International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 49 – 57
Received: 26 March 2026 ; *Accepted:* 03 November 2026 ; *Published:* 16 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P5
https://www.youtube.com/watch?v=pZc5QIrcrxA
https://archive.org/details/jsp-ijrdes-v8ri6p5
Original Paper : Applied Research – Experimental Research Paper

Closed-Form Characterization of Selection Bias in Cross-Validation for Feature-Selected Predictive Models using Machine Learning

Shabbeer Shaik

Department of Computing and Information Sciences, University of Technology and Applied Sciences, Muscat, Sultanate of Oman.

* *Corresponding Author : Shabbeer Shaik ;* *shabbeer.shaik@utas.edu.om*

Abstract: Choosing the most promising features on all the data and then cross-validating a classifier on them is a well-known mistake, but its size is usually shown by example, not predicted. We derive the expected accuracy reported by such leaky cross-validation when the features are pure noise. The selected features have effect sizes given by order statistics of $|N(0, 1)|$, which gives an explicit formula for the reported accuracy in terms of the sample size n, the number of features p, the number k kept and the number of folds K, and a closed form for the sample size at which the optimism falls below a chosen level. Across 27 simulated settings the formula was within 0.025 of the simulated value (mean error 0.007), while the simpler re substitution formula erred by up to 0.095. With 100 cases, 1,000 noise features and 5 kept, leaky cross-validation reported 74% accuracy for a classifier with true accuracy 50%; selection inside each fold reported 49%. Optimism fell as $1/\sqrt{n}$: bringing it below 5 points needed about 2,476 cases for 5 features of 1,000. With ten weak informative features among 10,000 and 50 cases, leaky cross-validation reported 98% against a true 52%. On the breast cancer and digits datasets, leakage was harmless when the signal was strong, but with permuted labels and 1,000 added noise features leaky cross-validation reported 80% and 81% against a true 50%. The price of the honest procedure is variance: with 50 cases the standard deviation of nested estimates was 4.6 times that of the leaky ones. Synthetic results are simulations; the real data are two small benchmark sets with added noise.

Keywords: Cross-Validation, Feature Selection, Selection Bias, Optimism, Nested Cross-Validation, Order Statistics.

1. Introduction

Many classifiers are built for data with far more features than cases: gene expression, imaging, text. A common first step is to keep the features that look most related to the label, and then to estimate accuracy by cross-validation on the reduced data. If the screening used all the cases, the cross-validation is not honest, because every held-out case has already helped decide which features are kept. This is described in textbooks [5] and in studies of gene-expression classifiers [1, 2], and its consequences for model selection have been measured [3, 4, 17]. What is usually missing is a prediction of how large the optimism will be before the data are collected.

We give one, in the simplest setting where the answer is exact enough to be useful: Gaussian features, a nearest-centroid classifier, and features that are unrelated to the label. The key fact is that the best of p noise features looks as informative as an order statistic of $|N(0, 1)|$ allows [11], and that a classifier fitted on those features reproduces the apparent signal on held-out cases whose labels were used to find it. We make five contributions.

- An explicit formula for the accuracy reported by leaky K-fold cross-validation on pure noise, as a function of n, p, k and K (Section 4).
- A closed form for the sample size at which the optimism falls below a chosen level, and the observation that optimism decreases as $1/\sqrt{n}$, so that halving it needs four times the data.
- A check of the formulas in 27 simulated settings and a measurement of the rate at which they break down.
- A comparison with nested cross-validation, including the variance it costs, and a case with weak real signal.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

- A real-data study on the breast cancer and digits datasets with added noise features and permuted labels.

All synthetic experiments are simulations from known models. The real-data experiment adds noise features to two benchmark datasets and, in a second version, permutes the training labels.

2. Background and Related Work

Selection bias. Ambroise and McLachlan showed that selecting genes on all the data before cross-validation gave strongly optimistic error rates, and that the bias was removed when selection was done inside each fold [1]. Simon and colleagues documented the same pattern in published microarray studies [2]. Varma and Simon and Cawley and Talbot measured the bias introduced by using the same cross-validation for tuning and for assessment, and recommended nested procedures [3, 4]. Krstajic and colleagues reviewed pitfalls for regression and classification [17], and Hastie, Tibshirani and Friedman give the standard account [5].

Cross-validation. The bias of K-fold cross-validation as an estimate of generalisation error and its dependence on K have been studied since Stone [13] and Kohavi [6]. Bengio and Grandvalet showed that there is no unbiased estimator of its variance [7], and Varoquaux showed how large the error bars are at the sample sizes common in neuroimaging [8]. Feature screening itself is a legitimate tool, with theory for when it keeps the true features [10], and a survey of the field is given by Guyon and Elisseeff [9].

What we add. The phenomenon is known. We contribute an explicit prediction of its size from order statistics, a closed form for the sample size needed to make it small, a check of both, and a measurement of what the honest alternative costs in variance.

