NATIVE LOCAL LINEAR EXPLANATIONS FOR TABULAR IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular in-context models adapt to unseen tasks without task-specific retraining, yet their cross-sample attention entangles feature representations into opaque predictions. We introduce TabLLE (*Tabular Learning with Local Linear Explanations*), an intrinsically interpretable in-context model that recasts support-conditioned prediction as dynamic, query-specific affine hyperplanes. Conditioned on query samples and column-level support sets via a Feature-Context Adapter (FCA), an adaptive hypernetwork generates instance-specific weights and biases in a single forward pass. For the parameters generated in the same forward pass, each class logit decomposes into an exact sum of coordinate-indexed local-linear terms and a sample-dependent additive offset, yielding native instance-level attributions without perturbation-based post-hoc evaluation. To prevent the hypernetwork from bypassing feature pathways via bias shortcuts, TabLLE is trained with a three-term objective that explicitly enforces standalone feature discriminability. Across 20 OpenML benchmarks, TabLLE is competitive with tree ensembles and leading in-context baselines, with 1.5–5.15$\times$ faster inference than TabICLv2 at matched retrieval sizes, while attaining the highest AUC and accuracy among intrinsically interpretable baselines. Finally, we formalize parameter drift under feature interventions and show that a behavioral head reliably accounts for cross-pass masking effects.

1 INTRODUCTION

Tabular data, characterized by heterogeneous feature domains, extreme numerical skew, and absence of spatial invariance, underpin critical automated decisions in clinical diagnostics, financial risk assessment, and operational monitoring. While tree-based ensembles (XGBoost, LightGBM, Random Forest) remain strong baselines, they operate strictly per-dataset, requiring iterative hyperparameter tuning and full retraining whenever applied to a new task.

Recent advances reframe tabular prediction through in-context learning, deploying a single foundation model that generalizes across unseen datasets at inference time without gradient descent. Prior-Data Fitted Networks (PFNs) (Müller et al., 2022) formalize this process as amortized Bayesian inference over synthetic task priors, with architectures such as TabPFN (Hollmann et al., 2023), TabICL (Qu et al., 2025), and TabICLv2 (Qu et al., 2026) demonstrating that pre-trained sequence models can effectively scale to larger tables through hierarchical encoding (den Breejen et al., 2024; Ma et al., 2025; Kim et al., 2024; Xie et al., 2022; von Oswald et al., 2023). However, prevailing tabular in-context foundation models function as opaque black boxes. To process multi-row support contexts, standard implementations compress column-level representations into unified entity tokens, discarding feature-isolated computational pathways and relying on external, sampling-heavy post-hoc surrogates such as Kernel SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), or Integrated Gradients (Sundararajan et al., 2017) for interpretability.

To provide in-context generalization with intrinsic transparency, we propose TabLLE (Tabular Learning with Local Linear Explanations). TabLLE relies on a decoupled architecture where contextual task representation is separated from query-level prediction. Specifically, a Feature-Context Adapter (FCA) extracts column-resolved support representations once per fixed support context. An adaptive hypernetwork then ingests this support context alongside the query embedding, dynamically synthe-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sizing instance-specific linear weights and an offset bias in a single forward pass. Consequently, for a fixed unmasked forward pass, each class logit decomposes algebraically into coordinate-indexed local-linear terms and a sample-dependent additive offset during prediction, yielding native explanations without perturbation-based post-hoc surrogate evaluation. To prevent the hypernetwork from bypassing feature pathways via trivial bias shortcuts, we introduce an objective that explicitly enforces standalone feature discriminability. Furthermore, we formalize the mechanism of parameter drift under counterfactual feature masking, providing a behavioral head that ensures consistency across intervention passes. Overall, the primary contributions of this work are fourfold:

- We propose TabLLE, an intrinsically interpretable tabular in-context foundation model that adapts to unseen tasks without task-specific retraining and provides closed-form, query-specific local linear attributions within the same forward pass.
- We establish the FCA to amortize column-aware support representations across queries, decoupling task-level context encoding from instance-level affine parameter synthesis.
- We formalize the mechanism of parameter drift under counterfactual feature masking and show that a behavioral head substantially mitigates the divergence between static native attributions and active masking responses.
- Pre-trained on 1.44M synthetic tasks, the 31.16M-parameter TabLLE matches tree ensembles and top in-context baselines across 20 OpenML benchmarks, achieving 1.5–5.15$\times$ faster inference than TabICLv2.

2 Related Work

Task-specific tabular prediction. Gradient-boosted trees (XGBoost, LightGBM, CatBoost) dominate tabular benchmarks through axis-aligned splits and ensemble residual fitting, with systematic evaluations attributing their advantage over deep networks to properties of tabular tasks rather than tuning gaps (Grinsztajn et al., 2022; Borisov et al., 2024). Deep architectures attempt to capture flexible feature interactions: FT-Transformer (Gorishniy et al., 2021) applies column-level self-attention, TabR (Gorishniy et al., 2024) adds learned nearest-neighbor retrieval, and TabM (Gorishniy et al., 2025) pursues parameter-efficient ensembling. All train on one dataset at a time and require retraining when the task changes.

Tabular foundation models and in-context learning. PFNs (Müller et al., 2022) reframe learning by training on sampled tasks so that inference reads a support set directly. TabPFN (Hollmann et al., 2023) applied this to small tables. TabICL (Qu et al., 2025) and TabICLv2 (Qu et al., 2026) scale via hierarchical column-then-row encoding. TabPFNv2 (Hollmann et al., 2025) refines column tokenization, TabForestPFN (den Breejen et al., 2024) enriches synthetic priors with forest generators, TabDPT (Ma et al., 2025) mixes real data into pretraining, and CARTE (Kim et al., 2024) represents tables as graphs for cross-schema transfer. None of these models maps the support-conditioned decision rule back onto individual input features.

Interpretable tabular models. EBM (Nori et al., 2019), NAM (Agarwal et al., 2021a), and NodeGAM (Chang et al., 2022) learn inspectable feature-wise additive functions, while IMN (Kadra et al., 2024) generates sample-specific discriminative rules. Both paradigms, however, depend on task-specific training. In the in-context setting, ShapPFN (Sena & Azevedo, 2026) achieves interpretability via Shapley value regression, yet it does not expose explicit local linear coefficients over input features.

3 TabLLE

3.1 Problem Setup

We formulate tabular in-context classification as episodic meta-learning (Müller et al., 2022) over a synthetic task distribution. Each task $\tau \sim p(\tau)$ provides a labeled support set $\mathcal{S}_\tau = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{S}$ and a disjoint query set $\mathcal{Q}_\tau = \{(\mathbf{x}_j^q, y_j^q)\}_{j=1}^{N_q}$, where each instance $\mathbf{x} \in \mathbb{R}^M$ and labels belong to a task-specific active class set $\mathcal{C}_\tau = \{0, \ldots, K_\tau - 1\}$. To accommodate variable class cardinalities up

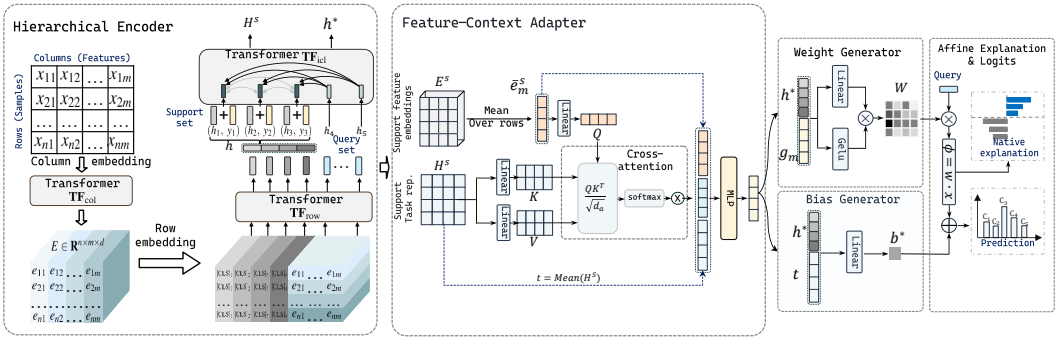


Figure 1: **Schematic illustration of the TabLLE framework.** Tabular entries are hierarchically processed by column, row, and asymmetric in-context transformers to yield support and query context embeddings. The FCA retrieves column-specific task evidence once per fixed support context. Conditioned on this context, a dual-stream hypernetwork dynamically instantiates instance-specific linear weights and biases in a single forward pass, providing an exact affine decomposition for native, zero-overhead feature attribution.

to a fixed maximum capacity $C_{\max} \geq K_\tau$, TabLLE maintains $C_{\max}$ structural output slots while restricting loss evaluation and softmax normalization strictly to $\mathcal{C}_\tau$. The model parameters θ are meta-trained over synthetic tasks drawn from a curriculum-driven prior $p_r(\tau)$. At meta-test time, given an unseen support set $\mathcal{S}_{\text{new}}$, TabLLE classifies query instances (denoted by superscript $*$) in a single forward pass with θ frozen, requiring zero gradient steps or hyperparameter adaptation.

