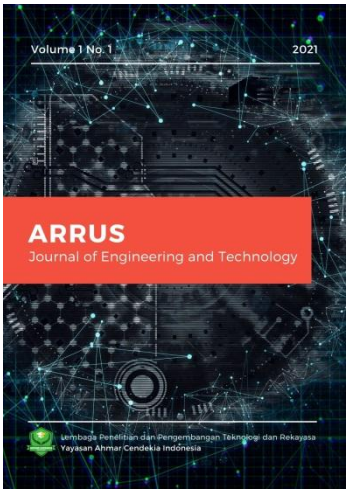
https://doi.org/10.35877/jetech5583

OPEN ACCESS

ISSN : 2776-7914 (Print) / 2807-3045 (Online)



*Corresponding author: Agung Tri Utomo, Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Negeri Makassar.

E-mail: agungtrituomo@unm.ac.id

RESEARCH ARTICLE

Sensitivity of Logistic Regression, Decision Tree, and Random Forest Performance to Random Seed Choice

Agung Tri Utomo[1], Muh. Qodri Alfairus[2], Nur Ikhwana[1],& Zulkifli Rais[1]

[1]Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Negeri Makassar, Makassar, Indonesia.
[2]Department of Accounting, Faculty of Economics and Business, Universitas Negeri Makassar, Makassar, Indonesia.

Abstract: This study examines how sensitive the predictive performance of logistic regression, a decision tree, and a random forest is to the value passed to *set.seed()* in R. The study uses four public binary datasets and eight synthetic datasets with a known data-generating process, 60 consecutive seeds, stratified 70/30 hold-out, and stratified 10-fold cross-validation, and measures accuracy, area under the ROC curve, and F1. A second experiment fixes the data split and varies only the model seed. The standard deviation of hold-out accuracy across seeds ranges from 0.012 to 0.089 and is close to the binomial error implied by test-set size; in synthetic data, test-set sampling accounts for a median of 94% of hold-out variance, and the deviation falls approximately as the inverse square root of sample size. Ten-fold cross-validation reduces it by a factor of 1.8 to 12. With a fixed split, logistic regression and an unpruned tree are deterministic, whereas random forest retains a standard deviation of 0.002 to 0.053. Model rankings reverse more often as true differences shrink, and choosing the best of 20 seeds by test accuracy inflates accuracy by about 0.08. The findings suggest that seed sensitivity under hold-out mainly reflects test-set sampling rather than algorithm-specific randomness.

Keywords: Random seed, Reproducibility, Logistic regression, Decision tree, Random forest

1. Introduction

The *set.seed()* function in R does not affect model quality. It only fixes the initial state of the random number generator (RNG) so that a computation can be repeated exactly. The scientifically meaningful question is therefore not which seed is best, but how much conclusions about model performance change when the seed changes. Choosing a seed because it yields good test performance is misleading, because it inflates the estimate without any basis for generalization.

Evidence from the machine learning literature shows that evaluation randomness can reverse conclusions. Bouthillier et al. (2019) distinguished reproducibility of methods from reproducibility of findings and showed that a conclusion drawn from one experiment can fail in another. Bouthillier et al. (2021) modeled the sources of variance in benchmarks, namely data sampling, parameter initialization, and hyperparameter choice, and proposed recommendations for comparing performance. Henderson et al. (2018) and Picard (2021) reported that results depend on the seed in deep reinforcement learning and computer vision,

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

respectively. Most of this evidence comes from deep learning. For the classical models that are common in data mining research in R, structured quantitative evidence is scarcer, even though the three models studied here differ fundamentally in their sources of randomness.

This study addresses five questions. (1) How large is the variation in accuracy, AUC, and F1 across seeds for each model and validation scheme? (2) How much of this variation comes from the data split, and how much from the internal randomness of the model? (3) Does the variance across seeds differ between models, and are model rankings stable across seeds? (4) How many seeds are needed for the mean performance to be reasonably precise? (5) Do these patterns hold on synthetic data for which the true accuracy of every fitted model is known?

2. Literature Review

Pineau et al. (2021) reported on the NeurIPS 2019 reproducibility program, including its checklist and code submission policy. Gundersen & Kjensmo (2018) surveyed reproducibility in artificial intelligence and stressed the importance of documentation. Madhyastha & Jain (2019) examined model stability as a function of the random seed in natural language processing. Heil et al. (2021) proposed reproducibility standards for machine learning in the life sciences, and Kapoor & Narayanan (2023) showed that data leakage aggravates the reproducibility crisis in machine-learning-based science. Lones (2024) summarized common mistakes in machine learning practice, including mistakes in comparing models.