3. Research Methodology and Gap

3.1. Data and classifier

There are n cases with balanced classes ($n/2$ each) and p features. Features are independent $N(0, 1)$ except for s informative features whose class means differ by δ. The classifier is nearest-centroid with unit variance: it predicts class 1 when $(x - (m_1 + m_0)/2)^T(m_1 - m_0) > 0$, where m_c are the class means of the selected features. Selection keeps the k features with the largest $|d_j|$, where $d_j = m_{1,j} - m_{0,j}$. For pure noise, $d_j \sim N(0, 4/n)$.

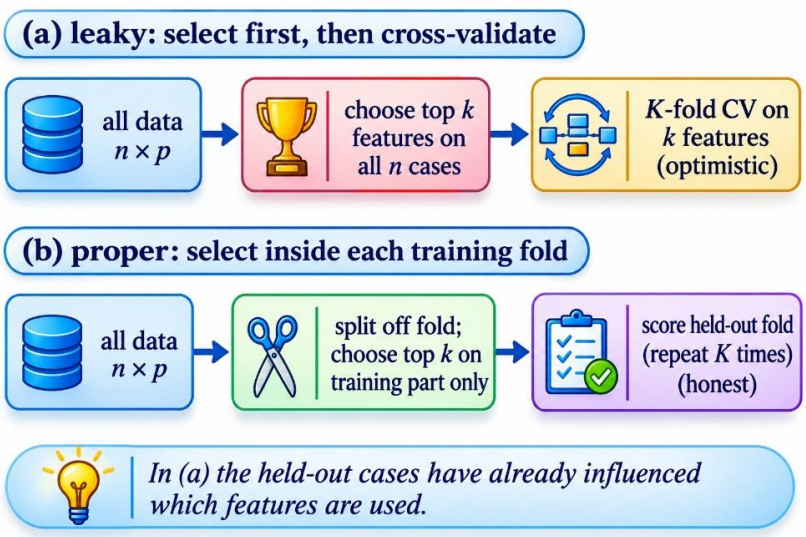

Figure 1: The binormal model. (a) Score densities for negatives and for positives at three AUC values. (b) The corresponding ROC curves. The AUC is the probability that a random positive scores above a random negative.

3.2. Two estimators of accuracy

Leaky CV: the k features are selected once on all n cases, and K-fold cross-validation ($K = 5$) is run on the reduced data.

Nested CV: in each fold the features are selected on the training part only, the classifier is fitted on the training part, and the held-out fold is scored. The reference is the **test accuracy**: the classifier fitted on all n cases with the features selected on all n cases, scored on fresh data. In the pure-noise case the true accuracy is 0.5 for every procedure.

4. Proposed System

Proposition 1 (selected effect sizes). *Let $z_{(1)} \geq \ldots \geq z_{(k)}$ be the k largest of p independent $|N(0, 1)|$ values. The selected differences are $d_{(i)} = (2/\sqrt{n})\, z_{(i)}$, and $z_{(i)}$ is approximated by*

$$z_i = \Phi^{-1}(\,1 - (i - 0.375) / (2(p + 0.25))\,), \qquad (1)$$

the Blom-type approximation to the expected order statistic [11, 12]. For $k = 1$ and $p = 1{,}000$, $z_1 \approx 3.3$.

Proposition 2 (resubstitution). *A nearest-centroid classifier scored on the data it was fitted to has accuracy about $\Phi(D/2)$, where*

$$D^2 = (4/n)\, \Sigma_{i=1}^{k}\, z_i^2. \qquad (2)$$

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

$$\hat{A}_{\text{resub}} \approx \Phi(D/2). \quad (3)$$

This follows because the class-conditional spread along the fitted direction is about 1 and the separation of the sample class means is D.

Proposition 3 (leaky K-fold). *Under leaky K-fold cross-validation the reported accuracy is about $\Phi(S/2)$, with*

$$S = (2/\sqrt{n}) \cdot \Sigma_i(z_i^2 - 1) \,/\, \sqrt{(\,\Sigma_i(z_i^2 + 1/(K-1))\,)}, \quad \hat{A}_{\text{leak}} \approx \Phi(S/2). \quad (4)$$

Proof. Condition on a selected feature's full-sample difference $a = d_{\text{full}}$. Since $d_{\text{full}} = (1 - 1/K)d_{\text{train}} + (1/K)d_{\text{test}}$, one finds $E[d_{\text{train}} \mid a] = a$ with conditional variance $4/(n(K-1))$, and $d_{\text{test}} = K\,a - (K-1)d_{\text{train}}$. The held-out separation along the fitted direction $w = d_{\text{train}}$ is therefore $E[w \cdot d_{\text{test}}] = \Sigma(a_j^2 - 4/n)$, while $\|w\|^2 \approx \Sigma(a_j^2 + 4/(n(K-1)))$. Substituting $a_j^2 \approx (4/n)z_j^2$ gives S.

Eq. (4) subtracts the part of the selection effect that is noise and keeps the part that is leaked, which is why it is lower than the resubstitution value (Eq. 3) and closer to the cross-validated one.

