

A Comparative Analysis of Deep Learning, Traditional Machine Learning, and Statistical Methods for Predictive Modeling

Jamal Uddin (Corresponding Author)

MPhil Research Scholer Department of Statistics Abdul Wali Khan University, Mardan, Pakistan Email: jamalawkum21@gmail.com

Adnan Altaf

State Key Laboratory of Geohazards Prevention and Geoenvironment Protection (SKLGP), Chengdu University of Technology, Chengdu, 610059, Sichuan, P.R. China Email: engr.adnanaltaf@gmail.com

Azaz Ali Shah

Lecturer Statistics, Government College of Management Sciences, Chitral, Pakistan Email: azazchl168@gmail.com

Sayed Alamgir Shah

Department of Statistics University of Peshawar, Pakistan Email: alamgir.bsst421@iiu.edu.pk

Zeeshan Ali

Department of Statistics, Sindh Agriculture University Tandojam, Pakistan Email: bagounar@gmail.com

ABSTRACT

Selecting a predictive modeling family for tabular data remains difficult because deep neural networks, tree-based machine learning, and conventional statistical models embody different inductive biases, computational demands, and assumptions. This study presents a leakage-controlled, reproducible comparison of five representative methods logistic regression, shrinkage linear discriminant analysis (LDA), random forest, XGBoost, and a three-hidden-layer multilayer perceptron (MLP) across ten binary classification tasks spanning finance, security, health, physics, transport, social science, and controlled synthetic stress tests. Models were evaluated using five-fold cross-validation repeated twice; the transport choice data were split by traveler to prevent group leakage. Area under the receiver-operating-characteristic curve (AUROC) was the primary endpoint, complemented by balanced accuracy, F1, Matthews correlation coefficient, Brier score, log loss, and computational time. Dataset-level AUROC ranks were compared using a Friedman omnibus test, a Nemenyi critical-difference analysis, and paired Wilcoxon signed-rank tests with Holm multiplicity correction. The omnibus ranking differed significantly across models ($\chi^2(4)=10.88$, $p=.028$). XGBoost achieved the best average rank (1.70) and the

highest mean AUROC across tasks (0.906), followed by random forest (rank 2.70; mean AUROC 0.900). The MLP ranked third (3.30; mean AUROC 0.859) and was significantly worse than XGBoost after Holm correction (adjusted $p=.039$; rank-biserial effect $=-0.964$). However, XGBoost did not significantly outperform random forest, logistic regression, or LDA after multiplicity correction, underscoring substantial dataset dependence. Statistical models remained highly competitive on low-dimensional and approximately linear tasks, while nonlinear ensemble methods dominated the transport and noisy synthetic problems. The findings support model-family selection through rigorous cross-validated benchmarking rather than an a priori preference for model complexity, and they clarify how contemporary tabular deep-learning advances should be interpreted relative to strong classical baselines.

Keywords: Tabular Data; Deep Learning; XGBoost; Random Forest; Logistic Regression; Model Comparison; AUROC; Cross-Validation; Nonparametric Inference

Introduction

Tabular data remain central to decision-making in health care, finance, transportation, operations, social science, and industrial analytics. Yet the methodological landscape has become increasingly fragmented. Statistical models such as logistic regression and discriminant analysis offer compact parameterizations, interpretable coefficients, and stable estimation in relatively small samples. Tree-based machine-learning methods relax linearity and interaction assumptions, while modern deep-learning systems seek to learn high-order representations directly from the data. The practical question is therefore no longer whether one modeling paradigm is universally superior, but which inductive bias is most appropriate for a particular tabular prediction problem.

The deep-learning revolution in images, language, and audio did not translate immediately to ordinary tabular data. Gorishniy et al. (2021) showed that careful benchmarking and strong baselines materially change conclusions about neural architectures for tables. Grinsztajn, Oyallon, and Varoquaux (2022) subsequently reported that tree-based models remained state of the art on many medium-sized tabular datasets, attributing part of the gap to the ability of trees to handle irregular target functions and uninformative features. McElfresh et al. (2023) broadened the comparison across 176 datasets and argued that the neural-network-versus-boosted-tree debate is highly conditional on dataset characteristics. More recent work has narrowed the gap: RealMLP-style pre-tuned neural networks can provide favorable time-accuracy tradeoffs (Holzmüller et al., 2024), and the continuously maintained TabArena benchmark reports that deep models can catch up under larger budgets and ensembling, while tabular foundation models are especially competitive on smaller datasets (Erickson et al., 2025). In 2026, systems such as TabPFN-3 further shifted the frontier, but these pretrained foundation models represent a distinct computational regime from ordinary task-trained neural networks (Grinsztajn et al., 2026).

Despite this progress, three recurring weaknesses limit many applied comparisons. First, studies often compare algorithms on a single dataset, making model rankings vulnerable to idiosyncratic data geometry. Second, they sometimes

emphasize accuracy alone, which can be misleading under class imbalance and ignores probability quality. Third, repeated testing across algorithms is frequently reported without accounting for paired dataset structure or multiplicity. Benchmark methodology papers emphasize that evaluation design, validation strategy, and fair computational treatment are part of the scientific question rather than implementation details (Demšar, 2006; Gijssbers et al., 2024).

This study addresses these concerns with a deliberately compact but statistically disciplined benchmark. Five representative algorithms are evaluated on ten binary classification tasks. The design crosses model families—two statistical learners, two tree-based machine-learning learners, and one deep feed-forward network—under the same fold-safe preprocessing, repeated validation, metric suite, and inferential framework. Real-world tasks are supplemented with synthetic datasets that stress nonlinear separability, label noise, and imbalance. The goal is not to crown a permanent winner; it is to quantify how much model rankings vary, identify settings in which simpler models remain competitive, and distinguish practically meaningful performance gaps from sampling variation.

