Calibrated channel omnibus feature selection for ultra-high-dimensional biomedical data

Why no single relevance criterion is safe when the effect composition is unknown

[1]Saswat Swain [2]Chandrasekhar panda [3]Pravat Satpathy
[1]Ph.D Scholar, [2]Ph.D Scholar, [3]R & D Division
[1,2]KSOM - KIIT Deemed to be University ,[1]Tekstein Solutions, India
[1]saswat.mck@gmail.com [2]contact@chandrasekharpanda.com [3]pravat@teksteinsolutions.com

Abstract—Feature selection on ultra-high-dimensional biomedical arrays is routinely benchmarked with a single relevance criterion, and comparative studies of methods such as mRMR and SVM-RFE report accuracy without asking whether the criterion matches the way the classes actually differ. This work shows that the question is not incidental. Classes may separate through a shift in location, a shift in dispersion at equal means, or a switch-like activation restricted to a subgroup, and a criterion consistent against one of these families can be no better than chance against another. Because the composition of effects in a real assay is unknown before selection, committing to one criterion is a wager rather than a design choice. We propose CALICO-FS, a single-stage omnibus ranker that maps three complementary dependence statistics onto a common permutation-calibrated scale, estimates the reliability of each from excess tail mass, and fuses them by a reliability-weighted mean. On a controlled benchmark of 12625 features and 21 samples with planted ground truth, CALICO-FS attains a worst-case recall over four effect regimes of 0.288 against 0.198 for the strongest of six baselines, and the best mean recall of 0.443. Across ninety-six paired comparisons it is never significantly worse than any baseline in any regime and is significantly better in twenty of twenty-four. We also report two negative findings that qualify the contribution: the method is less stable under resampling than classical filters, and its recovery advantage does not transfer to downstream classification accuracy.

Index Terms—feature selection, high-dimensional data, gene expression, permutation calibration, selection stability, omnibus testing

I. INTRODUCTION

High-throughput biomedical assays produce matrices in which the number of measured features exceeds the number of available samples by two or three orders of magnitude. A microarray study profiling twelve thousand probes across twenty-one patients is not unusual, and in that setting almost every downstream analysis begins by reducing the feature set. The reduction step therefore determines what the rest of the study can possibly discover.

The literature comparing selection methods in this regime is large and, on its own terms, mature. Filter approaches based on mutual information such as mRMR [1], wrapper approaches such as SVM-RFE [2], instance-based weighting such as ReliefF [3], sparse linear models [4], and ensemble tree procedures [5] have all been evaluated repeatedly, usually by the classification accuracy obtained from the features they return. What these comparisons hold fixed, and rarely examine, is the assumption that a single scalar notion of relevance is adequate.

That assumption is doing more work than it appears to. Two classes may differ in the mean of a feature; they may differ in its variance while sharing a mean; or a feature may be activated in only part of one class, producing a bimodal distribution. All three patterns are documented in cancer transcriptomics and methylation studies, the second as differential variability [12] and the third as outlier or switch-like expression [13]. A criterion built around a difference in means is consistent against the first and has essentially no power against the second. The practical difficulty is that the mixture of effect types present in a given dataset is unknown at selection time, which is precisely when the criterion must be chosen.

This paper makes that dependence explicit and measures it. We construct a benchmark in which the effect composition is controlled and the informative features are known by construction, and we show that every single-criterion method we test collapses in at least one composition regime. An analysis of variance filter recovers 0.608 of the planted features when effects are location-dominated and 0.132 when they are dispersion-dominated, a fall of more than a factor of four driven entirely by a property of the data that no practitioner observes in advance.

We then propose a remedy that is deliberately simple. CALICO-FS computes three classical statistics sensitive to different departures, places them on a common scale by permutation calibration, weights them by a per-dataset

estimate of how much signal each carries, and averages. It contains no redundancy-elimination and no ensemble stage; both were built, evaluated and removed because the evidence did not support them, and we report those results rather than omitting them.

We formalise three families of class-conditional departure and show empirically that criterion choice dominates method choice here; we introduce a calibrated omnibus ranker with no binning or bandwidth hyperparameter; we evaluate against six baselines over four regimes and thirty paired replicates with Holm-corrected inference; and we report two findings that run against our own proposal, arguing that they fix the method's scope rather than defeat it.

