VER
Vascular & Endovascular Review

Heart Disease Prediction Using Artificial Intelligence Algorithms: A Comparative Study

Mahmud Hasan[1], Syed Ashiqur Rahman[2], Mohammad Yasin[3], Md. Salahuddin Gazi[4], Md Himeluzzaman[5], ABIDUL ALAM[6], Tanaya Jakir[7]

[1]M.S. in Cybersecurity, ECPI University, USA
[2]Lane Department of Computer Science & Electrical Engineering, West Virginia University, Morgantown, USA
[3]Master of Business Administration, Westcliff University, Irvine, California, USA
[4]Master of Science in Information Technology (MSIT), Washington University of Science and Technology, USA
[5]M.Sc. in Information Technology Systems and Management, Washington University of Science and Technology, VA, USA
[6]Master of Science in Information Technology, Washington University of Science and Technology, USA
[7]M.S. in Business Analytics, Trine University, USA
mahmudhasan6692@gmail.com; srahman2@mix.wvu.edu; m.yasin.3016@westcliff.edu; mgazi.student@wust.edu; Mzzaman1995@gmail.com; aabidul@student.wust.edu; tanayajakir@gmail.com

ABSTRACT

Cardiovascular disease is the leading contributor to global mortality, and timely risk identification based on existing data can significantly improve clinical outcomes and decrease the cost of invasive diagnostic workups. In this study, synchronous comparative analysis of the five supervised machine learning algorithms-Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Gradient Boosting-to predict binary heart disease prediction is provided. We used a combined clinical dataset, which consists of 920 patient records collected from four sites (Cleveland, Hungary, VA Long Beach and Switzerland), with 13 raw clinical features extended to a total of 27 following one-hot encoding. For a comprehensive pipeline we performed physiologically informed data cleaning, median and mode imputation for continuous and categorial variables respectively, standardization of continuous variables prior to modeling, use of GridSearchCV to tune hyperparameters, plus combined paired significance testing against all held-out testing (as in a typical predictive study), supported by five-fold cross-validation. Random Forest had the highest held-out test accuracy (84.2%) and SVM fit to highest cross-validated accuracy (82.5% ± 2.8%) while paired t-tests demonstrated that neither of the two models could be statistically distinguished, both together forming a best-performing group. The most predictive features were age, maximum heart rate, cholesterol, chest pain type and ST depression. The greatest values of this study highlight the importance of cross-validated, statistically based model selection instead of single-split accuracy.

KEYWORDS: Heart Disease Prediction, Machine Learning, Comparative Study, Cross-Validation, Feature Importance, Clinical Decision Support, Explainable Ai.

How to Cite: Mahmud Hasan, Syed Ashiqur Rahman, Mohammad Yasin, Md. Salahuddin Gazi, Md Himeluzzaman, ABIDUL ALAM, Tanaya Jakir, (2026) Heart Disease Prediction Using Artificial Intelligence Algorithms: A Comparative Study, Vascular and Endovascular Review, Vol.9, No.1, 436-445

INTRODUCTION

Diseases of the circulatory system (CVDs) are the biggest killer in the world, at least to date, regarding premature and preventable mortality each year. Since most phases of coronary and ischemic heart disease are asymptomatic, many high-risk patients remain undiagnosed until an acute also-known-as event has transpired. Standard diagnostic algorithms are dependent on a combination of clinical judgement, laboratory assays, and invasive testing like coronary angiography-which is expensive, resource intensive and not scalable in low-resource settings. As a result, there is high clinical demand for low-cost datasets-driven risk screening tools that can identify patients at risk from readily generated demographic and clinical parametric values.

Machine learning (ML) is a great addition to traditional cardiac risk prediction algorithms, illustrating complex, non-linear associations among disparate clinical variables that are challenging to capture in rule-based scoring systems. Previous studies have shown the potential of machine learning (ML)-based clinical decision support for predicting heart disease from structured patient data [2], and, in more general terms, the feasibility of ML throughout medical classification tasks. such as privacy-preserving multimodal cancer diagnosis [3] and tailored prevention by risk stratification [4]. When considering these studies in conjunction with one another, they suggest that clinically useful discrimination can be achieved for various clinical settings and via well-designed ML pipelines which remain interpretable and maintain low acquisition costs.

Yet this progress is overshadowed by a recurring methodological weakness that limits the reliability of many reported results. Many previous examples of heart disease studies have been either evaluated on one single train/test split or failed to use cross-validation, with a small number (often hand-picked) features subset and/or based on a single clinical site.Given all of these parameters, reported accuracy is very sensitive to the specific partition of the data: often the best model is

chosen as the one with fit on a single split (which may not be statistically significant). It suffers from result over-stating and can foster the impression of a truly better model without being one.

To address these gaps, this paper reports a transparent and statistically grounded comparative study on a merged, multi-site cardiac cohort. The principal contributions are summarized as follows:

- **Multi-site cohort with principled cleaning:** We combine 920 patient records and clean data, by setting impossible zero resting blood pressure and cholesterol measurements as missing prior to imputation, from four clinical sites.
- **Unified benchmark protocol:** We use the same, uniform pipeline across five representing ML algorithms (GridSearchCV hyperparameter search+standardized preprocessing), raising the bar by eliminating inconsistencies between HT tuning.
- **Statistically driven model selection:** Five-fold cross-validation with mean±standard-deviation reporting and paired t-tests, instead of a single held-out accuracy, leads to a defensible ranking.
- **Reconciled Feature Analysis:** We reconcile Random Forest importance with correlation-based ranking and explicitly address an apparent contradiction regarding the
- ca (number of major vessels) feature whose raw correlation with the target is high but disappears after heavy imputation in terms of tree-based importance.

