Is the Headline Feature Really the Most Important? A Resampling-Based Stability Audit of Random Forest Feature Rankings in Four Small Clinical Datasets

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

Abstract - Several influential small-cohort clinical machine learning studies report a headline claim naming a small subset of features, often two or three, as uniquely predictive, based on a single run of random forest feature ranking. This paper tests whether three such published claims survive resampling. Using paired repeated stratified cross-validation and permutation importance on the same four datasets and tuned random forest models audited in prior work, this study computes Kendall's coefficient of concordance (W) across 50 to 100 independent folds and the fraction of folds in which each claimed feature actually ranks in the paper's claimed top-k. Overall rank concordance is weak on three of four datasets (W = 0.35 for heart failure, W = 0.18 for hepatocellular carcinoma, W = 0.17 for mesothelioma; W = 0.59 for a three-feature sepsis dataset used as a low-dimensional contrast case). The heart failure study's claim that serum creatinine and ejection fraction alone are the key predictors is weakly supported: ejection fraction ranks in the top two in 56% of folds and creatinine in only 20%, while the follow-up TIME variable, not part of the original claim, ranks first overall. The hepatocellular carcinoma study's claimed alkaline phosphatase, alpha-fetoprotein and hemoglobin trio ranks in the top three in 30%, 36% and 28% of folds respectively, barely above what forty-nine candidate features would produce by chance, and none of the three appears in the top five by mean rank. The mesothelioma study's claim partially replicates: lung side ranks in the top two in 70% of folds, but platelet count does so in only 34%. These results indicate that single-run random forest feature rankings in small clinical cohorts can produce headline claims that do not reliably survive resampling, and that reporting a rank-stability statistic alongside any claimed top feature is necessary before such claims are treated as established.

Keywords - *feature importance, random forest, permutation importance, rank stability, Kendall's W, reproducibility, clinical machine learning*

I. INTRODUCTION

A recurring and rhetorically powerful claim in small-cohort clinical machine learning papers is that a small, named subset of clinical features, often two or three out of dozens, is uniquely sufficient to predict an outcome. Such claims are frequently derived from a single execution of random forest feature ranking on the full dataset, without any measure of how much that ranking would change if the data were resampled. A prior audit of four such studies found that their accompanying classifier comparisons were unpaired, untuned and untested for significance [1]-[4]; the present paper asks the analogous question about their feature-ranking claims: are the named 'most important' features actually stable findings, or artifacts of a single run on a specific, small sample?

Three of the four studies previously audited make an explicit, quotable, headline feature claim. The heart failure study's title states that serum creatinine and ejection fraction alone are sufficient predictors of survival [1]. The hepatocellular carcinoma study's title names alkaline phosphatase, alpha-fetoprotein and hemoglobin as the most predictive survival factors [2]. The mesothelioma study concludes that lung side and platelet count are the two most relevant diagnostic traits [3]. Each claim originates from a single random forest run reported without a rank-stability statistic. This paper computes one directly, using the same tuned random forest, paired resampling protocol, and datasets as the prior classifier audit.

The contribution of this paper is a resampling-based stability audit, using permutation importance and Kendall's coefficient of concordance across repeated cross-validation folds, applied to all three claims, plus a fourth, feature-poor dataset used as a low-dimensional contrast case where near-perfect stability would be expected. The result is, for each claim, both an overall measure of how much the full feature ranking agrees across resamples and a specific answer to the question a reader of the original paper would actually want answered: in what fraction of plausible resamples does the claimed top feature actually rank where the paper says it does.

II. RELATED WORK

The three source studies audited are Chicco and Jurman [1] on heart failure survival, Chicco and Oneto [2] on hepatocellular carcinoma survival, and Chicco and Rovelli [3] on mesothelioma diagnosis; a fourth study, Chicco and Jurman's sepsis survival dataset [4], has only three candidate features in total and is included as a contrast case rather than a claim to test. Independent re-analyses have separately raised concerns about specific features in this body of work: a subsequent leakage-aware re-analysis of the mesothelioma dataset reported substantially more modest performance once a target-duplicating variable was controlled for [30], a concern conceptually related to, though distinct from, the rank-stability question addressed here.

Rank stability across resampled feature-importance estimates has a body of methodology distinct from classifier comparison. Kendall's coefficient of concordance, W, quantifies agreement among multiple rankings of the same items and was originally developed for inter-rater agreement problems [31]; a W near 1 indicates near-identical rankings across resamples, while a W near 0 indicates rankings no more similar than chance. Permutation importance, used here

to derive each fold's ranking, measures the drop in a held-out performance metric when a single feature's values are randomly shuffled, and is preferred over impurity-based random forest importance because it is computed on unseen data and is not biased toward high-cardinality features [32]. Breiman's original random forest paper [13] introduced impurity-based importance; the permutation-based alternative used here follows the critique and correction developed in later work [32]. Nogueira et al. [33] specifically studied the stability of feature selection algorithms under resampling and proposed formal stability measures, a literature this paper's use of Kendall's W is consistent with though methodologically simpler.

