronary arteries. This pathological process involves the progressive accumulation of lipids, cholesterol, and other deposits on the arterial wall, leading to the narrowing or blockage of blood flow to the myocardium. As a result, the heart receives an insufficient supply of oxygen and nutrients, which can clinically manifest as angina pectoris, arrhythmias, or acute myocardial infarction. It is important to note that coronary heart disease is the leading cause of death in developed countries, and its impact is closely associated with the increase of risk factors such as hypertension [59], obesity [60], and sedentary lifestyle [61]. Furthermore, early detection through tools based on clinical data and artificial intelligence techniques represents a promising approach to improve the diagnosis and prevention of this disease.

Table II presents the percentage distribution by provinces and regions of Peru, ordered from highest to lowest value. It can be observed that the Constitutional Province of Callao tops the list with 18.2%, followed by Metropolitan Lima and the Department of Lima with 16.3% each. In the intermediate range, Lambayeque, Cajamarca, Ica, and Arequipa stand out, with values ranging from 15.1% to 16.0%. Conversely, the regions with the lowest percentages are Ucayali (8.3%), Huancavelica (9.2%), and Junín (9.5%), falling below the reference threshold of 14.2%. This classification allows a clear differentiation between regions with higher concentration and those with lower relative representation.

III. Methodology

A. Definition of the CRISP-DM Methodology

Regarding the definition of the research topic, the CRISP-DM methodology emerges as a highly viable option for its application [63]. This methodological approach, widely accepted in data mining projects, provides a systematic and flexible framework that facilitates problem understanding, data analysis, and the orderly construction of predictive models. Furthermore, several authors highlight that the CRISP-DM methodology constitutes a fundamental tool for structuring data mining projects, as it enables systematic analysis of information and organizes the process into clearly defined

TABLE I. Comparison of ML Algorithms

Algorithm	Description	Advantages	Limitations
Logistic Regression	A statistical model used for binary classification problems, estimating the probability of a categorical outcome.	- Simple and interpretable - Efficient with linearly separable data - Low computational cost	- Limited to linear relationships - Sensitive to multicollinearity - Poor performance with complex data
Random Forest	An ensemble method based on multiple decision trees combined using bagging.	- Handles high-dimensional data - Reduces overfitting compared to single decision trees - Provides feature importance	- Less interpretable than single trees - Computationally intensive with large datasets - May require tuning of hyperparameters
XGBoost	A gradient boosting framework optimized for efficiency and accuracy in classification and regression tasks.	- High predictive performance - Handles missing values well - Scales efficiently to large datasets	- Complex to tune (many hyperparameters) - Higher computational cost than simpler models - Less interpretable than linear models

TABLE II. Percentage Distribution by Provinces and Regions in Peru [62].

Rank	Province/Region	%	Above/Below 14.2%
1	Constitutional Province of Callao	18.2	Above
2	Metropolitan Lima	16.3	Above
3	Department of Lima	16.3	Above
4	Lambayeque	16.0	Above
5	Cajamarca	15.5	Above
6	Ica	15.5	Above
7	Arequipa	15.1	Above
8	Tumbes	14.6	Above
9	Piura	14.3	Above
10	Moquegua	13.9	Below
11	Tacna	13.6	Below
12	La Libertad	13.4	Below
13	Cusco	11.5	Below
14	San Martín	11.3	Below
15	Áncash	11.2	Below
16	Amazonas	11.1	Below
17	Apurímac	10.7	Below
18	Ayacucho	10.4	Below
19	Pasco	10.3	Below
20	Huánuco	10.3	Below
21	Puno	10.2	Below
22	Madre de Dios	9.9	Below
23	Loreto	9.6	Below
24	Junín	9.5	Below
25	Huancavelica	9.2	Below
26	Ucayali	8.3	Below

phases: Business Understanding [64], Data Understanding [65], Data Preparation, Modeling, and Evaluation. This methodological framework not only provides an orderly and repeatable workflow but also allows each stage to be adapted to the specific characteristics of the data and the study objectives, as illustrated in Fig. 1. In health-related research, its application is particularly valuable, as it ensures process traceability and reinforces the validity of the generated predictive models. On the other hand, Fig. 2 describes a data analysis system for cardiac diseases driven by ML. The core process involves an ETL for enriched data ingestion, creating an interoperable database with clinical variables (such as blood pressure and BMI). The Data Assets manage the reuse of these resources through APIs. Data is stored in a Cardiac Disease Repository (with fragments and enterprise caching) and is subsequently processed and modeled using tools such as Python and R. The result is a trained ML catalog ready to be applied in various business scenarios.

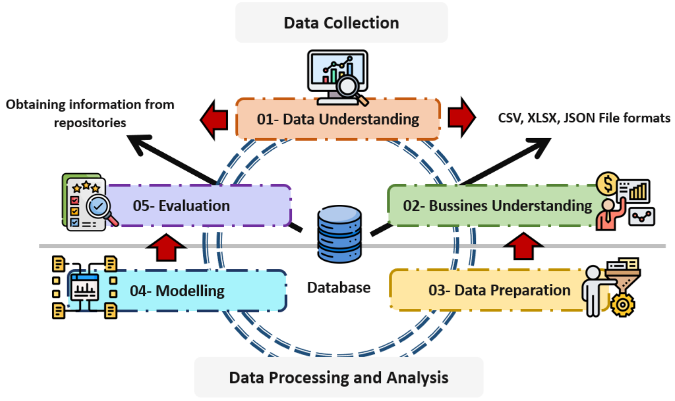

Data Collection
Obtaining information from repositories
01- Data Understanding
CSV, XLSX, JSON File formats
05- Evaluation
02- Bussines Understanding
Database
04- Modelling
03- Data Preparation
Data Processing and Analysis

Fig. 1. CRISP-DM methodology.

1) Business understanding: The initialization stage in model development consists of understanding the project's requirements based on the analysis of the problem to be addressed. It focuses on a deep comprehension of the situation and the clear formulation of objectives that will guide the subsequent development of the model. In this regard, Table III presents the Business Understanding phase of the study [66],[67],[68]. The main problem is identified as the high prevalence and late diagnosis of coronary heart disease. The business objective is to support early diagnosis through ML tools, while the project objective is to compare LR, RF, and XGBoost to determine the most efficient model. Success criteria encompass clinical, technical, and social aspects, also considering constraints such as data availability and risks like overfitting. Finally, expected benefits are highlighted, including support for early diagnosis, resource optimization, and reduction of cardiovascular mortality.

2) Data understanding: In this second stage, corresponding to the Data Understanding phase, a systematic process was carried out aimed at the initial exploration and analysis of the available data. First, the dataset was collected in its raw form, providing the necessary clinical and demographic information for the study. On one hand, Table IV defines the activities of the Data Understanding phase of CRISP-DM, which include dataset collection and description, exploration through statistical analyses and visualizations, quality verification to detect inconsistencies, analysis of relationships between variables, and generation of initial hypotheses, applied in this study to the use of clinical and demographic data to