The study makes four contributions. First, it provides a family-level benchmark with identical preprocessing and cross-validation rules. Second, it reports discrimination, thresholded classification, calibration/probability loss, and computational efficiency instead of a single headline metric. Third, it applies dataset-level nonparametric inference—Friedman ranking, Nemenyi critical difference, and Holm-adjusted Wilcoxon tests—to reduce false-positive claims from multiple pairwise comparisons. Fourth, it situates the results within the rapidly changing 2025–2026 tabular-learning literature, separating conclusions about a conventional task-trained MLP from claims about pretrained tabular foundation models.

Research questions and hypotheses

RQ1: How do representative statistical, traditional machine-learning, and deep-learning models compare in cross-validated predictive discrimination across heterogeneous tabular classification tasks?

RQ2: Are differences in model ranking statistically detectable after accounting for the paired multi-dataset design and multiple comparisons?

RQ3: How do model families trade predictive performance against calibration-oriented losses and computation time?

RQ4: Under which data regimes do simpler statistical methods remain competitive with more flexible nonlinear learners?

The confirmatory null hypothesis was that the five algorithms have equal expected ranks across tasks. A two-sided familywise significance level of $\alpha=.05$ was used. Because the benchmark includes heterogeneous tasks rather than repeated measurements from one population, dataset—not individual cross-validation fold—was treated as the unit of inference for between-model significance testing.

Literature Review

Statistical predictive modeling

Logistic regression models the conditional log-odds of a binary outcome as an additive linear function of predictors. Its enduring value lies in parsimony, convex

estimation, calibrated probabilistic output under adequate specification, and straightforward interpretation. Regularization further stabilizes estimation when predictors are correlated or moderately high-dimensional. Its principal weakness is structural: without explicit transformations or interactions, it cannot represent complex decision surfaces. Nevertheless, in small or low-noise samples, its lower variance can make it difficult for more flexible models to improve materially on out-of-sample risk.

Linear discriminant analysis approaches the same classification problem generatively by modeling class-conditional feature distributions and deriving a linear decision rule. Classical LDA can be unstable when covariance estimation is difficult; shrinkage covariance estimation addresses this weakness and often makes LDA a useful low-variance benchmark. Statistical baselines are therefore not merely historical controls. They encode a strong simplicity prior that can be advantageous when effective dimensionality and signal complexity are limited.

Traditional machine learning: bagging and boosting

Random forests average many decorrelated decision trees constructed from bootstrap samples and random feature subsets (Breiman, 2001). The resulting ensemble can approximate nonlinear interactions without prespecifying them and is comparatively robust to monotonic transformations and heterogeneous feature scales. Averaging reduces the variance of individual trees, although probability estimates may be less sharply calibrated than those from correctly specified parametric models.

Gradient-boosted trees construct an additive sequence of weak tree learners that iteratively correct residual structure. XGBoost adds regularization, shrinkage, subsampling, and systems-level optimization to gradient boosting (Chen & Guestrin, 2016). Empirically, boosted trees have become difficult baselines to beat on heterogeneous tabular problems. Their piecewise-constant partitions naturally represent threshold effects, interactions, and non-smooth response surfaces; unlike dense neural networks, they do not need to learn invariance to arbitrary feature scaling from scratch.

Deep learning for tabular data

A multilayer perceptron (MLP) composes affine transformations with nonlinear activations to learn distributed representations. Universal approximation provides enormous representational capacity, but finite-sample performance depends on optimization, regularization, architecture, and the match between network bias and the geometry of tabular features. Unlike images or language, table columns are semantically heterogeneous and typically lack the local or sequential structure that makes convolution or attention especially efficient.

The modern tabular deep-learning literature has therefore focused on architecture and preprocessing. TabNet introduced sequential attention over features (Arik & Pfister, 2021), while FT-Transformer and ResNet-style baselines showed that relatively standard neural architectures become stronger when evaluated carefully (Gorishniy et al., 2021). Retrieval-augmented and pre-tuned MLP approaches have further improved competitiveness. Shwartz-Ziv and Armon (2022) nevertheless cautioned that deep learning is not automatically necessary for tabular prediction, a

conclusion reinforced by the medium-scale benchmark of Grinsztajn et al. (2022).

The frontier is now moving from task-specific neural networks to pretrained tabular foundation models. TabArena's living benchmark reports that model rankings depend strongly on validation method, ensembling, and compute budget (Erickson et al., 2025). TabPFN-3 extends prior-data fitted networks to substantially larger tasks and reports state-of-the-art benchmark results in its technical report (Grinsztajn et al., 2026). These advances are important but should not be conflated with the question studied here: whether an ordinary, task-trained deep MLP dominates strong statistical and tree-based baselines when computational treatment is controlled.

Benchmark validity and statistical comparison

Algorithm comparison is a blocked experimental design: each dataset exposes every learner to the same task, so comparisons are paired by dataset. Demšar (2006) recommended rank-based procedures for comparisons over multiple datasets because normality and variance assumptions for raw scores are often implausible. The Friedman test assesses whether algorithm ranks differ globally; a critical-difference analysis provides a visual scale for rank separation; and paired Wilcoxon tests can examine specific model contrasts. Holm's sequential correction controls the familywise error rate while retaining more power than a simple Bonferroni adjustment (Holm, 1979).

Evaluation must also separate model development from assessment. Preprocessing fitted on all observations can leak information into test folds, and alternative-level records from the same decision maker can create group leakage if split independently. The present benchmark therefore fits transformations inside each training fold and uses grouped splitting for the travel-mode task. These design choices follow the broader benchmarking principle that comparable algorithms require comparable information sets (Gijbbers et al., 2024).