Corollary 1 (sample size). *The optimism $\hat{A}_{\text{leak}} - 1/2$ is at most ε when*

$$n \geq [\,\Sigma_i(z_i^2 - 1) \,/\, (\,\Phi^{-1}(\tfrac{1}{2} + \varepsilon) \cdot \sqrt{(\Sigma_i(z_i^2 + 1/(K-1)))}\,)\,]^2. \quad (5)$$

Since S is proportional to $1/\sqrt{n}$, the optimism for small values is proportional to $1/\sqrt{n}$, and halving it requires four times as many cases. The requirement grows with $\ln p$ and with the number of features kept.

Proposition 4 (nested CV on noise). *If labels are independent of the features and classes are balanced, nested cross-validation with any selection rule that uses only training labels has expected accuracy exactly 1/2.*

Proof. The prediction for a held-out case depends on its features and on the training data, and is independent of its own label. The label is equally likely to be either class, so the probability of a correct prediction is 1/2.

With real signal, nested cross-validation is not unbiased: it estimates the accuracy of a procedure trained on $n(1 - 1/K)$ cases, which is slightly lower than the accuracy of the final classifier.

5. Experimental Setup and Implementation

Synthetic experiments use the binormal model with $\theta \in$ The formula check uses pure noise with $n \in \{40, 100, 300\}$, $p \in \{100, 1{,}000, 10{,}000\}$ and $k \in \{1, 5, 20\}$, with 300 repetitions (200 for p = 10,000) of leaky and nested 5-fold CV and a fresh test set of 2,000 cases. The sample-size study uses p = 10,000 and n from 50 to 1,600 with 120 repetitions, and p = 1,000 with n from 100 to 6,400 with 100 repetitions (60 for the two largest n). The signal study has s = 10 informative features with δ = 0.5 among p = 10,000, k = 20, n from 50 to 800, 150 repetitions and a test set of 4,000 cases. Equation (1) is evaluated in closed form; no simulation is used for the formulas. Real data are the Wisconsin breast cancer set (569 cases) [15] and digits 3 against 8 (352 cases) [16], from scikit-learn [14]. Features are standardised. To each dataset we append $p_{\text{noise}} \in \{0, 100, 1{,}000, 5{,}000\}$ independent $N(0, 1)$ features. In each repetition a balanced subsample of $n \in \{40, 80, 160\}$ cases is the training set and all other cases are the test set (scored by balanced accuracy). Features are selected by the Welch t-statistic (k = 10) and a logistic regression (C = 1) is fitted. We run each setting with true labels and with the training labels permuted, so that the true accuracy is 0.5. There are 100 repetitions per setting.

6. Results and Analysis

6.1. Does the formula predict the optimism?

Table 1 compares the exact variance of Eq. (2) with the Hanley–McNeil approximation and with simulation. The exact formula matches simulation within 3.3%, which is consistent with the sampling error of a variance estimated from 20,000 repetitions. The Hanley–McNeil approximation is accurate at AUC 0.7 but overstates the variance with the AUC (Figure 2): by 12% at AUC 0.9 with 20 and 20 cases, and by a factor of 3.1 at AUC 0.98 with 20 positives and 380 negatives. A Wald interval that uses it is therefore too wide in that regime, which we see as over-coverage in Section 6.4

Table 1: Pure noise, n = 100: accuracy (%) reported by leaky CV, the two formulas, nested CV and a fresh test set.

p	k	Leaky CV	Eq. (4)	Eq. (3)	Nested CV	Test
100	1	60.9	59.2	60.8	50.6	50.1
100	5	66.6	65.9	69.7	49.9	50.0
100	20	70.6	70.0	78.7	50.4	50.0
1000	1	63.5	62.2	63.4	50.0	50.0
1000	5	73.5	72.8	75.4	49.1	50.0
1000	20	84.3	84.3	88.4	49.1	50.1
10000	1	65.7	64.5	65.6	50.6	50.0
10000	5	78.1	77.7	79.6	49.7	50.1
10000	20	91.2	91.3	93.4	49.1	50.1

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

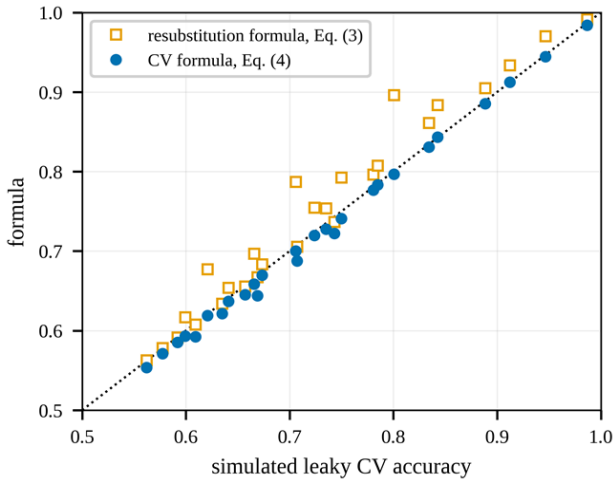

Figure 2: Simulated leaky cross-validation accuracy against the two formulas for 27 settings ($n \in \{40, 100, 300\}$, $p \in \{100, 1{,}000, 10{,}000\}$, $k \in \{1, 5, 20\}$). The dotted line is equality.