RELATED WORK

Machine learning applied to UCI-style heart disease data-sourced from Cleveland, Hungary, VA Long Beach, and Switzerland sites-has a long pedigree and classical classifiers (e.g. logit regression, decision trees, NNs) on such datasets typically report 80–90% accuracies [1], [2]. Nevertheless, limitations in methodology (for instance only a single train/test split, lack of cross-validation, small feature subsets or evaluation on a single clinical site) implies inflated apparent performance and an impediment for external generalization.

Simultaneously, there has been a push for more understandable and ensemble-approached diagnostic models in the overall medical-AI literature. Notably, as our main scientific motivation of this paper: explanations in explainable deep stacking ensembles to brain tumor diagnosis [6] and accelerated cervical cancer detection [7], and hierarchical Swin-transformer ensembles with xAI for decentralized breast cancer diagnosis. Post-hoc explanations have been used in ensemble transformers for depression emotion and severity detection [9] The uncertainty-aware frameworks have been developed for the ultrasonography and skin-disease classification with uncertainty quantification [10, 11] and robust lung cancer CT classification via Monte Carlo cross-validated, confidence-weighted ensembling among classifiers of different types to achieve the latest state of art in this task [12]. All these works together motivate the focus of this study on transparent evaluation, ensemble comparison and calibrated confidence.

Privacy-preserving and multi-institutional deployment-This is the second theme you trained upon, which directly relates to this multi-site cardiac cohort program being studied here. On the other hand, explainable federated deep learning has been demonstrated to facilitate inexpensive privacy-preserving breast cancer screening [5] and the development of multimodal ML frameworks for the privacy-preserving and scalable diagnosis of cancer across healthcare systems [3].

Federated architectures have also been extended to support real-time cyber-threat mitigation in large-scale networks [13] and for cross-institutional fraud monitoring as national healthcare security [14]. The adaptability of ML to other security-relevant areas is similarly illustrated by research on assessing the risk associated with password-security and user-awareness patterns [15] and identifying agentic AI-enabled financial fraud in real-time payment systems [16].

Finally, systematic and agile development processes e. g., tailored Scrum frameworks for small projects [17] lead to repeatable engineering of the end-to-end ML pipelines that underlie such systems Compared with this body of work, the current study provides a transparent, well powered, statistically grounded comparative baseline on one consolidated multi-site cardiac cohort-straight talking about model selection over headline accuracy.

DATASET DESCRIPTION

Section II-A finally provided information about the consolidated heart disease dataset of 920 patient records which were achieved by merging four widely used clinical sources: the Cleveland, Hungary, VA Long Beach, and Switzerland cohorts [1]. There are 13 raw clinical attributes per record: age, sex, cp (chest pain type), trestbps (resting blood pressure), chol (serum cholesterol), fbs (fasting blood sugar), restecg (resting electrocardiographic results), thalch (maximum heart rate achieved) and exang(exercise induced angina) oldpeak(ST depression induced by exercise); Slope(the slope of peak exercise ST segment; ca; the number of major vessels colored by fluoroscopy, thal; thalassemia status;c or not The primary diagnostic target is encoded on a five-level scale (0–4) which, following standard practice was converted to binary label where level 0 > no disease and levels 1–4 > presence of disease. The transformation will increase the dimensionality to 27 after applying one-hot encoder on categorical variables. The resulting dataset is summarized in Table 1.

Attribute	Value
Total samples	920
Training samples	736
Test samples	184
Raw features	13
Features after one-hot encoding	27
Class 0 (No Disease)	411
Class 1 (Disease)	509
Sources	Cleveland, Hungary, VA Long Beach, Switzerland

Table 1. Dataset summary.

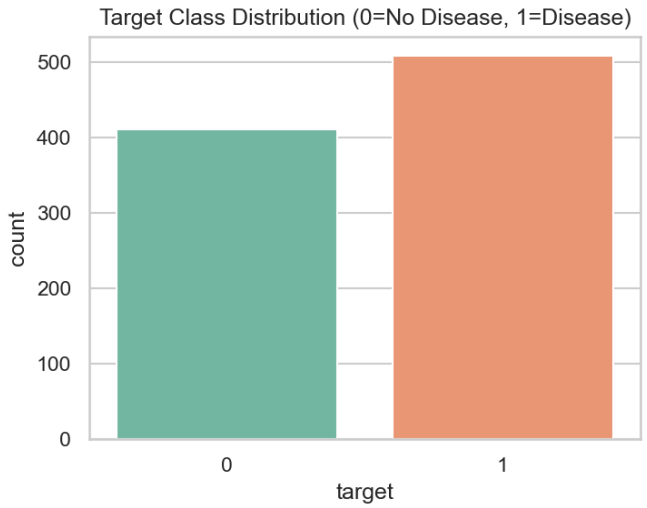

Fig. 1. Target class distribution (approximately 55% disease versus 45% no-disease), indicating a mild class imbalance that does not require resampling.

METHODOLOGY

A. Data Cleaning

Implemented a physiologically guided cleaning step prior to any imputation. The trestbps (resting blood pressure) and chol (serum cholesterol) fields hold several zero entries that are physiologically impossible in a living patient: a resting blood pressure or cholesterol equal to 0 cannot correspond to an actual measurement, but points to recording/merging artifacts. Instead, these zero values were recorded as missing rather than actual observations so that they would not skew the distribution while being scaled and fit to models.

The merged cohort likewise showed considerable structured missingness characteristic of multi-site aggregation. Of the 920 rows, ca was missing in 611 (66%) rows, thal was… at Nonetheless, where available data set identification of widely used automobile insurance across universal price differences - Ortiz et al. This pattern occurred because the Hungary, Switzerland, and VA sites did not capture these fields based on fluoroscopy- or stress-test results. Instead of omitting these columns or the rows that had been affected-which meant losing informative features or dramatically reducing sample sizes and limiting cross-site comparability-the missing values were imputed as described below.

