Selecting Probability Calibrators for Tabular Classification

Haojie Li
Southwest University, Chongqing, 400000, China
HaojieLi.1125@gmail.com

ABSTRACT

Reliable probabilities matter in risk-sensitive tabular decision making. In credit scoring, fraud detection, medical triage, and review prioritization, confidence quality often matters as much as label accuracy. Yet post-hoc calibrators such as temperature scaling, sigmoid scaling, isotonic regression, Dirichlet calibration, and beta calibration do not behave uniformly across model families, class structures, calibration-set sizes, or evaluation protocols. Rather than searching for a universally best calibrator, this work treats tabular probability calibration as a selection problem. Under a strictly separated Train / Cal / Select / Test protocol, uncalibrated outputs and candidate calibrators are compared on 20 OpenML datasets, three baseline model families, and three random seeds. Selection on the held-out Select split uses a guarded policy, the Phase-8 Guarded Calibration Selector (GCS-8), with fallback conditions and isotonic-specific constraints. Across the multiseed benchmark, the frozen selector reduces mean test log-loss from 0.332157 to 0.298072; On held-out validation datasets, mean test log-loss reaches 0.189940. Gains persist under class imbalance and small calibration splits, while isotonic emerges as the least stable option in low-data calibration regimes and beta provides a useful binary-only extension. Overall, tabular probability calibration is better framed as choosing among candidate post-hoc calibrators under a strict protocol than as seeking a single globally best method.

Keywords: Probability Calibration, Tabular Classification, Post-Hoc Calibration, Model Selection, Openml.

1. INTRODUCTION

Reliable probabilities matter wherever classifier outputs drive downstream action. Credit approval, fraud screening, manual-review queues, and medical triage all depend not only on which label is predicted, but also on how much confidence that prediction deserves. Overconfident probabilities distort thresholds, misallocate intervention effort, and obscure operational risk even when top-1 accuracy remains high. In such settings, trustworthy uncertainty is not a cosmetic property. It is part of the decision mechanism itself.

The issue is especially sharp for tabular classification. Tabular data remain central in industrial machine learning, yet the modeling landscape differs markedly from imagedominated settings. Recent benchmark studies show that treebased methods remain unusually strong on many real-world tabular tasks, while deep tabular models improve steadily but do not dominate uniformly [1][2]. Probability calibration in this regime therefore cannot be reduced to the familiar story of neural networks plus temperature scaling [3]. Mixed feature types, uneven class structures, modest sample sizes, and heterogeneous model families all shape calibration behavior.

Post-hoc calibration methods already form a substantial toolbox. Early work introduced Platt scaling and isotonicstyle mappings from classifier scores to probabilities [4][5][6]. Subsequent studies showed that probability quality varies systematically across learner families, even when predictive accuracy is similar [7]. More recent work extended the toolbox with temperature scaling, beta calibration, and Dirichlet calibration [3] [8] [9]. A consistent message follows from this line of work: no single calibrator remains uniformly best across architectures, label spaces, data sizes, and evaluation protocols.

That instability becomes more problematic in tabular settings. Protocol choices matter. Strong baselines matter. Small calibration splits matter. Class imbalance matters. Under these conditions, a calibrator that looks favorable on one split may degrade on another, and aggregate metrics may conceal deterioration on minority-class probabilities. The practical question, then, is not simply how to fit a calibration map. A key challenge lies in choosing when to calibrate, which calibrator to trust, and when to keep the original scores instead.

This work treats tabular probability calibration as a selection problem under a strict evaluation protocol. Uncalibrated outputs and candidate post-hoc calibrators are compared under a fully separated Train / Cal / Select / Test design, with calibrator choice performed on a held-out Select split rather than absorbed into calibration or test evaluation. The candidate

pool covers temperature scaling, sigmoid scaling, isotonic regression, Dirichlet calibration, and beta calibration. On top of that pool, a guarded selector, GCS-8, combines empirical selection with fallback rules and structural constraints designed to limit overfitting in small or unstable regimes.