To achieve support-conditioned prediction while maintaining exact, native transparency, TabLLE organizes computation into three coordinated stages (Figure 1): (i) a hierarchical encoder that converts tabular entries into decoupled column tokens and cross-instance representations (§ 3.2); (ii) a FCA that aggregates column-level support evidence once per fixed support context (§ 3.3); and (iii) a dual-stream hypernetwork that dynamically synthesizes instance-specific affine classification parameters in a single forward pass (§ 3.4). We detail each stage below.

3.2 Support-Conditioned Tabular Encoding

Building upon hierarchical tabular transformers (Gorishniy et al., 2021; Qu et al., 2025), the encoding backbone (Figure 1) decomposes computation into column, row, and in-context attention stages. Compressing intra-row features prior to inter-instance attention ensures that computational complexity scales quadratically with sample count rather than total tabular cell count (Hollmann et al., 2023).

Given the concatenated support and query matrix $\mathbf{X} = [\mathbf{X}^s; (\mathbf{x}^*)^\top]$, a target-aware column encoder first projects numerical and categorical entries into dense feature tokens while embedding support labels:

$$\mathbf{E} = \mathrm{TF}_{\mathrm{col}}(\mathbf{X}; \mathbf{y}^s). \tag{1}$$

Within each column-attention block, inducing tokens attend only to support rows, and each row attends to these tokens. Label embeddings are added exclusively to support rows. The row-level transformer subsequently aggregates intra-row feature interactions into unified instance representations:

$$[\mathbf{R}^s; \mathbf{r}^*] = \mathrm{TF}_{\mathrm{row}}(\mathbf{E}). \tag{2}$$

Before inter-instance modeling, ground-truth support classes are reinforced through additive categorical embeddings:

$$\tilde{\mathbf{r}}_i^s = \mathbf{r}_i^s + \mathbf{e}_y(y_i^s), \quad i = 1, \ldots, S. \tag{3}$$

Here $\mathbf{e}_y(\cdot)$ is a learnable class-embedding table. This row-level pathway is distinct from the label-conditioned column encoding in Equation (1): the latter shapes per-column tokens, whereas Equation (3) equips each support instance with an explicit class anchor before inter-instance attention. In $\mathrm{TF}_{\mathrm{icl}}$, all positions serve as queries, but keys and values are restricted to support rows:

$$[\mathbf{H}^s; \mathbf{h}^*] = \mathrm{TF}_{\mathrm{icl}}\Big(Q = [\tilde{\mathbf{R}}^s; \mathbf{r}^*],\ K = V = \tilde{\mathbf{R}}^s\Big). \tag{4}$$

Together with support-only column aggregation, this asymmetry enforces support-to-query conditioning. Queries strictly read fixed support representations, ensuring invariance to query-batch composition and enabling support-side cache reuse.

3.3 Feature-Context Adapter

While column-level encoding preserves distinct feature identities, it lacks the task-level relational context captured by $\mathrm{TF}_{\mathrm{icl}}$. Conversely, the support representation $\mathbf{H}^s$ captures cross-sample task evidence but collapses individual column channels. The FCA aligns these complementary representations by retrieving dataset-level context conditioned on feature-specific embeddings.

Let $\mathbf{e}^s_{im}$ denote the feature-group token at column m in support row i. Its support-averaged representation is $\bar{\mathbf{e}}^s_m = \frac{1}{S}\sum_{i=1}^{S} \mathbf{e}^s_{im}$. A global task summary is formed as $\mathbf{t} = \frac{1}{S}\sum_{i=1}^{S} \mathbf{h}^s_i$. Learnable projections produce attention queries from $\bar{\mathbf{e}}^s_m$ and keys/values from $\mathbf{h}^s_i$:

$$\mathbf{a}_m = \mathrm{Attn}(W_q \bar{\mathbf{e}}^s_m,\ \{W_k \mathbf{h}^s_i\},\ \{W_v \mathbf{h}^s_i\}), \quad \mathbf{g}_m = \mathrm{MLP}_{\mathrm{FCA}}([\bar{\mathbf{e}}^s_m; \mathbf{a}_m; \mathbf{t}]). \tag{5}$$

Here $W_q \in \mathbb{R}^{A\times E}$ and $W_k, W_v \in \mathbb{R}^{A\times D}$ map the E=128-dimensional column-group tokens and the D=512-dimensional row-level support states into a shared A=256-dimensional attention space, aligning feature-level queries with task-level context. The resulting token $\mathbf{g}_m$ integrates explicit feature identity with dataset-level discriminative evidence. For a fixed support context, $\mathbf{g}_m$ is query-independent and reused across queries. Changing the support context requires recomputation.

3.4 Query-Specific Affine Prediction and Native Explanation

Conditioned on the contextual representations, a dual-stream hypernetwork (Ha et al., 2017) generates the decision parameters for each query instance. Specifically, a gated weight generator $\mathcal{H}_W$ predicts class-specific linear coefficients from $\mathbf{h}^*$ and $\mathbf{g}_m$, while an ungated bias generator $\mathcal{H}_b$ synthesizes the offset bias from $\mathbf{h}^*$ and $\mathbf{t}$:

$$w^*_{cm} = \big[\mathcal{H}_W(\mathbf{h}^*, \mathbf{g}_m)\big]_c, \qquad \mathbf{b}^* = \mathcal{H}_b(\mathbf{h}^*, \mathbf{t}), \tag{6}$$

forming the instance-adaptive parameter matrix $\mathbf{W}^* \in \mathbb{R}^{C_{\max}\times M}$.

To handle irregular feature structures and missing entries, let $a^{\mathrm{feat}}_m \in \{0,1\}$ denote feature validity and $\mu^{\mathrm{miss}}_m \in \{0,1\}$ denote entry missingness, yielding effective query values $\tilde{x}^*_m = a^{\mathrm{feat}}_m(1-\mu^{\mathrm{miss}}_m)x^*_m$ where invalid padded channels and missing entries are structurally masked. The resulting query-adaptive affine classification model defines a per-class decomposition of the output logits:

$$\phi^*_{cm} \triangleq w^*_{cm}\,\tilde{x}^*_m, \qquad z_c = \sum_{m=1}^{M} \phi^*_{cm} + b^*_c, \quad \forall c \in \mathcal{C}_\tau. \tag{7}$$

The native explanation for the predicted class $\hat{c}$ is the corresponding slice, and we write $\phi^*_m \equiv \phi^*_{\hat{c}m}$ throughout. This formulation yields an algebraically exact affine decomposition conditioned on the instantaneous forward pass. Drawing conceptual motivation from self-explaining neural networks (SENN) (Alvarez-Melis & Jaakkola, 2018b), the identical parameters that compute the prediction simultaneously define its native explanation, eliminating perturbation sampling and surrogate approximations (Ribeiro et al., 2016).

3.5 Training Objective

TabLLE is meta-trained on synthetic tasks sampled from a curriculum-driven prior distribution $p_r(\tau)$ spanning diverse structural causal models (SCM), tree ensembles, and sparse linear functions (detailed in Section B). If optimization relies solely on the full logit predictions, the hypernetwork tends to suffer from shortcut learning (Geirhos et al., 2020), routing dominant discriminative signals through the unconstrained bias vector $\mathbf{b}^*$ alone and causing the feature-weight pathways to collapse into uninformative constants. We prevent this shortcut with a three-term objective. For each query $(\mathbf{x}^q_j, y^q_j)$, define the feature-only score $z^{\mathrm{feat}}_c = \sum_m \phi^*_{cm}$. The loss is:

$$\mathcal{L}_{\tau,j}(\theta; r) = \lambda_{\mathrm{full}}\ \mathrm{CE}_{\mathcal{C}_\tau}(\mathbf{z}, y^q_j) + \lambda_{\mathrm{feat}}\ \mathrm{CE}_{\mathcal{C}_\tau}(\mathbf{z}^{\mathrm{feat}}, y^q_j) + \lambda_{\mathrm{bias}}\,\frac{1}{K_\tau}\sum_{c\in\mathcal{C}_\tau}(b^*_c)^2. \tag{8}$$

Table 1: Benchmark evaluation across 20 OpenML datasets (5-fold × 5-repeat). Entries are IQM across the 20 dataset-level scores, with 95% confidence limits as subscript/superscript estimated by 10,000 bootstrap resamples of the datasets. Models are categorized by inductive paradigm, native interpretability (Interp.), and zero-shot in-context transfer capability (0-Shot).

