THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 13 Issue 1

RESEARCH ARTICLE

Machine Learning Strategies for Improved Cardiovascular Disease Detection

Lionel Chong [1], Gazi Husain [1], Daniel Nasef [2], Prince Vathappallil [1], Mihir Matalia [3], Milan Toma [2]

[1] Department of Anatomy, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA.
[2] Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA.
[3] Academic Technologies Group, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA.

OPEN ACCESS

PUBLISHED
31 January 2025

CITATION
Chong, L., Husain, G., et al., 2025. Machine Learning Strategies for Improved Cardiovascular Disease Detection. Medical Research Archives, [online] 13(1). https://doi.org/10.18103/mra.v13i1.6245

DOI
https://doi.org/10.18103/mra.v13i1.6245

ISSN
2375-1924

ABSTRACT

Machine learning offers potential to enhance cardiovascular diagnostics by analyzing various types of data, including but not limited to medical imaging. However, selecting the appropriate ML algorithm for predicting specific cardiovascular diseases is complex and depends on factors such as the type of data, the imaging modality, and the characteristics of the disease. This review aims to provide guidance on selecting suitable machine learning algorithms for diagnosing and predicting specific cardiovascular conditions using various medical imaging modalities. Recent studies were reviewed, focusing on machine learning algorithms applied to cardiovascular imaging for the classification and prediction of cardiovascular diseases. Performance metrics such as accuracy and area under the curve were considered to evaluate the models. A summary table was created to compare the effectiveness of different machine learning algorithms across various cardiovascular conditions and imaging techniques. The review demonstrates that different machine learning algorithms have unique strengths depending on the imaging modality and the specific cardiovascular disease. For example, convolutional neural networks are effective for processing image data like echocardiograms, while support vector machines and random forests are suitable for structured, tabular data. Many studies lack external validation and have issues such as data leakage, raising concerns about the generalizability of their results. Guidelines are provided to help clinicians and researchers select the most appropriate machine learning model based on the medical condition and imaging modality. Additionally, the importance of proper evaluation of machine learning studies is emphasized, including data splitting strategies, learning curves, and validation methods, to ensure the reliability of machine learning models in cardiovascular diagnostics. Selecting the appropriate machine learning algorithm for cardiovascular diagnostics is important and should be guided by the specific disease, imaging modality, and data characteristics. Robust evaluation and validation of machine learning models are essential to ensure their generalizability and clinical utility. Collaboration between medical professionals and machine learning experts is necessary to develop transparent, robust models that can improve patient outcomes in cardiovascular care.

Introduction

The selection of a machine learning (ML) algorithm for predicting specific cardiovascular disorders is a complex process that necessitates careful consideration of various factors.[1] This selection is typically guided by the algorithm's performance relative to other ML algorithms,[2] evaluated through a series of tests and validations using real-world cardiovascular data.[3] Different ML algorithms exhibit varying effectiveness depending on the type of data being analyzed.[4] For example, Convolutional Neural Networks (CNNs) are particularly well-suited for processing image data, making them a popular choice for tasks in cardiovascular medical imaging. These are designed to automatically and adaptively learn spatial hierarchies of features from input images, which is particularly advantageous in the realm of cardiovascular imaging, where the data is often complex and high-dimensional.[5]

Conversely, algorithms like Support Vector Machines (SVMs) or Random Forests may prove to be more effective when analyzing structured, tabular data related to cardiovascular disorders. These algorithms excel at handling high-dimensional data and are less prone to overfitting, making them particularly suitable for tasks such as predicting disease outcomes based on patient records, including risk assessments for conditions like coronary artery disease or heart failure. The specific cardiovascular disorder in question significantly influences the choice of an ML algorithm, as certain diseases may exhibit distinct patterns or characteristics that are more readily detected by specific algorithms. Therefore, the selection of the most appropriate algorithm is contingent upon the specific task, the nature of the data, and the cardiovascular disease (CVD) being addressed.

Performance metrics are a crucial factor in the decision-making process for selecting ML algorithms for cardiovascular disorders.[6] These metrics provide a quantitative assessment of an algorithm's effectiveness in detecting and diagnosing CVDs. Common metrics include accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The choice of metric often depends on the specific task and the relative importance of different types of errors. For example, in scenarios where false negatives can have severe consequences, such as in the detection and classification of CVDs, recall may be prioritized over overall accuracy to ensure that critical cases are not overlooked.

This review highlights a variety of ML algorithms utilized for processing medical images specifically aimed at diagnosing and predicting cardiovascular disorders. Furthermore, this paper presents a systematic framework for selecting a ML algorithm for diagnosis of specific cardiac pathologies. The methodology takes a comprehensive approach by considering three key factors: the particular cardiovascular condition being studied, the chosen imaging modality, and the algorithm's performance metrics. By integrating these elements, the framework offers clear guidance for selecting and optimizing ML models best suited for specific cardiovascular diagnostic tasks.

These algorithms include Decision Trees (DT), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Deep Learning techniques, Convolutional Neural Networks (CNN), Light Gradient Boosting Machine (LightGBM), Linear Discriminant Analysis (LDA), Google Teachable Machine, Naive Bayes (NB), Random Forests (RF), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost). Each algorithm possesses unique strengths and is selected based on the specific requirements of the cardiovascular diagnostic task at hand. The integration of these algorithms in cardiovascular medical diagnostics holds great promise for enhancing both accuracy and efficiency in the detection of heart-related diseases.

Medical Imaging Modalities for Cardiovascular Disorders

Medical imaging plays a pivotal role in the diagnosis, management, and treatment of cardiovascular diseases. Non-invasive and minimally invasive imaging techniques provide valuable insights into the structure and function of the heart and blood vessels, enabling healthcare professionals to accurately diagnose conditions, assess disease progression, and plan appropriate interventions.

The mind map presented in Figure 1 showcases the key imaging modalities commonly used in the assessment of cardiovascular diseases, including X-rays, ultrasound, and MRI. This diverse landscape of medical imaging modalities is intricately linked to various medical conditions, offering a comprehensive overview that will inform our subsequent exploration of this complex field and its intersection with ML.

The development of medical imaging modalities has transformed the landscape of cardiovascular healthcare, offering insights into various cardiovascular disorders and facilitating accurate diagnoses. These advanced imaging techniques have become essential in understanding the complexities of conditions such as coronary artery disease, heart failure, and arrhythmias. With the emergence of ML, the potential to enhance these imaging modalities is immense, promising to elevate diagnostic capabilities and redefine the future of cardiovascular diagnostics.

Figure 1: This mind map illustrates the diverse medical imaging modalities available for cardiac imaging. This paper focuses on the most commonly used imaging techniques, each represented by a distinct color. It also provides a comprehensive overview of these imaging tools, highlighting their specific applications in different cardiac pathologies.

Each of the cardiovascular disorders previously mentioned, ranging from coronary artery disease and heart failure diagnosed through advanced imaging techniques to arrhythmias identified via echocardiography, can now be detected with greater speed and precision. This enhanced capability is achieved through the integration of ML algorithms with these imaging modalities, enabling the analysis of vast amounts of data and the identification of patterns that may be overlooked by the human eye, ultimately leading to more accurate diagnoses and improved patient outcomes.

