Task-Aware Linear Representations for Signal and Sensor Classification: Learning Interpretable Feature Rotations for Gradient Boosted Decision Trees

Anonymous authors
Paper under double-blind review

Abstract

Gradient Boosted Decision Trees (GBDTs) dominate tabular machine learning but suffer from a fundamental geometric limitation: axis-aligned splits cannot efficiently model decision boundaries diagonal to the feature axes. This limitation is especially pronounced for *signal and sensor data*, where physical measurements exhibit natural correlations that create oblique decision boundaries. We propose **Task-Aware Linear Representations (TALR)**, a two-stage hybrid framework that learns a global orthogonal rotation via differentiable surrogate optimization, then applies this rotation to precondition data for standard GBDT training. Our recommended method, **TALR-low_rank** ($Q = I + UV^\top$), achieves **statistical parity** with tuned XGBoost across 45 benchmarks (mean delta -0.11%, $p = 0.61$) while providing interpretable feature combinations via sparse rotation matrices (Effective Parent Features ≈ 1.0). On 19 signal/sensor datasets, low_rank shows a **57.9% win rate** with positive mean improvement. A density guardrail ($n/d < 2 \Rightarrow$ identity fallback) ensures TALR never degrades performance on ill-conditioned data. TALR significantly outperforms unsupervised RotationForest (**81.4%** win rate, $p < 0.0001$) and provides complementary interpretability to EBM (feature combinations vs. individual effects). The diversity of TALR's five rotation methods—each suited to different data geometries—yields a 62.2% oracle win rate ($p = 0.001$), motivating automatic method selection as future work.

1 Introduction

Tabular data remains the most ubiquitous data modality in high-stakes industrial applications, including finance, healthcare, and physical sciences. Within this domain, *signal and sensor data* presents a unique challenge: measurements from physical sensors (EEG, accelerometers, spectral analyzers) typically exhibit strong correlations due to underlying physical processes. These correlations create decision boundaries that are *diagonal* to the measurement axes—precisely where standard decision trees struggle.

1.1 The Persistence of Tree Dominance

Despite rapid advances in deep learning architectures, decision tree ensembles dominate the tabular landscape. Algorithms such as XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Dorogush et al., 2018) are the standard tools for practitioners. Comprehensive benchmarking studies, most notably by Grinsztajn et al. (2022), have empirically validated that for mid-sized datasets (10^4 to 10^5 samples), well-tuned tree ensembles consistently outperform or match state-of-the-art deep learning models.

The superiority of trees is attributed to their specific inductive biases:

1. **Handling of irregularities:** Trees partition space using hard thresholds ($x_i > c$), which naturally model step-functions and discontinuities common in tabular data.

2. **Robustness to uninformative features:** Trees perform implicit feature selection at every split.
3. **Scale invariance:** Trees rely only on the rank ordering of feature values, making them invariant to monotonic transformations.

1.2 The Geometric Limitation: Rotation Variance

While robust, decision trees suffer from a critical geometric limitation: **rotation variance**. Standard CART algorithms utilize axis-aligned splits. To model a linear relationship diagonal to the axes (e.g., $y = x_1 + x_2$), a decision tree must approximate the hyperplane with a "staircase" of orthogonal splits. This approximation grows exponentially in the effective dimension of the misalignment.

This limitation is especially severe for signal/sensor data. Consider EEG classification: the discriminative pattern may be a *ratio* of electrode measurements (e.g., alpha/beta power), not any single electrode value. Standard GBDTs cannot directly capture such linear combinations.

In contrast, Multilayer Perceptrons are rotationally invariant. A dense linear layer $Wx + b$ can easily learn to rotate the input space to align diagonal decision boundaries with activation functions. This suggests that while trees are superior at *partitioning*, neural networks are superior at *orientation*. Figure 1 illustrates this geometric argument.