Three contributions frame the study. First, a strict four-way protocol that separates model fitting, calibrator fitting, calibrator selection, and final testing. Second, a guarded selection framework that treats uncalibrated and calibrated outputs as competing candidates under explicit fallback rules. Third, an imbalance-aware evaluation design that examines calibration quality beyond aggregate averages and into minority-class behavior.

2. RELATED WORK

Post-hoc calibration has long been framed as a mapping problem: fit a predictive model first, then learn a transformation from scores to probabilities on held-out data. Platt scaling remains the canonical parametric instance of this idea [4]. Zadrozny and Elkan extended the same logic to decision trees, naive Bayes models, and multiclass settings, helping establish post-hoc calibration as a practical add-on rather than a modelspecific redesign [5][6]. Niculescu-Mizil and Caruana further showed that probability quality differs substantially across learner families even when ranking quality is similar [7]. Two themes already appear in this classical literature: calibration cannot be discussed independently of the underlying model, and the choice of calibrator reflects a bias-variance tradeoff rather than a universally dominant rule.

Later work expanded both the calibrator family and the way calibration itself is assessed. Temperature scaling became a widely adopted baseline after Guo et al. showed that modern neural networks are often sharply miscalibrated despite strong predictive performance [3]. Beta calibration addressed limitations of logistic-style mappings in binary classification, especially when score distributions are skewed or the identity map should remain representable [8]. Dirichlet calibration extended the post-hoc toolbox to native multiclass probability transformations rather than reducing calibration to separate binary problems [9]. On the evaluation side, Kumar et al. argued that calibration can itself be modeled through trainable measures [10], while Nixon et al. showed that expected calibration error depends strongly on binning and aggregation choices [11]. Taken together, this literature suggests that richer calibrators alone do not resolve the problem. Conclusions remain sensitive to metric design, class structure, and protocol.

That sensitivity is especially relevant for tabular learning. Benchmark studies in this area have emphasized strong baselines, unified evaluation pipelines, and careful protocol control, precisely because conclusions can shift under seemingly minor experimental changes [1][2]. The same concern carries over to probability calibration. Existing representative studies mostly introduce, analyze, or compare individual calibrators [3][8][9]. Much less attention has gone to the practical question raised by heterogeneous tabular tasks: when model family, class imbalance, and calibration-set size vary, how should one choose among several plausible post-hoc calibrators, and when should calibration be skipped altogether? The present work addresses that gap by treating calibration not as the search for a single best mapping, but as a guarded selection problem under a strictly separated protocol.

3. PROBLEM SETTING AND METHOD

3.1 Problem Formulation

Consider a tabular classification task with a trained base classifier that outputs an uncalibrated probability vector $p(\mathbf{x})$ for input x. The objective here is not to redesign that classifier, but to improve the reliability of its probability output through a post-hoc map g. For any candidate calibrator, the calibrated prediction becomes $g(p(\mathbf{x}))$.

The central object is therefore not a single calibration map, but a candidate set

$$C = \{\text{uncalibrated}, g_1, g_2, \dots, g_m\} \tag{1}$$

where the original model output is kept as an explicit option. Given a fixed dataset-model pair, calibrator choice is treated as a held-out selection problem:

$$\mathrm{c}^* = \underset{\mathrm{c}\in\mathrm{C}}{\operatorname{argmin}} \mathrm{LogLoss}_{\mathrm{Select}}(\mathrm{c}) \tag{2}$$

The candidate pool in Eqs. (1) and (2) shifts the focus from fitting one mapping well to choosing among several plausible mappings under a strict protocol.

3.2 Four-Way Protocol and Candidate Pool

To prevent leakage between model fitting, calibrator fitting, and calibrator choice, every dataset follows a strictly separated Train / Cal / Select / Test split with fixed ratios $60\%/20\%/10\%/10\%$. The Train split is used only for fitting the base classifier. The Cal split is used only for fitting posthoc calibrators. The Select split is used only for choosing among candidates. The Test split is touched only once for final reporting.

