Revisiting Ensemble Superiority Claims in Heart Disease Prediction: A Statistically Rigorous Re-Evaluation with Bayesian-Optimized Classical and Ensemble Models

Nezar Abdilah Prakasa

Master of Artificial Intelligence, Department of Computer Science and Electronics, FMIPA, Universitas Gadjah Mada, Yogyakarta, Indonesia

nezarabdilahprakasa@mail.ugm.ac.id

Abstract: Recent studies on machine learning-based heart disease prediction frequently claim superior performance for complex ensemble and quantum-inspired classifiers over conventional models, typically based on a single train-test split without statistical validation. This study re-examines such claims using the Cleveland Heart Disease dataset (303 samples, UCI Machine Learning Repository) through a rigorous statistical validation framework. We replace the quantum classifiers used in a prior explainable ensemble-quantum study with classical gradient-boosting ensembles (XGBoost, LightGBM, and a stacking ensemble), and compare them against a Bayesian-optimized Support Vector Classifier (SVC) baseline using Optuna. All models were evaluated with 10-times-repeated 10-fold stratified cross-validation (100 F1-score observations per model), followed by the Friedman test, Nemenyi post-hoc analysis, and pairwise Wilcoxon signed-rank tests. Results show that although the stacking ensemble achieved the highest mean F1-score (0.8605 ± 0.0521), the difference from the tuned SVC (0.8584 ± 0.0509) was not statistically significant (Wilcoxon $p = 0.786$; Nemenyi $p = 0.993$), while the SVC required approximately 98% less computation time (0.113 s vs. 5.927 s per 10-fold run). Both models were, however, significantly superior to a multilayer perceptron baseline ($p < 0.001$). These findings indicate that model complexity does not guarantee statistically meaningful performance gains on small clinical tabular datasets and highlight the necessity of proper significance testing before claiming model superiority in the heart disease prediction literature.

Index Terms: *heart disease prediction, statistical significance testing, ensemble learning, hyperparameter optimization, Friedman test, Wilcoxon signed-rank test, Cleveland dataset*

I. INTRODUCTION

Cardiovascular disease (CVD), including coronary and ischemic heart disease, remains the leading cause of mortality worldwide, and early risk prediction is critical for timely clinical intervention [23]. Machine learning (ML) has therefore been extensively applied to structured clinical datasets such as the UCI Cleveland Heart Disease dataset for automated risk classification [4], [7], [17].

A large body of recent literature reports that complex ensemble architectures, hybrid classifiers, or exotic learners such as quantum-inspired classifiers outperform conventional models on this task [1], [3], [8], [15], [21], [22]. However, the majority of these studies base their superiority claims on a single train–test split or a single run of k-fold cross-validation, reporting only point estimates of accuracy or F1-score without any test of statistical significance [5], [6], [9], [14], [19]. In small clinical tabular datasets, where the Cleveland dataset contains only 303 samples, such point estimates are highly sensitive to sampling variance, and apparent performance differences may not reflect a real, reproducible advantage of one model over another.

This study directly builds on the ensemble-quantum machine learning study of Abdulsalam et al. [1], which proposed a bagging ensemble of Quantum Support Vector Classifiers (Bagging-QSVC) for heart disease prediction on the Cleveland dataset and reported it as superior to several classical baselines. We replace the quantum classifiers, which require specialized simulators and are impractical for routine clinical deployment, with classical gradient-boosting ensembles, and subject every model, including the baseline, to an identical, statistically rigorous evaluation protocol.

The contributions of this paper are as follows:

1) We re-examine ensemble superiority claims in heart disease prediction using a repeated stratified cross-validation protocol combined with the Friedman test, Nemenyi post-hoc analysis, and pairwise Wilcoxon signed-rank tests [24]–[26], which are largely absent from prior heart disease prediction studies.

2) We ensure a fair comparison by applying Bayesian hyperparameter optimization (Optuna [32]) to every model, including the classical baseline, rather than tuning only the proposed ensemble.

3) We introduce computational efficiency as an additional, practically relevant evaluation criterion alongside predictive performance.

4) We provide evidence-based, rather than accuracy-only, recommendations for model selection in resource-constrained clinical deployment settings.