On validation and evaluation, Varoquaux (2018) showed that cross-validation with small samples yields large error bars, and Vabalas et al. (2019) examined the bias of validation schemes under limited sample sizes. Bates et al. (2024) proved that cross-validation estimates the average prediction error of models fitted to other training sets rather than the error of the model fitted to the data at hand, and that the usual variance estimate for cross-validation tends to be too small. Raschka (2018) reviewed model evaluation and selection, and Benavoli et al. (2017) criticized null hypothesis significance testing for comparing multiple classifiers and proposed Bayesian analysis instead.

On the models, Couronné et al. (2018) compared random forest with logistic regression on 243 datasets and found random forest to be more accurate on about 69% of them. Christodoulou et al. (2019) found in a systematic review no performance benefit of machine learning over logistic regression for clinical prediction models. Probst & Boulesteix (2018) studied the number of trees, and Probst et al. (2019) reviewed hyperparameters and tuning strategies for random forest. Wright & Ziegler (2017)presented the ranger implementation. The studies above do not directly decompose the contribution of the seed for logistic regression, decision tree, and random forest under the same two validation schemes. This study fills that gap with four real and eight synthetic datasets.

3. Research Method and Materials

3.1. Data and models

Four real binary classification datasets from the mlbench package and eight synthetic datasets were used (Table 1). Rows with missing values and constant columns were removed, and ordinal predictors were converted to numeric. Three models were fitted with default settings and no hyperparameter tuning: logistic regression (glm, binomial family), a classification tree (rpart, default parameters, without cost-complexity pruning), and a random forest (randomForest, default parameters except the number of trees, which was set to 50 or 500). Predicted probabilities of the positive class (the second factor level) were converted to classes with a threshold of 0.5.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Table 1: Datasets used in the experiments (Bayes accuracy is the best accuracy achievable on the synthetic process)

Dataset	n	p	Description
Pima	392	8	Real: diabetes diagnosis
Breast Cancer	683	9	Real: malignant versus benign
Ionosphere	351	33	Real: radar returns; one constant column removed
Sonar	208	60	Real: small sample, many predictors
SYN_n100	100	10	Synthetic: linear signal, Bayes accuracy 0.80
SYN_n200	200	10	Synthetic: same process as SYN_n100
SYN_n400	400	10	Synthetic: same process as SYN_n100
SYN_n700	700	10	Synthetic: same process as SYN_n100
SYN_n1000	1000	10	Synthetic: same process as SYN_n100
SYN_nonlinear	500	10	Synthetic: interaction and quadratic signal, Bayes accuracy 0.85
SYN_weak	300	10	Synthetic: weak linear signal, Bayes accuracy 0.65
SYN_highdim	150	60	Synthetic: 5 informative of 60 predictors, Bayes accuracy 0.85

In the synthetic datasets, the predictors are independent standard normal variables and $P(y = 1 \mid x)$ is the logistic function of a signal scaled so that the Bayes accuracy equals the target value. For each synthetic dataset, one sample of the stated size was drawn with a fixed generation seed, and an independent population sample of 5,000 observations was drawn from the same process. The true accuracy of a fitted model is its accuracy on the population sample. The first five synthetic datasets share one process and differ only in n, which allows the effect of sample size to be isolated.

3.2. *Seed design and validation*

Two validation schemes were used: a stratified 70/30 hold-out split, and stratified 10-fold cross-validation (CV) with metrics computed from pooled out-of-fold predictions. In Experiment 1, a single call to *set.seed(s)* was made immediately before the data split, so the seed controlled both the split and the internal randomness of the model, which mirrors common practice. Seeds s = 1, 2, ..., 60 were used for every combination of dataset, validation scheme, and model configuration, giving 5,760 runs. In Experiment 2, the hold-out split was fixed with seed 123 and only the seed set before model fitting varied (60 seeds, 2,880 runs). This experiment separates the internal randomness of the model from the randomness of the split.