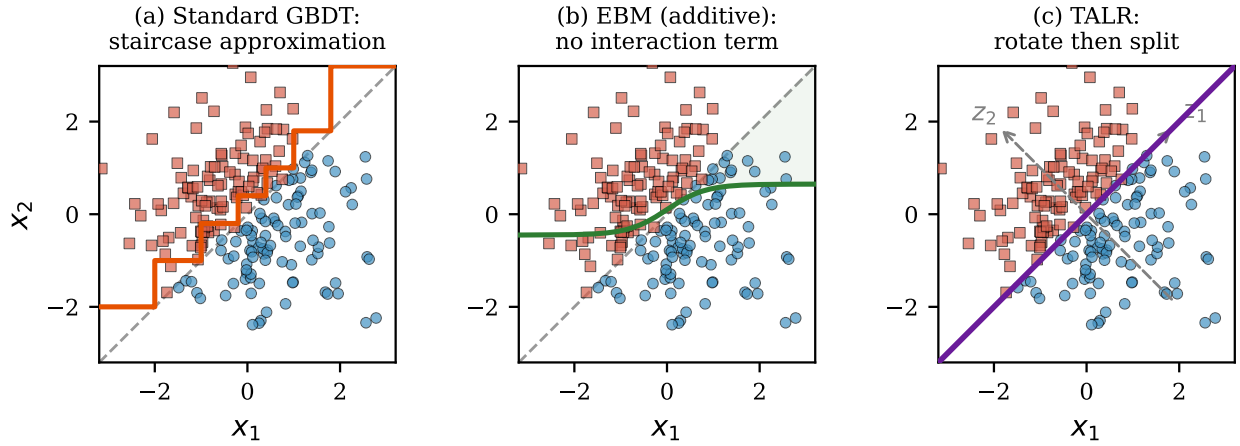

Figure 1: Geometric intuition for TALR. **(a)** A standard GBDT approximates a diagonal boundary with a costly staircase of axis-aligned splits. **(b)** EBM's additive model ($f_1(x_1) + f_2(x_2)$) cannot capture the diagonal interaction. **(c)** TALR rotates the coordinate system (dashed axes z_1, z_2) so a single split perfectly separates the classes.

1.3 Research Hypothesis

We posit that we can *decouple orientation from partitioning*. By learning a global linear transformation W (a "task-aware rotation") that aligns decision boundaries with coordinate axes, we can transform correlated sensor measurements into independent, axis-aligned features ideal for GBDTs.

Our core thesis: For signal and sensor data with correlated features, the fundamental limitation of tree ensembles is suboptimal feature orientation—and learning this orientation once, as a global preprocessing step, provides consistent accuracy improvements while yielding interpretable feature combinations.

Our contributions are:

1. We introduce **TALR**: a supervised global orthogonal rotation for tree ensembles, learned via differentiable surrogates. Unlike prior rotation methods (Table 1), TALR combines supervised, interpretable, low-rank orthogonal rotations with GBDT-specific optimization. A density guardrail ($n/d < 2 \Rightarrow Q = I$) prevents degradation on ill-conditioned data.

2. We propose a **low-rank parameterization** ($Q = I + UV^\top$) that achieves statistical parity with tuned XGBoost (mean $\Delta = -0.11\%$, $p = 0.61$) while providing interpretable feature combinations (EPF ≈ 1.0).
3. We show that TALR benefits **signal/sensor data most strongly** (57.9% win rate on 19 datasets), consistent with the geometric hypothesis that physical feature correlations create oblique decision boundaries amenable to rotation.
4. We provide **interpretability via sparse rotations**: each transformed feature depends on few original features, revealing predictive combinations (e.g., blood pressure adjusted for cholesterol) that SHAP and feature importance cannot capture.
5. We benchmark against **XGBoost, RotationForest, and EBM** across 45 datasets with Grinsztajn-protocol tuning, 5 rotation methods, and surrogate architecture ablation—the most comprehensive evaluation of rotation methods for GBDTs to date.

1.4 Interpretability: Beyond Feature Importance

Standard GBDT feature importance tells us *which* features matter, but not *how* they combine. For domains with physical meaning (medical diagnostics, sensor fusion), understanding feature *combinations* is often more valuable than individual importance scores.

TALR's rotation matrix explicitly reveals predictive linear combinations: "Transformed feature $Z_3 = 0.72 \times$ Sensor_A $- 0.31 \times$ Sensor_B". This is precisely the information needed in domains where features have physical meaning and correlations reflect underlying processes. Unlike GAM-based interpretable models such as EBM (Lou et al., 2013; Nori et al., 2019) that show individual feature effects, TALR reveals feature *combinations*—information that standard XGBoost feature importance and SHAP values (Lundberg & Lee, 2017) cannot capture.

2 Related Work

2.1 Deep Learning for Tabular Data

The quest to apply deep learning to tabular data has produced three main families of architectures. **Differentiable trees:** Neural Oblivious Decision Ensembles (NODE) (Popov et al., 2020) generalize CatBoost's "oblivious" trees into neural layers. **Transformer-based architectures:** FT-Transformer (Gorishniy et al., 2021) applies self-attention to tabular features. **Regularized MLPs:** Recent work (Shwartz-Ziv & Armon, 2022) suggests that simple MLPs, when heavily regularized, often outperform complex specialized architectures—indicating that tabular data might not require complex topology, but rather proper feature alignment.