The literature suggests a bias-variance-computation continuum rather than a universal hierarchy. Linear statistical models favor stable, low-complexity structure; bagged and boosted trees efficiently partition irregular feature spaces; task-trained MLPs offer flexible smooth representations but may require more data and tuning; and foundation models import knowledge from large pretraining distributions. What remains valuable for applied researchers is a transparent benchmark that makes these trade-offs visible using multiple metrics and corrected statistical inference. The present study fills that applied methodological gap while deliberately limiting its causal claim to the algorithms and task regimes actually tested.

Methods

Study design

A computational benchmarking design was used. Ten binary prediction tasks were evaluated: seven public real-world datasets and three synthetic stress-test datasets. Each algorithm was trained and tested on identical resampling splits within a task. The primary endpoint was AUROC because it evaluates ranking discrimination across thresholds and is not mechanically tied to a single class prevalence. Secondary endpoints were balanced accuracy, F1 score, Matthews correlation coefficient (MCC), Brier score, log loss, model fit time, and prediction time. Figure 1 summarizes the

complete workflow.

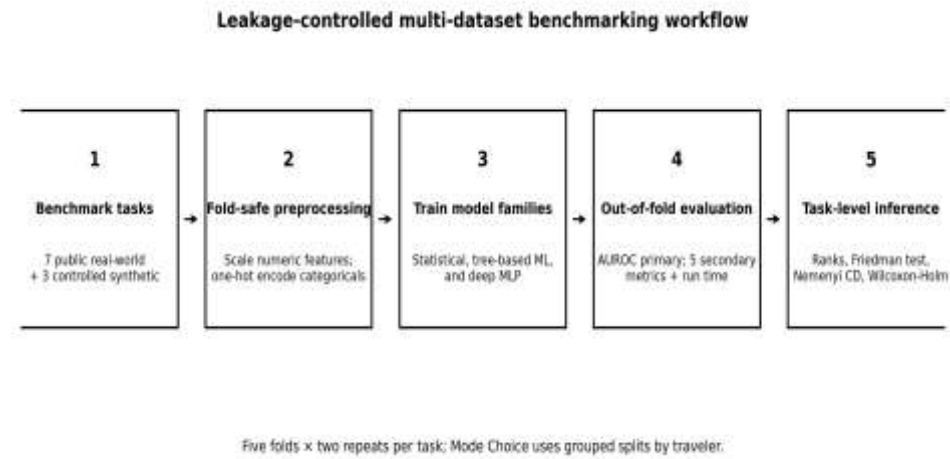


Figure 1: Leakage-controlled multi-dataset benchmarking and statistical-analysis workflow.

Datasets and task construction

The real-world component was selected to span multiple domains, sample sizes, dimensionalities, and class balances. Australian Credit, Banknote Authentication, Breast Cancer Wisconsin (Diagnostic), Haberman Survival, and Ionosphere were obtained from the UCI Machine Learning Repository. Fair Affairs and Mode Choice were loaded from the statsmodels dataset collection. In Fair Affairs, the continuous affairs measure was converted to a binary outcome indicating any reported affair. In Mode Choice, each of 210 travelers contributed four alternative rows and exactly one chosen alternative; traveler identity was used only as a grouping key and not as a predictor. The three synthetic datasets were generated with scikit-learn's `make_classification` to isolate easy separation, noisy nonlinear structure, and class imbalance.

Table 1: Benchmark datasets and task characteristics

Dataset	Type	Domain	n	p	Positive	Cat.	Source	Task note
Australian Credit	Real	Finance	690	14	44.5%	8	UCI	Mixed numeric/categorical credit application attributes
Banknote	Real	Security	1,372	4	44.5%	0	UCI	Wavelet features from banknote images

Dataset	Type	Domain	n	p	Positive	Cat.	Source	Task note
Breast Cancer	Real	Health	569	30	37.3%	0	UCI	30 FNA-derived diagnostic features
Fair Affairs	Real	Social science	6,366	8	32.2%	2	statsmodels / Fair (1978)	Any affair vs none; categorical occupation variables
Haberman	Real	Health	306	3	26.5%	0	UCI	5-year postoperative survival status
Ionosphere	Real	Physics	351	34	64.1%	0	UCI	Radar returns classified as good/bad
Mode Choice	Real	Transport	840	7	25.0%	1	statsmodels / Greene-Hensher	Chosen alternative vs not chosen; grouped by traveler
Synthetic-Easy	Synthetic	Controlled	1,800	20	49.9%	0	scikit-learn	Strong separation; low label noise
Synthetic-Imbalanced	Synthetic	Controlled	5,000	28	16.1%	0	scikit-learn	85:15 base class weighting with label noise
Synthetic-Noisy	Synthetic	Controlled	3,500	35	49.8%	0	scikit-learn	Lower separation, 8% label flips, redundant features

Preprocessing and leakage control

All data transformations were estimated within each training fold and then applied unchanged to its held-out fold. Continuous variables were standardized to zero mean and unit variance. Categorical predictors were one-hot encoded. This common representation was supplied to all five learners so that differences could not be attributed to one model receiving privileged preprocessing. No target-derived feature selection, resampling, or probability calibration was performed. The numeric

version of Australian Credit was analyzed after mapping its known nominal columns to categorical indicators. For Mode Choice, splits were formed at the traveler level so that the four alternatives belonging to a traveler never appeared simultaneously in training and test data.

