STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, December 2026, pp 5609–5629.
Published online in International Academic Press (www.IAPress.org)

Performance-Weighted Resampling Consensus Variable Selection for High-Dimensional Cancer Classification

Noor Abdulmunim, Omar Q. Alshebly*

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq
First-author email: nooor.24csp48@student.uomosul.edu.iq

Abstract The colon and prostate gene-expression matrices contain, respectively, 2,000 variables for 62 samples and 12,600 variables for 102 samples. In such small-sample settings, a useful selector must be judged by more than the accuracy of one fitted model. We developed Performance-Weighted Resampling Consensus Variable Selection (PWRCVS) to examine whether a short gene list can remain competitive after repeated perturbation of the training data. In each outer-training sample, Fisher Score and Information Gain were crossed with Random Forest, Gradient Boosting, and Extreme Gradient Boosting. Thirty stratified resamples produced six filter–learner paths per resample. A path contributed in proportion to its internally validated Matthews correlation coefficient (MCC), and variables were scored from their repeated positive permutation importance. Correlation components were then reduced to one representative before LASSO, Adaptive LASSO, Elastic Net, SCAD, MCP, RF, GBOOST, and XGBOOST were refitted. Evaluation used five repeats of outer stratified five-fold cross-validation; no test-fold observation entered filtering, tuning, or threshold selection. PWRCVS retained 7.76 ± 2.24 variables for colon cancer and 14.12 ± 4.00 for prostate cancer, reductions of 99.612% and 99.888%. On colon data, PWRCVS-XGBOOST gave MCC $= 0.663$ versus 0.666 for FS+IG-ENET, while the proposed list was 56.5% smaller and won eight of eleven matched MCC comparisons. On prostate data, PWRCVS-RF gave MCC $= 0.853$ versus 0.864 for IG-RF with 21.6% fewer variables. None of the 11 prespecified classifier-matched MCC contrasts was significant after Holm adjustment in either dataset. Stability selection, bootstrap selection and recursive feature elimination were then run inside the identical consensus machinery; none of the 12 resulting contrasts was significant either, and the proposed selector cost about four to six times more per fold. Thus, the empirical gain is parsimony with little loss of discrimination in these two datasets, not uniform predictive superiority.

Keywords High-dimensional data, Gene selection, PWRCVS, Regularized logistic regression, Ensemble learning, Matthews correlation coefficient, Cancer classification

AMS 2010 subject classifications 62H30, 62J07, 62P10, 68T05

DOI: 10.19139/soic-2310-5070-4546

1. Introduction

The two benchmark matrices used here illustrate the scale of the problem directly. The colon data contain about 32 measured genes per sample, and the corresponding ratio in the prostate data is about 124 [1, 2]. With these ratios, very different coefficient vectors or tree structures may fit the same training observations. A high training score can therefore coexist with an unstable selected set and a much lower score on held-out samples. This is the practical consequence of the $p \gg n$ regime addressed in the present analysis [3].

Recent work in this journal illustrates three complementary directions relevant to that problem: leakage-free statistical comparison of selector–classifier pairs in cancer data [4], factor-augmented dimension reduction for high-dimensional regression [5], and optimized feature selection within an attention-based clinical classifier [6]. These studies reinforce that predictive performance, selection stability, dimensional reduction, and computational

*Correspondence to: Omar Q. Alshebly (Email: omarqusay@uomosul.edu.iq; ORCID: 0000-0002-4126-3259), Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

feasibility are distinct objectives. The motivation for PWRCVS is therefore not to replace those frameworks, but to address a narrower operational question: when several low-cost selection paths disagree in a small training sample, can their validation performance be used to produce a shorter consensus list without claiming a predictive gain that the data do not support?

In this setting, the selected variables and the fitted classifier form one analytical procedure. Ranking genes once on the full matrix and then cross-validating only the classifier would allow the class labels of the nominal test folds to influence the model. Ambroise and McLachlan [7] and Varma and Simon [8] showed why this practice yields optimistic error estimates. We consequently repeated the filter calculation, subset-size decision, model tuning, and probability-threshold choice inside every outer-training sample.

We used two inexpensive filters that encode different evidence. Fisher Score (FS) favors a large between-class mean difference relative to within-class variation; Information Gain (IG) ranks a gene by the reduction in class entropy [9, 10]. Their candidates were examined with five penalized logistic estimators—LASSO, ALASSO, ENET, SCAD, and MCP—and with RF, GBOOST, and XGBOOST [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. The penalties differ in sparsity, grouping, and large-coefficient bias, while the tree ensembles allow nonlinear splits and interactions. This diversity is useful only if every path is evaluated under the same data partitions.

FS and IG were chosen because they are fast enough to be recomputed in every resample and encode complementary mean-separation and entropy evidence without estimating thousands of simultaneous parameters. RF, GBOOST, and XGBOOST were chosen as nonlinear importance generators with different randomization, sequential-fitting, and regularization mechanisms. An SVM could supply model-agnostic permutation importance, but its repeated tuning would introduce additional kernel and scale choices; a neural network would add many parameters to training portions containing fewer than 102 observations. Recent graph-neural feature selection demonstrates the promise of learned gene-dependency representations [21], but also requires a different data and tuning regime. The six selected paths are therefore a deliberate, computationally tractable design for these small samples, not a claim that they exhaust all possible filters or learners.

The difficulty is not limited to prediction error. If two correlated genes carry similar information, a small change in the training observations can exchange their ranks without materially changing classification. Exact list overlap may then be low even when the underlying signal is similar. Stability-selection and ensemble-selection studies treat this reproducibility as a separate outcome [22, 23, 24, 25]. Work on high-dimensional, low-sample-size data has likewise shown that a small near-best subset can be preferable when measurement and validation costs are considered [26, 27, 28]. We therefore report test performance, selected-set size, and stability together.

Our earlier procedures used screening followed by regularized logistic regression or SCAD-based resampling [29, 30]. PWRCVS changes the unit being aggregated. Instead of resampling one selector, it creates six FS/IG–tree paths in every resample. Their contributions are unequal: a path receives more weight only when its internal validation MCC is higher. Positive permutation importance and repeated selection are then combined, and a correlation graph removes redundant representatives before the final models are fitted. This construction tests whether diversity among selectors is useful after accounting for their observed internal performance.

The contributions are threefold. First, PWRCVS defines a performance-weighted selection consensus in which validation MCC and positive permutation importance jointly determine how much evidence each filter–learner path contributes. Second, the complete selector—including screening, weighting, consensus-threshold choice, correlation pruning, hyperparameter tuning, and decision-threshold choice—is evaluated inside a repeated outer cross-validation design. Third, the empirical claim is evaluated as a performance–parsimony trade-off, supported by corrected paired comparisons, component ablations, sensitivity analyses, recurrence, stability, and runtime rather than by accuracy alone. The novelty is this integrated and auditable selection strategy; it is not presented as a new asymptotic theory or as a fundamentally new base learner.

The empirical question is deliberately narrow: on the colon and prostate matrices, does PWRCVS reduce the number of retained genes without a material deterioration in outer-test performance? Conventional FS, IG, and unweighted FS+IG paths provide the reference. We also ask whether the selected lists recur across outer-training samples. No simulation evidence or third dataset is used in this article; conclusions are restricted to the two observed applications.

2. Statistical Background and Classifiers

2.1. High-dimensional binary classification

Let $\mathbf{X}$ denote the $n \times p$ expression matrix and let $Y = 1$ identify tumor tissue. For both applications p exceeds n by more than an order of magnitude, so the ordinary logistic likelihood cannot provide a stable unrestricted coefficient vector. We write the conditional tumor probability as

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\beta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})}{1 + \exp(\beta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})}. \tag{1}$$

All five logistic competitors minimize the same mean loss and differ through $P(\boldsymbol{\beta})$. Their tuning parameter λ was selected in the inner training procedure. A coefficient retained by this optimization is used as a selection indicator; it is not interpreted as a causal effect or as evidence from a conventional hypothesis test:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} [y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)] + \lambda P(\boldsymbol{\beta}) \right\}. \tag{2}$$

2.2. Fisher Score and Information Gain

FS and IG were recalculated after each resample rather than imported from a ranking of the complete dataset. For gene j, FS compares the class-specific means with the pooled within-class variation. IG measures the change in entropy after observing that gene; its discretization was also estimated from the current resample. The two scores are

$$\mathrm{FS}(j) = \frac{\sum_c n_c(\mu_{cj} - \mu_j)^2}{\sum_c n_c s_{cj}^2 + \varepsilon}, \tag{3}$$

$$\mathrm{IG}(X_j; Y) = H(Y) - H(Y \mid X_j). \tag{4}$$

Within an outer-training sample of size n_{tr}, each filter retained at most d_n genes. The bound below follows the screening scale of Fan and Lv [10]; here it controls computation and does not claim to estimate the true number of associated genes:

$$d_n = \min\left\{ p, \left\lfloor \frac{n_{\mathrm{tr}}}{\log(n_{\mathrm{tr}})} \right\rfloor \right\}. \tag{5}$$

2.3. Regularized logistic regression and tree ensembles

The five penalized models provide different responses to a correlated candidate list. LASSO gives a convex sparse fit, ALASSO changes the penalty gene by gene, and ENET can retain groups of related variables. SCAD and MCP use non-convex penalties to reduce shrinkage of large coefficients [31, 32]. We kept all five because no one penalty is uniformly preferable for the two data structures.

RF averages randomized trees; GBOOST and XGBOOST build trees sequentially, with XGBOOST adding explicit complexity control. Their raw importance values are not directly comparable and can be divided among correlated genes [33]. We therefore recomputed permutation importance on internal validation observations, converted importance to within-path ranks, and accumulated evidence across resamples.

3. Proposed PWRCVS Method

3.1. Rationale

PWRCVS was designed around two instabilities observed in $p \gg n$ selection: the ranking changes after a small perturbation of the sample, and the preferred ranking may depend on the learner used after filtering. Accordingly,

Table 1. Regularized logistic regression models used in the study

Model	Penalty $P(\boldsymbol{\beta})$	Applied statistical role
LASSO	$\sum_j \lvert\beta_j\rvert$	Direct sparsity, although correlated variables may exchange membership across samples.
ALASSO	$\sum_j w_j \lvert\beta_j\rvert$	Adaptive coefficient-specific penalties derived from an initial estimate.
ENET	$\alpha \sum_j \lvert\beta_j\rvert + (1-\alpha)\sum_j \beta_j^2/2$	Sparsity with a grouping tendency for correlated variables.
SCAD	$\sum_j p_{\lambda,\gamma}(\lvert\beta_j\rvert)$	Non-convex shrinkage designed to reduce bias for large coefficients.
MCP	$\sum_j p_{\lambda,\gamma}(\lvert\beta_j\rvert)$	A concave penalty balancing sparsity and reduced estimation bias.