Inclusion Criteria

This review includes recent studies on cardiovascular disorders that used metrics such as the Area Under the Curve (AUC) and accuracy to evaluate the performance of ML models designed to predict disease presence. These metrics assess different aspects of model performance and are applied in various contexts within CVD prediction. Accuracy is a straightforward metric commonly used in binary classification problems, such as determining whether a patient has a CVD or not. It calculates the proportion of correct predictions–both true positives (diseased patients correctly identified) and true negatives (healthy patients correctly identified)–out of all predictions made. However, accuracy can be misleading in cases where there is a significant class imbalance, which is often the case in medical datasets. For example, if 95% of patients do not have a particular cardiovascular condition (Class A) and only 5% do (Class B), a model that always predicts "no disease" will achieve an accuracy of 95%. Despite this seemingly high accuracy, the model fails to identify any actual cases of the disease, rendering it ineffective for diagnostic purposes.

The AUC, on the other hand, is derived from the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold settings. It measures the model's ability to distinguish between classes across all thresholds, providing an aggregate performance metric that is not sensitive to class imbalance. In the context of CVD prediction, where positive cases are relatively rare compared to negative cases, AUC offers a more reliable assessment of a model's discriminative ability. It's important to note that a model can have high accuracy but low AUC, and vice versa. For instance, a model that predicts every patient as healthy (the majority class) may have high accuracy due to the prevalence of healthy individuals, but it will have a low AUC because it cannot distinguish between patients with and without the disease. Conversely, a model that effectively ranks patients by their risk of CVD may achieve a high AUC but might have

lower accuracy if the classification threshold is not optimally set for prediction purposes.

Therefore, while both AUC and accuracy are useful metrics in evaluating ML models for cardiovascular disorders, they provide different insights into model performance. Accuracy reflects the overall rate of correct predictions but can be misleading in imbalanced datasets. AUC assesses the model's ability to distinguish between diseased and healthy patients across all thresholds and is better suited for evaluating performance when positive cases are scarce. The choice of metric should align with the specific objectives of the study and consider the class distribution within the dataset to ensure that the model is effective in identifying patients with CVDs.

Guided Selection of Machine Learning Models

While each dataset is distinct, this section outlines the process for selecting an appropriate ML algorithm tailored for predicting specific cardiovascular conditions based on various medical imaging modalities. The table below (Table 1) summarizes recent relevant studies published between 2015 and 2024. However, in prioritizing studies for inclusion in this review, preference was given to the most recent publications. Most of these studies performed comparative analyses of multiple ML models to identify the most effective ones, and only the top-performing algorithms are included in the table.

Table 1: Summary of ML algorithms utilized in cardiovascular imaging classification tasks described following this table. (Note that the reported values may be overly optimistic when generalized to a broader patient population due to inherent limitations within the studies.)

Medical Condition	Imaging Modality	ML Algorithm	Accuracy	AUC	Ref.	Year
Coronary artery disease	Stress Echocardiogram	SVM	N/A	0.934	[7]	2022
		RF	N/A	0.934	[7]	2022
		LR	N/A	0.934	[7]	2022
	Echocardiogram	Gradient boosting	85.2%	0.852	[8]	2023
		LGBoost	>80%	0.824	[8]	2023
		RF	>80%	0.817	[8]	2023
		Catboost	>80%	0.829	[8]	2023
		XGBoost	>80%	0.824	[8]	2023
	Cardiac CT	ABoosted Ensemble	N/A	0.790	[9]	2016
		SVM	94%	0.940	[10]	2015
		DL	76%	0.780	[11]	2019
	PET	CNN	N/A	0.900	[12]	2024
Cardiomyopathy	Echocardiogram	RF	~75%	0.900	[13]	2023
		LR	~75%	0.925	[13]	2023
		XGBoost	~75%	0.934	[13]	2023
	Cardiac MRI	BRL	80.72%	0.796	[14]	2015
Adverse cardiovascular events	Risk factors	KNN	85.94%	N/A	[15]	2020
		LR	87.5%	N/A	[15]	2020
		SVC	86.72%	N/A	[15]	2020
		DT	71.88%	N/A	[15]	2020
		MLP	62.5%	N/A	[15]	2020
		RF	86.72%	N/A	[15]	2020
		LightGBM	79.69%	N/A	[15]	2020
		Gradient boosting	83.59%	N/A	[15]	2020
Cardiomegaly	X-ray	CNN	81.2%	0.900	[16]	2024
Cardiac Function Assessment	Echocardiogram	CNN	97.8%	0.996	[17]	2017
		DL	98.8%	N/A	[18]	2021
		DL	92.5%	N/A	[18]	2021
		DeepNN	N/A	0.880	[19]	2021
Hypertrophic cardiomyopathy	Cardiac MRI	3D CNN	98%	N/A	[20]	2019

These studies focus on particular cardiovascular diseases and utilize specific imaging techniques, capturing essential details such as the disease being investigated, and the imaging modality employed. The table also presents key performance metrics, including accuracy and AUC, for the validation cohort; however, metrics for the testing cohort are not included. The validation set, which is used for model tuning and selection, serves as a more reliable indicator of the model's expected performance on new, unseen data.

Accuracy measures the model's overall ability to correctly classify cases, while AUC evaluates its discriminative power. However, both metrics are crucial for assessing a model's effectiveness on unseen data. It is important to note that some studies have reported either accuracy or AUC, which is a significant oversight. Since both metrics provide complementary insights into model performance, they should be reported together to enable a thorough and reliable evaluation. Future studies should ensure that both accuracy and AUC are included in their findings to deliver a complete assessment of model performance and facilitate accurate comparisons across different ML approaches.

The following guidelines delineate how physicians or researchers can utilize this summary to select the most appropriate ML model for diagnostic purposes, contingent upon the specific medical condition and the imaging modality employed:

- Step 1 (Identify the cardiovascular disease): Determine the specific CVD that necessitates diagnosis. For example, when diagnosing coronary artery disease, it is essential to focus on the entries in Table 1 that pertain to this condition.
- Step 2 (Identify the imaging modality): Ascertain the imaging modality that is available for use. This may include modalities such as computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound. In Table 1, examine the rows that correspond to both the identified medical condition and the available imaging modality.
- Step 3 (Refer to comparative analysis): Consult the comparative analysis presented in Table 1. This table details the ML models that have been employed for the specified medical condition and imaging modality, along with their associated performance metrics, including accuracy and area under the curve (AUC).
- Step 4 (Choose the ML model): Select the ML model based on the comparative analysis. For instance, if the diagnosis pertains to coronary artery disease and the imaging modality is CT, one would consider models such as linear discriminant analysis (LDA) and logistic regression (LR), which exhibit comparable accuracy and AUC. If the imaging modality is MRI, the choice may be between logistic regression and XGBoost. In the case of ultrasound, the study referenced in the last column indicates that convolutional neural networks (CNN) are the most effective algorithm.
- Step 5 (Apply the chosen ML model): After selecting the appropriate ML model, proceed to implement it within the diagnostic process.

Table 1 provides a summary of comparative analyses of various ML models applied to different CVDs and imaging modalities. It eliminates the necessity for conducting new comparative evaluations each time a diagnosis is needed for a cardiovascular condition. By referencing this table, physicians and researchers can swiftly identify the most suitable ML algorithm for predicting specific cardiovascular conditions using a designated imaging modality. This approach streamlines the decision-making process and enhances the efficiency of diagnostics.