II. RELATED WORK AND POSITIONING

Reviews of selection in bioinformatics [16] organise methods into filters, wrappers, and embedded procedures. This taxonomy is orthogonal to our concern: a filter and a wrapper can share the same underlying notion of relevance and therefore share the same blind spot. mRMR [1] pairs a mutual-information relevance term with a redundancy penalty. SVM-RFE [2] eliminates features by squared weight in a linear model, which ties its sensitivity to linear separability. ReliefF [3] scores by near-hit and near-miss distances and is more general, though its behaviour under pure dispersion shifts has not to our knowledge been characterised. Elastic-net logistic regression [4] inherits the linear model's assumptions. Shadow-feature procedures in the style of Boruta [5] test importances against permuted copies, an idea we adopt for calibration but apply per statistic rather than per model.

A second line of work treats stability as a first-class property. Kuncheva [7] proposed a chance-corrected consistency index, and Nogueira, Sechidis and Brown [6] derived an estimator satisfying a set of desirable axioms, which we adopt. Haury, Gestraud and Vert [17] observed that accuracy and stability rank methods differently on molecular signatures. Our stability results extend that observation in an uncomfortable direction: we find configurations in which a method is highly stable and substantially wrong, which suggests stability should not be read as a quality measure on its own.

The calibration machinery draws on large-scale inference. Efron's empirical null [9] motivates estimating a statistic's null distribution from the data rather than assuming it, and higher criticism [8] uses excess counts in the tail of a test statistic to detect sparse signal. We use an excess-tail count not to test a global hypothesis but to weight channels against one another.

Relative to the review motivating this study, which compared mRMR and SVM-RFE on B-cell chronic lymphocytic leukaemia arrays by classification accuracy alone, we hold that such comparisons are underdetermined: without knowing the effect composition, an accuracy ranking describes the dataset at least as much as the methods.

Table 1 Positioning against representative selection methods. Consistency is with respect to the three effect families of Section III.

Method	Loc.	Disp.	Switch	Calib.	Hyper-par.
mRMR [1]	yes	weak	partial	no	bins, k
SVM-RFE [2]	yes	no	partial	no	C, step
ReliefF [3]	yes	weak	partial	no	neighbours
Elastic-net [4]	yes	no	partial	no	alpha, lambda
RF-shadow [5]	yes	weak	yes	yes	trees, iter.
ANOVA-F	yes	no	partial	no	none
CALICO-FS	yes	yes	yes	yes	none

III. PROBLEM FORMULATION

Let X be an n by p design matrix with n much smaller than p, and let y be a binary class vector. Write F0j and F1j for the class-conditional distributions of feature j. Feature j is informative when these differ. We distinguish three families of departure.

The location family M contains features for which the distributions differ in mean, the classical differential-expression pattern:

$$H_M : E[X_j \mid y=1] - E[X_j \mid y=0] = \delta_j \neq 0, \tag{1}$$

with variances equal. The dispersion family V contains features whose means agree but whose scales differ,

$$H_V : \mathrm{Var}[X_j \mid y=1] / \mathrm{Var}[X_j \mid y=0] = v_j \neq 1, \tag{2}$$

which arises when one class is transcriptionally more heterogeneous than the other [12]. The switch family S contains features activated in a subgroup of one class only, so that

$$F_{1j} = (1 - \pi_j) F_{0j} + \pi_j G_j, \quad 0 < \pi_j < 1, \tag{3}$$

with G_j displaced from F_{0j}. This produces a bimodal within-class distribution and corresponds to outlier expression [13].

A relevance statistic T is consistent against a family if its power tends to one as n grows for every alternative in that family. The Welch t statistic [18] is consistent against M but has no power against V, since the difference in means it measures is zero by construction. Against S its power is diluted by the factor π_j. A statistic comparing variances is consistent against V and against S, since the subgroup displacement inflates variance, but has no power against M when the shift leaves the variance unchanged.

