Cardiac Disease Prediction Using Machine Learning Survey of Classification Algorithms and Trends

[1]Purnima Pandey, [2]Shivank Kumar Soni
M.Tech Scholar, Department of Computer Science & Engineering, Oriental Institute of Science & Technology
Assistant Professor, Department of Computer Science & Engineering, Oriental Institute of Science & Technology
[1]acppurnimaoct@gmail.com, [2]shivanksoni@gmail.com

Abstract Heart disease is a worldwide health issue and a top cause of mortality globally. Early diagnosis is critical for improving patient outcomes and lightening the burden of healthcare systems. With the growing role of big data in medicine, machine learning (ML) has been the new promise for heart disease prediction and diagnosis by uncovering hidden patterns in complex data sets. This study aims at proposing a comprehensive review of machine learning algorithms and classification techniques used in heart disease prediction. We talk about the performance of the models such as Logistic Regression, Decision Trees, Support Vector Machines, Random Forests, K-Nearest Neighbours, and ensemble methods. We also mention the significance of data preprocessing, feature selection, cross-validation, and hyperparameter tuning for improving model accuracy. Specific emphasis is laid on interpretability, where explainable AI (XAI) techniques such as SHAP and LIME contribute towards transparency and clinician confidence. The issues of data imbalance, model generalizability, and clinical adoption are also discussed in the study. Comparative examination of current study emphasizes the efficacy of ML in early detection and its potential implications for transforming cardiac care. For scientists and clinicians interested in putting into practice reliable, accurate, and interpretable machine learning-based systems for cardiac disease prediction, this study serves as a basis.

Keywords: Heart disease prediction, machine learning, classification algorithms, healthcare analytics, logistic regression, support vector machine, feature selection, data imbalance, explainable AI, SHAP, LIME, early diagnosis, cardiovascular disease.

I. Introduction

Heart disease continues to be one of the most common causes of death globally, claiming millions of lives annually. The rising incidence of cardiovascular disease creates an immediate need for early and precise diagnosis in order to enhance patient outcomes and minimize healthcare costs. Conventional diagnostic techniques, though successful, are often time-consuming, qualitative, and lacking in predictive value. This has fuelled enthusiasm for the use of machine learning (ML) technologies, which can analyse complex patterns within large medical data to guide clinical decision-making [1]. ML offers the promise of detecting heart disease at earlier stages by sensing weak signals that might not be picked up by conventional means, thus assuming a revolutionary function in modern healthcare by enhancing the speed, accuracy, and efficacy of diagnosis and prognosis. One of the foremost reasons for cardiovascular sickness and mortality in the world is valvular heart disease, and the prevalence of this illness is only anticipated to increase in the next few decades. The aortic, mitral, pulmonic, and tricuspid valves are the four heart valves. The valves ensure that backflow is averted in the four chambers of the heart, thereby maintaining pressure gradients needed for hemodynamic circulation necessary for life. Valvular heart disease is the secondary etiology of either insufficiency or regurgitation of the valves, both of which permit backward flow and have the capacity to equalize pressure in manners incompatible with cardiovascular performance [2]. Coronary artery disease, the most common etiology of heart disease, is resultant from coronary arteries failing to supply sufficient blood to the heart muscle. Extended myocardial ischemia is a potentially lethal condition that may lead to myocardial infarction or sudden cardiac death. Heart failure, or abnormal or compromised cardiac function, can also occur without coronary artery disease. Pathological alterations in the structure of the heart muscle result in this type of non-ischemic heart disease. Due to inefficient muscular contraction (systolic heart failure) or relaxation (diastolic heart failure), the heart cannot pump blood effectively when it has heart failure [3].

According to the World Health Organization's 2020 forecast, cardiovascular diseases (CVD) account for 17.9 million deaths annually, making them the leading cause of mortality globally. One crucial strategy to lower this toll is early CVD identification. Data mining is one of the numerous methods for enhancing disease detection and diagnosis. These related techniques are a potential approach for CVD classification because they enable the extraction of hidden knowledge and the identification of correlations among parameters within the dataset [4]. Globally, cardiovascular diseases account for about one-third of all fatalities. Ischemic heart disease (IHD) is the most common of the cardiovascular ailments. In fact, IHD is recognized as a significant risk to 21st-century sustainable development. IHD, also known as coronary artery disease (CAD) and atherosclerotic cardiovascular disease (ACD), presents clinically as ischemic cardiomyopathy and myocardial infarction. More and more people with non-fatal IHD have chronic disability and a lower quality of life. Atherosclerosis, an inflammatory artery disease linked to cholesterol buildup and metabolic changes brought on by a number of risk factors, is the main

pathological mechanism that causes IHD. Only 2% to 7% of the general population have no risk factors for IHD, but over 70% of at-risk persons have numerous risk factors [5]. To diagnose CVD, an electrocardiogram (ECG) is used. However, it requires time and effort to visually identify long-term ECG abnormalities. Many academics and practitioners have discovered machine learning-based heart disease diagnostics (MLBHDD) systems to be affordable and adaptable methods since the introduction of machine learning (ML) applications in the medical field. Consequently, a number of research that used various heart disease datasets proposed MLBHDD [6].

The creation of vast volumes of data in the medical field has been made possible by developments in computer technology, digital data storage, and communication technologies. Medical professionals can diagnose patients more accurately by identifying patterns in medical data. Features such as demographics, test results, pictures, video clips, and more make up a patient's data. Given the magnitude and wide dimensions of the data, manually extracting the needed information from the massive amount of data is an enormous undertaking. Therefore, automated analysis methods are needed to examine the data. Data mining techniques are useful because they can manage big datasets and automate analysis [7]. A decision support system for the simple and economical detection of heart disease from clinical data can be designed using machine learning techniques. An early diagnosis of the condition can be aided by this type of decision-making system. This kind of decision support system can be transformed into a medical chatbot that the patient can use to identify cardiac trouble early. A chatbot will accept healthcare data as input and provide patients with appropriate advice. To guarantee dependability and safety, these medical chatbot systems must be subject to laws. Therefore, in order to reduce safety risks, risk management actions must be carried out [8].