A considerable number of the studies reviewed lack external prospective validation, and several demonstrate significant issues related to data leakage. These limitations raise concerns regarding the accuracy values reported, suggesting they may be overly optimistic when generalized to a broader patient population. External prospective validation is an essential component in the assessment of ML models, as it entails evaluating the model on a completely independent dataset that was not utilized during the training or internal validation phases.[21]. This independent dataset is typically gathered in a different context or at a later time, thereby providing a more realistic assessment of the model's performance in real-world scenarios. The absence of such validation in many studies listed in Table 1 casts doubt on the generalizability of their findings across diverse clinical settings.

Data leakage, conversely, occurs when information from outside the training dataset is inadvertently incorporated into the model, resulting in inflated performance metrics.[22] This situation can arise when features from validation or test sets are utilized during the training process, granting the model access to information that would not be available in actual clinical practice. The presence of significant data leakage issues in several studies further undermines the reliability of their reported accuracy values, thereby compromising the true diagnostic potential of the models.[23]

While the selection of an appropriate ML algorithm for medical image classification is undeniably critical, it is equally important to first clearly define the clinical question at hand. This initial step not only informs the choice of algorithm but also guides decisions regarding data collection, feature selection, and the evaluation metrics employed in the study.[24] Furthermore, despite the aggregation of numerous studies, it is essential to acknowledge that each dataset possesses unique characteristics and complexities. These inherent differences can significantly influence the selection of the most suitable ML algorithm.

In medical imaging, the choice of algorithm can profoundly affect the performance and accuracy of image analysis. Specific attributes of the imaging dataset, such as its size, complexity, and type, must be considered when determining the appropriate algorithm. Additionally, the nature of the data (whether structured or unstructured) can play a decisive role in guiding the algorithm selection process. Tailoring the ML approach to the specific characteristics of the dataset is important for achieving optimal diagnostic outcomes.

Figure 2 illustrates a flowchart that delineates the process of selecting and implementing an ML algorithm specifically for the classification of CVDs. The flowchart commences with the initial input conditions, which encompass the specific cardiovascular condition being addressed and the selected imaging modality. Subsequently, the process advances to the selection of an appropriate ML algorithm, based on the information summarized in Table 1.

Afterwards, the model undergoes training and evaluation, leading to a critical decision point where performance metrics are assessed. If the performance metrics meet the established thresholds, the process proceeds to the final diagnostic step. Conversely, if the metrics are deemed unsatisfactory, the process enters an algorithm optimization loop. During this phase, alternative ML algorithms are tested and evaluated before returning to the model training stage for further refinement. This iterative approach ensures the selection of the most effective algorithm for accurate and reliable classification of CVDs.

In Figure 2, "satisfactory metrics" refers to the performance measures of the ML model, such as accuracy,

AUC, sensitivity, and specificity, that are considered adequate for the intended diagnostic application in CVDs. The precise threshold for what constitutes "satisfactory" varies based on the specific cardiovascular condition being addressed and the imaging modality employed. To evaluate whether the metrics are satisfactory, researchers and clinicians can compare their model's performance to the metrics reported in Table 1 for similar studies focused on CVDs. This comparative analysis enables them to assess their model's efficacy in relation to other published work within the same domain. The process entails identifying studies in Table 1 that concentrate on the same cardiovascular condition and imaging modality as the current research, and subsequently comparing the model's accuracy and AUC (when available) to those reported in analogous studies.

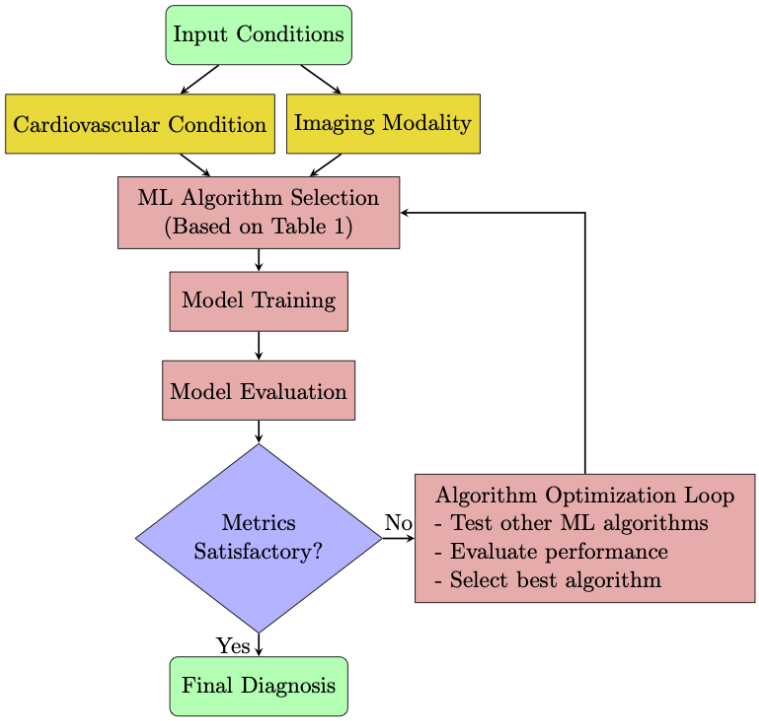

Figure 2: Flowchart for the selection and implementation of ML algorithms in the classification of cardiovascular diseases. This process integrates the specific cardiovascular condition, the chosen imaging modality, and the performance metrics of the algorithms to guide the selection and optimization of ML models for diagnostic tasks.

Several factors should be considered when evaluating whether the performance metrics of a model for CVD classification are satisfactory:

- Contextual Relevance: A model with slightly lower performance metrics may still be deemed satisfactory if it offers additional advantages, such as reduced processing time or lower computational demands, which can be critical in clinical settings.
- Clinical Significance: Even if a model's metrics fall below those reported in some published studies, they may still be acceptable if they meet the minimum thresholds for clinical utility specific to CVD diagnostics.
- Dataset Characteristics: If the dataset utilized is more complex or diverse than those employed in previous studies, achieving metrics comparable to those reported in the literature could be considered satisfactory.
- Performance Trade-offs: A minor reduction in one metric (e.g., accuracy) may be acceptable if it results in substantial improvements in another critical metric (e.g., AUC) that is particularly relevant for CVD applications.

By comparing a model's performance to the metrics reported in Table 1 and taking these factors into account, researchers can make informed decisions regarding the adequacy of the model's metrics for their specific CVD classification tasks. This approach facilitates a more nuanced evaluation that reflects the current state of the art in the field.

Evaluating Machine Learning Studies for Image Classification

The application of ML in cardiovascular imaging is necessitating a critical evaluation of the validity and reliability of these studies. A thorough understanding of how to assess ML studies is essential to ensure that their results are trustworthy and that the models developed are robust and generalizable. This section focuses on methods to evaluate ML studies in medical imaging classification, emphasizing the identification of potential issues such as overfitting or underfitting, evaluating performance metrics, and recognizing the significance of proper data splitting and validation strategies.

Data Splitting Strategies

A fundamental aspect of developing and evaluating ML models is the division of data into distinct subsets: training, validation, and testing sets. Proper data splitting is crucial to prevent overfitting and to ensure that the model's performance is generalizable to new, unseen data. The training set is used to train the ML model, allowing it to learn patterns from the data. The validation set is used during model development to tune hyperparameters and make decisions about model architecture, providing an unbiased evaluation of the model's performance during training. The testing set is used to provide an unbiased

evaluation of the final model's performance after training and validation and should only be used once the model is fully trained.