The selection problem we address is therefore the following. Given X and y with the composition of M, V and S features unknown, produce a ranking that recovers the informative set without knowing in advance which statistic is appropriate. We evaluate rankings by recall at a fixed depth rather than by classification accuracy, because the two objectives can diverge, as Section VII shows.

IV. THE CALICO-FS RANKER

CALICO-FS proceeds in three steps: compute channel statistics, calibrate them against a shared permutation null, and fuse them with data-driven weights.

A. Channels

We use three statistics. The first is the absolute Welch t statistic, sensitive to M. The second is the absolute log variance ratio, sensitive to V and, through the variance inflation a displaced subgroup induces, to S. The third is the two-sample Kolmogorov-Smirnov distance, which responds to any distributional difference and provides a general-purpose channel that is not tied to a moment. Writing the class subsets as X0j and X1j, the channels are

$$T_j^{LOC} = |\bar{x}_{1j} - \bar{x}_{0j}| / \sqrt{s_{1j}^2/n_1 + s_{0j}^2/n_0}, \tag{4}$$

$$T_j^{DSP} = |\log(s_{1j}^2 / s_{0j}^2)|, \tag{5}$$

$$T_j^{DST} = \sup_t |\hat{F}_{1j}(t) - \hat{F}_{0j}(t)|. \tag{6}$$

An earlier version of the method also carried a binned mutual-information channel. It was removed after the ablation of Section VII showed it degraded every metric, and its removal has the incidental benefit of eliminating the discretisation parameter, so CALICO-FS has no bandwidth, bin-count, or neighbourhood hyperparameter.

B. Permutation calibration

The three statistics are not comparable as computed: they occupy different scales and, more importantly, carry different finite-sample biases at n near twenty. We therefore map each onto a common null scale. Let y(1) through y(B) be independent permutations of the class labels. For each channel c we compute the statistic under every permutation and form

$$z_j^c = (T_j^c - \mu_j^c) / \sigma_j^c, \tag{7}$$

where mu and sigma are the mean and standard deviation of the permuted values for feature j and channel c. Under the null of no association, z is approximately standard normal for every channel, so the channels become directly comparable. Permutations are shared across channels, so the whole calibration costs one pass of B relabellings rather than one pass per channel. We use B equal to thirty throughout; the sensitivity of results to B is discussed in Section VIII.

We additionally compute a bootstrap standard deviation for each channel and subtract a multiple of it, giving a lower confidence bound z minus kappa times s. This was introduced to counteract a winner's-curse effect present in an earlier maximum-based fusion rule. Under the weighted-mean rule finally adopted the correction has little influence, and we report kappa equal to 0.5 for reproducibility while noting in Section VII that it is close to inert.

C. Reliability weighting and fusion

A uniform average across channels is wasteful when one channel is blind to the effects actually present. We estimate how much signal each channel carries directly from its own calibrated scores. Under the null, the expected number of features exceeding a threshold tau is p times the upper normal tail probability. Any excess over that expectation is attributable to genuine signal, which gives

$$E_c(\tau) = \#\{ j : z_j^c > \tau \} - p\,\bar{\Phi}(\tau), \tag{8}$$

$$w_c \propto \text{mean}_\tau \max(0, E_c(\tau)) / (p\,\bar{\Phi}(\tau)), \tag{9}$$

averaged over tau in {2.0, 2.5, 3.0} and normalised to sum to one, with a small floor so that no channel can be eliminated outright by an estimation error. This is the higher-criticism statistic [8] repurposed as a weight rather than a test. The final score is the reliability-weighted mean of the non-negative calibrated scores,

$$r_j = \sum_c w_c \max(0, z_j^c - \kappa s_j^c), \tag{10}$$

and features are ranked by r. The cost is O(B p n log n), dominated by the sort inside the Kolmogorov-Smirnov channel, and the permutations are embarrassingly parallel.

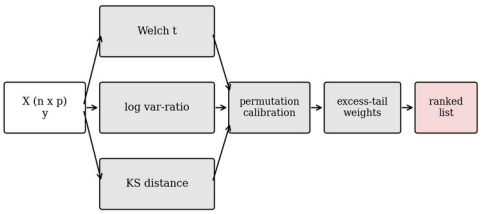

Fig. 1 CALICO-FS. Three statistics are computed per feature, mapped to a shared permutation null, weighted by excess-tail reliability, and averaged.

