Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 08 August-2026, Page No.- 11214- 11230

DOI: 10.47191/etj/v11i08.14, I.F. – 8.595

How Classification Models Degrade as Data Diminish: A Multi-Criteria Analysis of Tabular Classifiers Under Controlled Data Scarcity

Jairo Devon A. Daquioag[1], Eliza B. Ayo[2]

[1,2]Centro Escolar University, Manila, Philippines

ABSTRACT: Machine learning research is frequently conducted under an implicit assumption of data abundance, yet applied software development often proceeds under severe data constraints. This study evaluated how five widely used classification algorithms—logistic regression, support vector machines, Gaussian naive Bayes, decision trees, and random forests—degrade when training data are systematically reduced. Eight tabular datasets drawn from the UCI Machine Learning Repository, spanning small, medium, and large volume tiers, were analysed within a quantitative comparative experimental design. Training folds were reduced to 100%, 75%, 50%, and 25% of their original size through stratified fractional sampling inside a stratified five-fold cross-validation loop, while validation folds were retained at full volume. Ten criteria were recorded: accuracy, precision, recall, F1 score, AUC-ROC, AUC-PR, mean absolute error and mean squared error of predicted probabilities, training time, and interpretability. One-way analyses of variance and Tukey honestly significant difference tests were applied at an alpha level of .05. Between-model differences were statistically significant in 31 of 32 dataset-by-volume configurations. Tree-based models exhibited the largest reductions in accuracy and the steepest increases in calibration error under scarcity, whereas logistic regression preserved threshold balance and probability calibration at negligible computational cost. Random forests attained the highest scores only when data were abundant. Pairwise comparisons aggregated across datasets were not significant, indicating that dataset context, rather than architectural complexity alone, governs absolute performance.

KEYWORDS: machine learning, data scarcity, model selection, classification, probability calibration, bias–variance trade-off,tabular data

INTRODUCTION

Machine learning systems now mediate consequential decisions in clinical diagnostics, financial fraud detection, and routine commercial software. Advances in computing capacity and inexpensive storage have permitted an accompanying growth in algorithmic complexity. Academic evaluations of these systems have conventionally been conducted on large, curated corpora (Banko & Brill, 2001), and the resulting literature has often concluded that additional data reliably yield additional predictive capability (Halevy et al., 2009). Under conditions of data abundance, high-capacity models are able to recover fine-grained structure without incurring prohibitive generalisation penalties.

The assumption of unlimited data is rarely satisfied in applied settings. Data acquisition consumes time and financial resources, privacy regulation restricts the collection and retention of personal records (Zantvoort et al., 2024), and target events of interest may simply be rare. A clinical team studying an uncommon disease may assemble only a few dozen records, and a small commercial enterprise may possess only a few hundred transactions. When a high-capacity model is fitted to samples of this size, the estimated decision boundary becomes unstable and the resulting system fails on deployment.

Learning from limited samples therefore constitutes a distinct methodological problem. As the number of observations decreases, the probability of overfitting—the memorization of sample-specific noise rather than the recovery of generalizable structure—increases substantially. Under such conditions, the structural assumptions encoded within an algorithm exert greater influence on out-of-sample performance than the quantity of information supplied to it.

Despite its prevalence in practice, data scarcity remains comparatively neglected in mainstream machine learning research (Naser, 2026). A substantial portion of published work pursues incremental accuracy gains by scaling both model capacity and corpus size, which widens the distance between academic benchmarks and applied software engineering. Practitioners confronting a small dataset must nonetheless choose between a constrained model that may fail to capture complex structure but remains stable, and a flexible model that may capture such structure but may also fail unpredictably. In the absence of controlled comparison under explicit volume constraints, this choice remains largely intuitive. Because the UCI Machine Learning Repository contains numerous datasets with modest sample sizes (Kelly et al., n.d.), a controlled reduction procedure applied to such datasets permits direct

observation of the point at which particular architectures cease to generalize.

Existing empirical evidence indicates that simpler classifiers reach their performance ceilings at smaller sample sizes than complex alternatives (Fernández-Delgado et al., 2014), and that naive Bayes, logistic regression, and support vector machines remain comparatively stable when data are limited (Al-Bahadili et al., 2021). Higher-capacity architectures typically require substantially larger corpora before attaining comparable accuracy. Characterising these minimum data requirements is consequential for practice, since failures attributable to data starvation may otherwise be misattributed to the algorithm itself. Computational efficiency compounds this concern: fitting high-capacity models to small samples expends processing time and energy for limited benefit, and in instructional, small-enterprise, and edge-deployment contexts, data-efficient models constitute a practical necessity (Naser, 2026).

The present study addressed this gap by systematically evaluating five classification algorithms under controlled reductions of training-data volume. Rather than restricting evaluation to accuracy, the study assessed performance across ten criteria encompassing threshold-dependent exactness, threshold-independent separability, probability calibration, computational cost, and interpretability. The objective was to identify the conditions under which each architecture ceases to perform reliably and to translate these observations into evidence-based guidance for model selection.

Research Questions

The study was guided by the following general question: Which classification model achieves the most favourable trade-off among predictive stability, exactness, and computational efficiency as the volume of available training data varies? Four specific questions were derived from this general question:

1. When trained under small-dataset constraints, how do the five models perform across accuracy, precision, recall, F1 score, AUC-ROC, AUC-PR, mean absolute error, mean squared error, computational cost, and interpretability?
2. How do the same models perform on medium-sized datasets across the same ten criteria?
3. How do the same models perform on large datasets across the same ten criteria?
4. How do the observed trade-offs inform the suitability of these algorithms for implementation in applied software development projects?

Hypotheses

Null hypotheses

- **H01** — Decreasing the dataset size produces no significant difference in how the selected machine learning models degrade in predictive performance.
- **H02** — Dataset volume does not significantly affect the performance stability and variance of the classification models.
- **H03** — Complex models and simple models show no significant difference in data efficiency under constrained data conditions.

Alternative hypotheses

- **H11** — A significant difference exists in performance degradation among the selected models as the dataset size decreases.
- **H12** — The volume of available training data significantly alters the stability of the machine learning models.
- **H13** — Simple models demonstrate significantly greater data efficiency and stability than complex models under strict data scarcity.

LITERATURE REVIEW

The literature reviewed below is organised into five themes: the sensitivity of specific algorithmic structures to data scarcity, the multidimensional evaluation of classification performance, probability calibration and practical deployment constraints, the statistical comparison of competing models, and empirical evidence concerning performance under limited data. The section concludes with a synthesis identifying the gap addressed by the present study.

Algorithmic Structure and Sensitivity to Data Scarcity

The structural design of an algorithm governs its response to reductions in sample size. Logistic regression functions as the conventional baseline for classification tasks (Park, 2013). Because it fits a smooth and comparatively rigid boundary, it exhibits low variance and requires relatively few observations to stabilise its parameters. Its functional form precludes the highly irregular boundaries through which more flexible models accommodate outliers, and this restriction confers resistance to overfitting when data are limited (Lynam et al., 2020).

Support vector machines identify a maximally separating margin, projecting observations into higher-dimensional spaces where necessary (Cortes & Vapnik, 1995). Although effective on complex data, the resulting boundary depends on a small subset of observations located near the margin. When a reduced sample omits these critical support vectors, the estimated boundary may be substantially misplaced, producing abrupt reductions in out-of-sample accuracy.

Naive Bayes proceeds from a conditional independence assumption, treating each feature as contributing independently to the posterior probability (Domingos & Pazzani, 1997). Although this assumption is rarely satisfied, the resulting simplification confers considerable resilience under sparse sampling (McCallum & Nigam, 1998). Because estimation reduces to counting conditional frequencies rather than optimising a high-dimensional objective, the classifier is comparatively resistant to noise memorisation and provides a robust baseline where samples are insufficient for higher-capacity methods.

Decision trees partition the feature space through recursive logical splits (Quinlan, 1986). Unconstrained trees are well documented as prone to overfitting: given a small sample, recursive partitioning may continue until individual training observations are isolated, and the resulting structure generalises poorly (Breiman et al., 1984). Random forests were developed specifically to address the variance of single trees by aggregating predictions across many trees fitted to randomised subsamples (Breiman, 2001). This mechanism nonetheless presupposes sufficient data to generate genuinely diverse constituent trees; where samples are small, the ensemble may lack the variation required for effective aggregation, and its computational overhead yields diminishing returns.