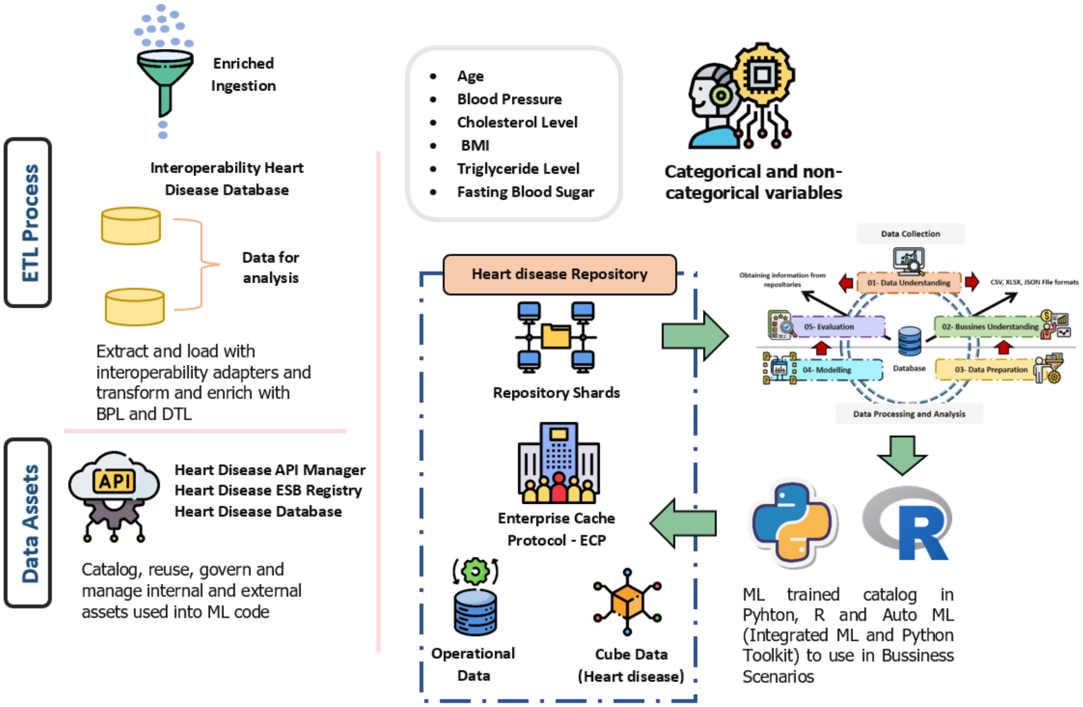

Enriched Ingestion
ETL Process
Interoperability Heart Disease Database
Data for analysis
Extract and load with interoperability adapters and transform and enrich with BPL and DTL
Data Assets
API
Heart Disease API Manager
Heart Disease ESB Registry
Heart Disease Database
Catalog, reuse, govern and manage internal and external assets used into ML code
Age
Blood Pressure
Cholesterol Level
BMI
Triglyceride Level
Fasting Blood Sugar
Categorical and non-categorical variables
Heart disease Repository
Repository Shards
Enterprise Cache Protocol - ECP
Operational Data
Cube Data (Heart disease)
Data Collection
Obtaining information from repositories
01- Data Understanding
CSV, XLSX, JSON File formats
05- Evaluation
02- Bussines Understanding
04- Modelling
Database
03- Data Preparation
Data Processing and Analysis
ML trained catalog in Pyhton, R and Auto ML (Integrated ML and Python Toolkit) to use in Bussiness Scenarios

Fig. 2. ML architecture.

TABLE III. Business Understanding Phase According to CRISP-DM

Element	Description
Business / Clinical Problem	Coronary heart disease is one of the leading causes of mortality worldwide. Its late diagnosis increases the risk of severe complications and places a significant burden on healthcare systems.
Business Objective	To contribute to the early diagnosis of coronary heart disease through decision-support tools based on ML, enabling optimization of healthcare resources and reduction of cardiovascular mortality.
Project Objective	To evaluate and compare the performance of three ML algorithms (LR, RF, and XGBoost) in predicting coronary heart disease, in order to identify the most efficient and clinically applicable model.
Success Criteria	- **Clinical**: support early identification of at-risk patients. - **Technical**: achieve satisfactory performance metrics (accuracy, sensitivity, specificity, AUC). - **Social**: provide evidence applicable to local contexts, with potential impact on reducing cardiovascular mortality.
Constraints	Availability and quality of clinical data, potential imbalance in class distribution, and computational limitations for training complex models.
Risks	Risk of model overfitting, bias in clinical data, and lack of generalization to other populations or healthcare settings.
Expected Benefits	Support for early diagnosis, optimization of hospital resources, reduction of complications and mortality, and contribution of evidence to the field of medical informatics.

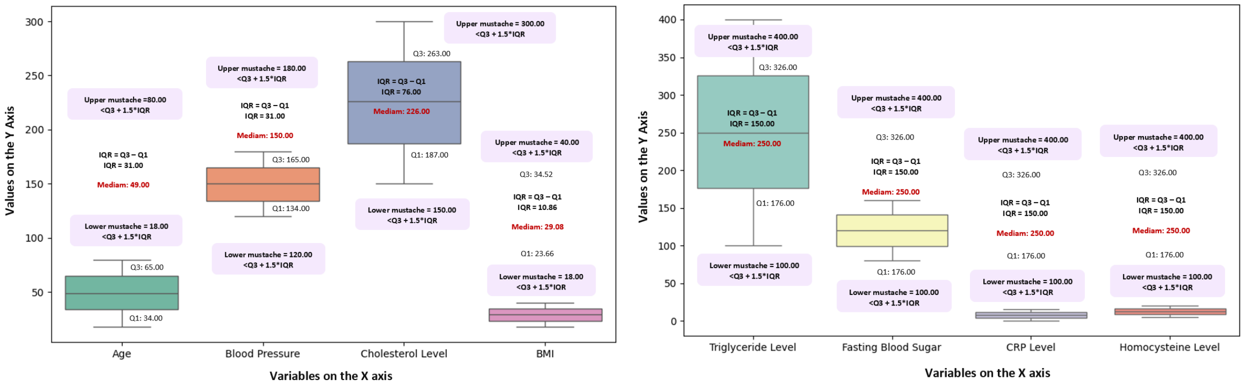
identify risk factors associated with heart disease [69],[70]. On the other hand, Fig. 3 shows box plots summarizing the distribution of various clinical and laboratory variables associated with heart disease. In Fig. 3a, direct risk factors such as age, blood pressure, cholesterol level, and body mass index (BMI) are presented, showing medians, quartiles, and dispersion ranges, with cholesterol and blood pressure exhibiting higher variability. In Fig. 3b, laboratory variables such as triglycerides, fasting glucose, C-reactive protein (CRP), and homocysteine are included, also showing outliers and greater dispersion in triglycerides, while fasting glucose exhibits a more concentrated range. These results allow the identification of central patterns and potential extreme values in the main factors associated with cardiovascular risk.

3) Data preparation: Table V presents the descriptive results of the variables included in the study: Table V(A) includes sociodemographic variables such as age and gender, which allow characterization of the population; Table V(B) covers relevant clinical variables such as blood pressure, cholesterol level, presence of diabetes, and body mass index, which are key factors in predicting heart disease; Table V(C) addresses habits and lifestyle factors, including alcohol and sugar consumption, stress level, and hours of sleep, which directly influence cardiovascular health; and Table V(D) presents the variable of interest, Heart Disease Status [71],[72], along with standardized measures of central tendency and dispersion, facilitating fair comparison across variables.

On the other hand, Table VI is divided into two sections: Table VI(A) General Information of the Dataset and Table VI(B) Missing Values per Column. The table confirms that the dataset contains a total of 21 attributes (20 predictor variables plus the target variable, Heart Disease

TABLE IV. ACTIVITIES OF THE DATA UNDERSTANDING PHASE IN CRISP-DM

Activity	Description	Application in the Study
Initial Data Collection	Obtain the dataset from the selected source and verify its structure.	Download of the dataset from Kaggle with clinical, demographic, and lifestyle variables.
Data Description	Identify the number of records, variables, and data types.	Review of 20 variables (Age, Gender, Blood Pressure, Cholesterol, Diabetes, etc.).
Data Exploration	Analyze relevant variables statistically and through visualizations.	Histograms of age, blood pressure, and cholesterol distribution; correlation plots.
Quality Verification	Detect missing, duplicate, outlier, or inconsistent values.	Identification of null values in clinical variables such as BMI and blood pressure.
Relationship Analysis	Examine links between attributes and the target variable.	Study of correlations between risk factors (cholesterol, diabetes, smoking) and *Heart Disease Status*.
Hypothesis Generation	Formulate initial assumptions to guide modeling.	Hypothesis: elevated cholesterol levels and hypertension are associated with a higher risk of coronary heart disease.



