

Measuring Model Durability under Progressive Tabular Distribution Shift: Absolute Performance, Operational Failure, and Label- Free Monitoring

Aras Masood Ismael

`aras.masood@spu.edu.iq`

Technical College of Informatics, Sulaimani Polytechnic University

Research Article

Keywords: distribution shift, model durability, balanced accuracy, label-free performance estimation, natural geographic shift, tabular machine learning

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10940907/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Measuring Model Durability under Progressive Tabular Distribution Shift: Absolute Performance, Operational Failure, and Label-Free Monitoring

Aras Masood Ismael

Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani 46001, Iraq

Corresponding author: aras.masood@spu.edu.iq

Abstract

Performance measured on a fixed test set gives little information about how long a classifier remains useful after deployment conditions change. We evaluated logistic regression, random forest, XGBoost, and a multilayer perceptron on eight public tabular datasets using five outer partitions, three blocked model/shift-path replications, eight progressive shift mechanisms, and eleven severity levels. The synthetic benchmark comprised 480 fitted predictors and 42,240 shifted evaluations; two Folktables tasks provided a descriptive natural geographic-shift assessment. Absolute area under the balanced-accuracy curve and eligibility-gated operational survival were primary endpoints, while relative retention was secondary. Clean balanced accuracy was positively associated with absolute durability after dataset-cluster adjustment (Spearman $\rho=0.822$, 95% cluster-bootstrap interval 0.366–0.978; Holm-adjusted blocked-permutation $p=0.0004$). Relative retention showed the opposite association ($\rho=-0.680$), demonstrating a normalization effect when weak clean baselines have little performance to lose. Random forest had the largest mean absolute area, but no pairwise model comparison survived multiplicity correction. In deployment-realistic monitoring without nominal severity, DoC had the lowest mean error when complete datasets were withheld, whereas the random-forest regressor had the lowest mean error when shift families were withheld. ATC-MaxConf and ATC-NegEntropy had the lowest observed error on the two evaluated ACS tasks; this two-task result is descriptive. These findings support reporting absolute performance, application-defined failure, relative degradation, uncertainty, and natural-domain evidence as distinct quantities.

Keywords: distribution shift; model durability; balanced accuracy; label-free performance estimation; natural geographic shift; tabular machine learning

1. Introduction

Machine-learning systems are commonly selected by performance on a held-out sample drawn from the same data-generating process as the training set. This procedure is appropriate for estimating ordinary generalization error, but it does not describe performance after the population, measurement process, class balance, or operating environment changes. Such changes are collectively described as dataset shift. The statistical distinctions among covariate shift, prior-probability shift, selection bias, and concept shift are well established [31,34]. In streaming settings, related work on concept drift emphasizes that change may develop gradually and that its consequences depend on both the data process and the response strategy [17].

The practical difficulty is not merely to detect that two samples differ. A distributional discrepancy may be statistically detectable yet harmless to a predictor, while a change concentrated near a decision boundary may be damaging even when a global discrepancy is modest. Kernel two-sample tests and black-box shift detectors address the presence of distributional change [21,35]. Performance monitoring addresses a different estimand: the predictive quality that remains in the target condition. Explaining which variables contributed to a shift is another distinct task [28]. Conflating these tasks can produce operational alarms without a defensible connection to model utility.

Robustness benchmarks have made controlled stress testing a standard part of model evaluation. ImageNet-C introduced repeated corruption severities, WILDS formalized evaluation across naturally occurring domains, and later work showed that gains under synthetic perturbations need not transfer to natural distribution shifts [24,27,37]. Tabular studies have now documented similar heterogeneity. TableShift evaluates real distribution shifts in tabular tasks, TabFSBench studies feature shifts in open environments, and clinical validation work has shown that confident predictions can remain fragile after population change [9,11,18]. These studies establish the importance of the problem, but they leave an important measurement choice unresolved: whether durability should be expressed on an absolute performance scale or relative to each model's clean baseline.

Relative normalization is attractive because it makes curves appear comparable across models with different starting points. It can also create a floor effect. A classifier with balanced accuracy near 0.50 has little useful performance to lose and can therefore retain almost all of its baseline under severe shift. A stronger classifier can lose a larger fraction and still outperform the weak classifier at every severity. A high relative-retention score cannot consequently be interpreted as evidence that a model is useful. The present study makes this issue testable by separating three quantities: absolute performance accumulated

across the stress path, operational survival above prespecified utility thresholds, and proportional retention relative to the clean score.

A second objective concerns monitoring before target labels arrive. Recent methods estimate target accuracy from confidence, output distributions, feature summaries, or learned relationships across auxiliary shifts [19,22,39]. Their assumptions differ, and performance can depend strongly on whether evaluation withholds a shift type, a complete dataset, or a natural domain. Overlap-aware and covariate-shift estimators further demonstrate that support mismatch limits what can be inferred from unlabeled target data [2,8,30]. We therefore evaluate label-free estimators under three transfer regimes rather than reporting only random condition-level splits.

This study contributes an evaluation framework rather than a new predictive architecture. It combines progressive tabular stress paths with raw sustained-failure detection, explicit right-censoring, absolute and operational areas, sensitivity to severity parameterization, dataset-cluster-aware inference, repeated outer partitions, and label-free performance estimation. The empirical analysis addresses whether the association between clean performance and durability depends on the endpoint definition, whether conclusions are stable to the severity scale, how sampling variability compares with joint model/shift-path variability, and which monitoring approaches transfer beyond familiar datasets and shifts.

2. Related work

2.1 Distribution shift and progressive evaluation

Dataset-shift theory describes changes in components of the joint distribution of predictors and outcomes [31,34]. Label-shift methods focus on changing class proportions under comparatively restrictive assumptions about class-conditional distributions [29]. RLSbench relaxes parts of that setting and evaluates adaptation across more realistic deviations [20]. Our label-prevalence perturbation belongs to this broader family, but it is only one of eight mechanisms and is not used as a proxy for feature or conditional change.

Progressive evaluation provides a curve rather than one arbitrary target condition. Corruption benchmarks use ordered severities to reveal nonlinear degradation and failure thresholds [24]. The approach is useful because a model can be superior under mild change and inferior under stronger change. It is also sensitive to how severity is parameterized. A unit of missingness, noise, scaling, and contamination has no common physical meaning. For that reason, the present analysis reports the physical endpoint of each transformation and repeats integration under nonlinear nominal axes and an empirical distribution-distance axis.

2.2 Absolute, relative, and operational robustness

Absolute performance and relative degradation answer different scientific questions. Absolute scores preserve the metric scale and directly represent predictive quality. Relative scores describe proportional change from a model-specific baseline. Natural-shift studies have shown that in-distribution accuracy often predicts out-of-distribution accuracy, which makes uncontrolled comparisons of relative or effective robustness difficult to interpret [36,37]. Our analysis does not treat one summary as universally correct. Instead, it assigns each summary a bounded interpretation and gives primary status to endpoints that cannot reward a model merely for beginning near chance.

The survival analogy is used only for ordered threshold crossing. Kaplan and Meier formalized estimation with censored event times [26]. Here, the event coordinate is shift severity rather than chronological time. A trajectory that does not cross the operational boundary is right-censored at the largest tested severity. This language is useful for preserving information about non-failure, but it does not imply a stochastic time-to-event process or a clinical survival model.

