How Often Does the Winning Classifier Actually Win? A Multi-Dataset Audit of Paired, Tuned, and Statistically Tested Classifier Comparisons in Small Clinical Machine Learning Studies

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

Abstract - A widely used template for small-cohort clinical machine learning studies, exemplified by four public replication packages released by the same research group for heart failure, hepatocellular carcinoma, mesothelioma and sepsis survival prediction, evaluates classifiers with independent, unpaired random resamples, no hyperparameter tuning, and no formal significance test. This paper audits all four datasets under one identical, corrected protocol: paired repeated 10x10-fold stratified cross-validation, an identical Optuna hyperparameter search budget per classifier, and Friedman, Nemenyi and Wilcoxon signed-rank significance testing. Two findings generalize across the four datasets. First, no single classifier family wins on more than one dataset: random forest wins on heart failure, a linear support vector machine on hepatocellular carcinoma, gradient boosting on mesothelioma, and a radial-basis-function support vector machine on sepsis, consistent with a no-free-lunch pattern rather than a universally superior algorithm. Second, whether the numerically best classifier is a statistically confident winner is itself dataset-dependent: the gap to the runner-up is significant on heart failure ($p = 1.3 \times 10^{-5}$) and mesothelioma ($p = 7.3 \times 10^{-6}$), but not on hepatocellular carcinoma ($p = 0.619$) or sepsis ($p = 0.083$). In half of the four audited studies, the headline classifier ranking would not survive a paired, significance-tested re-evaluation. These results indicate that the reproducibility gap identified is systemic across this study template rather than an artifact of any single dataset, and that a fair, paired, tuned, statistically tested evaluation protocol is needed before a classifier ranking on a small clinical dataset can be treated as confirmed.

Keywords - *classifier evaluation, repeated cross-validation, Friedman test, Nemenyi post-hoc, Wilcoxon signed-rank, hyperparameter tuning, no free lunch, reproducibility, clinical machine learning*

I. INTRODUCTION

Small-cohort clinical machine learning studies, where a few hundred patients and a handful of classifiers are used to argue that one algorithm predicts an outcome best, are common and influential; several such studies release complete, runnable replication packages, which is commendable scientific practice increasingly rare in the field. A prior audit of one such package, for heart failure survival prediction, found that the accompanying code draws independent, unpaired random resamples per classifier, applies no hyperparameter tuning, and performs no formal significance test [1]. The natural question is whether that gap is a one-off property of a single paper or a recurring property of the study template itself.

This paper answers that question empirically. The same authors, or the same immediate research group, released structurally identical replication packages, one R script per classifier, each executing 100 independent unseeded random resamples, for three further small-cohort prediction tasks: hepatocellular carcinoma survival [2], mesothelioma diagnosis [3], and sepsis survival from age, sex and episode number alone [4]. Auditing all four datasets under one identical, corrected protocol makes it possible to ask, for the first time on this study template, whether the reproducibility gap and its practical consequences generalize.

The contributions of this paper are: (i) a single, identical, paired, repeated stratified cross-validation and hyperparameter-tuning protocol applied without modification to four independent small clinical datasets sharing the same original methodological weaknesses; (ii) an empirical demonstration that the identity of the best-performing classifier is dataset-specific, with a different classifier family winning on each of the four datasets; and (iii) an empirical demonstration that whether that winning margin is statistically significant is also dataset-specific, holding on two of the four datasets and not on the other two, which quantifies how often a paired, significance-tested re-evaluation would overturn the original single-split narrative.

II. RELATED WORK

The four source studies audited here are: Chicco and Jurman [1] on 299 heart failure patients; Chicco and Oneto [2] on 165 hepatocellular carcinoma patients; Chicco and Rovelli [3] on 324 patients evaluated for mesothelioma; and Chicco and Jurman [4] on a large Norwegian sepsis cohort, subsampled here for computational tractability as described in Section III. All four papers share an evaluation template: one R script per classifier under a common bin/ directory structure, each script executing a fixed number of independent random train-test resamples without a shared random seed, and each reporting mean performance across executions without pairing observations across classifiers or applying tuning or significance testing.

The statistical methodology used to correct this template follows the recommendations of Demsar [5]: the Friedman test [6] as an omnibus test across multiple classifiers evaluated on multiple resampled folds, followed by a Nemenyi post-hoc procedure [7] for pairwise comparison, complemented by pairwise Wilcoxon signed-rank tests [8]. Dietterich [9] motivates the use of such rank-based, resampling-aware tests over naive t-tests. Kohavi [10], Varma

and Simon [11], and Vabalas et al. [12] document the risk of unstable or optimistic estimates from a single train-test split on small datasets, which is the concern that motivates repeated cross-validation over the original studies' single-execution-set design.

