(Research Article)

Model Selection Stability under Limited Training Samples: An Empirical Evaluation across Multiple Datasets

Wannaporn Sarapugdi[1,*]

[1]Program in Management of Digital Innovation, Faculty of Science and Technology, Rajabhat Maha Sarakham University, Maha Sarakham 44000, Thailand; Email: [1]wannaporn.sa@rmu.ac.th

[*]*Corresponding author:* Wannaporn Sarapugdi (wannaporn.sa@rmu.ac.th)

Received: 22 August 2026; **Revised:** 8 October 2026; **Accepted:** 12 October 2026; **Published:** 15 November 2026

Abstract

Comparative machine-learning studies often treat the algorithm with the highest observed score as the preferred model, although that choice may depend on the observations available for training. This study examined model-selection stability under limited training samples across six public classification datasets and five algorithms: Logistic Regression, an RBF-kernel SVM, Random Forest, XGBoost, and LightGBM. For each dataset, 30% of the observations were retained as a fixed evaluation set and the remaining 70% formed a development pool. Training fractions of 10%, 20%, 40%, 60%, 80%, and 100% were evaluated in 30 runs per condition. For fractions below 100%, training composition varied across stratified subsets; the 100% condition reused the complete development pool and served as the reference condition. Macro-F1 was the primary model-selection criterion. Dominant winner rate and normalised selection entropy summarized selection stability, while selection regret compared limited-data choices with the full-data reference. Paired Wilcoxon signed-rank tests with Holm correction were used to assess support for the mean-leading algorithm within the experimental protocol. Within-fraction winner agreement was comparatively high in several datasets, but this did not always mean that the same algorithm remained dominant as more training data became available. Landsat changed from Random Forest at 10–20% to LightGBM from 40% onward, whereas Dry Bean changed from Logistic Regression at 10% to SVM-RBF at intermediate fractions and XGBoost at 80–100%. Mean selection regret declined from 0.0097 to 0.0002 for Landsat and from 0.0058 to 0.0008 for Dry Bean between the 10% and 80% conditions. In five limited-data conditions, the mean-leading algorithm was not statistically supported against all competitors after multiplicity correction. The findings show that both training-sample size and composition can change the selected model and the evidence supporting that choice. They describe model-selection stability under training-sample variation conditional on a single fixed evaluation set.

Keywords: Limited Training Data, Model-Selection Stability, Repeated Sampling, Selection Regret, Statistical Comparison

Copyright © 2026 Journal of Computer and Creative Technology. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-ND).

Citation:
Sarapugdi, W. (2026). Model Selection Stability under Limited Training Samples: An Empirical Evaluation across Multiple Datasets. *Journal of Computer and Creative Technology, 4*(3), e4409. https://doi.org/10.65205/jcct.2026.e4409

(Research Article)

1. Introduction

Machine-learning research frequently compares several algorithms and reports the one with the highest Accuracy, F1-score, or related metric as the preferred model. Such rankings are easy to communicate, but they implicitly assume that the selected winner would remain similar if the experiment were repeated with another training sample from the same data source. This assumption is particularly uncertain when available training data are limited because a small sample may underrepresent difficult cases, minority classes, or local feature patterns. Consequently, sampling variation can affect not only absolute performance but also the identity of the first-ranked algorithm.

Previous work has established that sample size affects the variability and reliability of predictive performance. Rajput et al. (2023) showed that classification results can vary substantially at small sample sizes, while Silvey & Liu (2024) demonstrated that the amount of data needed to approach full-data performance differs across algorithms and datasets. Bornschein et al. (2020) further examined model selection in the small-data regime and showed that relative rankings can contain useful information even when only subsets are available. More recent work has considered efficient validation and method selection under restricted observations (Savvides et al., 2023; Alwash et al., 2025). That leads to a more specific question: if an equally sized training sample is drawn from the same source, does the selected algorithm change?

To examine this question, we use a controlled repeated-sampling design in which the evaluation set and model configurations remain fixed while the training sample changes. In this paper, model-selection stability (Reproducibility of Model Selection under Training-Sample Variation) refers to whether the selected algorithm remains the same when training samples of the same nominal size are repeatedly drawn from the same development pool. It does not refer to reproducibility across datasets, researchers, software implementations, computational environments, or alternative evaluation partitions. Accordingly, the results are conditional on the fixed evaluation set used in the experiment.

We present an empirical framework for examining model-selection stability under limited training data from three complementary perspectives. Winner frequency and normalised selection entropy describe how consistently an algorithm is selected; selection regret quantifies the difference from the full-data reference; and corrected paired comparisons assess statistical support for the mean-leading algorithm. The framework is applied across multiple classification datasets to distinguish numerical leadership from selection stability and statistical evidence rather than to introduce a new classifier or learning algorithm.

2. Research Objectives

1. To examine how training-sample size and composition affect model-selection stability, including winner agreement across repeated samples and changes in the dominant algorithm as the amount of training data increases.

2. To quantify, through selection regret, how far limited-data selections fall below the full-data reference model in full-data Macro-F1.

3. To assess the statistical support for the mean-leading algorithm through paired comparisons with multiplicity correction under the specified experimental protocol.

(Research Article)

3. Literature Review

3.1 Limited Training Data and Performance Variability

Learning-curve and sample-size studies show that predictive performance depends on both the amount and composition of the available training data. Rajput et al. (2023) reported substantial variation in classification performance at small sample sizes, while Silvey & Liu (2024) found that the amount of data required to approach full-data performance differed across algorithms and datasets. Uddin & Lu (2024) likewise showed that dataset-level characteristics affect learning algorithms differently. These studies establish that limited data can increase performance variability, but performance variability alone does not indicate whether the identity of the preferred algorithm remains stable. We therefore ask a related but distinct question: when samples of the same nominal size are repeatedly drawn, does the model-selection decision remain the same?

3.2 Model Selection and Ranking Reliability

Model selection under restricted data has been examined from the perspectives of ranking reliability and efficient validation. Bornschein et al. (2020) showed that relative model rankings obtained from small subsets can still provide useful information, although their reliability changes with the amount of available data. Savvides et al. (2023) considered the validation data required for reliable model selection and proposed bootstrap validation as a data-efficient strategy. Alwash et al. (2025) further demonstrated that method rankings based on limited observations may be unstable or misleading. In comparative benchmarking, Wiesenfarth et al. (2021) emphasized that rankings should be interpreted together with measures of ranking stability and statistical uncertainty rather than from rank order alone. These studies show why numerical rankings need to be considered alongside their variability. However, comparatively little attention has been given to examining, within one controlled design, winner agreement across repeated training samples, changes in the dominant model as training size increases, selection regret relative to a full-data reference, and statistical support for the mean-leading algorithm. We examine these four aspects together.

