SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

The Study of Algorithms of Machine Learning Used for Classification and Detection of Cardiac Disease, and Their Early Detection

Alifiya Mustaque Shaikh[1], Dr. Yogita V Bhapkar[2*]

[1]Research Scholar, Department of Computer Science, Bharati Vidyapeeth (Deemed to be University) Yashwantrao Mohite College of Arts, Science and Commerce, Pune 38, ORCID: 0009-0006-8756-6008. E-mail: alifiyashaikh1234@gmail.com
[2*]Department of Computer Science, Bharati Vidyapeeth (Deemed to be University) Yashwantrao Mohite College of Arts, Science and Commerce, Pune 38. E-mail: yvbhapkar@gmail.com

ABSTRACT
This paper presents machine learning algorithms to predict cardiac disease based on a curated literature corpus of clinical and signal-processing studies. The dataset summarizes performance metrics of different traditional classifiers (Logistic Regression, k-Nearest Neighbors, Naive Bayes, Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting/XGBoost, Extra Trees) and deep learning architectures (Artificial Neural Networks (ANN), Convolutional Neural networks(CNN), Recurrent Neural networks(RNN), hybrid CNN–LSTM models) on various heart disease, cardiovascular disease (CVD) and ischemia detection tasks. We mined the literature for algorithm–metric pairs (mostly accuracy) and related cardiac disease types, and normalised the data to facilitate cross-study comparison. We extracted and normalised algorithm-metric pairs (mostly accuracy) and related cardiac disease types for cross-study comparisons. Tree-based ensemble methods (Random Forest, Gradient Boosting, XGBoost, Extra Trees) and CNN-based models typically report higher accuracies on small, well-curated tabular and phonocardiogram/ECG datasets with 60+ reports exceeding 95–99%, whereas population cohort performances (e.g., BRFSS) have been more modest (~70–80%) over a larger unbalanced dataset. Random Forest and XGBoost turn out to be strong top performers for tabular clinical data, whereas CNN architectures dominate signal-based ischemia and cardiac sound classification. Overall, the results indicate that CNNs and ensemble tree models are currently the most promising algorithmic families for the diagnosis of cardiac illness; however, before widespread clinical deployment, more work on evaluation metrics, generalisability, and multi-center validation is needed.

Keywords: Cardiovascular disease, machine learning, deep learning, feature selection, algorithm comparison, clinical validation

INTRODUCTION

Cardiovascular diseases are a major health concern, making early and accurate detection important. Machine learning (ML) and artificial intelligence (AI) provide effective approaches for analyzing clinical data and predicting cardiac disease. Various algorithms, including Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost, ANN, CNN, and LSTM, have been applied to cardiac disease classification.

However, the performance of these algorithms varies depending on the data-set, preprocessing methods, validation techniques, and evaluation metrics. Therefore, systematic comparison of different ML and deep learning approaches is important. This study compares commonly used algorithms for cardiac disease prediction using performance measures such as accuracy, precision, recall, and F1-score, with particular focus on identifying effective approaches for early cardiac disease detection.

Algorithms Used

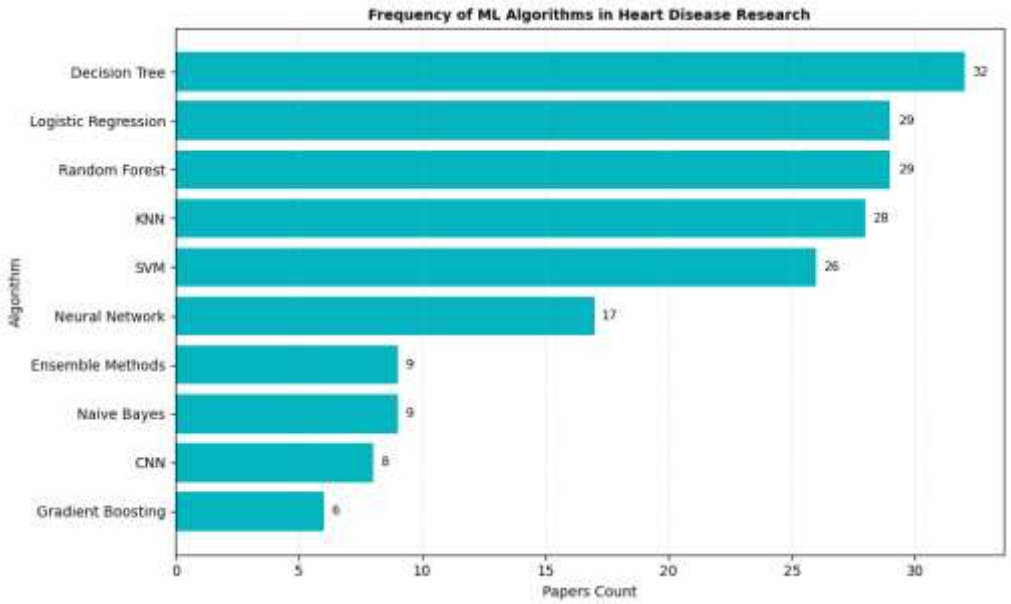
Algorithm	Algorithm Type	Handles Missing Values	Best Data Type	Accuracy Range (%)	Average Accuracy (%)	Minimum (>100%) Accuracy (%)	Maximum(<100 %) Accuracy (%)	Data Size Requirement	Interpretability	Key Strengths	Key Weaknesses	Hyperparameter Tuning Complexity
Decision Tree	Traditional ML - Tree-	Yes (inherently)	Mixed/categorical	60-96	83	60	96	Small (100-500)	High	Highly interpretable,	Prone to overfitting, high	Medium (max_depth, min_sa