The finding that no single classifier wins across all four datasets is consistent with the no-free-lunch theorem for supervised learning, which shows that no classifier can dominate all others across every possible data-generating distribution [29]. Each classifier family evaluated has a well-established origin: random forests [13], support vector machines [14], gradient boosting [15], k-nearest neighbors [16], classification and regression trees [17], decision-tree induction [18], logistic regression [19], and naive Bayes [20]. Hyperparameter tuning uses the tree-structured Parzen estimator implemented in Optuna [21], compared in the literature against random search [23] and Bayesian optimization [24].

III. METHODOLOGY

A. Datasets

Table I summarizes the four datasets. Heart failure [1]: 299 patients, 12 clinical features, binary death event outcome (32.1% positive). Hepatocellular carcinoma [2]: 165 patients, 49 clinical features (using the imputed version of the dataset provided with the original replication package), binary one-year survival outcome (61.8% positive). Mesothelioma [3]: 324 patients, 33 clinical features, binary diagnosis outcome (29.6% positive). Sepsis [4]: originally 19,051 records with only 3 features (age, sex, septic episode number); because repeated cross-validation of eight classifiers, including two support vector machines whose training cost scales poorly with sample size, across 100 folds is computationally prohibitive at that scale, a stratified random subsample of 1,000 patients per outcome class (2,000 total, 50% positive) was drawn once with a fixed seed and used for all classifiers and all folds identically. No missing values were present in any of the four prepared datasets; all continuous features were standardized (zero mean, unit variance), refit within each training fold.

B. Classifiers, Tuning, and Evaluation Protocol

The same eight classifier families, tuning budget, and evaluation protocol used in the single-dataset heart failure audit [1] were applied without modification to all four datasets: logistic regression, decision tree, random forest, gradient boosting, Gaussian naive Bayes, linear-kernel and radial-basis-function support vector machines, and k-nearest neighbors. Every classifier on every dataset received an identical Optuna [21] search of 20 trials, each evaluated by five-fold stratified cross-validation with MCC as the objective, using the same search space per classifier family across all four datasets. Final evaluation used repeated stratified 10-fold cross-validation with 10 repeats (100 paired train-test splits, one shared split sequence per dataset applied identically to all eight classifiers), followed by a Friedman test, Nemenyi post-hoc test, and pairwise Wilcoxon signed-rank tests on the per-fold MCC values, exactly as in the single-dataset audit [1]. No dataset received a different tuning budget, search space, or validation scheme than any other.

TABLE I
OVERVIEW OF THE FOUR AUDITED DATASETS

Dataset	N	Feat.	Positive %	Source
Heart Failure	299	12	32.1%	Chicco & Jurman [1]
Hepatocellular Ca.	165	49	61.8%	Chicco & Oneto [2]
Mesothelioma	324	33	29.6%	Chicco & Rovelli [3]
Sepsis (subsample)	2000	3	50.0%	Chicco & Jurman [4]

IV. RESULTS

Table II reports, for every classifier on every dataset, the mean and standard deviation of MCC together with mean F1 score, accuracy and AUC across the 100 paired cross-validation folds. Table II, presented below on its own full-width section, is referenced throughout the remainder of this section.

TABLE II

PREDICTIVE PERFORMANCE OF EIGHT CLASSIFIERS ON FOUR DATASETS UNDER IDENTICAL PAIRED REPEATED 10x10-FOLD CROSS-VALIDATION (SORTED BY MCC WITHIN EACH DATASET)