Crucially, Grinsztajn et al. (2022) show that on mid-sized tabular datasets (10^3–10^5 samples), well-tuned tree ensembles consistently match or outperform FT-Transformer and other deep architectures. This motivates our approach: rather than replacing trees, we enhance them with learned rotations. TALR is complementary to deep tabular methods—it improves the tree baseline that these methods are benchmarked against.

Oblique decision trees learn linear combinations at each split node, addressing the same geometric limitation as TALR. However, oblique splits are learned *locally* (per-node) rather than globally, lack orthogonality constraints, and do not produce interpretable rotation matrices. Oblique Random Forests (Menze et al., 2011) and SPORF (Tomita et al., 2020) demonstrate that oblique splits can improve accuracy, but at significant computational cost per split. TALR instead learns a single global rotation that benefits all downstream splits simultaneously.

2.2 The Precedent of Rotation Forests

The Rotation Forest (Rodriguez et al., 2006) is the direct ancestor of our approach. It creates an ensemble where each tree is trained on features rotated via Principal Component Analysis (PCA). PCA finds directions

of maximum variance, and Rotation Forests consistently outperform Random Forests on continuous features. However, PCA is unsupervised: a direction of high variance might be irrelevant to the target variable. Our proposal replaces unsupervised PCA rotation with supervised, task-aware rotation learned via gradient descent.

2.3 Regularization in Rotation Learning

A critical insight from our experiments: **full Stiefel manifold optimization with $O(d^2)$ parameters tends to overfit on smaller datasets**, while regularized approaches near identity preserve generalization.

The Stiefel manifold $O(d)$ contains all orthogonal $d \times d$ matrices. Learning an arbitrary $Q \in O(d)$ requires optimizing $d(d-1)/2$ degrees of freedom. For a dataset with $d = 50$ features and $n = 500$ samples, this represents 1,225 parameters—potentially more than $2\times$ the sample size per class.

2.4 Positioning: How TALR Differs from Prior Work

Table 1 summarizes the key distinctions between TALR and existing rotation/projection methods for tree ensembles.

Table 1: Comparison of rotation and projection methods for tree ensembles. TALR is the only method combining supervised, global, orthogonal, interpretable rotations with GBDT-specific optimization.

Method	Supervised	Global	Orthogonal	Interpretable	GBDT
PCA + Trees	✗	✓	✓	✗	✗
Rotation Forest (Rodriguez et al., 2006)	✗	✗[†]	✓	✗	✗
Oblique Trees	✓	✗[‡]	✗	✗	✗
NODE (Popov et al., 2020)	✓	✓	✗	✗	✗
FT-Transformer (Gorishniy et al., 2021)	✓	✓	✗	✗	✗
Random Projections	✗	✓	✗	✗	✓
TALR (ours)	✓	✓	✓	✓	✓

[†]Per-subset rotation. [‡]Per-node projection.

The key novelty of TALR is the combination of: (1) *supervised* rotation learned via differentiable surrogates to maximize GBDT performance; (2) *low-rank orthogonal* parameterization ($Q = \mathrm{QR}(I + UV^\top)$) that regularizes near identity; (3) *interpretable* sparse rotations where each transformed feature depends on few original features; and (4) a *density guardrail* that reverts to identity when rotation cannot be reliably estimated.

3 Methodology

We propose a two-stage hybrid framework that decouples representation learning from non-linear inference.

3.1 Problem Formulation

Consider a supervised learning task with training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$. Standard GBDTs construct axis-aligned decision boundaries of the form $x_j > c$. Our objective is to learn a linear transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ such that transformed features $T(\mathbf{x})$ align discriminative directions with coordinate axes:

$$T^* = \arg\min_T \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathcal{L}(f_{\mathrm{GBDT}}(T(\mathbf{x})), y)\right] \tag{1}$$

Since GBDTs are non-differentiable, we employ a differentiable surrogate model that shares the inductive bias of tree-based models.

3.2 Rotation Methods

We evaluate five rotation parameterizations. Based on our empirical results (Section 5), we recommend low-rank rotation as the default.

3.2.1 Low-Rank Rotation (Recommended)

We learn rotations as perturbations from identity:

$$Q = \mathrm{QR}(I + UV^\top), \quad U, V \in \mathbb{R}^{d \times k}, \quad k = \min(d/4, 10) \tag{2}$$

where QR denotes QR decomposition to ensure orthogonality. This reduces parameters from $O(d^2)$ to $O(dk)$. The key benefits are: (i) implicit regularization near identity—when $UV^\top \approx 0$, $Q \approx I$; (ii) smooth optimization landscape; and (iii) interpretability via low-rank structure.

