Ensembling Tabular Foundation Models: A Diversity Ceiling and a Calibration Trap

Anonymous Authors[1]

Abstract

Tabular foundation models (TFMs) now match or beat tuned gradient-boosted trees on a growing fraction of tabular tasks, but no single TFM wins on every dataset. Ensembling is the textbook fix, and it works less well than expected. Six modern TFMs form a near-redundant pool: their mean pairwise Q-statistic is 0.961, close enough to 1 that any convex combination is bounded above. We benchmark six ensemble strategies over six TFMs on 153 OpenML classification tasks. The best ensemble, two-level cascade stacking, buys $+0.18\%$ accuracy over the strongest single TFM at $253\times$ the compute. A Friedman–Nemenyi analysis places three ensembles and the best base TFM in a single equivalence group; three other ensembles are significantly *worse* than the best base. Stacking with a logistic-regression meta-learner is the most striking case: competitive accuracy and ROC-AUC, the worst log-loss rank among the ensembles. The meta-learner improves accuracy by sharpening class boundaries, which destroys calibration. We recommend greedy selection as the practical default.

1. Introduction

Tabular foundation models (TFMs) have advanced rapidly, with growing architectural diversity. We show that this diversity is mostly nominal: a pool of six modern TFMs produces near-redundant predictions, and ensembling cannot exploit diversity that is not there. TabPFN (Hollmann et al., 2023; 2025), TabICL (Qu et al., 2025), and variants like Mitra (Zhang et al., 2025a), TabDPT (Ma et al., 2024), and CARTE (Kim et al., 2024) all perform in-context learning (ICL) over synthetic or curated priors. Continued pre-training on real-world tables (Garg et al., 2025) now rivals heavy AutoML in a single forward pass: TabPFNv2.5 matches AutoGluon 1.4 on TabArena (Erickson et al., 2025). What these models do not do, however, is win uniformly. Across our 153-dataset benchmark, the per-dataset accuracy wins among six TFMs split as follows: TabPFNv2.5 52, TabICLv2 46, TabPFNv2.6 25, TabICL 13, LimiX 12, OrionMSPv1.5 5. The median accuracy spread between best and worst TFM on a single dataset is 1.95%; on 24% of datasets it exceeds 5%, so committing to one TFM in advance loses meaningful accuracy on roughly a quarter of the benchmark. Ensembling is the textbook response (Breiman, 1996; Dietterich, 2000; Caruana et al., 2004).

The cost-benefit story for TFMs differs from the gradient-boosted decision tree (GBDT) setting where ensemble methods originated. Per-task TFM inference is cheap once the model is pretrained (under one second on most benchmark datasets), though pretraining itself is a substantial fixed cost. Cascade-style stacking layers K-fold OOF inference across all bases, dwarfing any single forward pass. Beyond compute, there is a structural concern. TFMs trained with ICL on synthetic priors approximate Bayesian model averaging at inference (Müller et al., 2022; Zhang et al., 2023); broadly similar priors yield broadly similar posteriors, leaving little for a convex combiner to recover.

Position relative to existing work. Recent work ensembles TFMs differently. The post-hoc ensemble protocol bundled with TabPFN (Hollmann et al., 2025) ensembles hyperparameter configurations of *one* TFM; TabICLv2 (Qu et al., 2025) does the same internally with column and class shuffles. TabArena (Erickson et al., 2025) measures post-hoc ensembling across *heterogeneous* model classes (TFMs, GBDTs, neural baselines) and reports that some classes are overrepresented in cross-class ensembles due to validation overfitting. HAPEns (Maier & Purucker, 2026) adds hardware-aware multi-objective selection on a general model pool. TabM (Gorishniy et al., 2025) pursues parameter-efficient ensembling at the architecture level. Recent architectural work in the same TFM family includes Orion-BiX (Bouadi et al., 2025a), while Tanna et al. (2026) study fine-tuning protocols (zero-shot, meta-learning, SFT, PEFT) across the same CC18/TALENT/TabZilla benchmarks we use here. We isolate a different question: holding the model *class*

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

fixed, what does ensembling six different TFMs buy on its own?

Contributions.

1. A diversity-ceiling diagnostic. Six modern TFMs share an ICL-on-synthetic-priors recipe and produce near-redundant predictions ($Q = 0.961$), bounding the gain available to any convex combiner.
2. An empirical accuracy/compute frontier with a calibration overlay across six ensemble strategies on 153 OpenML tasks. The best ensemble buys a sub-percent accuracy gain at $253\times$ the compute of the strongest base; three of six ensembles are significantly worse than the best base.
3. Calibration findings. Stacking with a logistic-regression meta-learner has competitive accuracy and ROC-AUC ranks but the worst log-loss rank among the ensembles; the meta-learner sharpens class boundaries, improving accuracy at the cost of probability quality.

2. Ensemble Strategies

Six strategies share a common `fit`/`predict`/`predict_proba` interface over a fixed pool of K base TFMs producing class-probability vectors $p_k(x)$.

Weighted Averaging (WA). $\hat{p} = \sum_k w_k p_k$ with $w_k \propto \text{score}_k$ on validation. No second stage. Cheapest combiner.

Greedy Selection (Caruana et al., 2004). Forward selection with replacement: at each of S iterations, the base whose addition maximises validation accuracy is added. Final weight equals selection count over S. We use $S = 50$, matching the AutoGluon (Erickson et al., 2020) `WeightedEnsembleModel` default.

Stacking (Wolpert, 1992; Ting & Witten, 1999). Bases produce 5-fold out-of-fold (OOF) predictions. A logistic-regression meta-learner is trained on the OOF features.

Temperature-Scaled Blending (Guo et al., 2017). Per-base temperature T_k is fit on the validation set by minimising negative log-likelihood (NLL) of $\text{softmax}(\log p_k/T_k)$; calibrated probabilities are then averaged uniformly.

Cascade Stacking. Two-level stacking with skip connections, modifying AutoGluon's high-quality preset (Erickson et al., 2020). Level-1 OOF predictions concatenate with raw features and feed level-2 base models, also with K-fold OOF. A final greedy-selection layer combines all level outputs. We use 2 levels, 3-fold OOF, $S = 50$.