(a) Direct risk factors.

(b) Laboratory variables.

Fig. 3. Measures of central tendency according to dimensions.

Status), including both numerical (float64) and categorical (object) data types. However, the absence of the Statistical Summary section in the image prevents verification of claims regarding the mean age of 49 years, the average blood pressure of 149 mmHg, and cholesterol of 225 mg/dl, as well as the purported lack of calculated statistics for the BMI and Triglycerides variables.

4) Modelling: In this stage, predictive models were built using the data prepared in previous phases. For this purpose, different ML algorithms were selected and trained using the training and validation datasets. Fig. 4 presents the distribution of prediction errors for three classification models. In Fig. 4a, the performance of *RF* is shown, where errors are more concentrated around zero, indicating good predictive capability. In Fig. 4b, corresponding to *LR*, a greater dispersion of errors is observed, reflecting lower accuracy compared to the other methods [73]. Finally, Fig. 4c illustrates the distribution for *XGBoost*, which, like *RF*, keeps most errors close to zero, although with slight asymmetry. In all cases, the shaded areas highlight the most critical errors: false negatives on the left and false positives on the right, making it evident that the tree-based models (a) and (c) provide better control of these errors compared to the linear model (b).

B. Mathematical Formulas in ML Algorithms

The following formulas are commonly used in classical ML algorithms, such as LR, RF, and XGBoost. They formally describe the underlying mechanisms, optimization, and evaluation metrics of these models [74][75].

1) Logistic regression: Probability function: The logistic function models the probability of the positive class $Y = 1$ given predictors X [see Eq. (1)]. It transforms a linear combination of input features into a value between 0 and 1, representing probability.

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}} \tag{1}$$

Transforms a linear combination of predictors into a probability between 0 and 1.

2) Logistic regression: Log-loss function: The log-loss function is minimized to estimate the coefficients β [see Eq. (2)]. It penalizes predictions that are confident but incorrect, guiding the optimization of model parameters.

$$\mathcal{L}(\beta) = -\sum_{i=1}^{m} \Big[y_i \log(P(Y_i = 1 | X_i)) + (1 - y_i) \log(1 - P(Y_i = 1 | X_i)) \Big] \tag{2}$$

Penalizes incorrect predictions according to their probability, guiding parameter estimation.

TABLE V. Statistical Summary of Processed Variables

(a) Direct risk factors

Variable	count	mean	std	min	25%	50%	75%	max
Age	9971	-0.00	1.00	-1.72	-0.84	-0.02	0.86	1.69
Gender	10000	0.50	0.50	0.00	0.00	1.00	1.00	2.00
Blood Pressure	9981	-0.00	1.00	-1.69	-0.90	0.01	0.87	1.72
Cholesterol Level	9970	0.00	1.00	-1.71	-0.88	0.01	0.86	1.69
Diabetes	10000	0.50	0.51	0.00	0.00	0.00	1.00	2.00
BMI	10000	0.00	1.00	-1.72	-0.86	0.00	0.86	1.69
Smoking	10000	0.50	0.51	0.00	0.00	0.00	1.00	2.00
Family Heart Disease	10000	0.50	0.51	0.00	0.00	0.00	1.00	2.00

(b) Laboratory Variables

Variable	count	mean	std	min	25%	50%	75%	max
Triglyceride Level	9974	0.00	1.00	-1.72	-0.86	0.00	0.86	1.69
Fasting Blood Sugar	9978	0.00	1.00	-1.68	-0.88	0.01	0.88	1.68
CRP Level	9974	0.00	1.00	-1.72	-0.86	0.00	0.86	1.69
Homocysteine Level	9980	0.00	1.00	-1.72	-0.86	0.00	0.86	1.69

(c) Lifestyle Variables

Variable	count	mean	std	min	25%	50%	75%	max
Alcohol Consumption	10000	1.49	1.13	0.00	0.00	1.00	2.25	3.00
Stress Level	10000	1.02	0.82	0.00	0.00	1.00	2.00	3.00
Sleep Hours	9975	0.00	1.00	-1.68	-0.88	0.01	0.88	1.68
Sugar Consumption	10000	1.00	0.82	0.00	0.00	1.00	2.00	3.00

(d) Target Variable

Variable	count	mean	std	min	25%	50%	75%	max
Heart Disease Status	10000	0.20	0.40	0.00	0.00	0.00	0.00	1.00

3) Logistic regression: Odds ratio: The odds ratio expresses the effect of each predictor on the likelihood of $Y = 1$ [see Eq. (3)]. It quantifies how a unit change in a predictor multiplies the odds of the positive outcome, providing interpretability of model coefficients.

$$\text{Odds} = \frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta_0+\sum_{i=1}^{n}\beta_i X_i} \tag{3}$$

Represents the multiplicative change in odds for a unit change in each predictor.

4) Logistic regression: Regularized loss function (ridge): Regularization prevents overfitting by penalizing large coefficients [see Eq. (4)]. It adds an $L2$ penalty to the log-loss, encouraging smaller coefficients and improving the generalization ability of the model.

$$\mathcal{L}_{reg}(\beta) = \mathcal{L}(\beta) + \lambda \sum_{i=1}^{n} \beta_i^2 \tag{4}$$

Adds an $L2$ penalty to the log-loss, controlling model complexity.

5) Random forest: Classification prediction: For classification, the final prediction $\hat{y}$ is obtained by majority voting of B trees [see Eq. (5)]. Each tree casts a vote, and the class with the most votes is selected as the final prediction, improving robustness against individual tree errors.

$$\hat{y} = \text{mode}\{h_1(X), h_2(X), \ldots, h_B(X)\} \tag{5}$$

Each tree votes, and the most frequent class is chosen as the final prediction.

6) Random forest: Regression prediction: For regression problems, the prediction is the average of all B trees [see Eq. (6)]. Averaging reduces variance and improves the stability and generalization of the model, making it less sensitive to individual tree fluctuations.

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(X) \tag{6}$$

Reduces variance and improves generalization by averaging multiple models.

7) Random forest: Gini impurity: The Gini index measures the impurity of a node and is used for splitting [see Eq. (7)]. Lower Gini values indicate more homogeneous nodes, which helps the tree create purer child nodes and improve classification accuracy.

$$Gini = 1 - \sum_{k=1}^{K} p_k^2 \tag{7}$$

TABLE VI. SUMMARY AND CHARACTERISTICS OF THE DATASET

(A) GENERAL INFORMATION OF THE DATASET

Column	Non-null values	Data type
Age	9971	float64
Gender	9981	object
Blood Pressure	9981	float64
Cholesterol Level	9970	float64
Exercise Habits	9975	object
Smoking	9975	object
Family Heart Disease	9979	object
Diabetes	9970	object
BMI	9978	float64
High Blood Pressure	9974	object
Low HDL Cholesterol	9975	object
High LDL Cholesterol	9974	object
Alcohol Consumption	7414	object
Stress Level	9978	object
Sleep Hours	9975	float64
Sugar Consumption	9970	object
Triglyceride Level	9974	float64
Fasting Blood Sugar	9978	float64
CRP Level	9974	float64
Homocysteine Level	9980	float64
Heart Disease Status	10000	object

(B) MISSING VALUES PER COLUMN

Column	Missing values
Age	29
Gender	19
Blood Pressure	19
Cholesterol Level	30
Exercise Habits	25
Smoking	25
Family Heart Disease	21
Diabetes	30
BMI	22
High Blood Pressure	26
Low HDL Cholesterol	25
High LDL Cholesterol	26
Alcohol Consumption	2586
Stress Level	22
Sleep Hours	25
Sugar Consumption	30
Triglyceride Level	26
Fasting Blood Sugar	22
CRP Level	26
Homocysteine Level	20
Heart Disease Status	0

Minimized to create more homogeneous child nodes during tree construction.

