Fair and Statistically Rigorous Re-Evaluation of Machine Learning Classifiers for Heart Failure Survival Prediction: A Paired Repeated Cross-Validation Approach

Nezar Abdilah Prakasa
Master of Artificial Intelligence, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract - Chicco and Jurman (2020) demonstrated that random forests and other classifiers can predict survival of heart failure patients from a public 299-patient clinical dataset, releasing a complete replication package on GitHub. Inspection of that package reveals two methodological weaknesses that limit the validity of the reported classifier ranking: each classifier script draws its own 100 independent random resamples with no shared seed, so the reported comparisons are unpaired, and no classifier undergoes systematic hyperparameter tuning or formal significance testing. This study re-evaluates the same eight classifier families on the same data using an identical, paired, repeated 10x10-fold stratified cross-validation protocol, an identical Optuna hyperparameter search budget for every model, and a Friedman test with Nemenyi post-hoc and pairwise Wilcoxon signed-rank tests. Random forest achieves the highest mean Matthews Correlation Coefficient (MCC = 0.657), significantly outperforming six of the seven remaining classifiers (Nemenyi $p < 0.05$) but not logistic regression ($p = 0.236$), which is a nuance the original single-split protocol could not detect. Training-time analysis further shows decision tree and logistic regression reach within 0.05-0.07 MCC of random forest at under 2 ms training time per fold, a practical efficiency trade-off absent from the original study. The results confirm the original study's headline claim while substantially revising its confidence and completeness.

Keywords - *heart failure prediction, classifier evaluation, repeated cross-validation, Friedman test, Nemenyi post-hoc, Wilcoxon signed-rank, hyperparameter tuning, replication study*

I. INTRODUCTION

Cardiovascular diseases remain the leading cause of death worldwide, and heart failure (HF) is among their most severe manifestations, arising when the heart can no longer pump sufficient blood to meet the body's needs [3]. Predicting which HF patients face the highest mortality risk from routinely collected clinical variables has therefore attracted sustained interest from the machine learning community.

Chicco and Jurman [1] analyzed 299 HF patients using ten classifiers and concluded that serum creatinine and ejection fraction alone are sufficient predictors of survival, a finding that has since been cited widely and re-used as a benchmark task. Distinctively for a clinical machine learning study, the authors released a public, runnable replication package on GitHub containing one R script per classifier, each performing 100 independent random train-test executions and reporting the mean of several metrics.

Re-running and reading that package for the present study surfaced two properties that were not discussed in the original paper. First, because each script shuffles the dataset independently and no random seed is fixed, the 100 executions of one classifier are drawn from a different sequence of splits than the 100 executions of any other classifier; the resulting per-model averages are therefore not paired observations, which rules out valid application of paired tests such as Wilcoxon signed-rank even though the original references cite that very test [8]. Second, every classifier is trained with the default hyperparameters of its R implementation; no tuning stage exists, and the published ranking of classifiers is consequently confounded with how favorable each library's defaults happen to be for this particular dataset. Neither weakness invalidates the original conclusions, but both weaken the statistical basis on which classifier rankings were drawn.

This paper re-evaluates the same eight classifier families that overlap with common scikit-learn implementations of the original study's methods (random forest, decision tree, gradient boosting, logistic regression, naive Bayes, linear and radial-basis-function support vector machines, and k-nearest neighbors) on the same publicly released dataset, correcting both weaknesses while changing nothing else about the task. The specific contributions are: (i) an identical, paired, repeated stratified cross-validation protocol shared across all eight classifiers, replacing the unpaired single-execution-set design; (ii) an identical hyperparameter search budget applied to every classifier via Optuna [21], removing the default-hyperparameter confound; and (iii) a Friedman omnibus test with Nemenyi post-hoc comparison and pairwise Wilcoxon signed-rank tests, replacing the absence of any formal significance test in the original protocol.