Multidimensional Evaluation of Classification Performance

Reliance on any single performance metric produces an incomplete characterisation of model behaviour, as different metrics expose different failure modes (Sokolova & Lapalme, 2009). Accuracy, the ratio of correct predictions to total predictions, is the most widely reported criterion, yet it is unreliable under class imbalance: where a positive class occurs in 1% of cases, a classifier that predicts the majority class exclusively attains 99% accuracy while providing no diagnostic value.

Precision and recall address this limitation from complementary directions. Precision quantifies exactness by measuring the proportion of predicted positives that are correct and is therefore critical where false positives carry substantial cost. Recall quantifies completeness by measuring the proportion of actual positives that are detected and is critical in domains such as medical diagnostics, where undetected positive cases carry severe consequences. Because the two criteria are frequently in tension, the F1 score combines them through their harmonic mean (Powers, 2011), preventing a model from optimising one at the expense of the other and providing a more robust summary under skewed label distributions.

Threshold-independent criteria evaluate performance across the full range of decision thresholds. The area under the receiver operating characteristic curve measures class separability across all thresholds (Powers, 2011); however, it can be optimistic on small or imbalanced samples because correct classification of a large negative class contributes substantially to the statistic. The area under the precision–recall curve addresses this bias by excluding true negatives from consideration and concentrating on performance with respect to rare positive targets. Models that appear favourable under the receiver operating characteristic curve may therefore perform considerably less well under the precision–recall curve when data are scarce or imbalanced.

Probability Calibration and Practical Deployment Constraints

Calibration criteria evaluate the fidelity of predicted probabilities rather than the correctness of discrete labels (Van Smeden et al., 2024). The mean squared error of predicted class probabilities corresponds to the Brier score and penalises confident errors severely, whereas the mean absolute error reports the average magnitude of probability deviation. A correct prediction issued with low confidence and an incorrect prediction issued with high confidence are treated very differently under these criteria. High-capacity models fitted to small samples frequently become overconfident, producing extreme probability estimates for incorrect predictions and thereby degrading calibration even where discrete accuracy appears acceptable.

Non-statistical criteria carry comparable weight in deployment. Computational cost, operationalised as training time, constrains feasibility on limited hardware and increases the energy footprint of model development. Interpretability governs institutional trust: a linear model exposes a closed-form relationship between features and predictions, whereas an ensemble aggregates many internal decisions into an opaque prediction. In regulated domains such as medicine and finance, interpretability frequently constitutes a requirement rather than a preference.

Statistical Comparison of Machine Learning Models

Descriptive performance differences do not by themselves establish algorithmic superiority, since an apparent advantage of a few percentage points on a small dataset may reflect a favourable partition rather than a genuine structural difference. Hypothesis testing is therefore required to distinguish systematic effects from sampling variation. Analysis of variance provides the conventional procedure for comparing mean performance across several models simultaneously (Refaeilzadeh et al., 2009), assessing whether between-model variation exceeds within-model variation by more than chance would predict. Establishing this inferential baseline guards against the selection of computationally expensive architectures on the basis of incidental performance advantages.

Empirical Evidence Under Limited Data

A growing body of work examines the behaviour of learning systems as sample size decreases. Large neural architectures and heavily 11216arameterized boosting methods presuppose data abundance (Halevy et al., 2009) and generalise poorly when applied to small samples. Traditional methods, constrained by simpler functional forms, are comparatively well suited to such conditions. In clinical research, where privacy regulation frequently restricts samples to fewer than one thousand records, comparative work on drug intoxication mortality found that conventional classifiers performed at least as well as more complex alternatives (Choi & Boo, 2020). In ybersecurity, where labelled attack data are similarly scarce, comparisons between complex ensembles and simpler baselines reported improvements of up to six percentage points in favour of the simpler frameworks under extreme low-data conditions (Corea et al., 2024). Domain-specific learning-curve analyses have further indicated that predictive performance may not converge until sample sizes reach several hundred to several thousand observations (Zantvoort et al., 2024). The consistent implication

is that as data volume decreases, the structural bias of simpler models becomes an advantage rather than a limitation.

Synthesis and Research Gap

The reviewed literature establishes that algorithmic performance is conditioned on data availability, that high-capacity models require correspondingly large samples, and that traditional linear and probabilistic models resist overfitting because their functional forms cannot accommodate arbitrary noise. A methodological gap nonetheless persists. Most prior comparisons characterise degradation using accuracy alone and therefore do not document the simultaneous deterioration of exactness, completeness, threshold behaviour, and probability calibration. The present study addresses this gap by evaluating degradation across an integrated criterion suite that includes the F1 score, the imbalance-sensitive area under the precision–recall curve, and calibration error, and by combining repeated cross-validation with formal inferential testing to determine whether observed differences are attributable to algorithmic structure rather than sampling variation.

THEORETICAL FRAMEWORK

Statistical learning theory holds that the objective of a learning algorithm is accurate prediction on previously unseen observations rather than reproduction of the training sample (Cortes & Vapnik, 1995). Generalisation error quantifies the frequency of failure on new data. When samples are large, high-capacity models can approximate intricate structure without absorbing sampling noise (Banko & Brill, 2001); when samples are small, the same models increasingly fit noise rather than signal.

The bias–variance trade-off provides the central explanatory mechanism. Total prediction error decomposes into bias arising from restrictive modelling assumptions, variance arising from sensitivity to the particular training sample, and irreducible noise. Under limited data, the variance component dominates. High-capacity algorithms such as random forests exhibit low bias but high variance, so their performance fluctuates according to which few observations are sampled. Constrained models such as logistic regression and naive Bayes carry higher bias because they presuppose comparatively smooth and simple structure (Domingos & Pazzani, 1997); this bias functions as a stabilising constraint that suppresses variance and preserves predictive consistency when data are scarce.

Model capacity denotes the range of functional forms an algorithm can represent. Higher capacity requires proportionally more observations to identify the correct form; when sample size falls below the requirement implied by a model's capacity, overfitting becomes structurally unavoidable. Regularisation offers a partial remedy by penalising complexity and deliberately reducing variance, with the consequence that heavily regularised linear models frequently outperform unregularised high-capacity alternatives in low-sample regimes. The principle of parsimony expressed in Occam's razor extends this reasoning to model selection: when two models achieve comparable accuracy, the simpler model is preferable because it generalises through structural assumptions rather than through memorisation of individual records (Fernández-Delgado et al., 2014).