II. Types of Heart Disease

Heart disease is a general term for a number of conditions that involve the blood vessels and heart. Learning about the several forms is important to make an accurate diagnosis and proper treatment. The most prevalent forms are:

A. Coronary Artery Disease (CAD)

Atherosclerosis formation in the coronary arteries, which can occasionally be asymptomatic, is the hallmark of coronary artery disease (CAD). Acute coronary syndrome (ACS), silent myocardial ischemia, and stable angina are all included in coronary heart disease (CHD), commonly referred to as ischemic heart disease (IHD). The primary cause of CHD-related mortality is CAD. Myocardial infarction and unstable angina are examples of ACS, which is usually symptomatic. For the sake of clarity, we shall refer to CHD as "CAD" throughout this conversation. Globally, CAD is the main cause of death and the loss of disability-adjusted life years (DALYs). Low- and middle-income nations bear a disproportionate amount of this burden, which results in 129 million DALYs and around 7 million deaths yearly. Globally, CAD caused 164.0 million DALYs and 8.9 million deaths in 2015. Myocardial infarction survivors have a five to six times greater annual death rate than those without CAD, and they are at a much higher risk of experiencing [9]. Inadequate oxygen to the heart muscle can cause symptoms like pain or discomfort in the chest, which are also known as angina. In more serious ones, a plaque can burst and develop a blood clot, which in turn can clog the artery and lead to a heart attack, or what is also referred to as a myocardial infarction. Unless treated, the ongoing stress on the heart due to decreased blood supply can cause weakening of the heart muscle, leading to heart failure—a situation where the heart is unable to efficiently pump blood to supply the body's needs. CAD usually develops symptomatically and insidiously, so early detection and aggressive management are important in order to avoid life-threatening complications [10].

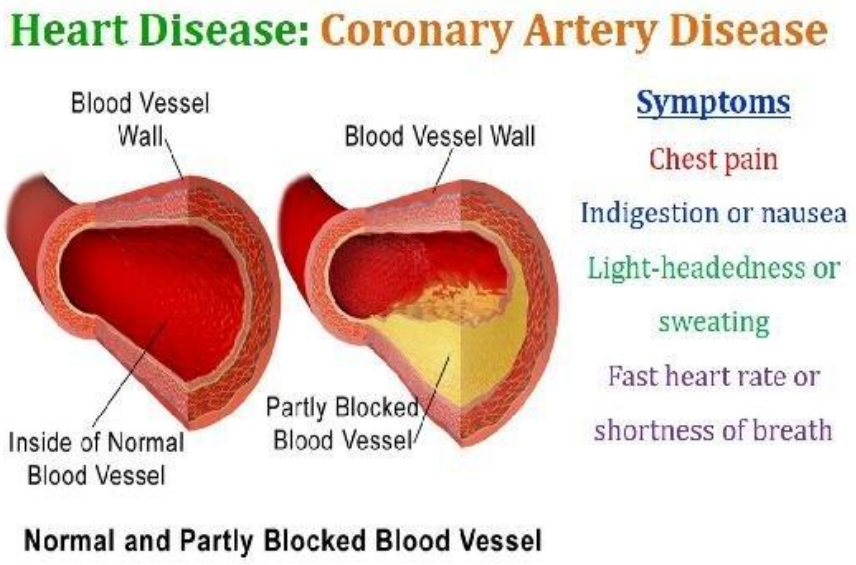

Figure 1. Coronary Artery Disease [11]

Figure 1 shows how providing nourishment to the heart's muscle and facilitating oxygen-rich blood circulation are the primary roles of the coronary arteries. Cholesterol can cause sickness or damage to the coronary arteries. The body receives less oxygen and nutrients from the coronary arteries as a result of the cholesterol [11].

B. Heart Failure

Heart failure is a situation where the heart cannot pump blood sufficiently to provide the body with what it needs. It usually comes as a complication of a pre-existing condition like coronary artery disease, high blood pressure, or past heart attacks that causes the heart muscle to become weak or damaged. Consequently, blood circulation is not sufficient, causing conditions such as tiredness, breathlessness, and swelling in the legs, abdomen, or lungs. Even though it is a progressive and chronic condition, timely diagnosis and appropriate treatment can control symptoms and improve quality of life [12]. Since heart failure is a diverse illness, it might be difficult to identify cases and classify patients in epidemiological studies. Most people agree that the left ventricular ejection fraction (LVEF) is a clinically valuable phenotypic trait that reveals underlying pathophysiological causes and treatment sensitivity. The three most common classifications for heart failure patients today are decreased (HFrEF; LVEF <40%), mid-range (HFmrEF; LVEF 40–49%), and preserved ejection fraction (HFpEF; LVEF ≥50%). 5 LVEF classification has been criticized for oversimplifying a complex disease, and cut-off values are subjective and vary throughout recommendations [13].

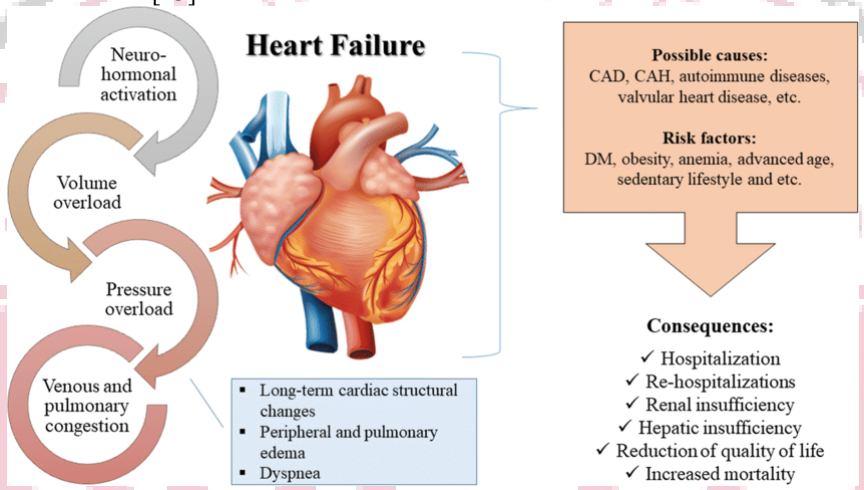

Figure 2 Evolution of Heart Failure: From risk factors to outcomes [14]