3.3. *Metrics and analysis*

Performance was measured by accuracy, the area under the ROC curve (AUC, computed from the Mann–Whitney statistic), and F1 for the positive class. For each cell, the mean, standard deviation (SD), range, and coefficient of variation were computed. Differences in variance between models were tested with the Fligner–Killeen test, and differences in mean performance with the Friedman test using seed as the block; p values were adjusted with the Holm method. Variance components (dataset, model, dataset × model, and a residual interpreted as seed) were estimated with a linear mixed model (lme4). The minimum number of seeds was computed as $n = (1.96 \times SD / \delta)^2 \; with \; \delta = 0.005$. As benchmarks, the hold-out SD was compared with the binomial error $\sqrt{(p(1-p)/n_{test})}$ with $n_{test} = 0.3n$, and with the same quantity multiplied by $\sqrt{((N - n_{test})/(N - 1))}$, which applies when the test set is drawn without replacement from a fixed sample of size N. On synthetic data, the share of hold-out variance due to the fitted model was computed as the variance across seeds of the population accuracy divided by the variance of the observed test accuracy, and the remainder was attributed to test-set sampling. Ranking stability was measured as the share of seeds in which the observed order of two models is opposite to their reference order (the sign of the all-seed mean CV difference), with ties counting one half. Seed selection was simulated by drawing K of the 60 seeds at random 2,000 times and keeping the seed with the highest hold-out test accuracy. Claims derived from the real-data results were tested on the synthetic data against criteria fixed before the full run (Table 8). Warnings issued by glm were counted rather than suppressed. All analyses were run in R 4.3.3 (Mersenne-Twister generator, Inversion normal generator, and Rejection sampling) with randomForest 4.7-1.1,

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

rpart 4.1.23, mlbench 2.1-3.1, and lme4 1.1-35.1. The R script, the raw outputs, and the checking script for the seed-selection analysis are provided as supplementary material.

4. Results and Discussion

4.1. *Results*

4.1.1. *Accuracy under hold-out and cross-validation*

Table 2 reports the mean and SD of accuracy over 60 seeds for the four real datasets. The hold-out SD is larger when the dataset is smaller and when accuracy is lower. The range of hold-out accuracy across seeds reaches 0.054 (Breast Cancer, random forest with 500 trees) to 0.258 (Sonar, logistic regression and decision tree). On Sonar, therefore, choosing a single hold-out seed can produce accuracy that differs by more than 20 percentage points from another seed for the same model.

Table 2: Mean (SD) accuracy over 60 seeds under hold-out and 10-fold cross-validation, real datasets

Dataset	Scheme	Logistic	Decision tree	RF, 50 trees	RF, 500 trees
Pima	Hold-out	0.781 (0.031)	0.756 (0.043)	0.780 (0.033)	0.781 (0.033)
	10-fold CV	0.779 (0.006)	0.758 (0.013)	0.779 (0.009)	0.783 (0.007)
Breast Cancer	Hold-out	0.964 (0.014)	0.943 (0.014)	0.968 (0.012)	0.969 (0.012)
	10-fold CV	0.967 (0.001)	0.946 (0.005)	0.971 (0.002)	0.973 (0.002)
Ionosphere	Hold-out	0.859 (0.031)	0.872 (0.025)	0.930 (0.022)	0.931 (0.020)
	10-fold CV	0.879 (0.010)	0.873 (0.010)	0.935 (0.005)	0.935 (0.003)
Sonar	Hold-out	0.693 (0.056)	0.702 (0.054)	0.808 (0.054)	0.820 (0.050)
	10-fold CV	0.724 (0.021)	0.716 (0.022)	0.834 (0.017)	0.841 (0.014)

For the random forest with 500 trees, the hold-out SD is larger than the CV SD by a factor of 6.2 (Breast Cancer), 6.7 (Ionosphere), 4.7 (Pima), and 3.7 (Sonar). For logistic regression the factors are 11.8, 3.2, 5.1, and 2.7. This reduction is not purely an effect of averaging over folds, because pooled CV evaluates all n observations whereas hold-out evaluates only 30% of them. Across all 12 datasets the factor ranges from 1.8 to 11.9 (2.4 to 11.9 on real data and 1.8 to 6.1 on synthetic data). Figure 1 shows that, for every real dataset, the hold-out boxes are wider than the CV boxes for the same model.