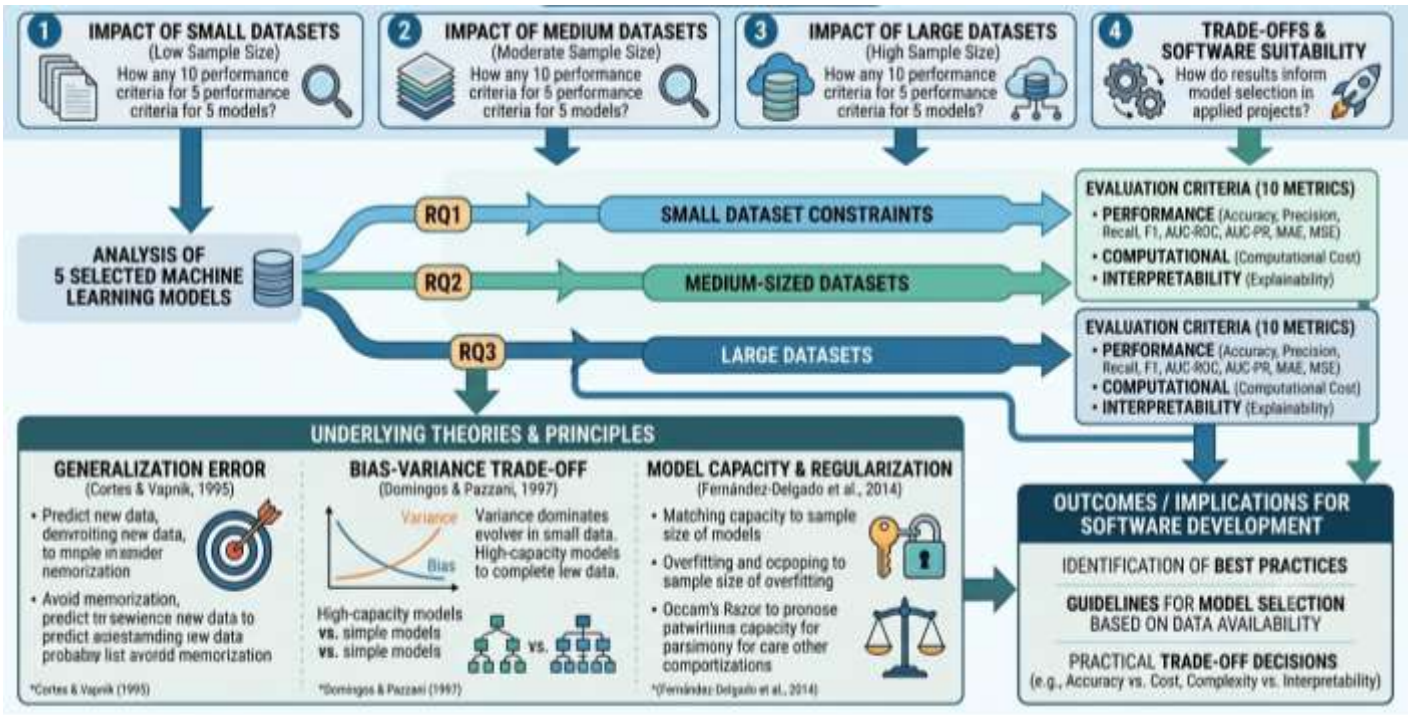

Figure 1. Conceptual Frameworks

METHOD

Research Design

The study employed a quantitative comparative experimental design organised according to an input–process–output framework (Feldman, 2025). The manipulated variable was the volume of training data supplied to each classifier; the measured variables were the ten evaluation criteria. Rather than fitting models to complete datasets only, the design imposed controlled data starvation by evaluating each algorithm at four training-volume levels—100%, 75%, 50%, and 25%—thereby permitting the reconstruction of empirical learning curves. Hardware, random seeds, preprocessing procedures, and evaluation protocols were held constant across all runs so that observed variation could be attributed to training volume and to algorithmic structure rather than to extraneous factors.

Within this framework, the inputs comprised the selected datasets, the four volume constraints, the five algorithms, and the cross-validation parameter. The process comprised preprocessing, partitioning, model fitting, cross-validation, extraction of predicted probabilities, timing of execution, and statistical analysis. The outputs comprised the mean and standard deviation of each evaluation criterion, together with the derived trade-off and degradation summaries reported below.

Datasets

Eight tabular datasets were obtained from the UCI Machine Learning Repository (Kelly et al., n.d.) and assigned to three volume tiers according to instance count. The selection spans heterogeneous application domains, including acoustic sensing, clinical diagnostics, financial security, and astrophysics, in order to reduce the influence of any single domain on the aggregate results. Small datasets were used to examine extreme scarcity and unfavourable ratios of features to observations; medium datasets provided intermediate conditions in which gradual degradation could be observed; and large datasets established empirical performance ceilings under minimal sampling variance. Table 1 summarises the characteristics of each dataset.

Table 1: Characteristics of the Eight Benchmark Datasets Drawn From the UCI Machine Learning Repository

Volume tier	Dataset	Instances (*N*)	Features	Application domain
Small	Breast Cancer Wisconsin (Diagnostic)	569	30	Healthcare diagnostics
Small	Wine Recognition	178	13	Chemical analysis
Small	Sonar (Mines vs. Rocks)	208	60	Acoustic sensing
Medium	Banknote Authentication	1,372	4	Financial security
Medium	Car Evaluation	1,728	6	Consumer analytics
Medium	Spambase	4,601	57	Text classification and cybersecurity
Large	MAGIC Gamma Telescope	19,020	10	Astrophysics
Large	Adult Census Income	48,842	14	Demographic economics

Note. *Instance and feature counts are those reported by the repository. Volume tiers were assigned on the basis of instance count and define the three analytical strata used throughout the Results section. Dataset sources: Aeberhard and Forina (1992); Becker and Kohavi (1996); Bock (2004); Bohanec (1988); Gorman and Sejnowski (1988); Hopkins et al. (1999); Lohweg (2012); Wolberg et al. (1993).*

Stratified Fractional Reduction Procedure

Model evaluation was conducted within a stratified five-fold cross-validation framework (Kohavi, 1995). Simple random partitioning is unsuitable under scarcity because a reduced training partition may exclude a target class entirely; stratification preserves the class distribution of the source dataset across all folds and therefore maintains the interpretability of precision and recall.

Within each cross-validation iteration, the training folds were reduced by stratified random sampling to the specified proportion of their original size, while the held-out validation fold was retained at full volume. Each model was therefore fitted to a progressively smaller training sample but evaluated against an identical and complete test partition, ensuring that

comparisons across volume levels reflect differences in learning conditions rather than differences in evaluation conditions.

Classifiers

Five classifiers implemented in the scikit-learn library (Pedregosa et al., 2011) were evaluated. Logistic regression served as the low-variance linear baseline (Park, 2013). The support vector classifier constructed boundaries using a radial basis function kernel (Cortes & Vapnik, 1995). Gaussian naive Bayes imposed conditional independence among features and functioned as the probabilistic reference model (Domingos & Pazzani, 1997). The decision tree was fitted without depth constraints and therefore represented an unregularised, high-variance structure (Quinlan, 1986). The random forest represented the ensemble condition, aggregating randomised trees through majority voting (Breiman, 2001).

All models were configured with the library's default hyperparameters. This decision was deliberate: because the study

sought to isolate the effect of training volume, hyperparameter optimisation was excluded to prevent tuning variance from confounding the manipulated variable. The design consequently estimates the relative trajectory of degradation under default configurations rather than the maximum attainable performance of each architecture.

Preprocessing and Control of Data Leakage

All transformations were executed within a scikit-learn pipeline combined with a column transformer (Pedregosa et al., 2011). Data leakage occurs when information derived from the evaluation partition influences model fitting, producing optimistically biased performance estimates. The pipeline architecture prevents this condition by estimating scaling and encoding parameters exclusively from the training partition within each cross-validation iteration and applying them to the held-out partition only at the point of prediction.

Categorical features were converted to binary indicator variables through one-hot encoding. Because reduction to 25% of the original training volume may eliminate rare categories from a training partition, the encoder was configured to ignore categories not observed during fitting, preventing runtime failures at prediction time. A global random seed of 42 was applied throughout so that variation in results is attributable to the reduction procedure rather than to stochastic initialisation. Numerical computation, data handling, and statistical testing were supported by the standard Python scientific stack (Harris et al., 2020; McKinney, 2010; Virtanen et al., 2020).

Evaluation Criteria

Ten criteria were recorded for every combination of dataset, model, and training volume. Nine were computed numerically from model predictions and predicted probabilities; interpretability was assessed qualitatively as a structural property of each architecture. Table 2 presents the operational definition of each criterion.

Table 2: Operational Definitions of the Ten Evaluation Criteria

Criterion	Operational definition
Accuracy	Ratio of correct predictions to total predictions; interpretable but potentially misleading under class imbalance.
Precision	Proportion of predicted positives that are correct; indexes exactness and the control of false positives.
Recall	Proportion of actual positives correctly identified; indexes completeness and the control of false negatives.
F1 score	Harmonic mean of precision and recall; prevents optimisation of one criterion at the expense of the other.
AUC-ROC	Area under the receiver operating characteristic curve; separability of classes across all decision thresholds.
AUC-PR	Area under the precision–recall curve; disregards true negatives and is therefore diagnostic for imbalanced targets.
MAE	Mean absolute deviation between predicted class probabilities and observed outcomes; a measure of probability calibration.
MSE	Mean squared deviation between predicted probabilities and observed outcomes (Brier score); penalises confident errors severely.
Computational cost	Wall-clock training time in seconds, recorded on identical hardware across all runs.

Criterion	Operational definition
Interpretability	Qualitative classification of structural transparency, distinguishing white-box models with directly inspectable parameters from black-box ensemble structures.