3.3 Repeated Experiments and Statistical Interpretation

Statistical comparison of learning algorithms becomes more difficult when performance estimates are obtained from repeated or overlapping samples. Dietterich (1998) showed that commonly used tests can produce misleading significance results when dependence induced by repeated train-test sampling is ignored. Nadeau & Bengio (2003) further demonstrated that variability arising from the choice of training set must be considered when making inference about generalization error and that repeated validation estimates cannot generally be treated as independent observations. The present experiment has a related dependence structure: training subsets drawn from the same development pool overlap across repetitions, and all models are evaluated on one fixed evaluation set. The paired Wilcoxon signed-rank tests used here are therefore interpreted as within-protocol comparative evidence rather than population-level significance tests. Holm adjustment was applied across the ten pairwise comparisons within each dataset-fraction condition; it does not remove dependence introduced by the sampling protocol.

4. Conceptual Framework

The framework links training-sample size and composition to variation in Macro-F1, algorithm rankings, and model-selection stability. Training-sample size is varied through six predefined fractions of the development pool (10%, 20%, 40%, 60%, 80%, and 100%), while training-sample composition varies

(Research Article)

across 30 repeated stratified samples at each fraction below 100%. All candidate algorithms are evaluated using the same fixed evaluation set and predefined model configurations.

Changes in training-sample size and composition may alter the Macro-F1 scores of the candidate algorithms and, consequently, their relative rankings and winner identity. We examine model-selection stability using several complementary measures. Dominant winner rate and normalised selection entropy describe agreement in winner identity across repeated samples. Changes in the dominant algorithm across training fractions show whether the same model persists as more training data become available. Selection regret measures how far a selected model lies from the full-data reference, while multiplicity-corrected paired comparisons assess statistical support for the mean-leading algorithm. All measures are interpreted within the fixed evaluation set and specified experimental protocol. These relationships are summarized in Figure 1.

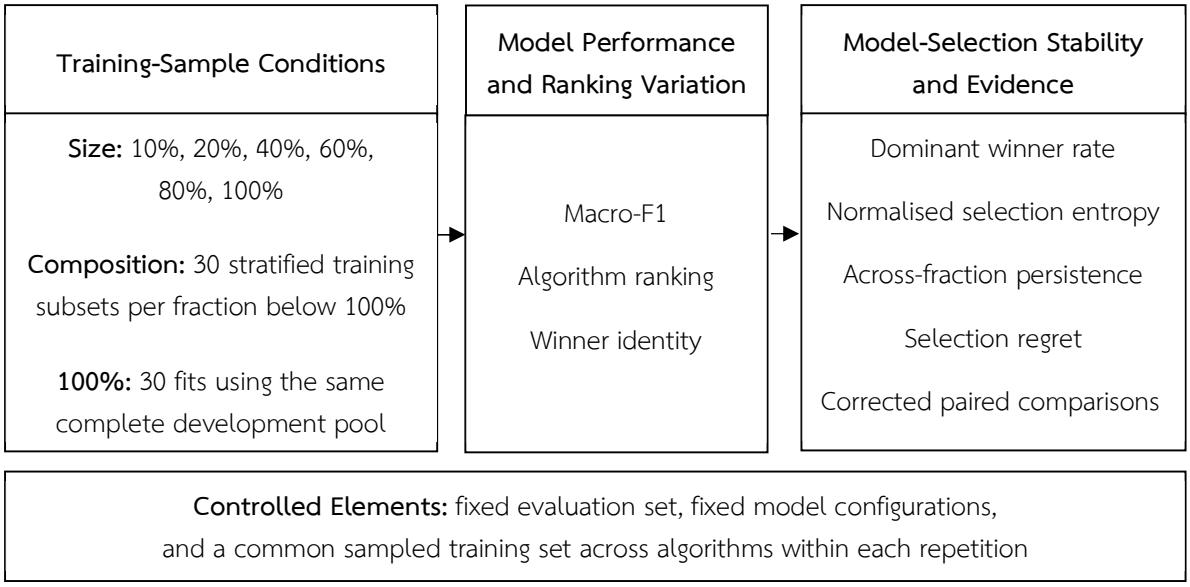


Figure 1. Conceptual Framework Linking Training Sample Size and Composition to Model-Performance Variation, Ranking Variation, and Model-Selection Stability

Note: Developed by the Author; Visual Layout Prepared with AI Assistance and Reviewed by the Author

5. Research Methodology

5.1 Datasets and Preprocessing

Six datasets from the UCI Machine Learning Repository were selected to vary sample size, dimensionality, class count, and application domain: Banknote Authentication (Lohweg, 2012), Rice (Cammeo and Osmancik) (UCI Machine Learning Repository, 2019), *Dry Bean* (UCI Machine Learning Repository, 2020b), Statlog Landsat Satellite (Srinivasan, 1993), Optical Recognition of Handwritten Digits (Alpaydin & Kaynak, 1998), and *AI4I 2020 Predictive Maintenance* (UCI Machine Learning Repository, 2020a). Dataset sizes ranged from 1,372 to 13,611 observations and from two to ten classes. All datasets were publicly available. Dataset characteristics and fixed split sizes are summarized in Table 1.

Table 1. Dataset Characteristics and Fixed Development/Evaluation Split Sizes

Dataset	N	Feat.	Cls.	Dev. pool	Eval. Set
Banknote	1,372	4	2	960	412
Rice	3,810	7	2	2,667	1,143
Dry Bean	13,611	16	7	9,527	4,084
Landsat	6,435	36	6	4,504	1,931
OptDigits	5,620	64	10	3,934	1,686
AI4I	10,000	6⟶8	2	7,000	3,000

Note: AI4I Contained Six Original Predictive Variables, and One-Hot Encoding of Type Produced Eight Final Model-Input Columns, while Failure-Mode Indicators (TWF, HDF, PWF, OSF, and RNF) were Not Used as Predictors

For AI4I, Machine failure was used as the binary target. The predictor matrix contained five continuous process variables-Air temperature, Process temperature, Rotational speed, Torque, and Tool wear-together with the categorical Type variable. One-hot encoding of Type produced three indicators (Type H, Type L, and Type M), giving eight final model-input columns. The five failure-mode indicators TWF, HDF, PWF, OSF, and RNF were contained in the target metadata and were not included as predictors. Identifier variables were also excluded from the model input. A direct preprocessing audit confirmed that none of these failure-mode or identifier variables entered the final feature matrix. Of the 10,000 observations, 339 (3.39%) had Machine failure = 1. StandardScaler was included in the LR and SVM-RBF pipelines, and Macro-F1 was retained as the primary metric so that each class contributed equally to the evaluation.