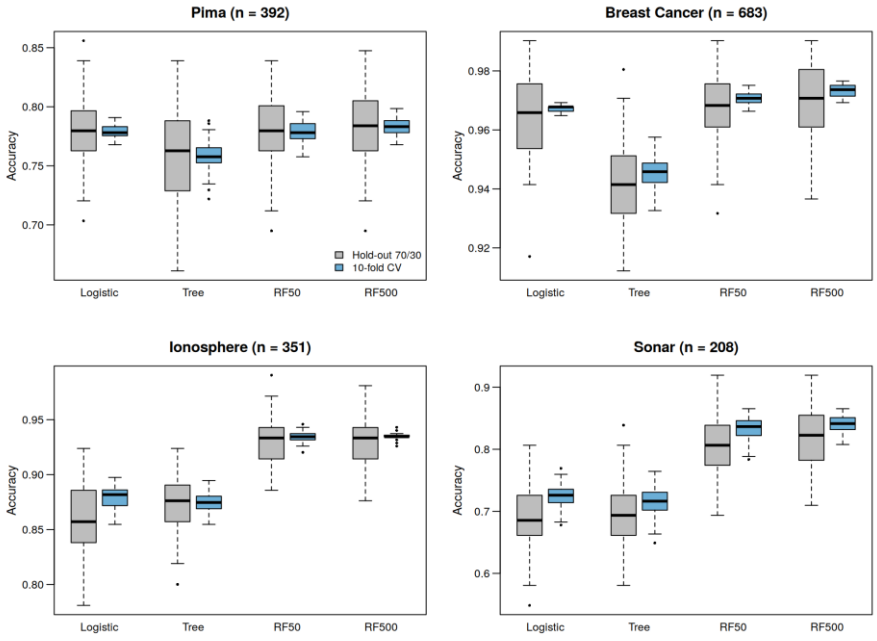


Figure 1: Distribution of accuracy across 60 seeds for each model on the real datasets under hold-out 70/30 (gray) and 10-fold CV (blue). The vertical axis is the same for both schemes within each dataset.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

4.1.2. Test-set sampling as the main source of seed variance

The ratio between the hold-out accuracy SD and the binomial error implied by the test-set size ranges from 0.76 to 1.09 across the 16 dataset × model combinations, with a mean of 0.93. In other words, under hold-out, the variation across seeds is almost exactly what would be expected from drawing a test set of limited size. Consistent with this, the Fligner–Killeen test on hold-out accuracy found no difference in variance between models after Holm adjustment (adjusted $p = 1.00$ for Breast Cancer, 0.065 for Ionosphere, 0.085 for Pima, and 1.00 for Sonar). Under 10-fold CV, variances differ between models for Breast Cancer, Ionosphere, and Pima (adjusted $p \leq 0.0001$) but not for Sonar (0.227). The binomial benchmark uses the approximation $n_test = 0.3n$ and assumes independent predictions, so it is only a rough reference. A refinement that accounts for the test set being drawn from a fixed sample is examined on the synthetic data below.

4.1.3. Internal randomness of the models

Table 3 shows the SD of accuracy when the split is fixed and only the model seed changes. Logistic regression and the decision tree give identical output. For the tree this is expected because it is not pruned using the internal cross-validation of rpart, so the internal folds do not affect the final tree. For the random forest, the internal SD falls when the number of trees rises from 50 to 500 on all four datasets. As a rough guide, the squared ratio of the fixed-split SD to the Experiment 1 hold-out SD puts internal randomness at 2% to 9% of total variance with 500 trees and 12% to 27% with 50 trees on the real datasets, but at 9% to 35% and 25% to 61% on the synthetic datasets, where samples are smaller or signals weaker. These figures come from a single split (seed 123) and should not be taken as general estimates.

Table 3: SD of accuracy with a fixed split; only the model seed varies, real datasets

Dataset	Logistic	Decision tree	RF, 50 trees	RF, 500 trees
Pima	0	0	0.0171	0.0089
Breast Cancer	0	0	0.0043	0.0019
Ionosphere	0	0	0.0089	0.0059
Sonar	0	0	0.0258	0.0121

4.1.4. Model comparison and ranking stability

The Friedman test (seed as block) rejects equal mean performance between models for every combination of dataset, scheme, and metric (all adjusted $p < 0.01$). Rankings within seeds, however, are not always stable (Table 4). On Pima, hold-out gives logistic regression higher accuracy than the 500-tree random forest in 40% of seeds and lower accuracy in 43%. A researcher using a single hold-out seed on this dataset could therefore reach either conclusion with almost equal probability. On Breast Cancer, logistic regression beats the random forest in 22% of hold-out seeds but in none of the CV seeds.