Figure 2 shows the pathophysiology, etiology, risk factors, and implications of heart failure. It underscores prominent mechanisms such as activation of the neurohormones, volume and pressure overload, and congestion in the venous/pulmonary system, which result in structural alterations to the heart, fluid retention, and dyspnea. Potential etiologies are coronary artery disease, congenital cardiac anomalies, autoimmune illnesses, and valvular disease. Risk factors including diabetes, obesity, age, and physical inactivity predispose to disease advancement. Eventually, heart failure leads to repeated hospitalization, organ dysfunction (renal and hepatic), decreased quality of life, and elevated mortality [14].

C. Arrhythmias

Arrhythmias are heart rhythm abnormalities caused by disruptions in the electrical impulses that regulate heartbeat. They can be detected by an electrocardiogram (ECG), which detects electrical impulses through electrodes on the chest or limbs. Arrhythmias can appear as bradycardia (abnormally slow heartbeat), tachycardia (abnormally fast heartbeat), or irregular beats arising from various heart chambers. Sinus arrhythmias are caused by irregularities of the sinus node, and sinus bradycardia describes slow rhythms while sinus tachycardia represents faster-than-normal rhythms. Other frequently occurring ones include atrial arrhythmias (such as atrial fibrillation and flutter) and ventricular arrhythmias (such as ventricular fibrillation and flutter), some of which represent life-threatening conditions such as cardiac arrest [15]. Advanced machine learning methods, including Support Vector Machines (SVM), have been used successfully to classify and identify arrhythmias based on ECG signals. These models learn based on labelled ECG data, including Normal Sinus Rhythm (NSR), Congestive Heart Failure (CHF), and Cardiac Arrhythmia, where features are extracted using techniques like Discrete Wavelet Transform (DWT). The research cited attained a superior accuracy level (about 95.92%) in the classification of arrhythmias, highlighting the promise of computerized systems in early and accurate identification of cardiac anomalies, ultimately contributing to timely medical intervention and curbing resultant mortality risks [16].

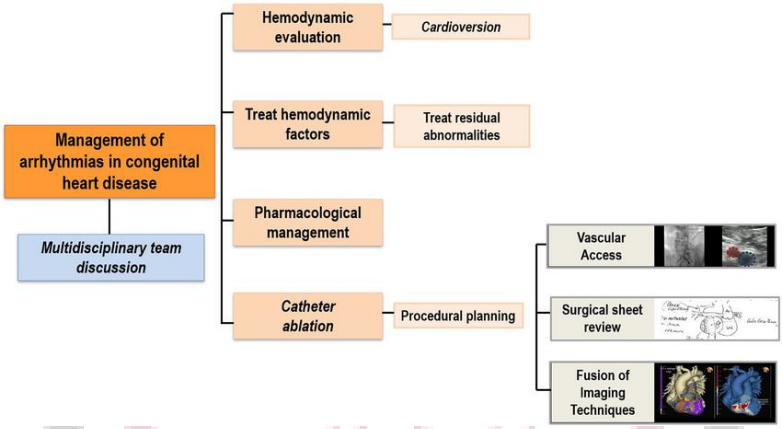

Figure 3 General arrhythmia workflow [17]

Figure 3 demonstrates a systematic method of the management of arrhythmias in patients with congenital heart disease, stressing the need for a multidisciplinary team discussion. Hemodynamic assessment, which can include cardioversion, as well as the treatment of hemodynamic factors and residual abnormalities, form part of the management strategies. Pharmacological treatment is also another important aspect. For procedural treatment, catheter ablation is scheduled with factors like vascular access, review of surgical sheet, and image integration to ensure better precision and outcomes. This combined approach provides individualized and integral care for complex congenital heart diseases [17].

D. Valvular Heart Disease

Valvular Heart Disease (VHD) is any malfunction or abnormality of any of the heart valves, usually caused by rheumatic fever or degenerative changes with age. While in the developing nations such as China, rheumatic heart disease (RHD) is still the major cause of VHD, especially among the elderly, although degenerative heart disease (DHD) is increasing immensely with growing populations and life-style changes. VHD encompasses conditions like aortic regurgitation, mitral regurgitation, and stenosis, usually advancing silently and elevating the risk of heart failure, atrial fibrillation, and mortality. Prevalence of the disease correlates highly with age and hypertension, thus necessitating prompt detection through echocardiography to avoid complications and enhance outcomes [18].

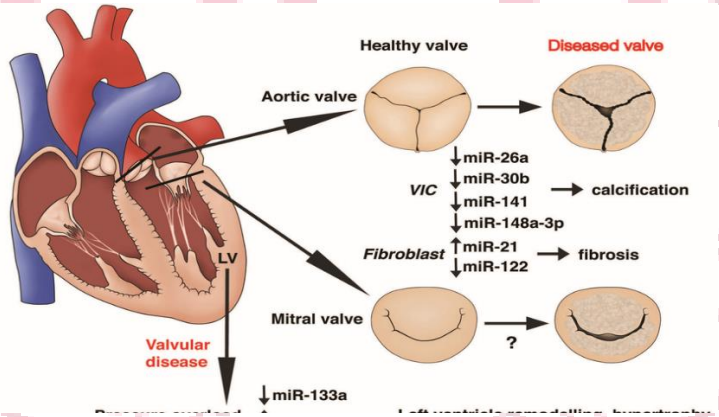

Figure 4 Molecular Pathogenesis of Valvular Heart Disease [19]

Figure 4 shows the development of diseased heart valves (aortic and mitral) from healthy ones and indicates the molecular function of certain microRNAs (miRNAs) in this process. Downregulation of miR-26a, miR-30b, miR-141, and miR-148a-3p in valve interstitial cells (VICs) results in calcification in aortic valve disease, whereas downregulation of miR-21 and miR-122 in fibroblasts causes fibrosis. Mitral valve disease seems to be associated with unknown factors (marked by "?"). Collectively, these alterations propel valvular disease and cause left ventricular (LV) remodelling and hypertrophy, in part modulated by the changed expression of miR-133a.