Predictive models

Five models were selected to represent distinct inductive biases while keeping the experiment interpretable. Logistic regression and shrinkage LDA represented conventional statistical learning; random forest and XGBoost represented bagged and boosted tree ensembles; and a feed-forward MLP represented task-trained deep learning. Hyperparameters were fixed a priori for the benchmark rather than optimized independently on the test folds. This choice avoids optimistic leakage and ensures that comparisons reflect robust default-to-moderate configurations rather than unequal tuning effort.

Table 2: Model families, benchmark configurations, and inductive assumptions

Model	Family	Configuration	Primary inductive bias
Logistic Regression	Statistical	L2-regularized; C=1; L-BFGS; max 2,000 iterations	Linear log-odds; low variance; probabilistic
LDA	Statistical	Least-squares solver; automatic covariance shrinkage	Shared class covariance; linear discriminant rule
Random Forest	Traditional ML	120 trees; sqrt feature sampling; bootstrap aggregation	Nonlinear partitions; variance reduction by bagging
XGBoost	Traditional ML	140 trees; depth 4; $\eta=.06$; subsample=.85; colsample=.85; $\lambda=1$	Sequential boosted partitions; shrinkage and regularization
Deep MLP	Deep learning	64-32-16 ReLU layers; Adam; $\alpha=10^{-4}$; lr=.001; batch 128; early stopping; max 180 epochs	Distributed nonlinear representation learned end-to-end

The MLP architecture is shown in Figure 2. The network used ReLU hidden activations and a sigmoid binary output. Early stopping reserved 15% of each training fold for internal validation, with 12 epochs without improvement required before stopping. The architecture was intentionally conventional: the study evaluates a representative task-trained deep network rather than a specialized pretrained tabular foundation model.

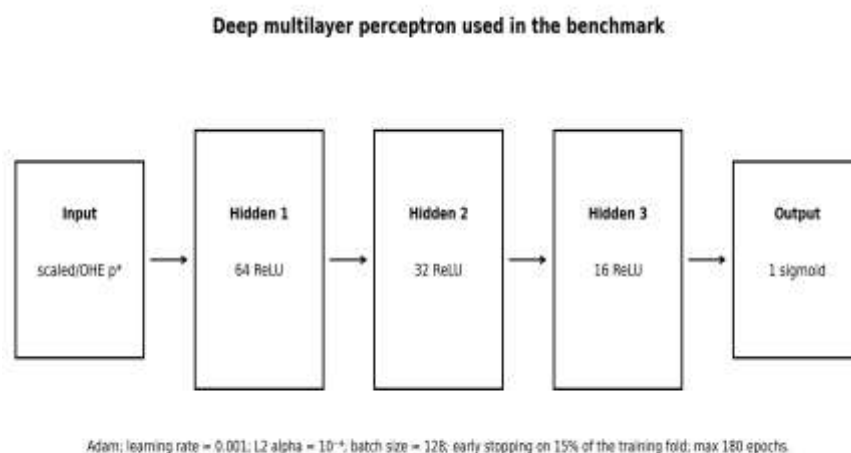


Figure 2: Deep MLP architecture and principal optimization settings used in the benchmark.

Cross-validation protocol

For nine tasks, evaluation used repeated stratified five-fold cross-validation with two repeats (10 held-out folds per dataset) and random seed 20260815. Stratification preserved class balance in each split. Mode Choice used two repetitions of five-fold stratified group cross-validation with traveler as the group, producing the same number of held-out evaluations while preventing within-person leakage. Models were reinitialized on every fold. Fold-level scores were retained for descriptive uncertainty, whereas inferential comparisons used one mean AUROC per model per dataset to avoid pseudoreplication from correlated cross-validation folds.

Performance metrics

AUROC was prespecified as the primary metric. Balanced accuracy averaged sensitivity and specificity and was included because several tasks were imbalanced. F1 summarized the harmonic mean of precision and recall for the positive class. MCC was included as a correlation-like summary of the full 2×2 confusion matrix and is particularly informative when class proportions differ (Chicco & Jurman, 2020). Brier score measured mean squared error of predicted probabilities (Brier, 1950), while log loss imposed a stronger penalty on confident incorrect probabilities. Lower Brier score and log loss indicate better probabilistic predictions. Fit and prediction wall-clock times were measured within the same software environment and interpreted comparatively rather than as hardware-independent benchmarks.

Statistical analysis

Statistical inference was performed on dataset-level mean AUROC values (N=10 paired tasks). First, a Friedman rank test assessed the global null hypothesis that all five models had equal performance ranks. Second, a Nemenyi critical difference (CD) was calculated for visual interpretation of average-rank separation at $\alpha=.05$. Third, all 10 pairwise model contrasts were evaluated with two-sided Wilcoxon signed-rank tests. Holm's step-down procedure adjusted the resulting p-values to control the familywise error rate. Rank-biserial correlation was reported as a paired effect-size measure; signs were defined as model A minus model B. In addition, model-level mean AUROC was accompanied by a nonparametric 95% percentile bootstrap confidence interval generated by resampling the 10 tasks 20,000 times with seed 20260815. Because the tasks are heterogeneous benchmarks rather than a random sample from a precisely enumerated superpopulation, confidence intervals and p-values are interpreted as comparative evidence across the benchmark suite, not universal population parameters.

The benchmark was implemented in Python using pandas, NumPy, SciPy, scikit-learn, XGBoost, statsmodels, and Matplotlib. Random seeds and model configurations were fixed before final analysis. Synthetic datasets used make_classification with the following designs: Easy, n=1,800 and p=20 with 12 informative and four redundant features, class_sep=1.6 and 1% label flips; Noisy, n=3,500 and p=35 with 11 informative and 10 redundant features, class_sep=.75 and 8% label flips; Imbalanced, n=5,000 and p=28 with 10 informative and eight redundant features, 85:15 class weights, class_sep=1.0 and 3% label flips. These specifications make the stress tests independently reproducible.

