Distilling Tabular Foundation Models for Structured Health Data

Anonymous Authors[1]

Abstract

Tabular foundation models (TFMs) achieve strong performance on health datasets, but their inference cost and infrastructure requirements limit practical use. We study whether their predictive behavior can be transferred to lightweight tabular models through knowledge distillation. Since in-context TFMs condition on the training set at inference time, naive distillation can introduce context leakage; we address this with stratified out-of-fold teacher labeling. Across 19 healthcare datasets, 6 TFM teachers, 4 student families, and several multi-teacher ensembles, We find that distilled students retain at least 90% of teacher AUC outperforming teachers in some cases while running at least 26× faster on CPU, preserving calibration and fairness critical for health applications. Moreover, multi-teacher averaging does not consistently improve over the best single teacher. Thus leakage-aware distillation is a viable route for bringing TFM-quality predictions into inference-constrained health settings.

1. Introduction

Structured data is central to health machine learning (Rajpurkar et al., 2022). Tasks such as risk stratification, disease screening, readmission prediction, and mortality estimation are often defined over EHR, laboratory values, demographics, and other tabular variables (Johnson et al., 2016). These datasets are frequently small, heterogeneous, and sparse, making data-efficient prediction essential. Healthcare deployment also requires more than high accuracy. Hospital systems are often CPU-based, air-gapped, and constrained by data-residency policies, limiting the use of cloud-GPU inference. Models must produce calibrated probabilities that support clinical decisions (Guo et al., 2017) and avoid amplifying demographic disparities that can affect access to care (Obermeyer et al., 2019; Hardt et al., 2016).

[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.

Tabular foundation models (TFMs) have recently emerged as a promising approach for such clinical datasets. Models such as TabPFNv2.5 (Grinsztajn et al., 2025), TabICLv2 (Qu et al., 2026), TabDPT (Ma et al., 2025), LimiX (Zhang et al., 2025), Orion-MSP (Bouadi et al., 2025)often match or outperform tuned gradient-boosted trees on datasets with fewer than ten thousand rows without per-task tuning. However, their inference paradigm requires conditioning on a task-specific context and running large GPU-based models, conflicting with the constraints above.

We therefore ask whether TFMs predictive behavior, including accuracy, calibration, and fairness, can be transferred to standard ML models that are easier to deploy. We study this question through knowledge distillation (Hinton et al., 2015) from TFMs into lightweight students. Since TFM teachers condition on labeled training examples at inference time, teacher outputs for examples included in the context can produce biased soft targets (Mansurov et al., 2024). We mitigate this using stratified out-of-fold teacher labeling. We run extensive experiments across 19 health datasets, (16 binary, 3 multiclass), using 6 TFMs as teachers, 4 standard models: LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), XGBoost (Chen & Guestrin, 2016), and MLP, as students, and 4 multi-teacher ensembles. We find that: (i) distilled students retain more than 90% of teacher AUC, outperforming teachers on some datasets; (ii) MLP students are 2× faster than LightGBM but are miscalibrated and amplify fairness gaps; (iii) Multi-teacher setting tend to be outperformed by the best single teacher.

We summarize our contributions as follows:

- We study knowledge distillation as a practical route for transferring TFM predictions into deployable healthcare tabular models.
- We propose a stratified out-of-fold teacher-labeling to prevent *teacher identity leakage* from ICL-based TFMs.
- We benchmark TFM distillation across 19 healthcare datasets, 6 teachers, 4 student families, and multi-teacher ensembles.
- We show that tree-based students provide the strongest accuracy–calibration–fairness tradeoff, while MLP distillation and naive multi-teacher averaging are less consistently beneficial.

2. Methodology

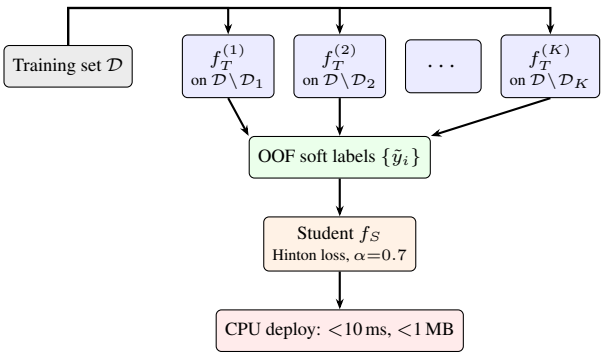

Figure 1. Out-of-fold (OOF) distillation pipeline. The training set is split into K=5 stratified folds; teacher $f_T^{(k)}$ is fit on $\mathcal{D}\setminus\mathcal{D}_k$ and produces soft labels only for $\mathcal{D}_k$, so no sample is ever scored by a teacher that conditioned on it. The student f_S is trained on the resulting leakage-free soft labels and deployed on CPU.