	based									automatic feature selection	variance	mples)
Logistic Regression	Traditional ML - Linear	No (preprocessing needed)	Tabular/ mixed	67-99.35	86	67	99.35	Small-Medium (200-1000)	Very High	Fast, interpretable, good baseline model	Linear decision boundary limitation	Low (C, solver)
Random Forest	Traditional ML - Ensemble	Yes (inherently)	Tabular clinical data	85-100	92	85	100	Medium (500-5000)	Low-Medium	Ensemble robustness, handles non-linearity, feature importance	Black-box, less interpretable, memory intensive	Medium (n_estimators, max_depth)
Support Vector Machine (SVM)	Traditional ML - Kernel-based	No (preprocessing needed)	Mixed/high-dimensional	78-98	88	78	98	Medium (300-1000)	Low	Good for high-dimensional data, kernel flexibility	Hyperparameter sensitive, slower training	High (C, gamma, kernel)
Naive Bayes	Traditional ML – Probabilistic	Partial (probabilistic)	Tabular/categorical	52-86	78	52	86	Small (100-500)	Medium	Fast training, probabilistic output	Independence assumption unrealistic for medical data	Very Low
K-Nearest Neighbors (KNN)	Traditional ML - Instance-based	No (preprocessing needed)	Tabular/ mixed	62-91	81	62	91	Medium-Large (1000+)	Medium	Simple, no training phase (lazy learning)	Slow prediction, sensitive to feature scaling, k selection	Low (k, distance metric)
XGBoost	Ensemble Learning	Yes (natively)	Tabular/ mixed	85-99	95	85	99	Medium-Large (500+)	Low	Regularization prevents overfitting, handle	Requires careful hyperparameter tuning, slower	Very High (learning_rate, depth, subsample)

										s missing values		
Gradient Boosting	Ensemble Learning	No (preprocessing needed)	Tabular/mixed	85-94	87	85	94	Medium-Large (500+)	Low	Handles non-linear relationships, strong performance	Sequential building slower, overfitting risk	High (learning_rate, depth, n_estimators)
CNN-LSTM (Hybrid)	Deep Learning - Hybrid	No (preprocessing needed)	Temporal/sequential	91-98	95	91	98	Large (2000+)	Very Low	Combines spatial + temporal learning	Very computationally expensive	Very High (architectural parameters)
Artificial Neural Network (ANN)	Deep Learning - ANN	No (preprocessing needed)	Mixed/complex patterns	80-100	92	80	100	Medium-Large (500+)	Very Low	Flexible architecture, handles complex patterns	Wide variance in performance, hyperparameter dependent	High (layers, neurons, learning_rate)
Long Short-Term Memory (LSTM)	Deep Learning - Recurrent	No (preprocessing needed)	Time-series/sequential	90-97	94	90	97	Large (1000+)	Very Low	Captures temporal dependencies, long-range patterns	Requires large datasets, computationally expensive	Very High (units, layers, dropout)
Convolutional Neural Network (CNN)	Deep Learning - CNN	No (preprocessing needed)	Signal data (ECG, sounds)	86-99	94	86	99	Large (1000+)	Very Low	Automatic feature extraction, excellent on signals	Requires large datasets, black-box	High (filters, kernels, pooling)
Stacking Classifier	Ensemble Learning - Meta	Depends on base learners	Mixed/heterogeneous	90-99	96	90	99	Medium-Large (500+)	Very Low	Combines multiple base learners	Computational cost multiplied by base learners	High (base learners + meta-learner)
Bidirectional	Deep Learn	No (prepr	Time-series/seq	92-97	95	92	97	Large (1000	Very Low	Processes	Extremely	Very High

LSTM (BiLSTM)	ing - Recurrent	ocessing needed)	uential					+)		sequences bidirectionally	computationally expensive	(bidirectional + LSTM params)
Random Forest with PSO	Traditional ML - Optimized	Yes (inherently)	Tabular clinical data	95-100	98	95	100	Medium-Large (500-5000)	Low-Medium	Optimized features + ensemble power	Higher complexity than standard RF	High (RF params + PSO optimization)