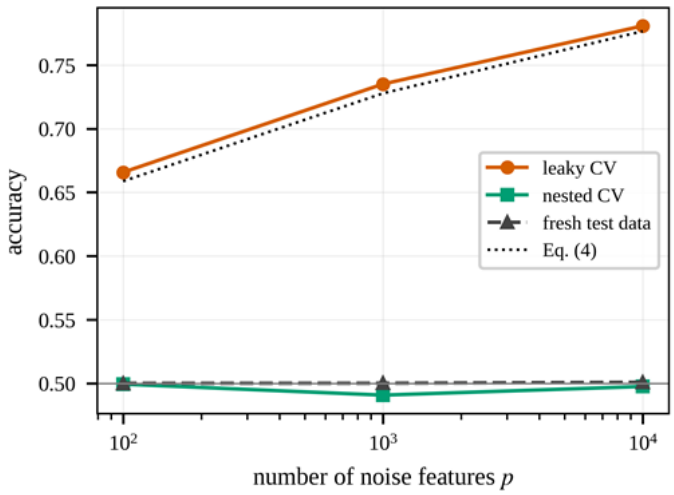

Figure 3: Pure-noise data, $n = 100$, $k = 5$. Leaky CV rises with the number of noise features, while nested CV and a fresh test set stay at 0.5. Dotted: Eq. (4).

6.2. How the optimism shrinks with n

Figure 4 shows the optimism as a function of n with $p = 10{,}000$. For $k = 20$ it is 47 points at $n = 50$, 42 at 100, 25 at 400 and 13 at 1,600, in good agreement with the formula (47, 41, 25 and 13). On the log scale the curves have slope close to −1/2: each fourfold increase in n halves the optimism. Even with 1,600 cases and 20 features chosen from 10,000, leaky cross-validation reports an accuracy 13 points too high.

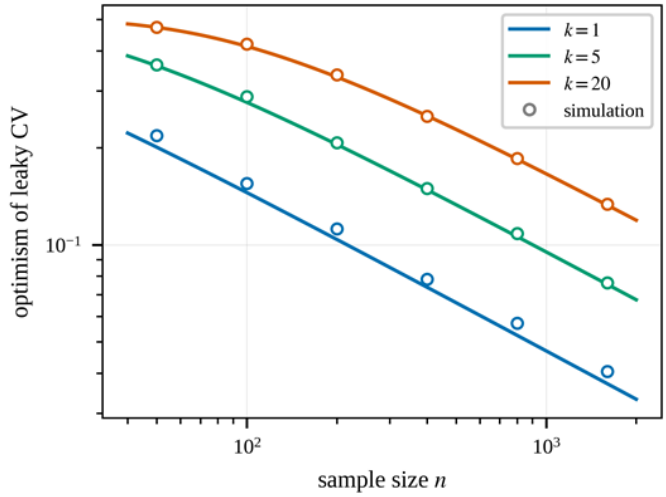

Figure 4: Optimism of leaky CV against sample size at $p = 10{,}000$ (pure noise, 120 repetitions). Lines: Eq. (4) minus 0.5. Circles: simulation.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

Table 2 gives the sample size at which the optimism falls below 5 or 2 points for $p = 1{,}000$ (Eq. 5), with the simulated value interpolated from a grid up to $n = 6{,}400$. The formula is within 24% of the simulation where the simulation reaches the target, and is always below it, so it is a lower bound in practice. For one feature, 607 cases are needed for an optimism of at most 5 points; for five features 2,326; for twenty features about 6,430. Figure 5 shows the requirement for other k.

Table 2: Sample size for optimism of leaky CV at most ε (pure noise, $p = 1{,}000$, $K = 5$): Eq. (5) and simulation. ">6,400" means the target was not reached on the simulated grid.

ε	k	Eq. (5)	Simulated
0.05	1	607	797
0.05	5	2,326	2,476
0.05	20	6,430	>6,400
0.02	1	3,809	4,922
0.02	5	14,600	>6,400

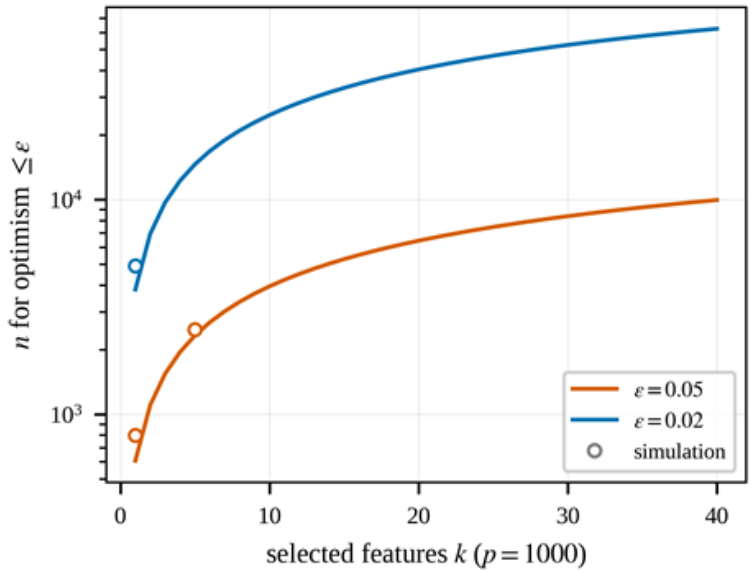

Figure 5: Sample size needed to bring the optimism of leaky CV below ε (Eq. 6), pure noise, $p = 1{,}000$. Circles: simulation, interpolated on a grid of n up to 6,400.