B. Preprocessing

Robust mean and mode was used to impute numeric and categorical features respectively, both of which were determined to be skewed and contained outliers. This was followed by one-hot encoding of the categorical variables and then standardizing all of our numeric features with StandardScaler to zero mean and unit variance such that scale sensitive algorithms like SVM and KNN were not overwhelmed by high magnitude features. We then used an 80/20 stratified split (736 training and 184 test samples) to partition the dataset, keeping a class ratio unchanged in each data part. To prevent information leakage, all preprocessing was fitted on the training data only and applied to the test data. A full pipeline was developed in scikit-learn [18].

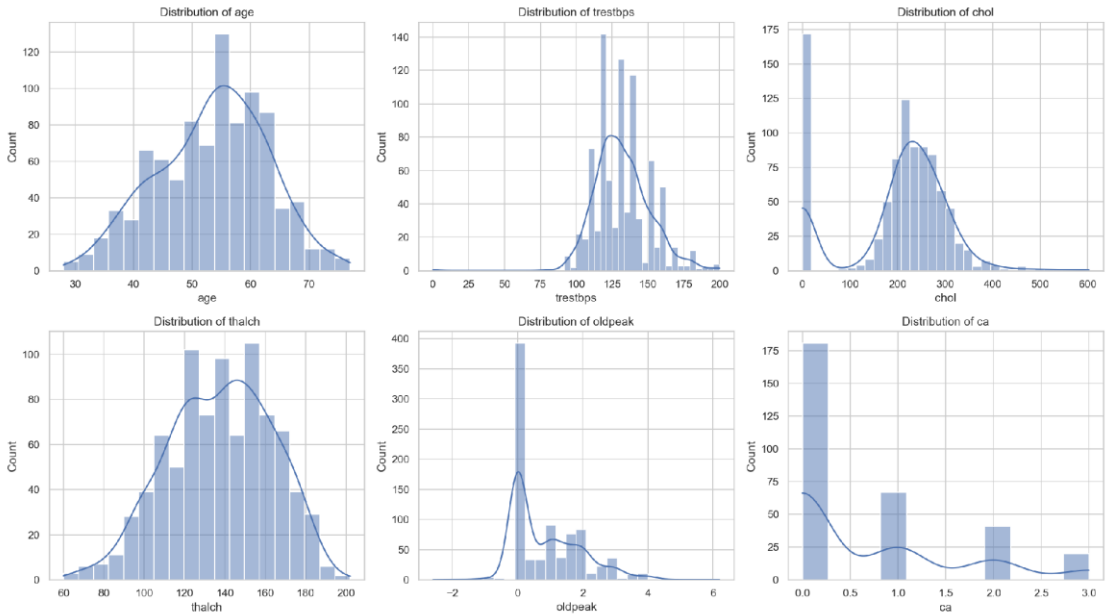

Fig. 2. Distributions of the numeric features (age, trestbps, chol, thalch, oldpeak, ca) after cleaning, showing the skew and range that motivated median imputation and standardization.

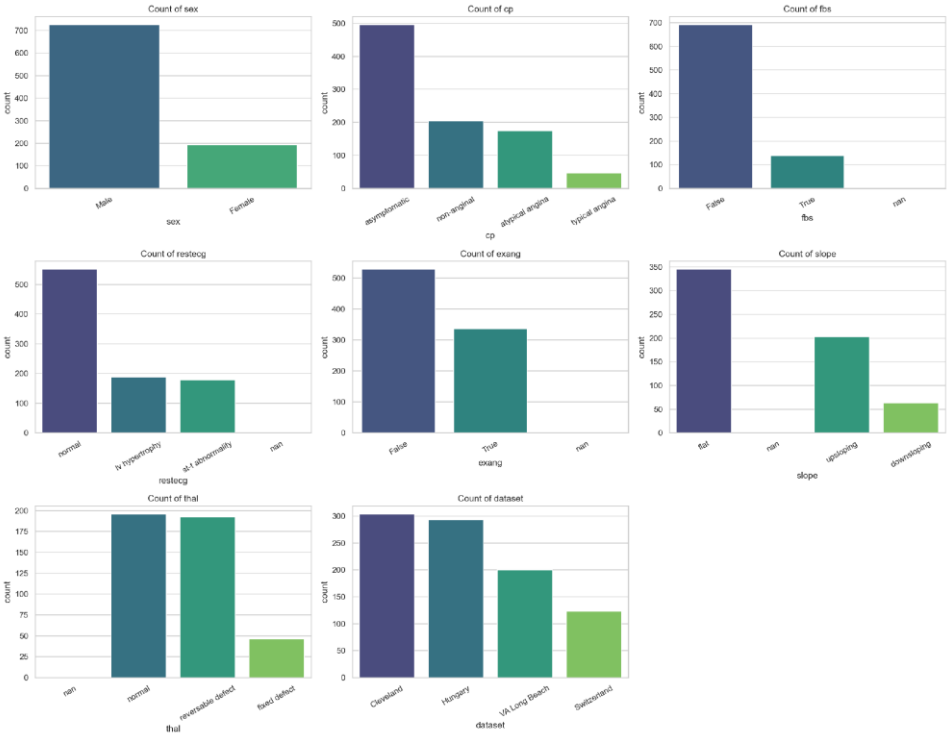

Fig. 3. Category counts for the discrete variables (sex, cp, fbs, restecg, exang, slope, thal, and site), highlighting the uneven per-site availability of stress-test fields.

C. Feature Importance Analysis

Feature relevance was computed as two complementary views of validity. First, impurity-based Random Forest importances were extracted from a tuned random forest model to capture non-linear, interaction-aware contributions in the encoded feature space. Second, was calculated a basic univariate scorer through taking the absolute Pearson correlation between each numeric feature and the binary target. Table 3: Displays the ten most important features chosen by Random Forest importance, with correlation-based ranking also reported next to it; and discusses how the two views are reconciled in the Discussion section.