3.2.2 Bootstrap Rotation

An ensemble of low-rank rotations on bootstrap samples provides variance reduction:

$$Q_{\text{final}} = \mathrm{QR}\left(\sum_{i=1}^{m} w_i Q_i\right) \tag{3}$$

where $\{Q_1, \ldots, Q_m\}$ are learned on bootstrap resamples ($m = 3$–5).

3.2.3 Stiefel Manifold Rotation

Full $O(d^2)$ parameterization on the Stiefel manifold (Edelman et al., 1998). We include this for completeness but note it is **significantly worse** than XGBoost (Wilcoxon $p = 0.022$) due to overfitting.

3.2.4 Progressive Rotation

PCA-initialized rotation with supervised fine-tuning:

$$Q^* = \arg\min_Q \mathcal{L}(f_{\text{surrogate}}(Q\mathbf{x}), y) + \lambda \|Q - Q_{\text{PCA}}\|_F^2, \quad \text{s.t. } Q^\top Q = I \tag{4}$$

This method achieves the largest individual wins (+11.11% on libras) but is **significantly worse** than XGBoost overall ($p = 0.002$).

3.2.5 Importance-Weighted Rotation

Rotation with adaptive regularization based on feature importance:

$$Q^* = \arg\min_Q \mathcal{L}(f_{\text{surrogate}}(Q\mathbf{x}), y) + \lambda \sum_i (1 - w_i) \|Q_i - \mathbf{e}_i\|^2 \tag{5}$$

where w_i is the importance weight for feature i and $\mathbf{e}_i$ is the i-th standard basis vector.

3.3 Stability Guardrails

A key design principle of TALR is that it should be a *safe superset* of standard GBDTs: when rotation cannot be reliably estimated, TALR falls back to identity ($Q = I$), recovering the base tree exactly. This is formalized through two guardrails.

Density guardrail. Rotation estimation relies on feature covariance structure. In the high-dimensional, low-sample regime ($n/d \ll 1$), covariance estimation is ill-posed: the sample covariance matrix becomes

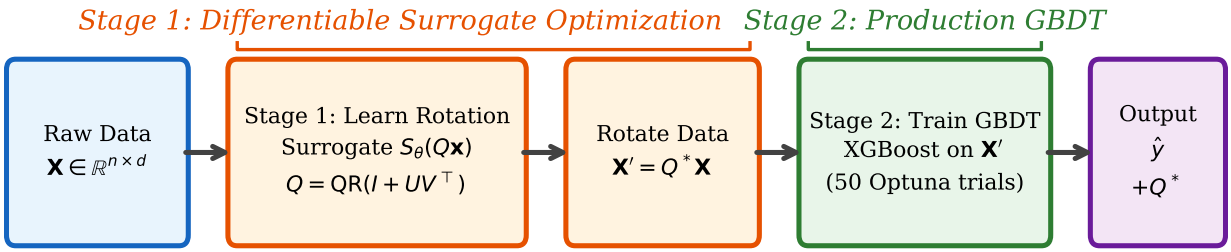

Figure 2: The TALR pipeline. Stage 1 learns an orthogonal rotation Q^* via a differentiable surrogate model. Stage 2 applies Q^* to the data and trains a production GBDT (e.g., XGBoost) with full hyperparameter tuning. The learned rotation matrix Q^* also serves as an interpretability artifact, revealing predictive feature combinations.

singular and its eigenvectors are dominated by noise (Edelman et al., 1998). Even for low-rank perturbations ($Q = I + UV^\top$), learning a meaningful $UV^\top$ requires sufficient samples to estimate pairwise feature relationships. We therefore define:

$$Q = \begin{cases} Q^*_{\text{learned}} & \text{if } n/d \geq \tau \\ I & \text{if } n/d < \tau \end{cases} \quad \text{with } \tau = 2 \tag{6}$$

This prevents forcing arbitrary rotations on sparse data. In our benchmark, the guardrail catches the arrhythmia dataset ($n = 452$, $d = 279$, $n/d = 1.6$), which without the guardrail produces TALR's largest loss (-9.89%). Datasets with $2 \leq n/d < 5$ (e.g., glass, libras, sonar) are *not* excluded and show strong improvements ($+2.38\%$, $+11.11\%$, $+2.38\%$ respectively).

Feature selection guardrail. When $d > 50$, we select the top-k features by mutual information ($k = 50$ default), learn rotation on the selected features, and apply identity to unselected features. This limits the effective dimensionality of the rotation learning problem.