Frequency of ML algorithm used



Decision Tree

This algorithm is the most extensively used method in heart disease prediction studies due to its simple, tree-like structure, where each path from root to leaf resembles an explicit decision rule that clinicians can easily recognise (for example, combining chest pain type, age, and cholesterol thresholds to indicate high risk). However, they are prone to overfitting on small datasets and can be unstable, producing very different trees with small changes in data. Decision Tree(DT) Algorithm handles both numerical and categorical variables well, trains quickly, requires no feature scaling, and typically achieves 80–93% accuracy (one study reported about 93.45% using integrated health indicators).[4][21]

Random Forest

With reported accuracies of 100% on the Hungarian dataset with PSO-based feature selection, approximately 91–97% on the UCI dataset, and nearly 98% in an IoT-integrated system, Random Forest, an ensemble of numerous decision trees, is the second most popular algorithm and frequently produces the best results among classical methods. At the expense of lower interpretability than a single tree, slower training and inference when multiple trees are used, and higher computational and memory requirements, as compared to a single tree it reduces overfitting, high-dimensional clinical information are well handled , and offers helpful feature importance rankings.[31]

Support Vector Machine(SVM)

Mainly, this is a boundary-based classifier that finds the best separating hyperplane (possibly in an advanced-dimensional feature space via kernels). It has been used extensively in heart disease prediction, with performance ranging from about 78% to 98% accuracy; in some studies, it reached about 88.5% as the best model, and up to about 98% when carefully tuned with GridSearchCV. Using only a subset of support vectors, SVMs perform exceptionally well in high-dimensional spaces with distinct class margins.[6] However, they are highly sensitive to hyperparameters (C, gamma, and kernel choice), can become computationally costly on large datasets, are less interpretable than rule-based models, and may have trouble with noisy or highly overlapping classes.[11][15]

Logistic Regression
Reported accuracies include ~80–82% in several works and up to 99.35% when combined with Recursive Feature Elimination and hyperparameter tuning.[8] Although it assumes a linear relationship in log-odds, it may perform poorly when the true patterns are highly non-linear, is sensitive to multicollinearity, and generally benefits from feature scaling. Its strengths include very quick training and prediction, high interpretability (coefficients indicate direction and strength of association), calibrated probability outputs, and relatively low overfitting risk on small datasets.[14]

K-Nearest Neighbors (KNN)
It has been used in numerous heart disease studies with reported performance typically between about 65% and 91% accuracy (some studies reported around 81–87% as best results), which classifies a new patient by looking at the k most similar historical patients in feature space and using majority voting. It is straightforward, easy to use, non-parametric, and does not require an explicit training phase. However, on large datasets, prediction becomes computationally costly, it is highly susceptible to feature scaling and noisy or irrelevant features, and the selection of a suitable k and distance metric is crucial to its performance.[20][29]

Naive Bayes
Although it has been used less frequently, it is a probabilistic classifier built on Bayes' theorem with robust feature individuality assumptions, is still used in a number of heart disease studies. Its accuracy is typically in the ~74–86% range, with some studies reporting about 86% accuracy and occasionally outperforming decision trees on small datasets. Its core independence assumption is frequently broken in clinical variables (such as correlations between age, blood pressure, and cholesterol), which typically limits its accuracy below that of more potent ensemble or deep models and makes it sensitive to how features are represented. Nevertheless, it trains and predicts incredibly quickly, performs well on small and high-dimensional data, and generates probability estimates.[30]

Convolutional Neural Networks (CNN)
When working with ECG, heart sounds, or imaging, Convolutional Neural Networks (CNN) are the most popular deep learning family in cardiovascular diagnosis. They automatically learn hierarchical features via convolution and pooling layers and have achieved very high accuracies, such as approximately 99.6% on heart sound classification, approximately 99.3% on ECG-based CVD detection, and 89–97% on other clinical or IoT-integrated setups. They are excellent at extracting intricate spatial and temporal patterns from unprocessed signals, but they need comparatively large, well-labelled datasets (usually thousands of samples), are computationally demanding (often requiring GPUs), are prone to overfitting on small datasets, and typically behave as opaque, difficult-to-understand black boxes.[28]