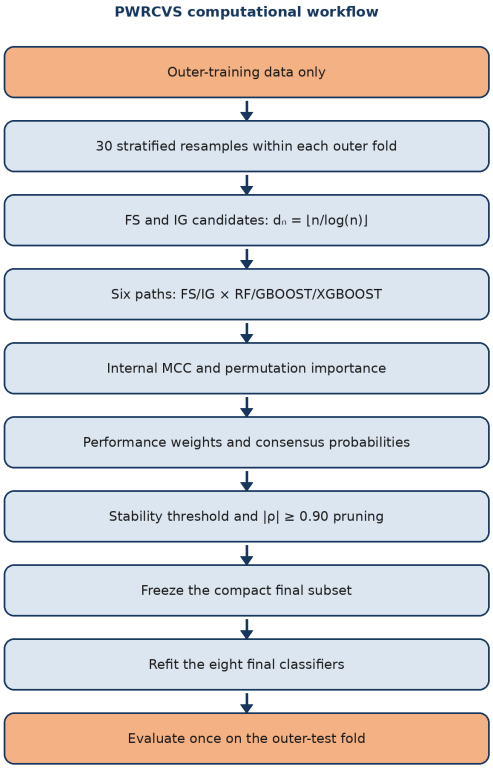


Figure 1. Leakage-controlled PWRCVS workflow. Every data-dependent operation, including filtering, path weighting, threshold selection, correlation pruning, and refitting, is completed within the outer-training sample before one evaluation on the reserved outer-test fold.

each resample produces six analyses—FS–RF, FS–GBOOST, FS–XGBOOST, IG–RF, IG–GBOOST, and IG–XGBOOST—rather than one definitive list.

For each analysis we stored three objects: the filtered candidate set, MCC on its internal validation observations, and the validation-set permutation importance of every candidate. Consequently, a path that fits its resampled training observations but fails on the associated validation portion receives little weight in the final consensus.

3.2. Performance weights and consensus probability

Index a resample by r, a filter by $f \in \{\mathrm{FS}, \mathrm{IG}\}$, and a tree learner by $m \in \{\mathrm{RF}, \mathrm{GBOOST}, \mathrm{XGBOOST}\}$. Let Q_{rfm} be the MCC obtained on the validation portion associated with that path. To construct nonnegative within-resample weights, we set negative MCC values to zero, add ε only to avoid an undefined denominator, and normalize over the

six paths. For gene j, permutation importance is the change in the same validation MCC after its values are shuffled:

$$w_{rfm} = \frac{\max(Q_{rfm}, 0) + \varepsilon}{\sum_{f,m}[\max(Q_{rfm}, 0) + \varepsilon]}, \tag{6}$$

$$\Delta_{rfm}(j) = Q_{rfm} - Q_{rfm}^{\mathrm{perm}(j)}. \tag{7}$$

Define $Z_{jrfm} = 1$ only when gene j survives filter f and has a positive importance rank retained for learner m; otherwise $Z_{jrfm} = 0$. Averaging the weighted indicators over the R resamples gives

$$\widehat{\Pi}_j^{\mathrm{PW}} = \frac{1}{R}\sum_{r=1}^{R}\sum_{f,m} w_{rfm} Z_{jrfm}. \tag{8}$$

Thus, neither frequent selection by poorly validating paths nor one high-scoring occurrence is enough to give a large $\widehat{\Pi}_j^{\mathrm{PW}}$.

MCC was specified before the outer-test results were examined. It is the phi correlation between the observed and predicted binary labels, incorporates all four cells of the confusion matrix, and is unchanged when the positive and negative labels are exchanged. These properties are useful here because the colon data are moderately imbalanced and both classes are scientifically important [34, 35]. Balanced accuracy averages sensitivity and specificity but does not incorporate the correctness of positive and negative predictions, whereas ROC–AUC summarizes ranking over all operating thresholds. By contrast, every PWRCVS path produces binary decisions at an internally selected threshold. Using MCC therefore aligns path validation, permutation importance, and the primary outer-test endpoint [36]. This alignment is a design rationale rather than a claim that MCC weighting must outperform every alternative weighting rule.

3.3. Stability threshold, correlation pruning, and refitting

The consensus threshold was selected separately inside every outer-training fold; the associated outer-test observations were not used. We evaluated $\pi_{\mathrm{thr}} \in \{0.05, 0.10, 0.20, 0.30, \ldots, 0.90\}$. For each distinct candidate subset, LASSO and RF were tuned by three-fold internal cross-validation, and the selection utility was their mean internal MCC. The candidate with the largest utility was retained; ties were resolved first in favor of fewer variables and then the larger threshold. To prevent a degenerate empty list, fewer than $d_{\min} = \max\{3, \lceil d_n/2 \rceil\}$ eligible genes triggered retention of the $d_{\min}$ largest consensus probabilities before correlation pruning. The chosen threshold and subset were then frozen for that outer fold. Among the genes above the threshold, an edge joined j and k when their outer-training correlation satisfied $|\rho_{jk}| \geq \rho_0$, with $\rho_0 = 0.90$. This conservative cutoff was specified to remove only near-collinear expression profiles while retaining moderate associations that may still contribute jointly to classification. We retained the member with the largest consensus probability from each connected component. All correlations were estimated without the outer-test observations. The omitted members remain recorded as correlated alternatives; the pruning rule is computational and is not a biological null finding.

After thresholding and pruning, the resulting $\widehat{S}_{\mathrm{PW}}$ was frozen. LASSO, ALASSO, ENET, SCAD, MCP, RF, GBOOST, and XGBOOST were then refitted from their initial state using only the outer-training observations. The three tree algorithms therefore have two separate roles: they generate importance within PWRCVS, and they may also serve as final classifiers. The test fold is accessed only after both roles have been completed:

$$\begin{aligned} \widehat{S}_0 &= \{j : \widehat{\Pi}_j^{\mathrm{PW}} \geq \pi_{\mathrm{thr}}\}, \\ \widehat{S}_{\mathrm{PW}} &= \left\{ \arg\max_{j \in G_h} \widehat{\Pi}_j^{\mathrm{PW}} : h = 1, \ldots, H \right\}. \end{aligned} \tag{9}$$

Table 2. PWRCVS implementation algorithm

Step	Operation
1	Form one stratified outer split; reserve its test observations before estimating any preprocessing quantity.
2	Tune the six filter–tree path configurations by seeded, stratified three-fold cross-validation in the outer-training sample; freeze the selected settings for the resamples.
3	Obtain 30 stratified subsamples without replacement, drawing 80% of each class and reserving 30% of each subsample for validation.
4	Recalculate FS and IG in each resample fit portion, retain the first d_n ranks per filter, and fit the six paths using the frozen settings.
5	On the internal validation portion, record path MCC and the mean MCC decrease over three independent permutations of each candidate.
6	Normalize the six MCC-derived weights and accumulate $w_{rfm}Z_{jrfm}$ for every gene.
7	Choose π_{thr} without the outer-test fold and retain one representative per $\lvert\rho\rvert \geq \rho_0 = 0.90$ component.
8	Refit the five penalized logistic and three tree classifiers on the frozen gene set.
9	Generate one probability vector for the reserved outer-test observations and store all evaluation quantities.
10	Repeat for 25 outer splits; summarize test performance, selected-set size, recurrence, and stability.

4. Data and Evaluation Design

4.1. Real gene-expression datasets

We analyzed the colon matrix of Alon et al. [1] and the prostate matrix of Singh et al. [2]. The analysis files contained no missing entries. Colon measurements were transformed on the $\log_2$ scale; the supplied prostate values were used without an additional transformation. Table 3 shows that the larger prostate sample is nevertheless more severely high-dimensional because its variable count rises to 12,600.

Table 3. Description of the two gene-expression datasets

Dataset	Samples	Genes	Class 0	Class 1	p/n	Transform
Colon	62	2,000	22 normal	40 tumor	32.26	$\log_2$
Prostate	102	12,600	50 normal	52 tumor	123.53	None added

4.2. Validation protocol and leakage control

For each dataset, five independently repeated stratified five-fold partitions yielded 25 outer test evaluations. In an outer-training set, we generated 30 stratified resamples and repeated every data-dependent operation: filter ranking, path weighting, π_{thr} selection, construction of correlation components, and classifier tuning. Only the finalized subset and fitted model were applied to the reserved fold. The complete run used master seed 20260722.

Two one-factor-at-a-time analyses examined design sensitivity without using an outer-test observation for tuning. First, 10, 20, 30, and 50 cumulative prefixes of the same seeded resample sequence were evaluated while d_n and all other rules were fixed. Second, the candidate cap was set to $0.5d_n$, $1.0d_n$, $1.5d_n$, and $2.0d_n$ while the original 30 resamples were retained. These values corresponded to 6, 12, 18, and 24 candidates for colon and 9, 18, 27, and 36 for prostate. The adaptive consensus threshold, correlation pruning, and LASSO final model were re-estimated inside each outer training fold for every setting. Paired MCC differences were assessed with the corrected resampled t standard error; Holm adjustment covered the three non-reference comparisons within each dataset and factor.

Training predictions were retained only to quantify apparent fit and the training–test gap. Rankings and conclusions use the outer-test summaries. We did not feed those summaries back into subset-size or hyperparameter choices. Computation used NumPy and scikit-learn [37, 38]; the tuning operations followed the inner-training principle described by Probst et al. [39].

4.3. Implementation details and computational cost

Within each PWRCVS resample, sampling was stratified and performed without replacement. Eighty percent of each class entered the subsample; 30% of that subsample was assigned to validation and the remainder to fitting (approximately 24% and 56% of the outer-training observations, respectively, subject to integer rounding). The 20% not drawn in a given resample was not used in that resample. The six path configurations were selected once per outer-training fold by seeded, stratified three-fold cross-validation and then held fixed during its 30 resamples. FS and the three-bin equal-frequency IG discretization were nevertheless recalculated from every resample fit portion. Each candidate's permutation importance was the mean validation-MCC decrease over three independent shuffles.

The analysis matrices contained no missing entries, so no imputation was performed. Standardization for the five penalized models was fitted on the applicable training portion and then applied to its validation or test portion; tree models used the stated input scale. Class imbalance was handled inside training only: inverse-frequency weights were used for the penalized models and GBOOST, balanced bootstrap weights for RF, and a fold-specific `scale_pos_weight` for XGBOOST. No synthetic oversampling was used. For a new dataset with missing values, any imputation rule must likewise be estimated inside the current training partition before filtering.