8) Random forest: Feature importance: Feature importance is calculated based on the total reduction in impurity [see Eq. (8)]. It quantifies how much each feature contributes to improving prediction accuracy, allowing interpretability and identification of key variables.

$$FI_j = \sum_{t=1}^{B} \sum_{\text{splits } s \text{ in tree } t} \Delta i(s) \cdot \mathbf{1}_{\{j \text{ used in } s\}} \tag{8}$$

Summarizes how much each feature contributes to prediction accuracy.

9) XGBoost: Prediction function: The final prediction is the sum of contributions from K sequential trees [see Eq. (9)]. Each tree iteratively corrects the errors of previous trees, allowing the model to capture complex patterns and improve overall predictive performance.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in \mathcal{F} \tag{9}$$

Each tree f_k corrects the errors of previous trees.

10) XGBoost: Regularized objective: XGBoost minimizes a regularized objective combining loss and complexity of trees [see Eq. (10)]. This approach balances predictive accuracy with model complexity, preventing overfitting while ensuring robust performance on unseen data.

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

Prevents overfitting while minimizing prediction error.

11) XGBoost: Gradient update: Each iteration updates predictions using the gradient and learning rate η [see Eq. (11)]. This sequential adjustment allows the model to correct previous errors gradually, improving convergence and prediction accuracy while controlling the step size.

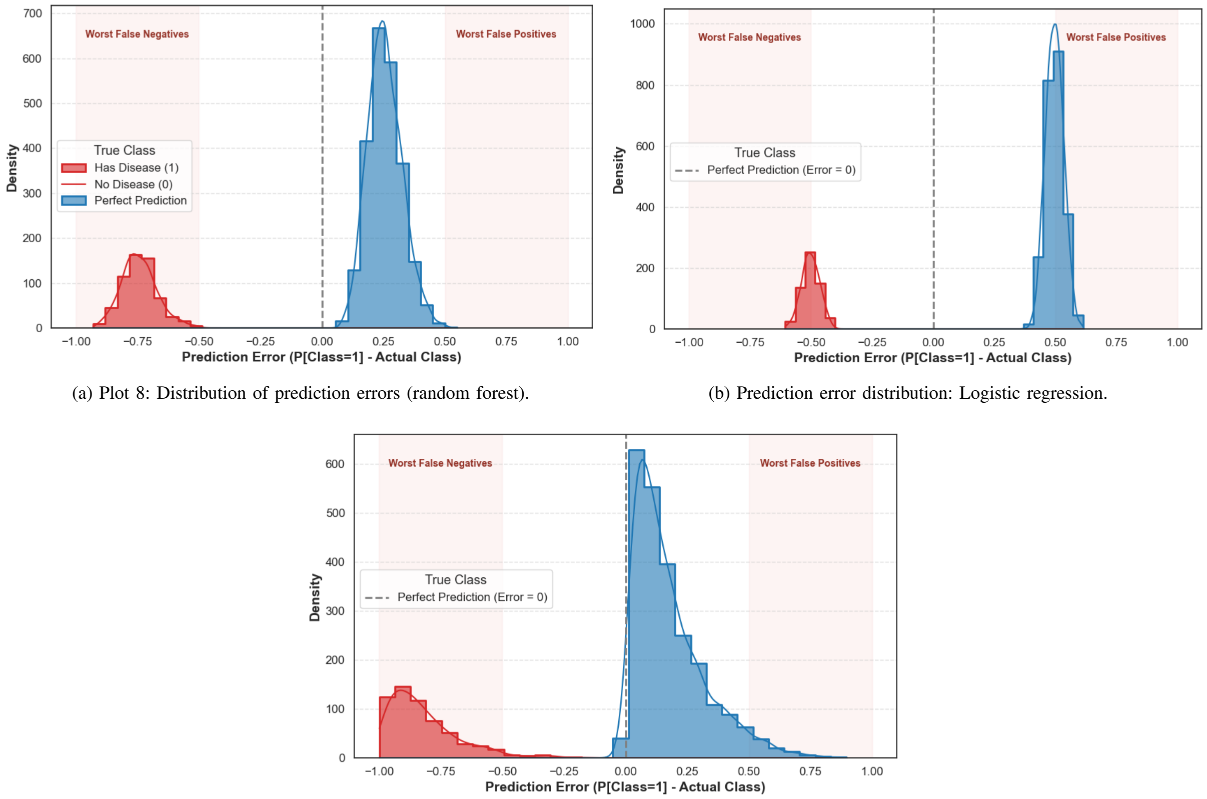
$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(X_i) \tag{11}$$

Sequentially corrects previous errors, controlling step size via η.

12) XGBoost: Second-order approximation: The second-order Taylor expansion allows efficient tree optimization [see Eq. (12)]. By using both first and second derivatives of the loss function, it provides precise gradient-based updates that accelerate learning and improve model accuracy.

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] + \Omega(f_t) \tag{12}$$

Uses first (g_i) and second (h_i) derivatives of the loss function for precise updates.



(a) Plot 8: Distribution of prediction errors (random forest).

(b) Prediction error distribution: Logistic regression.

(c) Prediction error distribution: XGBoost.

Fig. 4. Comparative analysis of the impact of class balancing and model diagnostics.

13) XGBoost: Leaf weight optimization: Optimal leaf weights are computed to minimize the regularized loss [see Eq. (13)]. This calculation ensures that each leaf contributes optimally to reducing the objective function while accounting for regularization, enhancing model stability and predictive performance.

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{13}$$

Assigns optimal output values to leaves of the tree considering regularization.

14) XGBoost: Split gain: The gain of a split measures its improvement in the objective function [see Eq. (14)]. Higher gain values indicate more effective splits, guiding the tree to partition data in a way that maximizes predictive accuracy and model efficiency.

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{14}$$

Quantifies how much a split improves the objective, guiding tree growth.