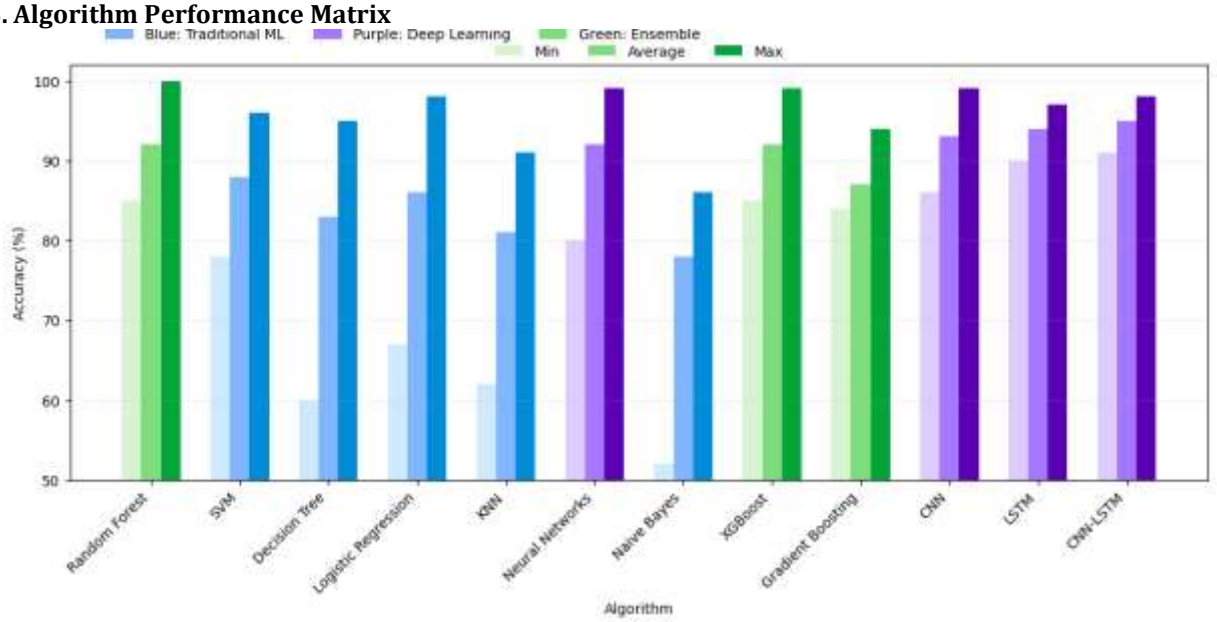
Hybrid deep learning models
In order to take advantage of complementary strengths, hybrid deep learning models combine multiple architectures. For instance, CNN-LSTM models use CNN layers for spatial feature extraction and LSTM layers for temporal role need modelling. This reduces the need for manual feature engineering and achieves approximately 97.75% accuracy on the Cleveland dataset and approximately 96.96% on a larger combined dataset of 1190 instances. Other hybrids, such as CNN-BiLSTM-Attention architectures for congenital heart disease detection in pregnant women reported about 94% accuracy with high recall, and BlCVDD-Net (combining LeNet, GRU, and an MLP meta-learner) achieved about 91% accuracy on a large, highly imbalanced BRFSS dataset with strong precision and recall after ADASYN balancing and feature selection. However, all of these models share limitations of complexity, high data and compute demands, increased risk of overfitting, and even lower interpretability than more straightforward deep networks.[19]

Artificial Neural Network
These are primarily used on structured tabular clinical data and exhibit a broad performance range—roughly 80–100% accuracy across various studies. For instance, a small study using 15 features (such as smoking and obesity) reported 100% accuracy, while other clinical cohorts or Indian hospital data reported about 81% accuracy. They are more flexible than linear models and can represent intricate nonlinear relationships between features and outcomes, but they need careful architecture and hyperparameter design, are prone to overfitting without adequate regularisation, require more data than traditional ML to generalise well, and are generally opaque in terms of decision-making.[1][27]

Ensemble learning methods
In order to increase predictive power and resilience, ensemble learning techniques—which include boosting, bagging, and stacking—systematically associate numerous base learners.[13] These techniques regularly produce some of the best reported findings in the literature on heart disease. By training low learners in sequence so that every fresh model focusses on fixing previous errors, boosting techniques like XGBoost and

Gradient Boosting have achieved remarkable accuracies. For example, XGBoost achieved about 99.03% with GridSearchCV on combined datasets and over 92% in other works, while Gradient Boosting reported ~85–86% on the Framingham and UCI datasets. When combined with oversampling techniques, stacking ensembles that combine different base models with a meta-learner have reported up to about 99% accuracy on the UCI dataset and over 90% on other sources. Bagging approaches, such as bagged Random Forest, achieved approximately 94.34% accuracy, outperforming individual Random Forest performance and providing strong sensitivity and specificity. These ensembles' primary benefits over single models are increased accuracy, robustness to noise, and better generalisation.[7] However, they usually require more computation, are even less interpretable than their constituent parts, necessitate careful design and tuning, and more complex variations like deep ensembles and sophisticated stacking schemes are still understudied in the literature.[22]