Random-Init (Deep) Ensemble (Lakshminarayanan et al., 2017). Each TFM is run with $M = 3$ different seeds. Per-base predictions are averaged across seeds, then cross-base averaging uses performance weights.

Strategies are implemented in Python on top of the public TabTune library (Tanna et al., 2025); code, configurations, and per-dataset logs are released with the paper.

3. Experimental Setup

Datasets. 153 OpenML classification tasks drawn from the CC18 (Bischl et al., 2021), TALENT (Liu et al., 2025), and TabZilla (McElfresh et al., 2023) pools. Selection criteria and the full dataset inventory are in Appendix H.

Base TFMs. Six models in inference mode: TabPFNv2.5 (Grinsztajn et al., 2025), TabPFNv2.6 (Hollmann et al., 2025), TabICL (Qu et al., 2025), TabICLv2 (Qu et al., 2025), LimiX (Zhang et al., 2025b), OrionMSPv1.5 (Bouadi et al., 2025b). We restricted the pool to publicly released checkpoints.

Protocol. Per dataset: an $80/20$ stratified train/test split; within train, a $75/25$ train/validation split for ensemble weight learning. Stacking and cascade levels use 5-fold and 3-fold internal CV respectively. A fixed seed controls splits and base-model initialisation.

Metrics. Accuracy, weighted F1, one-vs-rest ROC-AUC, multi-class log-loss, and total fit-time per dataset (seconds). For deeper analysis on the TabArena classification suite we additionally report expected calibration error (ECE) (Guo et al., 2017), the reliability component of the Brier decomposition (Brier, 1950), area under the risk–coverage curve (AURC) (Geifman & El-Yaniv, 2017), coverage at 95% accuracy, and worst-group accuracy (WGA) (Sagawa et al., 2020). Statistical significance is reported via Friedman (Friedman, 1940), Nemenyi (Nemenyi, 1963), and pairwise Wilcoxon signed-rank (Wilcoxon, 1945).

Hardware. A single H100 (80 GB) GPU per run.

4. Results

4.1. Aggregate performance

Table 1 reports per-method statistics across the 153 datasets in our benchmark. The accuracy spread among the top eight methods is 0.45 percentage points (TabICL at 0.872 to Cascade at 0.882). The Friedman test rejects equality of mean ranks across the 12 methods ($\chi^2 = 389.95$, $p < 10^{-30}$); methods are not exchangeable, but the question is which differences survive a per-pair correction. Calibration, selective-prediction, and group-robustness metrics on the TabArena suite are reported in Table 3 (Appendix C) and analysed in §4.3.

The Nemenyi critical difference at $\alpha = 0.05$ for $K = 12, N = 153$ is CD $= 1.347$. Three methods sit within CD of the top-ranked Cascade_2level: Stacking_LR ($\Delta = 0.48$), Greedy_Selection ($\Delta = 0.80$), and TabICLv2 ($\Delta = 0.90$).

Table 1. Mean performance over 153 OpenML classification tasks. Lower rank is better. Methods are sorted by accuracy rank; ensemble methods are italicized. Best in each rank column is bold.

Method	Acc.	Acc. rank	Log-loss	Log-loss rank	ROC-AUC	ROC-AUC rank	Fit (s)
Cascade_2level	0.882	**4.48**	0.289	4.55	0.907	5.01	178.5
Stacking_LR	0.880	4.96	0.288	8.13	0.907	**4.59**	6.6
Greedy_Selection	0.881	5.28	0.278	4.54	0.906	5.51	6.7
TabICLv2	0.881	5.39	0.274	**4.11**	0.907	4.73	**0.7**
DeepEnsemble_3seed	0.879	6.10	0.281	5.62	0.907	5.58	75.7
WA_performance	0.879	6.18	0.282	5.50	0.907	5.49	6.6
Temp_Scaled	0.879	6.41	0.283	6.47	0.907	6.14	6.6
TabPFNv2.5	0.878	6.44	0.286	6.64	0.904	7.69	1.2
TabPFNv2.6	0.876	6.78	0.283	6.26	0.903	6.89	1.6
LimiX	0.875	7.67	0.288	7.17	0.902	7.64	0.5
TabICL	0.872	7.96	0.289	7.61	0.900	8.02	1.0
OrionMSPv1.5	0.848	10.34	0.356	11.41	0.880	10.72	1.8

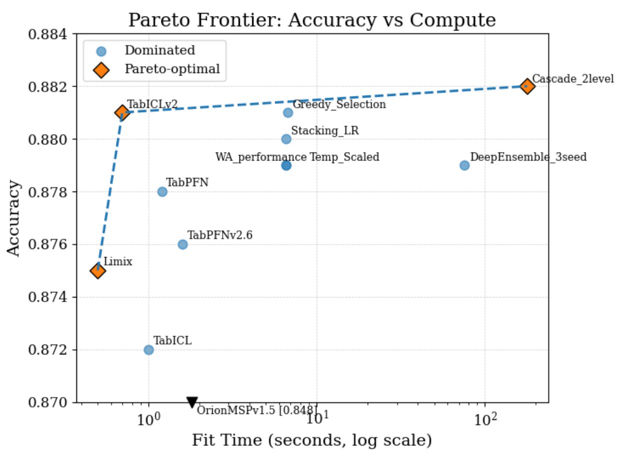

Figure 1. Mean accuracy vs. mean fit time (log scale, seconds per dataset) over 153 OpenML tasks. Pareto-optimal methods are marked. Cascade buys an accuracy advantage over TabICLv2 for 253× the compute.

Three ensembles and one base TFM are statistically indistinguishable on accuracy across 153 tasks; the remaining three ensembles cannot beat the best base. Pairwise Wilcoxon tests sharpen the picture: against TabICLv2, only Cascade_2level wins (+0.18%, $p = 0.008$); Greedy_Selection (+0.01%) and Stacking_LR (−0.03%) tie; WA, Temp_Scaled, and DeepEnsemble_3seed are all significantly worse ($p < 0.05$). One ensemble of six clears the bar of beating the strongest base.