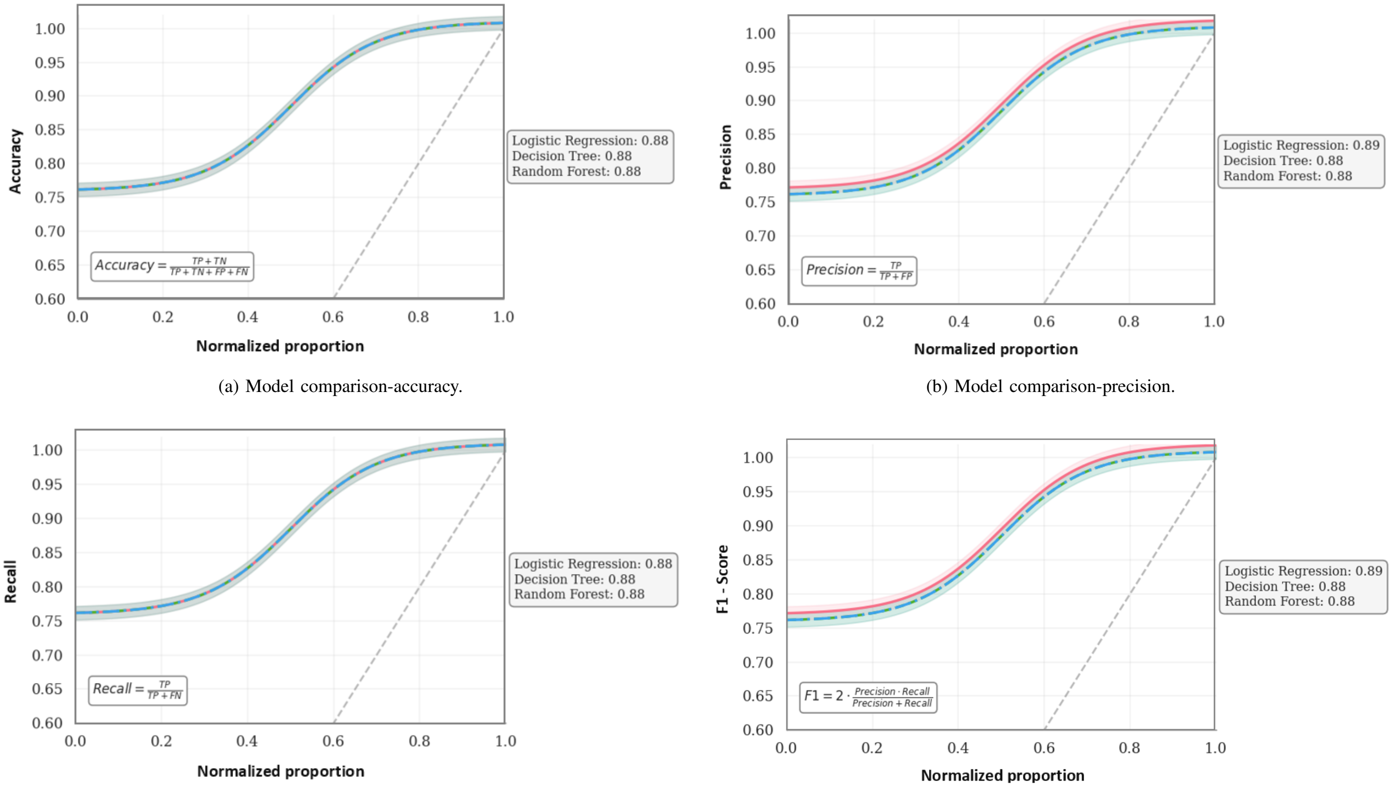
IV. Results

This section presents the results obtained from the comparison of metrics for the algorithms used, aiming to evaluate the performance of each model and select the one that offers the best results. Additionally, a comparison between different ML methodologies is carried out, highlighting their main characteristics and evaluation parameters. From this analysis, it was determined that the most suitable methodology for the development of the present study is CRISP-DM, as it provides a structured and flexible approach that facilitates both the construction and interpretation of the models.

A. Evaluation of Results

This section presents a comparative analysis of the predictive performance of three machine learning architectures: LR, RF, and XGBoost. The models were rigorously evaluated using standard binary classification metrics (Accuracy, Precision, Sensitivity (Recall), and F1-score), with the objective of predicting the presence of coronary artery disease (CAD) based on a standardized set of clinical and health habit features.

The XGBoost model consistently demonstrated superior performance across all evaluation metrics (Fig. 5). Its inherent ability to model nonlinear relationships and complex interactions translated into the best discrimination of the positive class, achieving the highest Accuracy ($\approx 91.2\%$) and the highest F1-score (F1 $\approx$ 90.8%) in Fig. 5d among the



(a) Model comparison-accuracy.

(b) Model comparison-precision.

(c) Model comparison-recall.

(d) Model comparison-F1-score.

Fig. 5. Comparison of ML models by performance metrics.

models. This result underscores the robustness of XGBoost and positions it as the reference model in this study. The RF algorithm exhibited notably competitive performance, maintaining a balanced trade-off between Precision (Fig. 5b) and Sensitivity (Fig. 5c). Its metrics remained very close to the top-performing model (Acc $\approx 90.5\%$, F1 $\approx 90.1\%$) (Fig. 5a), validating the effectiveness of bagging techniques in achieving strong generalization and mitigating the risk of overfitting without sacrificing predictive power. In contrast, the LR model, due to its inherently linear nature, yielded more modest values across all metrics (Acc $\approx 87.8\%$, F1 $\approx 86.9\%$). While LR provides a fundamental interpretative baseline, its performance was predictably outperformed by the tree ensemble models, reiterating its limitations in capturing the inherent complexity of the physiological phenomena underlying coronary disease.

B. Comparison of Methodologies

For the comparison of methodologies, Table VII presents a comparative evaluation of four widely recognized data mining methodologies: CRISP-DM, Knowledge Discovery in Databases (KDD), Sample, Explore, Modify, Model, Assess (SEMMA), and Analytics Solutions Unified Method for Data Mining (ASUM-DM). Each was analyzed under five evaluation criteria: clear structure and phases, practical applicability, integration with machine learning, reproducibility, and interpretability. The results, expressed as percentages, show that the CRISP-DM methodology achieves the highest performance across all evaluated aspects, with an approximate general average of 93%, standing out for its methodological clarity, ability to replicate results, and ease of communicating findings. In contrast, the KDD and SEMMA methodologies exhibit a more technical but less integrative approach, while ASUM-DM maintains a balanced performance albeit with less formal structuring. Overall, the results support the selection of CRISP-DM as the most robust and adaptable methodological framework for studies predicting cardiovascular diseases using machine learning algorithms.

V. Discussion

Previous research has identified multiple causes and characteristics associated with heart diseases within the medical field. In this context, the results obtained in the present study are consistent with trends described in recent literature, where ensemble-based algorithms, particularly XGBoost and Random Forest, have demonstrated superior performance in predicting cardiovascular diseases. In this study, the XGBoost model achieved the highest values across all evaluation metrics (*accuracy, precision, recall* and *F1-score*), confirming its ability to model nonlinear interactions between clinical variables and lifestyle-related factors. These findings align with those reported by [32], who demonstrated that XGBoost provides an optimal balance between sensitivity and specificity in detecting cardiac complications and identifying myocardial ischemia, reaching AUC values above 0.90. Likewise, the RF model showed competitive performance, particularly in the *precision* and *recall* metrics, suggesting adequate

TABLE VII. Comparative Evaluation of Methodologies Based on Performance

Evaluation Criteria	Criterion Description	CRISP-DM	KDD	SEMMA	ASUM-DM
Clear Structure and Phases (%)	Assesses the clarity, sequentiality, and organization of methodological stages.	**95**	88	82	90
Practical Applicability (%)	Measures the ease of implementation in real-world data mining projects.	**93**	85	87	89
Integration with ML (%)	Evaluates compatibility with modeling and validation processes for machine learning algorithms.	**90**	80	86	88
Reproducibility (%)	Analyzes the ability to replicate results and maintain consistency across different runs.	**92**	83	80	86
Interpretability (%)	Assesses the transparency of results and ease of communicating findings to different audiences.	**94**	82	78	84
Approximate general average (%)	Average value of overall performance considering the previous criteria.	**93**	84	83	87

generalization capability and lower susceptibility to overfitting. This behavior is consistent with the observations highlighted by [30], [38], and [47], who emphasized the efficiency of Random Forest in classifying heart diseases and its superiority over traditional methods such as LR. These studies revealed that RF achieved precision above 92% and AUC values close to 0.99, results comparable to those obtained in the present work.

In contrast, LR exhibited moderate performance, which was expected given its linear nature and lower ability to capture complex relationships among predictor variables. Nevertheless, its utility lies in providing an interpretable and clinically transparent baseline, an aspect also emphasized by [14] and [53], who acknowledge that despite its limitations, this model retains value in scenarios where interpretability is prioritized over predictive complexity. Moreover, the superiority of XGBoost in this study can also be attributed to its ability to integrate regularization mechanisms and sequential optimization, characteristics that [43] and [20] noted allow maintaining high accuracy even in clinical datasets with certain imbalance or noise. This behavior reinforces the idea that boosting models are more suitable for complex biomedical contexts, where relationships between predictors and outcome variables are not strictly linear.

More broadly, the results of this work confirm observations reported by several authors [51], [52], who reported that combining optimization techniques (such as Grey Wolf Optimization or CoCoSo) with ensemble algorithms significantly increases diagnostic accuracy, outperforming traditional models. Although the present study did not implement metaheuristic strategies, the robustness of the XGBoost model suggests that its incorporation in later stages could further enhance the system's predictive capability. Consequently, the findings align with the general assertion in the literature that ML methods—particularly those based on boosting and bagging—offer a more favorable balance between accuracy, generalization, and stability for the diagnosis and prediction of CHD [39], [19]. Therefore, the XGBoost model consolidates as the most suitable alternative for future external validation and integration into medical decision support systems, aligning with the current trend of incorporating artificial intelligence into clinical practice to strengthen early diagnosis and personalized cardiovascular risk management.

