Leveraging Foundation Models in Healthcare: A Distillation Approach to Interpretable Clinical Prediction

Hans Farrell Soegeng[1*], **Tristan Guérand**[1], **Thomas Peyrin**[1]

[1]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
hans0048@e.ntu.edu.sg, thomas.peyrin@ntu.edu.sg

Abstract

In data-scarce settings, learning accurate and interpretable models for high-stakes medical tabular classification remains a fundamental challenge, as healthcare decisions must be transparent and trustworthy. We propose a novel pipeline for few-shot tabular classification that distills the predictive power of large foundation models into globally interpretable student models. Given a tabular dataset, we generate synthetic data using CTGAN to approximate the underlying distribution. We then finetune high-capacity teacher models (TabPFN, TabM) on a small number of labeled examples and use them to pseudolabel the synthetic data. Finally, we train student explainer models (i.e., XGBoost, decision trees, Generalized Linear Rule Model (GLRM), and TTnet) on this pseudolabeled synthetic dataset. These student models are globally and exactly interpretable, yielding logical decision rules (e.g., disjunctive normal form) that fully reproduce their predictions. Evaluated across 7 clinical tabular tasks, our distilled models generally outperform baselines trained directly on the few-shot data, with improved ROC AUC scores across few-shot settings. This work demonstrates that foundation models can be effectively leveraged as teachers to produce small, transparent, and high-performing classifiers. Our approach advances the goal of reliable and interpretable machine learning in real-world settings where labeled data is limited.

Code — https://github.com/hansfarrell/clinicaldistill

Introduction

In high-stakes clinical settings, machine learning models must be transparent and trustworthy, yet the scarcity of large, labeled datasets presents a fundamental challenge (Rudin 2019; Doshi-Velez and Kim 2017). This data limitation creates an unacceptable trade-off: clinicians must either rely on traditional interpretable models that underperform with limited data or trust opaque "black-box" foundation models whose internal logic is hidden, restricting their use in healthcare.

While foundation models like TabPFN (Hollmann et al. 2025) and TabM (Gorishniy, Kotelnikov, and Babenko 2025) excel at few-shot tabular classification, their complex architectures prevent the direct scrutiny required for clinical adoption. We propose a concise distillation pipeline that bridges the gap between black-box accuracy and required transparency.

Our framework (illustrated in Figure 1) transfers the predictive power of a foundation model into a globally and exactly interpretable student model. The process begins by generating synthetic data with CTGAN (Xu et al. 2019) to mirror the underlying distribution of the clinical data. A foundation model, finetuned on as few as four real patient samples, then generates high-quality pseudolabels for this synthetic dataset. Finally, a lightweight, interpretable student model (e.g., a decision tree or TTnet rule extractor (Benamira et al. 2023)) is trained on this large, pseudolabeled dataset. The resulting classifier can be fully expressed as a set of human-readable logical rules (e.g., in Disjunctive Normal Form) that exactly reproduce its predictions.

This approach offers dual benefits critical for clinical deployment: it boosts predictive performance in data-scarce environments and replaces computationally expensive foundation models with efficient and predictable rule sets at inference time. We validate our method on 7 medical classification tasks using real-world clinical data sourced from the U.S. National Library of Medicine (ClinicalTrials.gov), MIT's PhysioNet, and Kaggle. Our results demonstrate that, in nearly all cases, distillation significantly improves test ROC AUC over baseline models without sacrificing the full interpretability required for healthcare applications.

Our contributions are:

- We propose a novel distillation framework tailored for creating interpretable models in few-shot clinical settings by combining foundation models with synthetic data generation.
- We deliver global, exact interpretability, ensuring student models can be converted into logical formulas that reproduce their behavior with perfect fidelity.
- We demonstrate the framework's effectiveness on a diverse set of real-world clinical datasets, confirming its practical value for developing trustworthy medical AI.

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

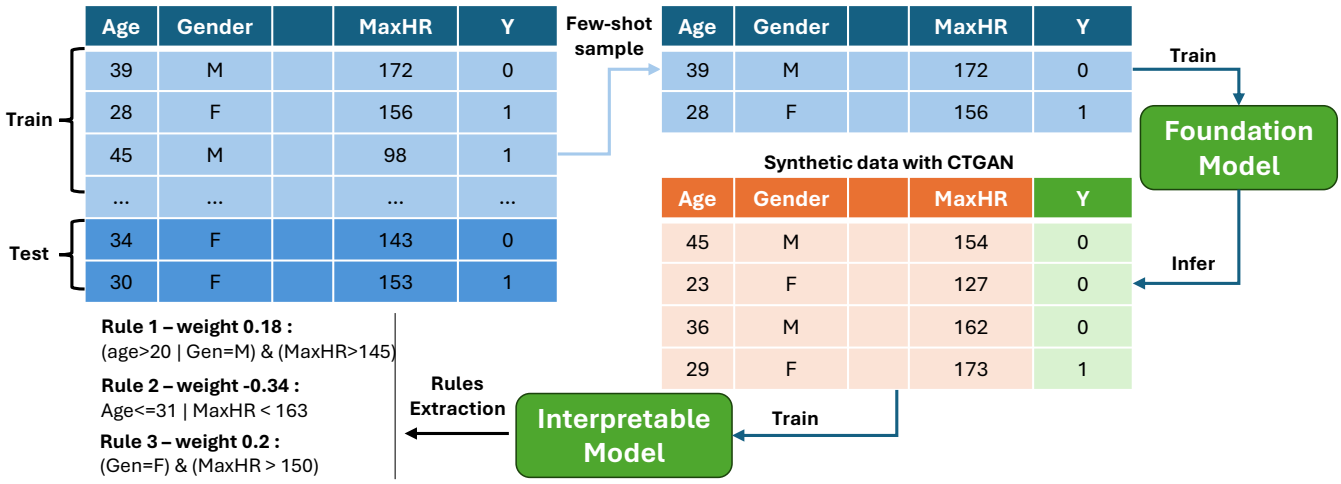

Figure 1: A high-level illustration of the pipeline.

Related Work

Foundation Tabular Models

Recent advances have produced powerful "foundation" models for tabular data, which are particularly promising for the few-shot learning challenges common in clinical research. Models like TabPFN (Hollmann et al. 2025), a transformer (Vaswani et al. 2017) pre-trained on millions of synthetic tasks, and TabM (Gorishniy, Kotelnikov, and Babenko 2025), a parameter-efficient deep learning architecture, have demonstrated state-of-the-art performance with limited data. Similarly, TabLLM (Hegselmann et al. 2023) leverages the semantic understanding of large language models by serializing data rows into text. MediTab (Wang et al. 2024) advances this approach by proposing a data-centric framework that not only consolidates heterogeneous medical tables into natural language but also employs a critical data auditing and enrichment pipeline to train a foundational model. While these models show immense potential for analyzing complex clinical datasets, their "black-box" nature, stemming from deep, opaque architectures, is a significant barrier to their direct adoption in healthcare, where transparency and trust are non-negotiable.