Table 4. Implementation and hyperparameter settings used in the submitted analysis

Component	Prespecified values or search set
Penalized logistic models	$\lambda/\lambda_{\max} \in \{0.05, 0.20\}$ for all five models; ALASSO $\gamma = 1$; ENET $\alpha = 0.50$; SCAD $a = 3.7$; MCP $\gamma = 3$; maximum 1,000 iterations and tolerance 10^{-7}.
Random Forest	300 trees; maximum depth in $\{\text{unlimited}, 3\}$; minimum leaf size 2; $\sqrt{p}$ variables considered at a split; balanced bootstrap class weights.
Gradient Boosting	100 trees; learning rate 0.05; maximum depth in $\{1, 2\}$; row subsampling 0.80; inverse-frequency observation weights.
XGBoost	100 trees; learning rate 0.05; maximum depth in $\{1, 2\}$; row and column subsampling 0.80; $\lambda = 1, \alpha = 0, \gamma = 0$, minimum child weight 1, histogram tree construction, and fold-specific class ratio.
Internal model selection	Seeded, shuffled, stratified three-fold cross-validation. Hyperparameters maximize out-of-fold MCC, with Brier score as the model-grid tie-breaker.
Decision threshold	Candidates comprise 0.20–0.80 in increments of 0.02 and the distinct clipped out-of-fold probabilities. MCC is maximized; ties favor balanced accuracy and then proximity to 0.50.
PWRCVS resampling	30 stratified subsamples without replacement; 80% of each class enters a subsample and 30% of that subsample is reserved for validation. Permutation importance uses three independent shuffles per candidate.

The reproducibility run used Python 3.12, NumPy 2.5.1, pandas 3.0.3, SciPy 1.18.0, scikit-learn 1.8.0, and XGBoost 3.0.5. The master seed was 20260722; learner fits were configured for one execution thread. We benchmarked the complete proposed pipeline on the first deterministic outer fold of each dataset using an Intel Xeon Platinum 8370C CPU at 2.80 GHz. Six-path tuning, 30-resample selection, and tuning/refitting of all eight final classifiers required 7.8, 75.2, and 6.4 seconds for colon (89.4 seconds total), and 8.3, 109.0, and 8.6 seconds for prostate (125.9 seconds total). Linear projection to the 25 outer folds gives approximately 37.3 and 52.5 minutes, respectively; peak resident memory was 218 and 221 MiB. These figures exclude the conventional comparators and revision-only sensitivity analyses and will vary with hardware. Re-execution reproduced the 16 archived PWRCVS metric and threshold values for the benchmarked folds to a maximum absolute difference of 1.11×10^{-16}.

4.4. Comparators and performance measures

The frozen PWRCVS set was paired with eight final classifiers. Eleven conventional paths used the same building blocks without the proposed consensus: the unweighted FS+IG union preceded each of the five penalized logistic models, and FS or IG separately preceded each of the three tree learners. This produced 19 evaluated paths and allowed a matched comparison for every final PWRCVS classifier.

For each outer fold we recorded accuracy, balanced accuracy, G-Mean, MCC, ROC-AUC, and PR-AUC. MCC is the primary classification summary because it uses all four confusion-matrix cells [34]; PR-AUC was retained because it focuses on positive-class retrieval [40]. Selection was described by its mean size, percentage reduction from p, pairwise Jaccard overlap, and Nogueira's chance-corrected stability coefficient [22].

Formal inference used 11 prespecified classifier-matched contrasts: each PWRCVS penalized model was paired with the same model after the FS+IG union, and each PWRCVS tree model was paired separately with its FS and IG counterparts. Let d_i be the paired outer-fold MCC difference (PWRCVS minus conventional). Because the 25 results arose from overlapping repeated five-fold partitions, we used the Nadeau–Bengio corrected resampled standard error $s_d\sqrt{1/25 + n_{\text{test}}/n_{\text{train}}}$, where $n_{\text{test}}/n_{\text{train}} = 1/4$, rather than treating the folds as independent observations [41]. Two-sided corrected t tests and 95% confidence intervals used 24 degrees of freedom. Holm adjustment controlled the family-wise error rate across the 11 contrasts within each dataset [42]. The mean paired MCC difference was the primary effect estimate; Hedges' small-sample-corrected g_z was added as a descriptive standardized effect because the folds still overlap.

5. Results

5.1. Training performance

Tables 5 and 6 are included to expose model optimism, not to estimate future performance. In colon training samples, PWRCVS-XGBOOST had the largest proposed MCC (0.912). In prostate training samples, the corresponding maximum was 0.986 for PWRCVS-GBOOST. Several tree AUC values were nearly one, a pattern that would be reassuring only if it persisted on the reserved observations.

The apparent fits were obtained with markedly different subset widths. Penalized conventional paths used means of 17.84 genes for colon and 24.68 for prostate, whereas the PWRCVS means were 7.76 and 14.12. We therefore use training values only alongside the outer-test values and do not rank procedures from Tables 5–6 alone.

Table 5. Training performance for the colon data; mean (standard deviation) across 25 outer folds

Method	Accuracy	BAcc	G-Mean	MCC	ROC-AUC	PR-AUC	No. vars
PWRCVS-LASSO	0.902 (0.054)	0.905 (0.056)	0.900 (0.067)	0.806 (0.091)	0.980 (0.012)	0.990 (0.006)	7.76 (2.24)
PWRCVS-ALASSO	0.917 (0.032)	0.922 (0.039)	0.919 (0.043)	0.833 (0.064)	0.981 (0.011)	0.990 (0.006)	7.76 (2.24)
PWRCVS-ENET	0.923 (0.023)	0.930 (0.022)	0.928 (0.022)	0.842 (0.044)	0.978 (0.013)	0.989 (0.007)	7.76 (2.24)
PWRCVS-SCAD	0.931 (0.042)	0.936 (0.040)	0.935 (0.041)	0.860 (0.082)	0.981 (0.014)	0.991 (0.007)	7.76 (2.24)
PWRCVS-MCP	0.934 (0.044)	0.936 (0.051)	0.934 (0.055)	0.864 (0.091)	0.982 (0.014)	0.991 (0.007)	7.76 (2.24)
PWRCVS-RF	0.956 (0.025)	0.960 (0.021)	0.959 (0.022)	0.909 (0.047)	0.997 (0.003)	0.999 (0.001)	7.76 (2.24)
PWRCVS-GBOOST	0.949 (0.058)	0.957 (0.049)	0.954 (0.053)	0.904 (0.101)	0.999 (0.002)	1.000 (0.001)	7.76 (2.24)
PWRCVS-XGBOOST	**0.957 (0.026)**	**0.962 (0.021)**	**0.962 (0.022)**	**0.912 (0.050)**	**0.995 (0.005)**	**0.998 (0.002)**	**7.76 (2.24)**
FS+IG-LASSO	0.912 (0.031)	0.919 (0.026)	0.917 (0.027)	0.822 (0.054)	0.978 (0.020)	0.988 (0.015)	17.84 (1.40)
FS+IG-ALASSO	0.923 (0.041)	0.930 (0.039)	0.927 (0.041)	0.847 (0.074)	0.982 (0.015)	0.991 (0.008)	17.84 (1.40)
FS+IG-ENET	0.913 (0.027)	0.919 (0.028)	0.918 (0.029)	0.822 (0.053)	0.979 (0.021)	0.989 (0.013)	17.84 (1.40)
FS+IG-SCAD	0.918 (0.079)	0.932 (0.063)	0.927 (0.076)	0.846 (0.121)	0.981 (0.027)	0.989 (0.021)	17.84 (1.40)
FS+IG-MCP	0.931 (0.077)	0.932 (0.079)	0.927 (0.092)	0.863 (0.140)	0.986 (0.014)	0.993 (0.007)	17.84 (1.40)
FS-RF	0.951 (0.033)	0.959 (0.025)	0.958 (0.026)	0.902 (0.060)	0.997 (0.003)	0.998 (0.002)	12.00 (0.00)
FS-GBOOST	0.928 (0.111)	0.941 (0.087)	0.933 (0.116)	0.876 (0.162)	1.000 (0.001)	1.000 (0.001)	12.00 (0.00)
FS-XGBOOST	0.940 (0.034)	0.947 (0.028)	0.946 (0.030)	0.881 (0.062)	0.994 (0.005)	0.997 (0.003)	12.00 (0.00)
IG-RF	0.948 (0.037)	0.957 (0.028)	0.956 (0.030)	0.898 (0.065)	0.997 (0.003)	0.998 (0.001)	12.00 (0.00)
IG-GBOOST	0.933 (0.077)	0.948 (0.059)	0.944 (0.065)	0.880 (0.128)	0.999 (0.002)	1.000 (0.001)	12.00 (0.00)
IG-XGBOOST	0.935 (0.032)	0.940 (0.028)	0.938 (0.029)	0.868 (0.059)	0.992 (0.006)	0.996 (0.003)	12.00 (0.00)

5.2. Outer-test performance

For colon cancer, the largest proposed MCC was 0.663 for PWRCVS-XGBOOST, followed by 0.662 for PWRCVS-MCP and 0.659 for PWRCVS-ENET. The best conventional path, FS+IG-ENET, achieved 0.666 - only 0.003 higher

Table 6. Training performance for the prostate data; mean (standard deviation) across 25 outer folds

Method	Accuracy	BAcc	G-Mean	MCC	ROC-AUC	PR-AUC	No. vars
PWRCVS-LASSO	0.974 (0.016)	0.974 (0.016)	0.974 (0.016)	0.948 (0.031)	0.995 (0.005)	0.996 (0.004)	14.12 (4.00)
PWRCVS-ALASSO	0.970 (0.018)	0.970 (0.018)	0.970 (0.019)	0.942 (0.035)	0.996 (0.004)	0.996 (0.004)	14.12 (4.00)
PWRCVS-ENET	0.973 (0.013)	0.973 (0.013)	0.973 (0.013)	0.946 (0.027)	0.996 (0.004)	0.996 (0.004)	14.12 (4.00)
PWRCVS-SCAD	0.978 (0.037)	0.978 (0.036)	0.977 (0.039)	0.958 (0.067)	0.997 (0.004)	0.998 (0.004)	14.12 (4.00)
PWRCVS-MCP	0.977 (0.036)	0.977 (0.035)	0.977 (0.037)	0.956 (0.066)	0.997 (0.004)	0.997 (0.004)	14.12 (4.00)
PWRCVS-RF	0.980 (0.009)	0.980 (0.009)	0.980 (0.009)	0.960 (0.018)	0.999 (0.001)	0.999 (0.001)	14.12 (4.00)
PWRCVS-GBOOST	**0.993 (0.011)**	**0.993 (0.011)**	**0.993 (0.011)**	**0.986 (0.021)**	**1.000 (0.000)**	**1.000 (0.000)**	**14.12 (4.00)**
PWRCVS-XGBOOST	0.987 (0.010)	0.987 (0.010)	0.987 (0.010)	0.974 (0.020)	0.999 (0.001)	0.999 (0.001)	14.12 (4.00)
FS+IG-LASSO	0.962 (0.017)	0.963 (0.017)	0.962 (0.017)	0.926 (0.034)	0.993 (0.005)	0.994 (0.005)	24.68 (1.60)
FS+IG-ALASSO	0.964 (0.024)	0.964 (0.023)	0.964 (0.024)	0.929 (0.045)	0.995 (0.004)	0.995 (0.004)	24.68 (1.60)
FS+IG-ENET	0.958 (0.018)	0.958 (0.018)	0.958 (0.018)	0.917 (0.035)	0.992 (0.005)	0.993 (0.004)	24.68 (1.60)
FS+IG-SCAD	0.982 (0.024)	0.982 (0.024)	0.982 (0.025)	0.965 (0.046)	0.996 (0.005)	0.996 (0.005)	24.68 (1.60)
FS+IG-MCP	0.976 (0.030)	0.977 (0.030)	0.976 (0.031)	0.954 (0.058)	0.996 (0.005)	0.997 (0.005)	24.68 (1.60)
FS-RF	0.975 (0.014)	0.975 (0.014)	0.975 (0.014)	0.950 (0.027)	0.999 (0.001)	0.999 (0.001)	18.00 (0.00)
FS-GBOOST	0.986 (0.015)	0.986 (0.015)	0.986 (0.015)	0.973 (0.029)	0.999 (0.001)	0.999 (0.001)	18.00 (0.00)
FS-XGBOOST	0.975 (0.015)	0.975 (0.015)	0.975 (0.015)	0.950 (0.030)	0.998 (0.002)	0.998 (0.002)	18.00 (0.00)
IG-RF	0.980 (0.009)	0.980 (0.009)	0.980 (0.009)	0.961 (0.017)	0.999 (0.001)	0.999 (0.001)	18.00 (0.00)
IG-GBOOST	0.984 (0.018)	0.985 (0.017)	0.984 (0.018)	0.970 (0.034)	1.000 (0.000)	1.000 (0.000)	18.00 (0.00)
IG-XGBOOST	0.975 (0.019)	0.975 (0.019)	0.975 (0.020)	0.951 (0.037)	0.998 (0.002)	0.998 (0.002)	18.00 (0.00)

- but used 17.84 rather than 7.76 variables. PWRCVS also improved MCC in eight of eleven matched comparisons, especially for boosting and MCP paths.