Formalization. In this work, we consider a set of tabular datasets $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbf{R}^{l_i \times c_i}$ and $y_i \in \mathbf{R}^{l_i}$, l_i is the sequence length and the c_i is the number of features for the i-th dataset. We also consider a Tabular Foundation Model (TFM) as a teacher f_T that predicts the labels using a function $f_T(.) : \mathbf{R}^{l \times c} \to \mathbf{R}^l$.

Given a set of datasets $\mathcal{D}$ and a TFM f_T, we aim to find a student $f_S : \mathbf{R}^{l \times c} \to \mathbf{R}^l$ that approximates f_T without duplicating teacher's weights, gradients, or hidden states.

Distillation approach. To transfer the predictive behavior of f_T to f_S, we employ knowledge distillation (Hinton et al., 2015), training f_S to match the outputs of f_T rather than hard ground-truth labels. This choice is motivated by recent empirical works showing that distillation leads to stronger student generalization (Tang et al., 2020; Zhou et al., 2021).

We thus define the student objective as the Hinton mixed loss:

$$\mathcal{L} = \alpha \sum_i w_i T_i^2 \, \mathrm{KL}(\hat{p}_i^{T_i} \| q_i^{T_i}) + (1-\alpha) \sum_i w_i \, \ell_{\mathrm{CE}}(y_i, q_i), \tag{1}$$

with α=0.7. For 3 students, the soft loss reduces to a per-sample weighted MSE on soft-label logits.

Adaptive temperature. Per-sample $T_i \in [T_{\min}, T_{\max}]$ scales each $\tilde{p}_i$ as a function of teacher entropy $H(\tilde{p}_i)$: confident predictions get $T_i \approx T_{\min}$=1, ambiguous ones get $T_i \approx T_{\max}$=5, following Guo et al. (2017). Confidence weight $w_i = \exp(-(H(\tilde{p}_i) - \mu)^2 / 2\sigma^2)$ peaks on moderate-entropy predictions ((μ, σ)=(0.7, 0.2)).

Teacher soft labels. In our work, we consider that all teachers are based on TFMs. While all these models rely on in-context learning for predictions, the teachers condition on the training set as part of their input context. More formally, we denote the teacher as $f_T(\cdot \mid C)$ where $C = \{x_i, y_i\}$ is the context set. Thus, predictions over the samples x_i are produced by the teacher that has already observed their targets y_i, yielding soft labels that are overconfident compared to the predicted distribution on unseen data.

To mitigate this, we apply a $k = 5$ stratified cross-validation, fitting each teacher $f_T^{(k)}$ on $\mathcal{D}/\mathcal{D}_k$ and predicting only over $\mathcal{D}_k$ ensuring that no sample's soft target is produced by f_T that conditioned on it. For an ensemble of M teachers we average the per-fold predictions as shown in Figure 1.

3. Experiments

3.1. Setup

Datasets. We used 19 health datasets from which 16 are binary and 3 multiclass, covering cardiology, oncology, nephrology, dermatology, and critical care, with sample sizes from 299 to 9105 and feature counts from 4 to 40. We refer the reader to Table 6 for more details.

Teacher TFMs. We consider 6 TFMs: TabPFNv2.5 (Hollmann et al., 2025), TabPFNv2.6 (Grinsztajn et al., 2025), TabICLv2 (Qu et al., 2026), TabDPT (Ma et al., 2025), LimiX (Zhang et al., 2025), and Orion-MSP v1.5 (Bouadi et al., 2025), each with its built-in missing-value handler. For multi-teacher (M>1) settings, we average per-fold soft labels with equal weights.

Students. We use LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), XGBoost (Chen & Guestrin, 2016) (300 trees, depth 6, patience-30 early stopping) and an MLP (embedding $\min(8d, 128)$, cosine LR with warmup, label smoothing 0.05, SWA on final 20% of training, collapse-detector restart at higher dropout).

Baselines. We use LogisticRegression, XGBoost, and LightGBM with a fixed configuration (300 trees, depth 6) on zero-imputed inputs.

Metrics. We report (i) average AUC; (ii) retention which is defined as $\frac{AUC(\text{Student})}{AUC(\text{Teacher})} \times 100$; (iii) model latency and memory consumption; (iv) calibration and fairness. All reported results are aggregated over 5 simulations. All models and runs use identical pre-processing, anonymization, and evaluation as implemented in TabTune (Tanna et al., 2025).