The candidate pool combines uncalibrated outputs with five post-hoc alternatives: temperature scaling, sigmoid scaling, isotonic regression, Dirichlet calibration, and beta calibration [3][8][9]. Beta calibration is enabled only for binary tasks, reflecting its natural scope rather than forcing a multiclass extension. This pool spans the main post-hoc design families relevant to tabular classification: low-variance parametric mappings, a higher-variance nonparametric mapping, and a native multiclass transformation.

Candidate comparison is anchored on Select-set log-loss. Brier score and ECE are not used as separate optimization targets, but as risk-sensitive diagnostics to avoid selecting a candidate that improves one metric only by making probability quality worse elsewhere.

3.3 Guarded Selection

The final selector, GCS-8, is intentionally conservative. Selection begins with the full legal candidate set for a given dataset-model pair. A candidate is adopted only if it beats the uncalibrated baseline by a meaningful margin on the Select split and does not introduce obvious probability degradation.

Two fallback rules define this conservative behavior. First, if the Select-set log-loss gain over the uncalibrated output is smaller than 0.001, selection falls back to the original probabilities. Second, if Brier score worsens by more than 0.0005, the candidate is rejected even if its log-loss is slightly better. These rules treat calibration as optional rather than mandatory: when evidence for improvement is weak, the default action is to keep the original model output.

An additional structural guardrail is applied to isotonic regression, the candidate that proved most unstable in early selection experiments. Isotonic is disallowed for all nonrandom-forest models. Even for random forests, it remains eligible only on binary tasks with at least 2000 rows. The purpose is not to dismiss isotonic regression altogether, but to restrict it to settings where its flexibility is less likely to turn into high-variance overfitting.

The resulting procedure is simple. Fit all legal calibrators on the Cal split. Compare them with the uncalibrated baseline on the Select split. Apply the minimum-gain rule, the Brier-risk rule, and the isotonic guardrail. Freeze the surviving winner and evaluate it once on the Test split. Methodologically, the contribution lies less in proposing a new calibration map than in turning post-hoc calibration into a reproducible, constrained selection protocol for tabular classification.

4. EXPERIMENTAL SETUP

4.1 Datasets

All datasets are drawn from OpenML [12]. The frozen benchmark contains 20 tabular classification datasets: 10 binary and 10 multiclass. Sample sizes range from 1000 to 48842, covering both small-to-medium and medium-scale tabular problems. The pool includes 3 datasets with missing values, 11 with categorical features, and 12 imbalanced datasets whose minority ratio falls below 0.20. Fixed split manifests are reused across all baselines, calibrators, and selectors so that every method is evaluated on the same Train / Cal / Select / Test partitions.

4.2 Preprocessing and Baselines

All preprocessing is performed only after the four-way split, preventing any transformation from using full-dataset information. Two preprocessing pipelines are used, matched to model family. Logistic regression uses a linear pipeline: median imputation and standardization for numeric features, plus mode imputation and one-hot encoding for categorical features. Random forest and HistGradientBoosting use a tree pipeline: median imputation for numeric features and constant imputation with ordinal encoding for categorical features.

Three baseline model families are evaluated: logistic regression (LR), random forest (RF), and HistGradientBoosting (HGB). The purpose is coverage rather than novelty. LR represents a linear probabilistic baseline, RF a standard tree ensemble, and HGB a modern gradient-boosted tree implementation. All baselines are fitted on the Train split only, then

applied to the Cal, Select, and Test splits to produce the uncalibrated probabilities used by all downstream calibrators and selectors.

4.3 Repeated Seeds and Metrics

Core experiments are repeated across three random seeds: 20260316, 20260317, and 20260318. For each seed, the full pipeline is rerun in the same order: split generation, preprocessing, baseline training, calibrator fitting, and selector evaluation. Reported performance therefore reflects multiseed aggregation rather than a single lucky split.