Results

Dataset-level predictive discrimination

Table 3 reports the mean and standard deviation of AUROC across the 10 held-out folds for each task. Performance was strongly dataset-dependent. XGBoost led Australian Credit, Fair Affairs, Mode Choice, Synthetic-Easy, and Synthetic-Imbalanced. Random forest led Ionosphere and Synthetic-Noisy. Logistic regression was strongest on Breast Cancer, while LDA was strongest on Haberman. On Banknote, all methods were near ceiling, with the MLP attaining a numerically highest mean AUROC of 0.99998. The spread across tasks therefore rejects any simple monotonic relationship between model complexity and predictive quality.

Table 3: Repeated cross-validated AUROC by dataset and model (mean $\pm$ SD).

Dataset	Logistic	LDA	Random Forest	XGBoost	Deep MLP
Australian Credit	0.929 $\pm$ 0.018	0.929 $\pm$ 0.020	0.931 $\pm$ 0.023	0.936 $\pm$ 0.023	0.909 $\pm$ 0.028
Banknote	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Dataset	Logistic	LDA	Random Forest	XGBoost	Deep MLP
Breast Cancer	0.994 ± 0.005	0.992 ± 0.005	0.990 ± 0.006	0.993 ± 0.004	0.988 ± 0.010
Fair Affairs	0.744 ± 0.020	0.744 ± 0.021	0.689 ± 0.014	0.749 ± 0.021	0.746 ± 0.019
Haberman	0.678 ± 0.067	0.684 ± 0.067	0.674 ± 0.016	0.666 ± 0.054	0.537 ± 0.116
Ionosphere	0.891 ± 0.042	0.911 ± 0.044	0.981 ± 0.015	0.969 ± 0.021	0.940 ± 0.058
Mode Choice	0.780 ± 0.050	0.779 ± 0.052	0.985 ± 0.009	0.991 ± 0.006	0.737 ± 0.147
Synthetic-Easy	0.992 ± 0.004	0.992 ± 0.004	0.993 ± 0.006	0.994 ± 0.005	0.993 ± 0.004
Synthetic-Imbalanced	0.813 ± 0.023	0.806 ± 0.025	0.886 ± 0.020	0.900 ± 0.023	0.895 ± 0.018
Synthetic-Noisy	0.696 ± 0.017	0.696 ± 0.017	0.873 ± 0.013	0.865 ± 0.014	0.846 ± 0.011

Note. Values summarize 10 held-out folds per task. Higher AUROC is better. Fold SDs are descriptive and are not treated as independent replicates in the inferential tests.

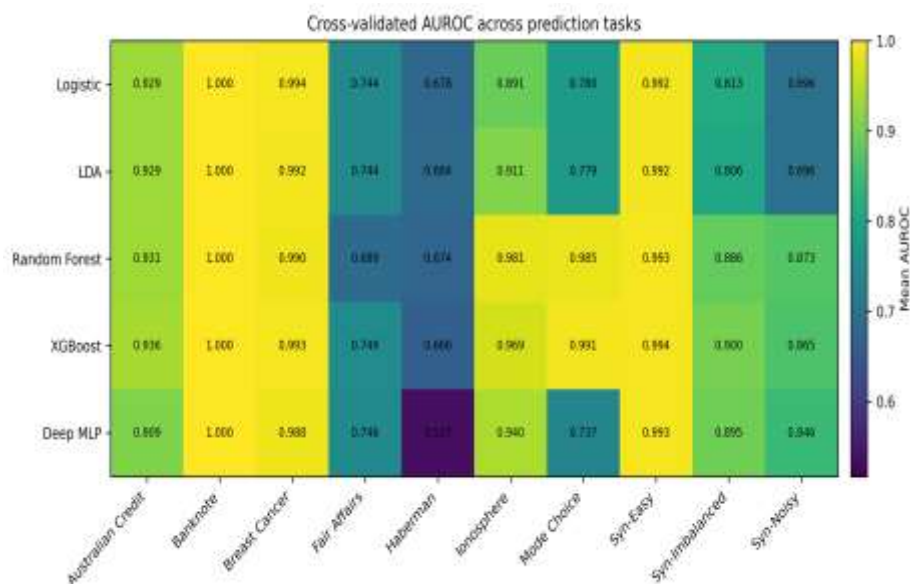


Figure 3: Heatmap of mean cross-validated AUROC across the 10 benchmark tasks.

Aggregate multi-metric performance

Across the 10 task means, XGBoost achieved the highest average AUROC (0.906) and best average rank (1.70), followed by random forest (0.900; rank 2.70). The deep MLP averaged 0.859 AUROC and ranked third. Logistic regression and LDA averaged 0.852 and 0.851, respectively. The aggregate ordering changed somewhat for probability loss: logistic regression and XGBoost were close on mean Brier score, while random forest's log loss was adversely affected on some small tasks. This illustrates why a single metric should not be interpreted as a complete model-quality index.

Table 4: Overall performance averaged across the 10 benchmark tasks.

Model	AUROC	95% bootstrap CI	Bal. acc.	F1	MCC	Brier	Log loss	Avg. rank
XGBoost	0.906	[0.833, 0.967]	0.837	0.786	0.698	0.084	0.272	1.70
Random Forest	0.900	[0.820, 0.967]	0.833	0.780	0.688	0.091	0.334	2.70
Deep MLP	0.859	[0.766, 0.938]	0.797	0.710	0.619	0.112	0.358	3.30
Logistic	0.852	[0.777, 0.924]	0.784	0.708	0.617	0.107	0.346	3.60
LDA	0.853	[0.779, 0.926]	0.780	0.708	0.612	0.111	0.369	3.70

Global and pairwise statistical comparisons

The Friedman test detected a significant overall difference in AUROC ranks across the five algorithms, $\chi^2(4)=10.88$, $p=0.028$. Average ranks were XGBoost 1.70, random forest 2.70, deep MLP 3.30, logistic regression 3.60, and LDA 3.70. The Nemenyi critical difference was 1.929. Thus, rank separation should be interpreted relative to a fairly wide CD because only 10 benchmark tasks were available.