Note. *AUC-ROC = area under the receiver operating characteristic curve; AUC-PR = area under the precision–recall curve; MAE = mean absolute error; MSE = mean squared error. MAE and MSE were computed on predicted class probabilities rather than on discrete predictions and therefore index calibration rather than classification correctness. Definitions follow Powers (2011), Sokolova and Lapalme (2009), and Van Smeden et al. (2024).*

Data Collection Procedure

Data collection was fully automated. For each dataset, the analysis script loaded the raw data, initialised the five classifiers, and entered the stratified five-fold cross-validation loop. Within each fold, the stratified reduction function removed training observations until the target proportion was reached; each model was then fitted exclusively to the surviving subset and evaluated against the intact validation fold. The script computed all ten criteria, recorded training time from the system clock, and exported aggregated results to structured comma-separated value files. Mean scores and standard deviations were computed across folds, the latter serving as the operational index of model stability.

Statistical Analysis

Two inferential procedures were applied at an alpha level of .05. First, a one-way analysis of variance was computed for each dataset-by-volume configuration, comparing mean accuracy across the five models by evaluating between-model variance relative to within-model variance. Because this procedure indicates only that at least one model differs from the others, a Tukey honestly significant difference test was subsequently applied to the results aggregated across all datasets at the 25% training volume in order to identify which specific pairs of models differed. Degradation was quantified descriptively as the change in each criterion between the 100% and 25% conditions, averaged across datasets.

Scope and Delimitations

The study was restricted to supervised binary and multiclass classification on structured tabular data. Deep learning architectures were excluded because their data requirements fall outside the range examined here, and hyperparameter optimisation was excluded for the reasons stated above. The findings therefore describe the comparative behaviour of five conventional classifiers under default configurations and controlled volume reduction; they should not be interpreted as estimates of the maximum performance attainable by any of these architectures after tuning.

RESULTS

Results are reported in four parts. Performance is first described within each volume tier, then summarised across all datasets in an aggregated degradation and cost matrix, then examined criterion by criterion, and finally evaluated inferentially through analysis of variance and post hoc comparison. Throughout, the trade-off figures plot mean accuracy against the standard deviation of accuracy across folds; the upper-left region of each plot represents the preferred combination of high accuracy and low instability.

Performance Under Extreme Scarcity: Small Datasets

Small datasets required models to estimate decision boundaries from severely limited observations. As shown in Figure 1, the ensemble condition shifted rightward as training volume decreased, indicating increased variance and heightened sensitivity to the composition of individual training folds. Logistic regression remained positioned in the upper-left region across all four volume levels, indicating that its decision boundary remained stable despite the reduction in available observations.

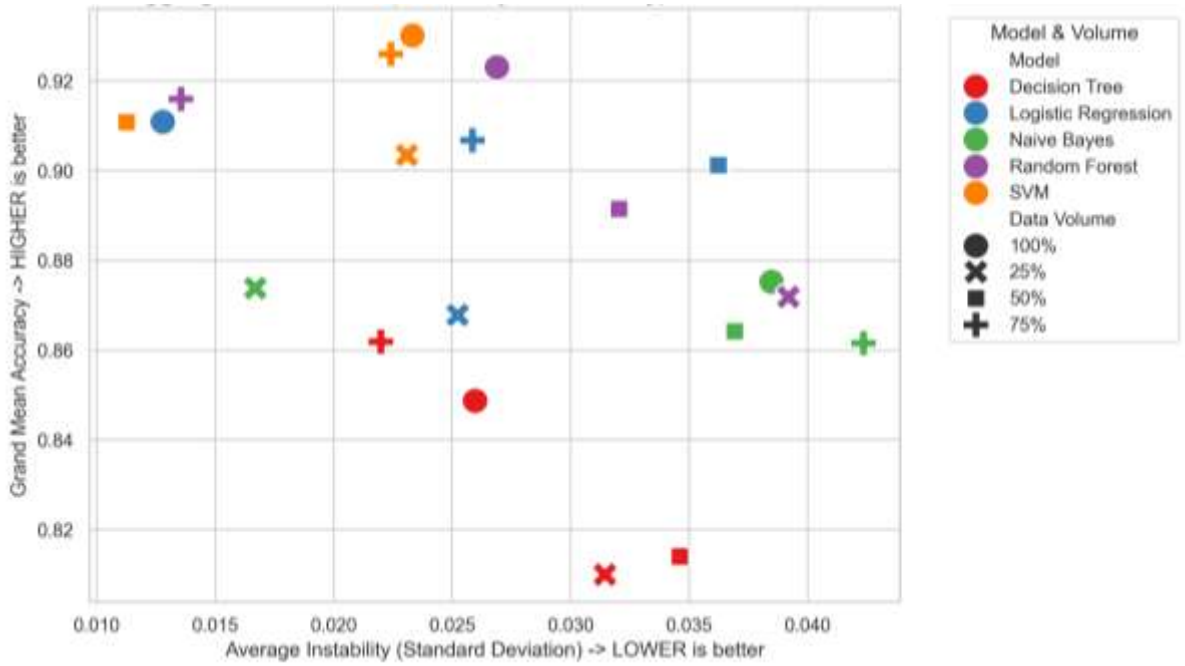

Figure 2. Aggregated Trade-Off Between Accuracy and Stability on Small Datasets

Figure 2 disaggregates performance at the 25% volume condition across the numerical criteria. Although the higher-capacity models retained competitive raw accuracy, their precision and F1 scores fluctuated more markedly than those of the constrained models. Naive Bayes and the support vector classifier maintained comparatively strong recall, identifying positive cases without requiring densely populated regions of the feature space to confirm the underlying pattern.

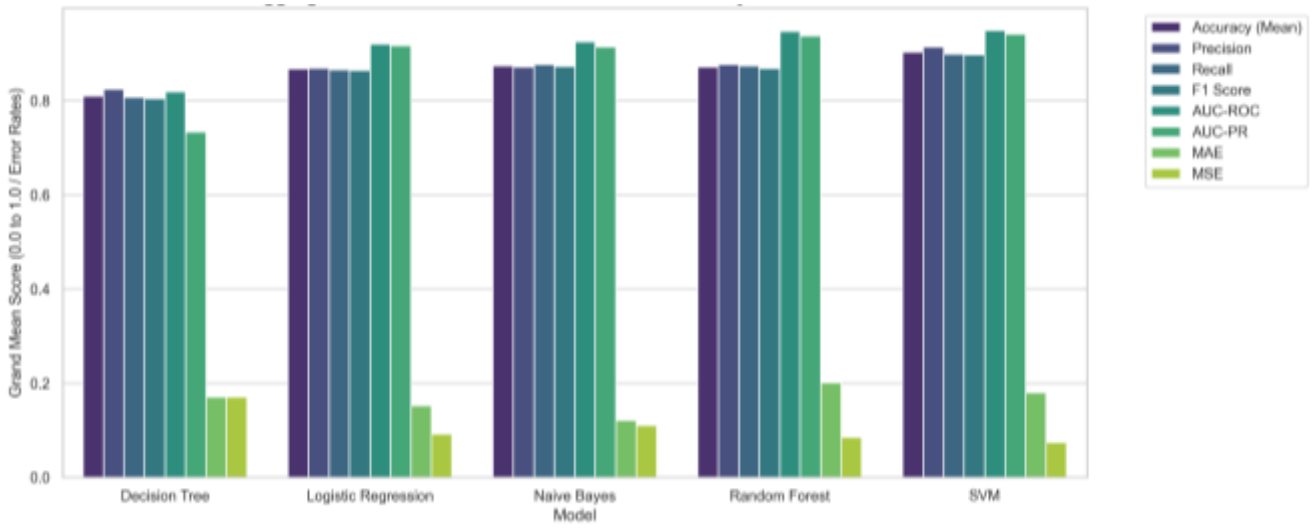

Figure 3. Aggregated Model Performance at 25% Training Volume on Small Datasets

Performance Under Moderate Constraint: Medium Datasets

Medium datasets supplied sufficient observations for the approximation of more complex boundaries while continuing to constrain high-capacity architectures. As shown in Figure 3, the support vector classifier occupied the upper-left region across volume levels, indicating that moderate sample sizes provided a sufficient number of support vectors to define stable margins without inducing the variance observed in the unregularised tree.