Table 4: Proportion of seeds in which each model has the highest accuracy (ties split equally): logistic regression / decision tree / random forest with 500 trees, real datasets

Dataset	Hold-out	10-fold CV
Pima	0.39 / 0.19 / 0.42	0.33 / 0.02 / 0.65
Breast Cancer	0.29 / 0.00 / 0.71	0.00 / 0.00 / 1.00
Ionosphere	0.04 / 0.00 / 0.96	0.00 / 0.00 / 1.00
Sonar	0.03 / 0.04 / 0.92	0.00 / 0.00 / 1.00

AUC and F1 follow the same pattern as accuracy but with larger spread on small data. Under hold-out, the SD of AUC for the 500-tree random forest is 0.004 (Breast Cancer), 0.010 (Ionosphere), 0.034 (Pima), and 0.032 (Sonar), and the SD of F1 on Pima ranges from 0.049 to 0.075 across models. For AUC, the Fligner–Killeen test shows differences in variance between models on all datasets and schemes (adjusted $p < 0.05$).

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

4.1.5. Variance components and number of seeds

After controlling for dataset and model, seed accounts for about 10% of the variance in hold-out accuracy and about 1% under CV on the real datasets (Table 5). On the synthetic datasets, which differ less from one another by design, the seed share is larger: 43% under hold-out and 10% under CV. The dataset components should not be interpreted as findings, because they depend on how different the chosen datasets happen to be; the contrast between hold-out and CV in the seed component is the stable result.

Table 5: Variance components of accuracy (percent), from the mixed model, for real and synthetic datasets

Component	Real: hold-out	Real: 10-fold CV	Synthetic: hold-out	Synthetic: 10-fold CV
Dataset	77.4	84.1	30.2	38.4
Dataset × model	7.1	7.9	13.7	29.9
Model	5.7	6.7	13.1	22.2
Seed (residual)	9.9	1.2	43.0	9.5

Table 6 gives the number of seeds needed for the 95% confidence interval of the mean accuracy to have a half-width of at most 0.005 on the real datasets. Under hold-out this requires 23 to 474 seeds, and under CV only 1 to 75. The calculation uses the sample SD from 60 seeds, so values far above 60 are extrapolations. Figure 2 shows why a small number of seeds is misleading: the cumulative mean of hold-out accuracy still fluctuates until roughly 20 to 30 seeds. On Sonar, the decision tree appears about 0.13 better than logistic regression at the first seed, but the gap shrinks to about 0.01 at seed 60. On Breast Cancer, logistic regression looks best over the first two or three seeds before the random forest overtakes it.

Table 6: Number of seeds for a 95% confidence interval half-width of at most 0.005 for mean accuracy, real datasets

Dataset	Scheme	Logistic	Decision tree	RF, 50 trees	RF, 500 trees
Pima	Hold-out	145	286	169	166
	10-fold CV	6	26	13	8
Breast Cancer	Hold-out	30	32	23	24
	10-fold CV	1	5	1	1
Ionosphere	Hold-out	146	95	75	63
	10-fold CV	15	15	4	2
Sonar	Hold-out	474	442	444	382
	10-fold CV	67	75	47	29

