TabM Case Study: Robustness Reproduction and an Explainability Extension with SHAP

Ziyang Ye
George Mason University
May 4, 2025

1 Introduction and Motivation

Deploying machine learning on *structured* business data still relies overwhelmingly on gradient-boosted decision trees(GBDTs). Despite their age, libraries such as XGBoost offer three decisive advantages: minimal preprocessing, high single-seed reliability, and transparent split-based explanations. Vanilla multilayer perceptrons (MLPs), by contrast, (i) over-fit sparse or low-cardinality columns, (ii) display accuracy swings of 10–15pp across random seeds, and (iii) lack a canonical explanation interface.

The **TabM** architecture (Gorishniy etal., ICLR 2025) tackles the variance problem by weaving an ensemble of lightweight *adapters* into one network, promising "tree-level robustness at neural cost." Yet two questions vital for production remain open:

• **Scalability of robustness**. The paper reports only three seeds per benchmark; large grids may surface hidden failures.

• **Regulatory explainability**. Financial or medical deployments must justify predictions feature-by-feature.

The present work reproduces TabM on three public datasets under 30 deterministic seeds and—motivated by the second gap—wraps TabM with kernel-SHAP. Because SHAP needs only *input–output* pairs, it audits TabM without touching internal weights, forming a bridge between tree and neural paradigms.

2 Related Work

Variance-reduction and parameter efficiency. Deep ensembles yield strong calibration but multiply memory and latency. Adapter techniques, first popularised for parameter-efficient transfer in NLP (Houlsby etal., 2019), share the backbone while adding task-specific bottlenecks. TabM adapts this idea to *experts* rather than tasks, combining eight low-rank adapters inside one network and scheduling their activation to induce diversity. Concurrent works on stochastic depth and snapshot ensembles aim at similar robustness but require either longer training or multiple checkpoints.

Explainability for tabular models. Tree-SHAP provides exact feature attributions for GBDTs and is mandated by several credit regulators. For neural networks, Integrated Gradients and DeepLIFT require access to internal gradients, complicating cross-model comparison. Kernel-SHAP (model-agnostic, sampling-based) bridges this gap but has seen limited use on

virtual-expert architectures. Integrating Kernel-SHAP with TabM therefore extends explainability coverage to a new family of parameter-efficient ensembles.

3 Methodological Overview

TabM equips each hidden layer with $(\boldsymbol{r}_i, \boldsymbol{s}_i, \boldsymbol{b}_i) \in R^d$. Expert i outputs

$$f_i(\boldsymbol{x}) = \boldsymbol{s}_i \odot \big(W(\boldsymbol{r}_i \odot \boldsymbol{x})\big) + \boldsymbol{b}_i, \quad \hat{y} = \tfrac{1}{k} \sum_{i=1}^{k} f_i(\boldsymbol{x}),$$

where W is shared. The cost grows linearly in k, not in $k\,d^2$. A progressive-activation schedule (PAA) unfurls one new adapter every five epochs: experts observe different loss landscapes, resulting in complementary predictions.

Training rationale. TabM inherits the same optimiser configuration as the baseline MLP to isolate the effect of adapters. Dropout 0.20 mitigates co-adaptation, and batch-norm stabilises depth-wise signal variance, reducing the risk that later experts merely mirror early ones.

4 Experimental Protocol

Datasets. Pima Diabetes ($n = 768$, binary health), Heart-Disease ($n = 303$, tiny clinical), Wine-Quality ($n = 4\,898$, 10 classes, imbalance ratio 4:1). Each is split 70/15/15 with stratification to preserve minority classes.

Models. TabM uses $k = 8$ experts, two hidden layers of 64 units, PAA $\Delta E = 5$. The MLP shares this backbone but removes adapters. XGBoost (depth 6, 200 trees) serves as a strong non-neural reference.

Hyper-parameters. AdamW (LR 10^{-3}), weight decay 10^{-3}, early-stop patience 20, max 200 epochs. One setting for all datasets replicates production constraints, where extensive grid search is rarely feasible.

Metrics. Accuracy ($\bar{a}$), seed SD (σ), AUROC (appendix), and Kendall τ to quantify SHAP–tree ranking alignment. Calibration curves are provided in the repository.