3.2. Empirical Results

Accuracy. Table 1 shows the avg AUC and retention on LGBM and MLP on the 16 binary datasets. We refer the reader to Table 7 for the results on the other student models. Similar results are observed on the 3 multiclass datasets and refer the reader to Table 8.

Distillation is strongly student-dependent. Tree-based students benefit most: LightGBM retains 98.8–99.6% of teacher AUC and improves over the default LightGBM base-

Table 1. Avg AUC and retention on LGBM and MLP over the binary datasets.

Type	Model	AUC	Ret.
Baseline	LogisticRegression	.857 ± .108	–
	XGBoost	.854 ± .133	–
	LightGBM	.853 ± .134	–
Teacher	LimiX	.875 ± .119	–
	TabICLv2	.873 ± .124	–
	TabPFNv2.6	.870 ± .125	–
	TabDPT	.868 ± .126	–
	TabPFNv2.5	.868 ± .125	–
	Orion-MSP v1.5	.864 ± .124	–
Distilled (LGBM)	LimiX→LGBM	.865 ± .119	98.8%
	TabICLv2→LGBM	.863 ± .122	98.9%
	TabPFNv2.6→LGBM	.862 ± .123	99.1%
	TabPFN→LGBM	.862 ± .124	99.4%
	TabDPT→LGBM	.861 ± .123	99.1%
	Orion-MSP→LGBM	.860 ± .119	99.6%
Distilled (LGBM, multi-teacher)	[PFN+OrionMSP+Limix]→LGBM	.864 ± .121	–
	[PFN+Limix]→LGBM	.864 ± .121	–
	[PFN+ICL+Limix]→LGBM	.863 ± .123	–
	[PFN+ICL+Limix+TabDPT]→LGBM	.862 ± .123	–
Distilled (MLP)	LimiX→MLP	.795 ± .173	90.8%
	TabPFN→MLP	.787 ± .184	90.7%
	TabPFNv2.6→MLP	.782 ± .189	89.9%
	TabICLv2→MLP	.782 ± .190	89.6%
	TabDPT→MLP	.781 ± .190	89.9%
	Orion-MSP→MLP	.775 ± .182	89.7%
Distilled (MLP, multi-teacher)	[PFN+OrionMSP+Limix]→MLP	.790 ± .181	–
	[PFN+Limix]→MLP	.787 ± .180	–
	[PFN+ICL+Limix+TabDPT]→MLP	.787 ± .182	–
	[PFN+ICL+Limix]→MLP	.785 ± .182	–

line, while Table 7 shows that CatBoost and XGBoost can reach or exceed 100% retention. This suggests that distilled students can sometimes outperform their teachers, consistent with distillation acting as a regularizer through softened targets rather than only as compression. In contrast, MLP students retain only 89.6–90.8% and remain below default baselines. Multi-teacher distillation also gives limited gains: except for CatBoost, the best multi-teacher student does not outperform the best single-teacher student, suggesting that uniform averaging can add noise when teachers disagree (Du et al., 2020).

Latency and memory. Table 2 reports latency on 6 datasets; full results are in Table 9. Students run on CPU, while TFM teachers run on GPU. Distilled LGBM and MLP students require about 7 ms and 3.8 ms, respectively, versus 187.4 ms for the fastest teacher, yielding roughly 26× and 49× lower latency while staying under 1 MB. This makes distilled students suitable for CPU-only, high-throughput healthcare deployment.

Calibration. Table 3 reports ECE and Brier scores on 12 binary datasets, with and without global temperature scaling (TS) (Guo et al., 2017; Platt et al., 1999). Distilled LGBM students preserve much of the teachers' calibration, with ECE around .058–.063, substantially better than default LightGBM and XGBoost. TS provides only small additional gains for LGBM students, suggesting that they are already reasonably calibrated. In contrast, MLP students

Table 2. Inference latency and throughput across datasets with test examples between 91 and 1821.

Model	Latency (ms)	Throughput (predictions/s)
TFM teachers (GPU)		
TabICLv2	187.4	3,221/s
LimiX	433.8	1,450/s
TabPFNv2.6	564.9	1,099/s
Orion-MSP v1.5	1,911.9	323/s
TabDPT	4,010.0	600/s
Distilled MLP students (CPU)		
LimiX→MLP	3.7	172K/s
TabPFNv2.6→MLP	3.8	170K/s
TabICLv2→MLP	3.8	169K/s
[PFNv2.6+Limix]→MLP	3.8	166K/s
Distilled LGBM students (CPU)		
LimiX→LGBM	7.0	87K/s
[PFNv2.6+ICL+Limix]→LGBM	7.1	92K/s
TabPFNv2.6→LGBM	7.3	82K/s
[PFNv2.6+Limix]→LGBM	7.7	98K/s

Table 3. Avg ECE and Brier on 12 binary datasets w/o global temperature scaling (TS).