Log-loss serves as the primary selection metric because it evaluates full predictive distributions and penalizes high-confidence mistakes sharply. Brier score and expected calibration error (ECE) are reported as complementary diagnostics of probability quality and reliability. Since ECE is known to depend on binning and aggregation choices [11], it is not treated as the sole decision criterion. For binary tasks, three additional minority-class diagnostics are recorded: minority-class test log-loss, minority-class test Brier score, and the gap between the mean predicted minority probability and the empirical minority-class rate. These support the imbalance-aware analysis reported later.

5. RESULTS AND ANALYSIS

5.1 Overall Multiseed Performance

Table 1 reports the frozen multiseed aggregate over 180 dataset-model-seed runs. GCS-8 improves all three-headline metrics over the uncalibrated baseline. Figure 1 visualizes the same overall selector ranking across the three-seed benchmark. Mean test log-loss drops from 0.332157 to 0.298072, mean test Brier score from 0.056893 to 0.054333, and mean test ECE from 0.054209 to 0.033324. Relative to the oracle over the same candidate pool, GCS-8 leaves a mean regret of 0.009438, recovering most of the available headroom.

Table 1. Overall multiseed comparison.

Selector	Log-loss	Brier	ECE	Regret
Baseline	0.332157	0.056893	0.054209	0.043523
GCS-8	0.298072	0.054333	0.033324	0.009438
Oracle	0.288634	0.053751	0.029763	0.000000

Minority-class diagnostics improve as well. On binary tasks, GCS-8 reduces mean minority test log-loss from 1.381740 to 1.176205 and mean minority test Brier score from 0.381818 to 0.361518. The mean minority probability gap also moves closer to zero, from-0.002790 to-0.001145. The gain is therefore not confined to aggregate averages. Probability quality improves where miscalibration is often most operationally costly.

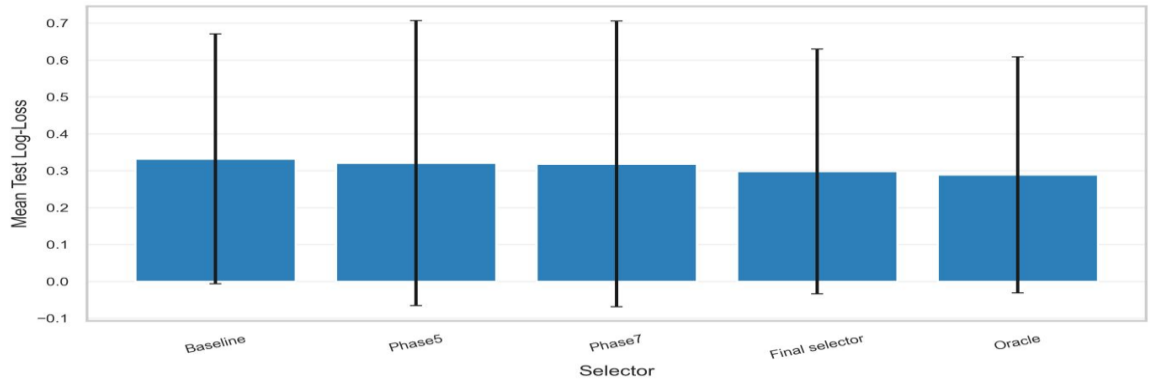

Figure 1. Overall selector comparison under three random seeds (Here, Final selector refers to GCS-8).

5.2 Validation-Set Freezing of the Final Policy

Pooled aggregates alone are not enough. A selector can still overfit dataset-level quirks. To test whether the rule survives beyond development data, the policy is frozen and evaluated on held-out validation datasets. Table 2 shows a clear progression: mean validation test log-loss decreases from 0.209899 for the baseline to 0.199736 for Phase5, 0.195150 for Phase7, and 0.189940 for GCS-8.

Table 2. Validation-set comparison after policy freezing.

