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: 78 – 87
Received: 01 March 2026 ; *Accepted:* 30 October 2026 ; *Published:* 19 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P8
https://www.youtube.com/watch?v=YKU5_6Z6ej8
https://archive.org/details/jsp-ijrdes-v8ri6p8
Original Paper : Applied Research – Experimental Research Paper

Reliable Model Comparison Under Dual Sources of Variability: Training Seeds and Evaluation Test Cases using Machine Learning

Seepana Praveenkumar

Department of Nuclear and Renewable Energy Sources , Ural Federal University, Russia

* *Corresponding Author : Seepana Praveenkumar ; pravinkumar.sipana@urfu.ru*

Abstract: Model comparisons often report the mean accuracy over K random seeds and judge the gap against the seed-to-seed spread. That spread measures one source of noise. The other is the test set, which is the same for every seed and so cannot be averaged away by running more seeds. We give the standard error of the difference, $\sqrt{(2\sigma_s^2/K + \sigma_d^2/N)}$, where σ_s is the seed standard deviation of accuracy, σ_d the standard deviation across test cases of the paired difference in correctness, and N the number of test cases. A seed-only interval then covers the true difference of the two procedures with probability $2\Phi(1.96\, a/\sqrt{(a^2 + b^2)}) - 1$, which falls toward zero as seeds are added; here a and b are the two noise terms. In a semi-synthetic simulation with 100 test cases, coverage fell from 70% to 37% as K went from 3 to 30, within 0.052 of the formula for $K \geq 5$. On the digits dataset, for a multilayer network and a random forest with the same population accuracy to within 0.1 points, the seed-only interval covered 41% with 20 seeds and 100 test cases, and rejected equality in 59% of draws; an interval that also includes test-set noise covered 96%. Pairing the comparison on the same test cases reduced the per-case standard deviation by factors of 4.1 to 7.3. Seeds beyond $K^* = 2\sigma_s^2 N/\sigma_d^2$ add little; for the real pairs K^* was between 0.6 and 10.8 at 300 test cases. Simulations are semi-synthetic; the real data are one small digits split with four model families.

Keywords: Random Seeds, Model Comparison, Statistical Power, Test-Set Noise, Paired Evaluation, Reproducibility.

1. Introduction

Machine-learning papers routinely claim that one method beats another because its accuracy, averaged over a handful of random seeds, is higher by some margin. The margin is judged against the spread over seeds, either informally or by a t-test [1, 2, 6, 8]. Variance across seeds is real and large [6, 7], and reporting it is good practice.

It is also incomplete. All the seeds are evaluated on the same test set. The test set is a finite sample from the population of cases, and the two methods may be better on different cases. Whatever difference they show on this test set contains an error that is the same for every seed, so averaging over seeds does not reduce it. A comparison that uses only the seed spread therefore answers a narrower question, which method is better on this test set, and can be badly overconfident about the broader one, which method is better.

We make the point quantitatively. The standard error of the comparison has two terms, one that falls with the number of seeds and one that falls with the number of test cases, and each is estimable. We derive what the seed-only interval covers, how many seeds and cases a comparison needs, and how much pairing helps, and we check the results in a semi-synthetic model and on real classifiers. We make five contributions.

- A two-term standard error for the difference between two training procedures, and the coverage of the seed-only interval as a function of those terms (Section 4).
- A closed-form power function and minimum detectable difference, with a floor that no number of seeds can lower.
- The seed count beyond which further seeds add little, $K^* = 2\sigma_s^2 N/\sigma_d^2$, and the corresponding test-set size.
- A quantification of the benefit of pairing the comparison on the same test cases.
- A real-data check using four model families on the digits dataset with 60 seeds each.

All semi-synthetic results are simulations from a known model. The real-data experiment uses

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-V8RI6P8

one small dataset and treats the 60-seed average on a held-out pool as the true accuracy of each procedure.

2. Background and Related Work

Comparing classifiers. Dietterich studied the Type I error of tests for comparing learning algorithms and the sources of variation, including the choice of test set and the randomness of training [1]. Demšar recommended nonparametric tests for comparisons across datasets [2]. McNemar's test compares two classifiers on the same cases through their disagreements [3], and Bouckaert and Frank examined the replicability of such tests [4]. Japkowicz and Shah survey evaluation methods [5].

Seeds and reproducibility. Henderson and colleagues showed that seed variation can change conclusions in deep reinforcement learning [6], Picard showed the spread of accuracy across seeds in computer vision [7], and Bouthillier and colleagues decomposed the variance in benchmarks into data splits, initialisation and other sources [8]. Colas and colleagues applied power analysis to the number of seeds [9], and Agarwal and colleagues argued for interval estimates [11]. Card and colleagues documented low power in natural-language experiments [17].

What we add. The decomposition is known. We give the coverage of the seed-only interval, the explicit floor on the detectable difference, the balance point for the number of seeds, and a measurement of these quantities on real classifiers.