5.2 Experimental Design

For each dataset, one stratified evaluation set containing 30% of the observations was created using master split seed 2026, while the remaining 70% formed the development pool. This split was fixed throughout the experiment. Training subsets corresponding to 10%, 20%, 40%, 60%, 80%, and 100% of the development pool were then evaluated. For each fraction below 100%, 30 stratified training subsets were generated using repetition seeds 42–71. Within a repetition, the same sampled observations were supplied to all five algorithms, and their predictions were evaluated on the same fixed evaluation set. The complete design comprised 6 datasets × 6 training fractions × 30 repetitions × 5 models = 5,400 model fits.

The repeated training subsets were drawn from the same development pool and were not mutually exclusive, so subsets from different repetitions could overlap. Because of this overlap, the repetitions are not statistically independent samples from a population. They were used to show how the selected model changed when training-sample composition varied under a controlled setting. The fixed evaluation set served as a common benchmark for Macro-F1, the numerical winner in each repetition, the full-data reference, and paired comparisons. It was not an independent post-selection test set. All comparisons are conditional on this fixed evaluation sample.

At the 100% condition, the complete development pool was used in every repetition. Training-sample composition therefore did not vary at this condition; repeated fits reflected only algorithmic

randomness where applicable. The 100% condition was treated as a reference condition, not as evidence of sampling-based model-selection stability, and was used primarily to define the full-data reference model for the regret analysis.

5.3 Candidate Algorithms

Five established algorithms were selected to represent linear, kernel, bagged-tree, and boosted-tree learning families: Logistic Regression (LR), SVM with an RBF kernel (Cortes & Vapnik, 1995), Random Forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), and LightGBM (Ke et al., 2017). Their fixed configurations are listed in Table 2.

Table 2. Fixed Configurations Used for the Candidate Algorithms

Model	Configuration
LR	StandardScaler; max_iter = 4,000
SVM-RBF	StandardScaler; C = 1.0; gamma = scale
RF	300 trees; default depth; n_jobs = -1
XGBoost	300 estimators; max_depth = 6; learning_rate = 0.05
LightGBM	300 estimators; learning_rate = 0.05

All configurations were fixed before the main experiment and were not changed in response to the observed results. Keeping these configurations fixed reduced variation arising from repeated hyperparameter optimization and allowed the analysis to focus on changes associated with training-sample size, training-sample composition, and algorithmic randomness. The conclusions therefore concern these predefined configurations rather than the globally optimized potential of each algorithm family.

5.4 Evaluation Measures

Macro-F1 was the primary selection metric because it assigns equal weight to each class. For class c, precision, recall, and class-specific F1 are defined as (1) – (4)

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{1}$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

$$F1_c = 2\frac{P_c R_c}{P_c + R_c} \tag{3}$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{4}$$

Four related terms are used throughout the analysis. A numerical winner is the algorithm attaining the highest Macro-F1 within a single repetition; exact ties are retained as a winner set. The dominant winner is the algorithm appearing most frequently in the numerical winner sets across the 30 repetitions at a given training fraction. Separately, the mean-leading algorithm is the algorithm with the highest mean

Macro-F1 across the 30 repetitions. Statistical support for the mean-leading algorithm is assessed in Section 5.5. These terms are kept separate because a frequently selected model, the model with the highest mean score, and a model supported by corrected paired comparisons need not be identical.

Let W(q,r) denote the set of algorithms attaining the highest Macro-F1 at training fraction q in repetition r. When an exact tie occurred, all tied algorithms were included in W(q,r). The dominant winner rate (DWR) was the largest proportion of the R = 30 repetitions in which any one algorithm belonged to the winner set; I(·) denotes the indicator function.

$$\mathrm{DWR}(q) = \max_{m \in M} \left[\frac{1}{R} \sum_{r=1}^{R} I\left(m \in W_{q,r}\right) \right] \tag{5}$$

Winner dispersion was additionally summarised by normalised Shannon entropy (Shannon, 1948). For model m at training fraction q, p(m,q) is the proportion of all winner assignments attributable to model m among K = 5 candidates. Under this definition, an exact tie contributes one assignment to each tied algorithm before the proportions are normalised.

$$\mathrm{H_{norm}(q)} = -\frac{\sum_{m=1}^{K} p_{m,q} \ln p_{m,q}}{\ln K} \tag{6}$$

Values near zero indicate concentration on one model; larger values indicate more dispersed winner identities. Entropy provides a descriptive summary of winner dispersion.

5.5 Full-Data Reference, Selection Regret, and Statistical Support

For each dataset, the full-data reference model m* was defined as the algorithm with the highest mean Macro-F1 across the repeated fits using 100% of the development pool. Selection regret quantified the consequence of a limited-data selection relative to the full-data reference.

$$\mathrm{Regret(q,r)} = \bar{F}^{\mathrm{full}}_{m^*} - \frac{1}{|W_{q,r}|} \sum_{m \in W_{q,r}} \bar{F}^{\mathrm{full}}_{m} \tag{7}$$

Here, W(q,r) is the set of numerical winners at training fraction q in repetition r, and the full-data mean for algorithm m is its mean Macro-F1 across the 30 fits using the complete development pool. For a unique winner, |W(q,r)| = 1; for an exact tie, the full-data reference performances of the tied winners are averaged. Smaller regret indicates that the algorithm selected under limited data has a full-data reference performance closer to that of the reference-best algorithm. Selection regret is a conditional, reference-based quantity and is not interpreted as an unbiased estimate of post-selection or deployment generalization loss.

Statistical support was assessed from paired Macro-F1 values because all five algorithms within a repetition were trained on the same sampled observations and evaluated on the same fixed evaluation set. For each dataset and training fraction, all ten model pairs were compared using two-sided Wilcoxon signed-rank tests (Wilcoxon, 1945), followed by Holm correction (Holm, 1979) across the ten comparisons at α = 0.05. The algorithm with the highest mean Macro-F1 was termed the mean-leading algorithm. Mean Macro-F1 identifies the leading algorithm; the signed-rank tests assess paired score differences rather than a hypothesis about arithmetic means. A mean-leading algorithm was described as statistically supported within the experimental protocol only when its mean Macro-F1 was higher than that of every

(Research Article)

competitor and all four corresponding Holm-adjusted pairwise comparisons satisfied the significance threshold.

The repeated training subsets overlap and the same evaluation set is reused across repetitions; therefore, these p-values are not interpreted as population-level significance tests based on independent experimental samples. They are used only as within-protocol comparative evidence. When the mean-leading algorithm did not satisfy all four corrected comparisons, the result is reported as insufficient evidence of superiority under the specified paired testing and multiplicity-correction procedure. This wording does not imply equivalence between models or establish that no model is better in the underlying population.