4.2. Accuracy/compute frontier

Fit times span four orders of magnitude. TabICLv2 averages 0.71 s per dataset; Greedy, Stacking_LR, WA, and Temp_Scaled all sit near 6.6 s, which is roughly the cost of one forward pass through the six bases plus a thin combination layer. DeepEnsemble_3seed costs 75.7 s, and Cascade_2level costs 178.5 s. Figure 1 plots the trade-off.

The Pareto frontier is dominated by TabICLv2 (cheapest competitive option) and Greedy_Selection (best accuracy at moderate cost). Cascade_2level sits on the frontier, but its marginal accuracy advantage over TabICLv2 corresponds to a 253× compute multiplier. DeepEnsemble_3seed is dominated outright: WA_performance and Stacking_LR achieve similar or better accuracy at one tenth its cost. Figure 2 (Appendix B) shows the corresponding critical-difference diagram.

4.3. Calibration and the diversity ceiling

Log-loss diverges from accuracy. The log-loss column of Table 1 tells a different story than the accuracy column. TabICLv2 has the lowest log-loss rank (4.11); Greedy_Selection (4.54) and Cascade_2level (4.55) sit close behind, both producing convex combinations of probability vectors. Stacking_LR ranks 8.13, the worst of any method tested except OrionMSPv1.5. Linear stacking still places the right class label, which is why its accuracy and ROC-AUC ranks stay competitive, but the cross-entropy objective on OOF predictions pushes the meta-learner toward sharper probability outputs than the bases produce, which degrades calibration.

Calibration tracks combination strategy, not compute. Table 3 (Appendix C) reports five complementary metrics on the TabArena classification suite: ECE, Brier reliability, AURC, coverage at 95% accuracy, and worst-group accuracy. TabICLv2 sets the calibration ceiling (ECE = 0.0236, Brier-REL = 0.0024). Greedy_Selection is the only ensemble that approaches it (ECE = 0.0253), and it never optimises for calibration directly. Stacking_LR records the worst calibration of any ensemble (ECE = 0.0272, Brier-REL = 0.0031), consistent with its log-loss rank. Temperature-Scaled Blending is equally instructive: per-base NLL minimisation gives an ECE of 0.0273, no better than Stacking_LR's.

Base-model diversity caps uncertainty quality. The mean pairwise Q-statistic (Kuncheva & Whitaker, 2003) across the six TFMs is 0.961 ($\sigma = 0.183$, Cohen's $\kappa = 0.856$, in the "almost perfect agreement" band of the conventional Landis–Koch scale). Q values close to 1 signal near-redundancy: the six bases share the ICL-on-synthetic-priors recipe and tend to fail on the same instances, so any convex combiner has little variance to absorb; Appendix D states this ceiling formally as a consensus-set bound on the ensemble-vs-base accuracy gap. The ceiling is most consequential for DeepEnsemble_3seed: three random seeds perturb context order but share the synthetic prior, producing an AURC of 0.0617 (28% above TabICLv2's 0.0483) and coverage at 95% accuracy of 62.1% versus 68.4%. The 75.7 s cost buys neither accuracy nor uncertainty improvement; for selective prediction, Greedy_Selection (AURC = 0.0484) is the appropriate choice.

Cascade earns its cost on group robustness. Worst-group accuracy is the one axis where heavy stacking earns its overhead. Cascade_2level reaches 0.803, on par with TabICLv2 (0.802) and outperforming all other ensembles by roughly three points (Greedy and Stacking_LR both at 0.776). The skip-connection architecture appears to implicitly downweight base models that are systematically biased on minority subgroups; simpler convex combiners do not replicate this. The fairness margin is narrow and confined to sensitive-attribute datasets, but it is the one regime in which cascade's 253× compute overhead translates into a qualitative advantage rather than a marginal one.

Per-dataset patterns. Aggregate means hide per-dataset behaviour. Cascade_2level beats the per-task oracle base on 42 of 153 tasks (ties on 36, loses on 75, mean $\Delta = -0.0025$); against TabICLv2 specifically, 70/41/42 (mean $\Delta = +0.0018$). Ensembles win on average not because they are uniformly better, but because no fixed base is best on every dataset. The Pearson correlation between ensemble gain and inter-base accuracy spread is -0.03, so the ceiling is roughly uniform across dataset shapes; ensembling helps no more on high-spread datasets than on low-spread ones.

5. Discussion

The shape of the result holds across every metric we report. TFMs trained with ICL on synthetic priors already approximate Bayesian model averaging at inference, so explicit downstream ensembling lands inside the noise floor of a strong single base. Cascade stacking buys the last 0.2% of accuracy by letting the meta-learner combine raw features alongside OOF predictions, but the cost-benefit ratio is poor outside competition settings. Greedy selection is the more honest default: roughly 10× the cost of the strongest single model, the same mean accuracy as the heaviest stack, and no calibration regression.

Two patterns matter for downstream work. First, calibration is not a free byproduct of accuracy ensembling. Stacking with logistic regression damages probability quality despite improving accuracy rank, and uniform averaging after per-base temperature scaling is no better than the best base. Calibration-aware meta-learners optimised for log-loss directly, or post-hoc recalibration applied to the ensemble output rather than to its members, both remain open. Second, the per-dataset variance in best-base identity is what ensembling is really being asked to solve. A small gating learner trained on dataset metafeatures to pick the best TFM per dataset might match cascade at far lower compute, and is a better fit to the structure of the problem than a stack.

Limitations. We use a single seed per dataset; statistical power comes from across-dataset variation rather than within-dataset replicates, which is what paired Wilcoxon and Friedman–Nemenyi tests assume on $N = 153$ tasks. A small number of the largest tasks were dropped for individual base TFMs due to memory constraints, and Table 1 reports the intersection of tasks that completed for every method. We do not include GBDT baselines; the comparison is between single TFMs and TFM-only ensembles, leaving open whether out-of-class diversity (TFM + GBDT) recovers gains the within-class pool cannot.