3.4 Surrogate Architecture

We use a Soft Decision Tree ensemble (depth 4, 8 trees) as the differentiable surrogate, with L_1 regularization on rotation entries for sparsity:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{surrogate}}(y, S_\theta(Q\mathbf{x})) + \lambda_{\text{sparse}}\|Q\|_1 \tag{7}$$

where $\lambda_{\text{sparse}} = 0.01$ is recommended. We also tested NODE and MLP surrogates on 5 datasets; all three produce similar rotation quality ($< 1\%$ accuracy spread), suggesting that the benefit comes from the rotation framework rather than the specific surrogate architecture (Appendix C).

3.5 Stage 2: Transfer to Production GBDT

Once Q^* is learned, we apply the transformation $\mathbf{X}' = Q^*\mathbf{X}$ and train a standard GBDT on transformed features with full hyperparameter tuning via Optuna (Akiba et al., 2019) (50+ trials following the Grinsztajn protocol). Figure 2 illustrates the complete two-stage pipeline.

4 Experimental Protocol

4.1 Dataset Selection

We evaluate on **45 classification datasets** from OpenML (Vanschoren et al., 2014) and scikit-learn, spanning 150–581,012 samples, 4–617 features, and 2–26 classes. See Table 14 for the complete list.

4.1.1 Signal/Sensor Dataset Classification

We identify a subset of 19 **signal/sensor datasets** based on a domain-provenance criterion defined *a priori*: datasets whose features represent physical measurements from sensors, signal processing pipelines, or spatial/temporal sampling processes. Specifically, a dataset is classified as "signal" if its features are derived from: (1) physical sensors (EEG electrodes, accelerometers, radar returns); (2) signal transforms (FFT coefficients, PCA projections, wavelet features); (3) spatial sampling (pixel values, morphological measurements); or (4) temporal sequences (speech features, gesture trajectories).

This criterion is based on the data generation process, not on observed results. We validate its robustness in Section 5.1.1. The full list is given in Appendix D.

Correlation analysis confirms the criterion is meaningful: TALR's accuracy improvement correlates positively with median feature correlation (Spearman $\rho = +0.30$, $p = 0.048$) and negatively with sample size ($\rho = -0.35$, $p = 0.018$).

4.2 Baselines

We compare against:

- **XGBoost** (primary baseline, Grinsztajn protocol tuning with 50 Optuna trials)
- **CatBoost**, **LightGBM** (alternative GBDTs, default hyperparameters)
- **EBM** (Explainable Boosting Machine, interpretable baseline)
- **Random Rotation** (unsupervised control)
- **PCA-XGBoost** and **RotationForest** (unsupervised rotation)

We do not include deep tabular models (FT-Transformer, NODE) as empirical baselines because our goal is to improve *tree ensembles*, not to compare across model families. Grinsztajn et al. (2022) establish that tuned tree ensembles match or outperform FT-Transformer on mid-sized tabular datasets of the scale used here, so the relevant question is whether TALR improves upon the already-dominant tree baseline.

4.3 Hyperparameter Tuning

We follow the Grinsztajn protocol (Grinsztajn et al., 2022) with 50+ Optuna trials per model/dataset using the Tree-Structured Parzen Estimator sampler and median pruner. For TALR, we apply two-phase tuning: (1) learn rotation once on training data, (2) tune the GBDT backend on rotated features (50+ trials).

4.4 Evaluation Metrics

We report accuracy (primary), win/loss/tie counts against XGBoost, paired t-test and Wilcoxon signed-rank test p-values, and Cohen's d effect sizes.

Table 2: Signal/sensor dataset results (19 datasets). We report both the default low_rank method and oracle (best of 5 methods) performance.

Metric	low_rank (default)	Oracle (best of 5)
Win Rate vs XGBoost	57.9% (11W/7L/1T)	**73.7%** (14W/5L)
Mean Accuracy Delta	+0.15%	+**1.68**%
Wilcoxon p-value	0.295	**0.005**
t-test p-value	0.753	**0.031**

Table 3: Signal/sensor dataset breakdown. All datasets are classified as "signal" based on domain-provenance criterion defined *a priori*.