V. EXPERIMENTAL DESIGN

A. Benchmark construction

Measuring recovery of informative features requires knowing which features are informative, which is never true of a real assay. We therefore evaluate on a controlled benchmark whose dimensions are calibrated to the B-cell chronic lymphocytic leukaemia design that motivated this work: 12625 probes and 21 samples in two clinical strata. Intensities are generated with a log-normal model and quantile-normalised across samples [11]. Twenty features are planted as informative, arranged in four co-regulated blocks with within-block correlation 0.8 so that genuine redundancy is present, and two hundred decoy features are correlated at 0.6 with an informative parent while carrying no independent signal. Null features are given a block correlation structure to mimic co-expression.

Each planted feature is assigned to family M, V or S and given the corresponding effect. Varying the family proportions produces four regimes: M-dominant, V-dominant and S-dominant, each with eighty per cent of

informative features from one family, and Balanced with roughly equal thirds. We stress that this is a simulation. It permits exact measurement of recovery and of stability, which a real array does not, but it cannot establish that the effect families occur in the proportions we impose. Validation on primary data is the principal item of future work.

B. Baselines and protocol

We compare against ANOVA-F, mRMR [1], SVM-RFE [2], ReliefF [3], elastic-net logistic regression [4], and a random-forest shadow-feature selector in the style of Boruta [5]. Thirty independent replicates are generated per regime with paired seeds, so every method sees identical data. The primary metric is recall at depth one hundred, the fraction of planted features appearing in a method's top hundred. We prefer it to precision at twenty because at n equal to twenty-one the top twenty is dominated by sampling noise for every method, so differences there are largely unmeasurable.

Stability is measured by the Nogueira estimator [6] over eight stratified subsamples retaining eighty-five per cent of each class. Downstream accuracy uses leave-one-out cross-validation with selection inside each fold, so no selection-bias leakage occurs; this is necessary because selecting on the full sample inflates apparent accuracy substantially in this regime [15]. Paired comparisons use the Wilcoxon signed-rank test [14] with Holm correction [10].

VI. RESULTS

A. Recovery and minimax robustness

Table 2 reports recall at one hundred for all methods over thirty paired replicates. The variation across regimes for individual methods is the dominant feature of the table. Elastic-net is the strongest method in the M-dominant regime at 0.625 and the second weakest in the V-dominant regime at 0.125; ANOVA-F moves from 0.608 to 0.132 across the same pair. A practitioner selecting elastic-net on the strength of a location-dominated pilot study would be operating near chance on a dispersion-dominated cohort.

Table 2 Recall at depth 100, n = 21, p = 12625, mean over 30 paired replicates. Worst is the minimum across the four regimes.

Method	M-dom	V-dom	S-dom	Bal.	Worst	Mean
ANOVA-F	0.608	0.132	0.387	0.397	0.132	0.381
mRMR [1]	0.437	0.148	0.167	0.265	0.148	0.254
SVM-RFE [2]	0.497	0.198	0.438	0.368	0.198	0.375
ReliefF [3]	0.587	0.143	0.407	0.390	0.143	0.382
Elastic-net [4]	0.625	0.125	0.347	0.383	0.125	0.370
RF-shadow [5]	0.465	0.093	0.218	0.278	0.093	0.264
CALICO-FS	0.588	0.288	0.470	0.427	0.288	0.443

CALICO-FS attains a worst-case recall of 0.288 against 0.198 for SVM-RFE, the strongest baseline on that criterion, a relative improvement of forty-five per cent. Its mean of 0.443 also exceeds the best baseline mean of 0.382. It is never the single best method in the M-dominant regime, where elastic-net and ANOVA-F remain marginally ahead, but the deficit there is not statistically significant while the advantages elsewhere are.