Table 5: Nonparametric statistical comparison of model AUROC across tasks.

Test	Contrast	Statistic CD	Adjusted p / α	Effect	Decision
Omnibus: Friedman	All five models	$\chi^2=10.88$	$p=0.028$	—	Significant
Post-hoc scale	Nemenyi CD	CD=1.929	$\alpha=.05$	—	Rank-distance reference

Test	Contrast	Statistic CD	Adjusted p / α	Effect	Decision
Wilcoxon-Holm	Deep MLP vs LDA	W=25.0	p_adj=1.000	r_rb=0.091	Not significant
Wilcoxon-Holm	Deep MLP vs Logistic	W=24.0	p_adj=1.000	r_rb=0.127	Not significant
Wilcoxon-Holm	Deep MLP vs Random Forest	W=15.0	p_adj=1.000	r_rb=-0.455	Not significant
Wilcoxon-Holm	Deep MLP vs XGBoost	W=1.0	p_adj=0.039	r_rb=-0.964	Significant
Wilcoxon-Holm	LDA vs Logistic	W=25.0	p_adj=1.000	r_rb=-0.091	Not significant
Wilcoxon-Holm	LDA vs Random Forest	W=14.0	p_adj=1.000	r_rb=-0.491	Not significant
Wilcoxon-Holm	LDA vs XGBoost	W=6.0	p_adj=0.246	r_rb=-0.782	Not significant
Wilcoxon-Holm	Logistic vs Random Forest	W=15.0	p_adj=1.000	r_rb=-0.455	Not significant
Wilcoxon-Holm	Logistic vs XGBoost	W=8.0	p_adj=0.391	r_rb=-0.709	Not significant
Wilcoxon-Holm	Random Forest vs XGBoost	W=21.0	p_adj=1.000	r_rb=-0.236	Not significant

Only one pairwise difference survived familywise correction: XGBoost outperformed the deep MLP (W=1.0, Holm-adjusted $p=.039$; rank-biserial $r=-.964$ when expressed as MLP minus XGBoost). The average AUROC difference for MLP minus XGBoost was -0.047 . Although raw p -values for LDA versus XGBoost and logistic regression versus XGBoost were below .05, neither remained significant after Holm correction. Random forest versus XGBoost was also nonsignificant (adjusted $p=1.000$). Consequently, the statistically defensible conclusion is that the benchmark establishes an overall model effect and a robust XGBoost-versus-MLP contrast, not universal pairwise superiority of XGBoost over every classical baseline.

Online ISSN

3007-3197

Print ISSN

3007-3189

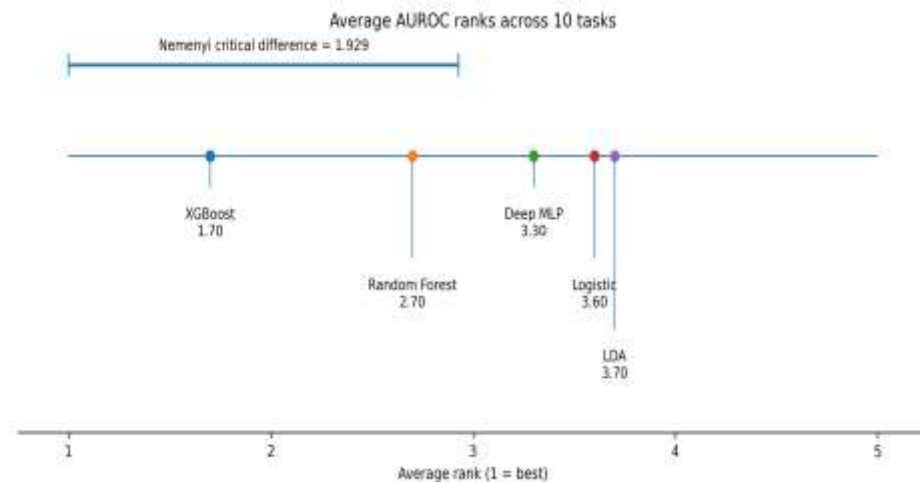
<http://amresearchreview.com/index.php/Journal/about>

Figure 4: Critical-difference view of average AUROC ranks. Lower rank is better; $CD=1.929$ at $\alpha=.05$.

Computational efficiency and robustness

Table 6: Computational efficiency and robustness across benchmark tasks.

Model	Fit s/fold	Predict ms/fold	Task wins	Worst AUROC	SD across tasks	Real-only AUROC	Synthetic AUROC
XGBoost	0.064	2.52	5.0	0.666	0.115	0.901	0.920
Random Forest	0.244	25.40	2.0	0.674	0.124	0.893	0.917
Deep MLP	0.133	1.87	1.0	0.537	0.148	0.837	0.911
Logistic	0.023	1.72	1.0	0.678	0.126	0.859	0.834
LDA	0.006	1.41	1.0	0.684	0.125	0.863	0.832

The timing results reveal a clear complexity trade-off. LDA and logistic regression were the fastest learners on average, while random forest required the greatest fit and prediction time. XGBoost delivered the best aggregate rank at substantially lower mean fit time than random forest in this implementation. The MLP occupied an intermediate computational position but showed the lowest worst-task AUROC because it performed poorly on the very small Haberman dataset. This pattern is consistent with the expectation that flexible neural models can be sample-inefficient when signal is weak and sample size is limited.