VI. Conclusion

Faced with the significant challenges posed by machine learning in the application of algorithms aimed at the diagnosis and prediction of heart diseases, researchers have intensified efforts to develop increasingly accurate, robust, and interpretable models. These models, resulting from the interaction between clinical, genetic, and lifestyle factors, demand computational approaches capable of capturing nonlinear relationships and hidden patterns in medical data. In this context, the present study conducted a comparison of various data mining methodologies oriented toward the development of predictive models in the field of cardiopathy. The approaches analyzed included KDD, SEMMA, ASUM-DM, and CRISP-DM, considering technical, methodological, and applicability criteria. The evaluation results indicated that the CRISP-DM methodology stood out as the most efficient and adaptable to the study context, due to its iterative structure, systematic approach to data understanding, and alignment with the analytical objectives of ML.

In line with the above, the CRISP-DM methodology is structured into five fundamental phases that guide the entire knowledge discovery process from data. The first phase, Business Understanding, establishes the analytical objectives and expected outcomes of the predictive model, taking into account the specific clinical context. The second phase, Data Understanding, enables the characterization of variables through descriptive statistics, identifying their structure and dimensionality without the need to examine each individual observation. The third phase, Data Preparation, encompasses the cleaning, transformation, and normalization of information, correcting missing, inconsistent, or outlier values that could affect modeling quality. In the fourth phase, Modeling, machine learning algorithms are trained with the processed data, enabling the identification of predictive patterns and the reduction of prediction errors. Finally, the fifth phase, Evaluation, compares the generated models using performance metrics such as Accuracy, Precision, Recall, and F1-Score, selecting the one with the greatest generalization capacity and predictive efficacy. The results obtained in this study support the effectiveness of boosting and bagging-based models, such as XGBoost and RF, in predicting cardiovascular diseases. These techniques demonstrated a superior balance between accuracy and generalization capacity compared to LR, which is critical in clinical contexts where classification errors can have significant consequences. In particular, the

XGBoost model emerges as the most promising alternative for application in later stages of the study, including external validation and integration into medical decision support systems, thus contributing to more preventive, personalized, and evidence-based care.

Finally, understanding the implications of coronary heart disease is essential to visualize its causes and guide the implementation of effective technological solutions. Based on the findings obtained through our model, it is suggested to complement the evaluated algorithms with additional technologies, such as deep learning models, integration of medical imaging data, or time-series analysis of vital signs, in order to improve prediction accuracy and efficacy. Future studies could explore the combination of these techniques with clinical decision support systems, as well as the validation of the model across different population contexts, strengthening both medical informatics and the practical application of early diagnostic tools, with a potential impact on reducing cardiovascular mortality and optimizing healthcare resources.

Acknowledgment

Thanks to the University of Sciences and Humanities, which made this project a reality.

References

[1] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, Álvaro García. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, M. S. V. Elkind, K. R. Evenson, C. Eze-Nliam, J. F. Ferguson, G. Generoso, J. E. Ho, R. Kalani, S. S. Khan, B. M. Kissela, K. L. Knutson, D. A. Levine, T. T. Lewis, J. Liu, M. S. Loop, J. Ma, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, R. Poudel, M. Rezk-Hanna, G. A. Roth, E. B. Schroeder, S. H. Shah, E. L. Thacker, L. B. VanWagner, S. S. Virani, J. H. Voecks, N. Y. Wang, K. Yaffe, and S. S. Martin, "Heart disease and stroke statistics-2022 update: A report from the american heart association," *Circulation*, vol. 145, no. 8, pp. E153–E639, 2022, doi:10.1161/CIR.0000000000001052.

[2] M. Bhushan, A. Pandit, and A. Garg, "Machine learning and deep learning techniques for the analysis of heart disease: a systematic literature review, open challenges and future directions," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 10 493–10 536, 2023, doi:10.1007/s10462-023-10493-5.

[3] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized xgboost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, 2022, doi:10.1016/j.jksuci.2020.10.013.

[4] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Hdpm: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, no. —, pp. 133 034–133 050, 2020, doi:10.1109/ACCESS.2020.3010511.

[5] D. Larassati, A. Zaidiah, and S. Afrizal, "Sistem prediksi penyakit jantung koroner menggunakan metode naive bayes," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 2, pp. 294–302, 2022, doi:10.29100/jipi.v7i2.2842.

[6] V. Gujrati, S. Joshi, and S. Nagpal, "Logistic regression to predict heart disease," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 7, no. 10, 2023, doi:10.55041/ijsrem25957.

[7] H. E. Hamdaoui, S. Boujraf, N. E. H. Chaoui, B. Alami, and M. Maaroufi, "Improving heart disease prediction using random forest and adaboost algorithms," *International Journal of Online and Biomedical Engineering*, vol. 17, no. 11, 2021, doi:10.3991/ijoe.v17i11.24781.

[8] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-xgboost algorithm," *Information (Switzerland)*, vol. 13, no. 10, p. 475, 2022, doi:10.3390/info13100475.

[9] M. Abubakar, A. H. Maidabara, Y. M. Malgwi, and A. Mohammed, "Web based heart disease prediction model using machine learning technique," *Computer Science & IT Research Journal*, vol. 5, no. 2, 2024, doi:10.51594/csitrj.v5i2.837.

[10] K. Brown, P. Roshanitabrizi, J. Rwebembera, E. Okello, A. Beaton, M. G. Linguraru, and C. A. Sable, "Using artificial intelligence for rheumatic heart disease detection by echocardiography: Focus on mitral regurgitation," *Journal of the American Heart Association*, vol. 13, no. 2, p. e031257, 2024, doi:10.1161/JAHA.123.031257.

[11] S. Agrawal, "Revolutionizing cardiovascular health: A machine learning approach for predictive analysis and personalized intervention in heart disease," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 3, 2024, doi:10.22214/ijraset.2024.58797.

[12] H. E. S. Alsafi and O. N. Ocan, "Retraction note: A novel intelligent machine learning system for coronary heart disease diagnosis," *Applied Nanoscience*, vol. 14, no. 3, 2024, doi:10.1007/s13204-024-03012-7.

[13] B. M. R. International, "Retracted: An effective machine learning-based model for an early heart disease prediction," *BioMed research international*, vol. 2024, 2024, doi:10.1155/2024/9754362.

[14] C.-Y. Ma, Y.-M. Luo, T.-Y. Zhang, Y.-D. Hao, X.-Q. Xie, X.-W. Liu, X.-L. Ren, X.-L. He, Y.-M. Han, K.-J. Deng, D. Yan, H. Yang, H. Tang, and H. Lin, "Predicting coronary heart disease in chinese diabetics using machine learning," *Computers in Biology and Medicine*, vol. 169, no. 7, p. 107952, 2024, doi:10.1016/j.compbiomed.2024.107952, artículo en línea con identificador 107952; sin rango de páginas tradicional.

[15] M. M. Mijwil, A. K. Faieq, and M. Aljanabi, "Early detection of cardiovascular disease utilizing machine learning techniques: Evaluating the predictive capabilities of seven algorithms," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 1, pp. 18–32, 2024, doi:10.52866/ijcsm.2024.05.01.018.

[16] P. K. Mall, S. Srivastava, M. M. Patel, A. Kumar, V. Narayan, S. Kumar, P. K. Singh, and D. S. Singh, "Optimizing heart attack prediction through ohe2lm: A hybrid modelling strategy," *Journal of Electrical Systems*, vol. 20, no. 1, pp. 66–75, 2024, doi:10.52783/jes.665.

[17] G. P. Raju, “Multiple disease prediction system,” *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 8, no. 3, pp. 1–7, 2024, doi:10.55041/ijsrem28972.