3. Research Methodology and Gap

3.1. Two procedures, many seeds, one test set

A training procedure A (an architecture, hyperparameters and a training set) produces a classifier for each random seed. A second procedure B does the same. Let μ_A and μ_B be the population accuracies of the procedures, averaged over seeds, and $\delta = \mu_A - \mu_B$ the quantity of interest. We train K seeds of each, score all of them on a test set T of N cases, and take the difference of mean accuracies $\hat{D}$. Let $P_{A,i}$ be the probability, over seeds, that procedure A is correct on case i, and σ_d^2 the variance across cases of $P_{A,i} - P_{B,i}$. Let σ_s^2 be the variance over seeds of the accuracy of one procedure on T.

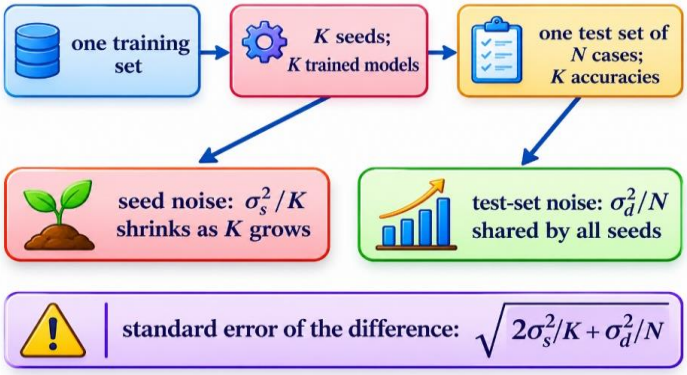


Figure. 1 Two sources of noise when two models are compared on one test set. Seeds change the trained models; the test set changes which cases are scored. Adding seeds only reduces the first.

3.2. Intervals

Seed-only interval: $\hat{D} \pm 1.96\ \hat{s}_{\text{seed}}$, with $\hat{s}_{\text{seed}}^2 = (s_A^2 + s_B^2)/K$, where s_A^2, s_B^2 are the sample variances of the seed accuracies. **Combined interval:** $\hat{D} \pm 1.96\ \sqrt{(\hat{s}_{\text{seed}}^2 + \hat{v}/N)}$, where $\hat{v}$ is the sample variance over test cases of the seed-averaged paired difference in correctness. **Paired comparison:** $\hat{D}$ computed on the same test cases for both procedures, so that the per-case difference, not the per-case outcomes, enters the variance.

4. Proposed System

Proposition 1 (standard error). Conditional on the training data, $\hat{D}$ has mean δ_T (the expected difference on the test set) and, over seeds, variance $2\sigma_s^2/K$. The test-set difference δ_T varies around δ with variance σ_d^2/N (times $1 - N/N_{pop}$ for a finite population). Hence

$$\text{Var}(\hat{D} - \delta) = 2\sigma_s^2/K + \sigma_d^2/N. \quad (2)$$

Proof. Write $\hat{D} - \delta = (\hat{D} - \delta_T) + (\delta_T - \delta)$. The first term has variance $2\sigma_s^2/K$ over seeds given T. The second is the mean of N independent case-level terms $P_{A,i} - P_{B,i}$, each with variance σ_d^2. The terms are uncorrelated.

The first term is the noise that seeds expose; the second is invisible to the seeds. As $K \to \infty$ the variance does not go to zero but to σ_d^2/N.

Corollary 1 (what the seed-only interval covers). *Let $a = \sigma_s\sqrt{(2/K)}$ and $b = \sigma_d/\sqrt{N}$. The seed-only interval $\hat{D} \pm 1.96\ a$ covers δ with probability*

$$\text{Cov} = 2\Phi(\ 1.96\ a\ /\ \sqrt{(a^2 + b^2)}\ ) - 1. \quad (3)$$

This is 95% only if $b = 0$. As K grows, a falls and 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-V8RI6P8

coverage falls toward 0. For $b = a$ it is 84%, and for $b = 3a$ it is 49%.

Corollary 2 (power and detectable difference). *A two-sided test of $\delta = 0$ at level 5% that uses the full standard error $S = \sqrt{(a^2 + b^2)}$ has power*

$$\text{Power}(\delta) = \Phi(\delta/S - 1.96) + \Phi(-\delta/S - 1.96), \quad \delta_{\min} = (1.96 + 0.84)\, S, \tag{4}$$

where $\delta_{\min}$ is the difference detected with 80% power. Even with unlimited seeds $\delta_{\min} \geq 2.8\, \sigma_d/\sqrt{N}$.

Corollary 3 (balance point). *The two terms of Eq. (2) are equal when*

$$K^* = 2\sigma_s^2\, N / \sigma_d^2. \tag{5}$$

Beyond about K^* seeds, further seeds shrink the variance by less than a factor of two. If σ_d is large relative to σ_s, one or two seeds suffice; if the two procedures agree on most cases, more seeds help.

Remark (pairing). If both procedures are scored on the same N cases, σ_d^2 is the variance of the paired difference, which is $\text{Var}(P_A) + \text{Var}(P_B) - 2\text{Cov}(P_A, P_B)$. If they are scored on independent test sets, σ_d^2 is replaced by $\text{Var}(P_A) + \text{Var}(P_B)$. The ratio is the variance reduction from pairing.