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

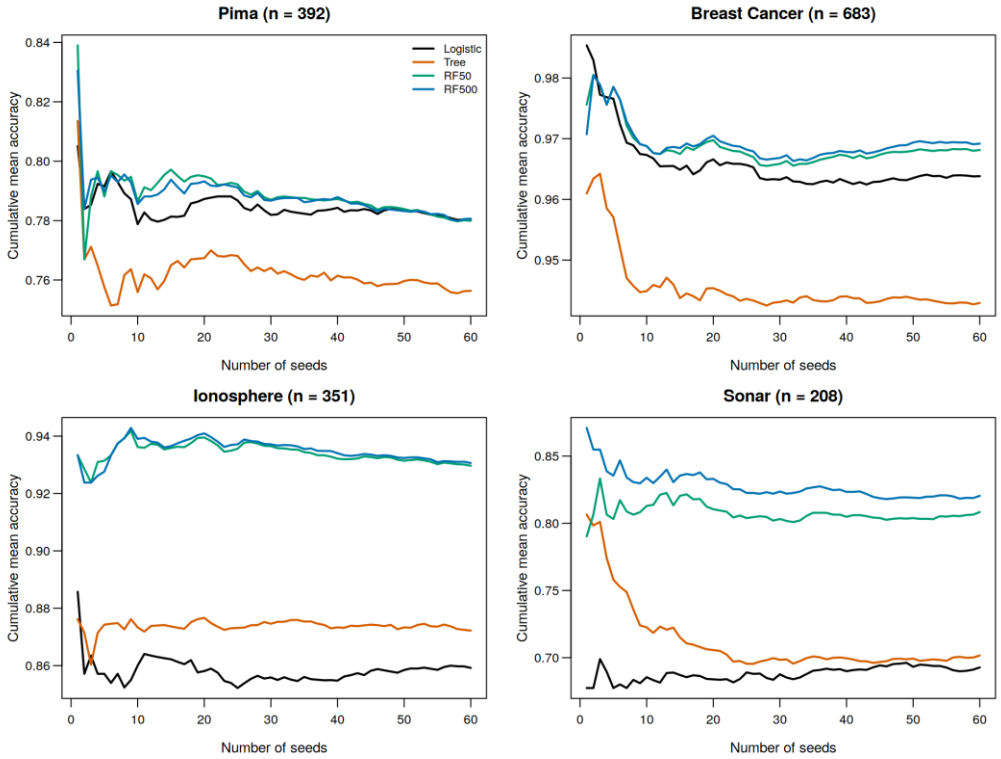


Figure 2: Cumulative mean hold-out accuracy against the number of seeds for each model on the real datasets.

4.1.6. Validation on synthetic data with known truth

In the synthetic datasets the true accuracy of every fitted model is measured on the independent population sample, so claims drawn from the real-data results can be tested directly. Table 7 decomposes the hold-out variance for each synthetic dataset. Test-set sampling accounts for 89% to 98% of the variance in hold-out accuracy across seeds. The observed SD is 0.81 to 0.96 times the plain binomial benchmark and 0.97 to 1.15 times the finite-pool benchmark, so the finite-pool version is the closer description and the small remainder is consistent with variation of the fitted model across training sets.

Table 7: Synthetic datasets: hold-out variance decomposition (medians over the four model configurations; RF share is for 500 trees)

Dataset	n	Bayes accuracy	SD / binomial	SD / finite-pool	Test-sampling share	RF internal share
SYN_n100	100	0.80	0.84	0.99	0.93	0.25
SYN_n200	200	0.80	0.89	1.06	0.97	0.35
SYN_n400	400	0.80	0.91	1.09	0.95	0.09
SYN_n700	700	0.80	0.96	1.15	0.98	0.11
SYN_n1000	1000	0.80	0.81	0.97	0.94	0.17
SYN_nonlinear	500	0.85	0.88	1.05	0.89	0.18
SYN_weak	300	0.65	0.86	1.03	0.93	0.27
SYN_highdim	150	0.85	0.96	1.14	0.91	0.22

Table 8 lists the claims, the criteria fixed before the full run, and the outcome. Nine of the ten pre-defined criteria were met. On the size ladder, the hold-out SD fell with n at a log-log slope of −0.53 (95% confidence interval −0.59 to −0.48), which includes the binomial value of −0.5 (Figure 3). On the weak-signal dataset, where the reference gap between logistic regression and the 500-tree random forest is only 0.004, the order of the two flips in 43% of hold-out seeds and in 45% of CV seeds, so cross-validation does not help when two models are truly equivalent. On the non-linear dataset, where the gap is 0.115, the order never flips under either scheme.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Table 8: Claims tested on synthetic data, with criteria fixed before the full run

Claim	Statistic	Value	Result
C1: hold-out SD is close to the binomial error	median SD / binomial (criterion 0.80 to 1.20)	0.88	Supported
C1b: same, finite-pool corrected	median SD / finite-pool (0.80 to 1.20)	1.05	Supported
C2: test-set sampling dominates hold-out variance	median share (at least 0.50)	0.94	Supported
C3: SD falls as $1/\sqrt{n}$	log-log slope, CI must contain −0.5	−0.53 (−0.59, −0.48)	Supported
C4: CV SD is below hold-out SD	share of cells (must be 1.00)	1.00	Supported
C5: logistic regression and unpruned tree are deterministic on a fixed split	maximum SD (below 1e-12)	0	Supported
C6: random forest internal randomness shrinks with more trees	share of datasets (at least 0.75), minimum SD above 0	1.00; 0.0019	Supported
C7: the smaller the gap, the more often hold-out flips the order	Spearman correlation (at most −0.60)	−0.69	Supported
C8: hold-out flips the order at least as often as CV	share of pairs (at least 0.80)	0.94	Supported
C9: best-of-K seed selection inflates accuracy; K = 1 is unbiased	optimism at K = 1 / 5 / 10 / 20 (increasing, |K = 1| below 0.01)	−0.015 / 0.034 / 0.049 / 0.063	Not supported
C9b (post hoc): increase over the K = 1 baseline	increase at K = 5 / 10 / 20 (threshold chosen post hoc)	0.049 / 0.064 / 0.077	Supported