II. RELATED WORK

Beyond the original study [1], the same 299-patient dataset originates from a survival analysis by Ahmad et al. [2], who used a Cox proportional-hazards model and identified age, renal function, blood pressure, ejection fraction and anemia as significant risk factors, providing a clinical statistics baseline against which later machine learning re-analyses can be compared.

The choice of evaluation metric matters for imbalanced clinical outcomes. Chicco and Jurman [4] argue that the Matthews Correlation Coefficient (MCC) is more informative

than accuracy or F1 score because it accounts for all four confusion-matrix cells and remains meaningful under class imbalance; this study adopts MCC as its primary metric, consistent with the original authors' own methodological recommendation, in addition to reporting F1, accuracy and AUC [25] for comparability with the wider literature.

The statistical machinery used here to compare classifiers is standard in the machine learning evaluation literature but was not applied in the original study. Demsar [5] recommends the Friedman test [6] followed by a Nemenyi post-hoc procedure [7] as the appropriate non-parametric approach when comparing multiple classifiers across multiple resampled datasets, and this recommendation has become a de-facto standard for classifier benchmarking papers. Dietterich [9] earlier showed that naive paired t-tests on resampled folds inflate false-positive rates because the folds are not independent, motivating the use of rank-based tests instead. Separately, Kohavi [10], Varma and Simon [11], and Vabalas et al. [12] each document how the choice of validation scheme, and in particular the danger of a single train-test split on small clinical datasets, can produce optimistic or unstable estimates of generalization performance, which is the concern that motivates replacing the original study's single-split-per-execution design with repeated cross-validation in this work.

Each classifier family evaluated here has a well-established origin: random forests [13], support vector machines [14], gradient boosting [15], k-nearest neighbors [16], classification and regression trees [17], the ID3/C4.5 family of decision-tree induction [18], logistic regression [19], and naive Bayes [20]. Hyperparameter search itself has a body of methodology distinct from the classifiers being tuned, including random search [23], Bayesian optimization [24], and the tree-structured Parzen estimator implemented in the Optuna framework [21] used in this study.

III. METHODOLOGY

A. Dataset

The dataset is the same 299-patient, 12-feature heart failure clinical records table released with the original replication package [1], sourced originally from Ahmad et al. [2]. The binary target is the death event (Event = 1) during the follow-up period; 96 of 299 patients (32.1%) experienced the event. No missing values are present. All continuous features are standardized (zero mean, unit variance) prior to model fitting; standardization parameters are refit on the training fold within every cross-validation split to avoid information leakage from the test fold.

B. Classifiers

Eight classifier families were selected to overlap with the majority of methods evaluated in the original study while remaining within widely available scikit-learn [22] implementations: logistic regression, decision tree, random forest, gradient boosting, Gaussian naive Bayes, linear-kernel support vector machine, radial-basis-function support vector machine, and k-nearest neighbors.

C. Identical Hyperparameter Tuning Budget

To remove the default-hyperparameter confound identified in Section I, every classifier was tuned with an identical Optuna [21] search of 30 trials, each trial evaluated by five-fold stratified cross-validation on the full dataset with MCC as the optimization objective. The search space for each classifier was defined once and used without modification; no classifier received more tuning trials, a wider search range, or a different inner validation scheme than any other.

D. Paired Repeated Stratified Cross-Validation

Final evaluation used repeated stratified 10-fold cross-validation with 10 repeats (100 total train-test splits). Critically, and unlike the original replication package, a single shared sequence of splits (fixed random seed) was generated once and applied identically to all eight classifiers, so that fold k of repeat r contains the same patients for every classifier. This pairing is what makes the subsequent Wilcoxon signed-rank comparisons valid, since each of the 100 paired observations reflects the same train-test partition evaluated under two different models.

E. Statistical Testing

A Friedman test [6] was first applied to the 8 x 100 matrix of per-fold MCC scores to test the omnibus null hypothesis that all eight classifiers perform equally. Given a significant Friedman result, a Nemenyi post-hoc test [7] was used to identify which pairs of classifiers differ significantly while controlling the family-wise comparison structure, following the procedure recommended by Demsar [5]. In addition, pairwise Wilcoxon signed-rank tests [8] were computed for every classifier pair on the paired per-fold MCC values, reporting both the p-value and the direction of the difference (which classifier had the higher mean), rather than only flagging significance without direction.