2.3 Label-free performance estimation

Confidence-based estimators rely on the relationship between predictive confidence and correctness. Calibration methods show why that relationship must be assessed rather than assumed [23]. Difference of Confidence adjusts source accuracy by the change in mean confidence. Average Thresholded Confidence learns a threshold on labeled source predictions and applies it to unlabeled target predictions. Learned meta-models instead map shift descriptors to observed performance across auxiliary environments [19,22,39]. The flexibility of learned estimators can improve interpolation but also increases the risk of failure when the deployment dataset lies outside the meta-training collection.

2.4 Statistical evaluation across datasets

Rows generated from the same dataset are dependent. Models share observations, severity points share a fitted predictor and a nested perturbation path, and replications within an outer split share the same partition. Algorithm comparisons therefore preserve the dataset as the inferential block [13]. Benchmark studies also show that random variation can alter apparent rankings [5]. We address these issues through dataset-level summaries, blocked permutation, cluster-bootstrap intervals, repeated outer partitions, explicit reporting of partition variability and joint model/shift-path variability, and multiplicity correction [15,25,33].

3. Materials and methods

3.1 Study design and datasets

The synthetic benchmark comprised Adult, Bank Marketing, Electricity, Covertypes, Numerai, Nomao, Phoneme, and MiniBooNE, identified by immutable OpenML data identifiers [38]. Datasets were specified for the revised analysis before the V6 rerun, after weaknesses in the earlier fixed-split study had been identified; the study was not preregistered, and we do not describe this choice as prospective prespecification. Inclusion required a public classification dataset, at least 5,000 observations, compatibility with repeated train-validation-test evaluation, and diversity in sample size, dimensionality, class count, feature type, and baseline difficulty. We excluded regression tasks, datasets below the size threshold, inaccessible or inactive versions, and duplicate or closely related versions of an included benchmark. Exact OpenML versions were fixed by data ID to prevent silent substitution. The resulting collection is a purposive stress-test panel rather than a probability sample of tabular learning problems.

Table 1. Dataset characteristics. IR, class-imbalance ratio (smallest class count divided by largest class count); Num., numerical features; Cat., categorical features; Miss., percentage of missing values.

Dataset	OpenML ID	N	Raw features	Classes	Num.	Cat.	IR	Miss.	Task
Adult	1590	48,842	14	2	6	8	0.315	1.0%	Binary
Bank Marketing	1461	45,211	16	2	7	9	0.132	0%	Binary
Electricity	151	45,312	8	2	8	0	0.738	0%	Binary
Covertypes	1596	581,012	54	7	54	0	0.010	0%	7-class
Numerai	23517	96,320	21	2	21	0	0.980	0%	Binary
Nomao	1486	34,465	118	2	89	29	0.380	38%	Binary
Phoneme	1489	5,404	5	2	5	0	0.415	0%	Binary
MiniBooNE	41150	130,064	50	2	50	0	0.390	0%	Binary

Natural validation used ACSIncome and ACSEmployment generated from 2018 American Community Survey public-use microdata through Folktables [14]. California was the source domain; Texas, New York, Florida, and Illinois were unordered target domains. Each state was deterministically limited to 50,000 observations. Because states do not constitute an intrinsic severity sequence, no area-under-geography endpoint is computed in this revision. We report source performance, mean target performance, worst-target performance, and source-to-target changes. Empirical sliced Wasserstein distance is shown only as a descriptive batch-distance covariate and is not used to impose a geographic order.

3.2 Data partitioning and leakage control

For Adult, Bank Marketing, Covertypes, Numerai, Nomao, Phoneme, and MiniBooNE, each outer partition used a stratified 70% training, 15% validation, and 15% test split. The five partitions were generated independently with random seeds 42, 123, 999, 2024, and 2025. A partition was created once and shared by all four model families and all three model seeds, enabling paired comparisons. Within every outer partition, constant-feature removal, numerical and categorical imputation, standardization, one-hot encoding, and binary threshold selection were refitted without access to the test set. The preprocessing pipeline used only the

training subset. Each fitted model selected its binary decision threshold on that partition's validation subset by maximizing balanced accuracy; the threshold was then fixed for clean and shifted test evaluation.

Electricity preserved source order. For the five rolling-origin partitions, the initial training windows contained 45%, 50%, 55%, 60%, and 65% of observations. In every partition, the next 15% formed validation data and the following 15% formed test data; observations after the test window were unused. The windows therefore overlap and are repeated temporal evaluations, not five independent cohorts. All models and model seeds used the same window boundaries.

For each Folktables task, California observations were deterministically capped at 50,000 and split with the same five stratified 70%/15%/15% outer partitions. Training preprocessing and model-specific validation threshold selection were repeated inside every California partition. The held-out California 15% was the source-domain test set. Texas, New York, Florida, and Illinois were each deterministically capped at 50,000 and used only as external target-domain evaluations; no target-state labels entered fitting, preprocessing, threshold selection, or label-free estimation.

These controls prevent direct test-label leakage. They do not make repeated outer partitions statistically independent. Inferential analyses therefore first aggregate to dataset-level blocks. Outer-partition variation and joint model/shift-path variation are reported descriptively rather than interpreted as independent sources of uncertainty.

3.3 Models and preprocessing

Preprocessing was fitted using training data only. Constant variables were removed. Numeric variables were median-imputed and standardized. Categorical variables were imputed by their most frequent category and one-hot encoded, with unseen levels ignored. Binary decision thresholds were chosen on validation data to maximize balanced accuracy and then fixed for all test conditions.

Logistic regression used an L2 penalty, $C=1.0$, L-BFGS optimization, balanced class weights, and a maximum of 1,000 iterations. Random forest used 300 trees, maximum depth 24, square-root feature sampling, minimum leaf size 2, balanced-subsample class weights, and unrestricted parallel CPU fitting. XGBoost used 400 histogram trees, maximum depth 8, learning rate 0.05, minimum child weight 1, row subsampling 0.80, column subsampling 0.80, L2 regularization $\lambda=1.0$, class-balanced sample weights, and binary logistic or multiclass soft-probability loss. XGBoost did not use early stopping; the tree count was fixed. The multilayer perceptron used hidden layers of 256 and 128 units with batch normalization, GELU activation, dropout 0.20, class-weighted cross-entropy, AdamW with learning rate 0.001 and weight decay 0.0001, batch size 1,024, gradient clipping at 5, mixed precision on GPU, ReduceLROnPlateau scheduling, and validation macro-F1 early stopping with patience 15 within 100 epochs [6,10,12,16].

The execution source records Python 3.13 and requires NumPy, pandas, SciPy, scikit-learn, PyTorch, XGBoost 2 or later, Matplotlib, Seaborn, and Folktables 0.0.12 or later. The original V6 archive did not retain a complete package lock file; therefore, exact minor versions of every library cannot be reconstructed retrospectively. The V10 package adds an environment-recording script and recommends running it before any future archival rerun. This limitation affects software-environment reconstruction, not the stored predictions or recomputed statistics.

3.4 Progressive shift construction

Shifts were applied to raw test features before imputation, standardization, and one-hot encoding; the training-fitted preprocessing pipeline was then applied unchanged. Gaussian noise, scaling, outlier contamination, and the score used for

subpopulation sampling operated only on columns classified as numerical in the training partition. Missingness and within-column permutation could affect every nonconstant raw predictor, including categorical predictors. No shifted numeric value was clipped before preprocessing. Unseen categorical levels were ignored by the fitted encoder. Labels were never used to construct covariate shifts; they were used only for the controlled label-prevalence resampling benchmark.