Claim C9 failed only through its K = 1 condition. The optimism rose with K as expected, but at K = 1 it averaged −0.015 (negative on seven of eight synthetic datasets, positive on one), whereas an unbiased procedure should give zero. Each synthetic dataset is a single fixed sample, so its idiosyncratic difficulty shifts the accuracy of all seeds in the same direction. A separate check on 300 independent samples per setting (supplementary script) gave a K = 1 deviation between −0.008 and 0.001 for logistic regression and the tree at n = 100, 200, and 400, with standard errors of 0.002 to 0.005, which supports this reading. The post hoc claim C9b removes the offset by measuring the increase over K = 1; its threshold was chosen after seeing the result and should be read with that in mind. On the real datasets, choosing the best of 20 seeds raised hold-out accuracy by 0.06 above the seed mean, against 0.08 on the synthetic datasets.

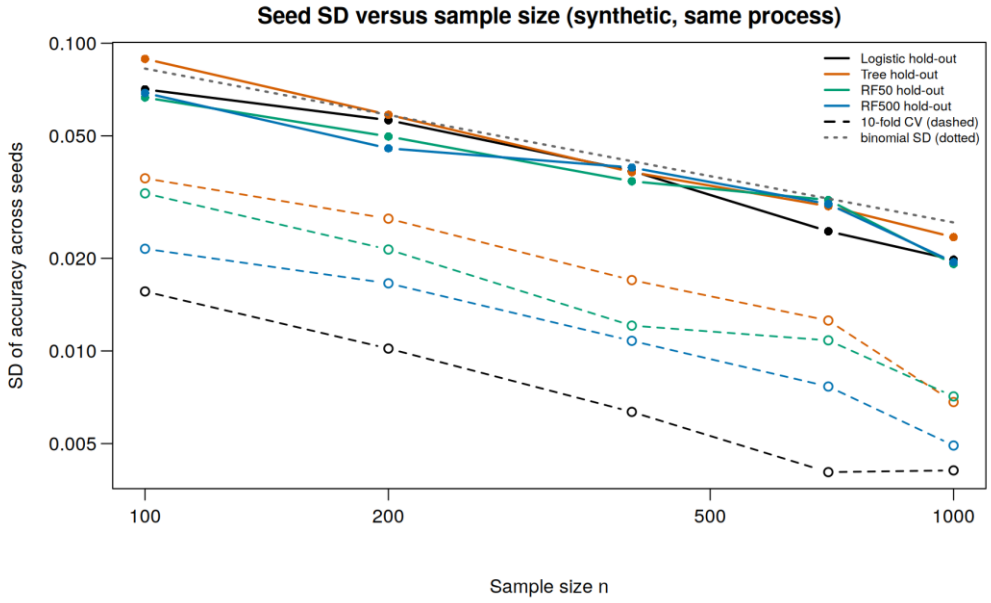


Figure 3: Standard deviation of accuracy across seeds against sample size for the synthetic size ladder (log scales). Solid lines: hold-out; dashed lines: 10-fold CV; dotted line: binomial SD.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

4.2. *Discussion*

Seed sensitivity in this experiment is determined mostly by the validation scheme and the size of the test set, not by the algorithm. Under hold-out, the SD across seeds is almost equal to the sampling error of the test set, and the synthetic data, where the truth is known, attribute a median of 94% of the variance to test-set sampling. This agrees with the argument of Varoquaux (2018) that small samples produce large evaluation errors, and with the finding of Bouthillier et al. (2021) that data sampling is a prominent source of variance. Replacing the model does not reduce this uncertainty; increasing the amount of evaluation does, through cross-validation, repetition, or a larger test set.

Where the true performance gap is large (random forest on Ionosphere and Sonar, or any model on the non-linear synthetic data), almost every seed yields the same ranking. Where it is small (Pima, or the weak-signal synthetic data), one hold-out seed is not enough to say which model is better, and cross-validation does not resolve a gap that is essentially zero, which reinforces the point of Bouthillier et al. (2019) that a conclusion from one experiment may not hold. The advantage of random forest on most datasets is consistent with Couronné et al. (2018), and the near tie on Pima is in line with Christodoulou et al. (2019), although both comparisons here are limited to four datasets and default parameters. The internal randomness of random forest is small relative to split randomness and shrinks with the number of trees, in line with Probst and Boulesteix (2018). With many trees and cross-validation, the model seed is a minor nuisance, but the seed of the data split cannot be ignored.