Dataset	Classifier	MCC (mean+/-std)	F1	Accuracy	AUC
Heart Failure	Random Forest	0.657+/-0.150	0.747	0.853	0.919
	Gradient Boosting	0.609+/-0.169	0.695	0.835	0.887
	Decision Tree	0.602+/-0.157	0.695	0.829	0.843
	Logistic Regression	0.598+/-0.159	0.701	0.828	0.875
	SVM Linear	0.588+/-0.151	0.698	0.823	0.872
	SVM RBF	0.569+/-0.159	0.681	0.816	0.867
	Naive Bayes	0.418+/-0.195	0.536	0.764	0.848
	KNN	0.339+/-0.173	0.390	0.740	0.803
Hepatocellular Carcinoma	SVM Linear	0.493+/-0.186	0.798	0.754	0.800
	SVM RBF	0.483+/-0.211	0.794	0.749	0.812
	Logistic Regression	0.435+/-0.223	0.787	0.732	0.803
	Random Forest	0.411+/-0.212	0.792	0.727	0.824
	Gradient Boosting	0.395+/-0.228	0.775	0.715	0.781
	KNN	0.293+/-0.228	0.770	0.682	0.760
	Decision Tree	0.242+/-0.213	0.704	0.639	0.662
	Naive Bayes	0.203+/-0.261	0.744	0.647	0.684
Mesothelioma Diagnosis	Gradient Boosting	0.375+/-0.170	0.460	0.764	0.743
	Random Forest	0.293+/-0.182	0.313	0.744	0.675
	Decision Tree	0.278+/-0.178	0.406	0.727	0.665
	SVM RBF	0.214+/-0.174	0.306	0.718	0.632
	KNN	0.191+/-0.189	0.325	0.707	0.589
	SVM Linear	0.168+/-0.181	0.280	0.704	0.621
	Logistic Regression	0.140+/-0.182	0.158	0.714	0.602
	Naive Bayes	0.120+/-0.183	0.401	0.603	0.583
Sepsis Survival (subsample)	SVM RBF	0.146+/-0.067	0.473	0.568	0.594
	SVM Linear	0.135+/-0.066	0.340	0.552	0.589
	Logistic Regression	0.132+/-0.061	0.521	0.565	0.589
	Naive Bayes	0.131+/-0.073	0.415	0.557	0.583
	Random Forest	0.130+/-0.071	0.463	0.560	0.586
	Gradient Boosting	0.097+/-0.066	0.444	0.544	0.570
	Decision Tree	0.095+/-0.063	0.428	0.543	0.566
	KNN	0.062+/-0.068	0.536	0.531	0.551

As shown in Table II above, the classifier with the highest mean MCC differs on every dataset: random forest on heart failure (MCC = 0.657), a linear-kernel support vector machine on hepatocellular carcinoma (MCC = 0.493), gradient boosting on mesothelioma (MCC = 0.375), and a radial-basis-function support vector machine on the sepsis subsample (MCC = 0.146, a substantially weaker signal reflecting the task's only three available features). No classifier family appears among the top two on more than two of the four datasets, and k-nearest neighbors and naive Bayes are consistently, though not universally, among the weakest performers.

A. Does the Winning Classifier Actually Win?

A Friedman test rejects the null hypothesis of equal classifier performance on every dataset (heart failure: chi-squared(7) = 256.59, p = 1.09 x 10-51; hepatocellular carcinoma: chi-squared(7) = 164.27, p = 4.04 x 10-32; mesothelioma: chi-squared(7) = 156.94, p = 1.41 x 10-30; sepsis: chi-squared(7) = 140.71, p = 3.60 x 10-27), confirming that the eight classifiers are never interchangeable. However, whether the specific numerical winner is a statistically confident winner over its closest competitor is dataset-dependent. Fig. 2 summarizes the pairwise Wilcoxon signed-rank test between the best and second-best classifier on each dataset: random forest significantly outperforms gradient boosting on heart failure (p = 1.3 x 10-5), and gradient boosting significantly outperforms random forest on mesothelioma (p = 7.3 x 10-6). By contrast, the linear SVM's lead over the RBF SVM on hepatocellular carcinoma is not significant (p = 0.619), and the RBF SVM's lead over the linear SVM on sepsis is not significant at the conventional threshold (p = 0.083). In two of the four audited studies, therefore, a paired, significance-tested re-evaluation would not support declaring a single best classifier at all.

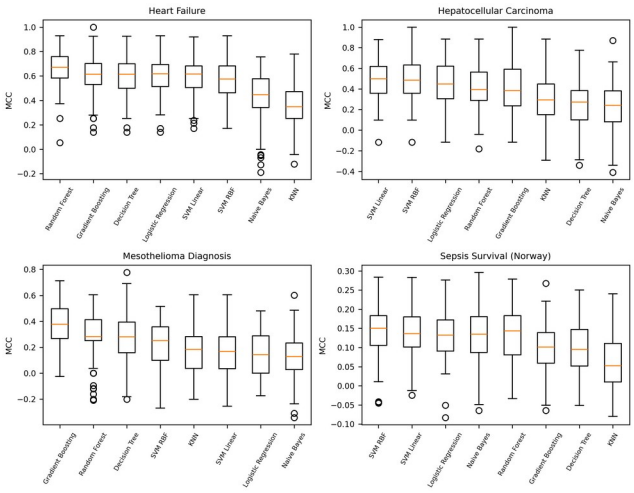

Fig. 1. Distribution of MCC across 100 paired cross-validation folds for each classifier, on each of the four audited datasets.