For prostate cancer, PWRCVS-RF was the best proposed model, with accuracy 0.924, balanced accuracy 0.924, G-Mean 0.922, and MCC 0.853. IG-RF provided the best conventional MCC of 0.864 using 18 variables, whereas PWRCVS used 14.12 on average. PWRCVS-GBOOST exceeded both FS-GBOOST and IG-GBOOST, and PWRCVS-XGBOOST exceeded FS-XGBOOST.

Table 7. Testing performance for the colon data; mean (standard deviation) across 25 outer folds

Method	Accuracy	BAcc	G-Mean	MCC	ROC-AUC	PR-AUC	No. vars
PWRCVS-LASSO	0.812 (0.086)	0.812 (0.092)	0.799 (0.103)	0.627 (0.183)	0.893 (0.080)	0.946 (0.045)	7.76 (2.24)
PWRCVS-ALASSO	0.799 (0.097)	0.789 (0.114)	0.774 (0.130)	0.585 (0.219)	0.895 (0.085)	0.948 (0.045)	7.76 (2.24)
PWRCVS-ENET	0.826 (0.070)	0.827 (0.078)	0.814 (0.086)	0.659 (0.137)	0.901 (0.084)	0.951 (0.042)	7.76 (2.24)
PWRCVS-SCAD	0.807 (0.097)	0.798 (0.111)	0.780 (0.133)	0.607 (0.210)	0.880 (0.091)	0.930 (0.069)	7.76 (2.24)
PWRCVS-MCP	0.833 (0.077)	0.826 (0.078)	0.816 (0.085)	0.662 (0.152)	0.870 (0.088)	0.921 (0.070)	7.76 (2.24)
PWRCVS-RF	0.826 (0.102)	0.816 (0.105)	0.806 (0.111)	0.644 (0.209)	0.895 (0.099)	0.949 (0.052)	7.76 (2.24)
PWRCVS-GBOOST	0.778 (0.127)	0.788 (0.122)	0.773 (0.131)	0.578 (0.239)	0.881 (0.104)	0.938 (0.063)	7.76 (2.24)
PWRCVS-XGBOOST	0.833 (0.111)	0.828 (0.113)	0.820 (0.117)	0.663 (0.220)	0.887 (0.103)	0.939 (0.065)	7.76 (2.24)
FS+IG-LASSO	0.797 (0.097)	0.799 (0.109)	0.785 (0.121)	0.598 (0.213)	0.892 (0.093)	0.945 (0.050)	17.84 (1.40)
FS+IG-ALASSO	0.820 (0.094)	0.813 (0.105)	0.805 (0.109)	0.629 (0.204)	0.883 (0.091)	0.942 (0.047)	17.84 (1.40)
FS+IG-ENET	**0.832 (0.092)**	**0.835 (0.093)**	**0.829 (0.094)**	**0.666 (0.188)**	**0.888 (0.098)**	**0.941 (0.056)**	**17.84 (1.40)**
FS+IG-SCAD	0.813 (0.127)	0.813 (0.127)	0.798 (0.145)	0.631 (0.246)	0.862 (0.113)	0.906 (0.089)	17.84 (1.40)
FS+IG-MCP	0.797 (0.103)	0.783 (0.110)	0.765 (0.130)	0.583 (0.227)	0.862 (0.097)	0.908 (0.079)	17.84 (1.40)
FS-RF	0.823 (0.110)	0.821 (0.115)	0.812 (0.121)	0.644 (0.226)	0.874 (0.116)	0.939 (0.056)	12.00 (0.00)
FS-GBOOST	0.755 (0.168)	0.757 (0.151)	0.715 (0.215)	0.538 (0.291)	0.871 (0.134)	0.939 (0.070)	12.00 (0.00)
FS-XGBOOST	0.804 (0.142)	0.799 (0.141)	0.790 (0.145)	0.609 (0.287)	0.876 (0.114)	0.941 (0.054)	12.00 (0.00)
IG-RF	0.781 (0.106)	0.765 (0.113)	0.744 (0.133)	0.555 (0.216)	0.842 (0.094)	0.916 (0.051)	12.00 (0.00)
IG-GBOOST	0.774 (0.114)	0.774 (0.116)	0.755 (0.137)	0.554 (0.227)	0.826 (0.103)	0.909 (0.059)	12.00 (0.00)
IG-XGBOOST	0.783 (0.096)	0.765 (0.128)	0.727 (0.203)	0.540 (0.241)	0.837 (0.111)	0.913 (0.067)	12.00 (0.00)

Table 8. Testing performance for the prostate data; mean (standard deviation) across 25 outer folds

Method	Accuracy	BAcc	G-Mean	MCC	ROC-AUC	PR-AUC	No. vars
PWRCVS-LASSO	0.894 (0.062)	0.894 (0.061)	0.892 (0.062)	0.794 (0.123)	0.946 (0.042)	0.949 (0.045)	14.12 (4.00)
PWRCVS-ALASSO	0.888 (0.060)	0.888 (0.060)	0.886 (0.061)	0.783 (0.118)	0.941 (0.047)	0.942 (0.053)	14.12 (4.00)
PWRCVS-ENET	0.898 (0.056)	0.898 (0.057)	0.895 (0.059)	0.804 (0.110)	0.954 (0.036)	0.961 (0.033)	14.12 (4.00)
PWRCVS-SCAD	0.853 (0.096)	0.852 (0.096)	0.848 (0.100)	0.713 (0.190)	0.896 (0.097)	0.878 (0.113)	14.12 (4.00)
PWRCVS-MCP	0.865 (0.075)	0.865 (0.075)	0.861 (0.078)	0.737 (0.147)	0.909 (0.067)	0.891 (0.086)	14.12 (4.00)
PWRCVS-RF	0.924 (0.050)	0.924 (0.049)	0.922 (0.050)	0.853 (0.097)	0.965 (0.046)	0.972 (0.037)	14.12 (4.00)
PWRCVS-GBOOST	0.917 (0.058)	0.918 (0.058)	0.915 (0.060)	0.841 (0.113)	0.967 (0.040)	0.972 (0.035)	14.12 (4.00)
PWRCVS-XGBOOST	0.911 (0.067)	0.912 (0.067)	0.910 (0.068)	0.828 (0.131)	0.967 (0.041)	0.974 (0.033)	14.12 (4.00)
FS+IG-LASSO	0.922 (0.053)	0.922 (0.053)	0.920 (0.054)	0.849 (0.104)	0.963 (0.046)	0.971 (0.038)	24.68 (1.60)
FS+IG-ALASSO	0.902 (0.065)	0.902 (0.065)	0.901 (0.065)	0.808 (0.131)	0.953 (0.050)	0.962 (0.043)	24.68 (1.60)
FS+IG-ENET	0.921 (0.063)	0.922 (0.063)	0.919 (0.066)	0.850 (0.119)	0.968 (0.043)	0.975 (0.034)	24.68 (1.60)
FS+IG-SCAD	0.882 (0.083)	0.883 (0.083)	0.880 (0.085)	0.772 (0.164)	0.926 (0.075)	0.917 (0.086)	24.68 (1.60)
FS+IG-MCP	0.879 (0.066)	0.879 (0.066)	0.877 (0.067)	0.765 (0.130)	0.932 (0.066)	0.912 (0.093)	24.68 (1.60)
FS-RF	0.928 (0.061)	0.928 (0.060)	0.926 (0.063)	0.861 (0.115)	0.961 (0.051)	0.969 (0.042)	18.00 (0.00)
FS-GBOOST	0.906 (0.058)	0.906 (0.058)	0.903 (0.062)	0.822 (0.107)	0.963 (0.047)	0.969 (0.041)	18.00 (0.00)
FS-XGBOOST	0.904 (0.053)	0.905 (0.053)	0.902 (0.054)	0.815 (0.103)	0.962 (0.045)	0.969 (0.038)	18.00 (0.00)
IG-RF	**0.929 (0.047)**	**0.930 (0.047)**	**0.929 (0.048)**	**0.864 (0.093)**	**0.966 (0.047)**	**0.972 (0.038)**	**18.00 (0.00)**
IG-GBOOST	0.912 (0.065)	0.912 (0.065)	0.909 (0.068)	0.830 (0.126)	0.967 (0.044)	0.974 (0.035)	18.00 (0.00)
IG-XGBOOST	0.917 (0.060)	0.918 (0.060)	0.915 (0.063)	0.842 (0.114)	0.966 (0.045)	0.972 (0.039)	18.00 (0.00)

5.3. Ablation of performance weighting

To isolate the weighting component, we replaced the six within-resample MCC-derived weights by equal weights of $1/6$. The resamples, candidate-set sizes, consensus threshold selected for the full rule, correlation pruning, outer partitions, and LASSO final-classifier family were otherwise unchanged. Across the 25 paired outer folds, dependence was accommodated with the corrected resampled t standard error $s_d\sqrt{1/25 + 1/(5-1)}$. Table 9 shows no evidence that either weighting rule was uniformly superior. For colon, the mean weighted-minus-equal MCC difference was -0.001 (corrected 95% CI: -0.168 to 0.166; $p = 0.992$). For prostate it was -0.021 (corrected 95% CI: -0.083 to 0.040; $p = 0.480$). The mean selected-set sizes were also close. Consequently, performance weighting is interpreted as an adaptive mechanism that reduces the vote of weakly validating paths; on these two datasets it did not guarantee higher predictive accuracy.