Globally Interpretable Tabular Models

For a model to be trusted in a clinical setting, its decision-making process must be fully transparent. This requires global interpretability, where the model's entire logic can be expressed as a set of human-readable rules, as opposed to local, post-hoc explanations from methods like LIME or SHAP, which are vulnerable to adversarial attacks and do not provide a complete picture of the model's reasoning (Slack et al. 2020).

Our work focuses on models that meet this high standard. Tree-based models like Decision Trees and XGBoost offer clear, hierarchical rule paths (Chen and Guestrin 2016). More explicitly, rule-based models such as GLRM (Wei et al. 2019) and TTnet (Benamira et al. 2023) are designed to output their logic as weighted conjunctive rules or Boolean formulas. These models produce the kind of verifiable, "glass-box" logic essential for clinical decision support.

Beyond TTnet and GLRM for tabular data, DiffLogicNet (Petersen et al. 2022) introduces differentiable logic gate networks that learn networks of binary logic operators (e.g., AND, XOR) via a continuous relaxation and then discretize them into exact logic gate circuits. Notably, DiffLogicNet and its convolutional extension for vision tasks can be compiled into Boolean circuits, enabling fast, hardware-friendly inference on image classification benchmarks while preserving a circuit-level representation of the model's decision logic.

Distillation

Knowledge distillation (KD) is a technique for transferring knowledge from a large teacher model to a smaller student model, often by training the student on the teacher's soft predictions (Hinton 2015). This concept has been extended to variants like task-specific (Jacob, Agarwal, and Stenger 2023), multi-teacher (Wen et al. 2024), and self-distillation (Gou et al. 2023). Combining KD with synthetic data has also been explored in other domains. For example, Nguyen et al. used synthesized images for few-shot KD with black-box teachers in computer vision (Nguyen et al. 2022), a concept similar to our pipeline. However, the combination of synthetic augmentation and interpretability has seen little work in the tabular domain. Our study builds on these approaches, using distillation to transfer a teacher's knowledge into an interpretable student model to achieve transparency.

Methodology

Our distillation pipeline begins with the few-shot clinical dataset, $(X_{\text{shot}}, y_{\text{shot}})$, which is preprocessed by one-hot encoding categorical features and normalizing continuous ones. The process unfolds in two main stages:

1. Teacher Training and Pseudolabel Generation: First, a foundation model (the "teacher") is adapted to the small,

few-shot dataset $(X_{\text{shot}}, y_{\text{shot}})$. Depending on the architecture, this is achieved either by leveraging the pretrained prior of generalist model (TabPFN) via finetuning, or by training the model directly on the limited samples (TabM). Concurrently, we employed CTGAN (Xu et al. 2019) as our synthetic data generator. While recent diffusion-based methods like TabDDPM (Kotelnikov et al. 2023) have demonstrated state-of-the-art fidelity, they incur significantly higher computational costs during the sampling phase. For our distillation pipeline, which requires generating large volumes of synthetic pseudolabeled data, CTGAN offers an optimal trade-off between distributional fidelity and computational efficiency. The trained teacher model then infers pseudolabels, y_{syn}, for this entire synthetic dataset. This results in a large, low-variance training set, $(X_{\text{syn}}, y_{\text{syn}})$, which encapsulates the teacher's learned knowledge.

2. Training Student Models: In the final stage, an interpretable student model is trained on the complete pseudolabeled dataset, $(X_{\text{syn}}, y_{\text{syn}})$. Each student model is designed to be globally and exactly interpretable, converting its internal logic into human-readable expressions. Decision Trees and XGBoost: The logic of these models can be fully represented by their tree structures. Each path from a root to a leaf node forms a conjunctive rule (a series of AND conditions), and in the case of XGBoost, the final prediction is an aggregation of the outputs from an ensemble of such trees. GLRM (Generalized Linear Rule Model): This model learns a classifier as a weighted sum of conjunctive rules (Wei et al. 2019). Its prediction is determined by the function: $f(x) = \text{intercept} + \sum_i \mathbb{1}_{\{\text{Rule}_i(x)\}} \cdot w_i > 0$, where $\text{Rule}_i(x)$ is a Boolean clause and w_i is its learned weight. An example of a distilled GLRM is shown in Table 1. TTnet: This model is explicitly designed to produce a set of rules in Disjunctive Normal Form (DNF), providing a clear logical expression for its predictions (Benamira et al. 2023).

Coefficient	Clause
+4.786	(intercept)
+4.170	`FastingBS` = 1
−2.509	`Age` ≤ 62.00 AND `ChestPainType` ≠ TA AND `Cholesterol` > 134.80
−2.210	`Cholesterol` ≤ 305.00
−2.053	`Age` ≤ 62.00 AND `Oldpeak` ≤ 1.80 AND `ST_Slope` = Up
+1.923	`Sex` = M AND `RestingBP` ≤ 160.00 AND `RestingBP` > 110.00
+1.803	`Age` ≤ 65.00 AND `ChestPainType` ≠ ATA AND `RestingBP` > 110.00
−1.711	`Oldpeak` ≤ 1.00
−1.590	`ChestPainType` = ATA
−1.262	`Cholesterol` > 0.00 AND `RestingECG` ≠ ST AND `Oldpeak` ≤ 1.80
−1.193	`ExerciseAngina` = N AND `Oldpeak` ≤ 2.30
+1.032	`RestingBP` > 110.00 AND `MaxHR` ≤ 170.00 AND `ST_Slope` = Flat
−0.929	`MaxHR` ≤ 103.00
−0.858	`RestingBP` ≤ 120.00
−0.832	`RestingBP` ≤ 130.00 AND `Oldpeak` ≤ 2.30
+0.820	`FastingBS` = 0 AND `Oldpeak` ≤ 2.30
−0.817	`Age` ≤ 65.00 AND `Cholesterol` > 0.00 AND `Oldpeak` ≤ 0.60
−0.812	`Age` ≤ 62.00 AND `ChestPainType` ≠ NAP AND `Cholesterol` > 0.00

Table 1: Example rule-based classification model coefficients and conditions from GLRM distilled from 128-shots fine-tuned TabPFN on the **Heart Disease** dataset. We predict whether a patient has heart disease or not.