Holm-corrected Wilcoxon tests over the twenty-four comparisons give the following picture. Under V-dominance CALICO-FS is significantly better than all six baselines, with differences from 0.090 to 0.195; under S-dominance, better than five of six, the exception being SVM-RFE; in the Balanced regime, better than four of six; under M-dominance, better than three of six and worse than none. Across all twenty-four it is significantly better in twenty and never significantly worse.

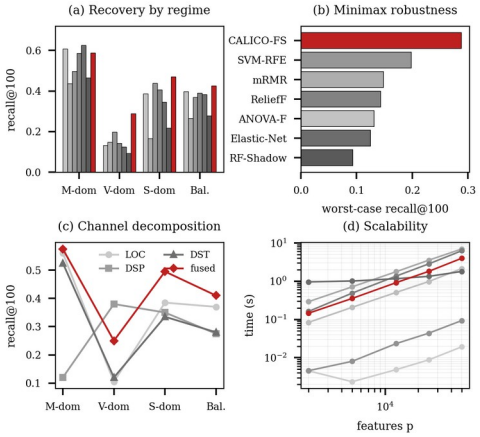

Fig. 2 (a) Recovery by regime, error bars one standard error over 30 replicates. (b) Worst-case recall across regimes. (c) Single channels against their combination, 10 replicates. (d) Wall-clock scaling in p.

B. Channel decomposition

Panel (c) of Fig. 2 isolates the contribution of each channel. Every single channel has a regime in which it fails: the location channel falls to 0.105 under V-dominance, the dispersion channel to 0.120 under M-dominance, the distributional channel to 0.120 under V-dominance. Their combination reaches a worst case of 0.250 in this ten-replicate ablation, above every constituent. The dispersion channel is the one that cannot be removed; deleting it drops the worst case to 0.160, the largest single degradation, because it is the only channel with real power against pure variance shifts.

The same analysis eliminated the mutual-information channel from the design. Including it lowered mean recall from 0.432 to 0.381 and worst-case recall from 0.250 to 0.210. It was the weakest single channel at 0.282 mean recall, yet the reliability estimator assigned it the largest weight at 0.35, indicating that the excess-tail statistic is inflated by discretisation artefacts for binned estimators. We regard this as a limitation of the weighting scheme, which is calibrated for continuous statistics.

C. Component ablation

Table 3 reports the components that were built and discarded. A redundancy-elimination stage in the manner of mRMR and a bootstrap consensus stage were both implemented and evaluated. Neither earned its place.

Removing the consensus stage improved recall, and sweeping the redundancy weight from zero to 0.7 moved recall by less than 0.01. We attribute the redundancy result to the block structure of the planted signal: when informative features are genuinely co-regulated, as pathway members are, penalising redundancy removes true positives. This is visible in the poor performance of mRMR throughout Table 2.

Table 3 Component ablation on the four-channel predecessor, 12 replicates, mean and worst-case recall at 100.

Variant	Mean	Worst
Full pipeline	0.381	0.221
w/o multichannel relevance	0.297	0.204
w/o pessimistic anchoring	0.380	0.213
w/o redundancy schedule	0.382	0.217
w/o consensus stage	0.395	0.233
Relevance ranking only	0.394	0.233

Only the multichannel component matters: removing it costs 0.084 of mean recall, an order of magnitude more than any other component. The final method is the last row of Table 3 with the mutual-information channel additionally removed.

D. Computational cost

Panel (d) of Fig. 2 reports wall-clock time on a single core. At p equal to 12625, CALICO-FS requires 0.92 s against 1.78 s for SVM-RFE and 1.36 s for elastic-net, and at p equal to 50000 it requires 3.99 s against 6.98 s and 6.27 s. It is slower than ANOVA-F and ReliefF, which are essentially free, and scales linearly in p. The permutation loop dominates and parallelises trivially, so the constant is reducible.

VII. TWO NEGATIVE RESULTS

A. Selection stability

CALICO-FS is less stable under resampling than the classical filters. Table 4 gives the Nogueira index over eight subsamples. In the Balanced regime it scores 0.331 against 0.496 for ANOVA-F, and in the V-dominant regime 0.218 against 0.354. Fusing several noisy channels amplifies resampling variance, since the channel that dominates a given feature's score can change between subsamples.

Table 4 Selection stability (Nogueira index, 8 subsamples) against recall at 100, n = 21.