Paradigm	Model	Interp.	0-Shot	AUC IQM	Acc IQM	Paradigm	Model	Interp.	0-Shot	AUC IQM	Acc IQM
In-Context Foundation	TabICLv2	×	✓	$.937_{.884}^{.976}$	$.879_{.817}^{.931}$	*Deep Tabular Networks*	ModernNCA	×	×	$.895_{.835}^{.950}$	$.844_{.773}^{.906}$
	TabPFN-kNN	×	✓	$.911_{.849}^{.964}$	$.863_{.787}^{.921}$		TabM	×	×	$.888_{.819}^{.951}$	$.842_{.766}^{.904}$
	TabForestPFN	×	✓	$.914_{.863}^{.963}$	$.859_{.778}^{.918}$		MambaTab	×	×	$.883_{.817}^{.947}$	$.847_{.765}^{.906}$
	ShapPFN	✓	✓	$.871_{.798}^{.937}$	$.818_{.727}^{.885}$		AutoInt	×	×	$.880_{.798}^{.947}$	$.835_{.753}^{.902}$
TabLLE (Ours)		✓	✓	$.916_{.852}^{.965}$	$.865_{.794}^{.919}$		FT-Trans.	×	×	$.879_{.799}^{.948}$	$.840_{.755}^{.907}$
Intrinsically Interpretable	EBM	✓	×	$.906_{.846}^{.960}$	$.862_{.781}^{.919}$		NODE	×	×	$.865_{.783}^{.941}$	$.844_{.766}^{.901}$
	NAM	✓	×	$.893_{.827}^{.947}$	$.839_{.772}^{.900}$		TabNet	×	×	$.876_{.812}^{.917}$	$.800_{.745}^{.860}$
	NodeGA2M	✓	×	$.887_{.820}^{.946}$	$.840_{.760}^{.900}$		TabR	×	×	$.865_{.779}^{.926}$	$.802_{.713}^{.870}$
	NodeGAM	✓	×	$.885_{.817}^{.943}$	$.836_{.751}^{.898}$						
	IMN	✓	×	$.889_{.811}^{.946}$	$.847_{.766}^{.906}$						
Tree Ensembles	XGBoost	×	×	$.914_{.848}^{.962}$	$.858_{.786}^{.913}$	*Linear Simple*	LR	✓	×	$.876_{.813}^{.939}$	$.800_{.731}^{.871}$
	RF	×	×	$.912_{.851}^{.965}$	$.862_{.790}^{.919}$		DT	✓	×	$.838_{.751}^{.914}$	$.820_{.733}^{.889}$
	LightGBM	×	×	$.916_{.851}^{.964}$	$.861_{.791}^{.917}$						

Here CE denotes the cross-entropy loss. The first term optimizes the complete predictor. The second term explicitly enforces standalone feature discriminability, ensuring that feature contributions independently sustain classification accuracy. The third term regularizes bias magnitude, preventing it from becoming a predictive shortcut. We use $\lambda_{\text{full}}=1$, $\lambda_{\text{feat}}=0.5$, $\lambda_{\text{bias}}=0.01$. The episodic objective averages query losses within each sampled task.

4 Experiments

Setup. We evaluate on 20 public OpenML classification benchmarks spanning binary and multi-class tasks, sample sizes from 150 to 150k, and feature counts from 4 to 59 (full list in Section A). All experiments use a 5-fold × 5-repeat protocol (25 evaluations per method per dataset). The primary metrics are macro one-vs-rest area under the ROC curve (AUC) and accuracy. We compare against 22 baselines across four families: tree ensembles (XGBoost, LightGBM, RF), in-context models (TabICLv2, TabPFN-kNN, TabForestPFN), deep tabular networks (FT-Transformer, TabNet, TabM, TabR, ModernNCA, and others), and interpretable models (EBM, NAM, NodeGAM, IMN, ShapPFN). TabLLE and the TabICL variants use no per-dataset tuning. Full details appear in Sections A and B.

4.1 Benchmark Performance

Table 1 details the interquartile mean (IQM) and dataset-bootstrap 95% confidence intervals for the 23 evaluated methods (Agarwal et al., 2021b). TabLLE attains the second-highest aggregate score with an AUC of 0.916 and an accuracy of 0.865. These metrics trail TabICLv2 by 0.021 and 0.014 respectively. Among intrinsically interpretable models, TabLLE improves upon the AUC IQM of EBM by 0.010 and exceeds the scores of NAM, NodeGAM, and IMN by wider margins. Applying the conservative criterion of 95% lower confidence limits preserves this overall ordering. The lower bounds of TabLLE rank second globally at 0.852 for AUC and 0.794 for accuracy. These values exceed the corresponding bounds of all remaining 21 baselines, including every tree ensemble variant. Within the interpretable subset, TabLLE yields the highest point estimates and lower bounds for both evaluation metrics. Section H provides the individual dataset results.

4.2 Empirical Analysis of Native Local Linear Explanations

Geometric validity of local linear decisions. A primary concern for locally linear models is whether query-dependent affine outputs can model complex decision surfaces without sacrificing feature interactions. To test this, we evaluate TabLLE on three 2D synthetic datasets (Moons, Circles, and XOR; 512 support and 2,048 test samples) alongside six baselines (Figure 2). TabLLE accurately traces the nonlinear contours across all tasks (AUCs of 0.994, 0.999, and 1.000). In particular, its exact solution on XOR demonstrates that dynamically generating instance-specific parameters

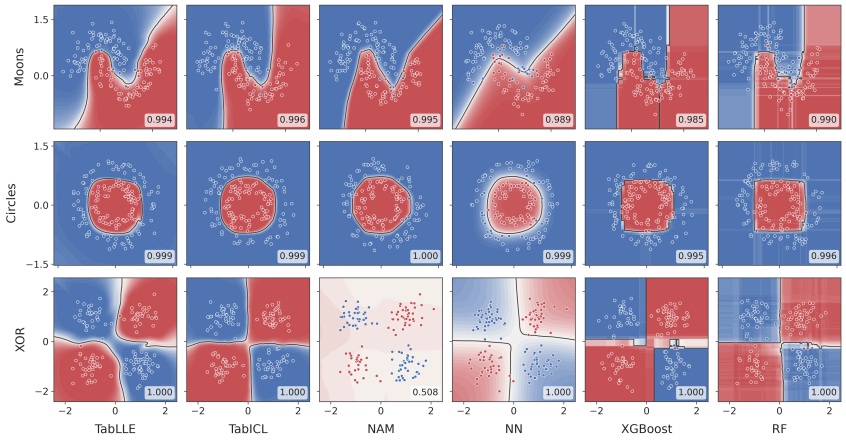


Figure 2: Decision boundaries on 2D synthetic tasks. Test AUC is shown per panel.

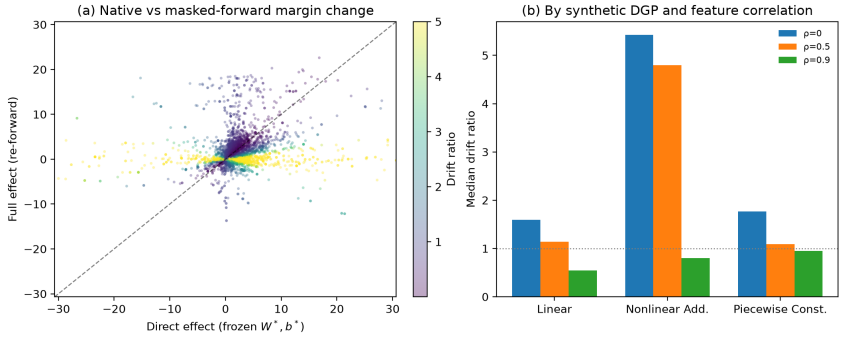


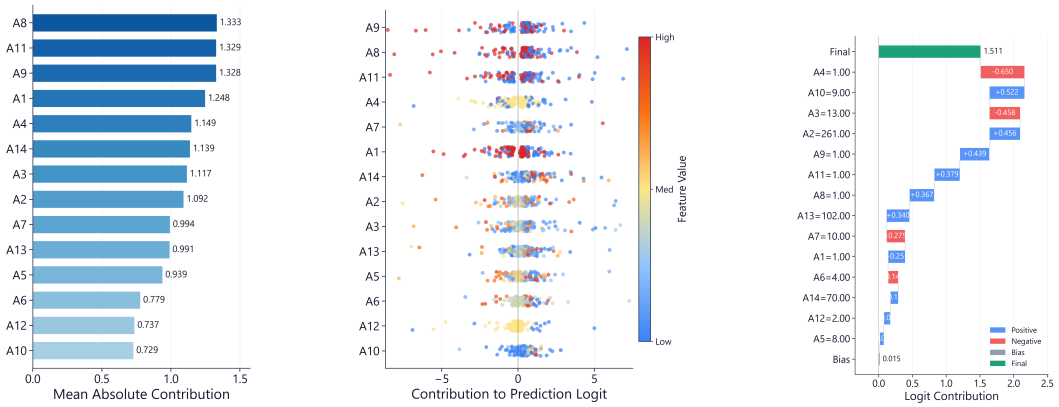
Figure 3: Left: scatter of Δm_{direct} vs. Δm_{full} (color = drift ratio, dashed $y{=}x$ = zero drift). Right: median drift ratio by DGP and ρ.

$(\mathbf{W}^*, \mathbf{b}^*)$ overcomes the structural limitation of generalized additive models (e.g., NAM), which fail when features interact nonlinearly. Geometrically, this confirms that TabLLE's local linear explanations reflect accurate first-order approximations of the underlying data manifold at each query point, rather than being forced a globally linear compromise. Locality radius measurements are detailed in Section D.