III. Risk Factors and Causes in heart disease

Heart disease arises from the culmination of lifestyle, genetic, and environmental risks that heighten a person's susceptibility over a period of time. The biggest contributor is poor lifestyle habits, including an unhealthy diet rich in saturated fats, salt, and sugar, physical inactivity, smoking, and heavy alcohol use. These habits can cause diseases such as obesity, high blood pressure, high cholesterol, and type 2 diabetes, all of which put an enormous stress on the blood vessels and heart. These diseases, over time, can cause damage to the arteries, decrease blood flow, and result in atherosclerosisa leading cause of heart attack and coronary artery disease [20]. Besides lifestyle determinants, medical comorbidities and conditions are also extremely important in the causation of heart disease. Chronic illnesses like high blood pressure, diabetes, and high cholesterol levels have a significant impact on

cardiovascular risk. For example, high blood pressure causes the heart to pump blood harder, which can make the heart muscle enlarge and become weakened. In the same manner, high blood sugar in diabetics also causes damage to the lining of blood vessels, leading to plaque deposition. Individuals with a history of heart attack, stroke, or other cardiovascular complications are also more likely to develop more aggressive or recurrent heart disease [21].Age, gender, family history, and genetics are non-modifiable factors affecting heart disease risk in an individual. Heart disease risk rises with age, especially for men above the age of 45 and women above the age of 55. Family history of heart problems, particularly among close relatives, indicates a genetic risk of heart problems. In addition, some inherited diseases such as familial hypercholesterolemia can lead to very high cholesterol levels at an early age, substantially increasing the risk of premature heart disease. Although these are unmodifiable, those who are aware of their genetic predispositions can implement preventive practices through frequent examinations and lifestyle changes to lower their overall risk [22].

IV. Machine Learning in Healthcare in health disease

Machine learning (ML), artificial intelligence's branch, increasingly affects healthcare by allowing computers to examine intricate medical data and aid in diagnosis and treatment. ML is able to identify subtle patterns in patient histories, images, and labs, facilitating early disease detection and customized care. ML enhances diagnostic accuracy, minimizes human error, and automates repetitive tasks such as image analysis. Yet its clinical uptake is threatened by requirements, such as the availability of high-quality data, interpretability of the models, possible algorithm bias, and integration into current healthcare systems, which demand planning, investment, and concerted effort [23]. Machine learning (ML) is a form of artificial intelligence where an algorithm is created that learns patterns from data and makes decisions or predictions without being programmed. It gets better with time as it is fed more data. For healthcare, ML allows computers to process large and complicated sets of medical datapatient histories, imaging test results, and gene mutations—with incredible speed and accuracy. Supervised learning, unsupervised learning, and reinforcement learning are frequently applied methods, every one of which finds application on various types of healthcare issues, such as classification, prediction, clustering, and treatment optimization [24].

A. Relevance to Medical Diagnosis

Machine learning is becoming a vital component in medical diagnosis because it is capable of identifying patterns and associations in data that are not obvious to human clinicians. With training on past patient information, ML models are capable of helping predict disease risks, identify diseases, and recommending treatments. For example, ML is capable of interpreting ECG signals, imaging, or blood test results to identify diseases such as heart disease, diabetes, or cancer at early stages. Not only do these tools assist physicians with more precise decisions, but they also assist in minimizing diagnostic errors and rationalizing care across diverse medical environments [25]. Machine learning integration in hospitals has a number of benefits. Diagnostic speed and accuracy are improved, early detection of disease is supported, and personalized treatment strategies can be developed from the patients' individual profiles. ML also supports automation of routine activities like image analysis, reducing the burden on healthcare providers and enabling them to spend more time on patient care. Furthermore, ML is able to reveal insights from big datasets that can result in enhanced clinical guidelines, risk-identification of populations, as well as more efficient allocation of hospital resources [26].Although promising, machine learning also has pitfalls to clinical application. One of the most important issues is the requirement of abundant, high-quality, and representative data sets, which are usually hard to access due to data fragmentation and privacy regulations. Additionally, ML algorithms can be "black boxes," leaving it hard to interpret their predictions and gain clinician trust. Concerns like algorithm bias, non-standardization, and the requirement for ongoing validation and observation further muddle implementation. Furthermore, implementing ML within current healthcare systems involves major investment, training, and coordination among clinicians, data scientists, and IT professionals [27].

V. Performance Metrics

Evaluation metrics like Accuracy, Precision, Recall, and F1-Score are crucial in measuring the performance of heart disease prediction classification models. Accuracy represents the ratio of instances that were correctly predicted, while Precision indicates the number of true positives among all predicted positivescritical when there is a need to reduce false positives. Recall (or sensitivity) measures the extent to which the model recognizes true positive instances, and F1-Score is a balance between Precision and Recall, particularly for class imbalance scenarios. Moreover, the ROC Curve visually describes a model's capacity to differentiate between classes by representing the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds. Its performance is summarized by the Area Under the Curve (AUC), with higher AUC values denoting improved discrimination of the model, an ideal 1.0, and a random guess of 0.5. These measures assist in determining not just how accurate a model is but also how robust and unbiased it is under different conditions.