Table 9. Ablation of MCC-derived versus equal path weighting with LASSO as the final-classifier family

Dataset	MCC-weighted	Equal-weight	Difference	Corrected 95% CI	Corrected t	p	Variables W/E
Colon	0.627	0.628	-0.001	$[-0.168, 0.166]$	-0.010	0.992	7.76 / 7.32
Prostate	0.794	0.815	-0.021	$[-0.083, 0.040]$	-0.718	0.480	14.12 / 14.04

5.4. Component ablation

Weighting is one of several parts that could produce the shorter lists, so we extended the same paired design to the whole construction. Five contrasts were formed, each removing one component while the outer partitions, resamples, candidate cap, threshold rule and final classifier stayed fixed: MCC weights replaced by equal weights, correlation pruning switched off, and each of the three tree families dropped from the six paths in turn. Table 10 collects them. The largest colon shift came from removing the XGBOOST paths (Δ MCC $= 0.042$) and the largest prostate shift from removing the weighting (-0.021), but every corrected interval contained zero and every Holm-adjusted p equalled 1.000. Mean list length moved little as well, from 7.76 to between 7.32 and 8.04 for colon and from 14.12 to between 13.88 and 14.40 for prostate. No single component therefore carries the result on these data. The one design scale that did change list width was the first-stage candidate cap of Section 5.7, which is consistent with reading the

parsimony as a property of the screening-plus-threshold-plus-pruning combination, to which the weighting step is one contributor among several.

Table 10. Component ablation of PWRCVS across 25 outer folds. Each row removes one component and holds every other rule fixed.

Dataset	Removed component	Full	Reduced	Δ MCC	g_z	Corrected 95% CI; p_{Holm}
Colon	Path weighting	0.627	0.628	−0.001	−0.005	[−0.168, 0.166]; 1.000
	Correlation pruning	0.627	0.656	−0.029	−0.194	[−0.192, 0.133]; 1.000
	RF paths	0.644	0.664	−0.020	−0.200	[−0.129, 0.089]; 1.000
	GBOOST paths	0.578	0.619	−0.041	−0.204	[−0.256, 0.174]; 1.000
	XGBOOST paths	0.663	0.621	0.042	0.254	[−0.135, 0.218]; 1.000
Prostate	Path weighting	0.794	0.815	−0.021	−0.374	[−0.083, 0.040]; 1.000
	Correlation pruning	0.794	0.795	−0.001	−0.011	[−0.095, 0.093]; 1.000
	RF paths	0.853	0.839	0.014	0.190	[−0.065, 0.093]; 1.000
	GBOOST paths	0.841	0.829	0.012	0.138	[−0.082, 0.107]; 1.000
	XGBOOST paths	0.828	0.854	−0.027	−0.389	[−0.100, 0.047]; 1.000

Full is the mean outer-test MCC of the complete rule and Reduced the mean after the component is removed; Δ MCC is full minus reduced. Weighting and pruning are evaluated with the LASSO final family, each path family with its own tree classifier. Confidence intervals and tests use the corrected resampled t standard error, with Holm adjustment across the five contrasts of each dataset. g_z is descriptive because the outer folds overlap.

5.5. *Sensitivity to the consensus threshold*

We repeated the outer-fold evaluation with fixed values of π_{thr} while leaving the resamples, path weights, correlation rule, partitions, and LASSO final-classifier family unchanged. The adaptive rule selected 0.05, 0.10, and 0.20 in 5, 8, and 12 colon folds, respectively, and in 14, 8, and 3 prostate folds. Table 11 shows that the main effect of the threshold was on list width rather than outer-test MCC. No fixed-threshold MCC differed significantly from the training-only adaptive rule under the corrected resampled t comparison. Values from 0.20 through 0.90 produced the same subset in every outer fold because the prespecified $d_{\min}$ safeguard was activated; correlation pruning explains why the colon mean can fall slightly below six. These plateaus therefore document the behaviour of the minimum-list rule. Whether arbitrarily high thresholds would behave alike under a different safeguard remains untested here.

Table 11. Sensitivity of PWRCVS-LASSO to the consensus threshold across 25 outer folds

Dataset	Threshold rule	Chosen folds	Variables	MCC	Δ MCC	Corrected 95% CI; p
Colon	Adaptive	0.05: 5; 0.10: 8; 0.20: 12	7.76 (2.24)	0.627 (0.183)	—	—
Colon	Fixed 0.05	—	11.52 (0.92)	0.631 (0.137)	0.004	[−0.208, 0.217]; 0.966
Colon	Fixed 0.10	—	8.92 (1.53)	0.637 (0.194)	0.010	[−0.118, 0.138]; 0.872
Colon	Fixed 0.20–0.90	—	5.88 (0.44)	0.641 (0.206)	0.014	[−0.118, 0.146]; 0.831
Prostate	Adaptive	0.05: 14; 0.10: 8; 0.20: 3	14.12 (4.00)	0.794 (0.123)	—	—
Prostate	Fixed 0.05	—	17.16 (1.28)	0.801 (0.120)	0.007	[−0.037, 0.051]; 0.742
Prostate	Fixed 0.10	—	11.20 (2.04)	0.796 (0.129)	0.002	[−0.130, 0.134]; 0.973
Prostate	Fixed 0.20–0.90	—	9.00 (0.00)	0.817 (0.137)	0.023	[−0.117, 0.163]; 0.737

Values are mean (standard deviation). Δ MCC is fixed minus adaptive. Confidence intervals and p values use the corrected resampled t standard error. Thresholds 0.20–0.90 yielded identical subsets within every fold because the minimum-list safeguard was reached.

5.6. *Sensitivity to correlation pruning*

We next reconstructed each fold-specific set immediately before correlation pruning and applied $\rho_0 \in \{0.80, 0.85, 0.90, 0.95\}$, together with a no-pruning reference. All other operations and the LASSO final-classifier family were held fixed. Table 12 shows that the lower cutoff of 0.80 pruned more aggressively, especially for colon,

but did not produce a statistically detectable MCC change. Relative to 0.90, the colon mean MCC differences were -0.005, 0.000, and 0.001 at 0.80, 0.85, and 0.95, respectively; the prostate differences were 0.000, 0.000, and 0.000 after rounding. Thus, 0.90 is a conservative redundancy setting rather than a value on which the reported discrimination critically depends in these two datasets.

Table 12. Sensitivity of PWRCVS-LASSO to the correlation-pruning cutoff across 25 outer folds

Dataset	ρ_0	Variables	Changed	Jaccard	MCC	Δ MCC	Corrected 95% CI; p
Colon	0.80	6.12 (2.09)	20	0.795	0.621 (0.177)	-0.005	$[-0.098, 0.088]$; 0.907
Colon	0.85	7.44 (2.00)	6	0.966	0.627 (0.183)	0.000	[0.000, 0.000]; 1.000
Colon	0.90	7.76 (2.24)	0	1.000	0.627 (0.183)	—	—
Colon	0.95	8.00 (2.45)	4	0.976	0.627 (0.183)	0.001	$[-0.003, 0.005]$; 0.714
Colon	No pruning	8.04 (2.51)	4	0.973	0.627 (0.183)	0.001	$[-0.003, 0.005]$; 0.714
Prostate	0.80	13.76 (3.80)	6	0.979	0.794 (0.126)	0.000	$[-0.029, 0.029]$; 0.978
Prostate	0.85	14.12 (4.00)	0	1.000	0.794 (0.123)	0.000	[0.000, 0.000]; 1.000
Prostate	0.90	14.12 (4.00)	0	1.000	0.794 (0.123)	—	—
Prostate	0.95	14.12 (4.00)	0	1.000	0.794 (0.123)	0.000	[0.000, 0.000]; 1.000
Prostate	No pruning	14.12 (4.00)	0	1.000	0.794 (0.123)	0.000	[0.000, 0.000]; 1.000

Values are mean (standard deviation). Changed is the number of folds whose selected set differed from $\rho_0 = 0.90$; Jaccard is the mean set overlap with that reference. Δ MCC is candidate minus 0.90. Confidence intervals and p values use the corrected resampled t standard error.

5.7. Sensitivity to resample count and candidate size

Table 13 separates the effects of two additional design choices. Increasing the resample count generally increased selection stability, from 0.446 at 10 resamples to 0.575 at 50 for colon and from 0.547 to 0.620 for prostate. It did not yield a monotone MCC change. Relative to 30 resamples, the colon MCC differences at 10, 20, and 50 were 0.028, -0.006, and -0.031; the prostate differences were 0.028, 0.040, and 0.019. None was statistically significant (all corrected-test $p \geq 0.467$; all Holm-adjusted $p = 1.000$).

The candidate cap had a direct effect on list width. For colon, increasing the cap from $0.5d_n$ to $2d_n$ increased the mean selected count from 4.56 to 15.12; for prostate it increased from 7.16 to 19.76. Relative to the prespecified d_n rule, no alternative produced a statistically detectable MCC difference (all corrected-test $p \geq 0.395$; all Holm-adjusted $p = 1.000$). Every alternative cap changed the selected set in all 25 folds, which shows that the reported gene lists are conditional on the screening width even when outer-test discrimination is similar. We retain 30 resamples and d_n because they provide an intermediate stability–list-size and computation trade-off, not because these data establish unique optimal values.

5.8. Generalization and the performance-parsimony trade-off

The change in ranking from training to test is visible in Figure 2. PWRCVS-GBOOST ranked first on prostate training MCC, but PWRCVS-RF ranked first on the outer folds. Colon MCC estimates also varied more across folds, as expected with only 62 samples. For that reason, the 0.003 separation between the best colon rows should be read against their standard deviations (0.220 and 0.188); it does not establish a stable ordering.

Figure 3 and Table 14 describe the practical exchange. For colon, reducing the mean list from 17.84 to 7.76 genes changed the best MCC from 0.666 to 0.663. For prostate, reducing 18.00 genes to 14.12 changed it from 0.864 to 0.853. These fold averages are descriptive; none of them shows PWRCVS dominating all alternatives. They identify the situations in which the proposed rule is useful: the analyst is willing to accept a small numerical loss to reduce the laboratory follow-up list.

5.9. Formal paired comparison with conventional paths

Table 15 reports inference for the 11 classifier-matched contrasts. Colon point estimates favored PWRCVS in eight contrasts and conventional selection in three; the corresponding counts for prostate were three and eight.