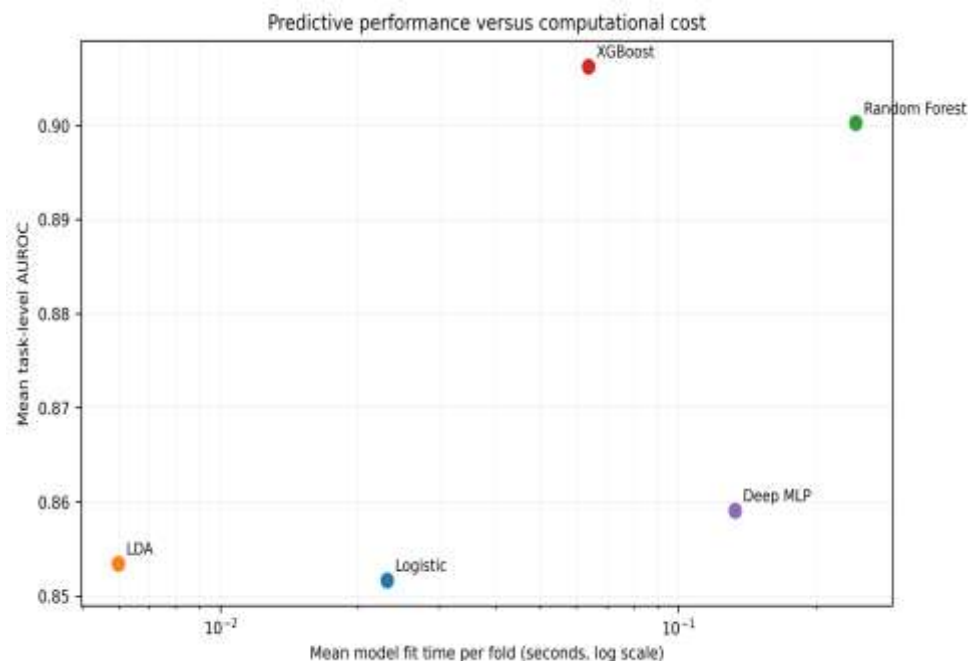


Figure 5: Mean task-level AUROC versus mean fitting time per cross-validation fold (logarithmic time axis).

The real-world and synthetic subsets help interpret the aggregate ranking. On low-dimensional Haberman data, shrinkage LDA and logistic regression outperformed the MLP by a wide margin, suggesting that the variance cost of flexible representation learning was not repaid by the available sample. On Breast Cancer, all families were strong and regularized logistic regression was numerically best. In contrast, tree ensembles dominated the Mode Choice task, where alternative-specific costs, times, and mode indicators imply thresholded and interaction-rich structure. On the noisy synthetic task, random forest achieved the highest AUROC, while XGBoost and the MLP remained well above the two linear statistical models. Under explicit imbalance, XGBoost was strongest by AUROC, whereas relying on plain accuracy would have obscured important minority-class behavior. These results support the view that model-family performance is conditional on data geometry, sample size, noise, and class prevalence.

Discussion

Principal findings

This benchmark produces a nuanced answer to the question posed in the title. Traditional tree-based machine learning was strongest overall, but the evidence does not support a simplistic claim that XGBoost significantly beats every alternative. The omnibus Friedman test established that ranks differed across algorithms, yet multiplicity-corrected pairwise inference identified only the XGBoost-versus-MLP contrast as statistically significant at the .05 familywise level. Random forest, logistic regression, and LDA each won at least one task, demonstrating meaningful heterogeneity in the best inductive bias.

The results align with the earlier large-scale tabular literature. Grinsztajn et al.

(2022) found a persistent advantage for trees on many medium-sized tabular problems, and McElfresh et al. (2023) showed that algorithm family interacts strongly with dataset properties. The present results reproduce that qualitative pattern in a smaller but tightly controlled suite: trees excel on tasks with nonlinear structure and heterogeneous predictors, whereas linear statistical models remain formidable when the signal is sufficiently smooth, low-dimensional, or sample-limited.

At the same time, the findings should not be used as evidence that contemporary deep tabular learning is inferior in general. The tested neural model was a conventional task-trained MLP. Since 2024, pre-tuned MLPs, retrieval-augmented architectures, and especially tabular foundation models have changed the frontier. TabArena reports deep-learning parity under larger budgets and ensembling, and TabPFN-3 reports major gains from synthetic pretraining and in-context prediction (Erickson et al., 2025; Grinsztajn et al., 2026). The appropriate interpretation is narrower: increasing representational depth alone did not guarantee superior tabular prediction under the controlled task-trained regime used here.

Why statistical models remained competitive

The competitiveness of logistic regression and LDA reflects the bias-variance trade-off. When the decision boundary is close to linear after scaling/encoding, flexible models spend capacity estimating interactions that may not generalize. Regularization further reduces variance. The Breast Cancer and Haberman results show why a statistical baseline should be retained even in modern benchmarking: an apparently 'simple' learner can be close to Bayes-relevant structure, and complexity cannot compensate for weak signal or small effective sample size. Moreover, statistical models are easier to audit, calibrate, and communicate in many regulated settings.

Why tree ensembles performed strongly

Random forest and XGBoost directly partition feature space and can capture thresholds and interactions without explicit feature engineering. This is particularly advantageous for mixed-scale tabular variables and irregular response surfaces. XGBoost's sequential error correction and regularization produced the best average rank, while random forest was strongest on Ionosphere and the noisy stress test. Their difference was not statistically significant after correction, so the choice between them should incorporate runtime, tuning tolerance, probability quality, and stability in addition to mean AUROC.