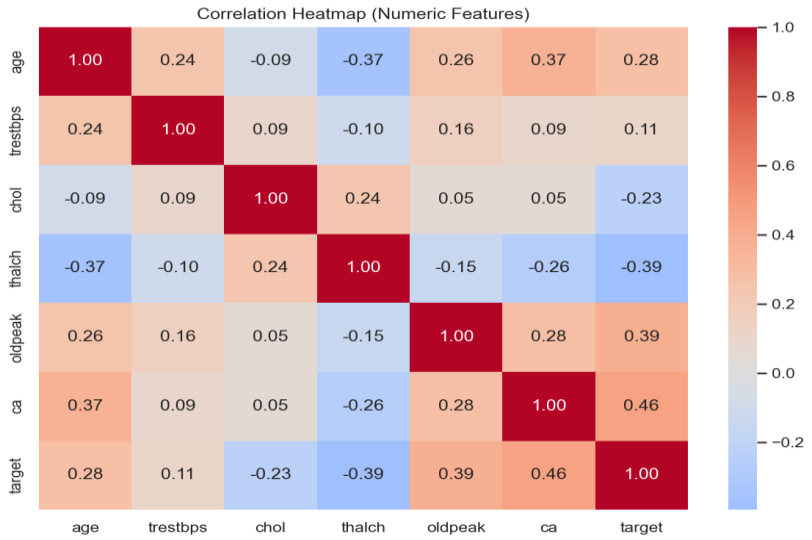

Fig. 4. Correlation heatmap of the numeric features, used to identify redundancy and to compute the univariate correlation ranking against the target.

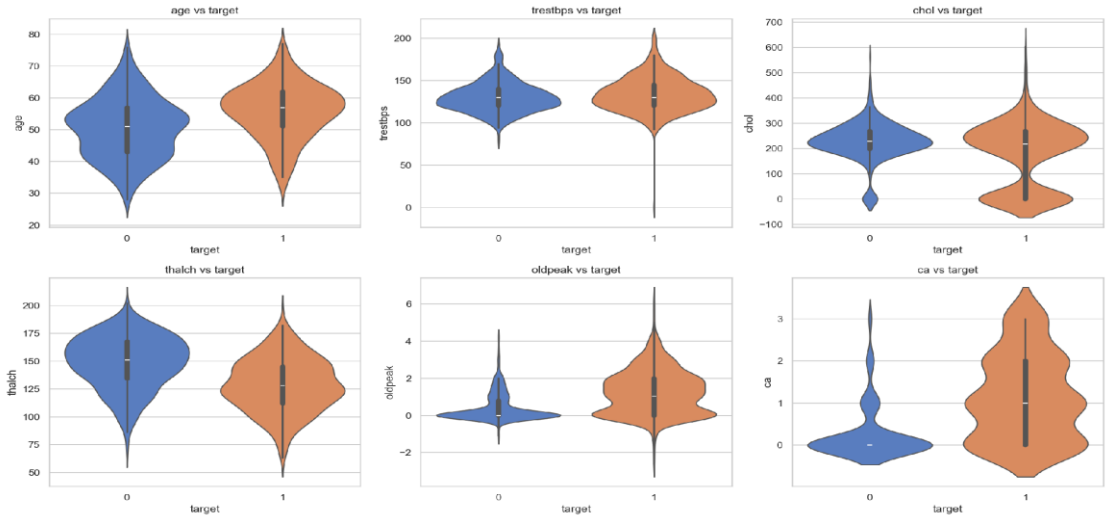

Fig. 5. Violin plots of each numeric feature split by target class, illustrating the class-separating power of thalch, oldpeak, and age.

D. Machine Learning Models

We evaluated five representative supervised classifiers across different inductive biases: Logistic Regression (a linear, highly interpretable baseline), Random Forest (a bagged tree ensemble that structures feature interactions and is resistant to noise), Support Vector Machine (SVM) with a radial basis function kernel (an instance-based method effective in high-dimension spaces), K-Nearest Neighbors (KNN, an interpretable non-parametric instance-based learner), Gradient Boosting(a sequential boosting ensemble). This allows for a within protocol comparison across linear, ensemble, margin-based and instance-based paradigms.

E. Hyperparameter Tuning

Each model was tuned using GridSearchCV with five-fold stratified cross-validation on the training set, optimizing classification accuracy. The searched grids and the selected optimal configurations were as follows:

- **Logistic Regression:** best C = 1; C ∈ {0.01, 0.1, 1, 10}
- **Random Forest:** n_estimators∈{100,200,300}, max_depth ∈{None,5,10}; bestn_estimators=200,max_depth = 10.
- **SVM:** C ∈ {0.1, 1, 10}, kernel ∈ {rbf, linear}; best C = 1, kernel = rbf
- **KNN :** n_neighbors ∈ {3, 5, 7, 9}, weights ∈ {uniform,distance}; best n_neighbors =7;weights= uniform.
- For the Gradient Boosting it was observed that the best parameters were n_estimators = 200, learning_rate = 0.1, max_depth = 2 (100 and 200 were also tried for n_estimators; and likewise 0.01 and 0.1 for learning_rate; layered in max_depth {2,3})

F. Evaluation Protocol

We evaluated model quality from three complementary perspectives. Begin by scoring each tuned model on the held out test set, using accuracy, precision, recall (sensitivity), specificity, F1- score, Matthews correlation coefficient (MCC) and area under the received operating characteristic curve (ROC-AUC). Second, five-fold stratified cross-validation was employed to report mean±standard-deviation accuracy and AUC, thus capturing stability over folds as opposed to quirks of a single split. Third, a paired t-test between the accuracies across all folds from cross-validation was thought of to more statistically test whether or not the differences were important enough to warrant selecting one model over another.