Contamination. Several base TFMs were pretrained on data overlapping with OpenML, a concern raised by the TabArena protocol (Erickson et al., 2025) and the workshop call. Reported deltas should be read as upper bounds on within-pool ensemble effects: under cleaner contamination protocols, the ensemble-vs-best-base gap is likely smaller, not larger. Contamination does not change the qualitative finding (a near-redundant pool produces a hard ceiling), but the precise accuracy delta is best treated as inflated.

Future work. Hybrid TFM+GBDT pools, per-dataset gating learners, time-series TFMs where base-model spread may be larger, and calibration-aware meta-learners.

6. Conclusion

Six tabular foundation models trained with ICL on synthetic priors form a near-redundant pool. A Q-statistic of 0.961 caps what any convex combiner can recover, and the empirical results follow: a top equivalence group of four methods (three ensembles plus the best base) statistically indistinguishable on accuracy; a best ensemble that recovers a sub-percent accuracy gain at 253× compute; and a calibration trap when a meta-learner is asked to manufacture extra accuracy by sharpening probabilities. Greedy selection is the practical default; cascade stacking is justifiable only when worst-group accuracy is a primary target. **The open question is whether out-of-class diversity (TFM + GBDT) breaks through the ceiling that within-class pools cannot.**

References

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Mantovani, R. G., van Rijn, J. N., and Vanschoren, J. OpenML benchmarking suites. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.

Bouadi, M., Seth, P., Tanna, A., and Sankarapu, V. K. Orion-Bix: Bi-axial attention for tabular in-context learning. *arXiv preprint arXiv:2512.00181*, 2025a.

Bouadi, M., Seth, P., Tanna, A., and Sankarapu, V. K. Orion-MSP: Multi-scale sparse attention for tabular in-context learning. *arXiv preprint arXiv:2511.02818*, 2025b.

Breiman, L. Bagging predictors. *Machine Learning*, 24: 123–140, 1996.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *International Conference on Machine Learning (ICML)*, 2004.

Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. AutoGluon-Tabular: Robust and accurate AutoML for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Mutalik Desai, P., Salinas, D., and Hutter, F. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=jZqCqpCLdU.

Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

Garg, A., Ali, M., Hollmann, N., Purucker, L., Müller, S., and Hutter, F. Real-TabPFN: Improving tabular foundation models via continued pre-training with real-world data. *arXiv preprint arXiv:2507.03971*, 2025.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Gorishniy, Y., Kotelnikov, A., and Babenko, A. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations (ICLR)*, 2025.

Grinsztajn, L., Floge, K., Key, O., Birkel, F., Jund, P., Roof, B., Jager, B., Safaric, D., Alessi, S., Hayler, A., Manium, M., Yu, R., Jablonski, F., Hoo, S. B., Garg, A., Robertson, J., Buhler, M., Moroshan, V., Purucker, L., Cornu, C., Wehrhahn, L. C., Bonetto, A., Scholkopf, B., Gambhir, S., Hollmann, N., and Hutter, F. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *ArXiv*, abs/2511.08667, 2025. URL https://api.semanticscholar.org/CorpusID:282939803.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

Kim, M. J., Grinsztajn, L., and Varoquaux, G. Carte: Pretraining and transfer for tabular learning, 2024. URL https://arxiv.org/abs/2402.16785.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Liu, S.-Y., Cai, H.-R., Zhou, Q.-L., Yin, H.-H., Zhou, T., Jiang, J.-P., and Ye, H.-J. Talent: A tabular analytics and learning toolbox. *Journal of Machine Learning Research*, 26(226):1–16, 2025. URL http://jmlr.org/papers/v26/25-0512.html.

Ma, J., Thomas, V., Hosseinzadeh, R., Kamkari, H., Labach, A., Cresswell, J. C., Golestan, K., Yu, G., Volkovs, M., and Caterini, A. L. TabDPT: Scaling tabular foundation models. *arXiv preprint arXiv:2410.18164*, 2024.

Maier, J. and Purucker, L. HAPEns: Hardware-aware post-hoc ensembling for tabular data. *arXiv preprint arXiv:2603.10582*, 2026.

McElfresh, D., Khandagale, S., Valverde, J., C, V. P., Feuer, B., Hegde, C., Ramakrishnan, G., Goldblum, M., and White, C. When do neural nets outperform boosted trees on tabular data?, 2023. URL https://arxiv.org/abs/2305.02997.

Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. Transformers can do bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Nemenyi, P. *Distribution-Free Multiple Comparisons*. PhD thesis, Princeton University, 1963.

Qu, J., Holzmüller, D., Varoquaux, G., and Le Morvan, M. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning (ICML)*, 2025.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Tanna, A., Seth, P., Bouadi, M., Avaiya, U., and Sankarapu, V. K. TabTune: A unified library for inference and fine-tuning tabular foundation models. *arXiv preprint arXiv:2511.02802*, 2025.

Tanna, A., Seth, P., Bouadi, M., and Sankarapu, V. K. Exploring fine-tuning for tabular foundation models. *arXiv preprint arXiv:2601.09654*, 2026.

Ting, K. M. and Witten, I. H. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10: 271–289, 1999.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Wolpert, D. H. Stacked generalization. *Neural Networks*, 5 (2):241–259, 1992.

Zhang, X., Maddix, D. C., Yin, J., Erickson, N., Ansari, A. F., Han, B., Zhang, S., Akoglu, L., Faloutsos, C., Mahoney, M. W., Hu, C., Rangwala, H., Karypis, G., and Wang, B. Mitra: Mixed synthetic priors for enhancing tabular foundation models. *arXiv preprint arXiv:2510.21204*, 2025a.

Zhang, X. et al. LimiX: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025b.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization, 2023. URL https://arxiv.org/abs/2305.19420.

A. Method-name glossary

The body and Tables 1–3 use compact short-form labels; Figures 3 and 4 render the same methods in long form. Table 2 reconciles the two.

Table 2. Short-form labels used in prose and Tables 1–3 mapped to the long-form labels rendered in Figures 3 and 4.