6.3. With real signal

Pure noise is the extreme case. Figure 6 and Table 3 add ten informative features, each shifted by 0.5, to 9,990 noise features ($k = 20$). The best accuracy attainable with the ten true features is 79%. With 50 cases the classifier selected on all data has a true accuracy of 52%, nearly chance, while leaky cross-validation reports 98%. At 200 cases the true accuracy is 65% and leaky CV reports 86%; at 800 cases the gap is 5 points. Nested cross-validation is close to the truth, and slightly below it at every size, by 3 points at 200 cases and 1 at 800, as expected for a procedure trained on 80% of the cases.

Table 3: Ten informative features among 10,000, $k = 20$: accuracy (%) from leaky CV, nested CV and fresh test data, and the standard deviation (points) of each estimate.

n	Leaky CV	Nested CV	Test	SD leaky	SD nested
50	97.6	50.0	52.0	1.9	8.7
100	92.2	53.4	56.5	2.2	6.7
200	86.2	62.3	65.4	2.2	5.1
400	82.1	71.1	73.1	1.8	3.2
800	80.5	75.2	75.9	1.4	1.8

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

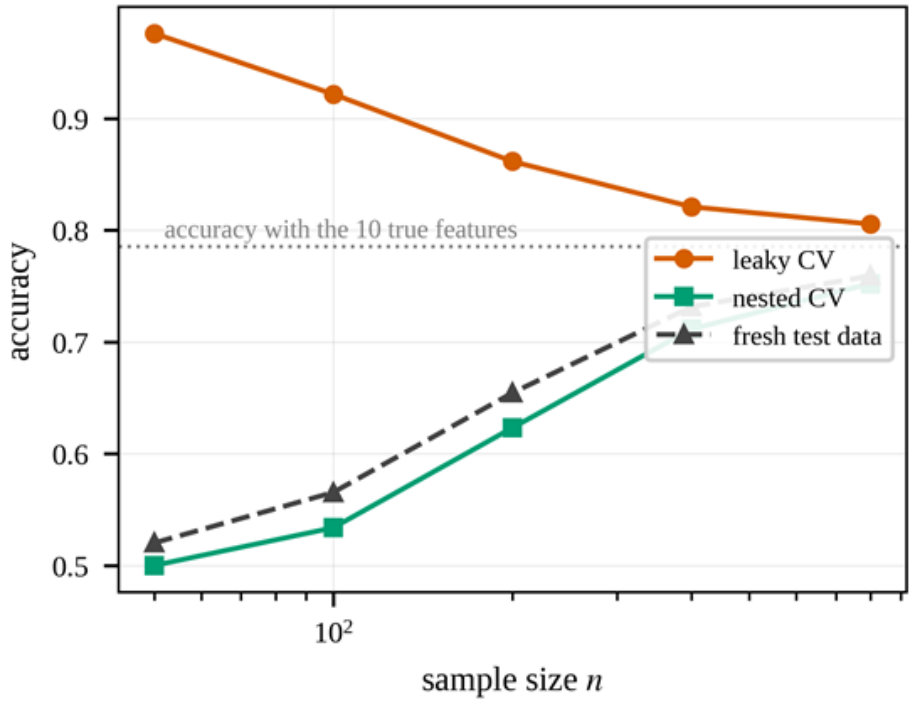

Figure 6: Ten informative features (shift 0.5 each) among 10,000, $k = 20$ selected (150 repetitions). Leaky CV is far above the test accuracy at small n; nested CV is slightly below it.

6.4. The price of honesty

Nested cross-validation is not free. Its standard deviation is larger than that of leaky cross-validation: 8.7 points against 1.9 at n = 50 in the signal study, a ratio of 4.6. The leaky estimate is more stable because it uses the same features in every fold, and stable here means stably wrong. In the pure-noise study at p = 1,000 and k = 5, the ratio of the nested to the leaky standard deviation is 2.0 at n = 40 (Figure 7). A nested estimate at small n has error bars of several points, which is itself the truthful message.