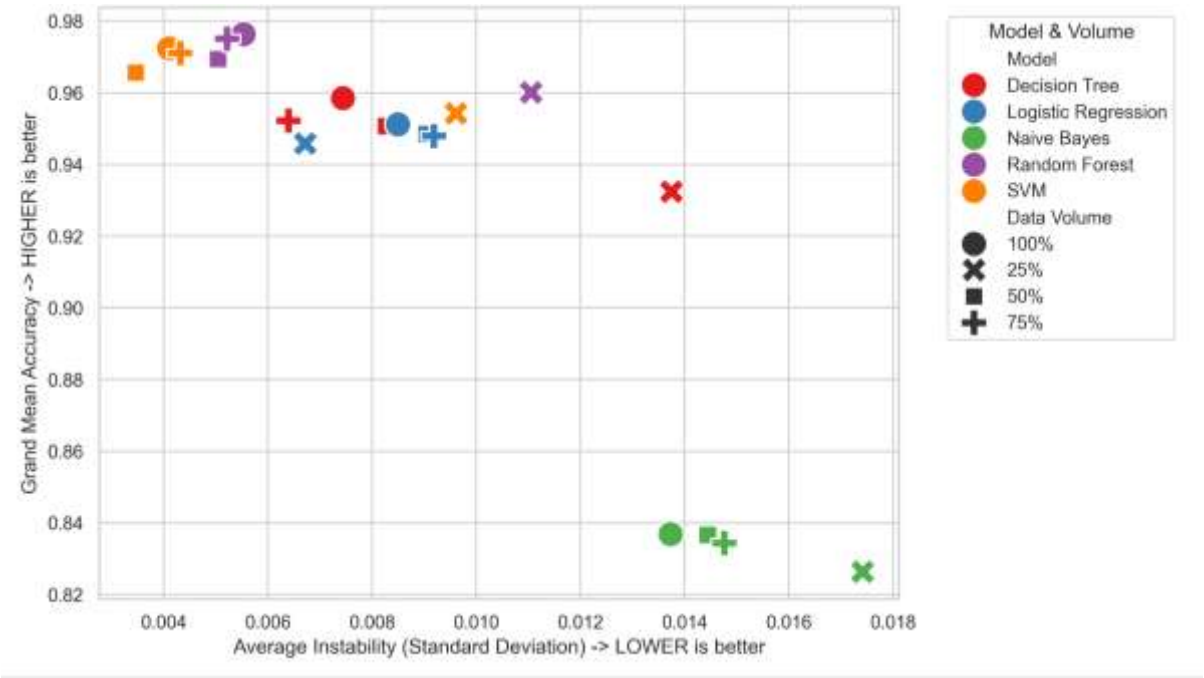

Figure 4: Aggregated Trade-Off Between Accuracy and Stability on Medium Datasets

At the 25% volume condition, the exactness criteria clustered more closely than on the small datasets, with the support vector classifier achieving the most favourable overall balance (see Figure 4). The decision tree continued to exhibit variance-related reductions in precision and recall, consistent with persistent difficulty in generalising beyond the training partition.

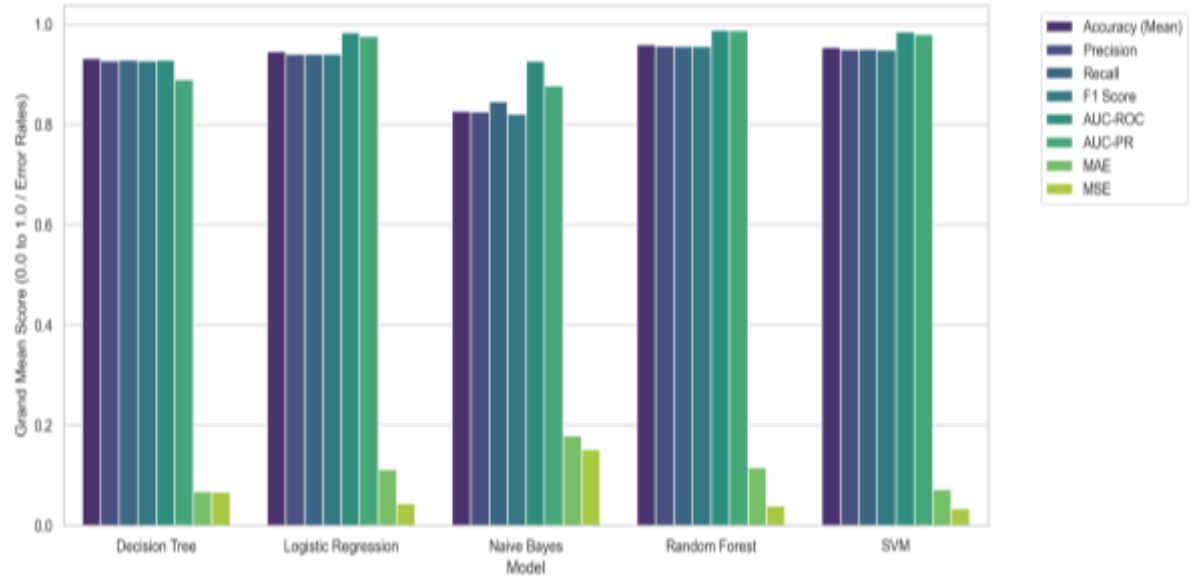

Figure 5. Aggregated Model Performance at 25% Training Volume on Medium Datasets

Performance Under Data Abundance: Large Datasets

Large datasets established the empirical performance ceilings of the five architectures under minimal sampling variance. The trade-off pattern changed substantially under these conditions (see Figure 5). The random forest moved into the upper-left region, indicating that access to abundant observations permitted the construction of genuinely diverse constituent trees and thereby suppressed the variance that had limited its performance on smaller datasets.

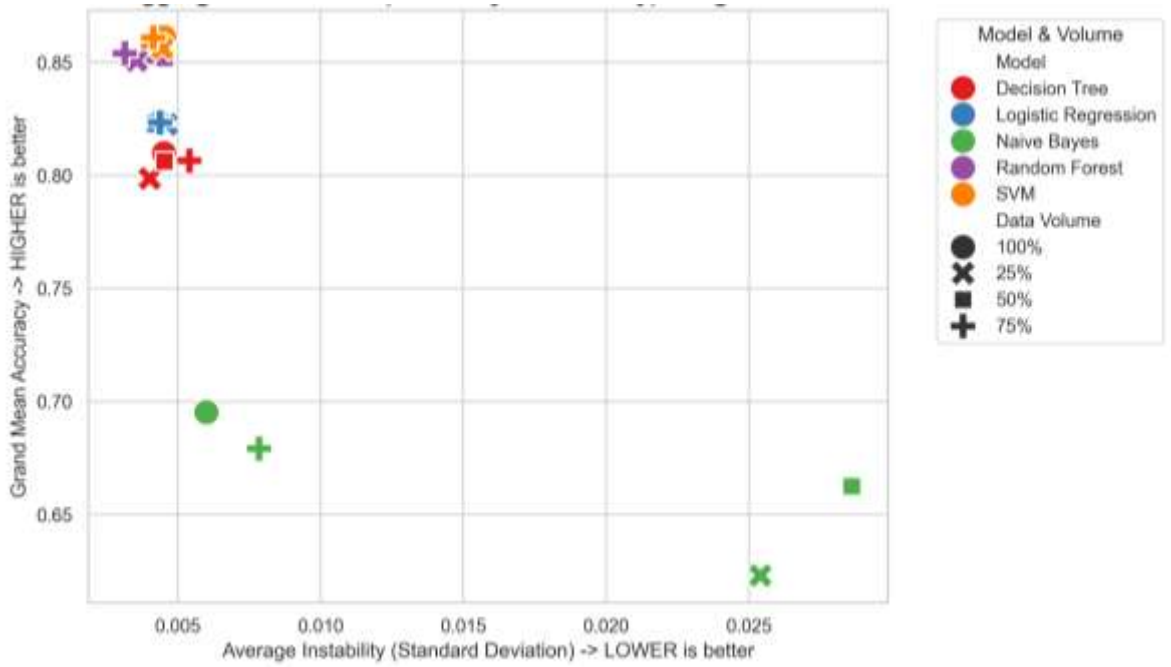

Figure 6. Aggregated Trade-Off Between Accuracy and Stability on Large Datasets

The criterion-level results for large datasets, presented in Figure 6, are consistent with conventional expectations: given abundant data, the random forest achieved the highest accuracy, precision, and area under the receiver operating characteristic curve among the models evaluated.