5. Experimental Setup and Implementation

The semi-synthetic model has a pool of 20,000 cases with latent margins $\mu_i \sim N(0.9, 1)$. Procedure j has a fixed case effect $a_{ij} \sim N(0, 0.6^2)$, a seed-level shift $b_{js} \sim N(0, 0.04^2)$ common to all cases, and independent case-seed noise $e_{ijs} \sim N(0, 0.5^2)$; a case is correct if the sum is positive. The true difference δ is set by shifting the margins of B. Accuracy is about 0.76, $\sigma_s = 0.0093$ over the full pool and $\sigma_d = 0.285$. For each of 800 repetitions we draw a test set of $N \in \{100, 300, 1{,}000, 3{,}000\}$ cases without replacement, and $K \in \{3, 5, 10, 30\}$ seeds of each procedure. The power study uses $N = 500$ and $\delta \in \{0, 0.005, \ldots, 0.03\}$.

Real data are the digits dataset (1,797 cases, ten classes) [14] from scikit-learn [13]. A fixed random 600 cases are the training set (standardised) and the remaining 1,197 are the pool. We train four procedures with 60 seeds each: multilayer networks with 32 and 64 hidden units (200 iterations), a random forest and an extra-trees ensemble with 50 trees each, and a deterministic logistic regression. The true accuracy of each procedure is its mean pool accuracy over 60 seeds, and the true difference of a pair is the difference. Each repetition draws a test set of $N \in \{100, 300\}$ cases from the pool and $K \in \{3, 5, 10, 20\}$ seeds from the 60, for 3,000 repetitions per cell. The seeds are therefore reused across repetitions, so the estimates condition on one training set and one set of 60 seeds.

6. Results and Analysis

6.1. What The Seed-Only Interval Covers

Table 1 and Figure 2 give coverage of the true difference in the semi-synthetic model, where the two procedures have equal accuracy and differ on individual cases ($\sigma_d = 0.285$). With 100 test cases, the nominal 95% seed-only interval covers 70% with 3 seeds, 57% with 10 and 37% with 30. More seeds make the interval narrower and wrong more often. Increasing the test set helps slowly: with 3,000 cases and 30 seeds the coverage is 68%. Eq. (3) matches the simulated coverage within 0.052 for $K \geq 5$; with $K = 3$ the formula is higher than the simulation by up to 0.08, because a variance estimated from three seeds is noisy and a t quantile with two degrees of freedom is wider than 1.96. The combined interval covers 95% to 98% in every cell, a little conservative because the case-level variance estimate also contains part of the seed noise.

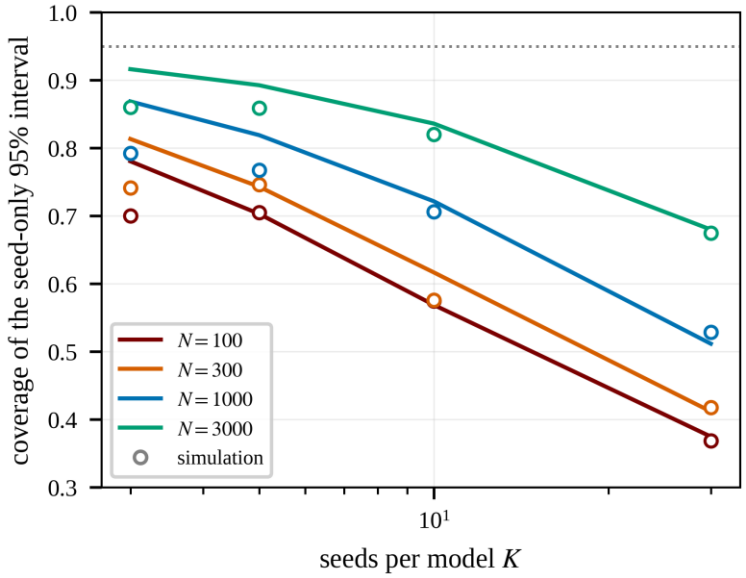


Figure. 2 Coverage of the nominal 95% seed-only interval for the true difference between two procedures of equal accuracy (semi-synthetic, 800 repetitions per cell). Lines: Eq. (3). Circles: simulation.

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-V8RI6P8

Table. 1 Coverage (%) of the nominal 95% seed-only interval (simulation, formula) and of the combined interval, semi-synthetic model.

N	*K*	Seed-only sim	Seed-only Eq. (3)	Combined sim.
100	3	70	78	97
100	5	71	70	97
100	10	57	57	96
100	30	37	37	95
300	3	74	81	97
300	5	75	74	96
300	10	58	62	96
300	30	42	41	95
1000	3	79	87	96
1000	5	77	82	96
1000	10	71	72	97
1000	30	53	51	96
3000	3	86	92	95
3000	5	86	89	97
3000	10	82	84	98
3000	30	68	68	97

Figure. 3 shows the standard deviation of the difference: it falls with K and then levels off at the floor $\sigma_d/\sqrt{N}$. With 300 test cases the standard deviation is 0.023 at $K = 3$, 0.017 at $K = 30$, and the floor is 0.016.