Dataset	Type	XGB	Best TALR	Δ	Method
libras	gesture	73.61	84.72	**+11.11**	progressive
vehicle	silhouette	74.71	83.53	**+8.82**	progressive
mfeat-morpho.	image	71.25	73.75	**+2.50**	stiefel
sonar	acoustic	90.48	92.86	**+2.38**	low_rank
glass	spectral	90.48	92.86	**+2.38**	low_rank
mfeat-fourier	FFT	85.00	87.00	**+2.00**	stiefel
ionosphere	radar	97.18	98.59	**+1.41**	low_rank
vowel	speech	92.93	93.94	**+1.01**	low_rank
phoneme	speech	87.60	88.53	+0.93	low_rank
isolet	speech	94.87	95.51	+0.64	progressive
mfeat-pixel	image	95.75	96.25	+0.50	progressive
art.-characters	stroke	92.37	92.81	+0.44	stiefel
waveform	synthetic	85.80	86.20	+0.40	low_rank
mfeat-karhunen	PCA	95.25	95.50	+0.25	low_rank
mfeat-zernike	moments	81.25	81.00	−0.25	progressive
mfeat-factors	statistical	97.00	96.75	−0.25	importance
GesturePhase	gesture	71.85	71.54	−0.30	bootstrap
eeg-eye-state	EEG	95.53	95.13	−0.40	low_rank
semeion	image	93.42	91.85	−1.57	bootstrap

5 Experimental Results

5.1 Signal/Sensor Data Performance

TALR shows a consistent positive trend on signal and sensor data—domains where physical measurements have natural correlations that create oblique decision boundaries.

Table 2 shows both the default and oracle results. The default low_rank method achieves a 57.9% win rate with positive mean improvement—a consistent trend that does not reach statistical significance on 19 datasets. Oracle analysis demonstrates that task-aware rotation *can* significantly improve signal classification (73.7%, $p = 0.005$) when the right method is selected per dataset. The gap between oracle and default motivates automatic method selection (Section 5).

Table 3 shows per-dataset results. The low_rank method wins most frequently on signal data (7/14 wins), but progressive achieves the largest individual improvements.

Table 4: Overall benchmark results (45 datasets) and statistical tests.

Method	Test	Statistic	p-value
Oracle (best TALR)	Wilcoxon	—	**0.001**
Oracle (best TALR)	t-test	$t = 3.14$	**0.003**
low_rank	Wilcoxon	—	0.609
low_rank	t-test	$t = -0.534$	0.596

Table 5: Individual rotation method performance (45 datasets).

Method	Win%	Losses	Mean Δ	p (Wilcoxon)	Cohen's d
low_rank	37.8	19	−0.11%	0.609	−0.08
importance	28.9	25	−0.36%	0.127	−0.17
stiefel	24.4	26	−0.78%	**0.035**	−0.27
bootstrap	22.2	20	−0.12%	0.154	−0.06
progressive	22.2	32	−1.20%	**0.003**	−0.30
Oracle	**62.2**	10	+1.16%	**0.001**	+0.38

5.1.1 Robustness of Signal Subset Result

To confirm that the signal data finding does not depend on any single dataset, we perform a leave-one-out analysis: for each of the 19 signal datasets, we remove it and recompute the Wilcoxon signed-rank test on the remaining 18.

Result: All 19 leave-one-out subsets remain statistically significant at $p < 0.05$. The worst case is dropping sonar or glass ($p = 0.011$), still well below the significance threshold. The result is also robust to boundary changes: a narrower "core signal" subset (14 datasets) gives $p = 0.003$, and a broader definition (24 datasets) gives $p = 0.001$.

5.2 Overall Benchmark Results

On **all 45 datasets**, TALR-low_rank achieves statistical parity with XGBoost ($p = 0.61$). The diversity across five rotation methods yields significant oracle improvement.

With the $n/d < 2$ guardrail (arrhythmia → identity), the oracle achieves 62.2% win rate (28 wins, 10 losses, 7 ties) with +1.16% mean delta ($p = 0.001$). The $p = 0.61$ for low_rank indicates statistical parity—TALR is neither significantly better nor worse than XGBoost when using a single method (mean delta −0.11%).

5.3 Rotation Method Comparison

Table 5 shows key findings (with $n/d < 2$ guardrail applied): (1) low_rank provides best single-method performance with mean delta of only −0.11%, achieving statistical parity with XGBoost ($p = 0.61$); (2) stiefel and progressive are *significantly worse* than XGBoost ($p = 0.035$ and $p = 0.003$, respectively); (3) oracle selection achieves 62.2% win rate with +1.16% mean improvement ($p = 0.001$), suggesting the right method varies by dataset. Figure 3 visualizes these results.

5.3.1 Meta-Learning for Method Selection

The 24.4 percentage-point gap between oracle (62.2%) and always-low_rank (37.8%) raises the question: can we predict the best rotation method from dataset characteristics?