Table 13. One-factor sensitivity of PWRCVS-LASSO to resample count and first-stage candidate size

Factor	Dataset	Setting	Variables	Stability	Jaccard (changed)	MCC	Δ MCC	Corrected 95% CI; p
Resamples	Colon	10	8.52 (2.84)	0.446	0.619 (23)	0.655 (0.139)	0.028	$[-0.140, 0.197]$; 0.731
	Colon	20	8.00 (2.36)	0.496	0.760 (19)	0.621 (0.207)	-0.006	$[-0.154, 0.142]$; 0.934
	Colon	30	7.76 (2.24)	0.538	1.000 (0)	0.627 (0.183)	—	—
	Colon	50	8.40 (2.47)	0.575	0.749 (20)	0.596 (0.164)	-0.031	$[-0.170, 0.109]$; 0.653
	Prostate	10	12.52 (3.95)	0.547	0.680 (25)	0.822 (0.108)	0.028	$[-0.085, 0.142]$; 0.612
	Prostate	20	12.40 (3.59)	0.569	0.742 (21)	0.833 (0.115)	0.040	$[-0.071, 0.150]$; 0.467
	Prostate	30	14.12 (4.00)	0.613	1.000 (0)	0.794 (0.123)	—	—
	Prostate	50	12.84 (4.05)	0.620	0.789 (21)	0.813 (0.125)	0.019	$[-0.059, 0.097]$; 0.620
Candidates	Colon	$0.5d_n$ (6)	4.56 (1.47)	0.462	0.521 (25)	0.614 (0.157)	-0.013	$[-0.195, 0.169]$; 0.884
	Colon	$1.0d_n$ (12)	7.76 (2.24)	0.538	1.000 (0)	0.627 (0.183)	—	—
	Colon	$1.5d_n$ (18)	11.84 (3.73)	0.560	0.560 (25)	0.664 (0.157)	0.037	$[-0.087, 0.162]$; 0.542
	Colon	$2.0d_n$ (24)	15.12 (4.06)	0.544	0.478 (25)	0.668 (0.159)	0.041	$[-0.128, 0.211]$; 0.620
	Prostate	$0.5d_n$ (9)	7.16 (1.86)	0.701	0.511 (25)	0.823 (0.137)	0.029	$[-0.074, 0.132]$; 0.565
	Prostate	$1.0d_n$ (18)	14.12 (4.00)	0.613	1.000 (0)	0.794 (0.123)	—	—
	Prostate	$1.5d_n$ (27)	17.04 (3.27)	0.651	0.632 (25)	0.813 (0.120)	0.019	$[-0.096, 0.134]$; 0.732
	Prostate	$2.0d_n$ (36)	19.76 (2.45)	0.624	0.602 (25)	0.837 (0.112)	0.044	$[-0.060, 0.148]$; 0.395

Values are mean (standard deviation). Stability is Nogueira's coefficient. Jaccard is mean overlap with the factor-specific reference (30 resamples or $1.0d_n$), followed in parentheses by the number of changed folds. Δ MCC is candidate minus reference. Confidence intervals and p values use the corrected resampled t standard error; all Holm-adjusted p values for the three comparisons within a dataset and factor equaled 1.000.

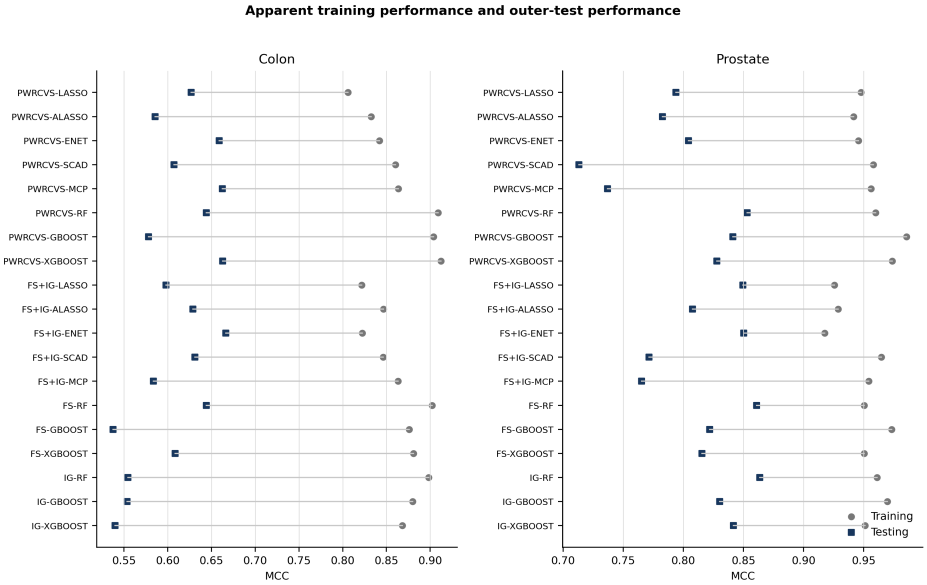


Figure 2. Apparent training MCC and mean outer-test MCC for the 19 evaluated paths. The connecting segments show the direction and magnitude of optimism, not confidence intervals; fold-to-fold standard deviations are reported in Tables 5–8.

Table 14. Best outer-test performance and the MCC-subset-size trade-off

Dataset	Best PWRCVS	MCC	Best conventional	MCC	Variables P/C	List reduction	Matched wins
Colon	PWRCVS-XGBOOST	0.663	FS+IG-ENET	0.666	7.76 / 17.84	56.5%	8/11
Prostate	PWRCVS-RF	0.853	IG-RF	0.864	14.12 / 18.00	21.6%	3/11

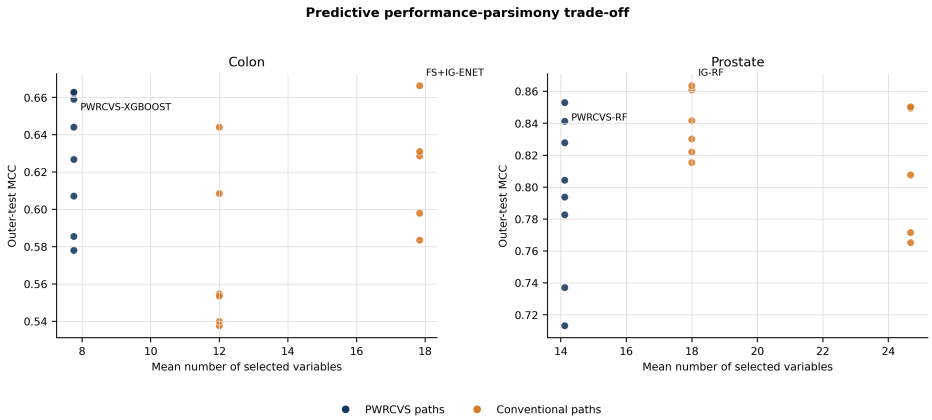


Figure 3. Mean outer-test MCC versus mean selected-set size. Labeled points identify the best mean PWRCVS and conventional rows in each dataset; corrected paired uncertainty is reported in Table 15.

Nevertheless, every corrected 95% confidence interval included zero and no contrast was significant after Holm adjustment (all adjusted $p = 1.000$). The smallest unadjusted p values were 0.380 for colon and 0.213 for prostate.

For the two headline best-mean rows in Table 14, the colon PWRCVS-XGBOOST minus FS+IG-ENET difference was -0.003 (corrected 95% CI: -0.272 to 0.265; $p = 0.979$; $g_z = -0.014$). The prostate PWRCVS-RF minus IG-RF difference was -0.011 (corrected 95% CI: -0.066 to 0.045; $p = 0.699$; $g_z = -0.204$). These headline contrasts quantify the uncertainty around the displayed ranking; because the best rows were identified from the observed fold means, the prespecified classifier-matched contrasts remain the primary formal analysis.

Table 15. Corrected paired MCC comparisons for the 11 prespecified classifier-matched contrasts

	Colon			Prostate		
PWRCVS / conventional path	Δ MCC (corrected 95% CI)	g_z	p_{Holm}	Δ MCC (corrected 95% CI)	g_z	p_{Holm}
LASSO / FS+IG-LASSO	0.029 [−0.212, 0.270]	0.128	1.000	−0.056 [−0.146, 0.034]	−0.667	1.000
ALASSO / FS+IG-ALASSO	−0.043 [−0.298, 0.212]	−0.182	1.000	−0.025 [−0.128, 0.078]	−0.262	1.000
ENET / FS+IG-ENET	−0.007 [−0.248, 0.233]	−0.033	1.000	−0.046 [−0.159, 0.067]	−0.439	1.000
SCAD / FS+IG-SCAD	−0.024 [−0.272, 0.224]	−0.104	1.000	−0.058 [−0.302, 0.185]	−0.258	1.000
MCP / FS+IG-MCP	0.079 [−0.152, 0.310]	0.367	1.000	−0.028 [−0.159, 0.103]	−0.232	1.000
RF / FS-RF	0.000 [−0.226, 0.226]	0.000	1.000	−0.008 [−0.114, 0.098]	−0.080	1.000
RF / IG-RF	0.089 [−0.117, 0.296]	0.466	1.000	−0.011 [−0.066, 0.045]	−0.204	1.000
GBOOST / FS-GBOOST	0.040 [−0.332, 0.413]	0.117	1.000	0.019 [−0.095, 0.134]	0.181	1.000
GBOOST / IG-GBOOST	0.024 [−0.261, 0.310]	0.092	1.000	0.011 [−0.084, 0.106]	0.126	1.000
XGBOOST / FS-XGBOOST	0.054 [−0.233, 0.341]	0.204	1.000	0.012 [−0.089, 0.114]	0.131	1.000
XGBOOST / IG-XGBOOST	0.123 [−0.223, 0.469]	0.382	1.000	−0.014 [−0.105, 0.077]	−0.163	1.000

Each contrast is the fold-matched PWRCVS MCC minus its conventional counterpart. Confidence intervals and two-sided tests use the Nadeau–Bengio corrected resampled t standard error. Holm adjustment covers the 11 contrasts within each dataset. g_z is Hedges' small-sample-corrected standardized paired difference and is descriptive because the outer folds overlap. Every corrected interval contains zero.

5.10. Comparison with established ensemble selectors

The conventional paths of Table 15 are single filter–classifier pipelines, so they leave open how PWRCVS behaves against selectors that also aggregate over perturbed samples. We added three: stability selection in the subsampling form of Meinshausen and Bühlmann [23], bootstrap selection, and recursive feature elimination with an L_2 logistic

base [43]. The first two each draw 100 LASSO supports of size d_n, by stratified half-sampling and by stratified bootstrap, and convert them into a per-gene selection frequency. Every selector then runs on the same 25 outer partitions and through the identical downstream code—the same π_{thr} grid, $d_{\min}$ safeguard, $|\rho| \geq 0.90$ pruning and refit—so only the aggregation rule differs. As in Section 5.3, the final family is fixed at LASSO and at RF.

Table 16 reports the twelve contrasts. None was significant after Holm adjustment, every corrected interval contained zero, and the smallest unadjusted p was 0.287. The point estimates run in both directions: on colon the proposed rule led all three under LASSO and trailed bootstrap selection under RF (Δ MCC $= -0.105$, corrected 95% CI -0.329 to 0.120), while on prostate it trailed all three under LASSO and came within 0.03 under RF. Colon intervals spanning some 0.45 MCC units mean these folds do not separate the four selectors in either direction.