II. RELATED WORK

Several studies have applied ensemble learning to the Cleveland dataset. Das and Sengur [9] evaluated ensemble methods for valvular heart disease diagnosis. Mohan et al. [4] proposed a hybrid random forest with linear model (HRFLM) achieving 88.7% accuracy. Ghosh et al. [3] combined Relief and LASSO feature selection with several classifiers, reporting up to 99.05% accuracy with a rotation forest model. Shorewala [5] compared bagging, boosting, and stacking techniques for early coronary heart disease detection. Majumder et al. [8] proposed a concatenated hybrid ensemble classifier and reported improved accuracy over individual base learners.

More recent work has explored quantum and quantum-inspired approaches. Abdulsalam et al. [1] proposed an explainable Bagging-QSVC ensemble combined with SHAP interpretation. Kavitha and Kaulgud [15] applied quantum k-means clustering for heart disease detection. While these studies report favorable point-estimate performance, none apply repeated cross-validation together with a formal statistical significance test to substantiate their superiority claims, a gap explicitly identified as a methodological weakness across the broader ML-for-healthcare literature [23].

Feature selection and preprocessing strategies have also been widely studied. Pathan et al. [6] analyzed the impact of feature selection on prediction accuracy. Alfadli and Almagrabi [17] investigated feature-limited prediction settings. Ozcan and Peker [18] applied classification and regression trees (CART), while Chandrasekhar and Peddakrishna [16] combined six classical classifiers with GridSearchCV-based hyperparameter tuning and 5-fold cross-validation, reporting logistic regression as the best performer on the Cleveland subset, a finding broadly consistent with the results of the present study, which likewise finds a simple, well-tuned model competitive with more complex ensembles.

The statistical methodology adopted in this paper follows established guidance for comparing multiple classifiers across multiple resampled folds. Demšar [24] recommends the Friedman test followed by the Nemenyi post-hoc test as the appropriate non-parametric procedure when comparing more than two classifiers over repeated folds, building on the original rank-based test of Friedman [25] and the pairwise signed-rank test of Wilcoxon [26]. These tests are standard in the broader machine learning literature but remain rarely applied in heart disease prediction studies specifically, motivating the present re-evaluation.

III. METHODOLOGY

A. Dataset

This study uses the Cleveland Heart Disease dataset from the UCI Machine Learning Repository, as redistributed in the replication package of Abdulsalam et al. [1]. The dataset contains 303 patient records with 13 clinical attributes (e.g., age, sex, chest pain type, resting blood pressure, cholesterol, and maximum heart rate) and a target variable indicating the presence of heart disease. Missing values, encoded as '?', were imputed using the median of each feature. The multi-class target was binarized into a two-class problem (0 = no disease, 1 = disease), following the original study's [1] and prior conventions [4], [7], resulting in 165 positive and 138 negative cases, reflecting a mild class imbalance.

B. Model Configuration

Three classical baseline models from the prior literature were retained: a Support Vector Classifier (SVC) with a radial basis function (RBF) kernel [27], a Multilayer Perceptron (MLP/ANN), and a Bagging ensemble of SVCs [28]. The quantum classifiers from the base study [1] (QSVC, QNN, VQC) were replaced with three classical gradient-boosting-based models: XGBoost [30], LightGBM [31], and a Stacking ensemble [29] combining XGBoost, LightGBM, and SVC with a logistic regression meta-learner, as these are computationally practical alternatives that are directly comparable in a standard (non-quantum-simulated) computing environment such as Google Colaboratory.

C. Hyperparameter Optimization

To ensure a fair comparison, every tunable model (SVC, XGBoost, LightGBM) was optimized using the Tree-structured Parzen Estimator sampler implemented in Optuna [32], with 50 trials per model and an inner 5-fold stratified cross-validation objective (mean F1-score). Critically, the SVC baseline was tuned under an identical budget to the ensemble models, avoiding the common methodological flaw of tuning only the proposed method. Table II reports the best hyperparameters found.