Experiments

In this section, we conduct experiments to evaluate our proposed pipeline and try to answer the following question: How accurate would the classification be with such models? How much does it improve the baseline of simply training the explainer model on the few-shot data?

Baseline and Evaluation. To assess the effectiveness of our distillation-based framework, we benchmark it against a baseline setting that uses no distillation. In the baseline pipeline, each student model is trained directly on the few-shot dataset $(X_{\text{shot}}, y_{\text{shot}})$ without access to the foundation model. This represents the standard approach to learning interpretable models under data-scarce conditions. For evaluation, we also include the 'all' shot setting, where the training set is not sampled but entirely included in the training of the foundation models for distillation and of the student models for the baseline method.

For evaluation, we measure the generalization performance of each trained student model on the held-out test set X_{test}, using the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) as the performance metric. Unlike accuracy, which depends on a fixed threshold, the ROC-AUC evaluates the quality of the model's probabilistic or soft predictions. Higher AUC values indicate better overall discriminative ability and generalization performance, i.e. an AUC of 1.0 indicates a perfect classifier whereas an AUC of 0.5 corresponds to random guessing.

We define a full evaluation unit as a quadruple $(\texttt{foundation\_model}, \texttt{dataset}, k, \texttt{student\_model})$ where the student model is trained via our distillation pipeline, i.e., on $(X_{\text{syn}}, y_{\text{syn}})$ pseudolabelled by the foundational model trained on $(X_{\text{shot}}, y_{\text{shot}})$. Each such unit is compared against a baseline triplet $(\texttt{dataset}, k, \texttt{student\_model})$ where the student model is trained on the few-shot data $(X_{\text{shot}}, y_{\text{shot}})$ instead.

The performance gap between the two setups quantifies the benefit of leveraging foundation models and synthetic data generation to support interpretable model learning in few-shot regimes. We report and compare test ROC-AUC scores across all combinations of:

- 7 medical tabular classification datasets (a description of the datasets is provided in Table 2),
- 8 values of k, number of shots (4, 8, 16, 32, 64, 128, 256, 'all'),
- 2 foundational models for the distillation pipeline (TabM, TabPFN),
- 4 student interpretable models (Decision Tree, Logistic Rule Regression, TTnet, XGBoost).

Overall, the distillation pipeline achieves higher ROC-AUC scores than the baseline in a majority of dataset–shot–student combinations, showing its broad applicability across domains. The gains are most prominent in low-shot settings (4-8 shots). These trends highlight both the strengths and the boundaries of the approach.

A summary of aggregate results is presented in Figure 2. More details on the experimental results are presented on the Extended Results section of the Appendix .

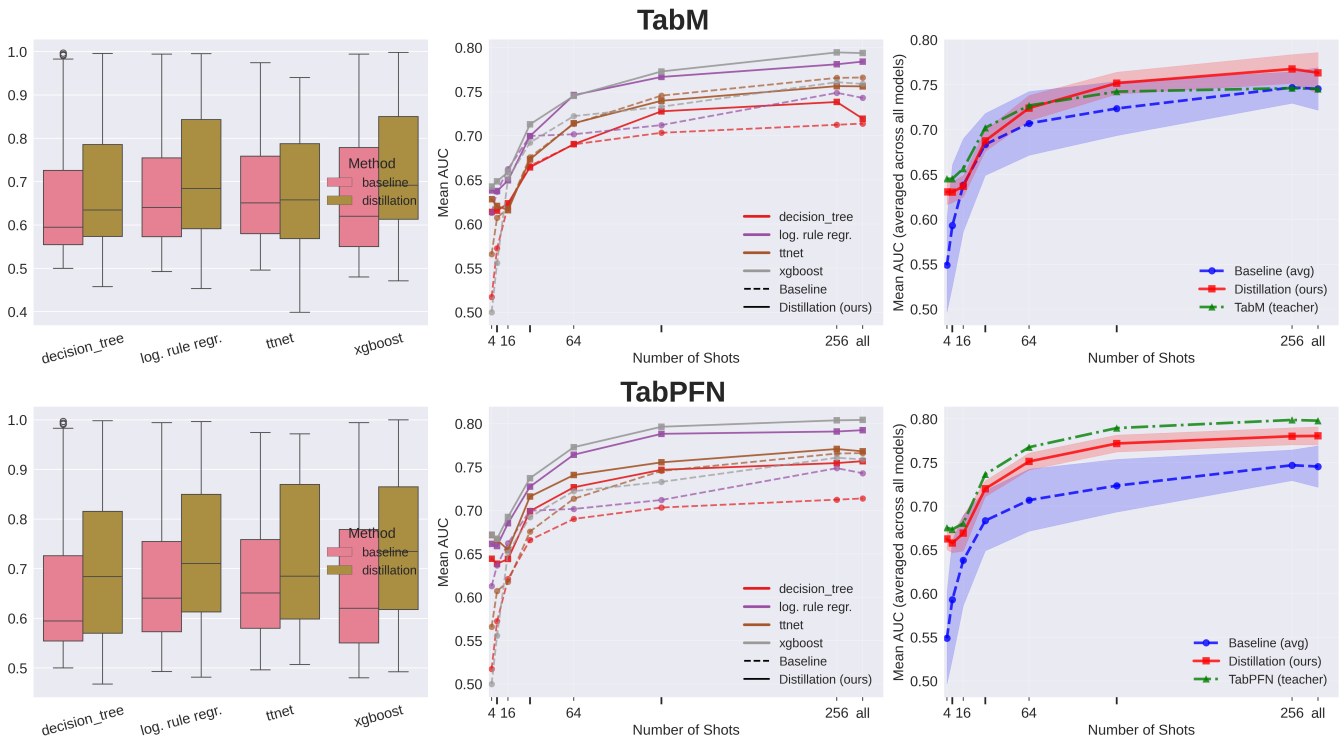

Figure 2: Each row of figures describes the performance of the TabM, TabPFN models (in descending order). For each parent model, we report the (i) distribution of the test ROC-AUC across all 15 datasets and 7 student models, (ii) mean ROC-AUC by shot size for each student model, (iii) the overall average ROC-AUC across students as a function of shot size, where the shaded regions represent the standard deviation across different student models at each shot size.

TabPFN and TabM consistently outperform the baseline across all shot settings, although the benefit narrows down for both on 16 shots. TabPFN, in particular, exhibits the most robust improvements across the board, with especially strong gains in ultra-low-shot scenarios. This reflects TabPFN's design for Bayesian in-context learning, leveraging its synthetic prior training to generalize from minimal data. By pseudolabeling CTGAN-generated samples, it provides the student models with a dense, low-variance training signal, critical in data-scarce settings.