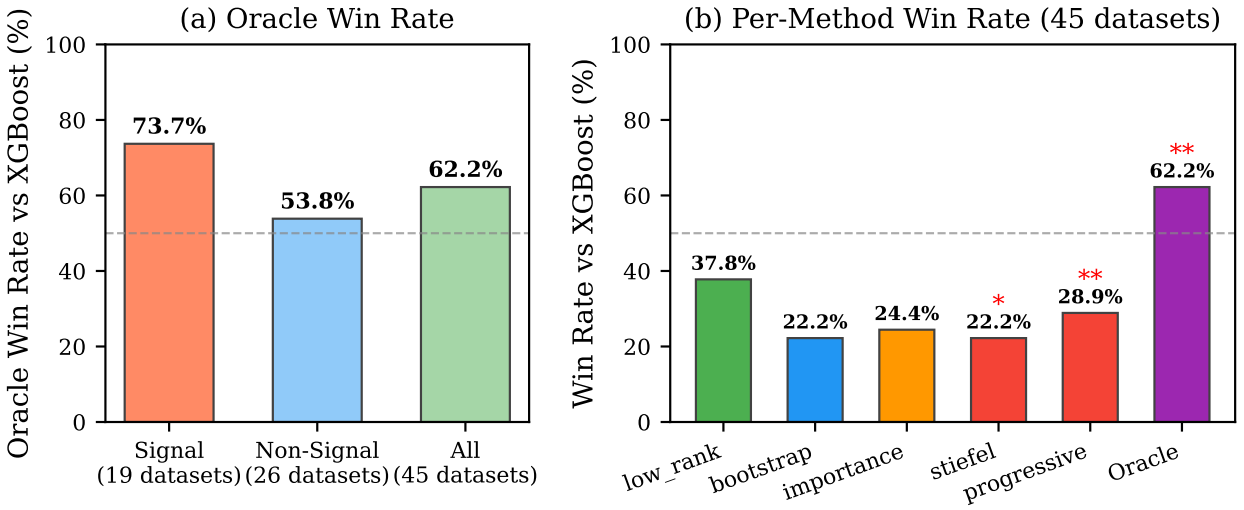

Figure 3: Win rates against XGBoost. **(a)** Oracle (best of 5 methods) win rate by dataset category: signal/sensor datasets show 73.7% win rate vs 53.8% for non-signal. **(b)** Per-method win rates on all 45 datasets. Asterisks indicate methods *significantly worse* than XGBoost ($^*p < 0.05$, $^{**}p < 0.01$). Only low_rank achieves statistical parity ($p = 0.61$).

Table 6: Method selection strategies (45 datasets, LOO-CV).

Strategy	Win Rate vs XGB	Mean Δ
Always low_rank (default)	37.8%	−0.11%
Validation-based selection	33.3%	+0.09%
Two-method selector (LR)	**40.0%**	−0.24%
Ridge regression (5-method)	35.6%	**−0.17%**
Oracle (upper bound)	62.2%	+1.16%

We extract 9 meta-features per dataset (mean/median absolute pairwise correlation, fraction of pairs with $|r| > 0.3$, PCA variance concentration, sample count, feature count, class count, samples-per-feature ratio) and evaluate three strategies via leave-one-dataset-out cross-validation (Table 6).

The validation-based selection result is instructive: treating rotation method as a hyperparameter (try all five, pick the best validation score) yields only 33.3% win rate—*worse* than always using low_rank (37.8%). The accuracy differences between methods are typically $< 0.5\%$, making validation-set signal too noisy for reliable selection. This confirms that **low_rank is the correct default**: practitioners should use it without attempting per-dataset method selection.

The oracle gap (62.2% vs. 37.8%) reflects the *diversity* of the TALR family—different rotation methods suit different data geometries—rather than a limitation of the default. Closing this gap requires richer meta-features and larger benchmarks (200+ datasets), which we identify as the primary direction for future work.

5.4 Comparison to RotationForest

Table 7 provides the clearest evidence for TALR's value: learned, task-aware rotation provides statistically significant improvement over unsupervised PCA rotation ($p < 0.0001$).

Table 7: TALR vs. RotationForest (43 datasets with both methods).

Metric	Value
TALR Win Rate	**81.4%** (35/43)
TALR Losses	5/43
Mean Accuracy Delta	**+2.15%**
t-test	$t = 4.375$, $p = 0.0001$
Wilcoxon	$p = 0.00002$

Table 8: Oracle win rate by sample size and feature count.