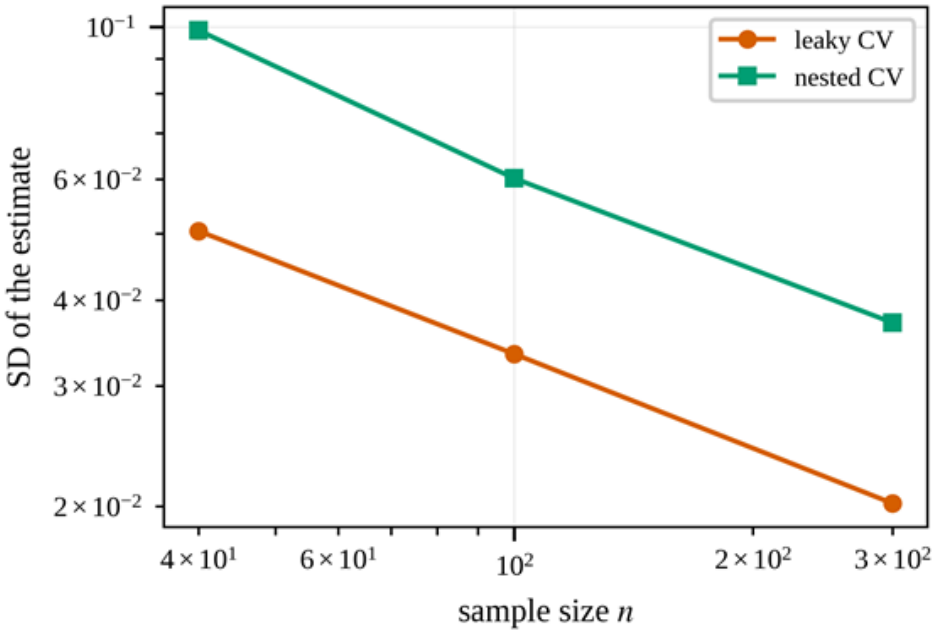

Figure 7: Standard deviation of the two estimates across repetitions (pure noise, $p = 1{,}000$, $k = 5$).

6.5. Real data

Table 4 and Figure 8 give the real-data results at $n = 80$. With true labels, leakage is harmless on these two datasets: leaky, nested and test accuracy agree within 0.5 points, even with 5,000 noise features added, because the real signal is strong and the selection finds the real features first. With permuted labels the picture is the one predicted. At 1,000 added noise features leaky cross-validation reports 80% on breast cancer and 81% on digits, and at 5,000 it reports 85% and 85%, while the test accuracy is 50% and nested CV 48% to 48%. Figure 9 shows that the leakage with true labels does not appear at 1,000 noise features for any of the training sizes.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

Table 4: Real data, $n = 80$: accuracy (%) from leaky CV, nested CV and held-out test data, with true and permuted training labels.

Data	Noise features	Labels	Leaky CV	Nested CV	Test
breast	0	true	93.8	93.5	93.5
breast	0	permuted	51.0	48.8	49.4
breast	1000	true	93.4	93.1	93.6
breast	1000	permuted	79.7	47.7	49.9
breast	5000	true	93.7	93.5	93.5
breast	5000	permuted	84.7	48.3	50.0
digits	0	true	96.4	95.7	96.2
digits	0	permuted	58.0	48.1	50.8
digits	1000	true	96.6	95.9	96.1
digits	1000	permuted	80.9	49.6	49.2
digits	5000	true	96.5	95.6	96.2
digits	5000	permuted	84.8	48.0	50.3

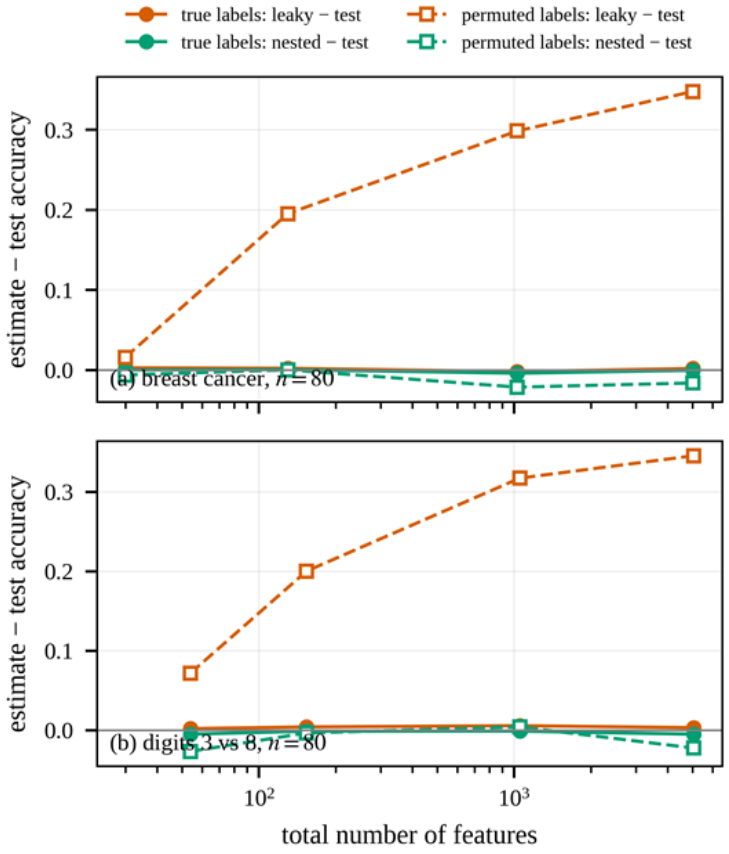

Figure 8: Real data with added noise features ($n = 80$, 100 repetitions). Estimate minus held-out test accuracy for leaky and nested CV, with true and permuted labels.