Short form (prose, tables)	Long form (figures)
`Cascade_2level`	Cascade 2-Level Stacking
`Stacking_LR`	Logistic Regression Stacking
`Greedy_Selection`	Greedy Selection
`WA_performance`	Weighted Averaging
`Temp_Scaled`	Temperature Scaling
`DeepEnsemble_3seed`	Deep Ensemble

B. Critical-difference diagram

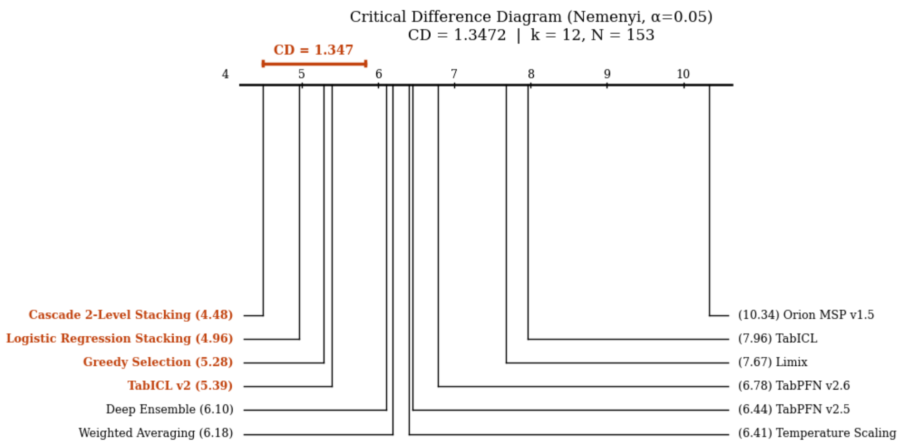

Figure 2. Critical-difference diagram for accuracy ranks (Friedman + Nemenyi, $\alpha = 0.05$, CD $= 1.347$). Methods connected by horizontal bars are statistically indistinguishable. The top equivalence group is Cascade_2level, Stacking_LR, Greedy_Selection, and TabICLv2.

C. Calibration, uncertainty, and group-robustness metrics

Table 3. Calibration, uncertainty, and group-robustness metrics on the TabArena classification suite. ECE: expected calibration error (15-bin). Brier-REL: reliability component of the Brier decomposition. AURC: area under the risk–coverage curve (max-probability scorer). Cov@95: fraction of samples predictable at $\geq 95\%$ accuracy. WGA: worst-group accuracy on sensitive-attribute datasets. Best ensemble and best base are bolded separately.

Method	**ECE↓**	**Brier-REL↓**	**AURC↓**	**Cov@95↑**	**WGA↑**
Cascade_2level	0.0261	**0.0025**	0.0485	0.677	**0.803**
Greedy_Selection	**0.0253**	0.0027	**0.0484**	0.679	0.776
Stacking_LR	0.0272	0.0031	0.0491	**0.684**	0.776
Temp_Scaled	0.0273	0.0030	0.0485	0.679	0.790
WA_performance	0.0277	0.0030	0.0485	0.679	0.790
DeepEnsemble_3seed	0.0275	0.0030	0.0617	0.621	0.790
TabICLv2	**0.0236**	**0.0024**	**0.0483**	**0.684**	**0.802**
TabICL	0.0261	0.0028	0.0498	0.673	0.775
TabPFNv2.6	0.0255	0.0028	0.0485	0.680	0.773
TabPFNv2.5	0.0258	0.0026	0.0490	0.670	0.779
LimiX	0.0252	0.0027	0.0496	0.673	0.789
OrionMSPv1.5	0.0419	0.0044	0.0603	0.629	0.758

D. Convex-combination ceiling: a formal statement

Proposition (Convex-combination ceiling). Let $f_1, \ldots, f_K$ be classifiers producing class-probability vectors $p_k(x)$ with hard predictions $\hat{y}_k(x) = \arg\max_c p_k(x)_c$, and let $\mathcal{C} = \{x : \hat{y}_1(x) = \cdots = \hat{y}_K(x)\}$ denote the consensus set. For any convex combiner $\hat{p}(x) = \sum_k w_k p_k(x)$ with $w_k \geq 0$ and $\sum_k w_k = 1$, the ensemble prediction $\arg\max_c \hat{p}(x)_c$ equals the unanimous label on $\mathcal{C}$. Consequently the accuracy gap between the ensemble and any single base classifier is bounded above by the disagreement-set fraction $1 - |\mathcal{C}|/|\mathcal{X}|$.

Proof sketch. On $x \in \mathcal{C}$ every $p_k(x)$ assigns its maximum to the same label y^*, so $\hat{p}(x)_{y^*} = \sum_k w_k p_k(x)_{y^*} \geq \sum_k w_k p_k(x)_c$ for every other class c, with equality only if every base ties at x. Hence $\hat{y}(x) = y^*$, matching every base on $\mathcal{C}$. Outside $\mathcal{C}$ the ensemble can differ from a base on at most $|\mathcal{X} \setminus \mathcal{C}|$ inputs, bounding the absolute accuracy gap.□

A high mean pairwise Q-statistic is a sufficient indicator that, when individual error rates are similar across the pool, the disagreement set is small (Kuncheva & Whitaker, 2003): errors concentrate on the same instances, so the consensus set is large and the bound above tightens. For our six-TFM pool the mean pairwise Q-statistic is $\bar{Q} = 0.961$ (computation in Appendix E), close to 1 and signalling a hard diversity ceiling any combiner must work against.

E. Q-statistic computation

The pairwise Q-statistic (Kuncheva & Whitaker, 2003) is computed from each pair of base TFMs' test-set predictions on each of the 153 tasks. For a pair (f_k, f_l) on a task, we form the 2×2 contingency table on test instances by (correct, wrong) under each classifier and define

$$Q_{kl}^{(\text{task})} = \frac{ad - bc}{ad + bc},$$

where a, b, c, d are the counts of (both correct), (k correct, l wrong), (k wrong, l correct), and (both wrong) respectively. Per-pair Q_{kl} is the unweighted mean of $Q_{kl}^{(\text{task})}$ across the 153 tasks. The pool-level $\bar{Q} = 0.961$ reported in §4.3 is the unweighted mean across the $\binom{6}{2} = 15$ unordered base-pair indices; $\sigma = 0.183$ is the standard deviation across the same set. Cohen's κ uses the same contingency tables and is reported using the standard chance-adjusted-agreement formula.