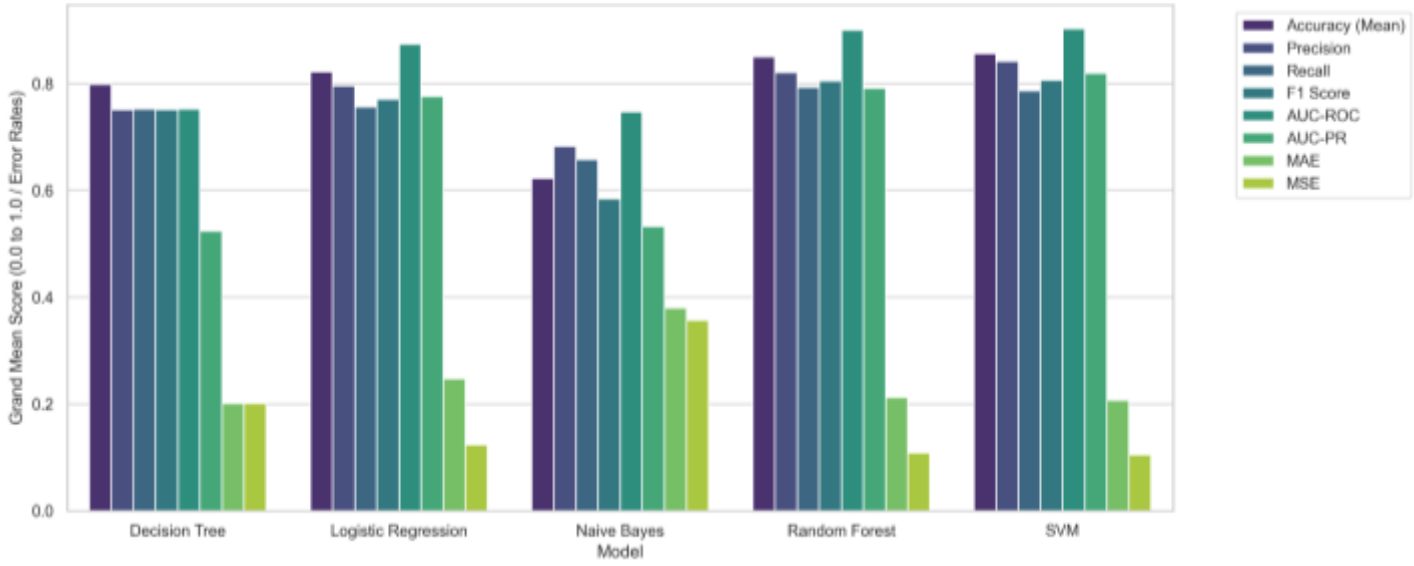

Figure 7. Aggregated Model Performance at 25% Training Volume on Large Datasets

Aggregated Degradation and Computational Cost

Table 3 reports the change in accuracy, area under the precision–recall curve, and calibration error observed as training volume was reduced from 100% to 25%, averaged across all eight datasets, together with instability at the 25% condition and mean training time at full volume.

Table 3: **Aggregated Performance Degradation and Computational Cost From 100% to 25% Training Volume**

Model	Accuracy change	AUC-PR change	MSE change	Instability at 25% (*SD*)	Training time at 100% (s)
Logistic regression	−0.0184	−0.0104	+0.0124	0.0131	0.098
Support vector machine	−0.0180	−0.0133	+0.0157	0.0133	105.479
Naive Bayes	−0.0225	−0.0405	+0.0217	0.0191	0.045
Decision tree	−0.0271	−0.0348	+0.0240	0.0179	0.142
Random forest	−0.0263	−0.0161	+0.0160	0.0197	1.247

Note. Values are averaged across all eight datasets. Negative values denote reductions and positive values denote increases relative to the 100% condition. AUC-PR = area under the precision–recall curve; MSE = mean squared error of predicted probabilities. Instability is the mean standard deviation of accuracy across the five cross-validation folds at the 25% condition, for which lower values indicate greater stability.

Criterion-Level Patterns

Predictive Correctness

High-capacity models carry limited structural bias and consequently require substantial data to validate the complex boundaries they generate. As shown in Table 3, the decision tree and random forest sustained the largest reductions in accuracy under scarcity (−0.0271 and −0.0263, respectively) and recorded the highest instability at the 25% condition (*SD*s = 0.0179 and 0.0197). Without sufficient observations to constrain them, 11224nregularized tree structures partitioned on sample-specific noise, with corresponding effects on precision and F1 score. Logistic regression sustained the smallest accuracy reduction (−0.0184) while maintaining balance between precision and recall.

Threshold Robustness

Whereas accuracy summarises discrete correctness, threshold-independent criteria characterise performance across the full range of decision thresholds. The area under the receiver operating characteristic curve remained comparatively stable for all models. The area under the precision–recall curve, which excludes true negatives and is therefore more sensitive to performance on rare positive targets, revealed a different pattern: naive Bayes and the decision tree sustained the largest reductions (−0.0405 and −0.0348), whereas logistic regression sustained a reduction of only −0.0104. The smooth probability function of logistic regression therefore preserved the ordering of imbalanced targets more effectively under scarcity than did recursive logical partitioning.

Probability Calibration

Calibration criteria evaluate whether a model's expressed confidence corresponds to its empirical accuracy. Under the 25% condition, the 11224arameterizat decision tree recorded the largest increase in mean squared error (+0.0240), indicating that its incorrect predictions were issued with inappropriately high confidence. Logistic regression recorded the smallest increase (+0.0124), indicating that its probability estimates remained approximately calibrated even where its discrete predictions were incorrect.

Computational Cost and Interpretability

Performance rankings based on statistical criteria alone omit the physical cost of computation. The support vector classifier matched the stability of the constrained models but required a mean of 105.479 s to fit at full volume, whereas logistic regression achieved comparable results in 0.098 s—a difference of approximately three orders of magnitude. The random forest required 1.247 s and, as an ensemble of many trees, does not expose an inspectable decision rule. Logistic regression, by contrast, yields a closed-form 11224arameterization that can be examined directly. Under conditions in which the accuracy differential between architectures is small, this disparity in cost and transparency is consequential for applied deployment.

Inferential Results

Analysis of Variance

A one-way analysis of variance was computed for each of the 32 dataset-by-volume configurations, comparing mean accuracy across the five models. Between-model differences were statistically significant in 31 of the 32 configurations. The single exception occurred for the Breast Cancer Wisconsin dataset at the 50% volume condition, $F = 0.90$, $p = .483$. Full results are reported in Table 4.

Table 4: One-Way Analysis of Variance Comparing Model Accuracy Within Each Dataset and Training-Volume Configuration

Dataset	Training volume (%)	*F*	*p*	Significant at α = .05
Breast Cancer (small)	100	6.10	.002	Yes
Breast Cancer (small)	75	3.75	.020	Yes
Breast Cancer (small)	50	0.90	.483	No
Breast Cancer (small)	25	4.96	.006	Yes
Wine Recognition (small)	100	9.79	< .001	Yes
Wine Recognition (small)	75	7.48	< .001	Yes
Wine Recognition (small)	50	14.36	1.11×10^{-5}	Yes
Wine Recognition (small)	25	20.34	8.08×10^{-7}	Yes
Sonar (small)	100	7.70	< .001	Yes
Sonar (small)	75	6.25	.002	Yes
Sonar (small)	50	4.23	.012	Yes
Sonar (small)	25	2.87	.050	Yes
Banknote Authentication (medium)	100	204.14	6.56×10^{-16}	Yes
Banknote Authentication (medium)	75	174.47	3.02×10^{-15}	Yes
Banknote Authentication (medium)	50	166.64	4.71×10^{-15}	Yes
Banknote Authentication (medium)	25	109.21	2.73×10^{-13}	Yes
Car Evaluation (medium)	100	118.76	1.23×10^{-13}	Yes
Car Evaluation (medium)	75	97.79	7.80×10^{-13}	Yes
Car Evaluation (medium)	50	126.41	6.75×10^{-14}	Yes
Car Evaluation (medium)	25	34.33	1.07×10^{-8}	Yes
Spambase (medium)	100	257.11	6.90×10^{-17}	Yes
Spambase (medium)	75	298.60	1.59×10^{-17}	Yes
Spambase (medium)	50	174.93	2.94×10^{-15}	Yes
Spambase (medium)	25	108.34	2.95×10^{-13}	Yes
MAGIC Gamma Telescope (large)	100	579.71	2.28×10^{-20}	Yes
MAGIC Gamma Telescope (large)	75	567.32	2.83×10^{-20}	Yes
MAGIC Gamma Telescope (large)	50	529.36	5.61×10^{-20}	Yes
MAGIC Gamma Telescope (large)	25	617.96	1.21×10^{-20}	Yes
Adult Census Income (large)	100	928.29	2.13×10^{-22}	Yes
Adult Census Income (large)	75	752.76	1.71×10^{-21}	Yes
Adult Census Income (large)	50	76.99	7.39×10^{-12}	Yes
Adult Census Income (large)	25	170.93	3.68×10^{-15}	Yes