RESULTS

Table 2 reports the full performance profiles of all five tuned models on the held-out test set together with the five-fold cross-validation statistics and the paired significance test against the best cross-validated model. Table 3 lists the ten most important features by Random Forest importance.

Model	Acc.	Prec.	Rec.	Spec.	F1	MCC	AUC	CV Acc. (±sd)	CV AUC (±sd)	p-val
Logistic Regression	0.826	0.824	0.873	0.768	0.848	0.647	0.901	0.817±0.014	0.889±0.017	0.524
Random Forest	0.842	0.841	0.882	0.793	0.861	0.680	0.918	0.810±0.032	0.875±0.029	0.198
SVM (RBF)	0.821	0.805	0.892	0.732	0.847	0.637	0.908	0.825±0.028	0.887±0.023	ref
K-Nearest Neighbors	0.810	0.813	0.853	0.756	0.833	0.614	0.864	0.799±0.029	0.859±0.029	0.049*
Gradient Boosting	0.794	0.796	0.843	0.732	0.819	0.580	0.893	0.796±0.027	0.859±0.026	0.030*

Table 2. Model performance comparison (held-out test set + 5-fold CV). "ref" = reference model with highest CV mean accuracy; * denotes p < 0.05 versus the reference in a paired t-test of CV accuracy.

Rank	Feature	Importance
1	age	0.114
2	thalch (max heart rate achieved)	0.107
3	chol (serum cholesterol)	0.098
4	cp = asymptomatic (chest pain type)	0.096
5	oldpeak (ST depression)	0.085
6	trestbps (resting blood pressure)	0.072
7	exang = False (exercise-induced angina)	0.063
8	exang = True	0.050
9	cp = atypical angina	0.044
10	sex = Female	0.026

Table 3. Top-10 features by Random Forest importance. Correlation-based ranking of numeric features (\|Pearson r\| vs. target): ca (0.456) > thalch (0.395) > oldpeak (0.386) > age (0.283) > chol (0.118) > trestbps (0.117).

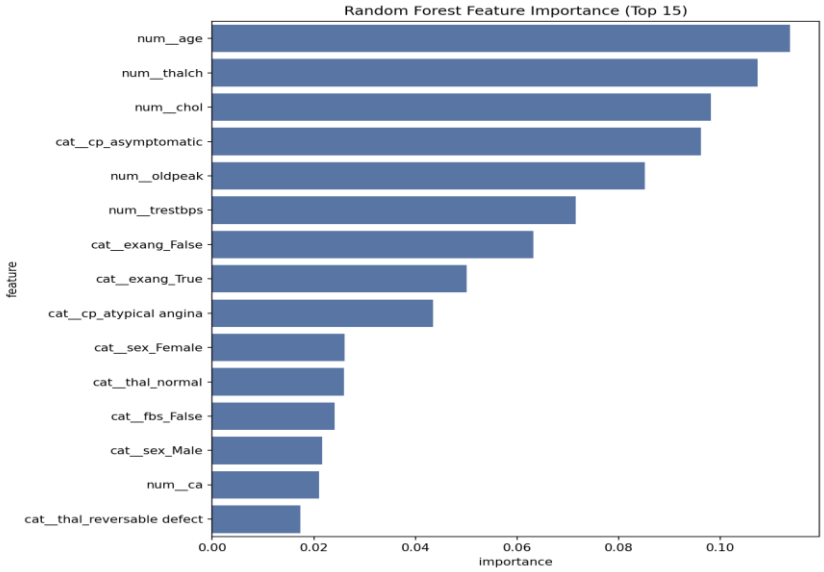

Fig. 6. Random Forest top-15 feature importance. Age, maximum heart rate, cholesterol, asymptomatic chest pain, and ST depression dominate the ranking.

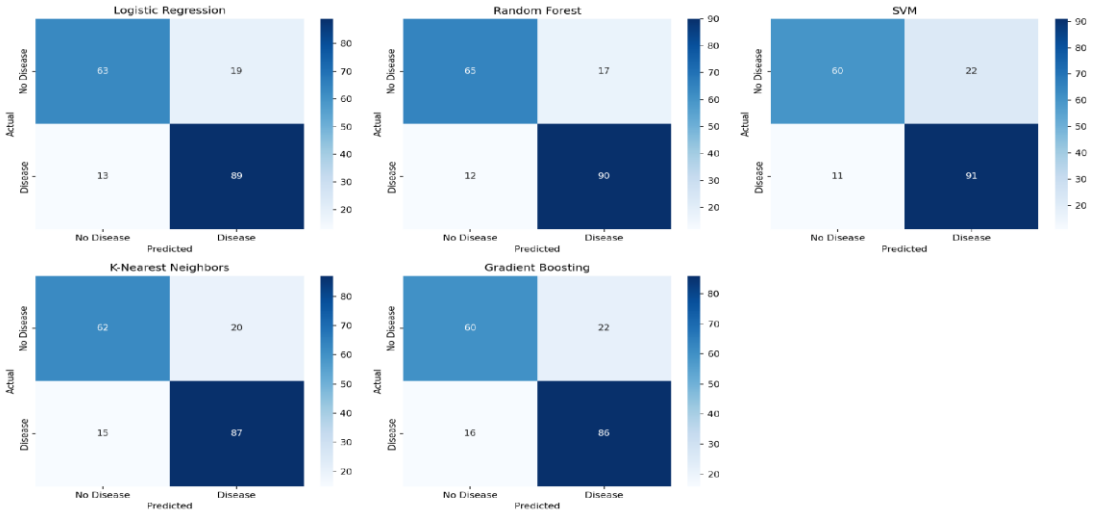

Fig. 7. Confusion matrices for the five tuned models on the held-out test set, showing the balance of false positives and false negatives per model.