Looking across student models, XGBoost (Chen and Guestrin 2016) and decision trees generally extract the largest improvements from distillation (they gain most in mean AUC and exploit richer signals from the pseudolabels), whereas TTnet (Benamira et al. 2023) generally shows the smallest incremental benefit (its CNN architecture limits its tabular learning capabilities). Logistic rule regression (Wei et al. 2019) get moderate, reliable gains as they are able to capture rule-like structure from the teacher.

Statistical Test. To formally evaluate whether the distillation pipeline yields higher predictive performance than the non-distilled baseline, we apply the paired Wilcoxon signed-rank test to the per-experiment ROC-AUC scores. Let x_i denote the AUC obtained by the distilled student model on experiment i and y_i the corresponding AUC for the baseline student model directly trained on few-shot data, where each paired index i corresponds to the same tuple $(\texttt{dataset}, k, \texttt{student\_model}, \texttt{seed})$. We test the one-sided hypothesis that distillation increases the median AUC.

Formally, we define paired differences

$$d_i = x_i - y_i, \text{ for } i = 1, 2, \ldots, N.$$

We test the one-sided alternative hypothesis:

$$\mathbf{H_0} : \text{median}(d_i) \leq 0 \text{ , vs. } \mathbf{H_1} : \text{median}(d_i) > 0.$$

Applying the test to the full set of matched AUC pairs produced a p-value of 5×10^{-65}. Given the extremely small value, we can confidently reject H_0 at conventional significance levels and conclude that the distillation pipeline produces significantly higher AUC performance than the baseline.

Aside from the Wilcoxon signed-rank test, we perform the Friedman statistical test with Nemenyi post-hoc to validate and rank the effect of distillation over the baseline method with TabM and TabPFN. The AUC results over the different $(\texttt{dataset}, k, \texttt{student\_model}, \texttt{seed})$ combination are grouped, and the 3 methods producing the groups (TabPFN, TabM, baseline) are compared. Thus, it is an all-pairs comparison, testing

$$\mathbf{H_0} : \text{the distribution of the 3 groups are the same, vs.}$$
$$\mathbf{H_1} : \text{there is a significant difference between the distribution of group } i, j$$

The pairwise p-values between the 3 methods are:

	Baseline	TabPFN	TabM
Baseline	1	0.000	0.0231
TabPFN	0.000	1	0.000
TabM	0.0231	0.000	1

At $\alpha = 0.05$, we can conclude that there is a significant difference between the distribution of the TabM group and the baseline group with a p-value of 0.023. With a p-value of close to 0 when statistically compared to both the TabM and baseline groups, the distillation effect with TabPFN is very evident, highlighting the effectiveness of the synthetic prior pre-training of TabPFN.

Overall, these results validate our core claim: distillation from foundation tabular models into globally interpretable students can enhance few-shot classification performance, particularly with TabPFN, provided the teacher's inductive biases align with the data regime.

Limitation and Future Work

The primary limitation of our framework is that the student model's performance is fundamentally bounded by the quality of its teacher. As the student learns exclusively from teacher-generated pseudolabels, it inevitably inherits any systematic biases, miscalibrations, or domain-specific weaknesses present in the foundation model. A student, therefore, cannot outperform a suboptimal teacher.

This dependency, however, underscores the importance of strategic teacher selection. Our framework is model-agnostic, allowing practitioners to choose a foundation model whose inductive biases are best aligned with the dataset's characteristics., for instance, leveraging TabPFN for extremely small datasets or TabM for numeric-heavy data. This points to promising directions for future work, including the development of automated teacher selection methods or the use of teacher ensembles to mitigate the weaknesses of any single model and enhance overall robustness.

Conclusion

In this work, we propose a novel framework for enabling globally interpretable few-shot tabular classification by distilling foundation models into rule-based student models. This approach is especially applicable for high-stakes fields like healthcare, where decisions must be transparent, trustworthy, and often rely on inherently data-scarce clinical datasets.

Our approach leverages the strengths of recent tabular foundation models (e.g. TabPFN, TabM) to generate high-quality pseudolabels on synthetic data, which are then used to train interpretable models like GLRM, TTnet, decision trees, and XGBoost.

Through experiments on 7 diverse clinical datasets across varying shot sizes, we demonstrate that distillation from foundation models improves the performance of interpretable student models compared to training them directly on few-shot data. In particular, TabPFN emerges as a consistently strong teacher, with distilled students outperforming their baselines across all shot settings and for all student models, often by substantial margins. By replacing them with distilled interpretable models, our pipeline not only enhances explainability but can also offer efficiency gains for deployment.

As tabular foundation models continue to advance, especially with the integration of ever more powerful LLM backbones, our distillation framework will serve as an essential tool to extract, inspect, and understand their learned representations. In this way, it not only addresses today's need for interpretable few-shot models, but also lays groundwork for mechanistic interpretability of future, large-scale tabular foundation models.

Our findings suggest that, when appropriately matched to dataset characteristics, foundation models can act as powerful teachers that elevate the practicality of interpretable models in low-data regimes, and in the case of TabPFN, can deliver consistently strong, universal gains across all settings tested.

References

Benamira, A.; Guérand, T.; Peyrin, T.; and Soegeng, H. 2023. Neural Network-Based Rule Models with Truth Tables. In Gal, K.; et al., eds., *ECAI 2023: 26th European Conference on Artificial Intelligence*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, 223–230. IOS Press.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning.

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

Gou, J.; Xiong, X.; Yu, B.; Du, L.; Zhan, Y.; and Tao, D. 2023. Multi-target knowledge distillation via student self-reflection. *International Journal of Computer Vision*, 131(7): 1857–1874.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 507–520. Curran Associates, Inc.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In Ruiz, F.; Dy, J.; and van de Meent, J.-W., eds., *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 5549–5581. PMLR.

Hinton, G. 2015. Distilling the Knowledge in a Neural Network.

Hollmann, N.; Müller, S.; Purucker, L.; et al. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637: 319–326.

Jacob, G. M.; Agarwal, V.; and Stenger, B. 2023. Online knowledge distillation for multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2359–2368.

Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; and Babenko, A. 2023. TabDDPM: modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Nguyen, D.; Gupta, S.; Do, K.; and Venkatesh, S. 2022. Black-box Few-shot Knowledge Distillation.