At severity s , Gaussian noise added s times the training-partition standard deviation using one fixed standard-normal matrix per path. Missingness revealed entries satisfying $U < 0.50s$ from one fixed uniform matrix. Scaling selected a fixed random half of numerical columns and transformed x to $m + (x-m)\exp(d s \log 2)$, where m is the training median and d is a fixed sign. Permutation used a fixed random column order, affected $\text{round}(0.50sd)$ columns, and used a fixed within-column row permutation. Outlier contamination selected the first $\text{round}(0.20sn)$ rows of a fixed row order and replaced each numerical value by m plus or minus $(3+3s)$ training standard deviations. Values were not clipped.

Label-prevalence shift sampled n rows with replacement, assigning the rarest clean-test class weight $\exp(s \log 9)$ and every other class weight one; in multiclass datasets only that rarest class received the elevated weight. Subpopulation shift averaged standardized values of the first eight numerical columns, converted the score to percentile ranks r , and sampled with weights $\exp(3s(r-0.5))$; when no numerical feature existed, the first categorical feature was factorized. Concurrent shift combined missingness $0.25s$, noise $0.60s$, and scaling strength $0.75s$. The path seed was the first 32 bits of SHA-256 applied to protocol-version, dataset-outer-partition, replication-seed, and shift-family identifiers. Consequently, all four model families within the same dataset $\times$ outer-partition $\times$ replication block were evaluated on identical shifted test data, preserving paired model comparisons. Across the three replication seeds, however, both model randomness and the shift realization changed. Variation across those replications is therefore termed joint model/shift-path variability and must not be interpreted as pure model-seed variability. Each severity reinitialized the generator with the same path seed; fixed base random arrays, fixed orders, and threshold revelation—not seed reuse alone—created nesting. Label and subpopulation resampling used common random-number streams but severity-dependent weights and are therefore coupled progressive paths rather than strict set inclusion.

3.5 Endpoints

All durability endpoints in Sections 3.5–3.6 estimate validation-threshold balanced accuracy $B(s)$, which protects interpretation under class imbalance. The separate label-free monitoring experiment estimates ordinary argmax accuracy because confidence-based estimators operate on the predicted class and its probability; these two estimands are intentionally not interchangeable. Balanced accuracy weights class-specific recall equally [7]. For K classes at severity s ,

$$B(s) = \frac{1}{K} \sum_{k=1}^K \frac{TP_k(s)}{TP_k(s) + FN_k(s)} \quad (1)$$

Absolute area under the balanced-accuracy curve was calculated by trapezoidal integration on severity axis a ,

$$\text{AUBA}^{(a)} = \sum_{i=0}^9 \frac{B(s_i) + B(s_{i+1})}{2} (a(s_{i+1}) - a(s_i)) \quad (2)$$

Relative retention was defined as the shifted score divided by the clean score. It was capped only to limit the influence of improvements under shift and was treated as secondary,

$$R(s_i) = \text{clip}\left(\frac{B(s_i)}{\max(B(0), \epsilon)}, 0, 1.25\right) \quad (3)$$

Operational eligibility was determined separately for every dataset $\times$ model $\times$ outer partition $\times$ model-seed replication from its clean balanced accuracy $B(0)$. The indicator was then inherited by all eight shift-family trajectories belonging to that fitted replication. Eligible replications satisfied $B(0) \geq 0.70$. Failure was located on the raw, unsmoothed trajectory and required two consecutive severity points for which $B(s) < 0.70$ or $R(s) < 0.80$. A nonfailing eligible trajectory was right-censored at severity 1.0; ineligible trajectories contributed neither operational area nor censoring. Adult, Bank Marketing, Coverttype, MiniBooNE, Nomao, and Phoneme had all four models eligible in every one of the 15 partition-seed blocks and formed the paired operational comparison. Electricity and Numerai had no eligible replications. Cumulative-minimum and decreasing-isotonic curves were sensitivity analyses only [1].

$$e = \min\{i: F_i = 1 \wedge F_{i+1} = 1, i = 0, \dots, 9\}, \quad F_i = \mathbf{1}\{B(s_i) < 0.70 \vee R(s_i) < 0.80\} \quad (4)$$

The event severity was assigned to the first point s_e of the qualifying pair, not the second point, and no interpolation was performed between grid values. Define $H_i=1$ for $i \leq e$ and $H_i=0$ for $i \geq e$. For a right-censored trajectory, $H_i=1$ at every tested interval. The operational survival area was

$$\text{OSA} = \sum_{i=0}^9 H_i (s_{i+1} - s_i) \quad (5)$$

Thus, an uncensored failure beginning at severity s_e contributes area s_e on the equally spaced grid, while a trajectory censored at severity 1.0 contributes area 1.0. Areas were first calculated separately for every dataset $\times$ model $\times$ outer partition $\times$ model seed $\times$ shift family. Dataset-model summaries were arithmetic means across eight shift families and 15 partition-seed replications after excluding ineligible replications; only datasets with complete eligibility for all four models entered paired operational tests.

3.6 Severity parameterization

The primary severity axis was linear. Early- and late-weighted sensitivity axes were defined as

$$a_{\text{linear}}(s) = s, \quad a_{\text{early}}(s) = \sqrt{s}, \quad a_{\text{late}}(s) = s^2 \quad (6)$$

A fourth axis used sliced Wasserstein distance in standardized feature space. The estimate averaged one-dimensional Wasserstein distances over 32 random projections and was normalized within each perturbation path [4]. This axis measures distributional displacement in the model input representation. It does not establish equal operational severity across shift mechanisms.

3.7 Label-free accuracy estimation

The monitoring target was ordinary argmax accuracy, not validation-threshold balanced accuracy. For target probabilities $\mathbf{p}_r$, maximum confidence was $c_r = \max_k p_{rk}$ and negative entropy was $g_r = -\sum_k p_{rk} \log(\max(p_{rk}, 10^{-8}))$. No change predicted source-validation accuracy A_{val} . Average Confidence (AC) predicted $n^{-1} \sum_r c_r$. Difference of Confidence (DoC) predicted $\text{clip}(A_{\text{val}} + \bar{c}_{\text{target}} - \bar{c}_{\text{val}}, 0, 1)$, where $\bar{c}_{\text{target}}$ and $\bar{c}_{\text{val}}$ are the target-batch and validation-set mean maximum confidences. For each

Average Thresholded Confidence estimator, a validation threshold q was the linear $(1-A_{val})$ quantile of the validation score. ATC-MaxConf predicted the target proportion with $c_r \geq q$; ATC-NegEntropy applied the same procedure to g_r [19,22,39].

$$\hat{A}_{DoC} = \text{clip}(A_{val} + \bar{c}_{target} - \bar{c}_{val}, 0, 1) \quad (7)$$

Learned estimators were ridge regression with $\alpha=1.0$, random-forest regression with 200 trees, maximum depth 12, square-root feature sampling, and minimum leaf size 3, and a feed-forward network (FFN) with 64- and 32-unit GELU layers and dropout 0.15. The FFN used Smooth-L1 loss, AdamW with learning rate 0.002 and weight decay 0.0001, gradient clipping at 5, a random 80%/20% internal fit-validation split, at most 100 full-batch epochs, and patience 15. Model identity was one-hot encoded. Continuous meta-features were standardized using means and standard deviations estimated only from the meta-training portion of each LODO or LOSO fold; zero-variance columns were assigned unit scale. Severity-zero rows were removed before meta-model fitting, so the eight repeated clean conditions did not receive extra weight. The primary deployment specification excluded nominal severity because that variable would ordinarily be unknown for an incoming deployment batch. Primary learned-model results are available for ridge and random-forest regression. The archived severity-assisted ridge, random-forest, and FFN analyses are retained only as a sensitivity benchmark in Supplementary Table S9; direct estimators never used severity.