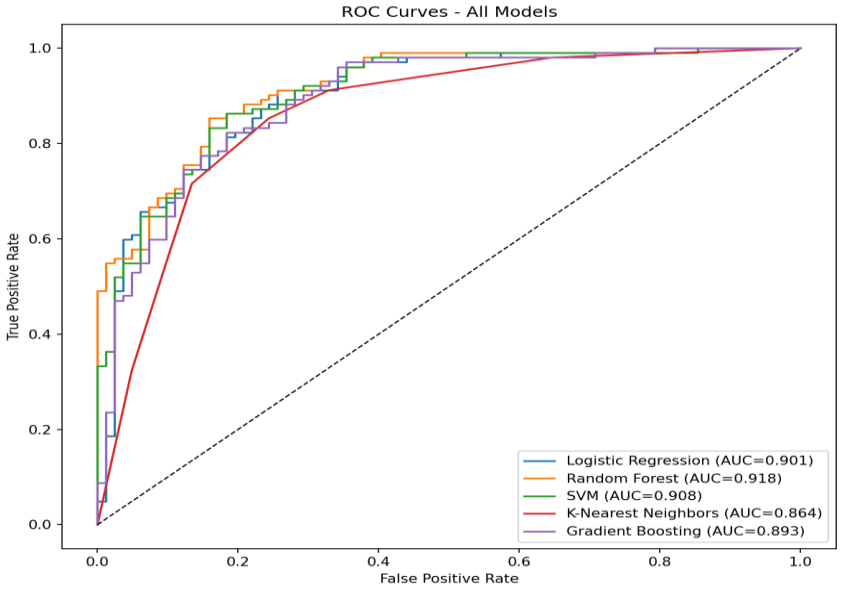

Fig. 8. ROC curves for all five models. Random Forest attains the highest ROC-AUC (0.918), followed by SVM and Logistic Regression.

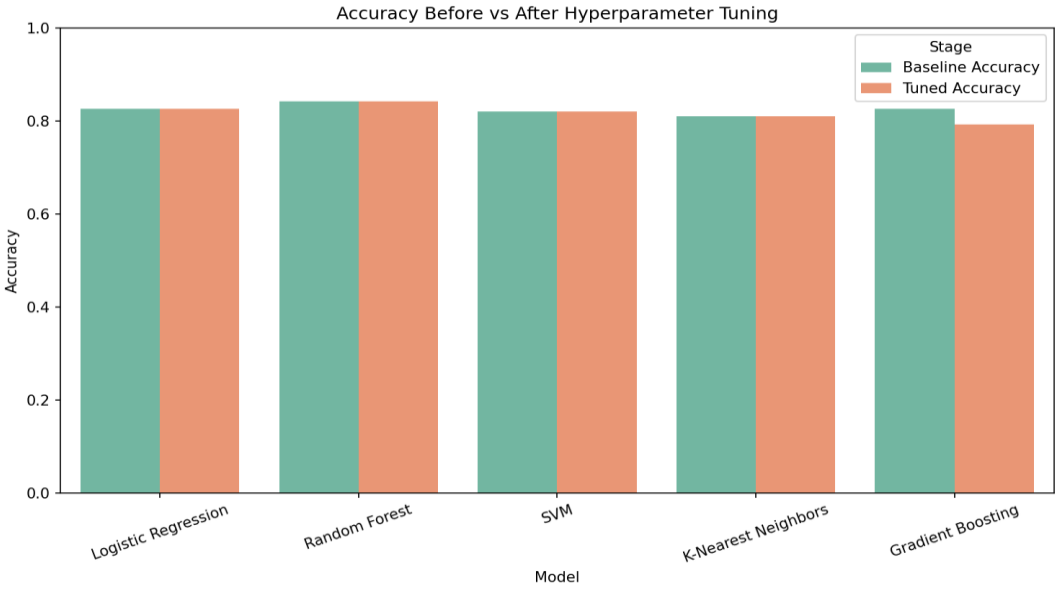

Fig. 9. Accuracy before and after GridSearchCV tuning per model, indicating that tuning yielded modest but consistent gains for most classifiers.

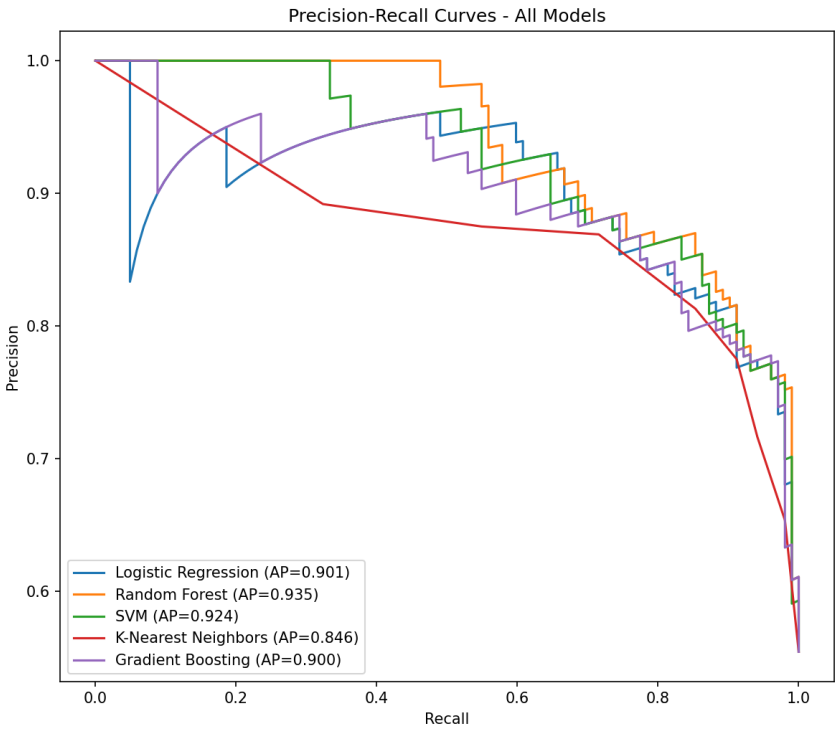

Fig. 10. Precision–recall curves for all five models, relevant given the mild class imbalance toward the disease class.