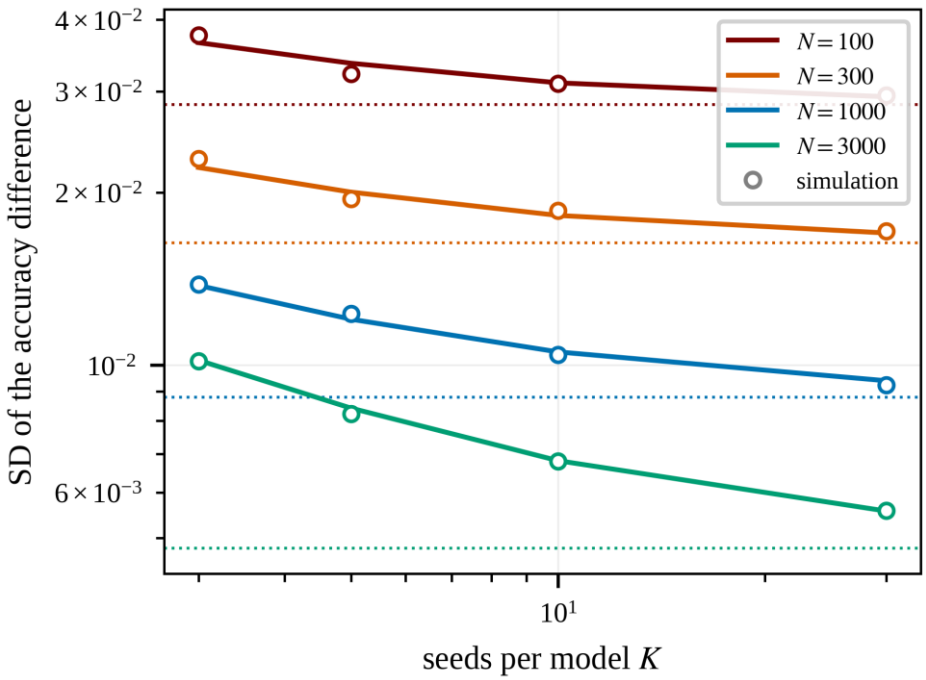


Figure. 3 Standard deviation of the accuracy difference against the number of seeds for four test-set sizes. Lines: Eq. (2). Circles: simulation. Dotted: the floor set by the test set.

6.2. Power

Table 2 and Figure 4 give the power of the combined test at $N = 500$. At $K = 10$ and $\delta = 0.02$ the power is 22% (formula 28%), and at $\delta = 0.03$ it is 47% (54%). Tripling the seeds to 30 raises the power at $\delta = 0.03$ only to 56%. The seed-only test is the opposite problem: with $\delta = 0$ it rejects 21% of the time at $K = 3$ and 56% at $K = 30$, against the nominal 5%; the combined test rejects 3% and 4%. The formula overstates the simulated power slightly because the estimated standard error is larger than the true one.

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-V8RI6P8

Table. 2 Rejection rate (%) of the combined test at $N = 500$: simulation and Eq. (4), and of the seed-only test.

K	δ	Combined Sim	Eq. (4)	Seed-only sim
3	0	3	5	21
3	0.005	3	6	25
3	0.01	5	9	29
3	0.015	10	13	35
3	0.02	13	20	44
3	0.03	27	38	62
10	0	4	5	39
10	0.005	5	6	40
10	0.01	7	11	44
10	0.015	14	18	59
10	0.02	22	28	70
10	0.03	47	54	85
30	0	4	5	56
30	0.005	5	7	58
30	0.01	11	12	66
30	0.015	17	20	73
30	0.02	28	32	86
30	0.03	56	60	96

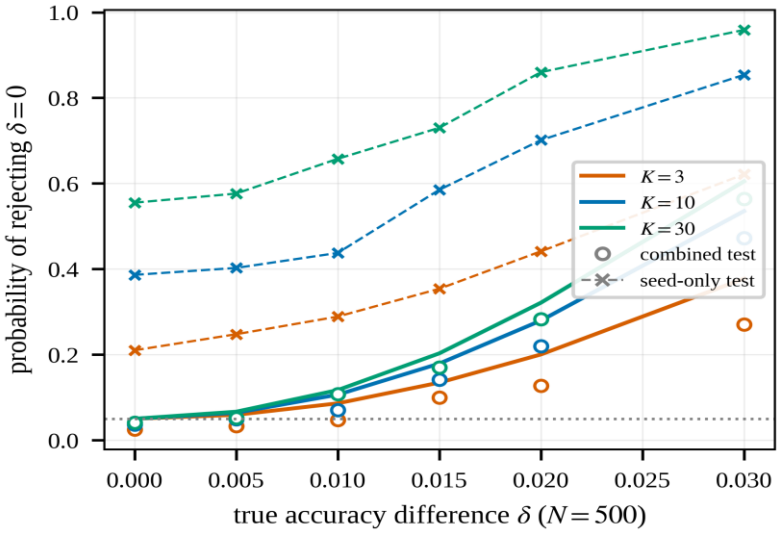


Figure. 4 Probability of rejecting equality against the true difference ($N = 500$). Lines: Eq. (4). Circles: the combined test; crosses: the seed-only test.