5.6 Experimental Records, Statistical Scope, and Ethical Considerations

The same master split seed, training fractions, repetition seeds, model configurations, primary metric, and statistical procedure were applied consistently across all six datasets. Run-level model scores, winner assignments, selection-regret values, pairwise comparisons, and summary statistics were retained as CSV outputs to support verification of the reported results. Agreement across alternative software environments, evaluation partitions, or independently collected datasets was not assessed. No new human-participant data were collected, and all experiments used publicly available benchmark datasets.

6. Results

6.1 Model-Selection Stability under Training-Sample Variation (Objective 1)

Model-selection stability varied across datasets and training fractions. At the 10% condition, the dominant winner appeared in the winner set in 26/30 repetitions (86.7%) for Banknote, Landsat, and AI4I, 23/30 (76.7%) for OptDigits, 18/30 (60.0%) for Rice, and 17/30 (56.7%) for Dry Bean. Rice showed its lowest within-fraction agreement at 20%, where LR was the dominant winner in 13/30 repetitions (43.3%). Dry Bean remained comparatively dispersed across several fractions, with dominant-winner counts of 20/30 (66.7%) at 20%, 19/30 (63.3%) at 40%, 16/30 (53.3%) at 60%, and 17/30 (56.7%) at 80%. Figure 2 shows the corresponding dominant winner rates across training fractions.

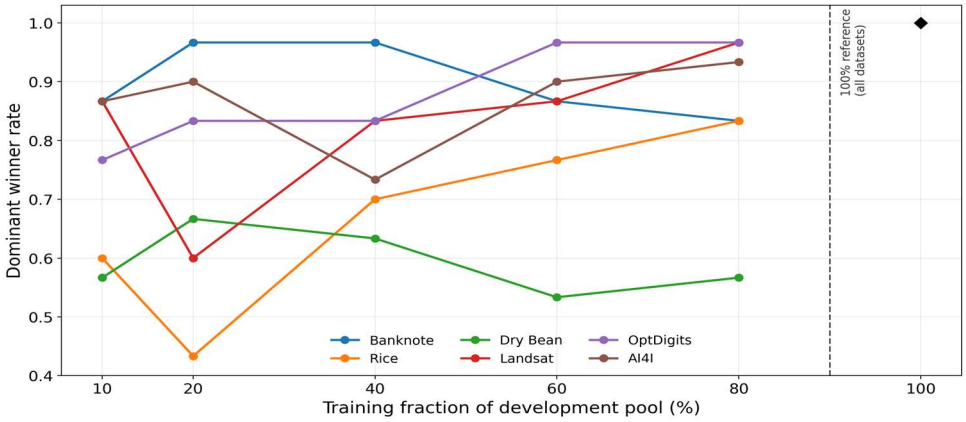


Figure 2. Dominant Winner Rate across Training-Sample Fractions, where Values from 10% to 80% Summarize Winner Agreement across Repeated Stratified Training Samples and the Neutral Marker at 100% denotes the Common Full-Development-Pool Reference for All Datasets

(Research Article)

Normalised selection entropy provided a complementary measure of winner dispersion. Rice reached an entropy of 0.638 at 20%, while Dry Bean reached 0.644 at 40%. For Landsat, entropy decreased from 0.418 at 20% to 0.091 at 80%, accompanying the increasing concentration on LightGBM. The trajectories were not uniformly monotonic across datasets, indicating that greater training-sample size did not produce a smooth increase in winner agreement in every case. Figure 3 shows the corresponding entropy trajectories.

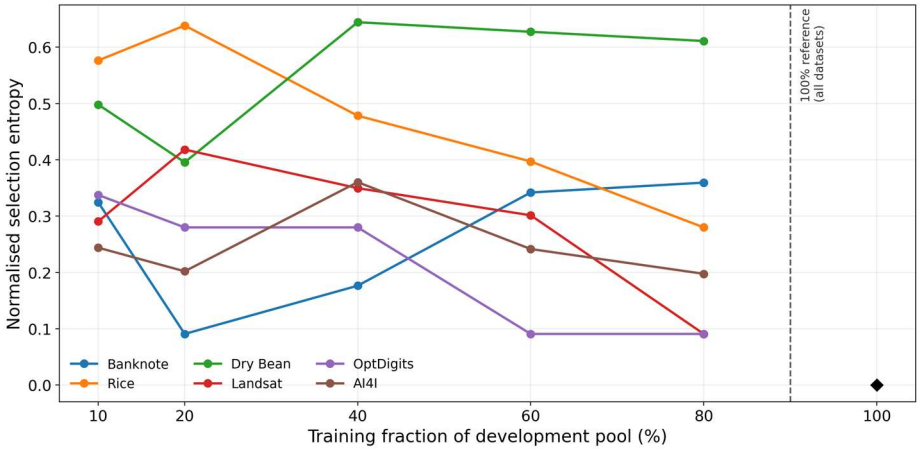


Figure 3. Normalised Selection Entropy across Training-Sample Fractions, where Values from 10% to 80% Reflect Dispersion of Winner Identity under Repeated Training-Sample Variation and the Neutral Marker at 100% denotes the Common Full-Development-Pool Reference for All Datasets

Within-fraction agreement did not necessarily imply persistence of the same dominant algorithm across training fractions. Dry Bean changed from LR at 10% to SVM-RBF at 20–60% and XGBoost at 80%. Landsat changed from RF at 10–20% to LightGBM from 40% onward. In contrast, SVM-RBF remained the dominant algorithm across the evaluated fractions for Banknote and OptDigits, LR for Rice, and LightGBM for AI4I. At the 100% reference condition, where the complete development pool was identical in all repetitions, each dataset selected the same dominant model in 30/30 runs. Exact ties were uncommon: Banknote had one tied repetition at each of the 10%, 40%, 60%, and 80% conditions, and AI4I had one tied repetition at 80%. Selected full-data references and limited-data summaries are provided in Table 3.

Table 3. Full-Data Reference Models, Dominant Winner Counts, and Selection-Regret Summaries at Selected Limited-Data Conditions, where Regret Values are Mean ± SD across 30 Repetitions

Dataset	Full-Data Ref.	10% Dominant Winner	Regret 10%, Mean ± SD	80% Dominant Winner	Regret 80%, Mean ± SD
Banknote	SVM-RBF (1.0000)	SVM-RBF 26/30 (86.7%)	0.0021 ± 0.0056	SVM-RBF 25/30 (83.3%)	0.0005 ± 0.0010
Rice	LR (0.9223)	LR 18/30 (60.0%)	0.0024 ± 0.0031	LR 25/30 (83.3%)	0.0009 ± 0.0020

(Research Article)

Table 3. Full-Data Reference Models, Dominant Winner Counts, and Selection-Regret Summaries at Selected Limited-Data Conditions, where Regret Values are Mean ± SD across 30 Repetitions (Cont.)