Model	ECE ↓	ECE+TS ↓	Brier ↓	Brier+TS ↓
Baselines				
LogisticRegression	.062	.063	.101	.100
LightGBM	.090	.069	.116	.106
XGBoost	.092	.069	.116	.107
Teachers				
TabPFNv2.6	.053	.053	.095	.094
TabDPT	.056	.053	.094	.093
LimiX	.056	.055	.093	.093
TabICLv2	.056	.053	.093	.093
Distilled LGBM students (single-teacher)				
TabDPT→LGBM	.060	.057	.100	.099
TabPFNv2.6→LGBM	.061	.060	.098	.098
Distilled LGBM students (multi-teacher)				
[PFNv2.6+Limix]→LGBM	.058	.061	.099	.099
[PFNv2.6+OrionMSP+Limix]→LGBM	.063	.059	.099	.098
Distilled MLP students (single-teacher)				
TabPFNv2.6→MLP	.123	.074	.148	.133
TabDPT→MLP	.124	.072	.149	.133
Distilled MLP students (multi-teacher)				
[PFNv2.6+Limix]→MLP	.124	.076	.150	.133
[PFNv2.6+ICL+Limix+TabDPT]→MLP	.125	.075	.149	.133

are poorly calibrated before TS, with ECE above .12, but improve markedly after scaling. Thus, calibration quality is strongly student-dependent: distilled trees are reliable without much post-hoc correction, whereas MLP students require calibration.

Fairness. Table 4 reports DP- and EO-difference across 8 datasets and 4 sensitive attributes. Negative Δ values mean the student has a smaller fairness gap than its teacher. Distilled LGBM students reduce DP gaps and EO gaps in most cases, with average $\Delta\text{DP} = -0.013$ and $\Delta\text{EO} = -0.014$, consistent with soft-label smoothing in a smaller hypothesis class. MLP students reduce DP more strongly ($\Delta\text{DP} = -0.034$), but often increase EO gaps ($\Delta\text{EO} = +0.041$). Thus, LGBM provides a more stable fairness tradeoff, while MLP can improve parity in prediction rates at the cost of subgroup error disparities.

Table 4. Group fairness across 8 datasets and 4 attributes. In multi-teacher settings, Δ are computed with TabPFNv2.6.

Model	DP-diff ↓	EO-diff ↓	ΔDP	ΔEO
Baselines				
LogisticRegression	.131	.183	–	–
XGBoost	.145	.165	–	–
LightGBM	.145	.167	–	–
Teachers				
LimiX	.134	.144	–	–
TabDPT	.143	.156	–	–
TabPFNv2.6	.145	.144	–	–
TabICLv2	.154	.166	–	–
Orion-MSP v1.5	.153	.202	–	–
Distilled LGBM students (single-teacher)				
LimiX→LGBM	.130	.125	−.005	−.018
OrionMSPv1.5→LGBM	.132	.173	−.021	−.029
TabPFNv2.6→LGBM	.131	.125	−.014	−.019
TabDPT→LGBM	.136	.171	−.007	+.015
TabICLv2→LGBM	.139	.136	−.015	−.031
Distilled LGBM students (multi-teacher)				
[PFNv2.6+OrionMSP+Limix]→LGBM	.129	.130	−.016	−.014
[PFNv2.6+Limix]→LGBM	.133	.126	−.012	−.017
[PFNv2.6+ICL+Limix]→LGBM	.135	.120	−.010	−.024
Distilled MLP students (single-teacher)				
OrionMSPv1.5→MLP	.104	.216	−.049	+.014
TabICLv2→MLP	.106	.165	−.048	−.001
LimiX→MLP	.108	.187	−.026	+.043
TabPFNv2.6→MLP	.108	.162	−.036	+.019
TabDPT→MLP	.117	.178	−.026	+.022
Distilled MLP students (multi-teacher)				
[PFNv2.6+OrionMSP+Limix]→MLP	.111	.193	−.034	+.049
[PFNv2.6+ICL+Limix]→MLP	.114	.207	−.030	+.063
[PFNv2.6+Limix]→MLP	.116	.219	−.029	+.075

Component ablation. Table 5 ablates MLP distillation with TabPFNv2.6 on 5 datasets. Hard-label training outperforms the full setup by 0.034 AUC (p=0.004), while removing adaptive temperature, confidence weighting, or augmentation changes AUC by at most 0.02. Thus, for MLPs, the soft-label machinery adds little on this benchmark; we leave tree-student ablations to future work.