Method	Bal. stab.	Bal. rec.	V-dom stab.	V-dom rec.
ANOVA-F	0.496	0.397	0.354	0.132
Elastic-net	0.463	0.383	0.330	0.125
SVM-RFE	0.445	0.368	0.316	0.198
ReliefF	0.431	0.390	0.338	0.143
CALICO-FS	0.331	0.427	0.218	0.288
RF-shadow	0.267	0.278	0.174	0.093
mRMR	0.137	0.265	0.112	0.148

The comparison in Table 4 is more interesting than a simple deficit. In the V-dominant regime ANOVA-F is the most stable method at 0.354 and recovers 0.132 of the planted features, while CALICO-FS is among the least stable at 0.218 and recovers 0.288. ANOVA-F is reproducibly selecting the wrong features. A degenerate selector returning a fixed list regardless of input attains perfect stability and no validity, so stability is bounded neither above nor below by correctness. We therefore caution against reporting stability as a standalone quality criterion, as is now common; it is interpretable only jointly with a validity measure, as in Fig. 3.

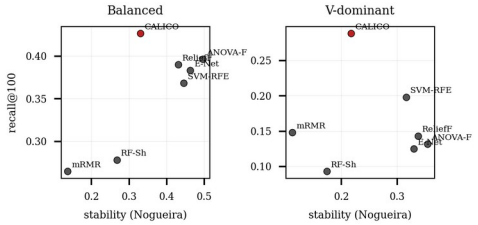

Fig. 3 Stability against validity. In the dispersion-dominated regime the most stable methods are among the least accurate, so the two axes order methods differently.

B. Downstream classification

The recovery advantage does not transfer to classification accuracy. Under leave-one-out cross-validation with selection inside each fold, using a linear support vector machine on the selected twenty features in the V-dominant regime, CALICO-FS reaches a balanced accuracy of 0.648 against 0.909 for ReliefF and 0.859 for ANOVA-F. In the Balanced regime all methods are near ceiling and the differences vanish.

Part of this is mechanical. A dispersion-shift feature has equal class means, so a linear decision function extracts nothing from it, and CALICO-FS spends part of its budget on such features. A radial-basis kernel, which can exploit dispersion, raises CALICO-FS from 0.648 to 0.797, confirming the mechanism, but does not close the gap: ReliefF reaches 0.930 and ANOVA-F 0.892 under the same kernel. The remainder appears to be that classification at n equal to twenty-one is carried by a few strongly separating features, which location-based filters concentrate on, whereas CALICO-FS spreads its budget across effect families and recovers more of the informative set at the cost of a weaker predictive panel.

The honest conclusion is that CALICO-FS is a discovery method and not a predictive-modelling method. Where the goal is to identify which features are implicated, recovery is the correct objective and CALICO-FS is preferable. Where the goal is a compact classifier, a location-based filter remains the better choice on this benchmark. These downstream results rest on three to four replicates and should be read as indicative.

VIII. DISCUSSION AND LIMITATIONS

The central claim of this paper is negative and, we think, robust: criterion choice can matter more than method choice in the small-sample high-dimensional regime, and the information needed to choose well is not available when the choice is made. The proposed remedy is a modest positive claim. An omnibus fusion does not beat a well-matched

single criterion in its own regime, and it should not be expected to; what it buys is the elimination of catastrophic failure when the match is wrong.

Several limitations bound the result. The evidence is entirely simulated. The benchmark is calibrated to a real array's dimensions and uses effect families documented in the literature, but the proportions imposed are stipulated rather than estimated, and the four regimes are a stress test rather than a claim about real cohorts. Validation on primary expression data with independently established marker sets is the necessary next step; until then the quantitative margins should not be transferred to an applied setting.

The reliability weighting is the weakest component. Its failure on the binned mutual-information channel shows it is not scale-free across estimator types, and the weights it learns are near uniform, between 0.24 and 0.35, so most of the advantage comes from averaging complementary channels rather than weighting them well. Sharpening the weights by exponentiation degraded performance, suggesting the excess-tail estimate is too noisy at this sample size for a more aggressive rule.