Dataset	Full-Data Ref.	10% Dominant Winner	Regret 10%, Mean ± SD	80% Dominant Winner	Regret 80%, Mean ± SD
Dry Bean	XGBoost (0.9358)	LR 17/30 (56.7%)	0.0058 ± 0.0033	XGBoost 17/30 (56.7%)	0.0008 ± 0.0010
Landsat	LightGBM (0.9112)	RF 26/30 (86.7%)	0.0097 ± 0.0052	LightGBM 29/30 (96.7%)	0.0002 ± 0.0013
OptDigits	SVM-RBF (0.9840)	SVM-RBF 23/30 (76.7%)	0.0007 ± 0.0013	SVM-RBF 29/30 (96.7%)	0.0001 ± 0.0005
AI4I	LightGBM (0.8855)	LightGBM 26/30 (86.7%)	0.0032 ± 0.0084	LightGBM 28/30 (93.3%)	0.0020 ± 0.0064

Note: The 100% Condition Defines the Full-Development-Pool Reference and is Not Included as a Repeated Training-Sample Condition in This Table

Complete mean Macro-F1 and standard deviation values for all five candidate algorithms across every dataset and training fraction are provided in Appendix Table A1.

Exact dominant-winner counts, normalised selection entropy, and selection-regret summaries for all training fractions are reported in Appendix Table A2.

6.2 Selection Regret Relative to the Full-Data Reference (Objective 2)

The result is shown in Figure 4.

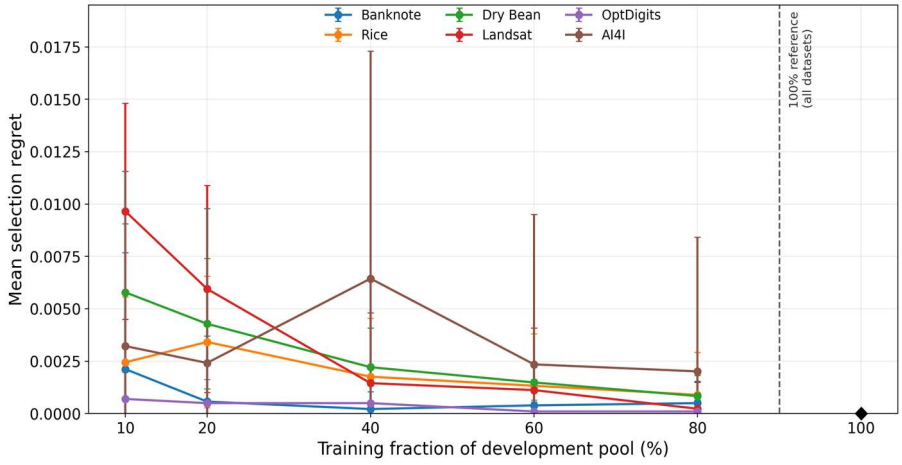


Figure 4. Mean Selection Regret Relative to the Full-Data Reference Model across Training-Sample Fractions, where Error Bars Show ±1 SD across 30 Repetitions, with Lower Error Bars Truncated at Zero for Display Because Selection Regret is Non-Negative, and the Neutral Marker at 100% denotes the Common Full-Development-Pool Reference Endpoint for All Datasets

(Research Article)

The selection regret associated with changes in winner identity differed across datasets. Landsat had the largest mean regret at the 10% condition, at 0.0097 ± 0.0052, which decreased to 0.0060 ± 0.0049 at 20%, 0.0015 ± 0.0034 at 40%, and 0.0002 ± 0.0013 at 80%. Dry Bean similarly decreased from 0.0058 ± 0.0033 at 10% to 0.0008 ± 0.0010 at 80%. In contrast, OptDigits had a mean regret of only 0.0007 ± 0.0013 at 10%, indicating that changes in the numerical winner corresponded to relatively small differences in full-data reference performance. Figure 4 shows the mean selection regret and its variability across fractions.

The relationship was not monotonic for every dataset. For AI4I, mean regret was 0.0032 ± 0.0084 at 10%, decreased to 0.0024 ± 0.0074 at 20%, increased to 0.0064 ± 0.0109 at 40%, and subsequently declined at the larger fractions. Model-selection agreement and selection regret should therefore be considered separately: a change in winner identity may correspond to either a small or a comparatively larger difference from the full-data reference. At 100%, observed selection regret was zero for all datasets because the same reference-best algorithm was selected in all 30 runs of the full-development-pool condition.

6.3 Statistical support for the mean-leading algorithm (Objective 3)

In five limited-data conditions, the mean-leading algorithm did not receive statistical support against all four competitors after Holm correction. For Rice at 20%, LR had the highest mean Macro-F1, but its paired comparison with SVM-RBF yielded a Holm-adjusted p = 0.6984. For Dry Bean at 10%, LR led in mean Macro-F1, but the comparison with SVM-RBF yielded p = 1.0000. At 60%, SVM-RBF was the mean-leading algorithm, while its comparison with XGBoost yielded p = 0.1534. At 80%, XGBoost led in mean Macro-F1, but the corrected comparisons with SVM-RBF (p = 0.1840) and LightGBM (p = 0.0936) did not meet the specified threshold. For Landsat at 20%, RF was the mean-leading algorithm, while its comparison with LightGBM yielded p = 0.1094. For these five conditions, the corrected comparisons provided insufficient evidence of superiority under the specified testing and multiplicity-correction procedure. Across the remaining 25 limited-data conditions, the mean-leading algorithm satisfied the specified criterion for statistical support within the experimental protocol. The five conditions lacking corrected statistical support are summarized in Table 4.

Table 4. Limited-Data Conditions in which the Mean-Leading Algorithm Lacked Corrected Statistical Support Against at Least One Competitor

Dataset	Training Fraction	Mean-Leading Algorithm	Comparison Lacking Corrected Support
Rice	20%	LR	SVM-RBF, p_Holm = 0.6984
Dry Bean	10%	LR	SVM-RBF, p_Holm = 1.0000
Dry Bean	60%	SVM-RBF	XGBoost, p_Holm = 0.1534
Dry Bean	80%	XGBoost	SVM-RBF, p_Holm = 0.1840; LightGBM, p_Holm = 0.0936
Landsat	20%	RF	LightGBM, p_Holm = 0.1094

Note: These Results Indicate Insufficient Evidence of Superiority under the Specified Paired Testing and Holm-Correction Procedure but Do Not Establish Model Equivalence

(Research Article)

7. Conclusion

We examined model-selection stability under controlled variation in training-sample size and composition across six tabular classification datasets. Agreement in the identity of the numerical winner varied across datasets and training fractions, and changes in winner identity differed in their selection regret relative to the full-data reference. In five limited-data conditions, the mean-leading algorithm also lacked corrected statistical support against at least one competitor. Under limited data, reporting only the highest score leaves out information about selection consistency, selection regret, and statistical support. The findings are conditional on one fixed evaluation set, overlapping training subsets drawn from the same development pool, and predefined model configurations; they do not establish which algorithm would be superior under new evaluation samples, alternative hyperparameter choices, or other task domains.