Potential data splitting strategies include hold-out validation, where the dataset is split once into training, validation, and testing sets. Common proportions for this split include 60% training, 20% validation, and 20% testing; however, other ratios such as 70% training, 15% validation, and 15% testing, or 80% training, 10% validation, and 10% testing are also frequently used, depending on the size of the dataset and the specific requirements of the study. Another strategy is k-fold cross-validation, where the dataset is partitioned into k equal-sized folds. The model is trained and validated k times, each time using a different fold as the validation set and the remaining folds as the training set, maximizing the use of available data. Additionally, stratified splitting ensures that each subset has the same class distribution as the original dataset, which is particularly important in cases of class imbalance. Proper data splitting techniques are essential to prevent information leakage between datasets, which can artificially inflate performance metrics and lead to overly optimistic conclusions about a model's capabilities.

Learning Curves and Model Convergence

Learning curves are valuable tools for diagnosing the learning behavior of ML models. They plot the model's performance on the training and validation sets over successive iterations or epochs, providing insights into how well the model is learning and whether it is overfitting or underfitting. The training curve illustrates the model's performance on the training data, while the validation curve reflects the model's performance on the validation data.

Healthy learning curves for a converged ML classification task typically exhibit convergence, where both training and validation curves approach a plateau, indicating that the model's performance is stabilizing. A small gap between the training and validation performance suggests good generalization, and plateauing values indicate that further training may not significantly improve performance. Signs of overfitting include the training curve showing high performance while the validation curve lags or declines, indicating that the model is learning noise specific to the training data and not generalizing well. Underfitting is indicated when both training and validation curves show poor performance and do not improve with more training, suggesting that the model is too simple to capture the underlying patterns in the data. Analyzing learning curves helps in assessing whether the model has been appropriately trained and whether it is likely to perform well on new data.

Receiver Operating Characteristic Curves and Performance Metrics

The Receiver Operating Characteristic (ROC) curve is an important tool for evaluating classification models, especially in medical diagnostics where the balance between sensitivity and specificity is critical. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 minus specificity) at various threshold settings. The Area Under the ROC Curve (i.e., AUC) is a scalar value summarizing the model's performance across all classification thresholds; a higher AUC indicates better discriminative ability.

Interpreting ROC curves for validation and testing sets involves looking for consistent performance, where similar ROC curves and AUC values for the validation and testing sets suggest that the model generalizes well. Discrepancies between curves, such as significant differences in ROC curves between validation and testing sets, may indicate overfitting, where the model performs well on validation data but poorly on unseen testing data. Large differences in AUC between validation and testing sets suggest overfitting, while consistently low AUC values on both sets indicate underfitting.

Confusion Matrices and Model Performance

A confusion matrix provides a detailed breakdown of a model's classification performance, showing the counts of true positives, true negatives, false positives, and false negatives. Analyzing confusion matrices entails comparing the validation and testing sets; the confusion matrix for the testing set should show performance metrics slightly lower than but close to those of the validation set. Significant discrepancies may indicate overfitting. Consistency in errors, such as similar patterns of errors across validation and testing sets, suggests that the model reliably captures data patterns.

Expected differences between validation and testing performance include a slight decrease in performance on the testing set, which is normal due to its nature as completely unseen data. A large drop in performance may indicate that the model has not generalized well and may have overfitted to the training and validation data.

Importance of External Validation

External validation involves testing the model on data from different settings or populations than those used in training and validation, providing insight into the model's generalizability. Benefits of external testing include assessing generalizability, demonstrating that the model performs well across different populations, imaging devices, or clinical settings, and enhancing credibility, as models validated externally are more likely to be trusted by the medical community.

Challenges in obtaining external data involve data access difficulties due to privacy concerns, data sharing restrictions, and lack of standardized data formats, as well as variability in data due to differences in data acquisition protocols, equipment, and patient demographics. While external validation is ideal, its absence necessitates careful consideration of the model's generalizability and applicability to broader clinical practice.

Identifying High-Quality Machine Learning Studies

In the assessment of ML studies within medical imaging classification, it is essential to identify the characteristics that distinguish robust models from those susceptible to overfitting or underfitting. Recognizing these indicators

aids in evaluating the validity, reliability, and generalizability of the ML models presented. Table 2 provides a comprehensive summary of key aspects to consider when determining whether an ML study represents a healthy model or exhibits signs of overfitting or underfitting. This table serves as a practical guide for researchers and practitioners to critically appraise ML studies and ensure the integrity of their findings.

Table 2: Summary of Indicators for Robust, Overfitting, and Underfitting ML Studies

Aspect	Robust ML Study	Overfitting ML	Underfitting ML
Data Splitting	Proper division into training, validation, and testing sets with no data leakage; appropriate proportion and stratification if necessary	Possible improper splitting or data leakage leading to artificially inflated performance on validation set	May have insufficient data or improper splitting, leading to poor learning of patterns
Performance Metrics	Consistent and high metrics (e.g., accuracy, precision, recall) across training, validation, and testing sets; metrics reflect true performance	Metrics are high on training and validation sets but substantially lower on testing set; indicates over-reliance on training data	Low metrics across all datasets; model fails to perform above baseline levels
Learning Curves	Training and validation curves converge and plateau at similar performance levels; small gap between curves indicating good generalization	Training curve shows high performance (e.g., low loss), but validation curve lags behind or worsens; large gap between curves indicating model is not generalizing	Both training and validation curves show poor performance; curves may plateau at low performance levels, indicating model is too simple
ROC Curves and AUC	Similar ROC curves and AUC values for validation and testing sets; high AUC indicating good discriminative ability	High AUC on validation set but significantly lower AUC on testing set; ROC curves differ markedly, suggesting model has learned noise from training data	Low AUC values on both validation and testing sets; ROC curves indicate poor ability to distinguish between classes
Confusion Matrices	Testing set confusion matrix shows slightly lower but comparable performance to validation set; errors are consistent across sets	Significant drop in performance from validation to testing set; confusion matrix shows model performs poorly on unseen data	Poor performance on both validation and testing sets; high misclassification rates indicating inability to capture patterns
Generalization Ability	Model generalizes well to unseen data; performs reliably across different datasets	Model does not generalize; performs well on training data but poorly on new, unseen data	Model cannot capture underlying patterns; fails to generalize due to oversimplification
Model Complexity	Appropriate complexity matching the problem; neither too simple nor too complex; uses regularization techniques if necessary	Model is overly complex; may have too many parameters leading to memorization of training data	Model is too simple; lacks the capacity to learn the necessary patterns in data
External Validation	Includes external validation using independent datasets from different settings; confirms model's generalizability	Lacks external validation; performance may be inflated due to overfitting on internal data	May not reach stage of external validation due to poor performance on internal data
Methodological Rigor	Detailed and transparent reporting of methods; proper use of cross-validation; avoidance of data leakage; reproducible results	Insufficient methodological details; potential data leakage; lack of transparency hindering assessment of validity	Methodology may be inadequate; poor model selection and training procedures

Red flags indicating potential issues include lack of validation or testing data, with studies reporting results only on training data likely overestimating performance. Unrealistically high performance, such as exceptionally high accuracy without corresponding validation on external data, may indicate overfitting. Insufficient methodological details, such as missing information on model training or evaluation, hinder assessment of validity.