Physical mechanism of parameter drift. Within a single forward pass, TabLLE yields exact feature attributions $\phi_m^* = w_{\hat{c}m}^* \tilde{x}_m^*$ through local logit decomposition (Equation (7)). Masking-based benchmarks such as XAI-Bench (Liu et al., 2021) evaluate explanations by masking individual features and repeating the forward pass. In models with input-independent parameters, masking alters only the input features. In TabLLE, the decision parameters $(\mathbf{W}^*, \mathbf{b}^*)$ are generated dynamically from the query representation $\mathbf{h}^*$. Masking a feature updates $\mathbf{h}^*$ and produces a new parameter pair $(\mathbf{W}^{*\prime}, \mathbf{b}^{*\prime})$. This re-estimation represents parameter drift. For a binary margin $m = z_{\hat{c}} - z_{\bar{c}}$, where $\bar{c}$ denotes the competing class, the margin change under frozen parameters Δm_{direct} diverges from the change obtained through a full re-forward pass Δm_{full}. Across 27,648 measurements spanning three data-generating processes, three correlation levels, and 128 random masks per configuration, the drift ratio $|\Delta m_{\text{full}} - \Delta m_{\text{direct}}|/(|\Delta m_{\text{full}}| + \varepsilon)$ (with ε a small constant for numerical stability) has a median of 1.42 and exceeds one in 58.4% of cases. Figure 3 visualizes the results: the left panel scatters Δm_{direct} against Δm_{full} for individual masks, and the right panel aggregates the median drift ratio by data-generating process and correlation level. This ratio peaks on nonlinear additive tasks and decreases as feature correlation rises, dropping from 2.39 at $\rho{=}0$ to 0.77 at $\rho{=}0.9$. These measurements clarify that masking-based metrics conflate direct feature removal with parameter re-estimation.

Table 2: XAI-Bench results (15 tasks). This masking-based protocol directly rewards modeling re-forward effects, amplifying parameter drift for native decomposition methods. Metric definitions and source references are given in Section E.

Method	Faithfulness ↑	Infidelity ↓	Top-3 Removal ↑	Stability ↑
Random	−0.045	3.811	3.206	−0.035
TabLLE (native)	0.263	7.343	4.565	0.983
LIME	0.516	4.525	5.533	0.833
TabLLE (behavioral head)	0.599	4.198	5.772	0.996
SHAP	0.678	4.211	6.207	0.963
Kernel SHAP	0.687	4.210	6.260	0.998



(a) Global importance. (b) Beeswarm. (c) Waterfall.

Figure 4: Hierarchical interpretation of TabLLE on the Australian credit benchmark across operational scales: (a) Global feature importance aggregated over support context; (b) Cohort-level beeswarm distribution of query attributions; (c) Instance-level exact logit decomposition, where feature attributions ϕ_m^* sum algebraically to the final margin.

Behavioral head and XAI-Bench evaluation. To align single-pass native attributions with multi-pass perturbation benchmarks without altering the underlying predictor, we introduce a lightweight behavioral head. Conditioned on $\mathbf{h}^*$, this head regresses the expected margin change under feature masking and re-inference. The head comprises a LayerNorm and a two-layer MLP ($512 \to 256 \to 6$, GELU activations), trained with a Huber loss on masked-subset attribution sums. In Table 2, native single-pass attributions obtain a Faithfulness score of 0.263 due to parameter drift, whereas the behavioral head achieves 0.599 and reduces Infidelity to 4.198, improving over LIME (4.525) and tracking Kernel SHAP (4.210). The behavioral head also produces deterministic attributions with 0.996 Stability, comparable to Kernel SHAP (0.998) and higher than LIME (0.833), without requiring hundreds of Monte Carlo sampling passes.

Hierarchical explanation on real-world benchmarks. To evaluate real-world utility, Figure 4 presents a hierarchical interpretation pipeline on the Australian credit benchmark across three operational scales: (i) *Dataset-level global ranking* (Figure 4a), identifying predominant risk drivers aggregated over the support context; (ii) *Cohort-level attribution distributions* (Figure 4b), exposing how feature directions and non-monotonic values modulate decisions across individual queries via a beeswarm plot; and (iii) *Instance-level exact logit decomposition* (Figure 4c), where positive and negative feature contributions ϕ_m^* sum strictly with the baseline bias b_c^* to match the predicted margin. Unlike sampling-based post-hoc surrogates that approximate explanations asymptotically, TabLLE delivers instantaneous, mathematically verifiable credit auditability in a single forward pass without sampling overhead.

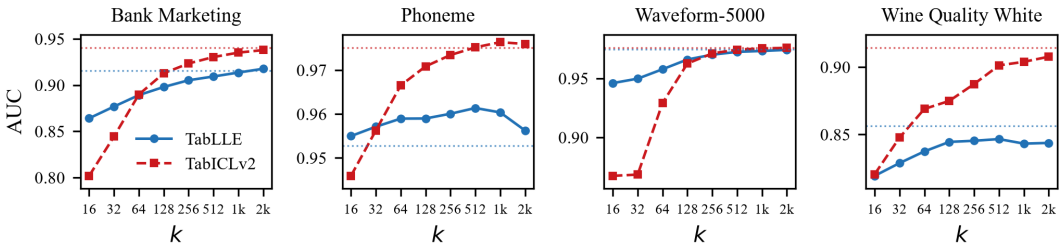


Figure 5: Inference robustness under local context compression ($k \in \{16, \ldots, 2048\}$) on four benchmarks. Solid and dashed curves denote TabLLE and TabICLv2, respectively, with dotted horizontal lines indicating full-support performance. TabLLE shows minimal degradation under small k and outperforms full support on phoneme at k=512 due to noise reduction.

Table 3: Inference speedup ratio ($\text{Time}_{\text{TabICLv2}}/\text{Time}_{\text{TabLLE}}$) under matched local context size k. Values $> 1.0\times$ indicate faster inference for TabLLE.

Dataset	k=16	k=512	k=2048
bank-marketing	1.50×	2.59×	3.95×
phoneme	1.51×	2.06×	2.88×
waveform-5000	2.40×	4.31×	5.15×
wine-quality-white	1.60×	2.54×	3.53×

4.3 Inference Efficiency and Context Scalability

Decoupling support encoding from query inference enables conditioning each query on a local neighborhood. We evaluate per-query k-nearest neighbors in standardized feature space against TabICLv2 across $k \in \{16, \ldots, 2048\}$ on four benchmarks under a 5-fold $\times$ 5-repeat protocol (Figure 5). Each query-specific neighborhood requires fresh support encoding and FCA computation without cache reuse, both included in the reported inference timings. When context is reduced to k=16, TabLLE loses 0.029 mean AUC relative to full support, whereas TabICLv2 drops by 0.093 (Figure 5). This resilience reflects the lower sample complexity of estimating a local affine hyperplane compared to learning non-parametric attention over sparse inputs. At k=512, the performance gap closes to 0.02, and local conditioning on *phoneme* exceeds full-support AUC by 0.008 by filtering distant label noise. In addition, localized conditioning bypasses monolithic cross-attention across full datasets, yielding $1.50\times$–$5.15\times$ inference speedups over TabICLv2 at matched k (Table 3). See Section F for context-dependent attribution stability.

4.4 Ablation Study

Three-term objective ablation. Without explicit constraints, the hypernetwork tends to route discriminative signals through the unconstrained bias $\mathbf{b}^*$, causing feature-weight pathways to collapse into uninformative values (§ 3.5). To validate the three-term objective (Equation (8)), we evaluate four loss configurations on 12 datasets $\times$ 3 seeds over 150k training steps (Figure 6). The standard objective ($\mathcal{L}_0$: full CE only) suffers from feature-path collapse (feature-only AUC < 0.55) on 92% of runs, with a bias-to-feature magnitude ratio exceeding 600. Partial objectives fail to prevent both failure modes simultaneously: adding feature CE alone ($\mathcal{L}_1$) reduces collapse to 11% but leaves the bias ratio near 1, whereas applying the L_2 bias penalty alone ($\mathcal{L}_2$) compresses the ratio to 0.026 while suppressing baseline expressiveness (3% collapse). The three-term objective $\mathcal{L}_3$ eliminates collapse entirely, drives the bias ratio to 0.019, and matches feature-only AUC with full prediction AUC, confirming that all three terms are necessary to secure faithful feature attributions while preserving predictive capacity.

Ablation on synthetic prior diversity. We compare two pre-training schedules branching from a shared 60k root (Section B) to isolate the source of zero-shot generalization. The *Control* branch trains exclusively on the base SCM prior. The *Broad* branch switches to structurally diverse generators, including Decision Surface Random Forests (DSRF), tree ensembles, and complex functions.

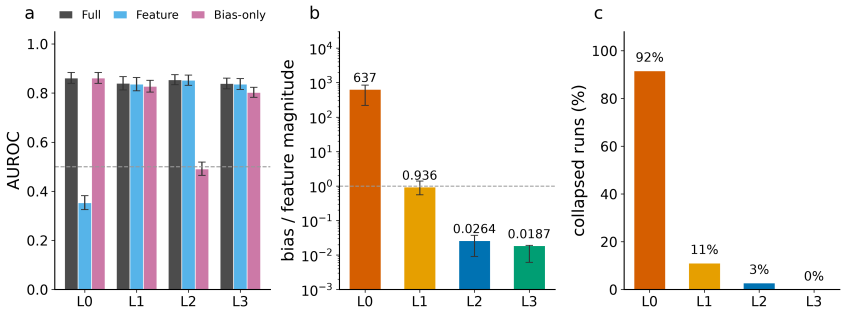


Figure 6: Ablation of the three-term objective across 12 datasets over 150k training steps: (a) Test AUC under full, feature-only, and bias-only logit components; (b) Bias-to-feature magnitude ratio (log scale); (c) Feature-path collapse rate. The three-term objective $\mathcal{L}_3$ eliminates shortcut collapse while preserving full predictive accuracy.