8. Discussion

In this experiment, model-selection stability under limited training data varied with both the dataset and the training fraction. Variation in training-sample composition affected not only Macro-F1 values but, in some settings, the identity of the selected algorithm. Dry Bean and Landsat provide the clearest examples of changes across training fractions. Dry Bean moved from LR at 10% to SVM-RBF at intermediate fractions and XGBoost at 80%, while Landsat moved from RF at 10–20% to LightGBM from 40% onward. In contrast, Banknote, Rice, OptDigits, and AI4I retained the same dominant algorithm across the evaluated fractions, although the degree of within-fraction winner agreement still varied. High agreement at one training fraction therefore does not necessarily imply that the same algorithm will remain dominant when more training data become available.

Winner frequency and normalised selection entropy describe how consistently an algorithm was selected, whereas selection regret shows how far a different selection lies from the full-data reference. OptDigits, for example, showed occasional changes in winner identity while maintaining very small regret. Landsat at 10%, in contrast, combined a relatively concentrated winner pattern with a larger difference from the full-data reference. Regret therefore adds information that winner consistency alone does not provide.

The paired comparisons separate numerical leadership from statistical support. Five limited-data conditions had a mean-leading algorithm that was not supported against all four competitors after Holm correction. These cases indicate insufficient evidence of superiority under the specified testing procedure; they do not show that the competing models are equivalent. Because the training subsets overlap across repetitions and every model is evaluated on the same fixed evaluation set, the corrected p-values provide comparative evidence within this protocol rather than population-level significance guarantees.
Agreement at the 100% condition reflects repeated fits on the same complete development pool rather than variation across alternative training samples. Its primary role here is to provide the full-data reference for comparison with the limited-data conditions.

The findings add to earlier work on sample-size effects and small-data model selection by focusing on the stability of the selection decision itself. Earlier studies have shown that predictive performance and data requirements vary across learning problems and algorithms (Rajput et al., 2023; Silvey & Liu, 2024; Uddin & Lu, 2024), while work on model selection has emphasized the difficulty of making reliable decisions from limited observations (Bornschein et al., 2020; Savvides et al., 2023; Alwash et al., 2025).

Here, repeated winner identity, winner dispersion, selection regret, and corrected pairwise evidence are considered within the same controlled setting.

The findings should be read within several limitations. First, one fixed evaluation set was reused throughout each dataset. This choice allowed training-sample variation to be studied under a common benchmark, but the reported rankings and statistical comparisons remain conditional on that evaluation partition. The study did not examine how the conclusions would change under alternative evaluation splits. Second, training subsets drawn from the same development pool can overlap, so repeated observations are not strictly independent. Third, the candidate configurations were fixed in advance and were not retuned within each repetition; the results compare the specified configurations rather than the maximum attainable performance of each algorithm family. Finally, all six datasets are tabular classification problems, so the findings should not be assumed to extend directly to deep learning, regression, time-series, or other high-dimensional settings. Future work should examine multiple outer evaluation partitions, nested hyperparameter optimisation, alternative performance metrics, and additional task families.

9. Recommendation

With limited training data, repeated training-sample evaluation can reveal whether model choice is sensitive to sample composition. Winner counts should be reported together with performance variability, and selection regret can help distinguish small ranking changes from more consequential ones. Statistical comparisons should be interpreted in light of the dependence introduced by overlapping training samples and reused evaluation data.

10. Declaration of Generative AI

ChatGPT (OpenAI) was used as an assistive tool for code-development support and, during manuscript preparation, for English-language editing, manuscript organisation, figure and table preparation, and consistency checking against the experimental outputs. AI-assisted code was reviewed, executed, and verified by the author before use. The author designed the study, defined the experimental protocol, conducted the experiments, generated and verified the numerical and statistical results, interpreted the findings, and prepared the final scientific conclusions. AI-assisted manuscript content was checked against the retained source data, code outputs, and cited literature. The author takes full responsibility for the final content of the manuscript.

11. CRediT Author Statement

Wannaporn Sarapugdi: Conceptualization; Methodology; Software; Validation; Formal Analysis; Investigation; Data Curation; Writing – Original Draft; Writing – Review & Editing; Visualization; Project Administration.

The author has read and approved the final manuscript.

12. References

Alpaydin, E. & Kaynak, C. (1998). *Optical Recognition of Handwritten Digits* [Data Set]. UCI Machine Learning Repository. https://doi.org/10.24432/c50p49

(Research Article)

Alwash, M., Al Hajj, G. S., Grytten, I., & Sandve, G. K. (2025). Meta Simulation Approach for Evaluating Machine Learning Method Selection in Data Limited Settings. *Scientific Reports*, *15*(1), 40766. https://doi.org/10.1038/s41598-025-24627-y

Bornschein, J., Visin, F., & Osindero, S. (2020, July 13-18). Small Data, Big Decisions: Model Selection in the Small-Data Regime. *The 37th International Conference on Machine Learning*, 119, 1035-1044. PMLR. https://proceedings.mlr.press/v119/bornschein20a.html

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016, August 13-17). XGBoost: A Scalable Tree Boosting System. *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895-1923. https://doi.org/10.1162/089976698300017197

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017, December 4-9). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *The 31st International Conference on Neural Information Processing Systems*, 3149-3157. https://dl.acm.org/doi/10.5555/3294996.3295074

Lohweg, V. (2012). *Banknote Authentication* [Data Set]. UCI Machine Learning Repository. https://doi.org/10.24432/c55p57

Nadeau, C., & Bengio, Y. (2003). Inference for the Generalization Error. *Machine Learning, 52*(3), 239–281. https://doi.org/10.1023/a:1024068626366

Rajput, D., Wang, W.-J., & Chen, C.-C. (2023). Evaluation of a Decided Sample Size in Machine Learning Applications. *BMC Bioinformatics*, *24*(1), 48. https://doi.org/10.1186/s12859-023-05156-9

Savvides, R., Mäkelä, J., & Puolamäki, K. (2023). Model Selection with Bootstrap Validation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *16*(2), 162–186. https://doi.org/10.1002/sam.11606

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(4), 623-656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Silvey, S., & Liu, J. (2024). Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study. *Journal of Medical Internet Research*, *26*, e60231. https://doi.org/10.2196/60231

Srinivasan, A. (1993). *Statlog (Landsat Satellite)* [Data Set]. UCI Machine Learning Repository. https://doi.org/10.24432/c55887

UCI Machine Learning Repository. (2019). *Rice (Cammeo and Osmancik)* [Data Set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MW4Z.