Note. Each test compared mean accuracy across the five classifiers within a single dataset-by-volume configuration. Probability values below .001 are reported in scientific notation to preserve the precision of the original computation. α = .05.

Post Hoc Comparison

Because the analysis of variance evaluated model differences within individual configurations, an additional test was required to assess whether any model demonstrated a 11226eneralized advantage across datasets. A Tukey honestly significant difference test was therefore computed on the results aggregated across all eight datasets at the 25% training volume. None of the ten pairwise comparisons reached significance, and every confidence interval included zero (see Table 5 and Figure 7).

Table 5 :Tukey Honestly Significant Difference Pairwise Comparisons of Model Accuracy at 25% Training Volume

Model 1	Model 2	Mean difference	*p*adj	95% CI lower	95% CI upper
Decision tree	Logistic regression	0.0327	.9682	−0.1152	0.1806
Decision tree	Naive Bayes	−0.0597	.7729	−0.2077	0.0882
Decision tree	Random forest	0.0466	.8927	−0.1013	0.1945
Decision tree	Support vector machine	0.0576	.7948	−0.0903	0.2056
Logistic regression	Naive Bayes	−0.0924	.3915	−0.2404	0.0555
Logistic regression	Random forest	0.0139	.9988	−0.1340	0.1618
Logistic regression	Support vector machine	0.0249	.9883	−0.1230	0.1729
Naive Bayes	Random forest	0.1064	.2569	−0.0416	0.2543
Naive Bayes	Support vector machine	0.1174	.1751	−0.0305	0.2653
Random forest	Support vector machine	0.0110	.9995	−0.1369	0.1590

Note. Mean differences are expressed in units of accuracy and computed as Model 1 minus Model 2, aggregated across all eight datasets. padj = probability value adjusted for multiple comparisons; CI = confidence interval. No comparison reached significance at α = .05.

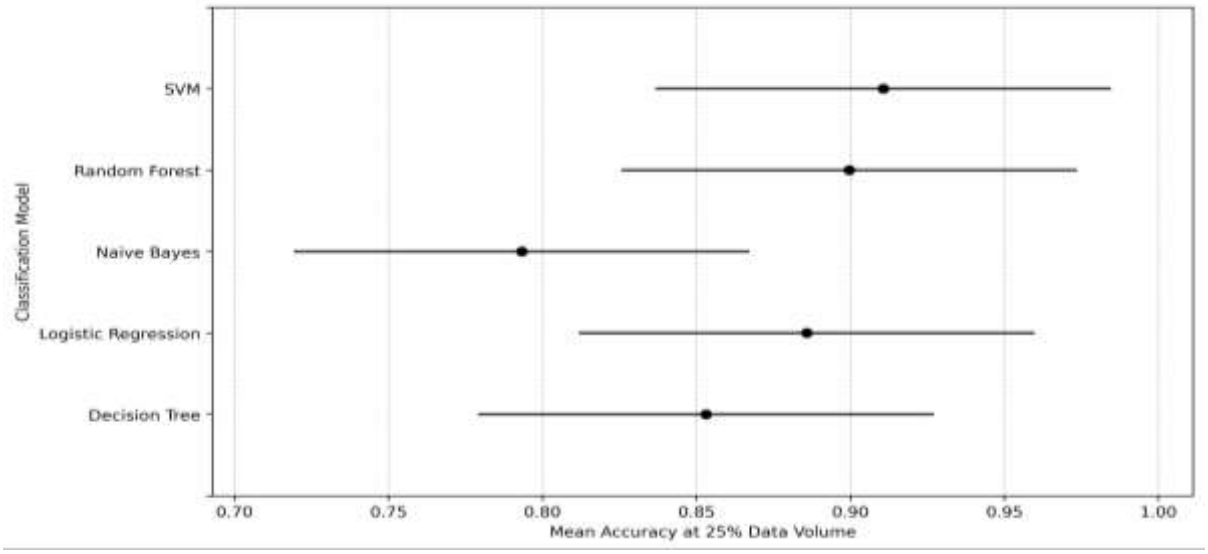

Figure 8: Confidence Intervals for Mean Model Accuracy at 25% Training Volume

The overlap among intervals reflects a substantive property of the aggregated data. Absolute accuracy is strongly conditioned by the intrinsic difficulty of each dataset, so that a model may attain markedly different scores across domains. This between-dataset variance is large relative to between-model variance and consequently obscures aggregate architectural comparison.

Hypothesis Testing Outcomes

The first null hypothesis (H01), which asserted that dataset size produces no significant difference in predictive performance,

was rejected. Accuracy declined across every model as training volume was reduced (see Table 3), and model performance
The second null hypothesis (H02), which asserted that dataset volume does not affect stability, was likewise rejected. Instability at the 25% condition varied systematically with model capacity, ranging from *SD* = 0.0131 for logistic regression to *SD* = 0.0197 for the random forest, and the trade-off plots document a consistent rightward displacement of high-capacity models as volume decreased.
The third null hypothesis (H03), which asserted no difference in data efficiency between simple and complex models, was rejected with respect to degradation magnitude, stability, and computational cost. The constrained models sustained smaller reductions in accuracy and in area under the precision–recall curve, smaller increases in calibration error, lower instability, and substantially lower training cost than the tree-based ensemble. It should be noted, however, that this conclusion rests on the pattern of degradation rather than on aggregate pairwise accuracy comparison, since the Tukey test detected no significant differences in mean accuracy between any pair of models at the 25% condition.

DISCUSSION

This study examined how five conventional classification algorithms degrade as training-data volume is systematically reduced, evaluating performance across ten criteria rather than accuracy alone. The results clarify the distance between benchmark evaluation conducted under data abundance and the conditions encountered in applied software development.

Degradation Patterns and the Bias–Variance Trade-Off

The observed degradation pattern is consistent with the bias–variance decomposition outlined in the theoretical framework. Algorithms with low structural bias require large samples to justify their representational capacity; when that requirement is not met, the variance component of prediction error dominates. The decision tree and random forest—the two architectures with the greatest capacity among those evaluated—sustained the largest accuracy reductions and the highest instability, whereas logistic regression, whose functional form constrains the admissible boundaries, sustained the smallest reduction. These findings align with prior evidence that simpler classifiers reach their performance ceilings at smaller sample sizes (Fernández-Delgado et al., 2014) and that conventional methods remain competitive with more complex alternatives in low-data clinical and security applications (Choi & Boo, 2020; Corea et al., 2024).
The results for the large datasets indicate that this advantage is conditional rather than absolute. When abundant data were available, the random forest achieved the highest scores across the principal predictive criteria, consistent with the mechanism by which ensemble aggregation suppresses variance when constituent models are genuinely diverse (Breiman, 2001). Model capacity is therefore best understood as a requirement to differed significantly within 31 of 32 configurations (see Table 4).
be matched against available data rather than as a property that is inherently advantageous or disadvantageous.