F. Head-to-head ensemble comparison

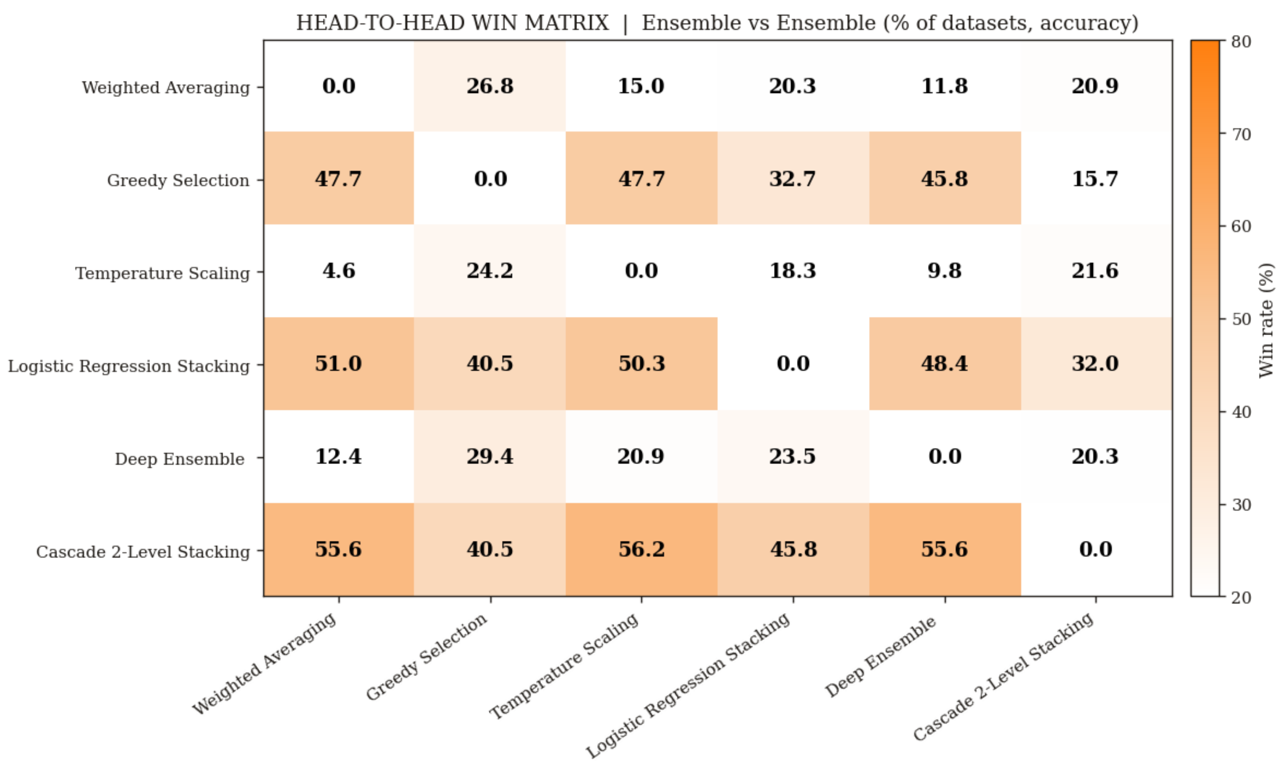

Figure 3. Head-to-head win matrix across 153 OpenML classification tasks (accuracy). Each cell (i, j) reports the percentage of datasets on which row method i outperforms column method j; the diagonal is zero by construction. Cascade 2-Level Stacking is the dominant strategy (55.6%, 56.2%, and 55.6% of datasets against Weighted Averaging, Temperature Scaling, and Deep Ensemble respectively). Logistic Regression Stacking ranks second. Weighted Averaging wins on at most 26.8% of datasets against any single opponent; Temperature Scaling never exceeds 24.2%.