A. Accuracy, Precision, Recall, F1-Score

These are simple evaluation measures utilized to determine the performance of models for classification:
Accuracy calculates the overall accuracy of the model in terms of correctly predicted instances divided by the total instances [28].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision indicates how many of the positively predicated cases were actually positive, which is crucial when false positives need to be minimized [28].

$$Precision = \frac{TP}{TP + FP}$$

Recall (also called Sensitivity or True Positive Rate) measures how many of the actual positive cases the model was able to identify [28].

$$Recall = \frac{TP}{TP + FN}$$

F1-Score is the harmonic mean of precision and recall, providing a balanced measure when there is an uneven class distribution or when both false positives and false negatives matter [28].

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

B. ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve is an effective graphical means to measure the performance of a classification model, especially for binary classification tasks. It depicts the True Positive Rate (TPR), or recall or sensitivity, versus the False Positive Rate (FPR) at different threshold levels. The TPR indicates how well the model is able to detect true positive cases, while the FPR represents the percentage of incorrect assignment of negative cases as positive. Through the use of the classification threshold, the ROC curve determines the way the model strikes the balance between detecting positives and preventing false alarms [29]. The Area Under the Curve (AUC) is a single scalar measure that captures the entire ROC curve, measuring the overall capacity of the model to discriminate between the two classes—positive and negative. A high AUC value of approximately 1.0 corresponds to superb discriminative performance, in which the model can nearly perfectly segregate positive cases from negative ones. On the other hand, an AUC of 0.5 would indicate that the model is no better than random guess with no significant predictive value. The ROC curve and AUC are hence particularly useful in cases of skewed datasets or differing classification costs since they give an idea of the model performance at all the decision thresholds and not at a single point like accuracy [30].

ROC Curve plots True Positive Rate (TPR) or False Positive Rate

TPR(Recall)

$$TPR = \frac{TP}{TP + FN}$$

FPR

$$FPR = \frac{FP}{FP + TN}$$

Recent study points to the increasing application of machine learning (ML) in the prediction of heart disease. A logistic regression model based on nine biochemical markers at Villa Scassi Hospital properly discriminated between heart failure and chronic-ischemic heart disease through 20-fold cross-validation [31]. A hybrid explainable AI model (HXAI-ML) that integrates data balancing techniques and interpretability methods such as SHAP and LIME obtained more than 99% accuracy over benchmark datasets [32]. Logistic regression also performed better than alternative models in coronary datasets [33], and accuracy was further improved with hyperparameter tuning to 91.80% [34]. Transparent ML methods have enhanced ECG-based diagnosis by improving clinical confidence [35], whereas feature selection and ensemble techniques, including Random Forest and Bagged Decision Tree, increased accuracy to more than 99% in certain scenarios [36–38]. Furthermore, ML also demonstrated potential for predicting cardiac risks due to conditions such as PCOS [39], while combining blockchain with schemes such as SCA_WKNN enhanced prediction performance as well as data protection [40]. Newer generation machine learning techniques come very highly in predicting and diagnosis of heart disease. Hence, among the different models, the XGBoost was the best with 91.8% accuracy after hyperparameter tuning through Bayesian optimization following one-hot encoding techniques on the Cleveland dataset [41]. The importance that was now attached to user-friendly design and dual-stage risk assessment, led to the development of a web-based system that utilizes XGBoost and Gradient Boosting with 85% and 93% accuracies, respectively, for heart disease and heart failure prediction from the UCI dataset [42]. AdaBoost was marginally better than XGBoost in predictive modeling with accuracies of 89% and 87%, respectively, and largely outperforming XGBoost in F1 and precision scores, reaffirming its consistencies in any kind of data class [43]. Gradient Boosting came out best, with 92.2% accuracy on the UCI dataset after aggressive preprocessing, including outlier removal and imputation, surpassing XGBoost and AdaBoost in overall predictive metrics [44]. It was identified that AdaBoost with 0.95 accuracy came first from among CatBoost, RF, LightGBM, and others, having a very good

NPV of 0.83 and very low rates of 0.04 for FPR and FNR after extensive experiments performed on a large 8763 global records dataset [45]. Another set of results that focused on single-parameter optimization showed that AdaBoost and the Extreme Learning Machine attained accuracies of 0.87 and 0.83 environments, respectively, with improvements obtained when the learning rate parameters were optimized [46]. Ensemble methods of LDA, CART, SVM, KNN, and Naïve Bayes with RF meta-classifier increased prediction accuracy from 85.53% to 87.64%, emphasizing the advantage of model stacking and feature optimization [47]. In Bangladesh, a comparative study of the classifiers on 391 patient records identified random forest achieving the highest accuracy (98.04%), precision (96.15%), and AUC (0.989), surpassing all others, including logistic regression and bagging tree, establishing its strong potential for clinical implementation in a low-resource setting [48].

Table 1 Comparative Analysis of Machine Learning Approaches for Heart Disease Prediction