Figure 9: Real data with 1,000 added noise features and true labels: accuracy estimates against the number of training cases.

6.6. The number of folds

Eq. (4) depends on the number of folds K only through the term $1/(K - 1)$, which matters little. Figure 10 shows leaky CV accuracy on noise for K from 2 to 20, with $n = 100$ or 40, $p = 1{,}000$ and $k = 5$ or 20. The formula is within 0.012 of the simulation in all 20 settings. Moving from 2 folds to 20 raises the reported accuracy by 1.9 points for $n = 100$ and $k = 20$, and by 1.0 points for $k = 5$. More folds increase the optimism slightly, because a larger training part reproduces more of the leaked effect, but the choice of K cannot repair a leaky pipeline: the leak is in the selection, not in the splitting.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

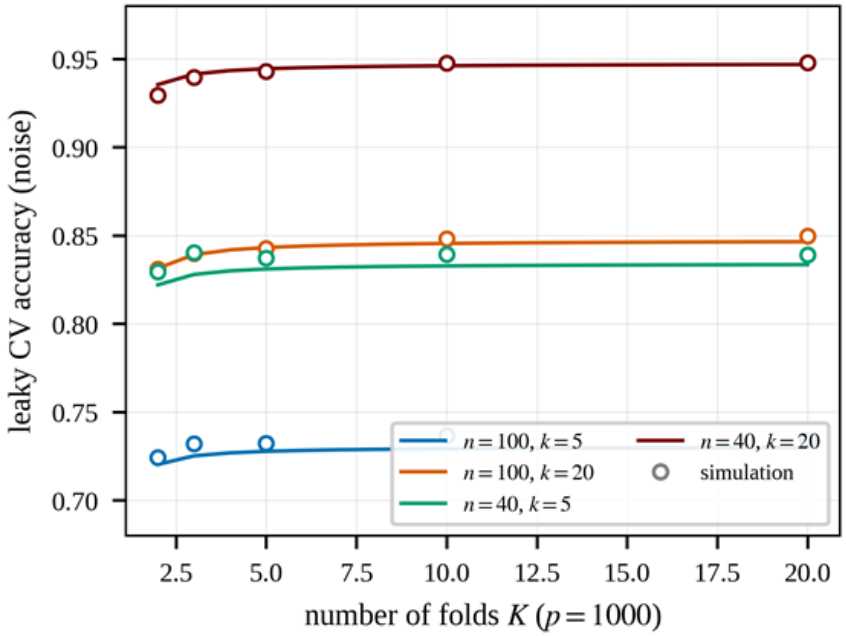

Figure 10: Leaky CV accuracy on pure noise as the number of folds grows (p = 1,000, 400 repetitions). Lines: Eq. (4). Circles: simulation.

6.7. A planning table

Table 5 evaluates the optimism of leaky cross-validation from Eq. (4) for k = 5 and K = 5 over a grid of sample sizes and numbers of noise features, in percentage points. A study with 100 cases and 10,000 candidate features would see 28 points of apparent accuracy from noise alone, and one with 1,000 cases and 100 features 5 points. Because the optimism depends on the number of features only through its logarithm, going from 100 to 100,000 features multiplies the squared effect sizes by about 2.7 for the best feature, which is a smaller change than going from 100 to 400 cases.

Table 5: Optimism (percentage points) of leaky CV on pure noise from Eq. (4), k = 5, K = 5.

n	p = 100	p = 1,000	p = 10,000	p = 100,000
50	22	30	36	40
100	16	23	28	31
200	11	17	20	24
500	7	11	13	15
1000	5	8	10	11
2000	4	5	7	8

6.8. A practical recipe

The results suggest a short procedure. First, never choose features on all the cases when the accuracy estimate will be reported: repeat the selection inside each fold, or hold out a test set before looking at the labels. Second, if you want to know how much damage a leaky pipeline has done, use Eq. (4) or Table 5: for n cases, p features and k kept, the formula gives the accuracy that leaky cross-validation would report on noise. If your reported accuracy is close to it, the result carries little information. Third, use Eq. (5) to plan sample size, remembering that halving the optimism takes four times the cases and that more features kept or screened makes it worse. Fourth, expect the honest estimate to be noisy at small n and report its standard deviation. Fifth, a permutation test with the whole pipeline inside it, as in the real-data study, is a direct way to see the optimism for a particular dataset.

7. Discussion and Limitations

What we think carries over. Three points seem general. The optimism of leaky selection is set by how large the best of many noise statistics is, which grows only with ln p, so more noise features hurt slowly and more cases help slowly. The optimism shrinks as $1/\sqrt{n}$, so reducing it by half needs four times the data. And leakage is most damaging exactly when the true signal is weak, which is when an accuracy estimate matters most.