Selector	Log-loss	Gain	Regret
Baseline	0.209899	0.000000	0.024957
Phase5	0.199736	0.010163	0.014794
Phase7	0.195150	0.014749	0.010208
rf_binary_rows_8000	0.190113	0.019786	0.005171
GCS-8	0.189940	0.019959	0.004998
Oracle	0.184942	0.024957	0.000000

These matters because it turns the guardrail from an exploratory heuristic into a frozen policy with held-out support. Figure 2 visualizes the same validation-set comparison and shows GCS-8 remaining slightly ahead of the development-derived alternative rule. The remaining gap to oracle is 0.004998 on validation datasets, much smaller than the baseline-oracle gap of 0.024957. Some room for improvement remains, but the main selection logic generalizes beyond the data used to shape it.

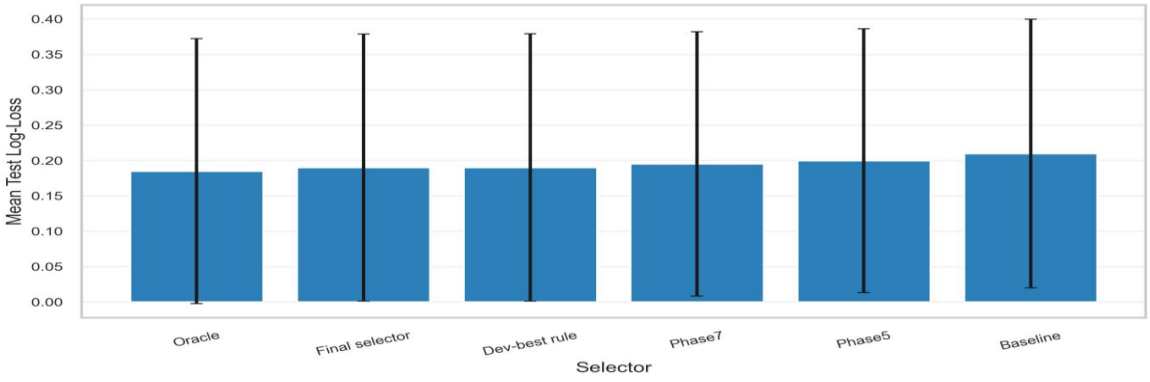

Figure 2. Selector comparison on held-out validation datasets.

5.3 Imbalance-Aware Gains Across Minority-Ratio Groups

The next question is whether overall gains hide failures on imbalanced tasks. Figure 3 suggests the opposite.

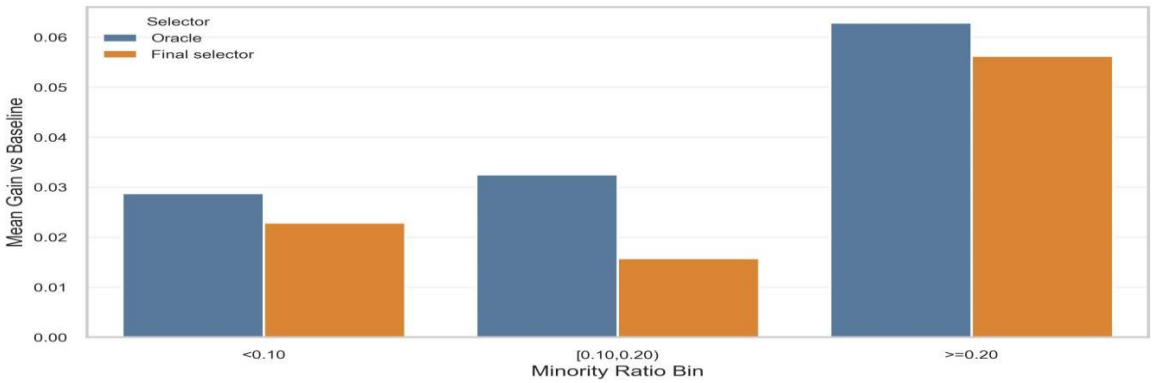

Figure 3. Grouped log-loss gain by minority-ratio bin.