UCI Machine Learning Repository. (2020a). *AI4I 2020 Predictive Maintenance Dataset* [Data Set]. UCI Machine Learning Repository. https://doi.org/10.24432/c5hs5c

UCI Machine Learning Repository. (2020b). *Dry Bean* [Data Set]. UCI Machine Learning Repository. https://doi.org/10.24432/c50s4b

(Research Article)

Uddin, S., & Lu, H. (2024). Dataset Meta-Level and Statistical Features Affect Machine Learning Performance. *Scientific Reports*, *14*(1), 1670. https://doi.org/10.1038/s41598-024-51825-x

Wiesenfarth, M., Reinke, A., Landman, B. A., Eisenmann, M., Saiz, L. A., Cardoso, M. J., Maier-Hein, L., & Kopp-Schneider, A. (2021). Methods and Open-Source Toolkit for Analyzing and Visualizing Challenge Results. *Scientific Reports*, *11*(1), 2369. https://doi.org/10.1038/s41598-021-82017-6

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80-83. https://doi.org/10.2307/3001968

13. Appendix A

Table A1. Mean Macro-F1 ± SD for All Candidate Algorithms Across All Datasets and Training Fractions

Dataset	Fraction	LR	SVM-RBF	RF	XGBoost	LightGBM
Banknote	10%	0.9732 ± 0.0076	0.9840 ± 0.0093	0.9411 ± 0.0180	0.9409 ± 0.0160	0.9537 ± 0.0141
Banknote	20%	0.9795 ± 0.0033	0.9919 ± 0.0046	0.9660 ± 0.0122	0.9676 ± 0.0133	0.9719 ± 0.0115
Banknote	40%	0.9818 ± 0.0021	0.9954 ± 0.0030	0.9863 ± 0.0057	0.9835 ± 0.0081	0.9843 ± 0.0058
Banknote	60%	0.9828 ± 0.0008	0.9975 ± 0.0026	0.9918 ± 0.0040	0.9903 ± 0.0052	0.9875 ± 0.0053
Banknote	80%	0.9829 ± 0.0000	0.9985 ± 0.0023	0.9945 ± 0.0026	0.9935 ± 0.0030	0.9885 ± 0.0042
Banknote	100%	0.9829 ± 0.0000	1.0000 ± 0.0000	0.9972 ± 0.0008	0.9975 ± 0.0000	0.9926 ± 0.0000
Rice	10%	0.9192 ± 0.0043	0.9169 ± 0.0047	0.9123 ± 0.0073	0.9046 ± 0.0095	0.9027 ± 0.0093
Rice	20%	0.9195 ± 0.0028	0.9188 ± 0.0033	0.9133 ± 0.0048	0.9059 ± 0.0061	0.9050 ± 0.0061
Rice	40%	0.9214 ± 0.0021	0.9192 ± 0.0035	0.9159 ± 0.0038	0.9079 ± 0.0039	0.9087 ± 0.0050
Rice	60%	0.9208 ± 0.0018	0.9185 ± 0.0034	0.9161 ± 0.0033	0.9112 ± 0.0036	0.9117 ± 0.0044
Rice	80%	0.9215 ± 0.0016	0.9178 ± 0.0025	0.9161 ± 0.0016	0.9122 ± 0.0033	0.9113 ± 0.0044
Rice	100%	0.9223 ± 0.0000	0.9170 ± 0.0000	0.9147 ± 0.0015	0.9151 ± 0.0000	0.9150 ± 0.0000
Dry Bean	10%	0.9228 ± 0.0021	0.9225 ± 0.0025	0.9155 ± 0.0036	0.9157 ± 0.0034	0.9157 ± 0.0044
Dry Bean	20%	0.9259 ± 0.0015	0.9270 ± 0.0017	0.9199 ± 0.0026	0.9220 ± 0.0033	0.9224 ± 0.0031

(Research Article)

Table A1. Mean Macro-F1 ± SD for All Candidate Algorithms Across All Datasets and Training Fractions (Cont.)

Dataset	Fraction	LR	SVM-RBF	RF	XGBoost	LightGBM
Dry Bean	40%	0.9268 ± 0.0015	0.9296 ± 0.0013	0.9240 ± 0.0021	0.9275 ± 0.0023	0.9271 ± 0.0028
Dry Bean	60%	0.9270 ± 0.0011	0.9310 ± 0.0013	0.9251 ± 0.0017	0.9302 ± 0.0020	0.9295 ± 0.0024
Dry Bean	80%	0.9275 ± 0.0010	0.9327 ± 0.0009	0.9268 ± 0.0009	0.9333 ± 0.0018	0.9320 ± 0.0017
Dry Bean	100%	0.9272 ± 0.0000	0.9337 ± 0.0000	0.9282 ± 0.0011	0.9358 ± 0.0000	0.9340 ± 0.0000
Landsat	10%	0.7876 ± 0.0114	0.8259 ± 0.0112	0.8506 ± 0.0064	0.8366 ± 0.0083	0.8413 ± 0.0094
Landsat	20%	0.7961 ± 0.0084	0.8429 ± 0.0079	0.8704 ± 0.0051	0.8636 ± 0.0081	0.8685 ± 0.0088
Landsat	40%	0.8025 ± 0.0056	0.8614 ± 0.0059	0.8845 ± 0.0045	0.8851 ± 0.0056	0.8909 ± 0.0062
Landsat	60%	0.8064 ± 0.0044	0.8705 ± 0.0038	0.8939 ± 0.0037	0.8955 ± 0.0033	0.9012 ± 0.0050
Landsat	80%	0.8077 ± 0.0037	0.8768 ± 0.0026	0.8982 ± 0.0028	0.9024 ± 0.0029	0.9074 ± 0.0027
Landsat	100%	0.8091 ± 0.0000	0.8795 ± 0.0000	0.9013 ± 0.0017	0.9044 ± 0.0000	0.9112 ± 0.0000
OptDigits	10%	0.9304 ± 0.0061	0.9463 ± 0.0076	0.9408 ± 0.0092	0.9026 ± 0.0095	0.9217 ± 0.0098
OptDigits	20%	0.9415 ± 0.0047	0.9628 ± 0.0049	0.9581 ± 0.0051	0.9347 ± 0.0061	0.9486 ± 0.0060
OptDigits	40%	0.9514 ± 0.0037	0.9737 ± 0.0026	0.9712 ± 0.0024	0.9538 ± 0.0038	0.9647 ± 0.0035
OptDigits	60%	0.9564 ± 0.0031	0.9793 ± 0.0015	0.9757 ± 0.0021	0.9611 ± 0.0025	0.9710 ± 0.0021
OptDigits	80%	0.9603 ± 0.0022	0.9820 ± 0.0012	0.9792 ± 0.0015	0.9663 ± 0.0022	0.9768 ± 0.0021
OptDigits	100%	0.9609 ± 0.0000	0.9840 ± 0.0000	0.9811 ± 0.0010	0.9681 ± 0.0000	0.9787 ± 0.0000
AI4I	10%	0.5755 ± 0.0366	0.5043 ± 0.0221	0.6268 ± 0.0362	0.6705 ± 0.0268	0.7145 ± 0.0416
AI4I	20%	0.5917 ± 0.0239	0.5317 ± 0.0382	0.6937 ± 0.0304	0.7322 ± 0.0338	0.7627 ± 0.0322
AI4I	40%	0.6071 ± 0.0151	0.6017 ± 0.0420	0.7766 ± 0.0192	0.8070 ± 0.0277	0.8285 ± 0.0233