Petersen, F.; Borgelt, C.; Kuehne, H.; and Deussen, O. 2022. Deep Differentiable Logic Gate Networks. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 2006–2018. Curran Associates, Inc.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1: 206–215.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 180–186. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5999–6009.

Wang, Z.; Gao, C.; Xiao, C.; and Sun, J. 2024. MediTab: Scaling Medical Tabular Data Predictors via Data Consolidation, Enrichment, and Refinement. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6062–6070. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wei, D.; Dash, S.; Gao, T.; and Gunluk, O. 2019. Generalized Linear Rule Models. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6687–6696. PMLR.

Wen, H.; Pan, L.; Dai, Y.; Qiu, H.; Wang, L.; Wu, Q.; and Li, H. 2024. Class Incremental Learning with Multi-Teacher Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28443–28452.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling Tabular data using Conditional GAN. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Datasets

We conduct experiments on 7 clinical binary classification tabular datasets, comprising 4 datasets from the US clinical trials library, 1 from the MIT physiology laboratory data bank, and 2 from Kaggle. We present the dataset names, the number of samples, the number of categorical and continuous features, the number of positive and negative instances, and the source link for the datasets.

Table 2: Number of samples, categorical features, continuous features, and the source of the 7 datasets used for the benchmark.

dataset_name	n_samples	cat.	cont.	1	0
breastcancer[1]	3871	14	2	266	3605
breastcancer2[2]	1651	13	6	320	1331
chemotherapy[3]	1604	31	11	714	890
coloncancer[4]	2968	14	3	357	2611
diabetes[5]	768	0	8	268	500
heart[6]	918	6	5	508	410
respiratory[7]	1776	22	24	283	1492

Implementation Details

We generated a fixed number of 5000 synthetic samples with the CTGAN generator for all datasets. This large sample size is selected to ensure sufficient density to capture the decision boundaries learned by the foundational model. All training and inference results for the parent models (TabPFN, TabM) uses seed 0 to pseudolabel the synthetic data. Afterwards, the distillation and baseline experiments of the student models across all 7 datasets and 7 shot settings uses 5 random seeds (0, 1, 6, 7, 8) to report the mean and standard deviation of metrics. We use a class-balanced sampling to obtain the few-shot data for parent model training for all $k \in [4, 8, 16, 32, 64, 128, 256]$.

Hardware.

For all experiments, we use 4 Nvidia GeForce RTX 3090 GPUs and 8 cores Intel(R) Core(TM) i7-8650U CPU clocked at 1.90 GHz, 16 GB RAM. The hardware chosen was based on availability; neural networks under PyTorch were run on GPU, while tree-based models from scikit-learn were run on CPU.

Foundation Models.

TabM: Trained up to 500 epochs with AdamW (Loshchilov and Hutter 2019) optimizer (learning rate 1e-3, weight decay 1e-4).

Source: https://clinicaltrials.gov/ct2/show/NCT00041119
Source: https://clinicaltrials.gov/ct2/show/NCT00312208
Source: https://clinicaltrials.gov/ct2/show/NCT00694382
Source: https://clinicaltrials.gov/ct2/show/NCT00079274
Source: https://www.kaggle.com/datasets/pritsheta/diabetes-dataset
Source: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
Source: https://physionet.org/content/mimic2-iaccd/1.0/

Explainer Models.

We perform standard hyperparameter tuning by seaching over the hyperparameter space from (Grinsztajn, Oyallon, and Varoquaux 2022).

Decision Tree:

Parameter	Distribution
Max Depth	UniformInt[1, 20]
Min Samples Split	UniformInt[2, 10]
Min Samples Leaf	UniformInt[1, 10]
Max features	[sqrt, log2, None, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
Criterion	["gini", "entropy"]

XGBoost:

Parameter	Distribution
Max depth	UniformInt[1, 11]
Num estimators	UniformInt[100, 6000]
Gamma	LogUniform[1e-8, 7]
Lambda	LogUniform[1, 4]
Alpha	LogUniform[1e-8, 1e2]

Logistic Rule Regression (Wei et al. 2019):

Parameter	Distribution
λ_0	[0.1, 0.05, 0.01]
λ_1	[0.01, 0.005, 0.001]

Extended Results

We present the extended results of the experiments in Tables 3 to 7 and Figures 3 to 9. Tables 3 to 5 show the mean AUC of the baseline vs distillation by TabPFN and TabM for all 7 datasets and 4 student models across $k = 4, 32, 256$ shots. The AUC is averaged across 5 random seeds and the standard deviation over the 5 seeds is shown as superscript.

Figures 3 to 9 plots these results for each dataset, showing the baseline vs distillation AUC comparison across all shot sizes for each student model. Only the TabPFN distillation results are shown in these figures for clarity, as it consistently outperforms TabM and the baseline method in our experiments, as seen in Figure 2.

We also present the model complexity of each student model in Table 7, measured as the number of boolean operators (& and |) + the number of rules that reproduces the model's logic exactly. An example rule set from GLRM is shown in Table 1, which consists of 18 rules and 19 boolean operators (&), adding up to a complexity of 37. The TTnet model can be converted into Disjunctive Normal Form (DNF) to obtain the exact number of rules and boolean operators, while decision trees and XGBoost models can be converted into sets of conjunctive rules by tracing each path from root to leaf.

Table 3: Results for $k = 4$ shots. Mean AUC with standard deviation shown as superscript.