Two design facts remain legible despite that. List length differed on colon, where PWRCVS retained 7.76 genes against 9.16 to 9.48, but not on prostate, where all four lay between 14.00 and 15.68; since the baselines inherit the same threshold and pruning rules, this places the parsimony of Table 14 largely in that shared machinery, in agreement with Table 10. The proposed selector is also the most expensive: measured as in Section 4.3, 83.0 seconds per colon fold and 117.3 per prostate fold against 14.7 and 29.7 for stability selection, factors of 5.6 and 3.9. All three comparators build on one sparse base learner, so the heterogeneity across filters and learners that the weighting exists to arbitrate is not exercised by these two datasets.

Table 16. PWRCVS against established ensemble selectors on the same 25 outer folds, with the final-classifier family held fixed.

Dataset	Family	Selector	MCC	Variables	Δ MCC	g_z	Corrected 95% CI; p_{Holm}
Colon	LASSO	PWRCVS	0.627	7.76	—	—	—
		Stability selection	0.618	9.40	0.009	0.053	[−0.173, 0.191]; 1.000
		Bootstrap selection	0.604	9.16	0.023	0.131	[−0.163, 0.208]; 1.000
		RFE	0.571	9.48	0.056	0.321	[−0.132, 0.244]; 1.000
	RF	PWRCVS	0.644	7.76	—	—	—
		Stability selection	0.694	9.40	−0.050	−0.223	[−0.291, 0.191]; 1.000
		Bootstrap selection	0.749	9.16	−0.105	−0.502	[−0.329, 0.120]; 1.000
		RFE	0.594	9.48	0.050	0.213	[−0.204, 0.305]; 1.000
Prostate	LASSO	PWRCVS	0.794	14.12	—	—	—
		Stability selection	0.831	14.04	−0.037	−0.330	[−0.158, 0.084]; 1.000
		Bootstrap selection	0.845	14.00	−0.052	−0.568	[−0.149, 0.046]; 1.000
		RFE	0.864	15.68	−0.070	−0.491	[−0.223, 0.083]; 1.000
	RF	PWRCVS	0.853	14.12	—	—	—
		Stability selection	0.824	14.04	0.029	0.350	[−0.060, 0.117]; 1.000
		Bootstrap selection	0.851	14.00	0.002	0.023	[−0.094, 0.098]; 1.000
		RFE	0.862	15.68	−0.009	−0.090	[−0.117, 0.099]; 1.000

Δ MCC is PWRCVS minus the comparison selector on the same fold. All four selectors use the identical outer partitions, candidate cap, consensus-threshold grid, minimum-list safeguard, $|\rho| \geq 0.90$ pruning and refitting code; only the way the per-gene selection evidence is produced differs. Stability selection and bootstrap selection draw 100 LASSO supports, by stratified half-sampling and by stratified bootstrap respectively. Confidence intervals use the corrected resampled t standard error with Holm adjustment across the six contrasts of each dataset.

5.11. Stability and recurrent variables

The Nogueira coefficients were 0.538 for colon and 0.613 for prostate (Table 17). In colon, this exceeded the coefficient for the FS+IG-ENET comparator; in prostate, the conventional lists showed greater exact overlap. List length affects this comparison: changing one gene has a larger effect when only a few genes are retained, and two correlated probes may alternate even though they convey similar predictive information.

Three colon identifiers—Hsa.6814, Hsa.37937, and Hsa.8147—were present in 23 of the 25 outer-training selections, and the prostate columns x6185 and x9172 were present in all 25. Recurrence can reflect a strong marginal signal, the correlation structure of the matrix, and preferences shared by the two filters or related tree learners. What it measures is internal algorithmic reproducibility; neither causation nor biomarker validity follows from it.

The legacy colon identifiers have appeared in an independent analysis of the same benchmark matrix: Hsa.6814 was annotated as collagen alpha-2(XI), Hsa.8147 as desmin, and Hsa.37937 with an older non-human myosin-heavy-chain description [44]. Separate proteomic work reported increased stromal desmin in advanced colorectal tumors and its localization around microvessels [45]. That finding supplies biological context for the desmin annotation, but it does not validate the present probe-level recurrence across platform, tissue compartment, or cohort. The supplied prostate matrix retains only positional names such as x6185 and x9172, without accession-level platform annotation; assigning them to named genes would therefore be speculative. We report the identifiers exactly as analyzed and require current platform re-annotation and independent specimen-level validation before any biological or clinical interpretation.

Table 17. Selection size, reduction, and stability

Dataset	Selection path	Variables	Reduction	Jaccard	Nogueira
Colon	PWRCVS	7.76 (2.24)	99.612%	0.380	0.538
Colon	FS-RF	12.00 (0.00)	99.400%	0.384	0.535
Colon	IG-RF	12.00 (0.00)	99.400%	0.283	0.426
Colon	FS+IG-ENET	17.84 (1.40)	99.108%	0.359	0.513
Prostate	PWRCVS	14.12 (4.00)	99.888%	0.445	0.613
Prostate	FS-RF	18.00 (0.00)	99.857%	0.502	0.661
Prostate	IG-RF	18.00 (0.00)	99.857%	0.504	0.666
Prostate	FS+IG-ENET	24.68 (1.60)	99.804%	0.519	0.679

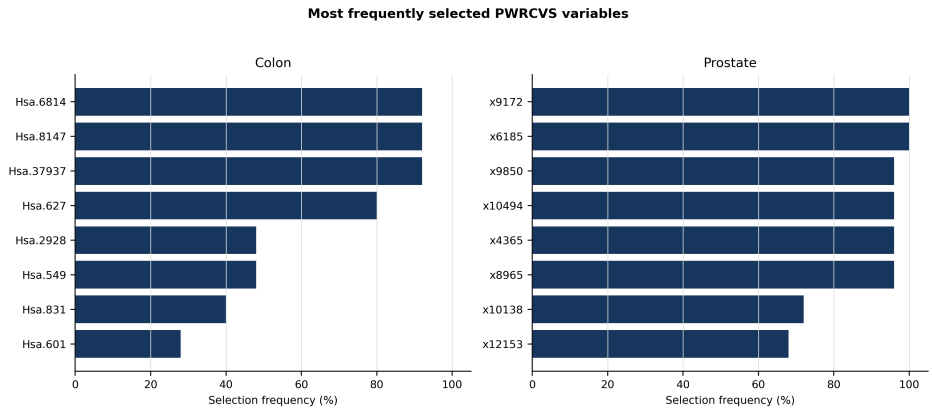


Figure 4. Selection frequency of the most recurrent PWRCVS variables across 25 outer-training selections. Frequency denotes conditional algorithmic recurrence and must not be interpreted as biological validation.

6. Discussion

The principal motivation and result should be read together. In a setting where measuring and validating every nominated gene is costly, a selector may be useful even when it does not improve the best mean predictive score, provided that it shortens the list without a detectable loss and makes the sources of selection variability visible. This is the contribution supported by the two applications; predictive superiority is not.

The colon analysis gives the strongest evidence for the proposed selector, but it also shows why the claim must be qualified. Eight matched MCC comparisons favored PWRCVS, yet the highest individual mean remained the conventional FS+IG-ENET value (0.666). The proposed XGBOOST path was lower by 0.003 and had substantial fold-to-fold variation. What PWRCVS changed most clearly was list width: its mean of 7.76 genes was less than half the 17.84-gene union used by ENET.

Table 18. Most frequently selected PWRCVS variables

Dataset	Variable index	Identifier	Selections	Frequency
Colon	249	Hsa.8147	23	92.0%
Colon	493	Hsa.37937	23	92.0%
Colon	1772	Hsa.6814	23	92.0%
Colon	1671	Hsa.627	20	80.0%
Colon	1042	Hsa.549	12	48.0%
Colon	1582	Hsa.2928	12	48.0%
Prostate	6185	x6185	25	100.0%
Prostate	9172	x9172	25	100.0%
Prostate	4365	x4365	24	96.0%
Prostate	8965	x8965	24	96.0%
Prostate	9850	x9850	24	96.0%
Prostate	10494	x10494	24	96.0%

Prostate results were more concentrated, but they did not reverse the predictive ranking. IG-RF remained the best conventional row at MCC 0.864, and PWRCVS-RF reached 0.853. The proposed set nevertheless reduced the mean number of genes from 18.00 to 14.12. This dataset dependence agrees with recent transcriptomic comparisons in which regularized and ensemble procedures both performed well and no fixed learner won across outcomes [27, 28].

The corrected paired comparisons sharpen this interpretation. They do not support a claim of predictive superiority for PWRCVS, but neither do they show a statistically detectable loss for any classifier-matched contrast after multiplicity correction. The intervals, especially for colon, remain wide because only 62 observations were repeatedly partitioned. Therefore, failure to reject a zero MCC difference must not be read as proof of equivalence. The supported practical result is narrower: the proposed rule often reduced the candidate list substantially while its observed discrimination remained within the uncertainty of the conventional path.

Three operations are responsible for the shorter PWRCVS lists. First, each resample limits both filters to d_n candidates. Second, a weak internal MCC reduces the contribution of the associated list. Third, only the largest-probability member of a $|\rho| \geq 0.90$ component is retained. Because the same frozen set was subsequently supplied to linear and tree classifiers, the reported comparison concerns the portability of the set as well as the performance of its internal tree learners.

The equal-weight ablation further qualifies the interpretation of the second operation. Its outer-test MCC was statistically indistinguishable from the MCC-weighted rule in both datasets. The contribution of the weighting step is therefore methodological—it makes path influence responsive to internal validation—while the present experiments do not establish an unconditional predictive advantage over simple voting.

The six paths are not statistically independent votes. Paths sharing FS or IG begin with overlapping candidate sets, and RF, GBOOST, and XGBOOST all use tree splits. Their dependence can concentrate evidence for variables favored by a filter or by tree-based partitioning. The weights are consequently normalized algorithmic contributions, not inverse-variance weights or independent meta-analytic evidence. Moreover, an MCC measured on a small internal validation portion is noisy and may reward a locally favorable path. Repeated splitting limits reliance on one validation allocation, and the equal-weight ablation exposes the issue, but neither device removes it completely.

The threshold analysis leads to a related qualification. Within the prespecified grid, lower fixed cutoffs retained more genes, but none produced a statistically detectable MCC change relative to the training-only adaptive choice. The threshold should therefore be viewed as a parsimony control whose effective range also depends on the minimum-list safeguard; the present data do not identify one universally optimal numerical cutoff.

The correlation sensitivity analysis likewise shows that the default 0.90 cutoff is not a sharp predictive boundary in these applications. Lowering it to 0.80 shortened the colon list more often, whereas cutoffs from 0.85 to 0.95 had small or no effects; none of these changes produced a statistically detectable MCC difference. This supports using correlation pruning to remove near substitutes while keeping its numerical cutoff configurable for a new dataset.

The remaining one-factor perturbations clarify the role of the first two design scales. Additional resamples mainly improved recurrence stability, whereas the candidate cap primarily controlled final list width and changed the

identity of retained genes. Prediction was non-monotone under both perturbations, and no alternative differed significantly from the submitted settings. Thus, 30 resamples and d_n are practical defaults for these small data rather than universal tuning constants; applications with different sample sizes or assay costs should repeat the same training-only sensitivity checks.