Reference	Key Focus	Algorithms Used	Dataset(s)	Performance Highlights	Performance Metrics Used	Challenges Addressed
[31]	Logistic regression to distinguish HF vs chronic-ischemic HD	Logistic Regression	Villa Scassi Hospital data	Identified strong predictors using LR with cross-validation	Logistic regression, 20-fold CV	Differentiating between similar cardiac conditions
[32]	XAI-based ML model with balancing + interpretation	Extra Trees, SHAP, LIME, PIA	Cleveland, Framingham	Accuracy up to 99.24%, with interpretability via XAI	Accuracy, Precision, XAI (SHAP, LIME, PIA)	Data imbalance, interpretability, generalization
[33]	Comparing ML classifiers on coronary HD data	Logistic Regression, others	Coronary Heart Disease dataset	LR had best performance on original dataset	Accuracy comparison of ML models	Class imbalance, algorithm comparison
[34]	Performance boost via tuning	LR, KNN, SVM, DT, RF	Heart disease dataset (unspecified)	Accuracy improved to 91.80%	Accuracy, Precision, Recall, F1-Score	Improving model accuracy with tuning
[35]	IML techniques for ECG signal-based diagnosis	Interpretable ML Models	ECG signal datasets	Focus on explainability, not accuracy	Emphasis on explainability over metrics	Complexity of ECG, model trust
[36]	Feature selection for predicting death events	LDA, RF, GBC, DT, SVM, KNN	Heart disease patient data	RF+SFS achieved 86.67% accuracy	Accuracy, Confusion Matrix, ROC, Precision, Recall, F1-score	Selecting key features for better predictions
[37]	Full ML pipeline with FS, tuning, and evaluation	DT, RF, SVM, NB + RFE, PCA	Kaggle, other real-world datasets	Up to 99.12% accuracy, robust model	Accuracy, Precision, Recall, ROC	High-dimension data, avoiding overfitting
[38]	Misclassified instance-based augmentation	Bagged Decision Tree	Cleveland, custom dataset	Accuracy up to 99.2% with 10-fold CV	Accuracy during CV, comparison with baseline	Boosting model with augmentation
[39]	Using disease symptoms to infer PCOS risk	Supervised + Unsupervised Learning	Merged disease datasets	Best performance using selected key features	Accuracy with full vs selected features	Feature minimization, disease interrelationship
[40]	Secure prediction using blockchain storage	SCA_WKNN	Blockchain-based medical data	Higher accuracy than KNN/WKNN, secure data	Accuracy, Precision, Recall, F1-score, RMSE	Secure data storage, performance, real-time monitoring
[41]	Heart disease prediction using optimized XGBoost	XGBoost (with Bayesian optimization), RF, ET	Cleveland Heart Disease Dataset	91.8% accuracy; outperforming RF and ET	Accuracy, Sensitivity, Specificity, F1-score, AUC	Hyperparameter tuning, categorical feature encoding
[42]	Web-based heart disease & failure prediction tool	XGBoost, Gradient Boosting	UCI Heart Disease Dataset	93% (heart failure), 85% (disease) accuracy	Accuracy	Dual-stage prediction, interface

						usability, 'time' feature exclusion
[43]	Comparative analysis of XGBoostvsAdaBoost	XGBoost, AdaBoost	-	AdaBoost: 89% accuracy vs. XGBoost: 87%	Accuracy, Precision, F1 Score	Algorithm selection, class-wise performance
[44]	Boosting algorithms for heart disease detection	Gradient Boost, XGBoost, AdaBoost	UCI ML Heart Dataset	Gradient Boost: 92.2% accuracy	Accuracy, Precision, Recall, F1 Score	Outlier & missing value handling, preprocessing impact
[45]	Early heart disease prediction using ensemble boosting	CatBoost, RF, Gradient Boost, LightGBM, AdaBoost	UCI ML Heart Dataset (8763 samples)	AdaBoost: 95% accuracy, high NPV and low FPR/FNR	Accuracy, NPV, FPR, FNR, FDR	High-dimensional data, robust prediction framework
[46]	Impact of single parameter tuning in ML prediction	Extreme Learning Machine, AdaBoost	Heart Failure Dataset	ELM: 83%, AdaBoost: 87% accuracy	Accuracy, Std. Dev.	Parameter tuning, model enhancement
[47]	Ensemble ML system for heart disease prediction	LDA, CART, SVM, KNN, NB + RF (meta classifier)	-	Accuracy improved from 85.53% to 87.64%	Accuracy	Ensemble learning, feature selection
[48]	CVD prediction in LMICs using ensemble methods	Logistic Reg., Naïve Bayes, Decision Tree, AdaBoost, RF, Bagging Tree	Dataset of 391 CVD patients + 260 controls	RF: 98.04% accuracy, AUC 0.989	Accuracy, Sensitivity, Specificity, AUC, Precision, F1	CVD risk analysis, class imbalance, practical deployment

VI. Challenges in Heart Disease

Coronary disease is a disease that poses substantial diagnostic and management problems because it is a complex and multifactorial disease. One of the biggest problems is the early detection of diseases such as coronary disease and heart failure, which usually develop quietly and are recognized by symptoms only in late stages. Conventional diagnostic tests like ECGs, echocardiogram, and angiogram need expert interpretation and can fail to pick up subtle signs of disease. Second, comorbid diseases such as diabetes, high blood pressure, and obesity can obscure or obscure cardiovascular symptoms, resulting in delayed treatment or misdiagnosis. Growing size and heterogeneity of patient data—everything from clinical documentation to imaging and genetic information—are also making human analysis slow and error-prone, highlighting the requirement for automated, knowledge-driven systems that can rapidly and accurately synthesize this data. Technologically, the inclusion of machine learning (ML) and artificial intelligence (AI) in clinical practice has its own challenges. ML models tend to have problems with datasets that are imbalanced, with disease instances underrepresented so that predictions tend to be biased. Additionally, the black-box nature of several sophisticated ML algorithms obstructs clinical trust and interpretability, which is important for life-or-death decision-making in medicine. There is also a lack of standardization between datasets, which makes it impossible to generalize models from one population to another reliably. Infrastructural constraints, like inadequate digital health systems and issues of data privacy, also make implementation difficult. To effectively leverage ML for the prediction and management of heart disease, data quality, transparency, regulatory compliance, and real-world validation issues must be tackled systematically.

VII. Conculsion

Heart disease remains a leading cause of death globally, highlighting the critical need for accurate, early and cost-effective diagnostic tools. Traditional diagnostic methods, though clinically valuable, often fall short in detecing subtle or early-stage cardiovasluar conditions. This gap has encouraged the speedy adoption of machine learning (ML) technologies for the prediction and classification of heart disease in particular. From this survey, it is clear that ML models like Logistic Regression, Random Forest, Support Vector Machine, and ensemble methods have demonstrated significant potential for correctly predicting heart disease, especially when utilized in conjunction with data preprocessing techniques, feature selection, and hyperparameter tuning. Additionally, explanation of AI models and integration of software packages such as SHAP, LIME, and ROC curves have greatly improved the interpretability and trustworthiness of ML predictions to make them acceptable to clinicians. Notwithstanding these developments, fundamental challenges of data imbalance, non-transparency of models, and generalizability across diverse populations continue to remain. Yet, constant innovations such as hybrid models, blockchain

integration, and interpretable ML methods are leading the way towards stronger, more secure, and more scalable diagnostic systems. In the end, machine learning is a valuable friend in the battle against heart disease, providing the possibility to revolutionize preventive cardiology and enhance patient outcomes across the globe.