A. Model Selection

One of the key takeaways is that the best model depends on the criterion chosen. Random Forest achieves the highest score of 0.842 based on single held-out test accuracy. Explicitly, via five-fold cross-validated mean accuracy-the more robust of those criteria as it is an average over folds rather than conditioned on one arbitrary split. SVM achieved the highest score, which was 0.825 ± 0.028 SVM does not differ significantly from Random Forest (p=0.198) or Logistic Regression (p=0.524), but performs better than KNN (p=0.049; p=- 0.030). Thus, Random Forest and SVM are best considered statistically tied for the top spot with Logistic Regression a close and much simpler runner-up, whereas KNN and Gradient Boosting lag behind for weaker cross-validation stability.

DISCUSSION

It is also interesting that at the top of the ranking we see Random Forest converge with SVM, despite their very distinct inductive biases. While Random Forest obtains the best held-out metrics (accuracy 0.842, MCC 0.680, ROC-AUC 0.918), it also has the largest standard deviation about held-out accuracy in cross-validation(±0.032); this variability is associated with bagged tree ensembles passage to success when training folds are perturbed. On the other hand, SVM has not only the highest cross-validation mean accuracy but also a narrow spread (±0.028) due to its relatively robustness gained through the margin maximization which is less sensitive to small perturbations of examples composing training sets. So, given that Random Forest is expected to generalize competitively while providing on-feature-importance estimates in a native way and SVM has slightly more stable average performance; it seems that the choice can be controlled by secondary considerations (e.g., interpretability and inference cost).

Logistic Regression is especially interesting: it is statistically indistinguishable from the leaders but also the simplest and most interpretable model that exists, making it a natural choice in medical scenarios where interpretability matters more than performance.

The analysis of the features is in accordance with known cardiological data. Random Forest ranking is led by age, maximal heart rate achieved (thalch), serum cholesterol, asymptomatic chest pain and ST depression (oldpeak)± all established correlates of ischemic heart disease.

A diminished maximal heart rate and significant exercise induced ST depression are classic signs of impaired cardiac reserve while asymptomatic chest pain is clinically important simply because it can be so easily missed. Prominence of these features is a support for the clinical plausibility of learned models and increases confidence that predictions arise from physiologically interpretable signals, rather than dataset artifacts.[18]

The significance testing only serves as a reinforcing warning about model selection. The lack of a statistically significant gap between the top three performing models suggests that accuracy at one or many top-k should not be reported as a single measure since, as is done in much previous research, it would misrepresent certainty in ranking-carrying on from empirical

to exaggerate-and leave out important qualitative differences across methods. In contrast, the higher cross-vallidation means of KNN and Gradient Boosting have evidential deficits: they are more plausible to be accurate due to their respective p-values < 0.05, meaning that they are actually surviving overfit on this cohort likely because univariate sensitivity due to curse of dimensionality in parametric model with the 27 np-1 (encoded) features plus n+1 constant terms,[19] while underparameterization of tuned Gradient Boosting depth-2 trees bears fruit - a tamed random forest that disentangles interaction limits discovery potential.

A less evident but much important difference regards the ca feature (number of major vessels). With univariate correlation, ca ranks as the single strongest numeric predictor of the target (|r| = 0.456), followed only by thalch and oldpeak; however it is not present in the top-10 importance plots for Random Forests. The intense missingness of ca (66% of rows) resolves this apparent contradiction. The strong correlation is calculated on the observed subset for which ca was present while most of the entries have been filled by one median constant. A feature which is constant across some 2/3rds of the training data provides very little split information when viewed from the point of view of the trees, so its purity-based importance collapses even though its observed values are highly predictive. This serves as a general warning against conflating univariate correlation and model-based importance under heavy imputation, as well as dropping either perspective in isolation.[20]

LIMITATIONS

Several limitations qualify these findings. For one, the cohort is heterogeneous: we pooled together four sites with different acquisition protocols, resulting in a distributional shift for which a single pooled model may not generalize fully well. Secondly, the analysis is based on heavy imputation of ca and thal fields, which are highly missing so a lot of these values have been inferred rather than measured, potentially declining their contribution.

Third, collapsing the five-level severity target into a binary label loses clinically relevant gradation in disease severity. Fourth, the evaluation is limited to a single dataset with no externally validated cohort and thus estimates may not generalise to independent populations. Lastly, the sample size of 920 records is decent for the classical models investigated, but diminutive for data-hungry deep learning methods

FUTURE WORK

Future Work will push in multiple directions. Cross validation on external cardiac cohorts is a major aim to assess transferability past the four merged sites. As the sample size increases, we will consider deep learning and ensemble-stacking strategies that have been successful for similar medical-imaging and diagnostic tasks [6]-[8].

We will hone model interpretability through per-patient, feature-level explanations using SHAP-based attribution to align with the explainable-AI trend seen in recent healthcare literature [5], [7], [9]. Instead of performing simple median or mode imputation, as we do in our regular model, here, missingness will instead be dealt with using a reliance on methods based again on principles from multiple imputation (e.g., MICE), which should restore more predictive value for the ca and thal fields. The prediction uncertainty through the recently proposed uncertainty-aware diagnostic frameworks [10]–[12] will then be calibrated for relaying the confidence to clinicians. Finally, we will consider site-level heterogeneity explicitly-potentially via federated [3] or privacy-preserving [5] institutional learning processes such that no sensitive patient data is centrally preserved as a shared model is learned, and the entire system will be architected under reproducibility-oriented iterative development processes [17].