6.3. The floor and the balance point

Eq. (4) shows that the detectable difference cannot be pushed below $2.8\,\sigma_d/\sqrt{N}$. Figure 5 and Table 3 give it for the MLP-64 against RF-50 pair of Section 6.4 ($\sigma_s = 0.0043$, $\sigma_d = 0.138$). With 300 test cases and 10 seeds the smallest detectable difference is 2.3 points, and the limit with unlimited seeds is 2.2 points. To detect 1 point, 10,000 cases would give a floor of 0.4 points. At the balance point K^* (Eq. 5) the two terms are equal: for these models $K^* = 0.6$ at 300 cases and 5.8 at 3,000, so on a small test set a single seed already captures most of the noise that seeds can reveal.

Table. 3 Smallest difference (points) detected with 80% power (Eq. 4) for $\sigma_s = 0.0043$ and $\sigma_d = 0.138$, by seeds K and test-set size N.

N	$K = 1$	$K = 3$	$K = 10$	$K = 30$	$K \to \infty$
100	4.2	4.0	3.9	3.9	3.9
300	2.8	2.4	2.3	2.3	2.2
1000	2.1	1.6	1.3	1.3	1.2
3000	1.8	1.2	0.9	0.8	0.7
10000	1.7	1.1	0.7	0.5	0.4

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-V8RI6P8

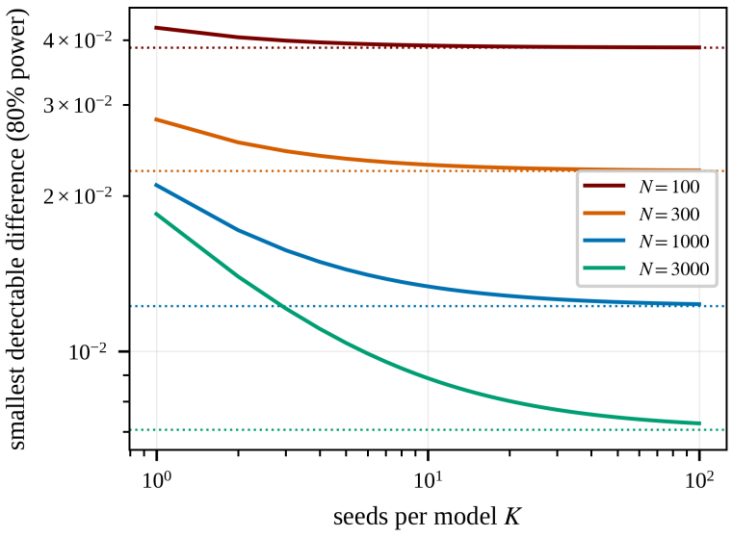


Figure. 5 Smallest difference detected with 80% power against seeds, for the MLP-64 against RF-50 pair of Section 6.4 ($\sigma_s = 0.0043$, $\sigma_d = 0.138$). Dotted: the limit as $K \to \infty$.

6.4. Real Classifiers

Figure 6 and Table 4 show the two sources of noise for real classifiers on digits. The seed standard deviation of accuracy is between 0.4 and 0.5 points for a single seed, while a test set of 300 cases has a sampling standard deviation of 1.1 points and one of 100 cases 2.1. The test set contributes more noise than the seeds unless the test set is large. The logistic regression is deterministic and has no seed variance, so no number of seeds could describe its uncertainty.

Table. 4 Real models on digits (training set 600, pool 1,197, 60 seeds): Accuracy, seed SD of accuracy, and sampling SD of accuracy for a test set of 100 and 300 cases (points).

Model	Accuracy (%)	Seed SD	Test SD $N = 100$	Test SD $N = 300$
MLP-32	94.1	0.5	2.2	1.2
MLP-64	94.9	0.5	2.1	1.1
RF-50	95.0	0.4	2.1	1.1
ET-50	96.1	0.4	1.8	1.0
LogReg	93.9	0.0	2.3	1.2

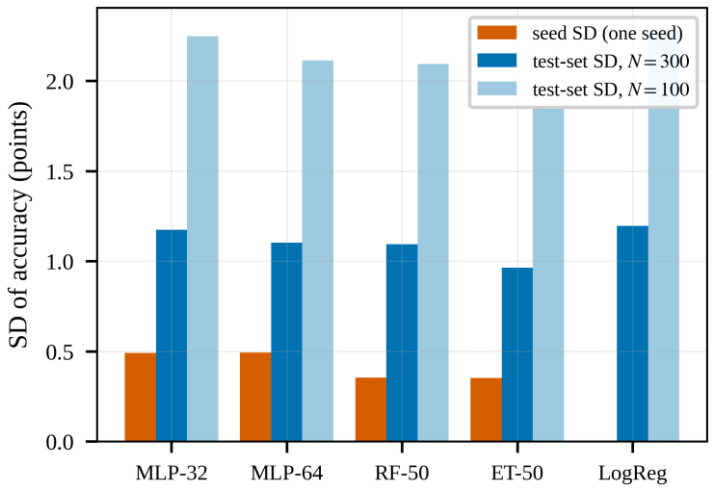


Figure. 6 Real models on digits: standard deviation of accuracy from one training seed (60 seeds) and from drawing a test set of 100 or 300 cases. LogReg is deterministic.