(Research Article)

Table A1. Mean Macro-F1 ± SD for All Candidate Algorithms Across All Datasets and Training Fractions (Cont.)

Dataset	Fraction	LR	SVM-RBF	RF	XGBoost	LightGBM
AI4I	60%	0.6065 ± 0.0145	0.6463 ± 0.0267	0.8232 ± 0.0139	0.8408 ± 0.0156	0.8566 ± 0.0129
AI4I	80%	0.6164 ± 0.0098	0.6764 ± 0.0140	0.8453 ± 0.0089	0.8547 ± 0.0154	0.8736 ± 0.0133
AI4I	100%	0.6226 ± 0.0000	0.6950 ± 0.0000	0.8633 ± 0.0065	0.8613 ± 0.0000	0.8855 ± 0.0000

Note: Values are Mean ± Sample Standard Deviation across 30 Repetitions, with the 100% Rows Representing Full-Development-Pool Reference Conditions in which Training-Sample Composition Did Not Vary across Those Repetitions

Table A2. Complete Model-Selection Stability and Selection-Regret Summary Across All Datasets and Training Fractions

Dataset	Fraction	Dominant Winner	DWR, k/30 (%)	Normalised Entropy	Selection Regret, Mean ± S.D.
Banknote	10%	SVM-RBF	26/30 (86.7%)	0.325	0.0021 ± 0.0056
Banknote	20%	SVM-RBF	29/30 (96.7%)	0.091	0.0006 ± 0.0031
Banknote	40%	SVM-RBF	29/30 (96.7%)	0.176	0.0002 ± 0.0008
Banknote	60%	SVM-RBF	26/30 (86.7%)	0.342	0.0004 ± 0.0009
Banknote	80%	SVM-RBF	25/30 (83.3%)	0.359	0.0005 ± 0.0010
Banknote	100% (reference)	SVM-RBF	30/30 (100.0%)	0.000	0.0000 ± 0.0000
Rice	10%	LR	18/30 (60.0%)	0.576	0.0024 ± 0.0031
Rice	20%	LR	13/30 (43.3%)	0.638	0.0034 ± 0.0031
Rice	40%	LR	21/30 (70.0%)	0.478	0.0018 ± 0.0028
Rice	60%	LR	23/30 (76.7%)	0.397	0.0013 ± 0.0025
Rice	80%	LR	25/30 (83.3%)	0.280	0.0009 ± 0.0020
Rice	100% (reference)	LR	30/30 (100.0%)	0.000	0.0000 ± 0.0000
Dry Bean	10%	LR	17/30 (56.7%)	0.498	0.0058 ± 0.0033
Dry Bean	20%	SVM-RBF	20/30 (66.7%)	0.395	0.0043 ± 0.0031
Dry Bean	40%	SVM-RBF	19/30 (63.3%)	0.644	0.0022 ± 0.0019
Dry Bean	60%	SVM-RBF	16/30 (53.3%)	0.627	0.0015 ± 0.0009
Dry Bean	80%	XGBoost	17/30 (56.7%)	0.611	0.0008 ± 0.0010
Dry Bean	100% (reference)	XGBoost	30/30 (100.0%)	0.000	0.0000 ± 0.0000
Landsat	10%	RF	26/30 (86.7%)	0.291	0.0097 ± 0.0052
Landsat	20%	RF	18/30 (60.0%)	0.418	0.0060 ± 0.0049
Landsat	40%	LightGBM	25/30 (83.3%)	0.350	0.0015 ± 0.0034
Landsat	60%	LightGBM	26/30 (86.7%)	0.301	0.0011 ± 0.0030
Landsat	80%	LightGBM	29/30 (96.7%)	0.091	0.0002 ± 0.0013
Landsat	100% (reference)	LightGBM	30/30 (100.0%)	0.000	0.0000 ± 0.0000
OptDigits	10%	SVM-RBF	23/30 (76.7%)	0.338	0.0007 ± 0.0013

(Research Article)

Table A2. Complete Model-Selection Stability and Selection-Regret Summary Across All Datasets and Training Fractions (Cont.)

Dataset	Fraction	Dominant Winner	DWR, k/30 (%)	Normalised Entropy	Selection Regret, Mean ± S.D.
OptDigits	20%	SVM-RBF	25/30 (83.3%)	0.280	0.0005 ± 0.0011
OptDigits	40%	SVM-RBF	25/30 (83.3%)	0.280	0.0005 ± 0.0011
OptDigits	60%	SVM-RBF	29/30 (96.7%)	0.091	0.0001 ± 0.0005
OptDigits	80%	SVM-RBF	29/30 (96.7%)	0.091	0.0001 ± 0.0005
OptDigits	100% (reference)	SVM-RBF	30/30 (100.0%)	0.000	0.0000 ± 0.0000
AI4I	10%	LightGBM	26/30 (86.7%)	0.244	0.0032 ± 0.0084
AI4I	20%	LightGBM	27/30 (90.0%)	0.202	0.0024 ± 0.0074
AI4I	40%	LightGBM	22/30 (73.3%)	0.360	0.0064 ± 0.0109
AI4I	60%	LightGBM	27/30 (90.0%)	0.242	0.0023 ± 0.0072
AI4I	80%	LightGBM	28/30 (93.3%)	0.198	0.0020 ± 0.0064
AI4I	100% (reference)	LightGBM	30/30 (100.0%)	0.000	0.0000 ± 0.0000

Note: Rows at 100% are Full-Development-Pool Reference Conditions in which Training-Sample Composition Did Not Vary across Those 30 Fits