4. Discussion and limitations

In this section, we discuss a few limitations of our work.

Benchmark scope. All datasets have less than 10K samples where most are under 1K. This is the regime where TFMs performs well and where distillation may be useful. In the future, we aim to run experiments on specific benchmarks like: MIMIC-IV in-hospital mortality, eICU mortality, and the UCI Diabetes 130-US-Hospitals readmission dataset.

Multiclass distillation gap. TabPFNv2.5 scores 0.985 AUC on the kidney disease dataset (3 classes, 400 rows), while its distilled TabPFN→LGBM version scores only 0.749. The reason is that the distillation breaks for 3+ class problems because of the LGBM regressor.

Training cost. It is dominated by cross-validated soft-label generation, which needs K=5 teacher passes. It takes ≈ 30 min per dataset on a single A100. While the cost is one-time, the student serves at sub-3 ms latency without a GPU.

Table 5. Ablation on the MLP student with TabPFNv2.6 teacher on 5 datasets. Δ is the paired difference vs. *full*; p from a Wilcoxon signed-rank test on per-(dataset, simulation) deltas.

Configuration	AUC	Δ vs. full	p
Full (all components)	.829	–	–
No adaptive temperature	.809	−.020	.49
No confidence weighting	.828	−.001	.93
No augmentation	.818	−.011	.34
Hard labels only (α=0)	.863	+.034	.004
Soft labels only (α=1)	.814	−.014	.54
Low temperature (T=1)	.855	+.027	.06
High temperature (T=5)	.810	−.018	.19

Hard-label ablation. Table 5 shows that hard-label training (α=0) outperforms the full pipeline (α=0.7) by 0.034 AUC (p=0.004, Wilcoxon). This does not affect the need for out-of-fold labeling: for ICL teachers, scoring examples that appear in the context can produce overconfident soft targets and leakage. Instead, the result suggests that, on these small and relatively clean datasets, the added soft-label components do not improve MLP training. Soft targets may be more useful in noisier or higher-dimensional clinical settings, where hard labels are less reliable.

Position to TabPFNv2.5 distillation engine. While PriorLabs released a closed-source distillation engine (Grinsztajn et al., 2025), we could not compare our experiments using their engine. Our work differs as *it is open-source, model-agnostic across multiple TFM families, and supports multi-teacher ensembling.*

5. Conclusion

In this work, we argue that TFMs tend to well predict on small health datasets, but are costly for inference- constrained health settings. We thus propose a methodology around knowledge distillation to transfer predictive behavior from TFMs to simpler standard models, and run extensive experiments on 19 healthcare datasets. Our results show that distilled models run, on CPU, at least 26× faster than the best TFMs on GPU. While knowledge transfer recovers at least 90% of TFM knowledge, it also preserves calibration and fairness. As a practical recommendation: use an LGBM student by default, and consider an MLP student only when sub-2 ms latency is a hard requirement.

For future work, we aim to (i) explore more complex multi-teacher distillation settings, as tested ones are outperformed by the single-teacher setting; (ii) evaluate the knowledge transfer on larger EHR benchmarks using more multi-class datasets; (iii) investigate more the value of the soft-label components when a hard-label baseline is in place.

References

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*, 2025.

Chen, T. and Guestrin, C. 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.

Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., and Zhang, C. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *advances in neural information processing systems*, 33:12345–12355, 2020.

Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Jäger, B., Safaric, D., et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NIPS Deep Learning Workshop*, 2015.

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.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ma, J., Rashidi, J., Olmos, P., and Caterini, A. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Mansurov, J., Sakip, A., and Aji, A. F. Data laundering: Artificially boosting benchmark results through knowledge distillation. *arXiv preprint arXiv:2412.15255*, 2024.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Qu, N., Yin, J., Yoo, H., Purucker, L., and Hutter, F. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. AI in health and medicine. *Nature Medicine*, 28(1):31–38, 2022. doi: 10.1038/s41591-021-01614-0.

Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E. H., and Jain, S. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 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.

Zhang, X., Ren, G., Yu, H., Yuan, H., Wang, H., Li, J., Wu, J., Mo, L., et al. LimiX: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., and Zhang, Q. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650*, 2021.