GCS-8 improves mean test log-loss in all three minority-ratio bins: from 0.145242 to 0.122350 for minority ratio < 0.10, from 0.441884 to 0.426098 for [0.10, 0.20), and from 0.390047 to 0.333844 for ≥ 0.20.

The minority-specific metrics reinforce that pattern. In the most imbalanced group, mean minority test log-loss decreases from 1.921842 to 1.539643. The practical interpretation is clear: GCS-8 does not gain merely by improving the majority class or smoothing easy cases. Its strongest gains appear where probability distortion is usually hardest to control.

5.4 Why the Guardrail Is Necessary

The final question is not whether GCS-8 looks better on average, but why the guardrail is needed at all. Table 3 answers that question by quantifying calibrator instability across calibration-set minority-count bins.

Table 3. Calibrator instability by minority-class size in the calibration split.

Cal bin	Candidate	Gain	Regret
< 50	isotonic	-0.034384	0.086804
< 50	temperature	0.040253	0.012166
< 50	sigmoid	0.041144	0.011275
< 50	dirichlet	0.035526	0.016893
< 50	beta	0.032097	0.004221
[50, 100)	isotonic	-0.138940	0.188309
[50, 100)	temperature	0.035968	0.013401
[50, 100)	sigmoid	0.039678	0.009691
[50, 100)	dirichlet	0.038392	0.010978
[50, 100)	beta	0.019936	0.006401
⩾ 100	isotonic	-0.022291	0.052778
⩾ 100	temperature	0.022251	0.008235
⩾ 100	sigmoid	0.022512	0.007974
⩾ 100	dirichlet	0.026815	0.003671
⩾ 100	beta	0.009517	0.002914

Isotonic regression is the least stable calibrator across all calibration-set minority-count bins. When the calibration split contains fewer than 50 minority examples, isotonic shows a mean gain of-0.034384 and a mean regret of 0.086804. In the [50, 100) range, instability becomes much larger, with mean gain-0.138940 and mean regret 0.188309. Even when the calibration split contains at least 100 minority examples, isotonic remains less stable than the parametric alternatives.

The contrast with the other candidates is sharp. Temperature scaling, sigmoid scaling, and Dirichlet calibration remain positive across all bins. Beta calibration, while binary-only, also remains stable and positive where it is applicable. The role of GCS-8 is therefore not to rely on a single dominant calibrator, but to choose conservatively among candidates whose stability changes across data regimes.

Taken together, the results support three claims. No single post-hoc calibrator remains best across tabular tasks. A strict four-way protocol plus validation-set freezing is necessary to distinguish attractive split-specific rules from genuinely transferable ones. Finally, class imbalance and calibration-set size must enter the selection logic explicitly; otherwise, mean improvements can be erased by a small number of high-regret failures.

6. THREATS TO VALIDITY

The first threat concerns dataset coverage. Although the benchmark spans 20 OpenML datasets, includes both binary and multiclass tasks, and explicitly covers missing values, categorical features, and class imbalance, it still represents static supervised tabular classification under moderate scale. The conclusions should not be read as automatically extending to extreme-scale industrial data, strongly temporal settings, heavy concept drift, or highly sparse feature spaces.

The second threat concerns the boundaries of the candidate pool. Three baseline families are included: logistic regression, random forest, and HistGradientBoosting. Together they cover linear models, standard tree ensembles, and modern gradientboosted trees, but they do not span the full space of competitive tabular learners. The same qualification applies to the calibration pool. Temperature scaling, sigmoid scaling, isotonic regression, Dirichlet calibration, and beta calibration provide a representative set of post-hoc options, yet methods such as Venn-Abers and ENIR remain outside the present study. The current evidence therefore supports the necessity of selector design within a strong and representative candidate pool, not a final statement over all possible calibrators.