G. Mean rank leaderboard

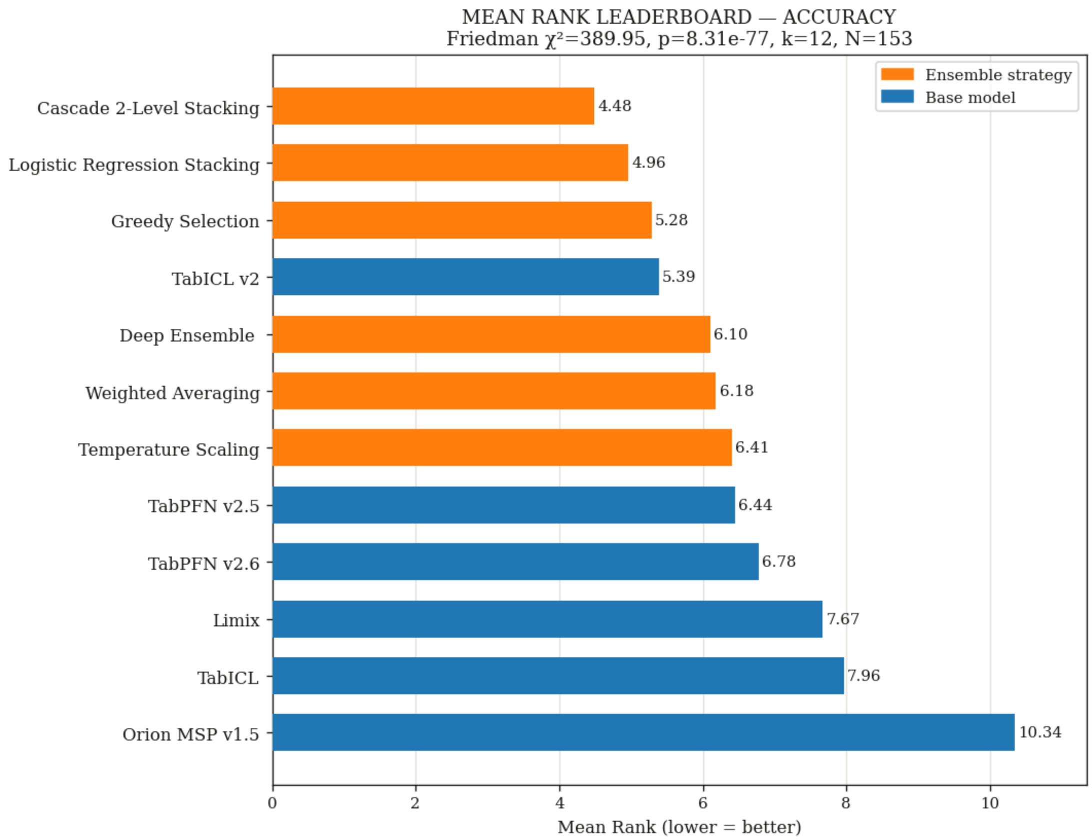

Figure 4. Mean rank leaderboard for the 12 methods evaluated across 153 datasets. Ensemble strategies (orange) cluster with the strongest base models (blue) at the top of the ranking; Cascade 2-Level Stacking achieves the best rank (4.48), followed by Logistic Regression Stacking (4.96) and Greedy Selection (5.28). Friedman $\chi^2 = 389.95$, $p \approx 8.31 \times 10^{-77}$.

H. Dataset details

This section lists the full inventory of OpenML benchmark datasets used in our study. Table 4 reports the OpenML identifier, dataset name, sample count, feature count, number of target classes, and source benchmark suite for each task.

Table 4. Full inventory of the 153 OpenML benchmark datasets used in this study. **ID**: OpenML dataset identifier. **Samples**: number of instances (range: 128–581,012; median: 3,196). **Feat.**: number of input features (range: 5–1,777; median: 22). **Cls.**: number of target classes (range: 2–10; median: 2). **Source**: benchmark suite from which the task was drawn. **CC18**: OpenML-CC18; **TA**: TabArena; **TL**: Talent; **TZ**: TabZilla. Datasets shared across suites carry combined source tags (e.g., CC18,TZ).

ID	Name	Samples	Feat.	Cls.	Source
3	`kr-vs-kp`	3,196	37	2	`CC18`
11	`balance-scale`	625	5	3	`CC18,TZ`
12	`mfeat-factors`	2,000	217	10	`CC18`
14	`mfeat-fourier`	2,000	77	10	`CC18,TZ`
15	`breast-w`	699	10	2	`CC18`
16	`mfeat-karhunen`	2,000	65	10	`CC18`
18	`mfeat-morphological`	2,000	7	10	`CC18`
21	`car`	1,728	7	4	`TL`
22	`mfeat-zernike`	2,000	48	10	`CC18,TZ`
23	`cmc`	1,473	10	3	`CC18`
27	`colic`	368	23	2	`TZ`
28	`optdigits`	5,620	65	10	`CC18`
29	`credit-approval`	690	16	2	`CC18,TZ`
30	`page-blocks`	5,473	11	5	`TL`
31	`credit-g`	1,000	21	2	`CC18,TZ`
32	`pendigits`	10,992	17	10	`CC18`
36	`segment`	2,310	20	7	`TL`
37	`diabetes`	768	9	2	`CC18`
38	`sick`	3,772	30	2	`CC18`
44	`spambase`	4,601	58	2	`CC18`
46	`splice`	3,190	61	3	`CC18,TZ`
50	`tic-tac-toe`	958	10	2	`CC18`
54	`vehicle`	846	19	4	`CC18,TZ`
60	`waveform-5000`	5,000	41	3	`TL`
151	`electricity`	45,312	9	2	`CC18`
179	`adult`	48,842	15	2	`TL`
180	`covertype`	110,393	55	7	`TL`
181	`yeast`	1,484	9	10	`TL`
182	`satimage`	6,430	37	6	`CC18`
188	`eucalyptus`	736	20	5	`CC18`
293	`covertype`	581,012	55	2	`TL`
333	`monks-problems-1`	556	7	2	`TZ`
458	`analcatdata_authorship`	841	71	4	`CC18`
469	`analcatdata_dmft`	797	5	6	`CC18`
470	`profb`	672	10	2	`TZ`
554	`mnist_784`	70,000	785	10	`CC18`
846	`elevators`	16,599	19	2	`TZ`
934	`socmob`	1,156	6	2	`TZ`
999	`audiology`	226	70	2	`TZ`
1038	`gina_agnostic`	3,468	971	2	`TL`
1043	`ada_agnostic`	4,562	49	2	`TZ`
1046	`mozilla4`	15,545	6	2	`TL`
1049	`pc4`	1,458	38	2	`CC18`