[18] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, “Machine learning-based predictive models for detection of cardiovascular diseases,” *Diagnostics*, vol. 14, no. 2, p. 144, 2024, doi:10.3390/diagnostics14020144.

[19] A. J. Marelli, C. Li, A. Liu, H. Nguyen, H. Moroz, J. M. Brophy, L. Guo, D. L. Buckeridge, J. Tang, A. Y. Yang, and Y. Li, “Machine learning informed diagnosis for congenital heart disease in large claims data source,” *JACC: Advances*, vol. 3, no. 2, p. 100801, 2024, doi:10.1016/j.jacadv.2023.100801.

[20] X. Li, Z. Wang, W. Zhao, R. Shi, Y. Zhu, H. Pan, and D. Wang, “Machine learning algorithm for predict the in-hospital mortality in critically ill patients with congestive heart failure combined with chronic kidney disease,” *Renal Failure*, vol. 46, no. 1, p. 2315298, 2024, doi:10.1080/0886022X.2024.2315298.

[21] M. Duan, Y. Zhang, Y. Liu, B. Mao, G. Li, D. Han, and X. Zhang, “Machine learning aided non-invasive diagnosis of coronary heart disease based on tongue features fusion,” *Technology and Health Care*, vol. 32, no. 1, pp. 441–457, 2024, doi:10.3233/THC-230590.

[22] F. Türk, “Investigation of machine learning algorithms on heart disease through dominant feature detection and feature selection,” *Signal, Image and Video Processing*, vol. 18, no. 4, pp. 3943–3955, 2024, doi:10.1007/s11760-024-03060-0.

[23] B. Duraisamy, R. Sunku, K. Selvaraj, V. V. R. Pilla, and M. Sanikala, “Heart disease prediction using support vector machine,” *Multidisciplinary Science Journal*, vol. 6, no. 1, pp. 104–112, 2024, doi:10.31893/multiscience.2024ss0104.

[24] A. Kumar, A. K. Singh, and A. Garg, “Evaluation of machine learning techniques for heart disease prediction using multi-criteria decision making,” *Journal of Intelligent and Fuzzy Systems*, vol. 46, no. 1, pp. 1–15, 2024, doi:10.3233/JIFS-233443.

[25] G. Saranya and A. Pravin, “Grid search based optimum feature selection by tuning hyperparameters for heart disease diagnosis in machine learning,” *The Open Biomedical Engineering Journal*, vol. 17, no. 1, pp. 4371–4385, 2024, doi:10.2174/18741207-v17-e230510-2022-ht28-4371-8.

[26] L. J. V. N and A. Das, “Forecasting the risk of coronary heart diseases using machine learning algorithms,” *Computology: Journal of Applied Computer Science and Intelligent Technologies*, vol. 3, no. 2, pp. 45–56, 2024, doi:10.17492/computology.v3i2.2307.

[27] D. D. Solomon, Sonia, K. Kumar, K. Kanwar, S. Iyer, and M. Kumar, “Extensive review on the role of machine learning for multifactorial genetic disorders prediction,” *Archives of Computational Methods in Engineering*, vol. 31, no. 2, pp. 789–810, 2024, doi:10.1007/s11831-023-09996-9.

[28] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, “A clinical decision support system for heart disease prediction using deep learning,” *IEEE Access*, vol. 11, no. 1, pp. 1–15, 2023, doi:10.1109/ACCESS.2023.3285247.

[29] M. Shokouhifar, M. Hasanvand, E. Moharamkhani, and F. Werner, “Ensemble heuristic–metaheuristic feature fusion learning for heart disease diagnosis using tabular data,” *Algorithms*, vol. 17, no. 1, p. 34, 2024, doi:10.3390/a17010034.

[30] N. S. Gupta, S. K. Rout, S. Barik, R. R. Kalangi, and B. Swapna, “Enhancing heart disease prediction accuracy through hybrid machine learning methods,” *EAI Endorsed Transactions on Internet of Things*, vol. 10, no. 1, pp. 1–12, 2024, doi:10.4108/eetiot.5367.

[31] S. H. Mulyani, N. Wijaya, and F. Trinidya, “Enhancing heart disease detection using convolutional neural networks and classic machine learning methods,” *Journal of Computer, Electronic, and Telecommunication*, vol. 4, no. 2, pp. 394–405, 2024, doi:10.52435/complete.v4i2.394.

[32] S. Srinivasan, S. Gunasekaran, S. K. Mathivanan, B. A. Malar, P. Jayagopal, and G. T. Dalu, “An active learning machine technique based prediction of cardiovascular heart disease from uci-repository database,” *Scientific Reports*, vol. 13, no. 1, pp. 1–15, 2023, doi:10.1038/s41598-023-40717-1.

[33] A. Hammoud, A. Karaki, R. Tafreshi, S. Abdulla, and M. Wahid, “Coronary heart disease prediction: A comparative study of machine learning algorithms,” *Journal of Advances in Information Technology*, vol. 15, no. 1, pp. 27–32, 2024, doi:10.12720/jait.15.1.27-32.

[34] C. Bernand, E. Mirand, and M. Aryun, “Coronary heart disease prediction models using machine learning and deep learning algorithms,” *AIP Conference Proceedings*, vol. 2838, no. 1, p. 020012, 2024, doi:10.1063/5.0179929.

[35] A. A. Maulani, S. Winarno, J. Zeniarja, R. T. E. Putri, and A. N. Cahyani, “Comparison of hyperparameter optimization techniques in hybrid cnn-lstm model for heart disease classification,” *Sinkron*, vol. 9, no. 1, p. 13219, 2024, doi:10.33395/sinkron.v9i1.13219.

[36] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, “Classification models combined with boruta feature selection for heart disease prediction,” *Informatics in Medicine Unlocked*, vol. 44, p. 101442, 2024, doi:10.1016/j.imu.2023.101442.

[37] A. H. Najim and N. Nasri, “Artificial intelligence for heart disease prediction and imputation of missing data in cardiovascular datasets,” *Cogent Engineering*, vol. 11, no. 1, pp. 1–18, 2024, doi:10.1080/23311916.2024.2325635.

[38] B. S.P. and D. M., “An optimized deep auto encoder with enhanced extreme learning machine model for heart disease prediction and classification,” *International Journal of System of Systems Engineering*, vol. 16, no. 2, pp. 145–160, 2026, doi:10.1504/ijsse.2026.10062277.

[39] P. Rani, R. Kumar, A. Jain, R. Lamba, R. K. Sachdeva, K. Kumar, and M. Kumar, “An extensive review of machine learning and deep learning techniques on heart disease classification and prediction,” *Archives of Computational Methods in Engineering*, vol. 31, no. 6, pp. 3871–3895, 2024, doi:10.1007/s11831-024-10075-w.

[40] S. Mondal, R. Maity, Y. Omo, S. Ghosh, and A. Nag, “An efficient computational risk prediction model

of heart diseases based on dual-stage stacked machine learning approaches," *IEEE Access*, vol. 12, no. 1, pp. 3 350 996–3 351 010, 2024, doi:10.1109/ACCESS.2024.3350996.

[41] B. A. Majeed, A. Y. Hardan, B. Y. Hardan, and D. F. Munaf, "Accurate ai-based chatbot to diagnose heart diseases pre-human doctor consultation," *Revue d'Intelligence Artificielle*, vol. 38, no. 1, pp. 121–135, 2024, doi:10.18280/ria.380121.

[42] J. S. Lee, E.-S. Choi, Y. Hwang, K.-S. Lee, and K. H. Ahn, "Preterm birth and maternal heart disease: A machine learning analysis using the korean national health insurance database," *PLoS ONE*, vol. 18, no. 3, p. e0283959, 2023, doi:10.1371/journal.pone.0283959.