A. Implementation details

Defaults: K=5 folds, $T_{\min}$=1, $T_{\max}$=5, α=0.7, confidence-weight $(\mu, \sigma) = (0.7, 0.2)$. LGBM students use 300 estimators with patience-30 early stopping. MLP students use embedding dimension $\min(8d, 128)$, hidden widths scaled to dataset size, warmup-plus-cosine LR, label smoothing 0.05, and SWA on the final 20% of training. A collapse detector monitors prediction entropy and restarts with higher dropout if degenerate predictions appear. Teachers run inference-only with no fine-tuning. Latency is measured on an Intel Xeon 8358 CPU for students and an NVIDIA A100 GPU for teachers, averaged over 50 runs (students) and 20 runs (teachers).

B. Dataset details

Table 6. The 19 health datasets used. Modality codes: C = cardiology, O = oncology, N = nephrology, D = dermatology, E = endocrinology, ICU = intensive care, H = hepatology, OG = obstetrics-gynaecology, W = women's health.

Dataset	Modality	Samples	Features	Classes	Use
heart_cleveland	C	303	13	2	accuracy, calibration, fairness
heart_failure_clinical	C	299	12	2	accuracy, calibration, fairness
south_african_heart	C	462	9	2	accuracy, calibration
breast_cancer_wisconsin	O	569	30	2	accuracy, calibration
wdbc	O	569	30	2	accuracy, calibration
breast_w	O	699	9	2	accuracy, calibration
indian_liver_patient	H	583	10	2	accuracy, calibration
blood_transfusion	—	748	4	2	accuracy, calibration
pima_diabetes	E	768	8	2	accuracy, calibration, fairness
cervical_cancer	W	858	35	2	accuracy
mammographic_mass	O	961	5	2	accuracy, calibration
sick	E	3,772	27	2	accuracy, calibration, fairness
support2	ICU	9,105	21	2	accuracy, calibration, fairness
framingham	C	4,238	15	2	accuracy, fairness (sex)
diabetes_130_us	E	101,766	47	2	accuracy, fairness (race)
mimic_iii_mortality	ICU	58,976	23	2	accuracy, fairness (gender)
chronic_kidney_disease	N	400	24	3	accuracy
dermatology	D	366	34	6	accuracy
analcatdata_dmft	OG	797	4	6	accuracy

Table 6 lists all 19 datasets with sample size, feature count, class count, and the role each plays in the evaluation. Three of the binary cohorts are versions of the Wisconsin breast-cancer dataset (`breast_cancer_wisconsin`, `wdbc`, `breast_w`); we kept all three for sanity checking but this overweights breast oncology in aggregate AUC by roughly 3×.

C. Additional Accuracy Results

Accuracy on CatBoost and XGBoost. Table 7 reports the avg AUC and retention on CatBoost and XGBoost over the 16 binary datasets. The results show that the strong retention observed for LightGBM extends to other tree-based students. CatBoost and XGBoost retain nearly all teacher AUC, with several settings exceeding 100% retention. The best CatBoost student reaches .877 AUC, slightly above the best teacher in Table 1, while the best XGBoost students reach .875 AUC. This supports the main-text conclusion that tree-based students are reliable distillation targets and that distillation can sometimes improve AUC rather than only compressing the teacher.

Accuracy on multi-class datasets. Table 8 reports the avg AUC and retention of all students over the 3 multi-class datasets. The results highlight that multiclass results are broadly consistent with the binary setting but with smaller differences between methods. Tree-based students retain high teacher performance, with LightGBM and XGBoost reaching up to .786–.787 macro-AUC and retention near or above 100% for some teachers. MLP students remain slightly weaker, with

Table 7. Avg AUC and retention on CatBoost and XGBoost over the 16 binary datasets.

Type	Model	AUC	Ret.
Distilled (CatBoost)	TabICLv2→CB	**.876** ± .109	100.4%
	LimiX→CB	.875 ± .111	100.0%
	TabPFNv2.6→CB	.875 ± .112	100.5%
	TabDPT→CB	.872 ± .113	100.5%
	Orion-MSP→CB	.870 ± .110	100.7%
Distilled (CatBoost, multi-teacher)	[PFNv2.6+ICL+Limix+TabDPT]→CB	**.877** ± .110	–
	[PFNv2.6+ICL+Limix]→CB	.876 ± .110	–
	[PFNv2.6+Limix]→CB	.876 ± .110	–
	[PFNv2.6+OrionMSP+Limix]→CB	.875 ± .110	–
Distilled (XGBoost)	TabICLv2→XGB	**.875** ± .109	100.3%
	TabPFNv2.6→XGB	.874 ± .112	100.5%
	LimiX→XGB	.874 ± .110	99.9%
	TabDPT→XGB	.869 ± .113	100.2%
	Orion-MSP→XGB	.868 ± .109	100.4%
Distilled (XGBoost, multi-teacher)	[PFNv2.6+ICL+Limix]→XGB	**.875** ± .110	–
	[PFNv2.6+ICL+Limix+TabDPT]→XGB	.875 ± .110	–
	[PFNv2.6+Limix]→XGB	.875 ± .110	–
	[PFNv2.6+OrionMSP+Limix]→XGB	.874 ± .110	–

retention between 96.9% and 99.9%. These results suggest that distillation transfers reasonably well to multiclass tasks, although the small number of datasets makes the trend less conclusive than in the binary benchmark.