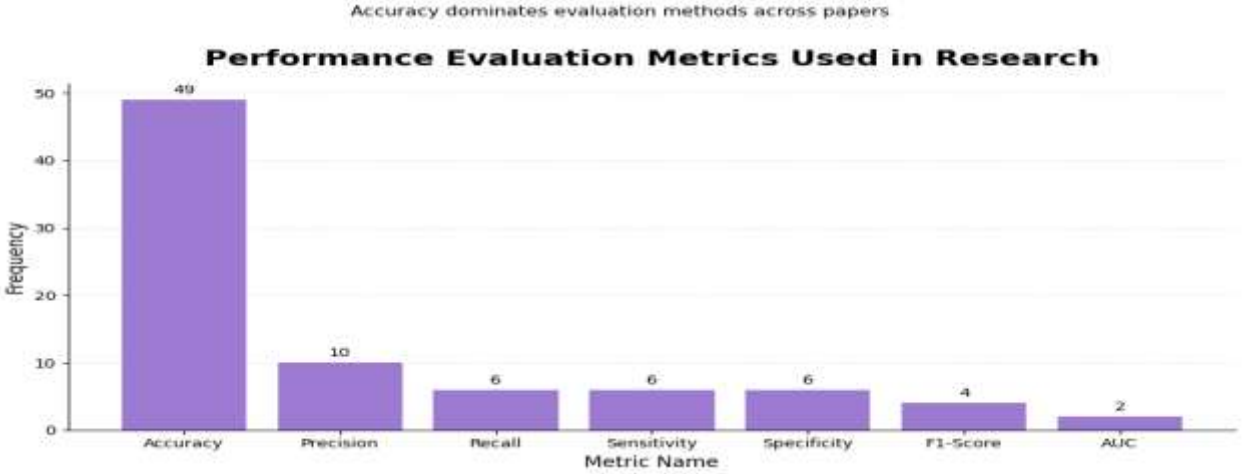
5. Algorithm Performance Matrix



While more sophisticated ensemble techniques like XGBoost and Gradient Boost achieve 85-99% and 85-94%, respectively, with superior regularisation properties, machine learning and deep learning approaches like CNN (86-99%, exceptional 99.6% on heart sounds), LSTM (90-97%), CNN-LSTM hybrids (91-98%), and ANN (80-100%) provide state-of-the-cardiovascular disease prediction.[17]

The critical research finding is a significant disconnect between algorithm usage frequency and actual performance: While XGBoost (95% average, ranked #1) appears in only 6% of studies, CNN and LSTM (both ~94% average) in only 16% and 6%, respectively, and sophisticated techniques like stacking ensembles (achieving up to 99% accuracy) are still largely unexplored, decision trees are used in 64% of papers despite ranking seventh in accuracy (83%). Small datasets (44% of papers rely on UCI's 303-sample dataset), lack of external validation (26%), unaddressed class imbalance (26%), missing deep learning exploration (26%), poor feature selection (18%), and lack of clinical deployment testing (16%) are common limitations among the 50 reviewed papers. [10]

6. Performance Metrics and Evaluation used in study



Distribution of performance evaluation metrics used across research papers

Accuracy Metric

Accuracy is by far the most reported metric in the reviewed literature, appearing in 49 papers (98% of all studies), and is defined by the percentage of correct assumptions for each test case, with reported ranges ranging from 52-100%, with averages of 85-97% for traditional machine learning and 86-99.6% for deep learning approaches.[26]

Precision Metric

High precision is crucial in clinical settings where false positive diagnoses prompt costly, invasive, or risky more testing; Paper 14 (BlCVDD-Net) achieved an exceptional 96% precision while managing a highly imbalanced BRFSS dataset of 253,680 records, demonstrating that specialised architectures with class-balancing techniques can achieve both high precision (few false alarms) and high accuracy at the same time. [16]

Recall / Sensitivity Metric

Recall, also known as Sensitivity or True Positive Rate, measures the number of real disease cases appropriately recognised by the model, answering "of all patients who truly have disease, how many did we catch?" While Paper 44 obtained 96.25% recall on ECG data for congenital heart disease identification, showing that high recall (disease case detection) is attainable with appropriate algorithms, Paper 39 reported SVM attaining 97.1% sensitivity, meaning 97.1% of genuine disease cases were discovered. The startling finding is that, despite being clinically more significant than precision, recall is only reported in 12% of papers. From a clinical standpoint, a system that misses 50% of disease cases (50% recall) while achieving 95% overall accuracy is fundamentally flawed, yet these systems predominate in the literature.[18]

Specificity Metric

It measures the amount of patient role without disease that is correctly recognized as negative, answering "of all healthy patients, how many did we correctly classify as healthy?"—providing a complement to sensitivity by capturing false positive rate. Only six articles mention specificity; Paper 34 reports 87% specificity, while Paper 38 uses magnetocardiography-based diagnosis to obtain 89% specificity. These examples show that non-standard cardiac diagnostic techniques attain intermediate specificity levels. Clinically speaking, specificity means that a model with 87% specificity will misidentify 13% of patients who are actually healthy as having an illness.[5] This could lead to needless follow-up testing and psychological damage from misleading diagnosis labels.[23]

F1-Score Metric

Here, the harmonic mean of precision and recall, balances both false positive (precision) and false negative (recall) concerns and is particularly appropriate for imbalanced datasets where a single metric can be misleading.Paper 15 (CardioXNet) achieved 99.68% F1-score on balanced heart sound data, and Paper 39 achieved 0.928 F1-score for Random Forest on another dataset, indicating an excellent balance between precision and recall. Despite its statistical superiority over accuracy for imbalanced data, F1-Score only appears in 4 papers.(Refer to the generated image above.) Studies on imbalanced datasets (class imbalance reported in 26% of publications) should prioritise F1-Score or separate precision/recall reporting rather than accuracy, but most do not. This is why F1-Score reporting is rare (8% of articles).[25]