Table 5 and Figure 7 give coverage of the 60-seed difference. The pair MLP-64 and RF-50 has nearly equal accuracy (difference 0.1 points) but disagrees on individual cases ($\sigma_d = 0.138$). With 20 seeds and 100 test cases the seed-

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-V8RI6P8

only interval covers 41% of the time (Eq. 3: 42%) and rejects equality in 59% of draws; the combined interval covers 96% and rejects 3%. For the pair MLP-32 and MLP-64 the procedures agree on most cases (σ_d = 0.036) and the seed-only interval is much closer to correct (89%), and for RF-50 and ET-50 it is 72%. The formula matches the simulation within 0.044 for $K \geq 5$. The size of the problem depends on σ_d, which can be measured.

Table. 5 Real pairs on digits: coverage (%) of the 60-seed difference by the seed-only interval (simulation, Eq. 3) and the combined interval; rejection rate (%) of equality by each test.

Pair	*N*	*K*	Seed sim	Seed Eq (3)	Comb	Rej-seed	Rej comb
MLP-32 / MLP-64	100	3	86	94	99	21	2
MLP-32 / MLP-64	100	20	89	87	98	41	9
MLP-32 / MLP-64	300	20	93	89	99	81	47
MLP-64 / RF-50	100	3	70	76	98	33	2
MLP-64 / RF-50	100	20	41	42	96	59	3
MLP-64 / RF-50	300	20	47	47	97	54	2
RF-50 / ET-50	100	3	82	91	98	36	4
RF-50 / ET-50	100	20	72	71	95	78	29
RF-50 / ET-50	300	20	77	75	98	99	88

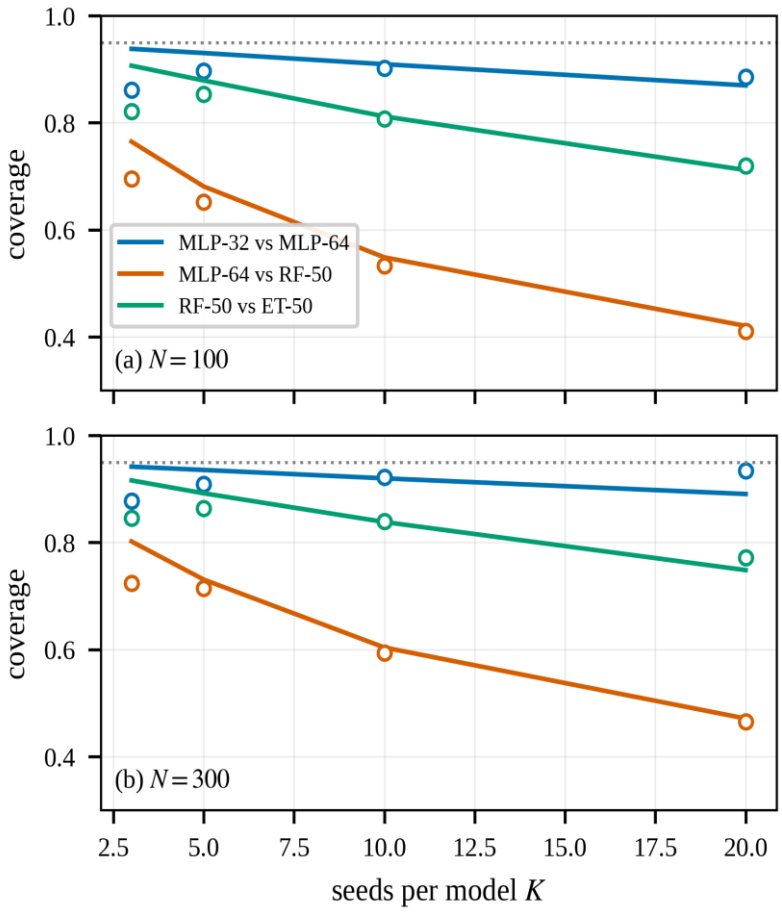


Figure. 7 Real digits models: coverage of the seed-only 95% interval for the 60-seed difference, by pair and test-set size (3,000 repetitions). Lines: Eq. (3). Circles: simulation.

6.5. Pairing

Pairing reduces σ_d greatly when the procedures agree. For MLP-32 against MLP-64 the per-case standard deviation falls from 0.265 unpaired to 0.036 paired, a variance reduction of 52.9; for RF-50 against ET-50 it falls from 0.243 to 0.059, and for MLP-64 against RF-50, whose errors are less aligned, from 0.261 to 0.138 (Figure 9). Comparing on separate test sets would need 52.9 times as many cases for the first pair. The test set size needed for a given power (Figure 8) is therefore much smaller for a paired comparison, and the pairing benefit is largest exactly where the seed-only analysis is most misleading.