D. Additional Latency Results

Figure 2 and Table 9 confirm that the latency gains hold across all distilled settings. All MLP students run between 3.7 and 4.5 ms, and all LGBM students between 7.0 and 10.7 ms on CPU, while the fastest GPU-based teacher, TabICLv2, requires 187.4 ms. Even at P99, distilled students remain below 17 ms, whereas teachers range from 202.6 ms to over 4 s. The figure therefore highlights a consistent deployment gap: distillation moves inference from hundreds or thousands of GPU milliseconds to single-digit CPU milliseconds, with MLP offering the lowest latency and LGBM remaining within a practical real-time range.

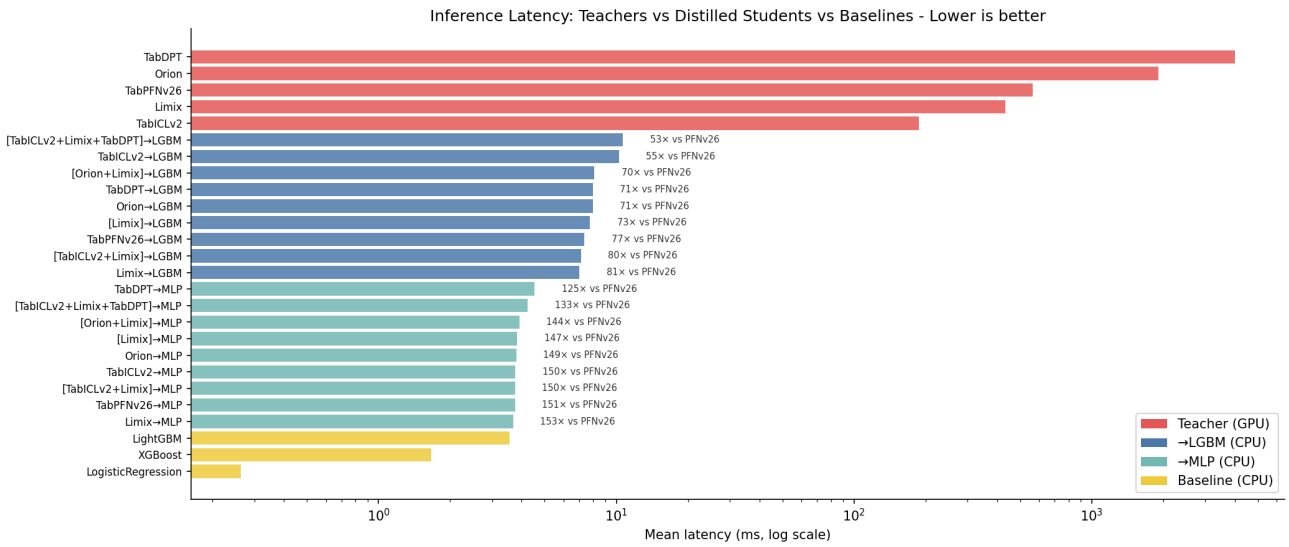

Figure 2. Inference latency, log scale. Teachers (red) sit at 200–4000 ms on GPU. Distilled students sit at 4–11 ms on CPU; tree baselines below 4 ms. Distillation collapses the latency gap.

Table 8. Multiclass panel results (3 datasets, macro-AUC). Per-dataset mean, ± std across datasets.