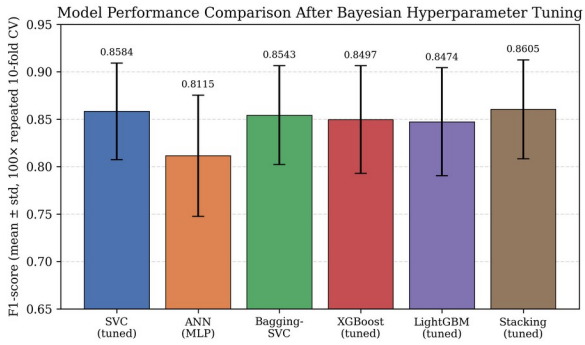

Fig. 1. Mean F1-score (± std) across six models over 100 repeated stratified 10-fold cross-validation runs, after Bayesian hyperparameter tuning.

D. Evaluation Protocol

All six models (three baselines, three ensemble/tuned variants) were evaluated using 10-times-repeated stratified 10-fold cross-validation (100 F1-score observations per model), which provides a substantially more robust performance estimate than the single-split evaluations used in [1], [3]–[5], [8]. Statistical comparison across all six models was performed using the Friedman test; where the Friedman test indicated a significant overall difference ($\alpha = 0.05$), the Nemenyi post-hoc test was applied to identify which specific pairs of models differed significantly. In addition, a paired Wilcoxon signed-rank test was applied directly between the top-ranked model (by mean F1-score) and the fastest competitive alternative, explicitly checking the direction of the difference in addition to its significance. Wall-clock training time across the 10 folds was also recorded as a secondary, practically motivated evaluation criterion.

IV. RESULTS

Table I summarizes the mean and standard deviation of the F1-score for all six models across 100 repeated cross-validation folds, along with the Nemenyi post-hoc p-value relative to the top-ranked model (Stacking).

TABLE I
F1-SCORE SUMMARY ACROSS 100 REPEATED CV FOLDS

Model	Mean F1	Std F1	Rank Nemenyi vs Stacking (p)
Stacking (tuned)	0.8605	0.0521	1.000 (ref)
SVC (tuned)	0.8584	0.0509	0.993
Bagging-SVC	0.8543	0.0522	0.531
XGBoost (tuned)	0.8497	0.0567	0.058
LightGBM (tuned)	0.8474	0.0569	0.038*
ANN (MLP)	0.8115	0.0640	0.000*

The Friedman test rejected the null hypothesis of equal performance across all six models ($\chi^2 = 82.84$, $p < 0.001$), confirming that at least one model differs significantly from the others. The Nemenyi post-hoc analysis (Fig. 2) shows that the Stacking ensemble (mean F1 = 0.8605) and the tuned SVC (mean F1 = 0.8584) are statistically indistinguishable ($p = 0.993$), as are SVC and Bagging-SVC ($p = 0.867$). In contrast, both the ANN/MLP baseline and, marginally, LightGBM were significantly outperformed by the top-ranked models ($p < 0.05$).

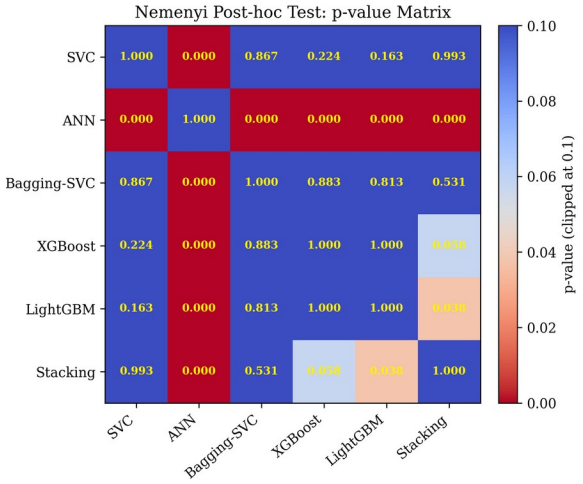

Fig. 2. Nemenyi post-hoc test p-value matrix across all six model pairs. Cells with $p < 0.05$ (darker) indicate a statistically significant performance difference.

A direct pairwise Wilcoxon signed-rank test between the Stacking ensemble and the tuned SVC confirmed the Nemenyi result: $W = 1196$, $p = 0.786$, with a mean F1 difference of only 0.0021, well within the range

attributable to sampling variance rather than a genuine model advantage.