Leave-one-dataset-out (LODO) held out one of eight complete datasets; leave-one-shift-out (LOSO) held out one of eight shift families. Synthetic-to-natural transfer trained learned models on synthetic conditions and evaluated the two ACS tasks. Predictions from the three learned-forecaster seeds were averaged for each exact target condition before error calculation. Fold MAE was then the unweighted mean absolute error across all conditions in the held-out dataset or shift family; natural evaluation used the task as the independent cluster. Cluster-bootstrap intervals were calculated by resampling the eight LODO or eight LOSO folds. Because the natural evaluation contained only two independent ACS tasks, natural-transfer uncertainty is reported descriptively using the observed task-level range rather than a confidence interval. Direct estimators have no meta-training and therefore use the same predictions under LODO and LOSO; only their fold grouping differs.

For the ACS targets, state labels were sequestered during estimation. Only unlabeled target features and model probabilities were available to the monitoring procedures. Labels were accessed afterward solely to compute ordinary argmax accuracy and the estimation errors reported in Results.

3.8 Statistical analysis

Four clean-performance association tests formed one multiplicity family because each tested the same predictor, clean balanced accuracy, against a related durability summary: linear-axis absolute AUBA, empirical-distance AUBA, operational survival area, and secondary relative STI. Linear AUBA and operational survival were the primary scientific endpoints; empirical-distance AUBA was a severity-parameterization sensitivity analysis, and relative STI was a diagnostic secondary endpoint. Holm correction was applied across all four to conservatively control family-wise error despite these different interpretive roles [25]. The overall Spearman coefficient used 32 dataset-model summaries. Its 95% cluster-bootstrap interval resampled eight datasets while retaining four-model blocks. A blocked permutation reassigned endpoint values among models within each dataset and tested the centered within-dataset association; Monte Carlo p values used the plus-one correction [33].

Forecasting uncertainty used the held-out unit rather than individual condition rows. Cluster-bootstrap intervals were calculated across eight dataset folds for LODO and eight shift-family folds for LOSO. Synthetic-to-natural transfer was summarized using the mean and observed range across the two ACS tasks; no population-level confidence interval or significance test was interpreted.

Model families were compared using a Friedman test across complete dataset blocks followed by paired Wilcoxon signed-rank tests with Holm adjustment [13]. For linear AUBA, $\chi^2(3)=9.45$, $N=8$; Kendall's W was calculated as $\chi^2/[N(k-1)]=0.394$. For operational survival, $\chi^2(3)=5.24$, $N=6$ and $W=0.291$. Pairwise summaries include the dataset-level mean and median difference and a dataset-bootstrap 95% confidence interval. Values shown as mean $\pm$ SD use the sample standard deviation between dataset-level model summaries unless a table states otherwise. Partition variability was the standard deviation across five outer-partition means after averaging the three replication blocks. Joint model/shift-path variability was the mean within-partition standard deviation across the three blocks; because both the trained-model seed and perturbation path changed, it does not isolate training instability.

4. Results

4.1 Execution and clean performance

All eight checks in Supplementary Validation Checklist S8 passed: exact synthetic-row count, complete outer partitions, complete model seeds, presence of both primary endpoints, explicit secondary status for relative STI, presence of cluster-aware inference, complete natural tasks, and complete natural domains. The synthetic analysis produced 42,240 condition rows and 3,840 endpoint rows without duplicate identifiers. Across datasets and repetitions, mean clean balanced accuracy was 0.824 for XGBoost, 0.817 for random forest, 0.809 for the multilayer perceptron, and 0.758 for logistic regression. The complete study contained 600 fitted predictors. Recorded fitting time was 20,061.6 seconds, or 5.57 hours; runtime is descriptive because hardware conditions were not controlled.

Table 2. Clean-test performance pooled across datasets and repetitions. MLP, multilayer perceptron; SD, standard deviation.

Calculation of Table 2: for each model and dataset, each metric was first averaged over the 15 fitted replications (five outer partitions $\times$ three blocked model/shift-path replications). The table reports the unweighted mean of the resulting eight dataset-level means. The value after $\pm$ is the sample SD across those eight dataset-level means; it is not the SD across all 120 fitted models and does not treat repeated partitions as independent observations.

Model	Balanced accuracy	Macro-F1	AUROC	Brier	ECE
Logistic regression	0.758 $\pm$ 0.138	0.689 $\pm$ 0.157	0.849 $\pm$ 0.145	0.334 $\pm$ 0.167	0.064 $\pm$ 0.072
MLP	0.809 $\pm$ 0.156	0.755 $\pm$ 0.152	0.875 $\pm$ 0.157	0.270 $\pm$ 0.174	0.059 $\pm$ 0.086
Random forest	0.817 $\pm$ 0.159	0.779 $\pm$ 0.159	0.879 $\pm$ 0.165	0.225 $\pm$ 0.163	0.067 $\pm$ 0.054
XGBoost	0.824 $\pm$ 0.160	0.777 $\pm$ 0.160	0.883 $\pm$ 0.165	0.243 $\pm$ 0.193	0.072 $\pm$ 0.087

4.2 Association between clean performance and durability

Clean balanced accuracy was positively associated with linear-axis AUBA. The overall Spearman coefficient was 0.822, the within-dataset coefficient was 0.865, the dataset-cluster bootstrap interval was 0.366 to 0.978, and the Holm-adjusted blocked-permutation probability was 0.0004. The empirical-distance result was similar, with an overall coefficient of 0.807 and an adjusted probability of 0.0004. The data therefore show that clean performance and absolute durability generally increased together in this benchmark.