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-V8RI6P8

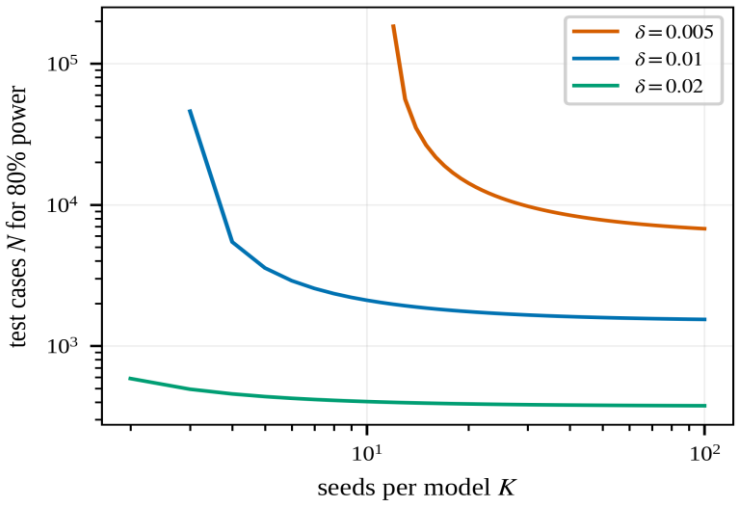


Figure. 8 Test cases needed to detect a difference δ with 80% power against the number of seeds ($\sigma_s = 0.0043$, $\sigma_d = 0.138$). The curves are vertical asymptotes at the minimum number of seeds.

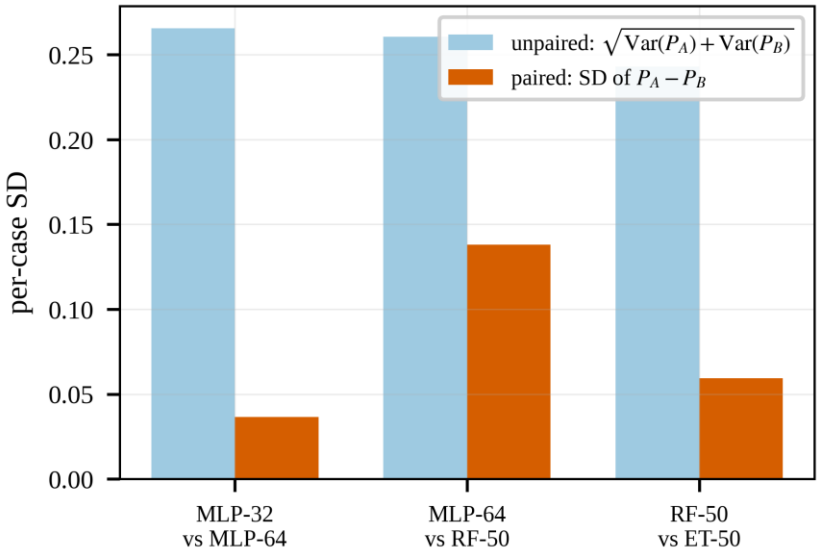


Figure. 9 Per-case standard deviation of the comparison with and without pairing, for three real pairs on digits.

6.6. Planning Table

Table 6 evaluates the test-set size needed to detect a difference δ with 80% power for the pair MLP-64 against RF-50. Detecting 2 points with 10 seeds needs 403 cases; with 3 seeds 493. Detecting 1 point needs 2,106 cases with 10 seeds, and 0.5 points needs never. For a pair that agrees more ($\sigma_d = 0.036$ as for the two networks) the numbers are about 14.7 times smaller.

Table. 6 Test cases needed for 80% power (Eq. 4) for $\sigma_s = 0.0043$ and $\sigma_d = 0.138$, by true difference δ and seeds K.

δ (points)	$K = 3$	$K = 10$	$K = 30$	$K \to \infty$
2	493	403	383	374
1	46,014	2,106	1,655	1,495
0.5	never	never	9,754	5,979

6.7. A Practical Recipe

The results suggest a short procedure. First, report the standard error with both terms: estimate σ_s from the seeds and σ_d from the per-case paired differences, and use $\sqrt{(2\sigma_s^2/K + \sigma_d^2/N)}$. Second, compute the balance point K^* and run about that many seeds, not more; put the effort into the test set instead. Third, always compare on the same test cases and report the paired standard deviation; it is the cheapest way to gain power. Fourth, calculate the smallest difference the experiment can detect before running it (Eq. 4), and do not interpret a smaller gap. Fifth, if the conclusion is meant to apply beyond the test set, do not use a seed-only interval; its coverage can be far below 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-V8RI6P8

nominal level.

Discussion and Limitations

What we think carries over. Three points seem general. Seeds and test cases are two separate sources of noise and only one of them can be reduced by running more training. The usual seed-only interval is overconfident when the procedures disagree on individual cases, and its coverage decreases as seeds are added. And pairing and a larger test set are cheaper and more effective than more seeds once K exceeds the balance point.