Critical evaluation of ML studies in cardiovascular imaging classification is essential to ensure that models are reliable, valid, and generalizable. By understanding data splitting strategies, interpreting learning curves and performance metrics, and assessing methodological rigor, it is possible to make informed judgments about the trustworthiness of reported results. As ML continues to advance in the medical field, such scrutiny plays an important role in effectively and responsibly translating these technologies into clinical practice.

Machine Learning Algorithms in Cardiovascular Diagnostics

From analyzing imaging data to predicting patient outcomes based on clinical parameters, ML has been applied across a wide range of cardiovascular diagnostic tasks. The diverse applications of ML algorithms in this field highlight their transformative potential to enhance diagnostic precision and improve patient care. By leveraging advanced computational techniques, ML enables the identification of complex patterns in cardiovascular data that may not be apparent through traditional methods. This section explores the implementation of various ML algorithms in cardiovascular diagnostics, emphasizing their versatility and impact.

Decision Trees (DT) were utilized in a study by Hsu et al., which developed a hemogram-based decision tree model to evaluate the relationship between the current probability of metabolic syndrome and the risk of future CVD, hypertension, and type 2 diabetes mellitus.[25] Support Vector Machines (SVM) were employed in a study by Lee et al., which developed intravascular ultrasound (IVUS)-based supervised ML algorithms to identify coronary lesions.[26] K-Nearest Neighbor (KNN) was applied in another study by Lee et al., which aimed to enhance the diagnostic accuracy of treadmill exercise tests using an ML-based algorithm.[27] Logistic Regression (LR) was used in a study by Cheng et al. to analyze serum interferon health data in coronary heart disease, comparing its performance with artificial neural network models.[28] Convolutional Neural Networks (CNNs) were explored in studies by Sun et al. and Mamun et al., which developed AI-based approaches for screening patients and predicting heart conditions.[29,30] Light Gradient Boosting Machine (LightGBM) was assessed in a study by Lee et al. for its predictive performance in evaluating risk scores in an Asian Brugada Syndrome population.[31] General Linear Model Boosting (GLMB) was examined in a study by Asadi et al., which identified the most effective tree-based ML method for detecting CVD.[32] A study conducted by Morguet et al. assessed 15 patients (ages 36–75, median age 59) with stable single-vessel disease (≥ 70% stenosis) and regional wall-motion abnormalities using multiple diagnostic techniques and analyzed with linear discriminant analysis.[33] A study conducted by Ciaccio et al. developed an algorithm to extract morphological components based on frequency, and its effectiveness in distinguishing complex fractionated atrial electrograms (CFAE) was evaluated.[34] Naïve Bayes (NB) was utilized in a study by Miranda et al. to develop a mining model for detecting CVD and assessing its risk level in adults. [35] Random Forests (RF) were applied in a study by Ambale-Venkatesh et al. to predict six cardiovascular outcomes using random survival forests.[36] Adaptive Boosting (AdaBoost) was explored in a study by Zhu et al. to assess the quality of ECG signals.[37] Extreme Gradient Boosting (XGBoost) was used in a study by Xie et al. to differentiate Left Ventricular Reverse Remodeling (LVRR) from non-LVRR in patients with newly diagnosed dilated cardiomyopathy.[38] Lastly, RoboFlow 3.0 Object Detection (RF3) was employed in a study by Ega et al., which developed a non-invasive blood pressure (NIBP) device reading system using ESP32-CAM and the YOLOv5 algorithm, leveraging Roboflow for NIBP digit recognition.[39] These studies collectively demonstrate the diverse applications of ML algorithms in cardiovascular diagnostics, emphasizing their potential to improve diagnostic accuracy and patient outcomes.

Discussion

Selecting an appropriate ML algorithm for predicting specific cardiovascular conditions is a multifaceted process that involves several challenges unique to the cardiovascular domain. These challenges include the need for large, annotated datasets specific to cardiovascular imaging modalities, ensuring the model's generalizability across diverse patient populations with varying cardiovascular risk profiles, and addressing issues related to overfitting and underfitting in model development.

One of the primary obstacles in developing robust ML models for CVD prediction is the requirement for extensive, high-quality datasets. Cardiovascular imaging data, such as echocardiograms, cardiac MRI, and CT scans, require detailed expert annotation to capture the nuances of cardiovascular pathology. The availability of such datasets is often limited due to privacy concerns, the labor-intensive nature of data annotation, and the need to represent diverse populations with different CVD patterns. This limitation can lead to models that perform well on training data but fail to generalize to new, unseen patient populations, highlighting the risk of overfitting.

Out of all the studies cited in this review, only two–Lee et al. and Ambale-Venkatesh et al.–utilized multicenter data, enhancing the generalizability of their findings. Lee et al. conducted a retrospective, multicenter cohort study assessing the predictive performance of various risk scores in an BrS population, including an intermediate-risk subgroup.[31] By incorporating data from multiple centers, this study addressed the variability inherent in patient populations and reduced the risk of overfitting the model to a single-center dataset. Similarly, Ambale-Venkatesh et al. assessed the ability of random survival forests to predict six cardiovascular outcomes using data from the Multi-Ethnic Study of Atherosclerosis (MESA), comparing its performance to traditional cardiovascular risk scores.[36]

The multicenter nature of the MESA dataset allowed for a more diverse representation of CVD manifestations, enhancing the robustness and applicability of the ML model across different populations. Including participants from multiple sites, each sites having distinct clinical practices, patient demographics, and treatment approaches, can provide a more accurate reflection of actual healthcare environments. By distributing the study population across diverse locations, the risk of local factors at any single hospital or clinic skewing the results is minimized. This broader scope may explain why the MESA dataset had improved the external validity of findings by more closely capturing the variability inherent in real-world clinical settings, thereby enhancing reliability. Furthermore, involving multiple centers can accelerate patient recruitment and yield a more heterogeneous participant pool, which strengthens the generalizability and practical relevance of the study's outcomes. As such, multicenter research is increasingly recognized as a critical strategy for enhancing both the impact and reproducibility of healthcare studies, particularly for ML applications that must remain robust across varied and evolving patient populations.

Recognizing robust ML studies is crucial for clinicians and researchers aiming to apply these models in practice. Robust studies often include key features such as the use of multicenter data, appropriate data splitting strategies to prevent data leakage, thorough validation using separate testing datasets, and external validation when possible. Overfitting and underfitting are significant red flags in ML studies. Overfitting occurs when a model learns the noise in the training data rather than the underlying pattern, leading to poor performance on new data. Underfitting happens when a model is too simplistic to capture the complexity of the data, resulting in poor performance both on training and unseen data. Evaluating ML studies in cardiovascular imaging requires careful scrutiny of their methodology. Studies should report comprehensive performance metrics, including both accuracy and AUC, to provide a complete picture of the model's predictive capabilities[40]. The lack of reporting either metric, as noted in some studies within this review, can hinder the assessment of a model's true performance. Moreover, proper data splitting into training, validation, and testing sets is essential to avoid data leakage and ensure that the model's performance on unseen data is a reliable indicator of its real-world applicability. Cross-validation techniques and the use of external datasets for validation, especially from multiple centers, can enhance the model's credibility and generalizability.

Moreover, transparency, explainability, and reproducibility are important factors to consider before implementing a diagnostic ML algorithm. Transparency about model design decisions, such as feature selection and data preprocessing, ensures other researchers can replicate results and clinicians can more readily interpret model outputs. Explainability, on the other hand, clarifies how a model arrives at its predictions and is vital for fostering trust and accountability in high-stakes scenarios like patient care, where clinicians must quickly assess whether an algorithm's output aligns with established clinical knowledge.