F. Efficiency Criterion

Wall-clock training and inference time were recorded for every fold and every classifier on identical hardware within the same process, and reported as a secondary efficiency criterion alongside predictive performance, since the original study did not report computational cost at all.

IV. RESULTS

Table I reports, for each of the eight classifiers, the mean and standard deviation of MCC, F1 score, accuracy and AUC across the 100 paired cross-validation folds, together with mean training and inference time per fold. Table I, presented below on its own full-width section, summarizes the complete set of predictive-performance and efficiency results referenced throughout the remainder of this section.

TABLE I

PREDICTIVE PERFORMANCE AND COMPUTATIONAL COST OF EIGHT CLASSIFIERS UNDER IDENTICAL PAIRED REPEATED 10x10-FOLD CROSS-VALIDATION (MEAN +/- STANDARD DEVIATION OVER 100 FOLDS)

Classifier	MCC	F1 Score	Accuracy	AUC	Train (ms)	Infer (ms)
Random Forest	0.657 +/- 0.150	0.747 +/- 0.111	0.853 +/- 0.063	0.919 +/- 0.049	126.79	6.22
Decision Tree	0.602 +/- 0.157	0.695 +/- 0.131	0.829 +/- 0.065	0.843 +/- 0.077	1.18	0.17
Logistic Regression	0.598 +/- 0.159	0.701 +/- 0.135	0.828 +/- 0.064	0.875 +/- 0.062	1.92	0.18
SVM Linear	0.586 +/- 0.145	0.697 +/- 0.125	0.822 +/- 0.058	0.870 +/- 0.065	6.17	0.31
SVM RBF	0.578 +/- 0.152	0.683 +/- 0.130	0.821 +/- 0.060	0.879 +/- 0.062	7.54	0.40
Gradient Boosting	0.562 +/- 0.167	0.676 +/- 0.130	0.813 +/- 0.067	0.891 +/- 0.065	101.11	0.49
Naive Bayes	0.414 +/- 0.199	0.532 +/- 0.175	0.763 +/- 0.067	0.848 +/- 0.068	0.87	0.19
K-Nearest Neighbors	0.339 +/- 0.173	0.390 +/- 0.165	0.741 +/- 0.053	0.803 +/- 0.074	0.72	1.70

As shown in Table I above, random forest attains the highest mean MCC (0.657), followed closely by decision tree (0.602) and logistic regression (0.598), while k-nearest neighbors is the weakest classifier by a wide margin (MCC = 0.339). The same ordering holds for F1 score and accuracy; AUC ranking differs slightly, with random forest still leading (0.919) but gradient boosting moving into second place (0.891) ahead of the tree- and margin-based alternatives, indicating that gradient boosting produces better-calibrated class-probability ranking than its point-prediction accuracy alone would suggest.

A. Omnibus and Post-Hoc Tests

The Friedman test on the 8 x 100 paired MCC matrix rejects the null hypothesis of equal performance overwhelmingly (chi-squared(7) = 264.06, p = 2.81 x 10-53), confirming that the eight classifiers are not interchangeable on this task. The Nemenyi post-hoc comparison shows that random forest differs significantly ($p < 0.05$) from decision tree ($p = 0.0041$), gradient boosting ($p < 0.001$), naive Bayes ($p < 0.001$), linear SVM ($p = 0.0257$), RBF SVM ($p = 0.0013$) and k-nearest neighbors ($p < 0.001$), but does not differ significantly from logistic regression ($p = 0.236$). This last result is a genuine refinement over the original single-split ranking: random forest's numerical lead over logistic regression, while consistent in direction, is not large enough relative to the fold-to-fold variance to be declared a statistically confident win.