CONCLUSION

In this dual-site study, we report an open-source and statistically grounded comparison of five binary machine learning algorithms for heart disease prediction on a merged cohort (920 records total) across four clinical sites. Random Forest achieved the best single-split test accuracy (84.2%) under a common pipeline of physiologically informed cleaning, imputation, tuning and multi-lens evaluation. SVM was the only model to achieve accuracy above 80% in cross-validation (82.5%) while significance testing indicated both SVM and RF were indistinguishable from Logistic Regression, so these three models formed a joint best-performing group. The main feature analysis showed that age, max heart rate, cholesterol, chest pain type and ST depression were the main predictors as expected from clinical practice and resolved a mismatch in ca importance due to imputation. The main takeaway is that model selection with cross-validation and significance testing can provide a more transparent foundation for selecting clinical predictors, compared to the single held-out accuracy number.

REFERENCES

1. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, “Heart Disease Data Set,” UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences.
2. S. U. Rashid, M. I. H. Siddiqui, F. U. Mahmud, M. S. Rahman, A. A. Kabir, and R. S. Shammah, “Machine learning based clinical decision support for heart disease prediction using structured patient data,” Journal of Computer Science and Technology Studies, vol. 6, no. 1, pp. 340–350, 2024.
3. A. A. Kabir, F. U. Mahmud, M. S. Rahman, S. U. Rashid, M. I. H. Siddiqui, and R. S. Shammah, “Multimodal

machine learning framework for privacy preserving and scalable cancer diagnosis across healthcare systems," Journal of Adaptive Learning Technologies, vol. 1, no. 6, 2024.

4. S. Khandakar, M. A. Al Mamun, M. M. Islam, M. Minhas, and N. Al Huda, "Unlocking cancer prevention in the era of AI: Machine learning models for risk stratification and personalized intervention," Educational Administration: Theory and Practice, vol. 30, no. 8, pp. 269–283, 2024.
5. M. I. H. Siddiqui, K. K. Bishnu, M. A. Al Mamun, M. Raihan, A. Islam, S. Akter, and I. Hossain, "Explainable federated deep learning for low-cost and privacy-preserving early breast cancer screening to reduce US healthcare burden," Vascular and Endovascular Review, vol. 6, no. 2, pp. 45–54, 2023.
6. R. Haque, M. A. Khan, H. Rahman, S. Khan, M. I. H. Siddiqui, Z. H. Limon, et al., "Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis," Computers in Biology and Medicine, vol. 191, art. 110166, 2025.
7. M. I. H. Siddiqui, S. Khan, Z. H. Limon, H. Rahman, M. A. Khan, A. Al Sakib, et al., "Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI," Informatics in Medicine Unlocked, vol. 56, art. 101657, 2025.
8. M. R. Ahmed, H. Rahman, Z. H. Limon, M. I. H. Siddiqui, M. A. Khan, A. S. U. K. Pranta, et al., "Hierarchical Swin transformer ensemble with explainable AI for robust and decentralized breast cancer diagnosis," Bioengineering, vol. 12, no. 6, art. 651, 2025.
9. S. Islam, R. Haque, M. A. Khan, A. B. Mohiuddin, M. I. H. Siddiqui, Z. H. Limon, et al., "Ensemble transformer with post-hoc explanations for depression emotion and severity detection," iScience, vol. 29, no. 2, 2026.
10. T. Mahmud, M. A. Al Mamun, N. Jahan, S. R. Madhu, M. S. Uddin, and M. M. H. Shimul, "OvCaNet: An uncertainty-calibrated deep learning framework for benign–malignant ovarian cancer classification from ultrasound imaging," Intelligence-Based Medicine, art. 100458, 2026.
11. F. R. Supti, M. K. Refat, M. M. Mamun, S. Dutta, A. Kader, F. M. Alamgir, et al., "Ensemble deep learning for automated skin disease classification: A multi-scale patient-level validation framework with uncertainty quantification."
12. N. Zuberi, "Robust lung cancer CT classification via Monte Carlo cross-validated multi-model benchmarking and confidence-weighted ensembling," Frontiers in Medicine, vol. 13, art. 1842385.
13. M. D. Hossain, F. Akter, M. Rahaman, B. Biswas, and H. M. Hasan, "Hierarchical federated learning with heavy-hitter detection for real-time cyber threat mitigation in large-scale networks," SSRN 6340898, 2026.
14. "Federated learning for cross-institutional fraud monitoring in national healthcare security," International Journal of AI, Engineering and Management Studies (IJAIEMS), vol. 1, no. 1, pp. 137–155, 2026.
15. Hosen, M. S., Al Mamun, M. A., Khandakar, S., Hossain, K., Islam, M. M., & Alkhayyat, A. (2024). Cybersecurity meets data science: a fusion of disciplines for enhanced threat protection. *Nanotechnology Perceptions*, 236-256.
16. B. Biswas, S. M. Faysal, S. Das, A. Hanif, T. Akter, R. A. M. Rashed, and S. Shamarukh, "A framework for agentic AI-driven financial fraud detection, investigation, and risk mitigation in real-time payment ecosystems," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 3s, pp. 1029–1040, 2026.
17. M. S. K. Chy, "Tailoring the Scrum framework for non-IT small projects: A longitudinal action research study on US business adoption," Global Social Sciences Review, vol. VIII, no. II, pp. 681–697, 2023.
18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
19. Shuchona Malek Orthi, Md Habibur Rahman, Kazi Bushra Siddiqa, Mukther Uddin, Sazzat Hossain, Abdullah Al Mamun, & Md Nazibullah Khan. (2025). Federated Learning with Privacy-Preserving Big Data Analytics for Distributed Healthcare Systems. *Journal of Computer Science and Technology Studies*, *7*(8), 269-281.
20. "Artificial Intelligence-Driven Early Detection of Neurological Disorders and Its Implications for Personalized Rehabilitation Strategies", *VER*, vol. 6, no. 2, pp. 104–111, Nov. 2023, doi: 10.64149/.