Reproducibility depends on proper validation strategies, including the use of separate training, validation, and testing datasets, as well as external validation through additional data sources. This approach lowers the risk of data leakage and guards against overfitting, where a model performs exceedingly well on training data but poorly on new data. Conversely, it also addresses underfitting by promoting iterative refinement of hyperparameters and model complexity until performance metrics stabilize or improve on both internal and external datasets.

Beyond these technical considerations, implementing regular audits can help detect unintended biases that may arise from imbalanced training data or flawed feature representations. Stress-testing models under challenging conditions, such as simulated data shifts or varying patient demographics, further helps ensure performance consistency before deployment in clinical settings. These steps, coupled with ongoing monitoring of real-world outcomes, can instill greater confidence in ML-driven tools.

The "black box" nature of complex ML models presents additional challenges in cardiovascular healthcare, where understanding the rationale behind a diagnostic decision is critical. Clinicians need transparent and interpretable models to trust and effectively integrate ML into patient care. Efforts to develop explainable AI and adhere to guidelines are necessary to bridge the gap between model complexity and clinical usability.[41]

Hence, while ML has the potential to significantly advance CVD diagnosis and prediction, it is imperative to develop models that are robust, generalizable, and transparent. Collaborations between medical professionals and ML experts are essential to address the intricacies of model development, prevent issues like overfitting and underfitting, and ensure that models are evaluated rigorously.[42] By focusing on these aspects, and giving particular attention to studies that utilize diverse, multicenter data, the cardiovascular field can better harness the power of ML to improve patient outcomes.[43]

Conclusion

The selection and implementation of a ML algorithm for cardiovascular diagnostics is closely linked to the specific characteristics of the data, imaging modality, and disease in question. While convolutional neural networks are particularly effective when dealing with spatially complex image data, other algorithms such as support vector machines and random forests may be more appropriate for structured clinical or tabular data. Each algorithm has its own strengths and weaknesses, underscoring the importance of selecting the most suitable method based on the specific condition and the imaging availabilities.

The wide range of ML tools available and the many upcoming models hold great promise for enhancing diagnostic accuracy, minimizing false negatives, and facilitating earlier detection of cardiovascular disease. However, the lack of external validation and the prevalence of issues like data leakage underscore the ongoing challenge of ensuring the generalizability and

practical utility of these models. To improve the reliability of ML approaches in cardiovascular practice and research, it is crucial to employ rigorous study designs that incorporate appropriate data splitting strategies, learning curves, and robust validation methods.

Future progress will require close collaboration between clinicians and ML experts. By combining their respective expertise in disease pathophysiology, real-world clinical workflows, and algorithmic optimization, these interdisciplinary teams can develop models that effectively address the unique challenges of processing complex cardiovascular data. Transparent reporting and careful consideration of ethical implications will also be essential for building trust and facilitating the successful integration of these models into routine clinical practice.

In the end, cutting-edge ML algorithms and high-quality cardiovascular data has the potential to revolutionize patient care. By identifying the most appropriate algorithm for each diagnostic or prognostic task, healthcare professionals can deliver more personalized treatment plans, optimize clinical decision-making, and usher in a new era of truly data-driven cardiovascular medicine.

Author Contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft preparation), writing (review and editing), visualization: L.C., G.H., D.N., P.V., M.M., M.T. Supervision, project administration, funding acquisition: M.T. All authors have read and agreed to the published version of the manuscript.

Funding Statement

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Competing Interest

The authors affirm that there are no financial or non-financial conflicts of interest associated with this study. We assert that we possess no personal, academic, or other interests that could be perceived as influencing the objectivity, integrity, or value of this research. Additionally, we declare that there are no factors that could potentially impact our decisions or actions concerning the presentation, analysis, or interpretation of the data, which might undermine or be perceived as undermining the objectivity, integrity, and value of this publication.

Institutional Review Board Statement

Not applicable.

Data Availability

The original contributions presented in the study are included in the article.

References

1. An Qi, Rahman Saifur, Zhou Jingwen, Kang James Jin. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges Sensors. 2023;23:4178; Doi: 10.3390/s23094178.
2. Decoux Antoine, Duron Loic, Habert Paul, et al. Comparative performances of machine learning algorithms in radiomics and impacting factors Scientific Reports. 2023;13; Doi: 10.1038/s41598-023-39738-7.
3. Huang Yinan, Li Jieni, Li Mai, Aparasu Rajender R.. Application of machine learning in predicting survival outcomes involving real-world data: a scoping review BMC Medical Research Methodology. 2023;23; Doi: 10.1186/s12874-023-02078-1.
4. Doulah Md. Siraj-Ud, Islam Md. Nazmul. Performance Evaluation of Machine Learn- ing Algorithm in Various Datasets Journal of Artificial Intelligence, Machine Learning and Neural Network. 2023;3:14–32; Doi: 10.55529/jaimlnn.32.14.32.
5. Pasa F., Golkov V., Pfeiffer F., Cremers D., Pfeiffer D.. Efficient Deep Network Archi- tectures for Fast Chest X-Ray Tuberculosis Screening and Visualization Scientific Reports. 2019;9; Doi: 10.1038/s41598-019-42557-4.
6. Thieu Nguyen Van. PerMetrics: A Framework of Performance Metrics for Machine Learning Models Journal of Open Source Software. 2024;9:6143; Doi: 10.21105/joss.06143.
7. Upton Ross, Mumith Angela, Beqiri Arian, et al. Automated Echocardiographic Detection of Severe Coronary Artery Disease Using Artificial Intelligence JACC: Cardiovascular Imaging. 2022;15:715–727; Doi: 10.1016/j.jcmg.2021.10.013.
8. Guo Ying, Xia Chenxi, Zhong You, et al. Machine learning-enhanced echocardiogra- phy for screening coronary artery disease BioMedical Engineering OnLine. 2023;22; Doi: 10.1186/s12938-023-01106-x.
9. Motwani Manish, Dey Damini, Berman Daniel S., et al. Machine learning for predic- tion of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis European Heart Journal. 2016:ehw188; Doi: 10.1093/eurheartj/ehw188.
10. Kang Dongwoo, Dey Damini, Slomka Piotr J., et al. Structured learning algorithm for detection of nonobstructive and obstructive coronary plaque lesions from com- puted tomography angiography Journal of Medical Imaging. 2015;2:014003; Doi: 10.1117/1.jmi.2.1.014003.
11. Kumamaru Kanako K, Fujimoto Shinichiro, Otsuka Yujiro, et al. Diagnostic accu- racy of 3D deep-learning-based fully automated estimation of patient-level minimum fractional flow reserve from coronary computed tomography angiography European Heart Journal - Cardiovascular Imaging. 2019; Doi: 10.1093/ehjci/jez160.
12. Hunter Chad, Moulton Eric, Chong Aun Yeong, Beanlands Rob, deKemp Robert. Deep Learning Improves Diagnosis of Obstructive CAD using Rb-82 PET Imaging of Myocardial Blood Flow Journal of Nuclear Medicine. 2024;65:241721.
13. Zhou Mei, Deng Yongjian, Liu Yi, Su Xiaolin, Zeng Xiaocong. Echocardiography- based machine learning algorithm for distinguishing ischemic cardiomyopathy from dilated cardiomyopathy BMC Cardiovascular Disorders. 2023;23; Doi: 10.1186/s12872-023-03520-4.
14. Gopalakrishnan Vanathi, Menon Prahlad G, Madan Shobhit. cMRI-BED: A novel informatics framework for cardiac MRI biomarker extraction and discovery applied to pediatric cardiomyopathy classification BioMedical Engineering OnLine. 2015;14:S7; Doi: 10.1186/1475-925x-14-s2-s7.
15. Zhang Caiwei, Qu Junhao, Li Weicheng, Zheng Lehan. Predicting Cardio- vascular Events by Machine Learning Journal of Physics: Conference Series. 2020;1693:012093; Doi: 10.1088/1742-6596/1693/1/012093.
16. Sengupta Sourya, Anastasio Mark A.. A Test Statistic Estimation-Based Approach for Establishing Self-Interpretable CNN-Based Binary Classifiers IEEE Transactions on Medical Imaging. 2024;43:1753-1765; Doi: 10.1109/tmi.2023.3348699.
17. Madani Ali, Arnaout Ramy, Mofrad Mohammad, Arnaout Rima. Fast and accu- rate view classification of echocardiograms using deep learning npj Digital Medicine. 2018;1; Doi: 10.1038/s41746-017-0013-1.
18. Narang Akhil, Bae Richard, Hong Ha, et al. Utility of a Deep-Learning Algorithm to Guide Novices to Acquire Echocardiograms for Limited Diagnostic Use JAMA Cardiology. 2021;6:624; Doi: 10.1001/jamacardio.2021.0185.
19. Pandey Ambarish, Kagiyama Nobuyuki, Yanamala Naveena, et al. Deep-Learning Models for the Echocardiographic Assessment of Diastolic Dysfunction JACC: Car- diovascular Imaging. 2021;14:1887–1900; Doi: 10.1016/j.jcmg.2021.04.010.
20. Fahmy Ahmed S., Neisius Ulf, Chan Raymond H., et al. Three-dimensional Deep Con- volutional Neural Networks for Automated Myocardial Scar Quantification in Hyper- trophic Cardiomyopathy: A Multicenter Multivendor Study Radiology. 2020;294:52–60; Doi: 10.1148/radiol.2019190737.
21. Riley Richard D, Archer Lucinda, Snell Kym I E, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study BMJ. 2024:e074820; Doi: 10.1136/bmj-2023-074820.
22. Kapoor Sayash, Narayanan Arvind. Leakage and the reproducibility crisis in machine- learning-based science Patterns. 2023;4:100804; Doi: 10.1016/j.patter.2023.100804.
23. Tampu Iulian Emil, Eklund Anders, Haj-Hosseini Neda. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images Scientific Data. 2022;9; Doi: 10.1038/s41597-022-01618-6.
24. Leek Jeffery T., Peng Roger D.. What is the question? Science. 2015;347:1314–1315; Doi: 10.1126/science.aaa6146.
25. Hsu C H, Chen Y L, Hsieh C H, Liang Y J, Liu S H, Pei D. Hemogram-based decision tree for predicting the metabolic syndrome and cardiovascular diseases in the elderly QJM: An International Journal of Medicine. 2020;114:363–373; Doi: 10.1093/qjmed/hcaa205.