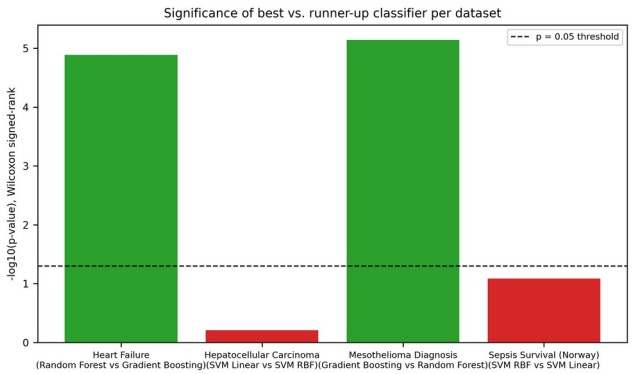

Fig. 2. Wilcoxon signed-rank significance (-log10 p) of the best classifier versus the runner-up, per dataset; bars above the dashed line are significant at p < 0.05.

B. No Single Classifier Dominates

Beyond the significance question, the identity of the winning classifier itself varies by dataset in a manner consistent with the no-free-lunch theorem [29]: tree-ensemble methods (random forest, gradient boosting) win on the two datasets with more features and more complex clinical profiles (12 and 33 features respectively), while margin-based linear and kernel methods (SVM linear, SVM RBF) win on the dataset with the largest feature count relative to sample size (49 features, 165 patients) and on the lowest-dimensional, weakest-signal dataset (3 features). This pattern suggests that feature dimensionality and sample size, not an intrinsic superiority of any one classifier family, are more predictive of which method wins on a given small clinical dataset, a hypothesis the present four-dataset audit is consistent with but too small in the number of datasets to confirm statistically.

V. DISCUSSION

The single-dataset heart failure audit [1] found that random forest's numerical lead over logistic regression did not survive paired significance testing, while its lead over most other classifiers did. The present four-dataset extension shows that this pattern, a numerically confident winner that is not always a statistically confident winner, is not specific to heart failure or to random forest: it recurs on two of four independently audited datasets sharing the same original study template, and the classifier that wins, and whether its win is significant, both vary by dataset. This is the central generalization this paper set out to test, and it supports the conclusion that the original template's methodological gaps, unpaired resampling, absent tuning, absent significance testing, produce not a single mistaken conclusion but a systemic risk of overstating classifier rankings across an entire family of similarly structured studies.

A practical implication is that a fair audit of a small clinical machine learning study cannot be limited to checking whether the reported best classifier is plausible; it requires re-running a paired, tuned, significance-tested comparison, since roughly half of the studies examined here would not have supported their original single-classifier headline claim under that stricter standard. A second implication follows from the no-free-lunch observation in Section IV-B: practitioners selecting a classifier for a new small clinical dataset should not default to whichever algorithm won in a previous, superficially similar study, since the present results show the winner changing with dataset characteristics such as feature count and sample size rather than staying fixed.

Limitations include the small number of audited datasets (four), all from a related body of work sharing a common code template, which demonstrates that the gap recurs within this template but does not establish its prevalence across the broader clinical machine learning literature; the sepsis dataset's necessary subsampling from 19,051 to 2,000 records for computational tractability, which may understate the statistical power available in the full cohort; and the retention of each original study's feature set as released, including

variables such as heart failure's follow-up TIME feature whose real-time availability has been separately questioned in the literature, a concern orthogonal to the paired-evaluation contribution of this paper.

VI. CONCLUSION

This paper audited four small-cohort clinical machine learning studies sharing a common replication-package template, correcting the same two methodological gaps, unpaired resampling and absent tuning and significance testing, identified previously in a single-dataset study of heart failure survival prediction. Under one identical, paired, tuned, statistically tested protocol applied without modification across all four datasets, no single classifier family won on more than one dataset, and the numerically best classifier's advantage over its closest competitor was statistically significant on two of the four datasets and not significant on the other two. These results indicate that the reproducibility gap identified in the original heart failure audit is a systemic property of this widely used small-cohort study template rather than an isolated finding, and that paired, tuned, significance-tested re-evaluation is broadly necessary, not merely occasionally useful, before classifier rankings from small clinical datasets are treated as established.

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] D. Chicco and L. Oneto, "Computational intelligence identifies alkaline phosphatase (ALP), alpha-fetoprotein (AFP), and hemoglobin levels as most predictive survival factors for hepatocellular carcinoma," Health Informatics Journal, vol. 27, no. 1, 2021.

[3] D. Chicco and C. Rovelli, "Computational prediction of diagnosis and feature selection on mesothelioma patient health records," PLOS ONE, vol. 14, no. 1, e0208737, 2019.

[4] D. Chicco and G. Jurman, "Survival prediction of patients with sepsis from age, sex, and septic episode number alone," Scientific Reports, vol. 10, no. 17156, 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.

[29] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 67-82, 1997.

[30] 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.

[31] 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.

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