The third threat is methodological. The final policy, GCS-8, is frozen through dataset-level validation, but its isotonic guardrail remains an empirical rule distilled from failure analysis rather than a theorem. The claim is therefore deliberately limited: the guardrail is stable and effective under the present protocol, not universally optimal. Future expansion of the dataset pool and candidate set may refine, relax, or partially replace that rule.

The fourth threat lies in the metrics themselves. Log-loss is used as the primary criterion because it evaluates full predictive distributions and penalizes high-confidence mistakes sharply. Brier score and ECE provide complementary information, but ECE is sensitive to binning and aggregation choices [11], and minority-class diagnostics can fluctuate when rare test examples are scarce. To reduce dependence on any single summary, the analysis combines headline metrics, minority-specific diagnostics, grouped comparisons, and heldout validation-set freezing. Even so, no finite benchmark removes metric-induced uncertainty entirely.

7. CONCLUSION

The central question in this work is not which post-hoc calibrator wins in the abstract, but how a tabular classification pipeline should choose among plausible calibrators under a strict protocol and fall back safely when evidence for improvement is weak. Framed this way, probability calibration becomes a selection problem rather than a single-method ranking exercise.

The empirical results support that shift in perspective. Posthoc calibration is useful on tabular tasks, but not because one method dominates across datasets, model families, and class structures. The gains come from treating the uncalibrated output as a valid candidate, enforcing separation between fitting, calibration, selection, and testing, and adding conservative guardrails where candidate instability is repeatedly observed.

Three conclusions follow. First, a strict four-way evaluation protocol makes it possible to separate model fitting, calibrator fitting, calibrator choice, and final testing without leakage. Second, a guarded selector turns post-hoc calibration into a constrained model-selection problem with fallback behavior rather than a forced post-processing step. Third, an imbalance-aware evaluation shows that the frozen selector, GCS-8, reduces mean test log-loss from 0.332157 to 0.298072 in the

multiseed benchmark, reaches 0.189940 on held-out validation datasets, and preserves gains under class imbalance while exposing isotonic regression as the least stable option under small calibration data.

Several extensions remain open. The current candidate pool can be broadened with methods such as Venn-Abers and ENIR, and stronger tabular deep baselines may further test the robustness of the selection logic. More broadly, the current isotonic guardrail should be understood as an empirically effective rule, not a universal theorem. Even within those limits, the main conclusion is clear: in tabular classification, reliable calibration depends as much on reliable selection as on the calibrators themselves.

REFERENCES

[1] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting Deep Learning Models for Tabular Data," in Advances in Neural Information Processing Systems, vol. 34, 2021.

[2] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?", in Advances in Neural Information Processing Systems, vol. 35, 2022.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proceedings of the 34th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 1321-1330.

[4] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in Advances in Large Margin Classifiers, A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. MIT Press, 1999, pp. 61-74.

[5] B. Zadrozny and C. Elkan, "Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers," in Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 609-616.

[6] B. Zadrozny and C. Elkan, "Transforming Classifier Scores into Accurate Multiclass Probability Estimates," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 694-699.

[7] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 625-632.

[8] M. Kull, T. S. Filho, and P. Flach, "Beta Calibration: A Well-Founded and Easily Implemented Improvement on Logistic Calibration for Binary Classifiers," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, vol. 54, 2017, pp. 623-631.

[9] M. Kull, M. Perello-Nieto, M. Kaengsepp, T. S. Filho, H. Song, and P. Flach, "Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration," in Advances in Neural Information Processing Systems, vol. 32, 2019.

[10] A. Kumar, S. Sarawagi, and U. Jain, "Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings," in Proceedings of the 35th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 2805-2814.

[11] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring Calibration in Deep Learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 38-41.

[12] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "OpenML: Networked Science in Machine Learning," in Proceedings of the 4th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, ser. Proceedings of Machine Learning Research, vol. 41, 2015, pp. 49-60.