PWRCVS also differs from established stability selection and bootstrap ensemble selection. Classical stability selection repeatedly applies a base sparse selector and thresholds its inclusion probabilities, with error-control results under stated assumptions [23]. Bootstrap and multi-criterion ensembles aggregate selections or ranks across perturbed samples [25]. PWRCVS instead uses subsamples without replacement, crosses two filters with three nonlinear importance generators, and lets internal MCC modify the contribution of each path before correlation pruning.

Section 5.10 measures that distinction. Given the same screening, thresholding and pruning machinery, none of the twelve contrasts against those selectors reached significance after Holm adjustment, and their selection cost was four to six times lower. Colon intervals spanning about 0.45 MCC units mean these data separate the selectors in neither direction, so they establish neither an advantage for the proposed weighting nor its redundancy. What they do support is the framework itself: it returns shorter lists than conventional filter–classifier paths, at a computational cost that should be weighed against that gain. All three comparators build on a single sparse base learner, so the heterogeneity the weighting exists to arbitrate is untested here; examining it requires data on which no single filter–learner pair is already well matched.

The moderate colon imbalance (22 normal and 40 tumor samples) was handled by stratified splitting and training-only class weights, without synthetic oversampling; the nearly balanced prostate data used the same prespecified procedure. MCC and balanced accuracy prevent the majority class from being the sole driver of interpretation. These safeguards address evaluation under the observed class proportions, but they do not make 62 colon observations equivalent to a large or externally representative cohort.

The recurrent variables are candidates for follow-up; they do not constitute a diagnostic panel. Both datasets contain a tumor/normal label but no disease stage, grade, treatment response, or Iraqi-population information. They cannot support a clinical decision rule or replace histopathology. Their immediate use is to reduce the number of assays that would enter a later laboratory and external-cohort validation study.

The absence of a clear predictive gain has plausible methodological causes. Several conventional paths were already strong; pruning deliberately trades some redundant predictive signal for a shorter list; and path weights are estimated from small validation portions. Thus, a more elaborate consensus need not improve discrimination when a single filter–learner pair is already well matched to the dataset. Together with the runtime estimates, this means PWRCVS is most defensible when the cost of laboratory follow-up makes parsimony valuable. When selection cost is low or compute is the dominant constraint, a simpler stability or single-filter procedure may be preferable.

The study has four important limits. The samples are small; both benchmark datasets are old; no independent platform was available for external validation; and probe annotations may have changed since the original releases. Repeated outer cross-validation estimates internal generalization while controlling selection leakage, but it cannot measure transport across cohorts, laboratories, populations, or assay platforms. A proper external validation would lock the discovery-cohort preprocessing, annotation map, selected signature, fitted classifier, and decision threshold before applying them once to an independent cohort with compatible endpoints. Re-selecting genes on that cohort would be a new internal analysis, not external validation.

The algorithm accepts a numerical predictor matrix and could in principle be applied to modern RNA-seq measurements, but count normalization, variance stabilization, low-count filtering, gene mapping, and batch handling would have to be estimated within each training partition. Performance on these legacy microarrays therefore does not establish performance on RNA-seq. Deep and attention-based selectors can model higher-order dependencies [21, 6]; traditional filter–wrapper paths remain attractive here because their parameter burden and runtime are compatible with very small training samples. The comparison between these paradigms remains an empirical question for larger modern cohorts.

The current formulation is binary. A multiclass extension could replace the binary filter statistics and path utility with multiclass mutual information and a generalized multiclass MCC, then use class probabilities with an argmax decision. A continuous-outcome extension would require regression filters, regression learners, and a prespecified

validation utility such as out-of-fold mean squared error or correlation; it is not obtained by simply reusing the present equations. Both extensions require independent evaluation.

Finally, the 25 test results come from repeated partitions of the same observations and are not 25 independent studies. Correlation components are estimated from each training fold because an unseen test correlation matrix is unavailable and using it would leak information; a distributional shift could nevertheless make those relationships different in deployment. A next study should use multicenter samples, add clinical covariates, assess pathway-level rather than probe-level recurrence, and publish complete selection lists with executable code.

7. Conclusions

Across the 25 outer folds, PWRCVS reduced the original colon and prostate spaces by more than 99.6%. Its strongest colon result used 56.5% fewer variables than the best conventional row, with MCC lower by 0.003. Its strongest prostate result used 21.6% fewer variables, with MCC lower by 0.011.

None of the 11 classifier-matched MCC contrasts was significant after Holm adjustment, although the corrected intervals were too wide to establish equivalence. These applications support a specific conclusion. PWRCVS is a compact selector that can preserve nearly the same discrimination as the best conventional path in the two benchmark datasets examined here. They do not show that performance weighting is universally superior. Additional datasets and independent platforms are needed before making a broader claim about prediction or gene-list stability.

Declarations

Funding. This research received no external funding.

Conflict of interest. The authors declare no conflict of interest.

Data availability. The colon and prostate datasets are published research benchmarks. The analysis code, pinned software requirements, fold-level outputs, and complete selection lists are provided with the revision as a supplementary reproducibility package.

Ethics. Only public de-identified data were used; no new ethical approval or patient consent was required.

Author contributions. Noor Abdulmunim contributed data organization, implementation, and the initial draft. Omar Q. Alshebly contributed methodology, supervision, validation, and critical revision. Both authors approved the final manuscript.

Thesis statement. This article is derived from the first author's master's thesis and has been reorganized and rewritten as a focused independent study.

AI-assisted language editing. OpenAI's ChatGPT was used to support English-language editing and LaTeX organization. The authors checked every numerical result, equation, citation, and interpretation against the analysis files and accept full responsibility for the submitted text.

REFERENCES

1. Uri Alon, Naama Barkai, Daniel A. Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999. https://doi.org/10.1073/pnas.96.12.6745.
2. Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, and William R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002. https://doi.org/10.1016/S1535-6108(02)00030-2.
3. Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. https://doi.org/10.1017/9781108627771.
4. Hanaa Elgohari, Mohamed Zakariaa Fouda, Mohamed Asim, and Omar M. Elzeki. A statistical evaluation of feature selection methods for breast cancer classification: A simulation and real-data study. *Statistics, Optimization & Information Computing*, 16(3):2826–2845, 2026. https://doi.org/10.19139/soic-2310-5070-4378.
5. Zainab Asaad Jasim and Emad Hazim Aboudi. The factor-augmented regression model (FARM) for solving the problem of high-dimensional data. *Statistics, Optimization & Information Computing*, 16(2):1033–1044, 2026. https://doi.org/10.19139/soic-2310-5070-3767.

6. Bajes Zeyad Aljunaeidi, Mohammed Tawfik, Issa M. Alsmadi, and Yasser Mohammad Al-Sharo. Cross-attention transformer networks with optimized feature selection for explainable respiratory disease classification. *Statistics, Optimization & Information Computing*, 15(5):3964–3985, 2026. https://doi.org/10.19139/soic-2310-5070-2912.
7. Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, 2002. https://doi.org/10.1073/pnas.102102699.
8. Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. https://doi.org/10.1186/1471-2105-7-91.
9. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
10. Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911, 2008. https://doi.org/10.1111/j.1467-9868.2008.00674.x.
11. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
12. Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. https://doi.org/10.1198/016214506000000735.
13. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
14. Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. https://doi.org/10.1198/016214501753382273.
15. Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. https://doi.org/10.1214/09-AOS729.
16. Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232–253, 2011. https://doi.org/10.1214/10-AOAS388.
17. Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. https://doi.org/10.1023/A:1010933404324.
18. Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. https://doi.org/10.1214/aos/1013203451.
19. Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002. https://doi.org/10.1016/S0167-9473(01)00065-2.
20. Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. https://doi.org/10.1145/2939672.2939785.
21. Weidong Xie, Shoujia Zhang, Linjie Wang, Kun Yu, and Wei Li. Feature selection of microarray data using multidimensional graph neural network and supernode hierarchical clustering. *Artificial Intelligence Review*, 57:63, 2024.
22. Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018.
23. Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010. https://doi.org/10.1111/j.1467-9868.2010.00740.x.
24. Vincent Hamer and Pierre Dupont. An importance weighted feature selection stability measure. *Journal of Machine Learning Research*, 22(116):1–57, 2021.
25. A. Wang, H. Liu, J. Yang, and G. Chen. Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. *Computers in Biology and Medicine*, 142:105208, 2022. https://doi.org/10.1016/j.compbiomed.2021.105208.
26. H. Shin and S. Oh. An effective heuristic for developing hybrid feature selection in high dimensional and low sample size datasets. *BMC Bioinformatics*, 25:390, 2024. https://doi.org/10.1186/s12859-024-06017-9.
27. E. Cantor, S. Guauque-Olarte, R. Leon, S. Chabert, and R. Salas. Knowledge-slanted random forest method for high-dimensional data and small sample size with a feature selection application for gene expression data. *BioData Mining*, 17:34, 2024. https://doi.org/10.1186/s13040-024-00388-8.
28. S. R. Oshternian, S. Loipfinger, A. Bhattacharya, and R. S. N. Fehrmann. Exploring combinations of dimensionality reduction, transfer learning, and regularization methods for predicting binary phenotypes with transcriptomic data. *BMC Bioinformatics*, 25:167, 2024. https://doi.org/10.1186/s12859-024-05795-6.
29. Omar Q. Alshebly and Suhail N. Abdullah. Gene selection and cancer type prediction via filtering and weighted regularized logistic regression. *Journal of System and Management Sciences*, 14(6):359–379, 2024. https://doi.org/10.33168/JSMS.2024.0623.
30. Omar Q. Alshebly and Suhail N. Abdullah. Proposed two-steps procedure of classification high dimensional data with regularized logistic regression. *Statistics, Optimization & Information Computing*, 12(2):325–342, 2024. https://doi.org/10.19139/soic-2310-5070-1846.
31. Zakariya Y. Algamal and Muhammad Hisyam Lee. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42:9326–9332, 2015. https://doi.org/10.1016/j.eswa.2015.08.016.
32. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. https://doi.org/10.18637/jss.v033.i01.
33. Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007. https://doi.org/10.1186/1471-2105-8-25.
34. Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6, 2020. https://doi.org/10.1186/s12864-019-6413-7.
35. Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14:13, 2021.
36. Davide Chicco and Giuseppe Jurman. The matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16:4, 2023.
37. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

38. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. https://doi.org/10.1038/s41586-020-2649-2.
39. Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019.
40. Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006. https://doi.org/10.1145/1143844.1143874.
41. Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52:239–281, 2003. https://doi.org/10.1023/A:1024068626366.
42. Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. https://www.jstor.org/stable/4615733.
43. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002. https://doi.org/10.1023/A:1012487302797.
44. Shuangge Ma, Xiao Song, and Jian Huang. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8:60, 2007.
45. Georgia Arentz, Tim Chataway, Timothy J. Price, et al. Desmin expression in colorectal cancer stroma correlates with advanced stage disease and marks angiogenic microvessels. *Clinical Proteomics*, 8:16, 2011.