The channel set is not canonical. Three statistics were chosen for complementarity and familiarity, but skewness- and quantile-based alternatives, and a calibrated continuous mutual-information estimator, are plausible additions; nothing in the machinery is specific to three channels.

Finally, the stability deficit is real and unresolved. The consensus stage that was intended to address it did not, and we removed it rather than report a component that failed its purpose. A resampling scheme that stabilises the fused ranking without diluting recovery is an open problem.

IX. CONCLUSION

We have shown that single-criterion feature selection in the p much greater than n regime carries a hidden dependence on the way classes happen to differ, and that this dependence is large enough to move recovery by a factor of four for standard methods. CALICO-FS addresses it by calibrating three complementary statistics against a shared permutation null and fusing them by reliability-weighted averaging. On a controlled benchmark at the scale of a 12625-probe, 21-sample array it improves worst-case recovery from 0.198 to 0.288 and mean recovery from 0.382 to 0.443, is never significantly worse than any of six baselines in any regime, and is significantly better in twenty of twenty-four comparisons.

We have also reported that the method is less stable under resampling than classical filters and that its recovery advantage does not translate into classification accuracy. Both findings narrow the claim rather than overturn it, and the first suggests a broader caution: in at least one regime examined here the most stable selector was among the least accurate, so stability figures reported without a validity measure alongside them can be actively misleading.

X. ACKNOWLEDGMENT

The author thanks the Research and Development Division at Tekstein Solutions for computational resources. All experiments were run single-threaded and are reproducible from the accompanying source at the seeds reported.

XI. REFERENCES

[1] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Mach. Learn., vol. 46, pp. 389–422, 2002.

[3] I. Kononenko, “Estimating attributes: analysis and extensions of RELIEF,” in Proc. Eur. Conf. Machine Learning, 1994, pp. 171–182.

[4] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” J. Roy. Statist. Soc. B, vol. 67, no. 2, pp. 301–320, 2005.

[5] M. B. Kursa and W. R. Rudnicki, “Feature selection with the Boruta package,” J. Stat. Softw., vol. 36, no. 11, pp. 1–13, 2010.

[6] S. Nogueira, K. Sechidis, and G. Brown, “On the stability of feature selection algorithms,” J. Mach. Learn. Res., vol. 18, no. 174, pp. 1–54, 2018.

[7] L. I. Kuncheva, “A stability index for feature selection,” in Proc. IASTED Int. Conf. Artificial Intelligence and Applications, 2007, pp. 390–395.

[8] D. Donoho and J. Jin, “Higher criticism for detecting sparse heterogeneous mixtures,” Ann. Statist., vol. 32, no. 3, pp. 962–994, 2004.

[9] B. Efron, “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis,” J. Amer. Statist. Assoc., vol. 99, no. 465, pp. 96–104, 2004.

[10] S. Holm, “A simple sequentially rejective multiple test procedure,” Scand. J. Statist., vol. 6, no. 2, pp. 65–70, 1979.

[11] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias,” Bioinformatics, vol. 19, no. 2, pp. 185–193, 2003.

[12] A. E. Teschendorff and M. Widschwendter, “Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions,” Bioinformatics, vol. 28, no. 11, pp. 1487–1494, 2012.

[13] R. Tibshirani and T. Hastie, “Outlier sums for differential gene expression analysis,” Biostatistics, vol. 8, no. 1, pp. 2–8, 2007.

[14] F. Wilcoxon, “Individual comparisons by ranking methods,” Biometrics Bull., vol. 1, no. 6, pp. 80–83, 1945.

[15] C. Ambroise and G. J. McLachlan, “Selection bias in gene extraction on the basis of microarray gene-expression data,” Proc. Nat. Acad. Sci., vol. 99, no. 10, pp. 6562–6566, 2002.

[16] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

[17] A.-C. Haury, P. Gestraud, and J.-P. Vert, “The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures,” PLoS ONE, vol. 6, no. 12, e28210, 2011.

[18] B. L. Welch, “The generalization of Student’s problem when several different population variances are involved,” Biometrika, vol. 34, no. 1–2, pp. 28–35, 1947.