Interpreting the deep MLP result

The MLP performed extremely well on Banknote and Synthetic-Easy and substantially outperformed the linear models on Synthetic-Noisy and Synthetic-Imbalanced. Its weakness was inconsistency on some small real datasets, particularly Haberman and Mode Choice. Neural networks must estimate many coupled parameters, and their optimization depends on internal validation, initialization, and regularization. These costs are most visible when sample size is small relative to function complexity. The result therefore illustrates conditional sample efficiency, not a fundamental inability of neural networks to model tables.

Always include a regularized statistical baseline. If a flexible model cannot materially improve repeated out-of-sample performance, the simpler model may offer better interpretability and governance.

Use boosted and bagged trees as default high-strength baselines for heterogeneous tabular data, especially when threshold effects and interactions are plausible.

Evaluate deep tabular models when sample size, compute budget, or transfer/pretraining resources justify them; distinguish ordinary task-trained MLPs from foundation-model approaches.

Report discrimination and probability-quality metrics together. AUROC alone cannot reveal whether predicted probabilities are well calibrated or whether threshold behavior is acceptable under imbalance.

Treat datasets as the unit of statistical inference in multi-dataset benchmarks and adjust pairwise tests for multiplicity. Cross-validation folds should not be mistaken for independent experimental units.

Prevent group and preprocessing leakage. A sophisticated learner evaluated with leaked splits can appear stronger than a genuinely better model evaluated correctly.

Limitations

The benchmark has five important limitations. First, it includes 10 tasks, enough to support blocked nonparametric testing but far fewer than modern benchmark suites containing dozens or hundreds of datasets. The wide Nemenyi critical difference reflects this limitation. Second, only one neural architecture was evaluated; specialized models such as TabNet, FT-Transformer, TabR, RealMLP, TabICL, and TabPFN may produce different rankings. Third, hyperparameters were intentionally fixed rather than nested-tuned. This strengthens fairness under a common budget but does not estimate the best attainable performance of each family. Fourth, the tasks are binary classifications; conclusions should not be extrapolated directly to regression, survival prediction, count data, multilabel problems, or time-dependent forecasting. Fifth, wall-clock timing is hardware- and software-dependent and should be interpreted as a relative implementation result.

Additional limitations concern the data themselves. Several UCI datasets are classic and comparatively small, and the Fair and Mode Choice datasets are historical. They are useful for controlled algorithm comparison but are not proxies for contemporary deployment distributions. Some real tasks also contain social or sensitive variables; this paper evaluates predictive methods, not the ethical suitability of deploying models trained on those attributes. Future work should use larger contemporary datasets, prespecified nested tuning budgets, external validation, calibration curves, subgroup robustness, and modern foundation-model baselines.

Conclusion

Across ten heterogeneous tabular binary-classification tasks, model family mattered, but the magnitude and direction of the advantage depended strongly on the dataset. XGBoost achieved the best average AUROC rank and was significantly better than the conventional deep MLP after Holm correction. Random forest was a close, statistically indistinguishable competitor, while logistic regression and shrinkage LDA remained highly competitive on several small or approximately linear tasks. The

key methodological conclusion is therefore stronger than a winner-takes-all ranking: reliable model selection requires leakage-controlled cross-validation, multiple performance dimensions, dataset-level paired inference, and multiplicity correction. For practitioners, strong statistical and tree-based baselines should precede claims that a more complex deep learner adds value. For researchers, the rise of tabular foundation models makes this disciplined baseline logic more—not less—important, because advances in pretraining, ensembling, and compute budget can otherwise be confounded with algorithmic superiority.

References

- Arik, S. O., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1–30.
- Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., & Hutter, F. (2025). TabArena: A living benchmark for machine learning on tabular data. *arXiv:2506.16791*.
- Fair, R. C. (1978). A theory of extramarital affairs. *Journal of Political Economy*, 86(1), 45–61.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.
- Gijsbers, P., Bueno, M. L., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., & Vanschoren, J. (2024). AMLB: An AutoML benchmark. *Journal of Machine Learning Research*, 25, 1–65.
- Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932–18943.
- Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., & Babenko, A. (2024). TabR: Unlocking the power of retrieval-augmented tabular deep learning. *International Conference on Learning Representations*.
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still

- outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520.
- Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., et al. (2026). TabPFN-3: Technical report. arXiv:2605.13986.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Holzmüller, D., Grinsztajn, L., & Steinwart, I. (2024). Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. *Advances in Neural Information Processing Systems*, 37.
- Liu, S.-Y., Cai, H.-R., Zhou, Q.-L., Yin, H.-H., Zhou, T., Jiang, J.-P., & Ye, H.-J. (2025). Talent: A tabular analytics and learning toolbox. *Journal of Machine Learning Research*, 26(226), 1–16.
- Lohweg, V. (2013). Banknote Authentication [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C55P57>
- McElfresh, D., Khandagale, S., Valverde, J., Prasad, V., Ramakrishnan, G., Goldblum, M., & White, C. (2023). When do neural nets outperform boosted trees on tabular data? *Advances in Neural Information Processing Systems*, 36.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Quinlan, R. (1987). Statlog (Australian Credit Approval) [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C59012>
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96.
- Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies*, 14(2), 1518-1527.
- Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90. <https://doi.org/10.1016/j.inffus.2021.11.011>
- Sigillito, V., Wing, S., Hutton, L., & Baker, K. (1989). Ionosphere [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5W01B>
- UCI Machine Learning Repository. (1999). Haberman's Survival [Dataset]. University of California, Irvine.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.
- Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5DW2B>