Continued on next page

Table 4 – *Continued from previous page*

ID	Name	Samples	Feat.	Cls.	Source
1050	pc3	1,563	38	2	CC18
1053	jm1	10,885	22	2	CC18
1063	kc2	522	22	2	CC18
1067	kc1	2,109	22	2	CC18,TZ
1068	pc1	1,109	22	2	CC18
1111	KDDCup09_appetency	50,000	231	2	TL
1112	KDDCup09_churn	50,000	231	2	TL
1114	KDDCup09_upselling	50,000	231	2	TL
1116	musk	6,598	168	2	TL
1119	adult-census	32,561	16	2	TL
1120	MagicTelescope	19,020	12	2	TL
1169	airlines	539,383	8	2	TZ
1459	artificial-characters	10,218	8	10	TZ
1461	bank-marketing	45,211	17	2	CC18
1462	banknote-authentication	1,372	5	2	CC18
1464	blood-transfusion-service-center	748	5	2	CC18
1467	climate-model-simulation-crashes	540	21	2	TL
1468	cnae-9	1,080	857	9	CC18,TZ
1471	eeg-eye-state	14,980	15	2	TL
1475	first-order-theorem-proving	6,118	52	6	CC18
1476	gas-drift	13,910	129	6	TL
1478	har	10,299	562	6	CC18
1480	ilpd	583	11	2	CC18
1485	madelon	2,600	501	2	CC18
1486	nomao	34,465	119	2	CC18,TZ
1487	ozone-level-8hr	2,534	73	2	CC18
1489	phoneme	5,404	6	2	CC18
1494	qsar-biodeg	1,055	42	2	CC18,TZ
1497	wall-robot-navigation	5,456	25	4	CC18
1501	semeion	1,593	257	10	CC18
1510	wdbc	569	31	2	CC18
1565	heart-h	294	14	5	TZ
1590	adult	48,842	15	2	CC18
1596	covertype	581,012	55	7	TL
4134	Bioresponse	3,751	1,777	2	CC18,TZ
4534	PhishingWebsites	11,055	31	2	CC18
4538	GesturePhaseSegmentationProcessed	9,873	33	5	CC18,TZ
6332	cylinder-bands	540	40	2	CC18
23381	dresses-sales	500	13	2	CC18
23512	higgs	98,050	29	2	TZ
23517	numerai28.6	96,320	22	2	CC18
40536	SpeedDating	8,378	121	2	TL
40646	GAMETES_Epistasis_2-Way_20atts_0.1H_EDM-1_1	1,600	21	2	TL
40647	GAMETES_Epistasis_2-Way_20atts_0.4H_EDM-1_1	1,600	21	2	TL
40648	GAMETES_Epistasis_3-Way_20atts_0.2H_EDM-1_1	1,600	21	2	TL
40649	GAMETES_Heterogeneity_20atts_1600_Het_0.4_0.2_50_EDM-2_001	1,600	21	2	TL
40650	GAMETES_Heterogeneity_20atts_1600_Het_0.4_0.2_75_EDM-2_001	1,600	21	2	TL
40668	connect-4	67,557	43	3	CC18
40670	dna	3,186	181	3	CC18
40680	mofn-3-7-10	1,324	11	2	TL

Continued on next page

Table 4 – *Continued from previous page*

ID	Name	Samples	Feat.	Cls.	Source
40681	mux6	128	7	2	TL
40682	thyroid-new	215	6	3	TL
40685	shuttle	58,000	10	7	TL
40701	churn	5,000	21	2	CC18
40900	Satellite	5,100	37	2	TL
40945	Titanic	1,309	14	2	TL
40966	MiceProtein	1,080	82	8	CC18
40975	car	1,728	7	4	CC18
40978	Internet-Advertisements	3,279	1,559	2	CC18
40979	mfeat-pixel	2,000	241	10	CC18
40981	Australian	690	15	2	TZ
40982	steel-plates-fault	1,941	28	7	CC18
40983	wilt	4,839	6	2	CC18
40984	segment	2,310	20	7	CC18
40994	climate-model-simulation-crashes	540	21	2	CC18
41027	jungle_chess_2pcs_raw_endgame_complete	44,819	7	3	CC18,TZ
41138	APSFailure	76,000	171	2	TL
41143	jasmine	2,984	145	2	TZ
41147	albert	425,240	79	2	TZ
41150	MiniBooNE	130,064	51	2	TZ
43945	electricity	38,474	9	2	TZ
43973	phoneme	3,172	6	2	TZ
46905	Amazon_employee_access	32,769	10	2	TA
46906	anneal	898	39	5	TA
46908	APSFailure	76,000	171	2	TA
46910	bank-marketing	45,211	14	2	TA
46911	Bank_Customer_Churn	10,000	11	2	TA
46912	Bioresponse	3,751	1,777	2	TA
46913	blood-transfusion-service-center	748	5	2	TA
46915	churn	5,000	20	2	TA
46916	coil2000_insurance_policies	9,822	86	2	TA
46918	credit-g	1,000	21	2	TA
46919	credit_card_clients_default	30,000	24	2	TA
46920	customer_satisfaction_in_airline	129,880	22	2	TA
46921	diabetes	768	9	2	TA
46922	Diabetes130US	71,518	48	2	TA
46924	E-CommereShippingData	10,999	11	2	TA
46927	Fitness_Club	1,500	7	2	TA
46929	GiveMeSomeCredit	150,000	11	2	TA
46930	hazelnut-spread-contaminant-detection	2,400	31	2	TA
46932	heloc	10,459	24	2	TA
46933	hiva_agnostic	3,845	1,618	3	TA
46935	HR_Analytics_Job_Change_of_Data_Scientists	19,158	13	2	TA
46937	in_vehicle_coupon_recommendation	12,684	25	2	TA
46938	Is-this-a-good-customer	1,723	14	2	TA
46939	kddcup09_appetency	50,000	213	2	TA
46940	Marketing_Campaign	2,240	26	2	TA
46941	maternal_health_risk	1,014	7	3	TA
46947	online_shoppers_intention	12,330	18	2	TA
46950	polish_companies_bankruptcy	5,910	65	2	TA

Continued on next page

Table 4 – *Continued from previous page*

ID	Name	Samples	Feat.	Cls.	Source
46952	qsar-biodeg	1,054	42	2	TA
46955	SDSS17	78,053	12	3	TA
46956	seismic-bumps	2,584	16	2	TA
46958	splice	3,190	61	3	TA
46960	students_dropout_and_academic_success	4,424	37	3	TA
46962	taiwanese_bankruptcy_prediction	6,819	95	2	TA
46963	website_phishing	1,353	10	3	TA
46969	NATICUSdroid	7,491	87	2	TA
46979	jm1	10,885	22	2	TA
46980	MIC	1,699	112	8	TA

I. Code and data availability

All ensemble strategies and base-TFM wrappers in this study were built on the open-source TabTune library (Tanna et al., 2025). Code, configurations, per-base inference logs, per-task scores, and the scripts that produce every figure and table in this paper will be released after acceptance via a public repository linked in the camera-ready version. Reviewers may contact the authors through the conference's anonymous channel for early access.