[43] N. R. Kolukula, P. N. Pothineni, V. M. K. Chinta, V. G. Boppana, R. P. Kalapala, and S. Duvvi, "Predictive analytics of heart disease presence with feature importance based on machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1070–1077, 2023, doi:10.11591/ijeecs.v32.i2.pp1070-1077.

[44] S. M. Muhammed, G. Abdul-Majeed, and M. S. Mahmoud, "Prediction of heart diseases by using supervised machine learning algorithms," *Wasit Journal for Pure Sciences*, vol. 2, no. 1, pp. 231–243, 2023, doi:10.31185/wjps.125.

[45] G. Murugesan, C. Kavitha, G. Jabakumar, and E. Swarnalatha, "Prediction of heart disease using machine learning algorithms with feature selection techniques," *CARDIOMETRY*, vol. 26, no. 26, pp. 778–786, 2023, doi:10.18137/cardiometry.2023.26.778786.

[46] S. Mishra, N. K. Tiwari, K. Kumari, and V. Kumawat, "Prediction of heart disease using machine learning," 2023, doi:10.1109/ICAAIC56838.2023.10140478.

[47] A. A. Ahmad and H. Polat, "Prediction of heart disease based on machine learning using jellyfish optimization algorithm," *Diagnostics*, vol. 13, no. 14, 2023, doi:10.3390/diagnostics13142392.

[48] P. Yu, M. Skinner, I. Esangbedo, J. J. Lasa, X. Li, S. Natarajan, and L. Raman, "Predicting cardiac arrest in children with heart disease: A novel machine learning algorithm," *Journal of Clinical Medicine*, vol. 12, no. 7, 2023, doi:10.3390/jcm12072728.

[49] G. A. Ansari, S. S. Bhat, M. D. Ansari, S. Ahmad, J. Nazeer, and A. E. Eljialy, "Performance evaluation of machine learning techniques (mlt) for heart disease prediction," *Computational and Mathematical Methods in Medicine*, vol. 2023, 2023, doi:10.1155/2023/8191261.

[50] S. Tuba, "Optimization heart disease prediction using machine learning models," *Fidelity : Jurnal Teknik Elektro*, vol. 5, no. 1, p. 142, 2023, doi:10.52005/fidelity.v5i1.142.

[51] K. M. M. Uddin, R. Ripa, N. Yeasmin, N. Biswas, and S. K. Dey, "Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset," *Intelligence-Based Medicine*, vol. 7, p. 100100, 2023.

[52] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, "Interpretable machine learning techniques in ecg-based heart disease classification: A systematic review," *Diagnostics*, vol. 13, no. 1, p. 111, 2023.

[53] L. M. Paladino, A. Hughes, A. Perera, O. Topsakal, and T. C. Akinci, "Evaluating the performance of automated machine learning (automl) tools for heart disease diagnosis and prediction," *AI (Switzerland)*, vol. 4, no. 4, pp. 53–68, 2023, doi:10.3390/ai4040053.

[54] A. J. Albert, R. Murugan, and T. Sripriya, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology," *Research on Biomedical Engineering*, vol. 39, no. 1, pp. 1–12, 2023, doi:10.1007/s42600-022-00253-9.

[55] L. V. Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy, M. Walczak, J. Garcke, C. Bauckhage, and J. Schuecker, "Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 1–20, 2023, doi:10.1109/TKDE.2021.3079836.

[56] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021, doi:10.1007/s42979-021-00592-x.

[57] Y. Xu, Y. Zhou, P. Sekula, and L. Ding, "Machine learning in construction: From shallow to deep learning," *Developments in the Built Environment*, vol. 6, p. 100045, 2021, doi:10.1016/j.dibe.2021.100045.

[58] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," *Information (Switzerland)*, vol. 11, no. 4, p. 193, 2020, doi:10.3390/info11040193.

[59] R. M. Radke, T. Frenzel, H. Baumgartner, and G. P. Diller, "Adult congenital heart disease and the covid-19 pandemic," *Heart*, vol. 106, no. 17, 2020, doi: 10.1136/heartjnl-2020-317258.

[60] P. Severino, A. D'Amato, M. Pucci, F. Infusino, L. I. Birtolo, M. V. Mariani, C. Lavalle, V. Maestrini, M. Mancone, and F. Fedele, "Ischemic heart disease and heart failure: Role of coronary ion channels," p. 3167, 2020, doi:10.3390/ijms21093167.

[61] R. H. Ritchie and E. D. Abel, "Basic mechanisms of diabetic heart disease," *Circulation Research*, vol. 126, no. 11, pp. 1501–1525, 2020, doi:10.1161/CIRCRESAHA.120.315913.

[62] Instituto Nacional de Estadística e Informática (INEI), "Perú: Enfermedades no transmisibles y transmisibles," 2025, [Accedido: 14-mayo-2025].

[63] N. Chandrasekhar and S. Peddakrishna, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," *Processes*, vol. 11, no. 4, pp. 1210–1221, 2023, doi:10.3390/pr11041210.

[64] V. Venkat, H. Abdelhalim, W. DeGroat, S. Zeeshan, and Z. Ahmed, "Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine," *Genomics*, vol. 115, no. 2, p. 110584, 2023, doi:10.1016/j.ygeno.2023.110584.

[65] N. Nissa, S. Jamwal, and M. Neshat, "A technical comparative heart disease prediction framework using boosting ensemble techniques," *Computation*, vol. 12, no. 1, p. 15, 2024, doi:10.3390/computation12010015.

[66] R. Canelón, C. Carrasco, and F. Rivera, "Design of a remote assistance model for truck maintenance in the mining industry," *Journal of Quality in Maintenance Engineering*, vol. 30, no. 1, pp. 1–337, 2024, doi:10.1108/JQME-02-2023-0024.

[67] S. Mirfakhraei, N. Abdolvand, and S. R. Harandi, "The rfmrv model for customer segmentation based on the referral value," *Interdisciplinary Journal of Management Studies*, vol. 17, no. 2, pp. 455–473, 2024, doi:10.22059/ijms.2023.329229.674722.

[68] C. Wang, A. Stupina, and S. Bezhitskiy, "Online sales prediction approach using methodology of crisp-dm," *ITM Web of Conferences*, vol. 59, p. 01006, 2024, doi:10.1051/itmconf/20245901006.

[69] S. Kanimozhi and N. Sivanandan, "Machine learning based heart disease prediction system," p. 03013, 2024, doi:10.1051/e3sconf/202449103013.

[70] P. K. Rajani, K. Patil, B. Marathe, P. Mhaisane, and A. Tundalwar, "Heart disease prediction using different machine learning algorithms," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9S, pp. 7430–7442, 2023.

[71] S. Roobini, M. S. Kavitha, and S. Karthik, "A systematic review on machine learning and neural network based models for disease prediction," *Journal of Integrated Science and Technology*, vol. 12, no. 4, pp. 787–787, 2024, doi:10.62110/sciencein.jist.2024.v12.787.

[72] P. Pachiyannan, M. Alsulami, D. Alsadie, A. K. J. Saudagar, M. AlKhathami, and R. C. Poonia, "A novel machine learning-based prediction method for early detection and diagnosis of congenital heart disease using ecg signal processing," *Technologies*, vol. 12, no. 1, pp. 1–15, 2024.

[73] S. Thangavel, S. Selvaraj, V. G. Karthikeyan, and K. Keerthika, "Analyzing machine learning classifiers for the diagnosis of heart disease," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 10, no. 1, pp. 1–12, 2024, doi:10.4108/eetpht.10.5244.

[74] S. V. Babu, P. Ramya, and J. Gracewell, "Revolutionizing heart disease prediction with quantum-enhanced machine learning," *Scientific Reports*, vol. 14, no. 1, p. 7453, 2024, doi:10.1038/s41598-024-55991-w.

[75] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," *IEEE Access*, vol. 11, no. 1, pp. 1–12, 2023, doi:10.1109/ACCESS.2023.3281484.