26. Lee June-Goo, Ko Jiyuon, Hae Hyeonyong, et al. Intravascular ultrasound- based machine learning for predicting fractional flow reserve in inter- mediate coronary artery lesions Atherosclerosis. 2020;292:171–177; Doi: 10.1016/j.atherosclerosis.2019.10.022.
27. Lee Yin-Hao, Tsai Tsung-Hsien, Chen Jun-Hong, et al. Machine learning of treadmill exercise test to improve selection for testing for coronary artery disease Atheroscle- rosis. 2022;340:23–27; Doi: 10.1016/j.atherosclerosis.2021.11.028.
28. Cheng Xiaobing, Han Weixing, Liang Youfeng, et al. Risk Prediction of Coronary Artery Stenosis in Patients with Coronary Heart Disease Based on Logistic Regression and Artificial Neural Network Computational and Mathematical Methods in Medicine. 2022;2022:1–8; Doi: 10.1155/2022/3684700.
29. Sun Jin-Yu, Qiu Yue, Guo Hong-Cheng, et al. A method to screen left ventric- ular dysfunction through ECG based on convolutional neural network Journal of Cardiovascular Electrophysiology. 2021;32:1095–1102; Doi: 10.1111/jce.14936.
30. Khan Mamun Mohammad Mahbubur Rahman, Elfouly Tarek. Detection of Cardio- vascular Disease from Clinical Parameters Using a One-Dimensional Convolutional Neural Network Bioengineering. 2023;10:796; Doi: 10.3390/bioengineering10070796.
31. Lee Sharen, Zhou Jiandong, Chung Cheuk To, et al. Comparing the Performance of Published Risk Scores in Brugada Syndrome: A Multi-center Cohort Study Current Problems in Cardiology. 2022;47:101381; Doi: 10.1016/j.cpcardiol.2022.101381.
32. Asadi Fariba, Homayounfar Reza, Mehrali Yaser, Masci Chiara, Talebi Samaneh, Zayeri Farid. Detection of cardiovascular disease cases using advanced tree-based machine learning algorithms Scientific Reports. 2024;14; Doi: 10.1038/s41598-024-72819-9.
33. Morguet Andreas J., Behrens Steffen, Kosch Olaf, et al. Myocardial viability evalu- ation using magnetocardiography in patients with coronary artery disease Coronary Artery Disease. 2004;15:155–162; Doi: 10.1097/00019501-200405000-00004.
34. Ciaccio Edward J., Biviano Angelo B., Garan Hasan. The dominant morphology of fractionated atrial electrograms has greater temporal stability in persistent as compared with paroxysmal atrial fibrillation Computers in Biology and Medicine. 2013;43:2127–2135; Doi: 10.1016/j.compbiomed.2013.08.027.
35. Miranda Eka, Irwansyah Edy, Amelga Alowisius Y., Maribondang Marco M., Salim Mulyadi. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier Healthcare Informatics Research. 2016;22:196; Doi: 10.4258/hir.2016.22.3.196.
36. Ambale-Venkatesh Bharath, Yang Xiaoying, Wu Colin O., et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis Circula- tion Research. 2017;121:1092–1101; Doi: 10.1161/circresaha.117.311312.
37. Zhu Zeyang, Liu Wenyan, Yao Yang, Chen Xuewei, Sun Yingxian, XU Lisheng. Adaboost Based ECG Signal Quality Evaluation in 2019 Computing in Cardiology Conference (CinC)CinC2019; Computing in Cardiology 2019; Doi: 10.22489/cinc.2019.151.
38. Xie Xiangkun, Yang Mingwei, Xie Shan, et al. Early Prediction of Left Ventricular Reverse Remodeling in First-Diagnosed Idiopathic Dilated Cardiomyopathy: A Com- parison of Linear Model, Random Forest, and Extreme Gradient Boosting Frontiers in Cardiovascular Medicine. 2021;8; Doi: 10.3389/fcvm.2021.684004.
39. Ega Adindra Vickar, Ginanjar Gigin, Azzumar Muhammad. Monitoring and data acquisition of automated non-invasive blood pressure reading with ESP32-CAM and YOLO algorithm in Proceedings of the 10th International Conference on Sustainable Energy Engineering and Application 2022 (ICSEEA2022);3069:020114; AIP Publishing 2024; Doi: 10.1063/5.0206265.
40. Thomas Anvin, Jose Rejath, Syed Faiz, Ong Chi Wei, Toma Milan. Machine learning-driven predictions and interventions for cardiovascular occlusions. Technology and Health Care. 2024;32(5):3535–3556; Doi: 10.3233/thc-240582.
41. Chen Haomin, Gomez Catalina, Huang Chien-Ming, Unberath Mathias. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review npj Digital Medicine. 2022;5:156; Doi: 10.1038/s41746-022-00699-2.
42. Gazi Husain, Jonathan Mayer, Molly Bekbolatova, Prince Vathappallil, Mihir Matalia, Milan Toma. Machine learning for medical image classification. Academia Medicine 2024;1. Doi: 10.20935/AcadMed7444.
43. Molly Bekbolatova, Jonathan Mayer, Chi Wei Ong, Milan Toma. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare. 2024;12(2):125; Doi: 10.3390/healthcare12020125.