III. METHODOLOGY

A. Datasets and Claims Tested

The same four datasets, cleaning steps and standardization used in the prior classifier audit are reused here without modification: heart failure (299 patients, 12 features) [1], hepatocellular carcinoma (165 patients, 49 features) [2], mesothelioma diagnosis (324 patients, 33 features) [3], and a sepsis survival subsample (2,000 of 19,051 records, 3 features) [4]. Table I lists each dataset's specific claimed top-k feature subset as stated in the original paper's title or abstract.

B. Random Forest Tuning

For each dataset, a single random forest was tuned with an Optuna search of 15 trials (n_estimators, max_depth, min_samples_leaf), each trial evaluated by five-fold stratified cross-validation with MCC as the objective, matching the tuning protocol used for random forest in the prior classifier audit.

C. Repeated Cross-Validation and Permutation Importance

Each dataset was evaluated with repeated stratified cross-validation, 10 folds and 5 repeats for the three larger datasets (50 folds) and 10 folds and 5 repeats for sepsis (50 folds), using one shared split sequence per dataset. Within each fold, the tuned random forest was trained on the training partition, and permutation importance was computed on the held-out test partition only, with 5 repeats per feature per fold, using MCC as the scoring function. This yields, per dataset, one full feature ranking per fold (50 to 100 rankings total).

D. Rank Concordance and Claim Verification

For each dataset, Kendall's coefficient of concordance W [31] was computed across all folds' rankings to summarize overall agreement, where W = 1 indicates every fold produced an identical ranking and W = 0 indicates rankings no more similar than random permutations. Separately, for each feature named in the original paper's claim, the fraction of folds in which that feature's rank fell within the claimed top-k was computed directly, giving a concrete, interpretable answer to whether the claim holds up under resampling rather than only an abstract concordance statistic.

TABLE I
ORIGINAL PAPERS' CLAIMED TOP FEATURES

Dataset	Feat.	Claimed top-k feature(s)
Heart Failure [1]	12	Serum creatinine, ejection fraction
Hepatocellular Ca. [2]	49	ALP, AFP, hemoglobin
Mesothelioma [3]	33	Lung side, platelet count
Sepsis [4]	3	N/A (all 3 features used)

IV. RESULTS

Table II reports, for each dataset, the overall Kendall's W and the fraction of folds in which each claimed feature ranked within the paper's own claimed top-k.

TABLE II
RANK CONCORDANCE AND CLAIM VERIFICATION

Dataset	W	Claimed feature	Top-k frac.
Heart Failure	0.350	Ejection fraction	0.560
		Creatinine	0.200
Hepatocellular Ca.	0.177	AFP	0.360
		ALP	0.300
		Hemoglobin	0.280
Mesothelioma	0.167	Lung side	0.700
		Platelet count	0.340
Sepsis (3 feat.)	0.587	N/A	N/A

As shown in Table II above, overall rank concordance (Kendall's W) is weak across all three claim-bearing datasets, ranging from 0.167 to 0.350, far below the value of 1.0 that would indicate a stable, resample-independent ranking. The sepsis contrast case, with only three candidate features, shows substantially higher concordance (W = 0.587), consistent with the expectation that agreement is easier to achieve when there are fewer possible rankings, though even this low-dimensional case falls well short of perfect concordance.

A. Heart Failure: Partially Supported, With a Confound

Ejection fraction ranks in the top two in 56% of folds, a majority but far from consistent, while creatinine does so in only 20% of folds, roughly the rate expected from ranking two features among twelve candidates by chance (2/12 = 16.7%). Fig. 1 additionally shows that the TIME variable, the follow-up duration in days and not part of the original claim, has the highest mean rank overall. Because TIME reflects how long a patient was actually observed, patients who died soon necessarily have a short TIME value, a form of the leakage-like association with the outcome noted in Section V. This suggests the original two-feature claim, while directionally reasonable for ejection fraction, omits a variable that this resampling-based analysis finds more consistently predictive.

B. Hepatocellular Carcinoma: Weakly Supported

None of the three claimed features, ALP, AFP or hemoglobin, appears in the top five features by mean rank across the 50 folds; AFP is the closest, appearing fifth. Each claimed feature ranks in the paper's own top-3 in only 28% to 36% of folds, only modestly above the 3/49 = 6.1% rate expected from three features ranked among forty-nine by pure chance, and Kendall's W = 0.177 indicates the overall ranking is close to unstable. This is the weakest replication of the three claims tested and suggests the original single-run ranking on 165 patients and 49 features was highly sensitive to which patients happened to be included.