Synthetic setting. The formulas assume independent Gaussian features, balanced classes, a nearest-centroid classifier with unit variance and feature screening by the

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

mean difference. Correlated features, other classifiers (which can fit noise more or less well) and other screening statistics will change the constants. The order-statistic approximation in Eq. (1) is crude for $k = 1$, where the formula underestimates the leaky accuracy by about 0.01 to 0.02. Eq. (5) was within the simulation by a margin that we measured only where the simulation reached the target.

Real data. We used two small datasets and a logistic regression with $k = 10$ features. Noise features were independent Gaussian, which is the best case for the formula, and the permuted-label experiment is a stress test and not a realistic dataset. We did not study tuning of hyperparameters inside the loop, repeated cross-validation, bootstrap estimates or other nested designs, although the same leakage arises there [3, 4].

8. Conclusion

Selecting features on all the cases before cross-validation reports accuracy that grows with the number of features screened and shrinks only as $1/\sqrt{n}$. In the pure-noise case the reported accuracy has an explicit formula, within 0.025 of simulation in 27 settings, and a closed form for the sample size needed to make the optimism small: about 2,326 cases for an optimism of 5 points with 5 of 1,000 features. Nested cross-validation removes the bias but costs variance, and the leakage is harmless only when the real signal dominates the screening noise. The next steps are to extend the formulas to correlated features and other classifiers, to tuning inside the loop, and to a diagnostic that estimates the optimism from the data.

References

[1]. C. Ambroise, G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, 99(10):6562–6566, 2002.

[2]. R. Simon, M. D. Radmacher, K. Dobbin, L. M. McShane. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.*, 95(1):14–18, 2003.

[3]. S. Varma, R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

[4]. G. C. Cawley, N. L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, 11:2079–2107, 2010.

[5]. T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[6]. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, 1995.

[7]. Y. Bengio, Y. Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.*, 5:1089–1105, 2004.

[8]. G. Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018.

[9]. I. Guyon, A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[10]. J. Fan, J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, 70(5):849–911, 2008.

[11]. H. A. David, H. N. Nagaraja. *Order Statistics*, 3rd ed. Wiley, 2003.

[12]. G. Blom. *Statistical Estimates and Transformed Beta-Variables*. Wiley, 1958.

[13]. M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B*, 36(2):111–147, 1974.

[14]. F. Pedregosa et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[15]. W. N. Street, W. H. Wolberg, O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *SPIE*, vol. 1905, 1993.

[16]. E. Alpaydin, C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

[17]. D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminformatics*, 6:10, 2014.

A Derivations

Eq. (2) and (3). The sample class means along a selected feature differ by d_j, and within class the sample variance is about 1, so the sample is approximately two unit-variance clusters separated by d_j. For k features with a nearest-centroid rule the separation along $m_1 - m_0$ is its length $D = \|d\|$, and the apparent accuracy is $\Phi(D/2)$. Substituting $d_{(i)}^2 = (4/n)z_i^2$ gives Eq. (2).

Eq. (4). See the proof of Proposition 3. For $K = 5$ the correction $1/(K-1)$ is 0.25. The mean separation $\Sigma(a_j^2 - 4/n)$ is smaller than the resubstitution one by $4k/n$, which is the part of the apparent effect that is noise in the held-out fold.

Eq. (5). Setting $\Phi(S/2) = 1/2 + \varepsilon$ gives $S = 2\Phi^{-1}(1/2 + \varepsilon)$. Since $S = c/\sqrt{n}$ with $c = 2\,\Sigma(z_i^2 - 1)/\sqrt{(\Sigma(z_i^2 + 1/(K-1)))}$, solving for n gives Eq. (5).

B Parameters

Synthetic: balanced classes, unit-variance Gaussian features, nearest-centroid classifier, 5-fold CV. Formula check: $n \in \{40, 100, 300\}$, $p \in \{100, 1{,}000, 10{,}000\}$, $k \in \{1, 5, 20\}$, 300 repetitions (200 for $p = 10{,}000$), test set of 2,000 cases. Sample-size study: $p = 10{,}000$, $n \in \{50, \ldots, 1{,}600\}$, 120 repetitions; $p = 1{,}000$, $n \in \{100, \ldots, 6{,}400\}$, 100 repetitions (60 for $n \geq 3{,}200$). Signal study: 10 informative features with shift 0.5, $p = 10{,}000$, $k = 20$, $n \in \{50, 100, 200, 400, 800\}$, 150

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P5

repetitions, test set of 4,000. Real data: standardised features, noise features $N(0, 1)$, $p_{noise} \in \{0, 100, 1{,}000, 5{,}000\}$, $n \in \{40, 80, 160\}$, Welch t selection with $k = 10$, logistic regression ($C = 1$), 5-fold CV, 100 repetitions, balanced accuracy on all remaining cases, with and without permuted training labels. All figures were produced by code at 600 dpi and are printed at their stored width. Code, results and figures are in the supplementary archive.

How To Cite:

Shabbeer Shaik , " Closed-Form Characterization of Selection Bias in Cross-Validation for Feature-Selected Predictive Models using Machine Learning ", International Journal of Research and Development in Engineering Sciences (IJRDES), Volume 8 , Issue 6 , pp 49 – 59 , November 2026. CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p5

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com