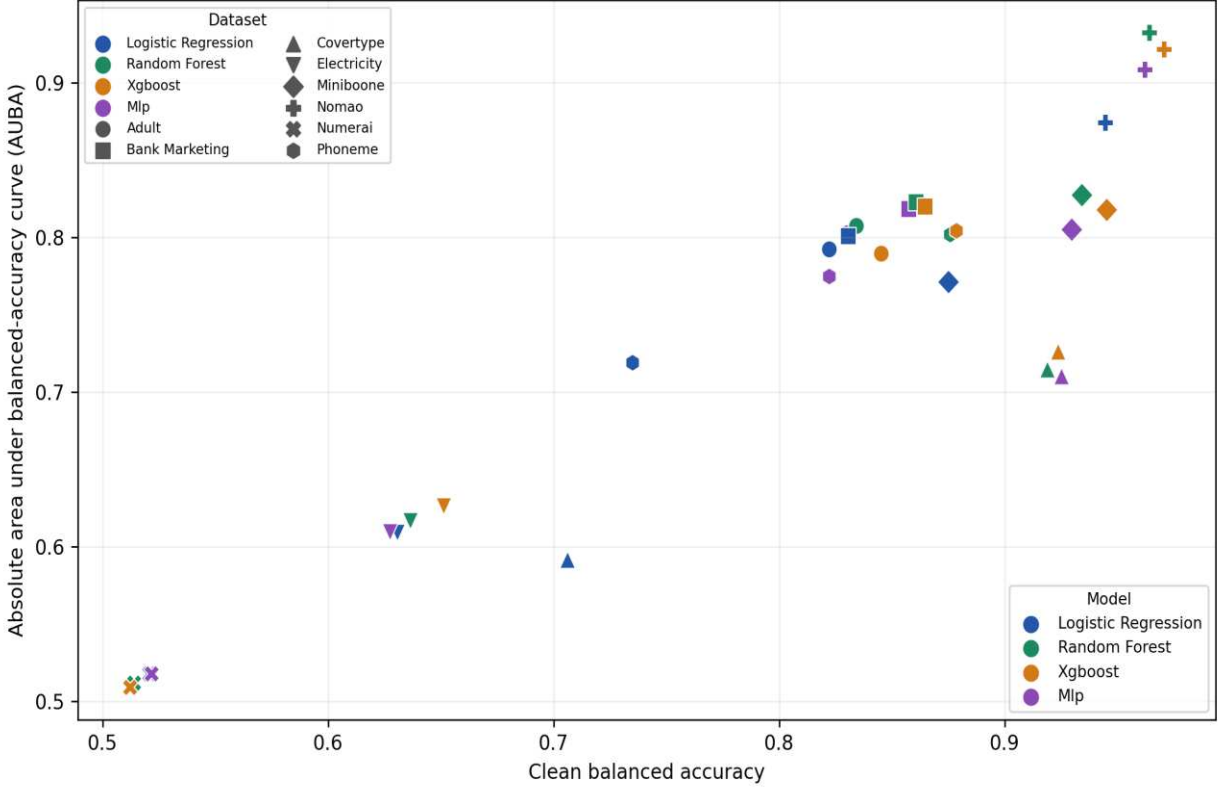


Figure 1. Clean balanced accuracy and linear-axis area under the balanced-accuracy curve (AUBA) for the 32 dataset–model summaries.

Relative STI produced the opposite conclusion. Its association with clean balanced accuracy was -0.680 , with a cluster interval from -0.919 to -0.086 and an adjusted probability of 0.0044 . Numerai provides the clearest example: model performance was close to chance, yet relative retention was nearly one. The reversal is therefore specific to normalization by the clean baseline and should not be described as evidence that low-performing models are operationally robust.

4.3 Model comparison and severity sensitivity

Mean linear AUBA was 0.754 for random forest, 0.752 for XGBoost, 0.743 for the multilayer perceptron, and 0.710 for logistic regression. The dataset-blocked omnibus comparison yielded $\chi^2(3)=9.45$, $p=0.0239$ and Kendall's $W=0.394$. None of six paired Wilcoxon comparisons survived Holm adjustment; the smallest adjusted p value was 0.0938 . For example, random forest minus XGBoost had mean AUBA difference 0.0025 with a dataset-bootstrap 95% interval from -0.0041 to 0.0088 . Operational survival yielded $\chi^2(3)=5.24$, $p=0.155$ and $W=0.291$ across the six fully eligible datasets. Tree ensembles therefore had the largest descriptive mean areas, but corrected pairwise superiority was not established.

Table 3. Operational survival by model. Mean operational survival area is calculated among eligible trajectories; censoring is the percentage of eligible trajectories without sustained failure by severity 1.0. MLP, multilayer perceptron.

Model	Mean operational area	Eligible trajectories	Censored among eligible
Logistic regression	0.7951	720/960	66.7%
MLP	0.8574	720/960	71.3%
Random forest	0.8642	720/960	75.1%
XGBoost	0.8490	720/960	69.7%

Rankings were stable when the severity axis changed. Spearman rank agreement with the linear axis was 0.989 for early weighting, 0.995 for late weighting, and 0.996 for empirical distance. The leading model agreed in seven of eight datasets under early weighting and in all eight datasets under late weighting and empirical distance. Absolute areas changed more than ranks, confirming that severity mapping affects magnitude even when comparative ordering is stable.

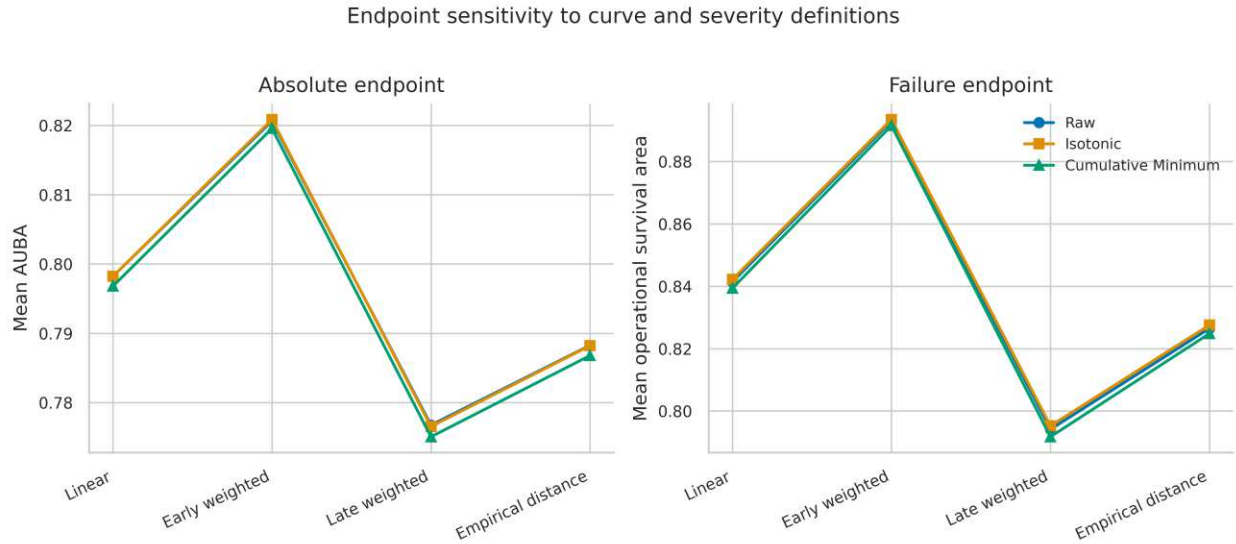


Figure 2. Sensitivity of absolute and operational areas to curve definition and severity parameterization.

4.4 Shift-family heterogeneity and replication

Concurrent perturbation and Gaussian noise were the most damaging shift families on the absolute scale, with pooled AUBA values of 0.681 and 0.684. Feature permutation and missingness followed at 0.698 and 0.714. Scaling, outlier contamination, subpopulation resampling, and label-prevalence resampling were less damaging under the implemented transformations. This ordering applies only to the specified physical perturbations. Balanced accuracy also reduces sensitivity to pure class-prevalence change when class-conditional behavior is preserved.

For linear AUBA, the median outer-partition standard deviation was 0.0050 and the median within-partition joint model/shift-path standard deviation was 0.0044. Outer variation was larger in 51.2% of dataset-model-shift cells. The corresponding

empirical-distance values were both approximately 0.0061. Operational areas often had zero estimated variation because the coarse grid and right-censoring produced identical threshold-crossing locations. A zero value in that setting indicates discrete measurement, not certainty.