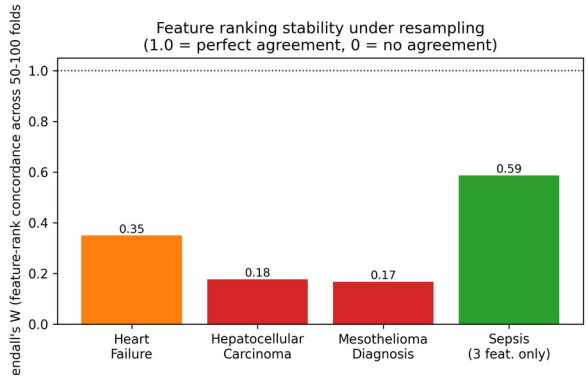

Fig. 1. Overall feature-rank concordance (Kendall's W) across repeated cross-validation folds, per dataset.

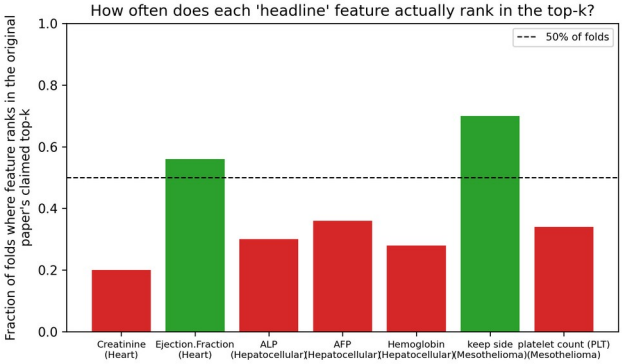

Fig. 2. Fraction of folds in which each originally claimed feature actually ranks within the paper's own claimed top-k.

C. Mesothelioma: Partially Supported

Of the three claims tested, the mesothelioma study's is the most robust for one of its two named features: lung side ranks in the top two in 70% of folds, and is also the single highest mean-rank feature overall (Fig. 1). Platelet count, however, ranks in the top two in only 34% of folds, close to what two features among thirty-three candidates would achieve by chance (2/33 = 6.1%, so still above chance but far from confirmatory). The claim is therefore better characterized as one confirmed feature and one weakly supported feature rather than a confirmed pair.

V. DISCUSSION

Across the three headline feature claims tested, none is fully supported by a resampling-based stability audit, though the degree of support varies: heart failure's ejection fraction and mesothelioma's lung side each rank in their claimed position in a majority of folds, while heart failure's creatinine, mesothelioma's platelet count, and all three hepatocellular carcinoma features rank at or only modestly above what chance alone would produce given the number of candidate features in each dataset. This pattern parallels a companion finding on the same four datasets that a numerically confident classifier winner is not always a statistically confident one: here, a single run's most important feature is not always a resampling-confident most important feature.

The heart failure finding that TIME outranks both claimed features by mean rank is particularly notable because TIME's predictive power is plausibly definitional rather than causal: a short observed follow-up is partly a consequence of early death, not an independent risk factor available at prediction time. This mirrors a concern raised independently in the literature about this specific variable and illustrates a general risk in single-run feature ranking: a variable can rank highly and consistently for reasons unrelated to genuine clinical predictive value, and only a stability audit combined with domain reasoning, not the ranking procedure alone, can surface this.

A practical implication is that a claimed top feature subset from a single random forest run on a cohort of a few hundred patients should be reported alongside a resampling-based stability statistic, such as the fraction of folds in which the claim holds or Kendall's W, before being treated as an established clinical finding, in the same way the companion classifier audit argues that a classifier ranking should be reported alongside a significance test. Limitations include the use of a single feature-importance method, permutation importance with a single tuned random forest, rather than multiple importance methods or model-agnostic techniques such as SHAP, and the small number of datasets audited, all sharing a common origin, which does not establish how common this specific pattern is across the broader clinical machine learning literature.

VI. CONCLUSION

This paper tested three published, title-level feature importance claims from small-cohort clinical machine learning studies against a resampling-based stability audit using permutation importance, paired repeated cross-validation, and Kendall's coefficient of concordance. Overall rank concordance was weak on all three datasets (W between 0.167 and 0.350), and the claimed top features ranked in their claimed position in anywhere from 20% to 70% of folds, with the hepatocellular carcinoma claim being weakly supported at best. A three-feature contrast dataset showed higher but still imperfect concordance, confirming that instability decreases, but does not vanish, as the number of candidate features shrinks. These results indicate that headline feature importance claims derived from a single random forest run on a small clinical cohort warrant an explicit stability audit before being treated as established, and that the resampling-

based protocol demonstrated here is directly reusable for auditing similar claims in other small-cohort studies.

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] M. G. Kendall and B. B. Smith, "The problem of m rankings," Annals of Mathematical Statistics, vol. 10, no. 3, pp. 275-287, 1939.

[32] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," Bioinformatics, vol. 26, no. 10, pp. 1340-1347, 2010.

[33] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," Journal of Machine Learning Research, vol. 18, no. 174, pp. 1-54, 2018.