The Value of Multidimensional Evaluation

Several findings would have been invisible under accuracy-only evaluation. The area under the receiver operating characteristic curve remained comparatively stable across models even as the area under the precision–recall curve declined markedly for naïve Bayes and the decision tree, corroborating the argument that the former criterion can be optimistic where negative cases are numerous (Powers, 2011; Sokolova & Lapalme, 2009). The calibration results are similarly informative: the increase in mean squared error observed for the unregularised tree indicates that scarcity affected not only whether predictions were correct but also whether the confidence attached to them was warranted. For applications in which predicted probabilities inform downstream decisions, such as clinical triage or risk scoring, this deterioration is consequential independently of discrete accuracy (Van Smeden et al., 2024).

The Absence of Universal Superiority

The post hoc analysis provides an important qualification to the preceding conclusions. Although degradation was systematic and between-model differences were significant within individual configurations, no pairwise comparison of aggregate accuracy at the 25% condition reached significance. Absolute performance is therefore governed substantially by dataset structure and domain difficulty, and between-dataset variance is sufficient to obscure architectural differences when results are pooled.
This outcome reinforces rather than undermines the principle of parsimony. Where a constrained linear model and a high-capacity ensemble produce statistically indistinguishable accuracy, the selection of the more complex architecture entails additional variance, degraded calibration, greater computational expenditure, and reduced transparency without a demonstrable predictive return. Under such conditions the simpler model is the better-justified choice.

Implications for Practice

Three implications follow for applied software development. First, model selection should be conditioned on data availability. In domains where labelled data are structurally limited—rare-disease diagnostics, niche fraud detection, and rare-event prediction—constrained models such as logistic regression provide stability, calibrated probability estimates, and interpretable parameters. Second, computational cost warrants explicit consideration in selection decisions. The support vector classifier required approximately three orders of magnitude more training time than logistic regression for comparable performance, an allocation that is difficult to justify where retraining is frequent or hardware is constrained. Third, simpler models should be benchmarked first and treated as the reference

against which additional complexity must demonstrate measurable benefit.

LIMITATIONS

Several limitations qualify the interpretation of these findings. All models were evaluated under default hyperparameter configurations; while this choice isolates the effect of training volume, it does not establish the performance attainable after tuning, and tuned high-capacity models might degrade differently. The analyses of variance compared models within each dataset-by-volume configuration rather than incorporating volume as an explicit factor, so degradation across volume levels was quantified descriptively rather than tested inferentially. Effect size estimates were not obtained, and their absence limits assessment of the practical magnitude of the reported differences. The study was also confined to structured tabular data and to eight datasets, only two of which occupied the large tier; the aggregate results for that tier therefore rest on a narrower base than those for the small and medium tiers. Finally, reduction proceeded to 25% of the original training volume, so behaviour under more extreme scarcity remains uncharacterized.

Directions for Future Research

Three extensions follow directly from these limitations. First, the volume-reduction protocol should be replicated on severely imbalanced datasets in order to determine whether scarcity disproportionately degrades minority-class recall in high-capacity tree models. Second, the algorithmic set should be expanded to include gradient boosting frameworks and shallow neural architectures, permitting their breaking points to be located relative to the conventional baselines examined here. Third, dimensionality reduction should be applied prior to volume reduction to test whether contraction of the feature space stabilises high-variance models under scarcity. Incorporating training volume as an explicit factor in the inferential model and reporting effect sizes would further strengthen the evidential basis of subsequent work.

CONCLUSION

This study evaluated the degradation of five conventional classification algorithms under systematic reduction of training-data volume across eight tabular datasets and ten evaluation criteria. Three conclusions follow from the results.

First, data volume conditions model viability. The assumption that greater architectural complexity yields better outcomes is not supported under scarcity: fitting high-capacity models to limited samples produced the largest accuracy reductions, the highest instability, and the greatest deterioration in probability calibration. Model capacity must be matched to the quantity of data available to support it.

Second, constrained models offer measurable advantages when data are scarce. Logistic regression sustained the smallest reductions in accuracy and in area under the precision–recall curve, the smallest increase in calibration error, and the lowest instability, while remaining directly interpretable. This pattern provides empirical support for the principle of parsimony in model selection.

Third, computational cost constitutes a substantive selection criterion. The marginal accuracy advantages achievable through ensemble and kernel methods were obtained at substantially greater computational expense, a trade-off that is difficult to justify in contexts prioritizing rapid retraining, low-power deployment, or energy efficiency.

The post hoc analysis nonetheless establishes an important boundary condition: no model demonstrated universal superiority across datasets, and absolute performance was governed substantially by dataset structure. The contribution of this study is therefore not the identification of a single optimal algorithm but the provision of an empirical account of how five widely used architectures degrade as data diminish, evaluated across correctness, threshold robustness, calibration, cost, and transparency. This account supplies applied practitioners with a defensible basis for matching algorithmic choice to data and hardware constraints, and it establishes a reproducible protocol against which additional architectures may be assessed.

REFERENCES

1. Aeberhard, S., & Forina, M. (1992). *Wine* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J
2. Al-Bahadili, H., Abu-Dalo, A., & Khreisat, L. (2021). Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences, 11*(2), 796. https://doi.org/10.3390/app11020796
3. Banko, M., & Brill, E. (2001). Scaling to very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 26–33).
4. Association for Computational Linguistics. https://aclanthology.org/P01-1005/
5. Becker, B., & Kohavi, R. (1996). *Adult* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20
6. Bock, R. (2004). *MAGIC gamma telescope* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C52C8B
7. Bohanec, M. (1988). *Car evaluation* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48
8. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

9. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.
10. Choi, Y., & Boo, Y. (2020). Comparing logistic regression models with alternative machine learning methods to predict the risk of drug intoxication mortality. *International Journal of Environmental Research and Public Health, 17*(3), 897. https://doi.org/10.3390/ijerph17030897
11. Corea, P. M., Liu, Y., Wang, J., Niu, S., & Song, H. (2024). *Explainable AI for comparative analysis of intrusion detection models*. arXiv. https://arxiv.org/abs/2406.09684
12. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018
13. Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*(2–3), 103–130. https://doi.org/10.1023/A:1007413511361
14. Feldman, K. (2025, January 23). A comprehensive guide to input-process-output models. *iSixSigma*. https://www.isixsigma.com/dictionary/input-process-output-i-p-o/
15. Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*(1), 3133–3181. https://jmlr.org/papers/v15/delgado14a.html
16. Gorman, R. P., & Sejnowski, T. J. (1988). *Connectionist bench (sonar, mines vs. rocks)* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5T01Q
17. Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36
18. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
19. Hopkins, M., Reeber, E., Forman, G., & Suermondt, J. (1999). *Spambase* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C53G6X
20. Kelly, M., Longjohn, R., & Nottingham, K. (n.d.). *UCI Machine Learning Repository*. https://archive.ics.uci.edu/
21. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1143). Morgan Kaufmann. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf
22. Lohweg, V. (2012). *Banknote authentication* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C55P57
23. Lynam, A. L., Dennis, J. M., Owen, K. R., Oram, R. A., Jones, A. G., Shields, B. M., & Ferrat, L. A. (2020). Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: Application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagnostic and Prognostic Research, 4*(1), 6. https://doi.org/10.1186/s41512-020-00075-2
24. McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (pp. 41–48). AAAI Press. https://cdn.aaai.org/Workshops/1998/WS-98-05/WS98-05-007.pdf
25. McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a
26. Naser, M. Z. (2026). A review of machine learning with small and limited data. *Journal of Big Data, 13*(1), 18. https://doi.org/10.1186/s40537-025-01346-9
27. Park, H.-A. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing, 43*(2), 154–164. https://doi.org/10.4040/jkan.2013.43.2.154
28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, R., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html
29. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.
30. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1023/A:1022643204877

31. Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532–538). Springer. https://doi.org/10.1007/978-0-387-39940-9_565
32. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
33. Van Smeden, M., Groenwold, R. H. H., Moons, K. G. M., Collins, G. S., & Riley, R. D. (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
34. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, P., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2
35. Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast cancer Wisconsin (diagnostic)* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B
36. Zantvoort, K., Nacke, B., Görlich, D., Hornstein, S., Jacobi, C., & Funk, B. (2024). Estimation of minimal data set sizes for machine learning predictions in digital mental health interventions. *npj Digital Medicine, 7*, Article 361. https://doi.org/10.1038/s41746-024-01360-w