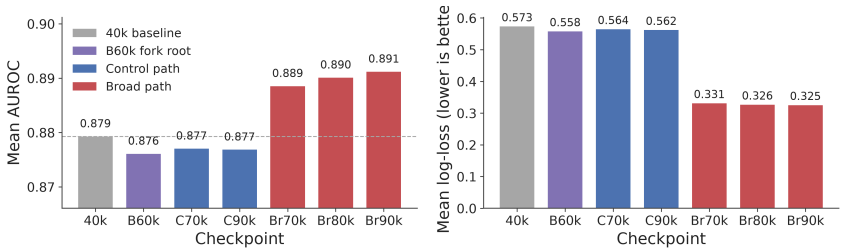


Figure 7: Zero-shot gains from prior diversity versus extended optimization. The panels display mean AUC and log-loss at each pre-training checkpoint. Evaluations use a 10-dataset suite under a full-support protocol with 5-fold cross-validation, 5 repeats, and single-ensemble inference. The dashed line marks the 40k baseline at an AUC of 0.879.

Evaluated on a 10-dataset validation suite (Figure 7), the Control branch yields no AUC gain. It scores 0.877 at both 70k and 90k, representing a slight decline from the 40k checkpoint ($\Delta\text{AUC} \approx -0.002$). Its log-loss shows only a marginal reduction from 0.573 to 0.562. The Broad branch demonstrates immediate improvements. It gains $+0.009$ AUC at 70k and reaches $+0.012$ at 90k. This trajectory raises the final mean AUC to 0.891, and cuts the log-loss to 0.325. Because these branches differ exclusively in their post-fork priors, the measurements establish that prior distributional diversity actively drives tabular in-context generalization. Prolonged optimization on homogeneous synthetic data provides no equivalent benefit. Extended learning curves are available in Section C.

5 Conclusion

TabLLE constructs a query-specific local affine classifier for each prediction. Each class logit decomposes exactly into per-feature contributions and a bias term within a single forward pass. Across 20 tabular benchmarks, its accuracy is statistically indistinguishable from tree ensembles and the strongest in-context baseline, while significantly outperforming most deep tabular networks. The three-term objective is required to maintain this formulation. Without separate feature-path supervision, the bias term absorbs the discriminative signal and collapses the native explanation.

Two limitations warrant emphasis. First, the affine output constraint reduces peak expressiveness on tasks with sharply curved boundaries. Second, native contributions can diverge from feature-masking effects once query-adaptive parameters shift under perturbation (median drift ratio 1.42). The behavioral head mitigates this in XAI-Bench evaluations without altering the predictor.

Future work includes scaling to larger feature dimensions, extending the affine head to regression and survival objectives, and investigating whether weights trained on interventional priors can carry causal meaning.

Reproducibility statement. The model architecture, training objective (Equation (8)), and loss coefficients are fully specified. The synthetic prior composition and training schedule are described in Section 3.5 and Section B. All evaluation datasets are publicly available through OpenML (Table 4). The 5-fold by 5-repeat protocol, preprocessing pipeline, and baseline-specific evaluation settings are detailed in Section 4 and Table 1.

Ethics statement. TabLLE is evaluated exclusively on public benchmarks with no personally identifiable information. The native explanations provide per-feature attributions that may inform decisions in high-stakes domains. We emphasize that these attributions reflect the model's learned computation, and their alignment with causal mechanisms depends on the data-generating process and should be validated by domain experts before deployment in consequential settings.

AI use statement. We used generative AI tools to edit the manuscript for readability and to assist with repository packaging. We did not use generative AI tools to generate theoretical claims, proofs, or experimental results. We reviewed all AI-assisted text and take responsibility for the final content.

REFERENCES

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4699–4711, 2021a.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Md Atik Ahamed and Qiang Cheng. MambaTab: A plug-and-play model for learning tabular data. In *2024 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018a. Presented at the ICML 2018 Workshop on Human Interpretability in Machine Learning (WHI).

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 7775–7784, 2018b.

Sercan O. Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6679–6687, 2021.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7499–7519, 2024. doi: 10.1109/TNNLS.2022.3229161. Published online in 2022.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Chun-Hao Chang, Rich Caruana, and Anna Goldenberg. NODE-GAM: Neural generalized additive model for interpretable deep learning. In *International Conference on Learning Representations*, 2022.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

David R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958. doi: 10.1111/j.2517-6161.1958.tb00292.x.

Felix den Breejen, Sangmin Bae, Stephen Cha, and Se-Young Yun. Fine-tuned in-context learning transformers are excellent tabular data classifiers. *arXiv preprint arXiv:2405.13396*, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. TabR: Tabular deep learning meets nearest neighbors. In *International Conference on Learning Representations*, 2024.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520, 2022.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations (ICLR)*, 2017.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Arlind Kadra, Sebastian Pineda Arango, and Josif Grabocka. Interpretable mesomorphic networks for tabular data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017.

Myung Jun Kim, Léo Grinsztajn, and Gaël Varoquaux. CARTE: Pretraining and transfer for tabular learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23843–23866. PMLR, 2024.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. In *International Conference on Learning Representations*, 2020.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Luan Borges Teodoro Reis Sena and Francisco Galuppo Azevedo. Real-time explanations for tabular foundation models. *arXiv preprint arXiv:2603.29946*, 2026.

Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. AutoInt: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1161–1170, 2019. doi: 10.1145/3357384.3357925.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Han-Jia Ye, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later. *arXiv preprint arXiv:2407.03257*, 2024.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 10967–10978, 2019.

A Datasets and Baselines

Table 4 lists the 20 OpenML benchmarks. Table 5 lists the 22 baselines included in the comparison.

Table 4: Datasets used in the main evaluation.

#	Dataset	n	M	K
1	jannis	83,733	54	4
2	waveform-5000	5,000	40	3
3	ionosphere	351	34	2
4	iris	150	4	3
5	sylvine	5,124	20	2
6	wine-quality-white	4,898	11	7
7	dresses-sales	500	33	2
8	steel-plates-fault	1,941	27	7
9	kc1	2,109	21	2
10	wine	178	13	3
11	blood-transfusion	748	4	2
12	phoneme	5,404	5	2
13	Australian	690	14	2
14	titanic	1,309	9	2
15	higgs	98,050	28	2
16	credit-g	1,000	59	2
17	segment	2,310	18	7
18	bank-marketing	45,211	26	2
19	give	150,000	10	2
20	breast-w	699	9	2

Table 5: Baseline models included in the comparison.

Abbreviation	Full name
TabICLv2	Tabular In-Context Learning v2 (Qu et al., 2026)
XGBoost	Extreme Gradient Boosting (Chen & Guestrin, 2016)
LightGBM	Light Gradient Boosting Machine (Ke et al., 2017)
RF	Random Forest (Breiman, 2001)
EBM	Explainable Boosting Machine (Nori et al., 2019)
TabForestPFN	Tab. Forest Prior-data Fitted Network (den Breejen et al., 2024)
TabPFN-kNN	TabPFN with k-Nearest Neighbors (Hollmann et al., 2023)
ModernNCA	Modern Neighborhood Component Analysis (Ye et al., 2024)
TabM	Parameter-Efficient Tabular Ensemble (Gorishniy et al., 2025)
LR	Logistic Regression (Cox, 1958)
NAM	Neural Additive Model (Agarwal et al., 2021a)
NODE	Neural Oblivious Decision Ensembles (Popov et al., 2020)
IMN	Interpretable Mesomorphic Networks (Kadra et al., 2024)
NodeGAM	NODE Generalized Additive Model (Chang et al., 2022)
NodeGA2M	NODE GAM Order 2 (Chang et al., 2022)
MambaTab	Mamba for Tabular Data (Ahamed & Cheng, 2024)
TabR	Retrieval-Augmented Tabular DL (Gorishniy et al., 2024)
FT-Trans.	Feature Tokenizer Transformer (Gorishniy et al., 2021)
TabNet	Attentive Tabular Learning (Arik & Pfister, 2021)
ShapPFN	Shapley-Regularized PFN (Sena & Azevedo, 2026)
DT	Decision Tree (Breiman et al., 1984)
AutoInt	Auto Feature Interaction (Song et al., 2019)

B Pre-training Details

Checkpoint. All main results use B-Broad-90k. The model has 31.16M parameters (details in Table 6). Pre-training proceeds in three phases: (1) steps 0–40k at lr 8×10^{-4} on the base MLP-SCM/Tree-SCM mixture, (2) steps 40k–60k at lr 10^{-5}, (3) steps 60k–90k under the Broad prior adding DSRF, tree-ensemble priors, and function-type priors.

Preprocessing. The preprocessing is fit on the support set and applied to queries. Non-binary numerics: median-filled, standardized. Binary numerics: mode-filled, unscaled. Categoricals: mode-filled, one-hot encoded.