4.5 Label-free monitoring

4.5.1 Synthetic transfer and withheld-group evaluation

Under LODO evaluation, DoC had the lowest mean fold MAE, 0.0536 (95% dataset-cluster interval 0.0306–0.0792), with mean fold RMSE 0.0798 and signed error 0.0455. No change had MAE 0.0602 (0.0358–0.0838). Among deployment-realistic learned estimators without nominal severity, random-forest regression had LODO MAE 0.0836 (0.0599–0.1111) and ridge had 0.1155 (0.0780–0.1718). Under LOSO evaluation, the no-severity random-forest regressor had the lowest mean fold MAE, 0.0308 (0.0246–0.0375), followed by no-severity ridge at 0.0532 (0.0486–0.0575) and DoC at 0.0536 (0.0361–0.0725). Direct estimators require no meta-training; their predictions are identical under LODO and LOSO, while only the held-out grouping used for fold summaries changes. Severity-assisted learned estimates are reported as a sensitivity analysis in Supplementary Table S9.

Table 4. Leave-one-dataset-out (LODO) label-free monitoring. Main learned-model rows exclude nominal severity. AC, Average Confidence; DoC, Difference of Confidence; ATC, Average Thresholded Confidence; RF, random forest; MAE, mean absolute error; RMSE, root-mean-square error.

Estimator	N	Mean MAE	95% cluster interval	Mean RMSE	Signed error
No change	8	0.0602	0.0358–0.0838	0.0844	0.0559
AC	8	0.0783	0.0463–0.1203	0.1017	0.0581
DoC	8	0.0536	0.0306–0.0792	0.0798	0.0455
ATC-MaxConf	8	0.0642	0.0368–0.0990	0.0928	0.0439
ATC-NegEntropy	8	0.0649	0.0374–0.0997	0.0933	0.0437
Ridge	8	0.1155	0.0780–0.1718	0.1292	-0.0641
RF regressor	8	0.0836	0.0599–0.1111	0.1017	-0.0065

Table 5. Leave-one-shift-out (LOSO) label-free monitoring. Main learned-model rows exclude nominal severity. AC, Average Confidence; DoC, Difference of Confidence; ATC, Average Thresholded Confidence; RF, random forest; MAE, mean absolute error; RMSE, root-mean-square error.

Estimator	N	Mean MAE	95% cluster interval	Mean RMSE	Signed error
No change	8	0.0602	0.0389–0.0821	0.0889	0.0559
AC	8	0.0783	0.0620–0.0961	0.1160	0.0581
DoC	8	0.0536	0.0361–0.0725	0.0878	0.0455
ATC-MaxConf	8	0.0642	0.0457–0.0842	0.1039	0.0439
ATC-NegEntropy	8	0.0649	0.0461–0.0846	0.1046	0.0437
Ridge	8	0.0532	0.0486–0.0575	0.0763	-0.0029
RF regressor	8	0.0308	0.0246–0.0375	0.0480	-0.0014

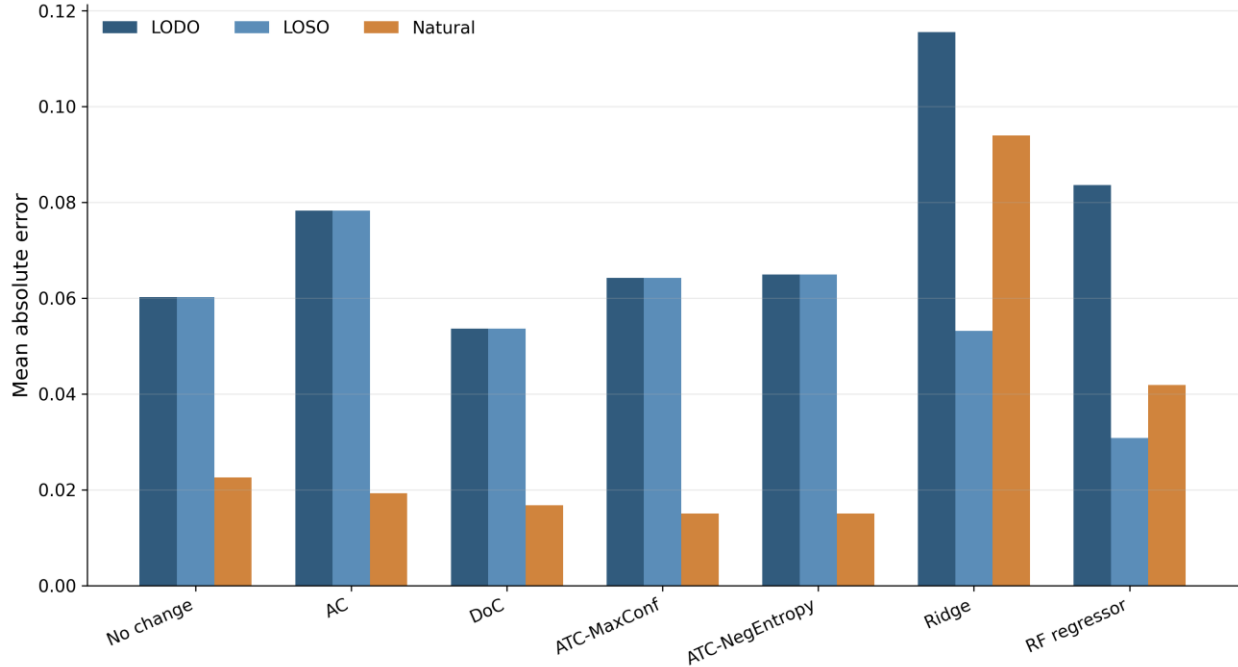


Figure 3. Mean absolute error for deployment-realistic label-free accuracy estimators under leave-one-dataset-out (LODO), leave-one-shift-out (LOSO), and descriptive natural-task evaluation. Learned ridge and random-forest (RF) regressors exclude nominal severity. AC, Average Confidence; DoC, Difference of Confidence; ATC, Average Thresholded Confidence.

4.5.2 Synthetic-to-natural transfer

ATC-MaxConf and ATC-NegEntropy each had mean task-level MAE 0.0150, with an observed two-task range of 0.0125–0.0175, mean RMSE 0.0183, and signed error 0.0025. DoC had MAE 0.0167, AC 0.0193, and no change 0.0226. For the deployment-realistic learned models, no-severity random-forest regression had MAE 0.0419 (observed range 0.0271–0.0567) and no-severity ridge had 0.0939 (0.0810–0.1068). Target-state labels were not used to construct any label-free estimate, tune confidence thresholds, fit the meta-models, or select a monitoring method. They were revealed only after prediction to calculate the true target accuracy and estimation error. These are descriptive observations from ACSIncome and ACSEmployment, not estimates of general natural-shift superiority.

Table 6. Synthetic-to-natural label-free monitoring across two independent American Community Survey (ACS) tasks. The interval column is the observed two-task range, not a confidence interval. AC, Average Confidence; DoC, Difference of Confidence; ATC, Average Thresholded Confidence; RF, random forest; MAE, mean absolute error; RMSE, root-mean-square error.