References

[1] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[2] Aluru, J. S., Barsouk, A., Saginala, K., Rawla, P., &Barsouk, A. (2022). Valvular heart disease epidemiology. *Medical sciences*, *10*(2), 32. https://doi.org/10.3390/medsci10020032

[3] Rolski, F., &Błyszczuk, P. (2020). Complexity of TNF-α signaling in heart disease. *Journal of clinical medicine*, *9*(10), 3267. https://doi.org/10.3390/jcm9103267

[4] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

[5] Khan, M. A., Hashim, M. J., Mustafa, H., Baniyas, M. Y., Al Suwaidi, S. K. B. M., AlKatheeri, R., ... & Lootah, S. N. A. H. (2020). Global epidemiology of ischemic heart disease: results from the global burden of disease study. *Cureus*, *12*(7). DOI: 10.7759/cureus.9349

[6] Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289.https://doi.org/10.1016/j.artmed.2022.102289

[7] Reddy, G.T., Reddy, M.P.K., Lakshmanna, K. et al. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evol. Intel. 13, 185–196 (2020). https://doi.org/10.1007/s12065-019-00327-1

[8] Rani, P., Kumar, R., Ahmed, N.M.O.S. et al. A decision support system for heart disease prediction based upon machine learning. J Reliable Intell Environ 7, 263–275 (2021). https://doi.org/10.1007/s40860-021-00133-6

[9] Shahjehan, R. D., Sharma, S., & Bhutta, B. S. (2024). Coronary artery disease. In *StatPearls [Internet]*. StatPearls Publishing.

[10] Pagliaro, B.R., Cannata, F., Stefanini, G.G. et al. Myocardial ischemia and coronary disease in heart failure. Heart Fail Rev 25, 53–65 (2020). https://doi.org/10.1007/s10741-019-09831-z

[11] Sandhya, Y. (2020). Prediction of heart diseases using support vector machine. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *8*(2), 2020.

[12] Arrigo, M., Jessup, M., Mullens, W., Reza, N., Shah, A. M., Sliwa, K., &Mebazaa, A. (2020). Acute heart failure. *Nature Reviews Disease Primers*, *6*(1), 16.

[13] Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. (2020). Epidemiology of heart failure. *European journal of heart failure*, *22*(8), 1342-1356.

[14] Durães, André & Hoffmann Filho, Conrado & Bitar, Yasmin & Gomes Neto, Mansueto. (2020). Heart Failure and Comorbidities—Part 1. Current Emergency and Hospital Medicine Reports. 8. 10.1007/s40138-020-00210-9.

[15] Kumari, C. U., Murthy, A. S. D., Prasanna, B. L., Reddy, M. P. P., &Panigrahy, A. K. (2021). An automated detection of heart arrhythmias using machine learning technique: SVM. *Materials Today: Proceedings*, *45*, 1393-1398. https://doi.org/10.1016/j.matpr.2020.07.088

[16] Ketu, S., Mishra, P.K. Empirical Analysis of Machine Learning Algorithms on Imbalance Electrocardiogram Based Arrhythmia Dataset for Heart Disease Detection. Arab J Sci Eng 47, 1447–1469 (2022). https://doi.org/10.1007/s13369-021-05972-2

[17] Francisco Pascual, Jaume & Vila, Núria & Santos-Ortega, Alba & Rivas-Gándara, Nuria. (2024). Tachyarrhythmias in congenital heart disease. Frontiers in Cardiovascular Medicine. 11. 10.3389/fcvm.2024.1395210.

[18] Santangelo, G., Bursi, F., Faggiano, A., Moscardelli, S., Simeoli, P. S., Guazzi, M., ... & Faggiano, P. (2023). The global burden of valvular heart disease: from clinical epidemiology to management. *Journal of clinical medicine*, *12*(6), 2178. https://doi.org/10.3390/jcm12062178

[19] Oury, Cécile &Servais, Laurence &Bouznad, Nassim&Hego, Alexandre&Nchimi, Alain &Lancellotti, Patrizio. (2016). MicroRNAs in Valvular Heart Diseases: Potential Role as Markers and Actors of Valvular and Cardiac Remodeling. International Journal of Molecular Sciences (IJMS). 17. 1120. 10.3390/ijms17071120.

[20] Osibogun, O., Ogunmoroti, O., & Michos, E. D. (2020). Polycystic ovary syndrome and cardiometabolic risk: Opportunities for cardiovascular disease prevention. *Trends in cardiovascular medicine*, *30*(7), 399-404. https://doi.org/10.1016/j.tcm.2019.08.010

[21] Kazemian, N., Mahmoudi, M., Halperin, F. et al. Gut microbiota and cardiovascular disease: opportunities and challenges. Microbiome 8, 36 (2020). https://doi.org/10.1186/s40168-020-00821-0

[22] Sharifi-Rad, J., Rodrigues, C. F., Sharopov, F., Docea, A. O., Can Karaca, A., Sharifi-Rad, M., ... & Calina, D. (2020). Diet, lifestyle and cardiovascular diseases: linking pathophysiology to cardioprotective effects of natural bioactive compounds. *International journal of environmental research and public health*, *17*(7), 2326. https://doi.org/10.3390/ijerph17072326

[23] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE access*, *8*, 107562-107582. https://doi.org/10.1109/ACCESS.2020.3001149

[24] Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, *2022*(1), 7351061. https://doi.org/10.1155/2022/7351061