Fig. 3 reports the computational cost of each model over the 10-fold training runs. The tuned SVC required only 0.113 s in total, compared to 5.927 s for the Stacking ensemble (approximately 52× slower) and 12.609 s for the ANN/MLP (approximately 111× slower), despite statistically equivalent or inferior predictive performance.

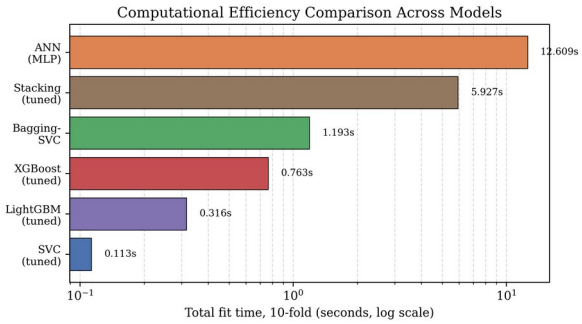

Fig. 3. Total model fitting time across 10 stratified folds (log scale). The tuned SVC is markedly faster than all ensemble alternatives at statistically comparable predictive performance.

TABLE II
BEST HYPERPARAMETERS FOUND VIA OPTUNA (50 TRIALS)

Model	Hyperparameter Terbaik (Optuna, 50 trial)
SVC	C=89.06, gamma=1.55e-4, kernel=rbf
XGBoost	n_estimators=265, max_depth=3, lr=0.168, subsample=0.564
LightGBM	n_estimators=391, max_depth=6, lr=0.0109, num_leaves=26

V. DISCUSSION

The central finding of this study is that, once subjected to rigorous statistical validation, the performance advantage typically attributed to complex ensemble architectures over simpler, properly tuned classifiers does not hold on the Cleveland dataset. This contrasts with the framing of several prior studies [1], [3], [8], [15], which report point-estimate accuracy gains for ensemble or quantum-inspired classifiers without testing whether those gains exceed what would be expected from cross-validation sampling variance alone.

This result is consistent with the broader observation in the general machine learning literature that, on small tabular datasets, the inductive bias of simpler models such as SVMs can match or exceed that of more complex gradient-boosting ensembles once both are given an equal hyperparameter tuning budget [24], [27]. Similar conclusions were reached by Chandrasekhar and Peddakrishna [16], who found logistic regression competitive with more elaborate pipelines on the same dataset family.

From a practical deployment perspective, the computational efficiency gap is arguably as clinically relevant as the (statistically insignificant) predictive performance gap. A model requiring 0.113 s to train is far more suitable for resource-constrained clinical settings, frequent retraining, or edge deployment than a five-model stacking ensemble requiring nearly 6 s, particularly given that the latter offers no statistically demonstrable accuracy benefit.

This study has several limitations. First, the Cleveland dataset is relatively small (303 samples) and drawn from a single clinical site, which may limit generalizability; validation on larger, multi-site cardiovascular datasets is an important direction for future work. Second, only F1-score was used as the primary CV metric for statistical testing; extending the same statistical protocol to AUC-ROC and calibration metrics would further strengthen the conclusions. Third, this study focused on tabular clinical features; extending the comparison to ECG-signal or imaging-based deep learning models is left for future work.

VI. CONCLUSION

This study revisited claims of ensemble superiority in heart disease prediction on the Cleveland dataset through a statistically rigorous re-evaluation protocol combining repeated stratified cross-validation, the Friedman–Nemenyi procedure, and pairwise Wilcoxon signed-rank testing, with all models, including the baseline, given an equal Bayesian hyperparameter tuning budget. The results show that a properly tuned Support Vector Classifier achieves predictive performance statistically indistinguishable from a stacking ensemble of gradient-boosting models, while requiring approximately 98% less computation time. Both substantially outperform a multilayer perceptron baseline. These findings underscore the necessity of formal statistical significance testing before claiming model superiority in clinical ML studies, and suggest that simpler, well-tuned models remain a strong, efficient choice for heart disease risk prediction in resource-constrained settings. Future

work will extend this statistical validation protocol to larger multi-site cardiovascular datasets and additional performance metrics.