Estimator	N	Mean MAE	Observed range	Mean RMSE	Signed error
No change	2	0.0226	0.0080–0.0372	0.0248	0.0150
AC	2	0.0193	0.0081–0.0304	0.0223	0.0154
DoC	2	0.0167	0.0092–0.0243	0.0189	0.0076
ATC-MaxConf	2	0.0150	0.0125–0.0175	0.0183	0.0025
ATC-NegEntropy	2	0.0150	0.0125–0.0175	0.0183	0.0025
Ridge	2	0.0939	0.0810–0.1068	0.0947	-0.0939
RF regressor	2	0.0419	0.0271–0.0567	0.0493	-0.0388

4.6 Natural geographic shift

Geographic domains were analyzed as unordered targets. On ACSIncome, all four models had lower mean balanced accuracy outside California; source-to-mean-target reductions ranged from approximately 0.020 to 0.028 across model families. On ACSEmployment, mean target performance was slightly higher than California performance for every model, although worst-state performance could still be lower within individual replications. Across both tasks, mean target balanced accuracy was 0.809 for XGBoost, 0.805 for random forest, 0.799 for the multilayer perceptron, and 0.771 for logistic regression. No “Natural AUBA” is reported, because the four states do not define an ordered severity path. The supplementary table reports source, mean target, worst target, source-to-mean change, and source-to-worst change for every model and replication.

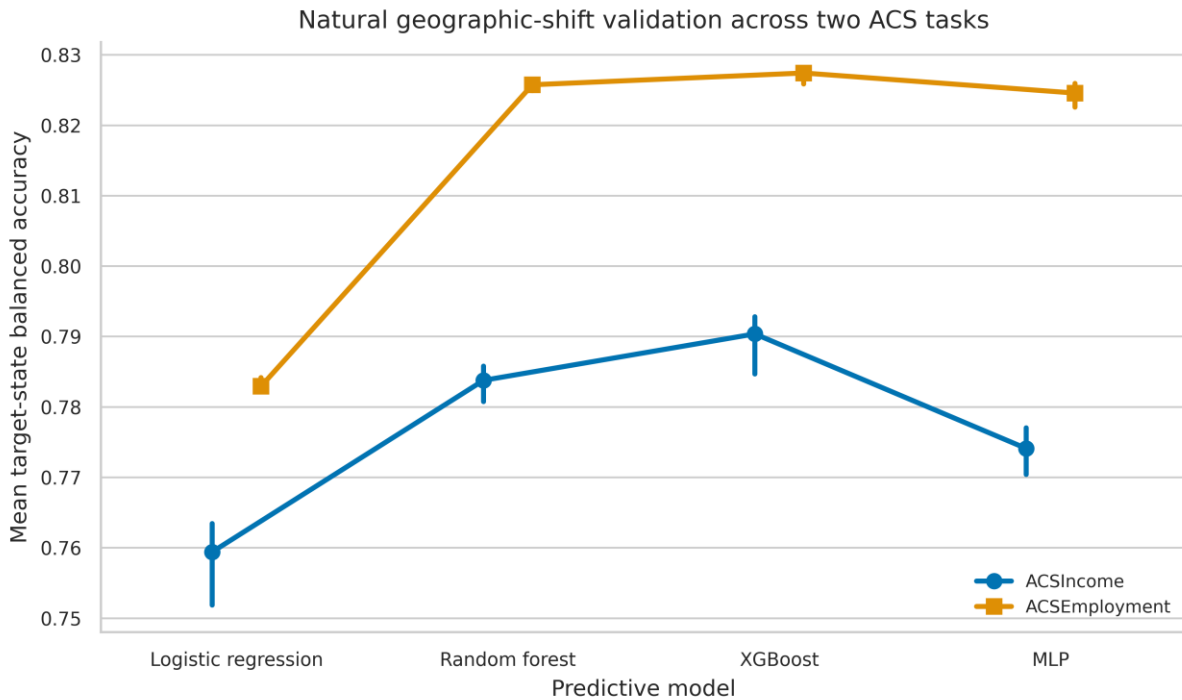


Figure 4. Mean target-state balanced accuracy for the American Community Survey (ACS) Income and Employment tasks.

5. Discussion

The principal result is a measurement result. The direction of the association between clean performance and durability changed when the endpoint changed from relative retention to absolute performance. Relative STI described proportional degradation correctly, but its denominator rewarded models with weak clean baselines. Absolute AUBA preserved the scale on which model usefulness was originally assessed. Operational survival added a utility floor and sustained failure rule. Taken together, these endpoints show why robustness claims must identify the quantity being measured rather than treating robustness as a single model property.

The positive association between clean performance and absolute AUBA is descriptive, not causal. Dataset difficulty, model capacity, and shift construction can affect both variables. Dataset centering and blocked inference preserve dependence among models on the same task but do not randomize architecture or remove all confounding. Operational survival illustrates a separate compositional effect: its overall correlation with clean balanced accuracy was -0.140 , whereas the within-dataset centered coefficient was 0.457 . The sign change occurs because eligibility removes two low-performing datasets and because datasets differ greatly in censoring and failure severity; between-dataset difficulty and within-dataset model ordering consequently act in different directions. The cluster interval, -0.616 to 0.698 , spans both substantial negative and positive associations, and the adjusted p value of 0.0737 is treated as inconclusive rather than suggestive evidence.

Severity parameterization remains a substantive limitation of any area-under-stress curve. Missingness probability, Gaussian standard deviation, scaling magnitude, and contamination fraction are not interchangeable units. High rank agreement across four axes suggests that model ordering was not fragile in this experiment. It does not make the axes physically equivalent. Reporting the transformation endpoints and repeating analysis on an empirical-distance axis are therefore necessary components of interpretation rather than optional supplementary details.

The monitoring results distinguish interpolation from extrapolation and oracle from deployable information. Random-forest regression remained effective in LOSO after nominal severity was removed, whereas simple confidence corrections were more competitive when complete datasets were withheld. The original severity feature and the natural normalization by the maximum target-state distance are retained only to document the oracle ablation and are not used for deployment claims. Random condition-level cross-validation would have overstated generalization because severities share perturbation paths, models share partitions, and states share source fits.

The study has several strengths: repeated outer partitions, blocked paired model comparisons, raw-trajectory event detection, explicit censoring, multiple severity axes, dataset-cluster inference, published monitoring baselines, natural-shift evaluation, and complete condition-level outputs. It also has important limitations. Across the three replication seeds, trained-model randomness and the synthetic shift path changed together; the reported within-partition variation is therefore joint model/shift-path variability and cannot isolate training instability. Eight synthetic datasets and two ACS tasks cannot represent all deployment environments. Seven datasets use overlapping repeated partitions, and the Electricity rolling origins are not independent cohorts. Synthetic perturbations are controlled stress tests rather than forecasts of future deployment change. Operational thresholds are application dependent. The empirical distance depends on preprocessing and projection sampling. The natural analysis covers geographic

variation in one survey year but not temporal or institutional change. Hyperparameters were fixed, and only four model families were compared.

A defensible deployment protocol follows from these findings. Model selection should retain an absolute performance curve, define a utility threshold before examining failures, detect sustained failure on the raw trajectory, report censoring, and repeat integration under plausible severity mappings. Relative retention may then be reported as a secondary description of proportional degradation. Label-free estimates should support investigation and prioritization rather than automatic intervention unless the estimator has been validated on comparable natural shifts. Future work should replace generic thresholds with domain-specific utility, add temporal and institutional shifts, use finer adaptive severity grids, and estimate hierarchical sources of variation across a larger collection of independent datasets.