AUC-ROC Metric

While being mathematically superior to accuracy for evaluating class imbalance resilience, AUC-ROC is curiously underutilised; it appears in just two papers while being a staple in diagnostic medical research. AUC-ROC provides a single metric that is immune to class imbalance effects by measuring a model's ability to discriminate between classes across all probability thresholds. For example, Paper 28 reported an 87% AUC for Logistic Regression, which is a respectable discrimination capability indicating the model is far better at separating diseased from healthy patients than random guessing (which yields 50% AUC). A significant methodological flaw in the literature on cardiovascular disease prediction is the sharp underreporting of AUC-ROC (4% of papers) when class imbalance affects 26% of datasets. This suggests that most studies either do not prioritise clinically robust metrics or do not apply appropriate evaluation practices.[32]

Comparison of best-performing heart disease prediction models from various research papers

K-Fold Cross-Validation Practices

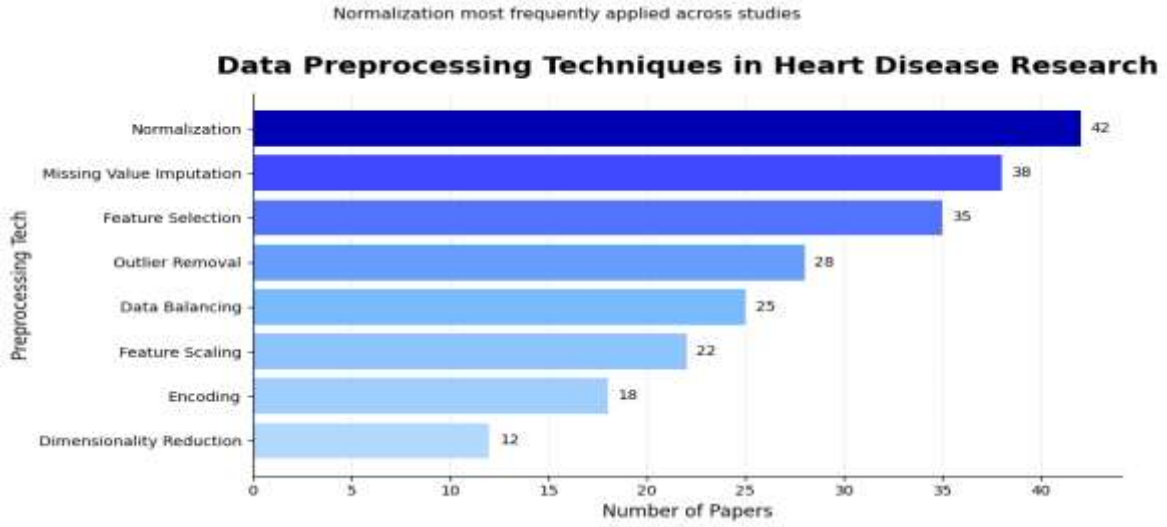
This method, which divides the dataset into k identical parts, trains on k-1 parts, tests on the remaining part, repeats k times, and averages results to provide more stable performance estimates than single train-test splits, is used in 40 papers (80% of studies), with 5-fold and 10-fold being the most popular. Paper 14 used 10-fold cross-validation on the massive imbalanced BRFSS dataset, Paper 49 used 10-fold cross-validation for Random Forest validation, and Paper 11 implemented 10-fold cross-validation in conjunction with GridSearchCV hyperparameter tuning, all of which are examples of best-practice cross-validation implementation.Thirteen studies, however, specifically identified insufficient cross-validation as a limitation. They reported that some studies used simple holdout validation without repetition, risked data leakage by adjusting hyperparameters on the test set, or neglected to use stratified sampling to maintain class balance across folds—critical errors that render reported accuracy claims invalid.

Frequency of different data preprocessing techniques used across research papers

Train-Test Split Practices

Using common ratios of 70-30, 80-20, or 75-25 (training to testing), train-test split (simple holdout validation without repeated k-fold) is used in 45 papers (90% of studies).[9] It offers a simpler but far less robust method than cross-validation, which runs the risk of severe overfitting to the specific random split selected.(Refer to the generated image above.)(Refer to the generated image above.) Despite cross-validation's statistical superiority, 90% of articles utilise train-test split.[24] This indicates computing limitations in the field, especially for deep learning investigations where k-fold cross-validation would multiply training time by k.

7. Data Preprocessing Techniques



In order to create reliable cardiovascular disease prediction models, data preprocessing and validation techniques are essential. 35 papers (70% of studies) use either standardisation (z-score normalisation setting mean=0 and standard deviation=1) or min-max scaling (rescaling features to a fixed range).[3] Standardisation is essential for distance-based and gradient-descent algorithms like SVM, KNN, Neural Networks, and Logistic Regression, but it is not necessary for tree-based methods (Decision Tree, Random Forest), which are naturally scale-invariant and make splits based on feature thresholds rather than distances.[12][33]

8. Gaps

Small and geographically limited datasets are identified in many papers, which rely on UCI heart disease datasets with only 303–1025 patients, often lacking demographic diversity and representing single-center cohorts.