Type	Model	AUC	Ret.
Baseline	LogisticRegression	.785 ± .303	–
	XGBoost	.770 ± .323	–
	LightGBM	.770 ± .323	–
Teacher	TabDPT	**.789** ± .297	–
	TabICLv2	.788 ± .299	–
	LimiX	.784 ± .304	–
	TabPFNv2.6	.782 ± .307	–
	TabPFN (v2.5)	.779 ± .312	–
	Orion-MSP v1.5	.773 ± .319	–
Distilled (CatBoost)	TabICLv2→CB	**.782** ± .304	99.2%
	LimiX→CB	.778 ± .312	99.2%
	Orion-MSP→CB	.777 ± .313	100.5%
	TabDPT→CB	.776 ± .314	98.4%
	TabPFNv2.6→CB	.774 ± .320	98.9%
Distilled (XGBoost)	TabICLv2→XGB	**.787** ± .299	99.9%
	TabDPT→XGB	.780 ± .309	98.8%
	LimiX→XGB	.777 ± .313	99.1%
	TabPFNv2.6→XGB	.777 ± .315	99.3%
	Orion-MSP→XGB	.776 ± .315	100.3%
Distilled (LightGBM)	TabICLv2→LGBM	**.786** ± .299	99.8%
	TabDPT→LGBM	.785 ± .302	99.5%
	Orion-MSP→LGBM	.785 ± .301	101.5%
	LimiX→LGBM	.784 ± .303	100.0%
	TabPFNv2.6→LGBM	.782 ± .305	100.0%
	TabPFN (v2.5)→LGBM	.779 ± .310	100.1%
Distilled (MLP)	TabPFN→MLP	**.778** ± .311	99.9%
	LimiX→MLP	.772 ± .319	98.5%
	TabPFNv2.6→MLP	.772 ± .320	98.7%
	TabDPT→MLP	.769 ± .323	97.5%
	Orion-MSP→MLP	.766 ± .328	99.0%
	TabICLv2→MLP	.763 ± .332	96.9%

E. Multi-teacher analysis with full coverage

Table 10 reports multi-teacher LGBM distillation on the full 19-dataset benchmark. None of the four ensembles that ran on all datasets exceeds the best single-teacher student (TabICLv2→LGBM at 0.906); they are tied at three decimal places. On the same subset, the best single teacher scores 0.916, so there is some real ensemble effect on those three. Whether it generalises requires running the ensemble on the remaining ten datasets, which we will do in a follow-up.

Table 9. Inference latency and throughput, for all settings of distilled students, averaged out across datasets ranging from 91 to 1821 test examples. All measurements on a single CPU core; teachers run on GPU.

Model	Mean ms	P99 ms	Throughput
Baselines			
LogisticRegression	**0.3**	0.3	2,249K/s
XGBoost	1.7	2.7	353K/s
LightGBM	3.6	4.1	162K/s
TFM teachers (GPU)			
TabICLv2	187.4	202.6	3,221/s
LimiX	433.8	457.1	1,450/s
TabPFNv2.6	564.9	577.8	1,099/s
Orion-MSP v1.5	1,911.9	1,941.1	323/s
TabDPT	4,010.0	4,024.1	600/s
Distilled MLP students			
LimiX→MLP	3.7	4.0	172K/s
TabPFNv2.6→MLP	3.8	4.1	170K/s
TabICLv2→MLP	3.8	4.4	169K/s
[PFNv2.6+Limix]→MLP	3.8	4.8	166K/s
[PFNv2.6+ICL+Limix]→MLP	3.8	4.3	166K/s
Orion-MSP→MLP	3.8	4.5	164K/s
[PFNv2.6+OrionMSP+Limix]→MLP	3.9	4.7	157K/s
[PFNv2.6+ICL+Limix+TabDPT]→MLP	4.2	4.7	150K/s
TabDPT→MLP	4.5	5.2	158K/s
Distilled LGBM students			
LimiX→LGBM	7.0	14.2	87K/s
[PFNv2.6+ICL+Limix]→LGBM	7.1	12.5	92K/s
TabPFNv2.6→LGBM	7.3	12.5	82K/s
[PFNv2.6+Limix]→LGBM	7.7	13.1	98K/s
TabDPT→LGBM	8.0	12.6	79K/s
Orion-MSP→LGBM	8.0	10.0	86K/s
[PFNv2.6+OrionMSP+Limix]→LGBM	8.1	15.7	78K/s
TabICLv2→LGBM	10.3	14.6	67K/s
[PFNv2.6+ICL+Limix+TabDPT]→LGBM	10.7	16.1	64K/s

Table 10. Multi-teacher LGBM distillation across binary datasets. The first four rows ran on all 16 datasets; the last ran on 3.

Teacher ensemble	AUC (mean)	Coverage (datasets)
[TabPFN + Limix]	.906	16
[TabPFN + TabICLv2 + Limix]	.906	16
[TabPFN + OrionMSP + Limix]	.906	16
[TabPFN + TabICLv2 + Limix + TabDPT]	.906	16
Best single-teacher LGBM (TabICLv2→LGBM)	.906	16
Best single teacher (TabICLv2)	.918	16