Synthetic setting. The semi-synthetic model has a simple margin structure, a single seed-level shift and independent case noise, and accuracy near 0.76. Real seed effects are correlated across cases in more complex ways. The formulas assume independent seeds and cases, and the t correction for small K is not included in Eq. (3).

Real data. We used one dataset, one training set of 600 cases and four model families with 60 seeds each, and a pool of 1,197 cases, so the "true" accuracy is itself an estimate and the same seeds are reused across repetitions. The results condition on one training set; variation due to the choice of training set is a third source of noise that we did not study. We did not examine other metrics, regression, or deep models.

7. Conclusion

The standard error of a model comparison has two terms, one from seeds and one from the test set, and only the first falls with more seeds. A seed-only interval for the true difference has coverage $2\Phi(1.96\ a/\sqrt{(a^2 + b^2)}) - 1$, which in our simulation fell from 70% to 37% as seeds rose from 3 to 30, and on real classifiers with the same accuracy fell to 41%. The smallest detectable difference has a floor of $2.8\ \sigma_d/\sqrt{N}$, seeds beyond $K^* = 2\sigma_s^2 N/\sigma_d^2$ add little, and pairing the comparison on the same cases can reduce the variance by a factor of 52.9 or more. The next steps are to include variation in the training set, to extend the analysis to several datasets, and to give intervals that combine all sources.

References

[1]. T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[2]. J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

[3]. Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[4]. R. R. Bouckaert, E. Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *PAKDD*, 2004.

[5]. N. Japkowicz, M. Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[6]. P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger. Deep reinforcement learning that matters. In *AAAI*, 2018.

[7]. D. Picard. torch.manual_seed(3407): influence of random seeds in computer vision. arXiv:2109.08203, 2021.

[8]. X. Bouthillier et al. Accounting for variance in machine learning benchmarks. In *MLSys*, 2021.

[9]. C. Colas, O. Sigaud, P.-Y. Oudeyer. How many random seeds? Statistical power analysis in deep reinforcement learning experiments. arXiv:1806.08295, 2018.

[10]. J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Erlbaum, 1988.

[11]. R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, M. G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *NeurIPS*, 2021.

[12]. B. L. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

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

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

[15]. B. Efron, R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[16]. S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808, 2018.

[17]. D. Card, P. Henderson, U. Khandelwal, R. Jia, K. Mahowald, D. Jurafsky. With little power comes great responsibility. In *EMNLP*, 2020.

A. Derivations

Eq. (2). The estimate is $\hat{D} = (1/N)\ \Sigma_{i \in T}\ (\bar{c}_{A,i} - \bar{c}_{B,i})$ with $\bar{c}$ the mean correctness over seeds. Given T, the seed randomness has variance $2\sigma_s^2/K$. The expected value given T is $(1/N)\Sigma(P_{A,i} - P_{B,i})$, whose variance over random T is σ_d^2/N, with the finite-population factor $1 - N/N_{pop}$ when sampling without replacement. The two parts are uncorrelated.

Eq. (3). $\hat{D} - \delta \sim N(0, a^2 + b^2)$ and the half-width is $1.96\ a$, so the coverage is $\Pr(|N(0, 1)| \le 1.96\ a/\sqrt{(a^2 + b^2)}) = 2\Phi(1.96\ a/\sqrt{(a^2 + b^2)}) - 1$.

Eq. (4). If $\hat{D} \sim N(\delta, S^2)$, the test rejects when $|\hat{D}| > 1.96\ S$, with probability $\Phi(\delta/S - 1.96) + \Phi(-\delta/S - 1.96)$. Setting this to 0.8 for $\delta > 0$ gives $\delta/S = 1.96 + 0.84$.

Eq. (5). Setting $2\sigma_s^2/K = \sigma_d^2/N$ gives $K = 2\sigma_s^2 N/\sigma_d^2$.

B. Parameters

Semi-synthetic: 20,000 cases, latent margin $N(0.9, 1)$, case effects $N(0, 0.6^2)$, seed shift $N(0, 0.04^2)$, case-seed noise $N(0, 0.5^2)$; $N \in \{100, 300, 1{,}000, 3{,}000\}$, $K \in \{3,$

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-V8RI6P8

5, 10, 30}, 800 repetitions; power study $N = 500$, $K \in \{3, 10, 30\}$, 800 repetitions per δ. Real data: digits, training 600 (standardised), pool 1,197; MLP with 32 or 64 hidden units, 200 iterations; random forest and extra trees with 50 trees; logistic regression ($C = 0.1$); 60 seeds each; test sets of 100 or 300 cases and 3, 5, 10 or 20 seeds drawn from the 60, 3,000 repetitions per cell. 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:

Seepana Praveenkumar , " Reliable Model Comparison Under Dual Sources of Variability: Training Seeds and Evaluation Test Cases using Machine Learning" , International Journal of Research and Development in Engineering Sciences (IJRDES), Volume 8 , Issue 6 , pp 78 – 87, November 2026.
CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri6p8

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