Concerns/Feedback	Summary of Revisions
"The title might be more detailed"	We appreciate your suggestion to make the title more detailed. Our original title, "Machine Learning for Cardiovascular Diagnostics," was chosen to encompass the broad application of machine learning in this field. We intended it to be concise yet indicative of the comprehensive review we've provided on how machine learning algorithms can enhance cardiovascular diagnostics through various imaging modalities. As requested, we have revised the title to convey the content and objectives of our manuscript more precisely. The manuscript's title has now been changed to: *"Machine Learning Strategies for Improved Cardiovascular Disease Detection."*
"This statement needs to be modified 'The review demonstrates' because review article are not designed or intended to demonstrate any fact but to compile and summarize current knowledge and state of the art in a given field"	We agree with the concern involving the term "demonstrates" and the implications of that in our review article. We have provided the following revision on page 1: *"This review discusses and complies the strengths of different machine learning algorithms based on the imaging modality used and the specific cardiovascular disease diagnosed."*
"Inline quotes don't meet AMA style. This needs to be addressed"	We appreciate this reminder. *Our inline citations were revised to meet the AMA style.*
"Aim and scope are not clear at all"	We appreciate this feedback and understand the concern. To address this, the following revisions were made on page 2 to provide a clear aim and focus for our article: *Furthermore, this paper presents a systematic framework for selecting a ML algorithm for diagnosis of specific cardiac pathologies. The methodology takes a comprehensive approach by considering three key factors: the particular cardiovascular condition being studied, the chosen imaging modality, and the algorithm's performance metrics. By integrating these elements, the framework offers clear guidance for selecting and optimizing ML models best suited for specific cardiovascular diagnostic tasks.*
"The manuscript covers too many topics. Usually this type of paper is focused in one or two key topics to be developed as deep as possible"	Thank you for this feedback. We understand your concern about the breadth of topics covered in our manuscript. Our intention was to provide a comprehensive overview of the current state of machine learning applications in cardiovascular diagnostics across various imaging modalities. To address your feedback, we have revised the manuscript by *refining our focus to emphasize the selection and evaluation of machine learning algorithms specific to cardiovascular imaging. We streamlined the content to eliminate redundancy and enhanced the depth of discussion in key areas, ensuring that each section directly contributes to the main objectives of the review.*
"The discussion section might be enriched and extended a little more"	Thank you for the comment. The Discussion was further elaborated and expanded on with the following inclusions: *(Page 16)*

	Including participants from multiple sites, each sites having distinct clinical practices, patient demographics, and treatment approaches, can provide a more accurate reflection of actual healthcare environments. By distributing the study population across diverse locations, the risk of local factors at any single hospital or clinic skewing the results is minimized. This broader scope may explain why the MESA dataset had improved the external validity of findings by more closely capturing the variability inherent in real-world clinical settings, thereby enhancing reliability. Furthermore, involving multiple centers can accelerate patient recruitment and yield a more heterogeneous participant pool, which strengthens the generalizability and practical relevance of the study's outcomes. As such, multicenter research is increasingly recognized as a critical strategy for enhancing both the impact and reproducibility of healthcare studies—particularly for machine learning applications that must remain robust across varied and evolving patient populations. *(Page 17)* *Additionally, transparency, explainability, and reproducibility are important factors to consider before implementing a diagnostic ML algorithm. Transparency about model design decisions, such as feature selection and data preprocessing, ensures other researchers can replicate results and clinicians can more readily interpret model outputs. Explainability, on the other hand, clarifies how a model arrives at its predictions and is vital for fostering trust and accountability in high-stakes scenarios like patient care, where clinicians must quickly assess whether an algorithm's output aligns with established clinical knowledge.* *Reproducibility depends on proper validation strategies, including the use of separate training, validation, and testing datasets, as well as external validation through additional data sources. This approach lowers the risk of data leakage and guards against overfitting, where a model performs exceedingly well on training data but poorly on new data. Conversely, it also addresses underfitting by promoting iterative refinement of hyperparameters and model complexity until performance metrics stabilize or improve on both internal and external datasets.* *Beyond these technical considerations, implementing regular audits can help detect unintended biases that may arise from imbalanced training data or flawed feature representations. Stress-testing models under challenging conditions—such as simulated data shifts or varying patient demographics—further helps ensure performance consistency before deployment in clinical settings. These steps, coupled with ongoing monitoring of real-world outcomes, can instill greater confidence in ML-driven tools.*
"It's necessary to include a Conclusion section"	We appreciate the feedback and concern for a lack of a conclusion section. *This section was added at the end of our article (page 16) to address this.*
"Avoid as much as possible the use of abbreviations in the abstract" "Verify that all abbreviations are accompanied by the full term when first used in both, the title and main body" "No sentence or paragraph must begin with an abbreviation"	Thank you for this feedback. *Abbreviations were removed from the abstract to address this.* Abbreviations in titles and subtitles were expanded to include the full form: *Receiver Operating Characteristic Curves and Performance Metrics (page 12)* *Machine Learning Algorithms in Cardiovascular Diagnostics (page 15)*

"Avoid the use of abbreviations in titles and subtitles"	*We have ensured that no abbreviations are used at the beginning of a sentence or paragraph.*
"Figures and tables legends should not be longer than three or four lines"	We appreciate this feedback. The following changes were made to Figure 1 legend on page 4: *This mind map illustrates the diverse medical imaging modalities available for cardiac imaging. This paper focuses on the most commonly used imaging techniques, each represented by a distinct color. It also provides a comprehensive overview of these imaging tools, highlighting their specific applications in different cardiac pathologies.*
"Figures fonts must be big enough to be read"	Thank you for this feedback. We understand the importance of ensuring that all elements within our figures are easily readable to enhance comprehension. The figure in question is a schematic that summarizes the medical imaging techniques utilized for detecting various cardiovascular diseases. Due to the complexity and the need to include numerous imaging modalities and conditions, some of the smaller bubbles contain text with reduced font sizes to fit all pertinent information within them. Considering that the manuscript will be accessed digitally, readers can zoom in on the figure to view all details clearly. Additionally, the accompanying text in the manuscript provides detailed explanations of the imaging modalities represented in the figure, which helps to clarify any information that might be less readable at first glance. We initially included more extensive descriptions within the manuscript, but due to the journal's strict word count limitations, we condensed the text to comply with the guidelines. This made the figure even more important, as it efficiently conveys comprehensive information that would otherwise exceed the word limit.