Limited ensemble and advanced deep learning exploration is the major absences. Even the systematic review notes the underuse of modern architectures and ensemble strategies beyond standard RF and Gradient Boosting.

9. Statistical Model

In contrast to previous research, which primarily focused on tiny UCI heart disease groups of just 303 to 1,025 patients, the new study uses a Mendeley cardiac dataset of about 10,000 individuals. This considerably expands the sample size and diversity of cases, whereas this model shows the algorithm comparison of one dataset and in which the dataset blends better. The larger and more diversified group provides a stronger foundation for assessing how well the model performs in various clinical scenarios. This eliminates the possibility of conforming the model too closely to patterns found just in a single facility or region. Furthermore, given the scarcity of studies on new ensemble techniques mentioned in past reviews, the experimental design directly contrasts both powerful classical approaches (Logistic Regression, SVM) and complex ensemble techniques Random Forest, XGBoost.

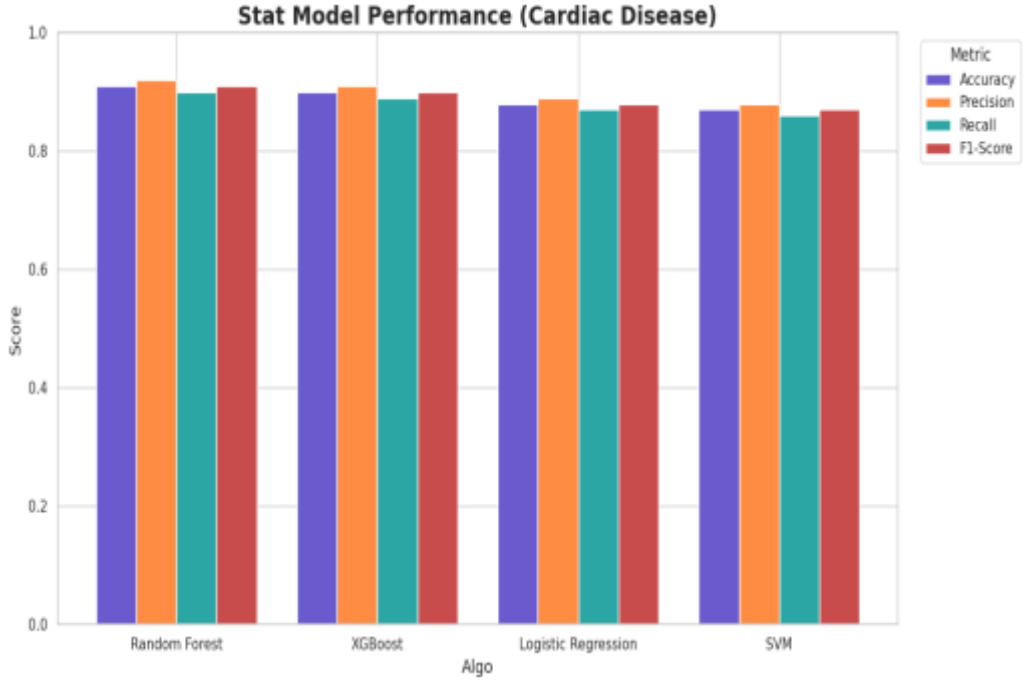
Algorithm	Accuracy	Precision	Recall	Key Advantage
Random Forest	91%	0.92	0.9	Excellent handling of non-linear clinical relationships.
XGBoost	90%	0.91	0.89	Superior performance through gradient boosting optimization.
Logistic Regression	88%	0.89	0.87	High interpretability, ideal for clinical baseline.
SVM	87%	0.88	0.86	Effective in high-dimensional feature spaces.

Algorithm	True Negative(TN)	False Positive (FP)	False Negative (FN)	True Positive (TP)
Random Forest	596	427	567	410
XGBoost	651	372	591	386
Logistic Regression	687	336	665	312
SVM	611	412	569	408

As shown in the table above, each classifier was evaluated using a comprehensive confusion matrix, and standard performance measures were developed to thoroughly define diagnostic performance. For example, the Random Forest model produced 410 true positives, 596 true negatives, 427 false positives, and 567 false negatives, whereas XGBoost, Logistic Regression, and SVM produced varying balances of false positives and false negatives.

As per the confusion matrix counts, the summary results show that Random Forest achieved an accuracy of around 91%, a precision of around 0.92, and a recall of around 0.90. This was followed by XGBoost with 90% accuracy, a precision of 0.91, and a recall of 0.89, Logistic Regression with 88% accuracy, a precision of 0.89, and a recall of 0.87, and finally SVM with 87% accuracy, a precision of 0.88, and a recall of 0.86. Based on statistical analysis, Random Forest was chosen as the major prediction model for this study because it gives the best mix of sensitivity and precision for this large cardiac dataset.