REFERENCES

[1] G. Abdulsalam et al., "Explainable Heart Disease Prediction Using Ensemble-Quantum Machine Learning Approach," Intell. Autom. Soft Comput., vol. 36, no. 1, pp. 761–779, 2023.

[2] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," IEEE Access, vol. 7, pp. 60134–60149, 2019.

[3] P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," IEEE Access, vol. 9, pp. 19304–19326, 2021.

[4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019.

[5] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," Inform. Med. Unlocked, vol. 26, p. 100655, 2021.

[6] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," Healthc. Anal., vol. 2, p. 100060, 2022.

[7] H. Khdair and N. Dasari, "Exploring Machine Learning Techniques for Coronary Heart Disease Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 5, 2021.

[8] A. B. Majumder, S. Gupta, D. Singh, B. Acharya, V. C. Gerogiannis, A. Kanavos, and P. Pintelas, "Heart Disease Prediction Using Concatenated Hybrid Ensemble Classifiers," Algorithms, vol. 16, no. 12, p. 538, 2023.

[9] R. Das and A. Sengur, "Evaluation of ensemble methods for diagnosing of valvular heart disease," Expert Syst. Appl., vol. 37, no. 7, pp. 5110–5115, 2010.

[10] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, and A. Elmaghraby, "Ensemble deep learning models for heart disease classification: A case study from Mexico," Information, vol. 11, no. 4, p. 207, 2020.

[11] F. Ali, S. El-Sappagh, S. M. R. Islam, D. Kwak, A. Ali, et al., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," Inf. Fusion, vol. 63, pp. 208–222, 2020.

[12] M. Dubey, J. Tembhurne, and R. Makhijani, "Heart Failure prediction on diversified datasets to improve generalizability using 2-Level Stacking," Multidiscip. Sci. J., vol. 6, no. 3, 2024.

[13] B. Paul and B. Karn, "Heart disease prediction using scaled conjugate gradient backpropagation of artificial neural network," Soft Comput., vol. 27, no. 10, pp. 6687–6702, 2023.

[14] T. Karadeniz, G. Tokdemir, and H. H. Maraş, "Ensemble Methods for Heart Disease Prediction," New Gener. Comput., vol. 39, no. 3, pp. 569–581, 2021.

[15] S. S. Kavitha and N. Kaulgud, "Quantum K-means clustering method for detecting heart disease using quantum circuit approach," Soft Comput., vol. 27, no. 18, pp. 13255–13268, 2023.

[16] A. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," Processes, vol. 11, no. 4, p. 1210, 2023.

[17] K. M. Alfadli and A. O. Almagrabi, "Feature-Limited Prediction on the UCI Heart Disease Dataset," Comput. Mater. Contin., vol. 74, no. 3, pp. 5871–5883, 2022.

[18] M. Ozcan and S. Peker, "A Classification and Regression Tree Algorithm for Heart Disease Modeling and Prediction," Healthc. Anal., vol. 3, p. 100130, 2023.

[19] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques," J. Healthc. Eng., vol. 2022, Art. no. 2585235, 2022.

[20] R. Yilmaz and F. H. Yağin, "Early Detection of Coronary Heart Disease Based on Machine Learning Methods," Med. Rec., vol. 4, pp. 1–6, 2021.

[21] B. A. Tama and K.-H. Rhee, "Tree-based classifier ensembles for early detection of heart disease," in Computational Intelligence in Biomedical Science and Engineering, Singapore: Springer, 2019, pp. 27–44.

[22] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," BioMed Res. Int., vol. 2020, 2020.

[23] S. J. Al'Aref et al., "Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging," Eur. Heart J., vol. 40, no. 24, pp. 1975–1986, 2019.

[24] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.

[25] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," J. Am. Stat. Assoc., vol. 32, no. 200, pp. 675–701, 1937.

[26] F. Wilcoxon, "Individual Comparisons by Ranking Methods," Biometrics Bull., vol. 1, no. 6, pp. 80–83, 1945.

[27] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[28] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.

[29] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241–259, 1992.

[30] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[31] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[32] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.