Pairwise Wilcoxon signed-rank tests corroborate the Nemenyi pattern: 21 of the 28 classifier pairs differ significantly at the uncorrected $p < 0.05$ threshold, and 19 of 28 remain significant after a conservative Bonferroni correction (alpha = 0.05 / 28 = 0.0018). Naive Bayes and k-nearest neighbors are not significantly different from each other ($p = 0.758$) and form a visibly weaker cluster in Fig. 1, while random forest, decision tree, logistic regression and the two SVM variants form a stronger, more tightly overlapping cluster whose members are frequently not distinguishable from one another once decision tree, logistic regression, linear SVM, RBF SVM and gradient boosting are compared pairwise.

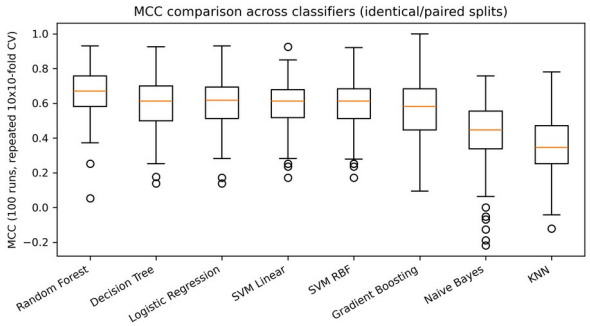

Fig. 1. Distribution of MCC across 100 paired cross-validation folds for each classifier.

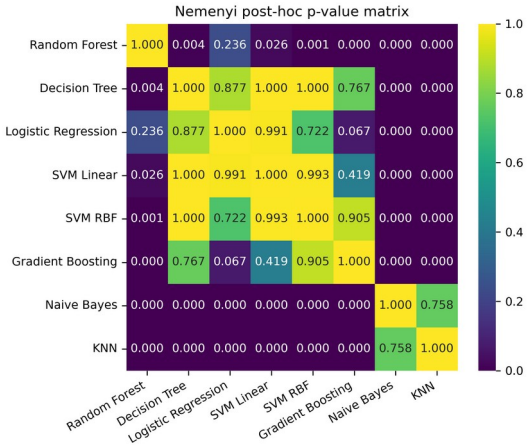

Fig. 2. Nemenyi post-hoc pairwise p-value matrix; darker cells indicate stronger evidence of a performance difference.

B. Efficiency

Mean training time per fold ranged from 0.72 ms (k-nearest neighbors, which performs no explicit training) to 126.79 ms (random forest). Decision tree and logistic regression, whose MCC is statistically indistinguishable from or only modestly below random forest, train in 1.18 ms and 1.92 ms respectively, roughly two orders of magnitude faster. Fig. 3

places this trade-off in context: for deployment settings where training or inference latency is constrained, logistic regression or decision tree offer most of random forest's predictive performance at a small fraction of its computational cost, a consideration entirely absent from the original study's default-hyperparameter, untimed comparison.

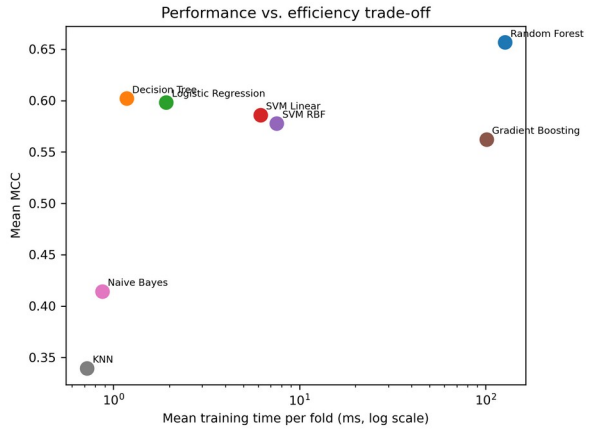

Fig. 3. Mean MCC versus mean training time per fold (log scale) for each classifier.