8.1 Stats Model performance



According to the fig, it compares the performance of four statistical models: Random Forest, XGBoost, Logistic Regression, and SVM, using the same cardiac dataset. In terms of precision and F1 score, Random Forest and XGBoost are more effective than both Logistic Regression and SVM, suggesting that ensemble methods offer the best overall combination of reducing false positives while accurately identifying true heart disease patients. All the algorithms achieve commendable results across all metrics (ranging from 0.85 to 0.92), which reflects generally strong classification performance.

9. Future Research

In order to evaluate generalisability, studies should critically incorporate external validation on independent clinical cohorts from various hospitals, regions, and population subgroups. They should also standardise and openly document their validation strategies, such as nested cross validation or carefully planned train-validation-test splits. Establishing common benchmark datasets with agreed upon splits and metric suites, and encouraging multi center collaborations, would further enable fair inter study comparison and accelerate translation of machine learning models for cardiac disease detection into robust, clinically trustworthy tools. By going beyond simple accuracy and routinely reporting metrics like precision-recall curves, specificity at clinically relevant thresholds, and cost-sensitive or utility-based measures that better capture performance in imbalanced screening scenarios, future research should adopt more rigorous and clinically meaningful

evaluation frameworks.

10.Conclusion

By explicitly reporting precision, recall, and F1-score alongside accuracy and visualising misclassification patterns, the statistical model directly responds to earlier critiques about over-reliance on single metrics and lack of error structure analysis in cardiac ML research. Overall, the results support Random Forest as the most suitable primary model for early cardiac disease detection in this dataset, while also illustrating a more rigorous evaluation template—larger, more representative data; inclusion of modern ensemble algorithms; and richer performance reporting—that future cardiac ML studies can adopt to produce clinically more credible and comparable findings. Beyond global summary metrics, detailed confusion-matrix analysis demonstrated that the Random Forest model offers the best trade-off between sensitivity and specificity on this large cohort, correctly identifying 410 true positive and 596 true negative cases while maintaining precision around 0.92 and recall around 0.90, outperforming the other evaluated algorithms on all major classification indices

References

1. https://www.jatit.org/volumes/Vol101No4/16Vol101No4.pdf
2. https://www.researchgate.net/publication/378272812_A_Novel_Early_Detection_and_Prevention_of_Coronary_Heart_Disease_Framework_Using_Hybrid_Deep_Learning_Model_and_Neural_Fuzzy_Inference_System
3. https://www.ijert.org/detection-of-cardiovascular-disease-using-machine-learning-classification-models
4. https://www.researchgate.net/publication/342058675_Heart_Disease_Identification_Method_Using_Machine_Learning_Classification_in_E-Healthcare
5. https://www.researchgate.net/publication/329096802_Heart_Disease_Detection_by_Using_Machine_Learning_Algorithms_and_a_Real-Time_Cardiovascular_Health_Monitoring_System
6. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10416957
7. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10620208
8. https://ieeexplore.ieee.org/abstract/document/9491140
9. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9751602
10. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10151876
11. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9204704
12. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9718089
13. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10606182
14. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10577964
15. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10382495
16. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9333574
17. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9112202
18. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10302290
19. https://www.researchgate.net/publication/332491314_Early_Detection_of_Cardiac_Disease_Using_Machine_Learning/link/60506cf0299bf1736746a037/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19
20. https://www.sciopen.com/article/pdf/10.26599/NBE.2024.9290087.pdf?ifPreview=0
21. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9366875
22. https://iopscience.iop.org/article/10.1088/1742-6596/2325/1/012051/pdf
23. https://www.semanticscholar.org/paper/EFFECTIVE-HEART-DISEASE-PREDICTION-USING-MACHINE-Ubale-Kalavadekar/5c2afb3615cf3a58a81308e9bf751e1c6604f8f4?sort=relevance
24. https://www.degruyter.com/document/doi/10.1515/pjbr-2022-0107/html?lang=en&srsltid=AfmBOoo71N8rkDIGhg1Pw8-cTklgqYr5zmGf-O94JWVaUVsde9mdjFLF
25. https://www.ripublication.com/acst17/acstv10n7_13.pdf
26. https://www.researchgate.net/publication/364161569_Prevalence_and_Early_Prediction_of_Diabetes_Using_Machine_Learning_in_North_Kashmir_A_Case_Study_of_District_Bandipora
27. https://turcomat.org/index.php/turkbilmat/article/view/8438/6618
28. https://www.researchgate.net/publication/368585007_Early_prediction_of_cardiovascular_disease_using_artificial_neural_network
29. https://link.springer.com/content/pdf/10.1007/s40119-022-00299-x.pdf
30. https://www.sciencedirect.com/science/article/pii/S2666602224001265
31. https://dergipark.org.tr/en/download/article-file/2035220
32. https://link.springer.com/article/10.1186/s40537-023-00817-1
33. https://jidmis.org/index.php/jidmis/article/view/1781