Table 6: Model architecture and parameter breakdown (B-Broad-90k). E and D are the column-token and row-state widths; A and H are the FCA attention and hidden widths.

Module	Structure	Params
$\mathrm{TF}_{\mathrm{col}}$	3 blocks, 8 heads, E=128, 128 inducers	0.88M
$\mathrm{TF}_{\mathrm{row}}$	3 blocks, 8 heads, 4 CLS, D=512, RoPE	0.4M
$\mathrm{TF}_{\mathrm{icl}}$	12 blocks, 8 heads, D=512, FFN 1024	25.7M
FCA	cross-attn, A=256, H=128	0.18M
$\mathcal{H}_W$	d_q=512, d_f=128, H=512, gated	2.7M
$\mathcal{H}_b$	d_q=d_t=512, H=512	1.3M
Total		**31.16M**

C Pre-training Dynamics

Figure 8 shows training loss and synthetic probe AUC for the Control (no prior expansion after 60k) and Broad branches. The Control branch plateaus at probe AUC 0.63. The Broad branch climbs after expansion, confirming gains from prior diversity.

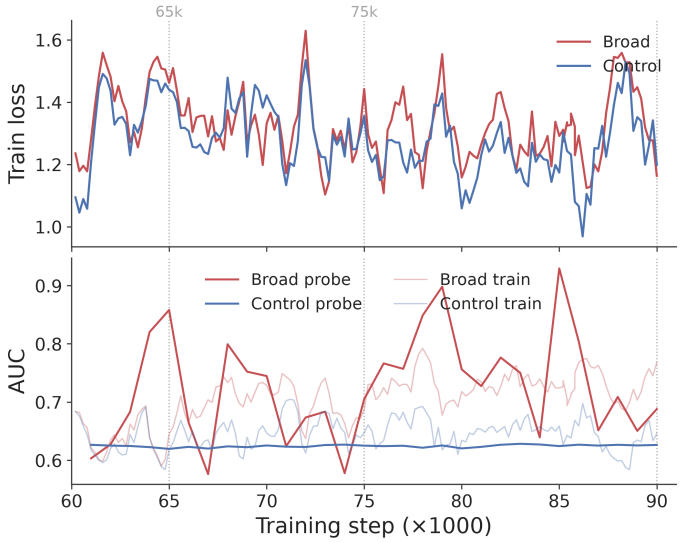


Figure 8: Training loss and probe AUC for Control and Broad branches (steps 60k–90k). The Broad branch probe AUC rises after prior expansion while Control stays flat.

D Local Linear Transferability

We probe the effective radius of each local linear explanation by extracting parameters $(\mathbf{W}_A^*, \mathbf{b}_A^*)$ at an anchor A and applying them to a nearby point B. The fraction of space achieving decision match ≥ 0.8 is 79.5% on SPARSE LINEAR, 68.8% on PIECEWISE, and 33.1% on PLATEAU. The effective radius adapts to boundary complexity: wide where the boundary is smooth, narrow where it bends sharply.

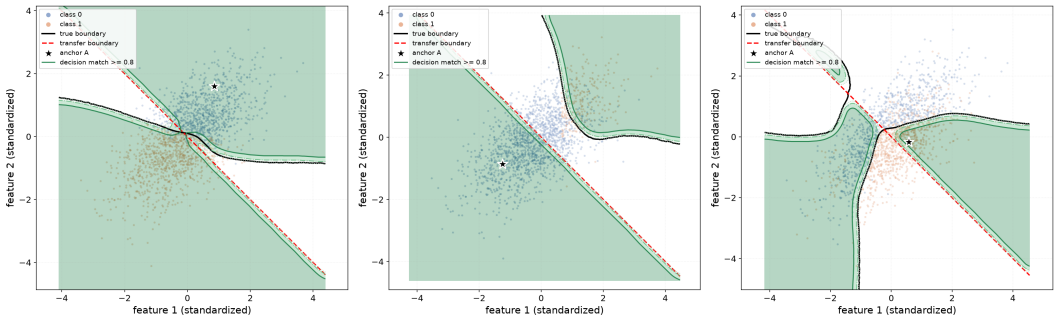


Figure 9: Local linear transferability on 2D tasks. Green regions satisfy decision match ≥ 0.8.

E XAI-Bench Metric Definitions

The four masking-based metrics reported in Table 2 follow the XAI-Bench protocol (Liu et al., 2021), adapted to a frozen in-context model. Masking is applied in raw query space: selected feature values are replaced by the corresponding column mean of the query batch, after which the support-only standardisation is re-applied and the complete model is re-run; no supervised re-training is performed. The benchmark comprises 15 synthetic tasks (three data-generating families $\times$ $\rho \in \{0, 0.25, 0.50, 0.75, 0.90\}$), and all metrics use the binary margin $m(\mathbf{x}) = z_1(\mathbf{x}) - z_0(\mathbf{x})$ and are averaged over query rows. Let $\mathbf{w} \in \mathbb{R}^M$ collect a method's attributions for a query $\mathbf{x}$.

- **Faithfulness** (Alvarez-Melis & Jaakkola, 2018b). Ablating each feature j (replacing it with its batch column mean) yields effects $e_j = |m(\mathbf{x}) - m(\mathbf{x}_{-j})|$; the score is the Pearson correlation between $|w_j|$ and e_j.
- **Infidelity** (Yeh et al., 2019). For Gaussian perturbations $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\, 0.1^2\mathbf{I})$, the score is $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left(m(\mathbf{x}) - m(\mathbf{x} - \boldsymbol{\epsilon}) - \boldsymbol{\epsilon}^\top \mathbf{w}\right)^2\right]$ with 32 samples per query.
- **Top-3 Removal.** Masking the three features with the largest $|w_j|$ yields the absolute margin change $|m(\mathbf{x}) - m(\mathbf{x}_{\text{mask}})|$. This is a PFN-adapted ROAR protocol (Hooker et al., 2019): the pre-trained model is frozen and no re-training follows removal.
- **Stability** (Alvarez-Melis & Jaakkola, 2018a). The score is the Pearson correlation between $\mathbf{w}(\mathbf{x})$ and the attributions recomputed on a Gaussian-perturbed copy $\mathbf{x} + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0},\, 0.01^2\mathbf{I})$.

F kNN Local Support Inference

We evaluate per-query k-nearest neighbor local support across $k \in \{16, 32, 64, 128, 256, 512, 1024, 2048\}$ on four datasets, comparing TabLLE against official TabICLv2 v2.0.3 under the same 5-fold $\times$ 5-repeat protocol with identical kNN indices. Table 7 reports mean AUC over 25 folds.

Table 7: kNN local-support AUC (mean over 25 folds). Bold marks the best kNN result per dataset per model.

Dataset	Model	k=16	k=32	k=64	k=128	k=256	k=512	k=1024	k=2048
bank-mkt.	TabLLE	.864	.877	.889	.898	.905	.909	.914	**.918**
	TabICLv2	.802	.844	.890	.913	.923	.930	.935	**.938**
phoneme	TabLLE	.955	.957	.959	.959	.960	**.961**	.960	.956
	TabICLv2	.946	.956	.967	.971	.973	.975	**.976**	.976
waveform	TabLLE	.946	.950	.958	.966	.970	.972	.973	**.974**
	TabICLv2	.868	.869	.929	.963	.971	.974	.975	**.976**
wine-q-w	TabLLE	.819	.829	.837	.845	.845	**.847**	.843	.844
	TabICLv2	.821	.848	.869	.875	.887	.901	.904	**.908**

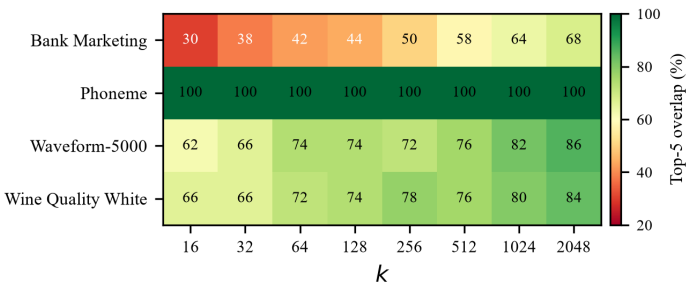


Figure 10: Top-5 feature attribution overlap (%) between kNN and full support. Higher values indicate more stable explanations. Overlap increases with k but remains below 100% in most cells, confirming context dependence of native contributions.



(a) AUC. CD=7.75, α=0.05.

(b) Accuracy. CD=7.75, α=0.05.

Figure 11: Critical-difference summaries for the 23-method comparison. Lower rank is better; TabLLE is highlighted in purple.