6. Conclusion

Model durability under distribution shift depends on how durability is defined. In this multi-dataset tabular benchmark, clean balanced accuracy was positively associated with absolute area under the shifted-performance curve, while a negative association remained only after normalization by each model's clean baseline. Operational survival contributed a minimum-utility interpretation but was less precise because eligibility and censoring reduced information. Severity remapping changed area magnitude more than model ordering. Among the two evaluated ACS tasks, thresholded-confidence estimators had the lowest observed natural-transfer error; the task count is too small to infer general superiority. Robustness reports should therefore separate absolute performance, operational failure, and relative degradation; preserve the dataset as the inferential unit; and include natural-shift evidence whenever deployment claims are made.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests: The author declares no competing interests.

Author contributions: Aras Masood Ismael completed the conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, original drafting, revision, and project administration.

Data and code availability: OpenML identifiers are reported in Table 1. Natural-shift data were generated from public ACS microdata through Folktables. The submission archive contains the executable workflow, condition-level outputs, derived tables, figures, logs, validation report, and checksums.

Clinical trial number: not applicable.

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable.

Ethics approval and consent to participate: Not applicable. This study did not involve human participants, human data requiring ethical approval, or animals. All analyses used publicly available, fully de-identified benchmark datasets (OpenML) and public-use United States Census Bureau microdata (ACS via Folktables). No recruitment, intervention, or collection of personally identifying information took place; therefore, ethics approval and consent to participate were not required.

Consent for publication: Not applicable. The manuscript does not contain data, images, or details relating to any individual person.

Acknowledgements: The author acknowledges the developers and maintainers of OpenML, Folktables, the original benchmark datasets, and the United States Census Bureau public-use data resources.

Declaration on generative artificial intelligence: During manuscript preparation, the author used a generative language system for language editing, structural suggestions, and document formatting. The system did not generate experimental observations or decide the reported analyses. The author inspected the source code, recomputed numerical statements from archived outputs, verified citations, revised the text, and accepts full responsibility for the manuscript.

References

- [1] Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression. Wiley, New York.
- [2] Bialek, J., Kivimaki, J.I., Kuberski, W., Perrakis, N., 2025. Estimating model performance under covariate shift without labels. *Advances in Neural Information Processing Systems* 38. <https://doi.org/10.52202/085713-5379>.
- [3] Blackard, J.A., Dean, D.J., 1999. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture* 24(3), 131-151. [https://doi.org/10.1016/S0168-1699\(99\)00046-0](https://doi.org/10.1016/S0168-1699(99)00046-0).
- [4] Bonneel, N., Rabin, J., Peyre, G., Pfister, H., 2015. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51(1), 22-45. <https://doi.org/10.1007/s10851-014-0506-3>.
- [5] Bouthillier, X., Delaunay, P., Bronzi, M., et al., 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems* 3.
- [6] Breiman, L., 2001. Random forests. *Machine Learning* 45, 5-32. <https://doi.org/10.1023/A:1010933404324>.
- [7] Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: *20th International Conference on Pattern Recognition*, pp. 3121-3124. <https://doi.org/10.1109/ICPR.2010.764>.
- [8] Cai, T., Namkoong, H., Yadlowsky, S., 2025. Diagnosing model performance under distribution shift. *Operations Research*. Advance online publication. <https://doi.org/10.1287/opre.2023.0217>.
- [9] Cauchois, M., Gupta, S., Ali, A., Duchi, J.C., 2024. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association* 119(548), 3033-3044. <https://doi.org/10.1080/01621459.2023.2298037>.
- [10] Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>.

- [11] Cheng, Z.J., Jia, Z., Zhou, Z., Li, Y.F., Guo, L.Z., 2025. TabFSBench: Tabular benchmark for feature shifts in open environments. *Proceedings of Machine Learning Research* 267, 10025-10089.
- [12] Cox, D.R., 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B* 20(2), 215-242. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>.
- [13] Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7, 1-30.
- [14] Ding, F., Hardt, M., Miller, J., Schmidt, L., 2021. Retiring Adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34, 6478-6490.
- [15] Efron, B., Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York. <https://doi.org/10.1201/9780429246593>.
- [16] Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>.
- [17] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A., 2014. A survey on concept drift adaptation. *ACM Computing Surveys* 46(4), Article 44. <https://doi.org/10.1145/2523813>.
- [18] Gardner, J., Popovic, Z., Schmidt, L., 2023. Benchmarking distribution shift in tabular data with TableShift. *Advances in Neural Information Processing Systems* 36. <https://doi.org/10.52202/075280-2324>.
- [19] Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H., 2022. Leveraging unlabeled data to predict out-of-distribution performance. *International Conference on Learning Representations*.
- [20] Garg, S., Erickson, N., Sharpnack, J., Smola, A., Balakrishnan, S., Lipton, Z.C., 2023. RLSbench: Domain adaptation under relaxed label shift. *Proceedings of Machine Learning Research* 202, 10879-10928.
- [21] Gretton, A., Borgwardt, K.M., Rasch, M.J., Scholkopf, B., Smola, A., 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13, 723-773.
- [22] Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L., 2021. Predicting with confidence on unseen distributions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1134-1144.
- [23] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. *Proceedings of Machine Learning Research* 70, 1321-1330.
- [24] Hendrycks, D., Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*.
- [25] Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65-70.
- [26] Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282), 457-481. <https://doi.org/10.1080/01621459.1958.10501452>.
- [27] Koh, P.W., Sagawa, S., Marklund, H., et al., 2021. WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of Machine Learning Research* 139, 5637-5664.
- [28] Kulinski, S., Inouye, D.I., 2023. Towards explaining distribution shifts. *Proceedings of Machine Learning Research* 202, 17931-17952.
- [29] Lipton, Z.C., Wang, Y.X., Smola, A., 2018. Detecting and correcting for label shift with black box predictors. *Proceedings of Machine Learning Research* 80, 3122-3130.
- [30] Mishra, A., Liu, A., 2025. ODD: Overlap-aware estimation of model performance under distribution shift. *Proceedings of Machine Learning Research* 286, 3031-3047.
- [31] Moreno-Torres, J.G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N.V., Herrera, F., 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45(1), 521-530. <https://doi.org/10.1016/j.patcog.2011.06.019>.

- [32] Moro, S., Cortez, P., Rita, P., 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62, 22-31. <https://doi.org/10.1016/j.dss.2014.03.001>.
- [33] Phipson, B., Smyth, G.K., 2010. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* 9(1), Article 39. <https://doi.org/10.2202/1544-6115.1585>.
- [34] Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D. (Eds.), 2009. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA. <https://doi.org/10.7551/mitpress/9780262170055.001.0001>.
- [35] Rabanser, S., Gunnemann, S., Lipton, Z.C., 2019. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems* 32.
- [36] Subbaswamy, A., Adams, R., Saria, S., 2021. Evaluating model robustness and stability to dataset shift. *Proceedings of Machine Learning Research* 130, 2611-2619.
- [37] Taori, R., Dave, A., Shankar, V., et al., 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems* 33, 18583-18599.
- [38] Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L., 2013. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter* 15(2), 49-60. <https://doi.org/10.1145/2641190.2641198>.
- [39] Yu, Y., Yang, Z., Wei, A., Ma, Y., Steinhardt, J., 2022. Predicting out-of-distribution error with the projection norm. *Proceedings of Machine Learning Research* 162, 25721-25746.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ModelSurvivalV5SupplementaryMaterial.docx](#)