Dataset	Student	Baseline	TabPFN	TabM
Breast	TT	$0.496^{0.045}$	$0.547^{0.020}$	$\mathbf{0.553^{0.011}}$
	XGB	$0.500^{0.000}$	$0.548^{0.020}$	$\mathbf{0.551^{0.013}}$
	LRR	$\mathbf{0.561^{0.052}}$	$0.543^{0.021}$	$0.546^{0.011}$
	DT	$0.514^{0.028}$	$\mathbf{0.540^{0.011}}$	$0.523^{0.006}$
Breast 2	TT	$\mathbf{0.529^{0.087}}$	$0.516^{0.032}$	$0.497^{0.039}$
	XGB	$\mathbf{0.500^{0.000}}$	$0.492^{0.009}$	$0.482^{0.013}$
	LRR	$\mathbf{0.532^{0.054}}$	$0.481^{0.010}$	$0.477^{0.012}$
	DT	$\mathbf{0.518^{0.036}}$	$0.487^{0.006}$	$0.485^{0.009}$
Chemo	TT	$0.521^{0.035}$	$\mathbf{0.528^{0.013}}$	$0.449^{0.009}$
	XGB	$0.500^{0.000}$	$\mathbf{0.507^{0.009}}$	$0.471^{0.007}$
	LRR	$\mathbf{0.572^{0.061}}$	$0.486^{0.006}$	$0.453^{0.008}$
	DT	$\mathbf{0.500^{0.000}}$	$0.468^{0.007}$	$0.458^{0.006}$
Colon	TT	$0.515^{0.032}$	$0.541^{0.028}$	$\mathbf{0.554^{0.014}}$
	XGB	$0.500^{0.000}$	$\mathbf{0.538^{0.009}}$	$0.527^{0.012}$
	LRR	$0.495^{0.050}$	$0.516^{0.009}$	$\mathbf{0.526^{0.014}}$
	DT	$0.500^{0.000}$	$0.532^{0.008}$	$\mathbf{0.536^{0.009}}$
Diabetes	TT	$0.602^{0.101}$	$\mathbf{0.785^{0.008}}$	$0.741^{0.036}$
	XGB	$0.500^{0.000}$	$\mathbf{0.790^{0.015}}$	$0.699^{0.016}$
	LRR	$0.653^{0.103}$	$\mathbf{0.800^{0.011}}$	$0.728^{0.010}$
	DT	$0.535^{0.069}$	$\mathbf{0.682^{0.014}}$	$0.681^{0.007}$
Heart	TT	$0.540^{0.214}$	$\mathbf{0.821^{0.011}}$	$0.694^{0.081}$
	XGB	$0.500^{0.000}$	$\mathbf{0.839^{0.006}}$	$0.813^{0.015}$
	LRR	$0.575^{0.234}$	$\mathbf{0.808^{0.009}}$	$0.785^{0.008}$
	DT	$0.554^{0.108}$	$\mathbf{0.815^{0.009}}$	$0.709^{0.005}$
Respiratory	TT	$0.758^{0.101}$	$\mathbf{0.964^{0.032}}$	$0.908^{0.014}$
	XGB	$0.500^{0.000}$	$\mathbf{0.992^{0.001}}$	$0.955^{0.002}$
	LRR	$0.902^{0.060}$	$\mathbf{0.995^{0.001}}$	$0.953^{0.003}$
	DT	$0.500^{0.000}$	$\mathbf{0.988^{0.007}}$	$0.902^{0.010}$

Table 4: Results for $k = 32$ shots. Mean AUC with standard deviation shown as superscript.

Dataset	Student	Baseline	TabPFN	TabM
Breast	TT	$0.578^{0.049}$	$\mathbf{0.578^{0.020}}$	$0.570^{0.021}$
	XGB	$0.564^{0.048}$	$0.575^{0.022}$	$\mathbf{0.578^{0.019}}$
	LRR	$0.572^{0.057}$	$\mathbf{0.583^{0.018}}$	$0.580^{0.021}$
	DT	$\mathbf{0.575^{0.030}}$	$0.551^{0.018}$	$0.567^{0.016}$
Breast 2	TT	$0.650^{0.044}$	$\mathbf{0.673^{0.010}}$	$0.630^{0.015}$
	XGB	$0.599^{0.051}$	$\mathbf{0.711^{0.008}}$	$0.655^{0.017}$
	LRR	$0.628^{0.024}$	$\mathbf{0.707^{0.006}}$	$0.647^{0.013}$
	DT	$0.567^{0.061}$	$\mathbf{0.644^{0.011}}$	$0.582^{0.009}$
Chemo	TT	$0.581^{0.045}$	$\mathbf{0.625^{0.005}}$	$0.609^{0.003}$
	XGB	$0.583^{0.014}$	$0.634^{0.007}$	$\mathbf{0.641^{0.004}}$
	LRR	$0.550^{0.037}$	$0.604^{0.006}$	$\mathbf{0.628^{0.006}}$
	DT	$0.568^{0.023}$	$\mathbf{0.589^{0.006}}$	$0.567^{0.004}$
Colon	TT	$0.557^{0.021}$	$\mathbf{0.560^{0.013}}$	$0.522^{0.018}$
	XGB	$0.536^{0.035}$	$\mathbf{0.550^{0.004}}$	$0.522^{0.008}$
	LRR	$\mathbf{0.569^{0.026}}$	$0.544^{0.004}$	$0.497^{0.010}$
	DT	$0.534^{0.008}$	$\mathbf{0.545^{0.010}}$	$0.510^{0.018}$
Diabetes	TT	$0.744^{0.032}$	$\mathbf{0.781^{0.016}}$	$0.588^{0.030}$
	XGB	$0.744^{0.040}$	$\mathbf{0.802^{0.007}}$	$0.725^{0.014}$
	LRR	$0.739^{0.024}$	$\mathbf{0.786^{0.008}}$	$0.695^{0.017}$
	DT	$0.646^{0.095}$	$\mathbf{0.737^{0.015}}$	$0.642^{0.017}$
Heart	TT	$0.857^{0.025}$	$\mathbf{0.883^{0.007}}$	$0.854^{0.007}$
	XGB	$0.847^{0.023}$	$\mathbf{0.890^{0.006}}$	$0.883^{0.004}$
	LRR	$0.854^{0.033}$	$\mathbf{0.876^{0.005}}$	$0.859^{0.005}$
	DT	$0.789^{0.041}$	$\mathbf{0.831^{0.005}}$	$0.823^{0.008}$
Respiratory	TT	$0.764^{0.049}$	$0.912^{0.016}$	$\mathbf{0.940^{0.009}}$
	XGB	$0.974^{0.016}$	$\mathbf{0.998^{0.000}}$	$0.987^{0.002}$
	LRR	$0.988^{0.007}$	$\mathbf{0.991^{0.000}}$	$0.991^{0.000}$
	DT	$0.983^{0.011}$	$\mathbf{0.997^{0.000}}$	$0.959^{0.014}$

Table 5: Results for $k = 256$ shots. Mean AUC with standard deviation shown as superscript.