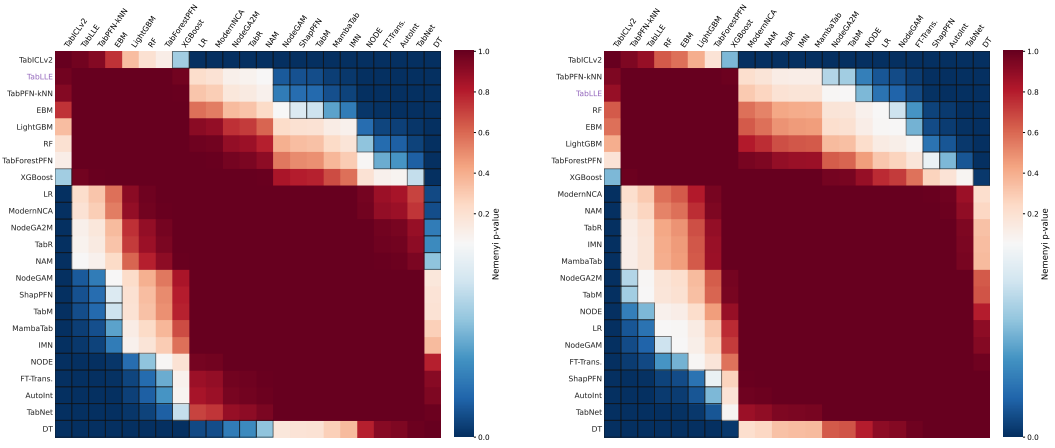
Explanation stability under context change. We track how native explanations shift when the support context changes from full to kNN. For 10 anchor queries per dataset (split 0, all 9 modes), we measure the overlap between the top-5 attributed features under kNN and under full support. Figure 10 shows that overlap increases with k but rarely reaches 100%, confirming that native contributions are genuinely context-conditioned. On bank-marketing, no prediction flips occur across any k, yet top-5 overlap at k=32 is only 38%, demonstrating that predictions can remain stable while underlying attributions shift. This connects directly to the parameter drift analysis in Section 4.2: reducing the support set re-estimates $\mathbf{W}^*$ and $\mathbf{b}^*$ through $\mathbf{h}^*$, altering per-feature contributions even when the final class decision is preserved.

G Statistical Significance

Across 23 methods and 20 datasets, Friedman tests reject performance equivalence with $p < 10^{-15}$ for both metrics. Under the Nemenyi post-hoc test, TabLLE attains the 2nd highest overall rank in AUC (average rank 5.38) and 3rd in accuracy (6.28). TabLLE significantly outperforms 10 baselines on AUC and eight on accuracy, establishing statistical advantages over both established deep tabular architectures (such as FT-Transformer, TabNet, NODE, and TabM) and interpretable baselines (such as NodeGAM and IMN). Crucially, no baseline achieves a statistically significant advantage over TabLLE on either metric.

H Per-Dataset Benchmark Results

Full per-dataset AUC and accuracy (mean $\pm$ std over 25 folds) for all 23 methods appear in Tables 8 to 11.



(a) AUC. (b) Accuracy.

Figure 12: Nemenyi p-value matrices for the 23-method comparison. Black-edged cells indicate $p < 0.05$.

Table 8: Per-dataset AUC (mean ± std), datasets 1–10.

Dataset	TabICLv2	TabLLE	XGBoost	LightGBM	RF	EBM	TabForestPFN	TabPFN-kNN	ModernNCA	TabM	LR	NAM	NODE	IMN	NodeGAM	NodeGA2M	MambaTab	TabR	FT-Trans.	TabNet	ShapPFN	DT	AutoInt
jannis	.912±.000	.877±.000	.873±.001	.873±.001	.854±.001	.859±.000	.875±.001	.840±.000	.853±.000	.812±.002	.814±.000	.850±.000	.781±.001	.855±.001	.822±.001	.827±.001	.816±.001	.874±.000	.805±.002	.865±.001	.749±.001	.741±.004	.806±.002
waveform-5000	.976±.000	.974±.000	.967±.001	.968±.001	.967±.001	.972±.000	.970±.000	.972±.000	.920±.002	.945±.001	.971±.000	.959±.000	.954±.002	.960±.001	.950±.007	.962±.001	.964±.002	.966±.001	.948±.001	.949±.003	.961±.000	.895±.004	.955±.001
ionosphere	.986±.001	.985±.001	.967±.007	.975±.007	.980±.001	.976±.004	.957±.007	.986±.001	.983±.001	.985±.002	.920±.007	.972±.005	.943±.008	.954±.013	.970±.006	.962±.007	.953±.007	.927±.010	.962±.012	.889±.015	.959±.004	.920±.005	.947±.011
iris	.998±.001	.998±.001	.978±.006	.983±.004	.994±.001	.993±.004	.993±.002	.997±.001	.993±.002	.989±.006	.999±.001	.989±.003	.998±.001	.982±.004	.976±.018	.985±.011	.997±.002	.409±.055	.983±.006	.932±.020	.997±.001	.977±.006	.991±.003
sylvine	.997±.000	.986±.001	.985±.000	.986±.001	.981±.001	.985±.000	.990±.001	.978±.001	.971±.001	.972±.002	.965±.000	.967±.001	.970±.001	.978±.000	.969±.003	.977±.002	.976±.002	.977±.001	.979±.001	.968±.000	.956±.000	.955±.003	.978±.001
wine-quality-white	.913±.001	.851±.002	.838±.002	.833±.007	.832±.013	.808±.004	.878±.005	.845±.006	.805±.008	.807±.004	.752±.006	.777±.005	.791±.007	.730±.009	.796±.005	.796±.002	.798±.008	.724±.016	.772±.014	.718±.012	.677±.002	.682±.023	.764±.019
dresses-sales	.616±.016	.609±.011	.561±.011	.572±.008	.598±.009	.625±.012	.564±.027	.578±.011	.514±.014	.518±.010	.594±.011	.596±.020	.517±.013	.518±.013	.588±.007	.599±.004	.545±.012	.557±.019	.528±.020	.520±.040	.621±.003	.500±.000	.513±.019
steel-plates-fault	.976±.001	.962±.001	.965±.001	.965±.001	.961±.001	.895±.070	.964±.002	.959±.001	.945±.002	.952±.002	.934±.001	.948±.001	.953±.001	.943±.003	.925±.000	.940±.003	.929±.001	.940±.003	.950±.002	.913±.004	.915±.001	.909±.002	.950±.001
kc1	.857±.005	.817±.003	.795±.009	.789±.006	.827±.002	.812±.006	.804±.006	.827±.002	.817±.002	.791±.007	.803±.002	.802±.002	.771±.009	.756±.007	.794±.021	.796±.018	.795±.002	.803±.002	.755±.009	.779±.012	.804±.001	.770±.005	.740±.012
wine	1.000±.000	1.000±.000	.997±.000	.999±.001	.999±.001	1.000±.000	.999±.001	1.000±.000	1.000±.000	1.000±.000	1.000±.000	.999±.001	1.000±.000	.996±.003	1.000±.001	.999±.001	.998±.001	.545±.056	.999±.001	.902±.049	1.000±.000	.954±.011	.999±.001

Table 9: Per-dataset AUC (mean ± std), datasets 11–20.

Dataset	TabICLv2	TabLLE	XGBoost	LightGBM	RF	EBM	TabForestPFN	TabPFN-kNN	ModernNCA	TabM	LR	NAM	NODE	IMN	NodeGAM	NodeGA2M	MambaTab	TabR	FT-Trans.	TabNet	ShapPFN	DT	AutoInt
blood-transfusion	.756±.005	.750±.004	.692±.010	.712±.008	.683±.011	.743±.004	.733±.010	.755±.003	.717±.011	.705±.012	.753±.003	.734±.008	.716±.007	.715±.037	.742±.005	.757±.013	.745±.005	.736±.006	.650±.016	.723±.024	.751±.003	.619±.018	.653±.017
phoneme	.976±.001	.953±.001	.954±.001	.953±.001	.962±.001	.948±.001	.957±.002	.959±.001	.943±.000	.943±.002	.813±.000	.898±.001	.944±.002	.916±.001	.869±.001	.872±.007	.949±.001	.950±.002	.942±.003	.922±.004	.858±.000	.901±.004	.937±.003
Australian	.942±.004	.936±.003	.935±.001	.933±.003	.939±.003	.942±.003	.913±.005	.937±.003	.916±.005	.904±.005	.932±.002	.928±.004	.896±.005	.913±.005	.936±.007	.935±.008	.910±.003	.920±.003	.899±.006	.874±.020	.923±.001	.910±.012	.911±.005
titanic	.881±.001	.855±.004	.853±.003	.858±.005	.867±.005	.860±.003	.833±.009	.855±.004	.835±.005	.827±.004	.841±.002	.836±.002	.819±.002	.841±.007	.835±.006	.836±.008	.845±.005	.839±.004	.815±.007	.822±.008	.834±.001	.837±.006	.801±.009
higgs	.823±.000	.795±.000	.800±.001	.804±.000	.794±.000	.793±.000	.801±.001	.766±.001	.784±.001	.754±.002	.682±.000	.771±.000	.727±.001	.761±.002	.694±.011	.673±.009	.738±.001	.799±.001	.742±.002	.799±.003	.712±.002	.676±.005	.746±.001
credit-g	.798±.003	.770±.002	.769±.010	.774±.008	.785±.004	.789±.006	.743±.010	.792±.007	.756±.009	.753±.006	.791±.003	.765±.006	.735±.006	.654±.032	.745±.015	.743±.008	.711±.007	.757±.004	.699±.017	.588±.040	.753±.002	.694±.006	.711±.005
segment	.999±.000	.998±.000	.999±.000	.999±.000	.998±.000	.999±.000	.999±.000	.999±.000	.998±.000	.998±.000	.996±.000	.998±.000	.998±.000	.994±.001	.997±.000	.998±.000	.998±.000	.995±.000	.997±.000	.995±.001	.989±.000	.990±.001	.998±.000
bank-marketing	.941±.001	.907±.000	.931±.001	.935±.000	.925±.001	.935±.000	.934±.001	.906±.001	.897±.001	.874±.002	.891±.000	.880±.001	.795±.002	.895±.001	.892±.005	.886±.004	.868±.005	.932±.001	.895±.003	.913±.003	.868±.001	.782±.013	.890±.002
give	.868±.000	.859±.000	.856±.001	.865±.000	.838±.001	.864±.000	.860±.001	.859±.000	.856±.001	.839±.001	.849±.000	.861±.000	.731±.004	.857±.000	.855±.001	.850±.002	.744±.006	.864±.000	.804±.003	.859±.001	.840±.001	.710±.017	.836±.001
breast-w	.995±.001	.993±.001	.990±.001	.992±.000	.991±.002	.994±.001	.992±.002	.995±.001	.992±.001	.992±.001	.995±.000	.989±.001	.989±.002	.994±.000	.993±.005	.993±.005	.988±.002	.989±.002	.989±.002	.987±.003	.993±.000	.970±.005	.989±.001

Table 10: Per-dataset accuracy (mean ± std), datasets 1–10.