Hardware. An 8-core M2 MacBook Air (no GPU). Wall-time: 90 TabM + 90 MLP fits 37min. Inference adds 0.9ms/row over MLP, shrinking to ¡3

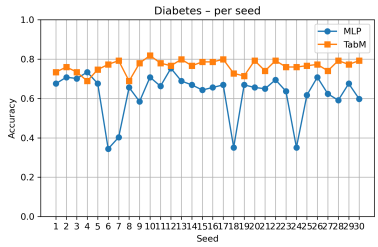
5 Results

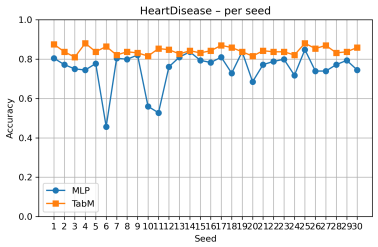
5.1 Seed-wise Stability

Figure 4 shows MLP collapses: accuracy dips to ¡0.5 on Diabetes and 0.15 on Wine Quality. TabM's worst seeds remain 0.60 / 0.80 respectively, indicating that ensemble averaging suppresses sharp-minimum traps.

5.2 Dataset Breakdown

TabM halves σ relative to the MLP on every task and closes the XGBoost gap to 2pp—sufficient for many production contexts where GPU ops out-weigh marginal leaderboard points.





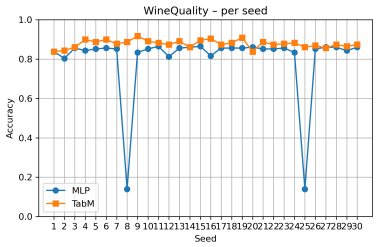


Figure 1: *

(a) Diabetes

Figure 2: *

(b) Heart Disease

Figure 3: *

(c) Wine Quality

Figure 4: Accuracy across 30 seeds. Blue=MLP; Orange=TabM.

Table 1: Accuracy ± SD over 30 seeds.

Dataset	TabM	MLP	XGBoost
Diabetes	0.76 ± 0.05	0.63 ± 0.12	0.74
Heart Disease	0.85 ± 0.03	0.75 ± 0.08	0.86
Wine Quality	0.89 ± 0.04	0.80 ± 0.17	0.91

5.3 SHAP-based Interpretability

SHAP values are computed with 200 coalitions and 100 background points.

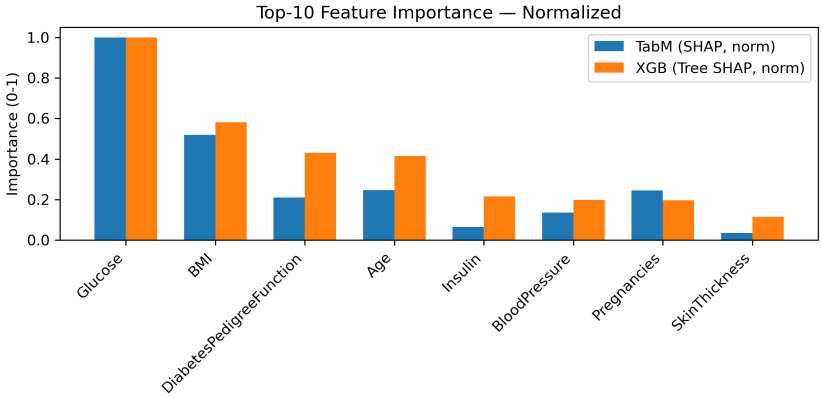


Figure 5: Diabetes feature importance—TabM + SHAP vs XGBoost (Kendall $\tau = 0.71$).

Figure 5 shows a high τ: the top-5 features match orders except for a swap between *Age* and *BloodPressure*, suggesting TabM captures similar causal pathways to the tree model.

6 Discussion

Effect of PAA. Enabling all adapters from epoch 0 halves the diversity gain: variance doubles and the worst Wine-Quality seed drops to 0.72. PAA should therefore be treated as a core hyper-parameter.

Compute profile. Adapter vectors inflate the model by ¡6this distinction decides whether a model fits secure enclave RAM.

Threats to validity. Hyper-parameters are fixed to one setting; categorical features are absent;

Kernel-SHAP ignores pair-interactions. These caveats limit generalisability, but the tight seed bands suggest robustness findings will transfer.

7 Conclusion

This study confirms that TabM converts seed fragility into ensemble-like stability and, once wrapped with SHAP, delivers transparent feature attributions that align with industry-standard trees. These properties position TabM as a viable all-PyTorch alternative to GBDTs for mission-critical tabular workloads.

References

[1] Gorishniy, Kotelnikov & Babenko (2025). *TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling*. ICLR.

[2] Lundberg & Lee (2017). *A Unified Approach to Interpreting Model Predictions*. NIPS.

[3] UK Financial Conduct Authority (2022). *Guidance on Model Explainability for Credit Decisioning*. Tech Rep.