Dataset	Student	Baseline	TabPFN	TabM
Breast	TT	$0.642^{0.017}$	$\mathbf{0.666^{0.016}}$	$0.627^{0.031}$
	XGB	$0.640^{0.017}$	$0.663^{0.025}$	$\mathbf{0.689^{0.041}}$
	LRR	$0.591^{0.017}$	$0.650^{0.024}$	$\mathbf{0.659^{0.023}}$
	DT	$0.584^{0.030}$	$0.621^{0.017}$	$\mathbf{0.641^{0.033}}$
Breast 2	TT	$0.717^{0.017}$	$\mathbf{0.754^{0.016}}$	$0.741^{0.017}$
	XGB	$0.709^{0.018}$	$\mathbf{0.777^{0.016}}$	$0.758^{0.026}$
	LRR	$0.687^{0.022}$	$\mathbf{0.773^{0.015}}$	$0.755^{0.019}$
	DT	$0.661^{0.016}$	$\mathbf{0.727^{0.017}}$	$0.702^{0.011}$
Chemo	TT	$0.665^{0.009}$	$\mathbf{0.693^{0.023}}$	$0.682^{0.006}$
	XGB	$0.656^{0.019}$	$\mathbf{0.735^{0.006}}$	$0.735^{0.014}$
	LRR	$0.638^{0.030}$	$\mathbf{0.695^{0.003}}$	$0.692^{0.004}$
	DT	$0.599^{0.031}$	$\mathbf{0.694^{0.009}}$	$0.630^{0.009}$
Colon	TT	$0.650^{0.015}$	$0.648^{0.005}$	$\mathbf{0.663^{0.010}}$
	XGB	$0.625^{0.029}$	$\mathbf{0.667^{0.008}}$	$0.664^{0.025}$
	LRR	$0.644^{0.035}$	$0.664^{0.005}$	$\mathbf{0.677^{0.005}}$
	DT	$0.590^{0.021}$	$0.622^{0.011}$	$\mathbf{0.640^{0.027}}$
Diabetes	TT	$\mathbf{0.803^{0.010}}$	$0.799^{0.010}$	$0.773^{0.011}$
	XGB	$0.800^{0.018}$	$\mathbf{0.856^{0.005}}$	$0.817^{0.026}$
	LRR	$0.786^{0.025}$	$\mathbf{0.840^{0.003}}$	$0.813^{0.011}$
	DT	$0.742^{0.008}$	$\mathbf{0.761^{0.002}}$	$0.729^{0.030}$
Heart	TT	$0.910^{0.011}$	$\mathbf{0.915^{0.001}}$	$0.881^{0.007}$
	XGB	$0.901^{0.009}$	$\mathbf{0.929^{0.009}}$	$0.902^{0.015}$
	LRR	$0.908^{0.015}$	$\mathbf{0.925^{0.004}}$	$0.887^{0.006}$
	DT	$0.815^{0.021}$	$\mathbf{0.860^{0.010}}$	$0.830^{0.024}$
Respiratory	TT	$\mathbf{0.973^{0.007}}$	$0.921^{0.012}$	$0.927^{0.005}$
	XGB	$0.994^{0.006}$	$\mathbf{1.000^{0.000}}$	$0.998^{0.001}$
	LRR	$0.988^{0.010}$	$\mathbf{0.991^{0.000}}$	$0.984^{0.010}$
	DT	$0.997^{0.005}$	$\mathbf{0.998^{0.000}}$	$0.996^{0.006}$

Table 6: Results for $k = all$ shots. Mean AUC with standard deviation shown as superscript.

Dataset	Student	Baseline	TabPFN	TabM
Breast	TT	$0.646^{0.018}$	$\mathbf{0.663^{0.018}}$	$0.636^{0.024}$
	XGB	$0.638^{0.017}$	$0.667^{0.023}$	$\mathbf{0.688^{0.042}}$
	LRR	$0.614^{0.027}$	$0.651^{0.023}$	$\mathbf{0.660^{0.023}}$
	DT	$0.592^{0.024}$	$0.620^{0.018}$	$\mathbf{0.637^{0.038}}$
Breast 2	TT	$0.718^{0.016}$	$\mathbf{0.754^{0.017}}$	$0.748^{0.015}$
	XGB	$0.712^{0.027}$	$\mathbf{0.777^{0.016}}$	$0.763^{0.026}$
	LRR	$0.702^{0.028}$	$\mathbf{0.770^{0.016}}$	$0.756^{0.017}$
	DT	$0.634^{0.027}$	$\mathbf{0.719^{0.014}}$	$0.690^{0.022}$
Chemo	TT	$0.665^{0.009}$	$\mathbf{0.677^{0.037}}$	$0.674^{0.005}$
	XGB	$0.656^{0.013}$	$\mathbf{0.735^{0.004}}$	$0.730^{0.012}$
	LRR	$0.623^{0.053}$	$\mathbf{0.705^{0.004}}$	$0.692^{0.004}$
	DT	$0.607^{0.047}$	$\mathbf{0.694^{0.005}}$	$0.621^{0.013}$
Colon	TT	$0.647^{0.012}$	$0.649^{0.008}$	$\mathbf{0.658^{0.010}}$
	XGB	$0.607^{0.022}$	$\mathbf{0.667^{0.007}}$	$0.658^{0.022}$
	LRR	$0.643^{0.035}$	$0.664^{0.005}$	$\mathbf{0.677^{0.005}}$
	DT	$0.595^{0.024}$	$0.622^{0.009}$	$\mathbf{0.632^{0.011}}$
Diabetes	TT	$\mathbf{0.804^{0.010}}$	$0.798^{0.009}$	$0.773^{0.010}$
	XGB	$0.798^{0.013}$	$\mathbf{0.855^{0.006}}$	$0.811^{0.028}$
	LRR	$0.739^{0.120}$	$\mathbf{0.841^{0.003}}$	$0.810^{0.009}$
	DT	$0.727^{0.024}$	$\mathbf{0.789^{0.013}}$	$0.745^{0.034}$
Heart	TT	$0.908^{0.013}$	$\mathbf{0.915^{0.001}}$	$0.877^{0.010}$
	XGB	$0.906^{0.006}$	$\mathbf{0.930^{0.010}}$	$0.907^{0.013}$
	LRR	$0.897^{0.030}$	$\mathbf{0.926^{0.004}}$	$0.900^{0.007}$
	DT	$0.845^{0.022}$	$\mathbf{0.859^{0.006}}$	$0.830^{0.019}$
Respiratory	TT	$\mathbf{0.974^{0.008}}$	$0.920^{0.011}$	$0.926^{0.003}$
	XGB	$0.994^{0.006}$	$\mathbf{1.000^{0.000}}$	$0.998^{0.000}$
	LRR	$0.983^{0.010}$	$0.991^{0.000}$	$\mathbf{0.995^{0.000}}$
	DT	$\mathbf{0.997^{0.005}}$	$0.996^{0.004}$	$0.882^{0.214}$

Table 7: Mean complexity of student models across different shot configurations.