[25] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208-222. https://doi.org/10.1016/j.inffus.2020.06.008

[26] Jaarsma, T., Hill, L., Bayes-Genis, A., La Rocca, H. P. B., Castiello, T., Čelutkienė, J., ... & Strömberg, A. (2021). Self-care of heart failure patients: practical management recommendations from the Heart Failure Association of the European Society of Cardiology. *European journal of heart failure*, *23*(1), 157-174. https://doi.org/10.1002/ejhf.2008

[27] Bader, F., Manla, Y., Atallah, B. et al. Heart failure and COVID-19. Heart Fail Rev 26, 1–10 (2021). https://doi.org/10.1007/s10741-020-10008-2

[28] Asif, M. A. A. R., Nishat, M. M., Faisal, F., Dip, R. R., Udoy, M. H., Shikder, M. F., & Ahsan, R. (2021). Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease. *Engineering Letters*, *29*(2).

[29] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., &Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Applied Sciences*, *11*(18), 8352. https://doi.org/10.3390/app11188352

[30] Drożdż, K., Nabrdalik, K., Kwiendacz, H. et al. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. Cardiovasc Diabetol 21, 240 (2022). https://doi.org/10.1186/s12933-022-01672-9

[31] Stojanov, D., Lazarova, E., Veljkova, E., Rubartelli, P., & Giacomini, M. (2023). Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population–Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy. *Journal of King Saud University-Science*, *35*(3), 102573. https://doi.org/10.1016/j.jksus.2023.102573

[32] Talukder, M. A., Talaat, A. S., & Kazi, M. (2025). HXAI-ML: a hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection. *Results in Engineering*, *25*, 104370.https://doi.org/10.1016/j.rineng.2025.104370

[33] Kwakye, K., & Dadzie, E. (2021). Machine learning-based classification algorithms for the prediction of coronary heart diseases. *arXiv preprint arXiv:2112.01503*. https://doi.org/10.48550/arXiv.2112.01503

[34] Hashi, E. K., & Zaman, M. S. U. (2020). Developing a hyperparameter tuning based machine learning approach of heart disease prediction. *Journal of Applied Science & Process Engineering*, *7*(2), 631-647.

[35] Ayano, Y. M., Schwenker, F., Dufera, B. D., &Debelee, T. G. (2022). Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review. *Diagnostics*, *13*(1), 111. https://doi.org/10.3390/diagnostics13010111

[36] Aggrawal, R., Pal, S. Sequential Feature Selection and Machine Learning Algorithm-Based Patient's Death Events Prediction and Diagnosis in Heart Disease. SN COMPUT. SCI. 1, 344 (2020). https://doi.org/10.1007/s42979-020-00370-1

[37] Islam, M. A., Majumder, M. Z. H., Miah, M. S., &Jannaty, S. (2024). Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction. *Computers in Biology and Medicine*, *176*, 108432. https://doi.org/10.1016/j.compbiomed.2024.108432

[38] Al-Ssulami, A.M., Alsorori, R.S., Azmi, A.M. et al. Improving Coronary Heart Disease Prediction Through Machine Learning and an Innovative Data Augmentation Technique. CognComput 15, 1687–1702 (2023). https://doi.org/10.1007/s12559-023-10151-6

[39] Aggarwal, S., & Pandey, K. (2023). Early identification of PCOS with commonly known diseases: obesity, diabetes, high blood pressure and heart disease using machine learning techniques. *Expert Systems with Applications*, *217*, 119532. https://doi.org/10.1016/j.eswa.2023.119532

[40] Hasanova, H., Tufail, M., Baek, U. J., Park, J. T., & Kim, M. S. (2022). A novel blockchain-enabled heart disease prediction mechanism using machine learning. *Computers and Electrical Engineering*, *101*, 108086. https://doi.org/10.1016/j.compeleceng.2022.108086

[41] Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 4514-4523. https://doi.org/10.1016/j.jksuci.2020.10.013

[42] Raj, S. N., Vani, R. S., Raja, B., Harsha, T. S., Drakshayani, T., & Charith, R. (2024). HEART DISEASE DETECTION USING XGB-CLASSIFIER AND FAILURE PREDICTION USING GRADIENT BOOSTING. *Journal of Nonlinear Analysis and Optimization*, *15*(1).

[43] Maji, K., Gupta, S., & Dutta, P. K. (2024, December). Enhancing heart disease prediction accuracy: comprehensive analysis of XGBoost and AdaBoost. In *IET Conference Proceedings CP913* (Vol. 2024, No. 37, pp. 130-136). Stevenage, UK: The Institution of Engineering and Technology. https://doi.org/10.1049/icp.2025.0833

[44] Ganie, S. M., Pramanik, P. K. D., Malik, M. B., Nayyar, A., & Kwak, K. S. (2023). An Improved Ensemble Learning Approach for Heart Disease Prediction Using Boosting Algorithms. *Comput. Syst. Sci. Eng.*, *46*(3), 3993-4006. http://dx.doi.org/10.32604/csse.2023.035244

[45] Nissa, N., Jamwal, S., & Neshat, M. (2024). A technical comparative heart disease prediction framework using boosting ensemble techniques. *Computation*, *12*(1), 15. https://doi.org/10.3390/computation12010015

[46] Mubaarok, M. N., Saragih, T. H., Indriani, F., Farmadi, A., & Rizal, A. (2024). Comparison of the Adaboost Method and the Extreme Learning Machine Method in Predicting Heart Failure. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(3), 253-263. https://doi.org/10.35882/jeeemi.v6i3.440

[47] Rajendran, N. A., & Vincent, D. R. (2021). Heart disease prediction system using ensemble of machine learning algorithms. *Recent Patents on Engineering*, *15*(2), 130-139. https://doi.org/10.2174/1872212113666190328220514

[48] Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., Hossain, R., & Hossain, K. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023. *BMC cardiovascular disorders*, *24*(1), 214. https://doi.org/10.1186/s12872-024-03883-2