Dataset	TabICLv2	TabLLE	XGBoost	LightGBM	RF	EBM	TabForestPFN	TabPFN-kNN	ModernNCA	TabM	LR	NAM	NODE	IMN	NodeGAM	NodeGA2M	MambaTab	TabR	FT-Trans.	TabNet	ShapPFN	DT	AutoInt
jannis	.769±.001	.724±.000	.716±.000	.715±.000	.704±.001	.690±.000	.718±.001	.676±.001	.693±.001	.656±.002	.644±.000	.680±.001	.625±.002	.694±.001	.651±.001	.664±.000	.673±.003	.719±.001	.650±.002	.709±.001	.549±.001	.588±.006	.641±.002
waveform-5000	.873±.001	.868±.001	.852±.003	.853±.005	.851±.003	.865±.001	.856±.003	.859±.001	.800±.002	.810±.001	.867±.001	.830±.006	.831±.004	.843±.002	.825±.017	.847±.009	.842±.004	.843±.002	.821±.003	.818±.004	.812±.001	.746±.007	.826±.003
ionosphere	.951±.004	.944±.003	.921±.002	.937±.004	.937±.006	.933±.010	.912±.012	.955±.007	.940±.007	.939±.004	.880±.003	.920±.003	.910±.005	.923±.010	.926±.011	.906±.020	.921±.011	.820±.015	.919±.011	.842±.018	.920±.004	.891±.008	.906±.007
iris	.964±.009	.956±.003	.939±.006	.940±.007	.953±.009	.953±.007	.959±.011	.972±.005	.957±.007	.956±.008	.971±.003	.959±.007	.956±.011	.945±.008	.907±.053	.920±.013	.963±.009	.339±.062	.956±.005	.739±.025	.957±.003	.948±.014	.952±.007
sylvine	.979±.001	.948±.001	.947±.001	.949±.001	.936±.001	.950±.001	.961±.001	.935±.001	.916±.002	.924±.002	.912±.000	.922±.001	.920±.001	.933±.001	.923±.006	.933±.008	.935±.004	.938±.002	.941±.002	.919±.002	.915±.000	.906±.005	.938±.002
wine-quality-white	.702±.002	.650±.003	.673±.003	.673±.004	.686±.003	.607±.003	.671±.004	.631±.002	.584±.003	.583±.010	.537±.001	.526±.004	.580±.008	.566±.002	.537±.006	.546±.007	.565±.007	.583±.004	.615±.009	.533±.004	.518±.002	.508±.004	.628±.006
dresses-sales	.619±.011	.590±.013	.565±.010	.563±.008	.602±.017	.611±.007	.565±.027	.584±.008	.533±.018	.540±.019	.580±.000	.584±.019	.537±.012	.530±.006	.600±.004	.618±.002	.558±.014	.582±.006	.533±.012	.566±.022	.632±.008	.580±.000	.524±.009
steel-plates-fault	.834±.003	.779±.006	.797±.009	.798±.007	.777±.006	.693±.099	.797±.006	.765±.004	.758±.004	.764±.005	.717±.003	.750±.001	.758±.004	.754±.003	.697±.015	.726±.016	.719±.004	.743±.009	.760±.007	.685±.007	.631±.004	.688±.008	.755±.006
kc1	.866±.002	.859±.001	.855±.003	.858±.005	.860±.002	.858±.002	.852±.009	.862±.002	.859±.002	.858±.002	.726±.002	.859±.002	.862±.003	.858±.004	.854±.003	.857±.008	.858±.002	.855±.003	.849±.007	.848±.003	.802±.002	.847±.003	.846±.002
wine	.994±.004	.987±.003	.957±.008	.972±.012	.979±.004	.980±.006	.974±.007	.994±.004	.980±.005	.979±.004	.981±.005	.971±.002	.977±.004	.971±.007	.972±.017	.961±.017	.971±.007	.323±.027	.974±.009	.804±.073	.975±.004	.926±.012	.965±.011

Table 11: Per-dataset accuracy (mean ± std), datasets 11–20.

Dataset	TabICLv2	TabLLE	XGBoost	LightGBM	RF	EBM	TabForestPFN	TabPFN-kNN	ModernNCA	TabM	LR	NAM	NODE	IMN	NodeGAM	NodeGA2M	MambaTab	TabR	FT-Trans.	TabNet	ShapPFN	DT	AutoInt
blood-transfusion	.796±.003	.798±.004	.741±.008	.754±.008	.745±.008	.782±.003	.788±.006	.793±.001	.775±.008	.764±.008	.774±.001	.788±.009	.783±.007	.758±.045	.769±.004	.794±.005	.800±.004	.762±.002	.738±.009	.779±.007	.780±.003	.762±.000	.732±.010
phoneme	.922±.003	.895±.001	.897±.002	.895±.002	.907±.002	.884±.001	.903±.003	.902±.001	.875±.002	.876±.002	.726±.001	.822±.002	.886±.005	.842±.003	.804±.007	.793±.013	.892±.002	.898±.002	.886±.006	.854±.006	.790±.001	.840±.002	.894±.004
Australian	.873±.005	.859±.005	.869±.006	.861±.004	.875±.005	.875±.003	.853±.009	.864±.005	.846±.008	.834±.007	.867±.006	.857±.004	.831±.006	.858±.009	.868±.010	.874±.013	.844±.008	.859±.007	.836±.008	.824±.019	.852±.005	.846±.014	.838±.003
titanic	.823±.004	.812±.004	.798±.004	.803±.005	.808±.009	.806±.002	.788±.004	.807±.005	.786±.002	.781±.003	.793±.002	.797±.007	.783±.002	.806±.003	.781±.008	.790±.003	.798±.004	.794±.006	.770±.004	.776±.013	.774±.003	.779±.008	.760±.008
higgs	.741±.000	.719±.001	.722±.001	.724±.000	.717±.000	.715±.000	.722±.001	.693±.001	.709±.001	.684±.001	.637±.000	.698±.000	.662±.001	.692±.001	.640±.011	.630±.008	.672±.002	.721±.001	.674±.002	.721±.002	.648±.002	.637±.003	.682±.001
credit-g	.764±.002	.733±.002	.745±.005	.751±.004	.763±.003	.766±.006	.728±.009	.758±.006	.729±.009	.739±.007	.739±.002	.746±.003	.734±.007	.690±.022	.735±.007	.724±.022	.712±.010	.723±.005	.694±.010	.679±.027	.687±.003	.692±.008	.700±.007
segment	.983±.002	.964±.001	.978±.001	.980±.002	.972±.002	.971±.002	.972±.002	.969±.002	.965±.001	.964±.003	.953±.002	.963±.003	.965±.002	.950±.003	.956±.007	.960±.003	.967±.003	.940±.001	.966±.003	.946±.006	.878±.003	.947±.003	.967±.001
bank-marketing	.911±.001	.899±.000	.906±.001	.908±.001	.904±.001	.910±.001	.908±.001	.901±.000	.889±.001	.876±.002	.831±.000	.849±.002	.878±.002	.900±.000	.898±.001	.897±.001	.879±.002	.907±.001	.897±.001	.902±.001	.889±.001	.894±.001	.870±.004
give	.937±.000	.937±.000	.935±.000	.937±.000	.935±.000	.937±.000	.936±.000	.937±.000	.935±.000	.931±.000	.795±.000	.937±.000	.921±.001	.937±.000	.936±.000	.936±.000	.918±.001	.937±.000	.925±.001	.936±.001	.936±.000	.934±.000	.932±.000
breast-w	.968±.002	.970±.002	.956±.004	.960±.003	.965±.002	.963±.004	.960±.005	.966±.002	.961±.003	.959±.002	.966±.001	.954±.003	.956±.003	.967±.003	.953±.022	.957±.026	.957±.003	.947±.012	.956±.004	.945±.013	.972±.001	.933±.004	.957±.005