Dataset	Student	Shots							
		4	8	16	32	64	128	256	all
Breast	TT	125	118	129	139	144	151	145	141
	XGB	1286	1769	2527	3065	3985	5706	6710	8012
	LRR	28	22	54	27	32	44	38	42
	DT	127	107	240	169	239	344	397	360
Breast 2	TT	115	139	141	139	159	163	165	154
	XGB	476	1700	1835	2111	3720	3989	3796	3756
	LRR	46	65	53	70	52	103	115	98
	DT	95	247	266	244	228	259	267	319
Chemo	TT	145	150	231	236	270	244	277	284
	XGB	1316	405	2054	1027	1867	2402	4823	3611
	LRR	59	21	68	79	75	99	83	95
	DT	38	46	163	224	168	142	183	191
Colon	TT	116	109	136	154	168	149	167	168
	XGB	1264	1184	2223	3151	7915	6806	5242	5735
	LRR	26	18	37	47	58	70	77	78
	DT	64	48	112	305	472	387	507	499
Diabetes	TT	65	61	72	75	83	84	78	75
	XGB	1542	1194	3934	2750	3502	4423	5973	5933
	LRR	41	52	43	62	93	76	77	71
	DT	121	211	212	263	353	313	351	292
Heart	TT	99	122	136	147	150	149	147	152
	XGB	1553	877	1398	1783	3579	2536	6208	3731
	LRR	15	23	39	48	57	35	48	54
	DT	31	58	176	233	223	166	357	316
Respiratory	TT	468	410	385	318	328	278	355	310
	XGB	2010	1978	1528	847	674	1853	987	701
	LRR	83	54	30	26	9	6	8	13
	DT	147	146	91	35	35	38	26	23

Figure 3: Performance comparison for Breast Cancer dataset.

Figure 4: Performance comparison for Breast Cancer 2 dataset.

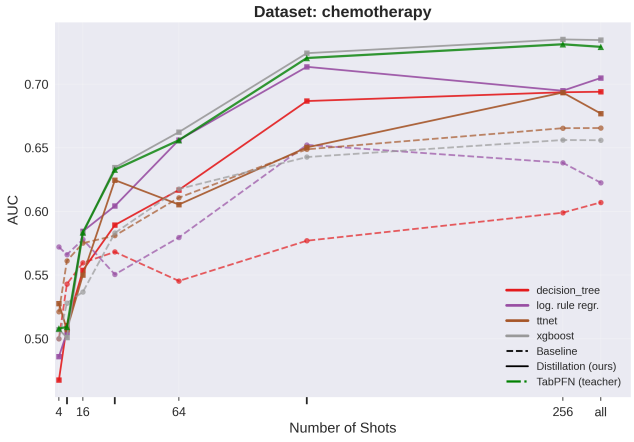

Figure 5: Performance comparison for Chemotherapy dataset.

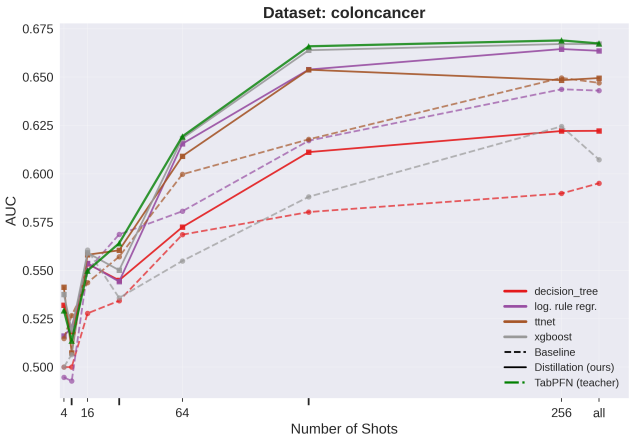

Figure 6: Performance comparison for Colon Cancer dataset.

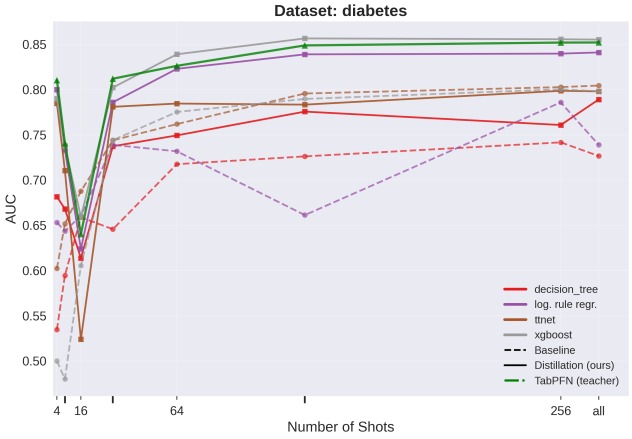

Figure 7: Performance comparison for Diabetes dataset.

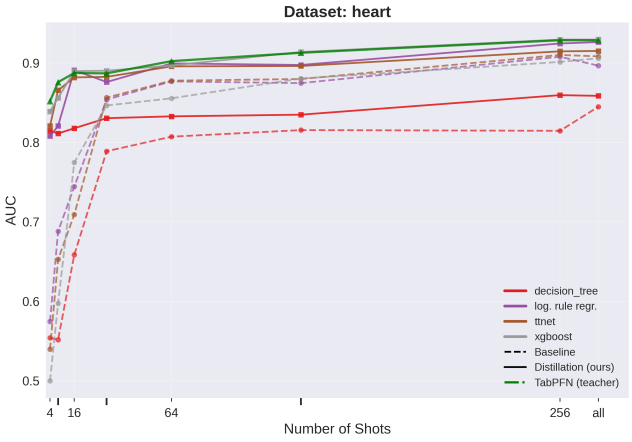

Figure 8: Performance comparison for Heart dataset.

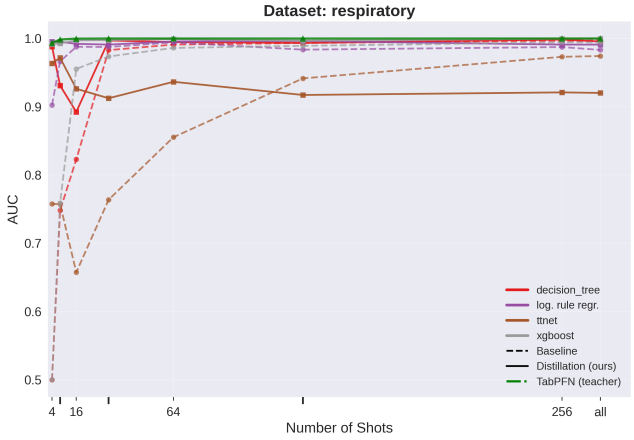

Figure 9: Performance comparison for Respiratory dataset.