V. DISCUSSION

The central claim of the original study, that ensemble tree methods such as random forest are the strongest classifiers for this dataset, survives the present re-evaluation: random forest retains the highest point estimate of MCC, F1, accuracy and AUC even after tuning every competing classifier with an identical budget and evaluating all of them on identical paired folds. In that sense, this replication is confirmatory rather than contradictory.

However, the added statistical rigor changes what can be claimed with confidence. Under the original unpaired, untuned, single-execution-set protocol, random forest's numerical lead over every other classifier, including logistic regression, would have been reported as the winning model without qualification. The paired repeated cross-validation and Nemenyi post-hoc procedure used here show that the lead over logistic regression specifically is not statistically significant, so a fair statement of the result is that random forest is among the best classifiers for this task together with logistic regression, decision tree, and both SVM variants, rather than uniquely best. This is precisely the kind of overstatement that repeated, paired, significance-tested evaluation is designed to catch, and it is consistent with the broader methodological literature on classifier comparison [5], [9] that motivated this study's design.

Two limitations should be noted. First, the dataset contains only 299 patients from a single cohort collected in 2015 [1], [2], so even a well-powered paired design has limited statistical power to detect small performance differences, which is itself part of why several classifier pairs in this study are not significantly distinguishable. Second, this study retained the original feature set as released, including the follow-up TIME variable; several later commentators have noted that TIME is not available at the moment a real-time prediction would be needed and may inflate apparent predictive performance, a concern that is orthogonal to the paired-evaluation and hyperparameter-fairness contributions of this paper and is left to future work.

VI. CONCLUSION

This study identified two concrete, verifiable methodological gaps in the widely used replication package accompanying Chicco and Jurman's heart failure survival prediction study: unpaired random resampling across classifiers, and the absence of any hyperparameter tuning or formal significance testing. Correcting both gaps while changing nothing else about the task, an identical Optuna tuning budget and a paired repeated 10x10-fold stratified cross-validation protocol with Friedman, Nemenyi and Wilcoxon signed-rank tests, confirms that random forest remains the strongest single classifier by point estimate, but reveals that its advantage over logistic regression is not statistically significant, and that decision tree and logistic regression offer a substantially more favorable performance-to-training-time trade-off than random forest or gradient boosting. These findings illustrate that fair, paired, statistically tested evaluation can both confirm and meaningfully qualify conclusions drawn from less rigorous single-split comparisons, and the full paired evaluation protocol used here is directly reusable for other small clinical tabular datasets facing the same methodological gaps.

REFERENCES

[1] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, vol. 20, no. 16, 2020.

[2] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," PLOS ONE, vol. 12, no. 7, e0181001, 2017.

[3] World Health Organization, "Cardiovascular diseases (CVDs) fact sheet," WHO, Geneva, Switzerland, 2021.

[4] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 6, 2020.

[5] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.

[6] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," Journal of the American Statistical Association, vol. 32, no. 200, pp. 675-701, 1937.

[7] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton University, Princeton, NJ, USA, 1963.

[8] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80-83, 1945.

[9] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895-1923, 1998.

[10] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI), 1995, pp. 1137-1143.

[11] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, no. 91, 2006.

[12] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," PLOS ONE, vol. 14, no. 11, e0224365, 2019.

[13] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[14] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273-297, 1995.

[15] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[16] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[17] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Boca Raton, FL, USA: CRC Press, 1984.

[18] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, pp. 81-106, 1986.

[19] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B, vol. 20, no. 2, pp. 215-242, 1958.

[20] I. Rish, "An empirical study of the naive Bayes classifier," in Proc. IJCAI Workshop on Empirical Methods in AI, 2001, pp. 41-46.

[21] 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. on Knowledge Discovery and Data Mining, 2019, pp. 2623-2631.

[22] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[23] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012.

[24] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 2951-2959.

[25] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145-1159, 1997.

[26] N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective. Cambridge, UK: Cambridge University Press, 2011.

[27] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," arXiv:1811.12808, 2018.

[28] D. Chicco, "Ten quick tips for machine learning in computational biology," BioData Mining, vol. 10, no. 35, 2017.