Seed selection was simulated explicitly: picking the best of 20 seeds by hold-out accuracy raised the apparent accuracy by 0.06 above the seed mean on real data and by 0.08 on synthetic data, and the hold-out accuracy ranges of 0.05 to 0.26 across seeds show how much room there is for bias. Kapoor & Narayanan (2023) and Lones (2024)

warn that leakage of test information into modeling decisions undermines findings, and choosing a seed by test performance is one such form.

Several limitations apply. Only four real and eight synthetic binary datasets of small to medium size were used, so generalization to large, multiclass, imbalanced, or high-dimensional data cannot be claimed, and each synthetic dataset is a single fixed sample. Sixty consecutive seeds were used, which is usually considered adequate for the Mersenne-Twister but remains an assumption; the relative precision of an SD estimated from 60 seeds is about 9%. Hyperparameters were left at defaults, and tuning would add a further source of randomness (Probst et al., 2019). The decision tree was not pruned by internal cross-validation; if it were, the seed would influence the final tree. Warnings from glm were issued in every run on Ionosphere, Sonar, and the high-dimensional synthetic dataset and in no run on the other nine datasets, so logistic regression on those three datasets was fitted under separation or non-convergence and its accuracy there should be read accordingly. The cross-validation scheme used no feature selection or tuning; studies of small samples report that K-fold CV can be biased when such steps are pooled with the data (Vabalas et al., 2019), which this design does not address. The classification threshold was fixed at 0.5, and F1 refers to the positive class only. The criteria in Table 8 were set by the author, one was loosened from 0.005 to 0.01 after a 2-seed trial and before the full run, and C9b is post hoc. The share of internal randomness comes from a single split.

5. Conclusion

Across four real and eight synthetic binary datasets, three models, and 60 seeds, the variation in hold-out performance across seeds largely reflects the randomness of the test set; on synthetic data with known truth, test-set sampling accounted for a median of 94% of it, and the SD fell approximately as the inverse square root of sample size. Ten-fold cross-validation reduces the variation by a factor of 1.8 to 12. Logistic regression and the unpruned decision tree have no internal randomness, while random forest has a smaller internal randomness

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

that shrinks as trees are added. Model rankings reverse across seeds when true differences are small, and cross-validation does not help when two models are truly equivalent. Choosing a seed by test performance inflates the apparent accuracy. Reporting should therefore rely on many seeds or repeated cross-validation, record the R version and RNG settings, and avoid conclusions drawn from a single seed.

References

Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? Journal of the American Statistical Association, 119(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. Journal of Machine Learning Research, 18(77), 1–36.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepahvand, M. P., Raff, E., Madan, K., Voleti, V., Kahou, S. E., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., & Vincent, P. (2021). Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems, 3, 447–469.

Bouthillier, X., Laurent, C., & Vincent, P. (2019). Unreproducible research is reproducible. Proceedings of the 36th International Conference on Machine Learning, 97, 725–734.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. BMC Bioinformatics, 19, 270. https://doi.org/10.1186/s12859-018-2264-5

Gundersen, O. E., & Kjensmo, S. (2018). State of the art: Reproducibility in artificial intelligence. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11503

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. Nature Methods, 18, 1132–1135. https://doi.org/10.1038/s41592-021-01256-7

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11694

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Lones, M. A. (2024). Avoiding common machine learning pitfalls. Patterns, 5(10), 101046. https://doi.org/10.1016/j.patter.2024.101046

Madhyastha, P., & Jain, R. (2019). On Model Stability as a Function of Random Seed. Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 929–939. https://doi.org/10.18653/v1/K19-1087

Picard, D. (2021). Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. ArXiv Preprint ArXiv:2109.08203.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). Journal of Machine Learning Research, 22(164), 1–20.

Probst, P., & Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. Journal of Machine Learning Research, 18(181), 1–18.

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.

https://doi.org/10.35877/jetech5583

ISSN : 2776-7914 (Print) / 2807-3045 (Online)

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/widm.1301

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. ArXiv Preprint ArXiv:1811.12808.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PLoS ONE, 14(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage, 180, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1–17. https://doi.org/10.18637/jss.v077.i01

This open access article is distributed under a Creative Commons Attribution (CC-BY-NC) 4.0 license.