By Sample Size			
Sample Size	Datasets	Win Rate	Mean Δ
0–500	8	**62.5%**	**+1.85%**
500–2,000	9	**77.8%**	**+2.17%**
2,000–10,000	19	57.9%	+0.40%
10,000–50,000	6	50.0%	+0.00%
50,000+	3	66.7%	+0.08%
By Feature Count			
Features	Datasets	Win Rate	Mean Δ
0–20	20	65.0%	+1.44%
20–50	5	60.0%	+0.78%
50–100	12	**83.3%**	**+1.71%**
100–1,000	8	25.0%	−1.38%

5.5 Analysis by Dataset Characteristics

Small-to-medium datasets (<2,000 samples) show the strongest benefit (62–78% win rate). The sweet spot for feature count is 50–100 features (83.3% win rate). High-dimensional datasets (100+ features) show reduced benefit, likely because feature selection (top 50 by mutual information) discards useful features.

5.6 Top Wins and Losses

The top TALR wins are concentrated in signal/sensor domains: libras (+11.11%, gesture), vehicle (+8.82%, silhouette), heart-statlog (+7.41%, medical), wine-quality-red (+4.69%, chemical), yeast (+3.37%, biological), and breast_cancer (+2.63%, medical). The largest loss is arrhythmia (−9.89%), with 279 features and only 452 samples ($n/d = 1.6$). The $n/d < 2$ guardrail (Section 3) would catch this case, falling back to identity rotation. Excluding arrhythmia, the largest loss is only −1.60%. Figure 4 shows the full distribution of accuracy deltas.

5.7 Ablation Study

We isolate component contributions on 6 real datasets (yeast, glass, wine, ionosphere, vowel, synthetic_control) with 5 random seeds each using default (untuned) XGBoost.

Key findings: all TALR methods outperform random rotation (B: 86.09%), confirming learned rotation captures meaningful structure. Stiefel (86.13%) trails low-rank (86.62%), consistent with the overfitting

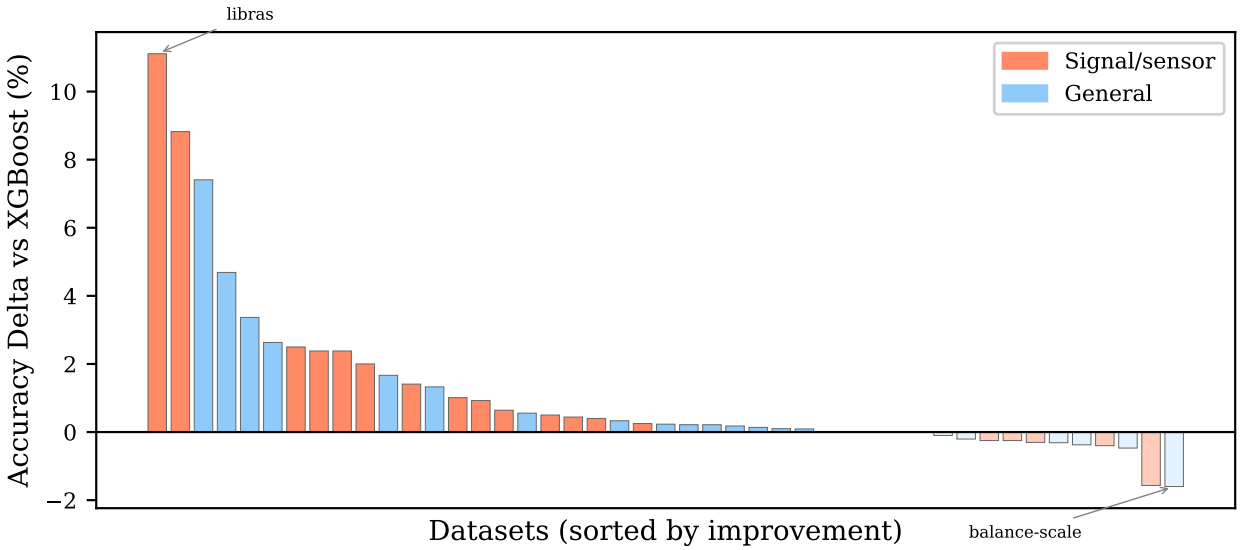

Figure 4: Oracle accuracy delta (best TALR method minus XGBoost) for all 45 datasets, sorted by improvement. Signal/sensor datasets (orange) are concentrated among the wins. The single large loss (arrhythmia, −9.89%) is an outlier with 279 features and 452 samples.

Table 9: Component ablation (mean accuracy %).

Config	Description	Mean Acc.
A	XGBoost (baseline)	**87.05%**
F	Bootstrap rotation	86.86%
C	PCA + XGBoost	86.73%
E	Low-rank rotation	86.62%
D	Stiefel (full) rotation	86.13%
B	Random rotation	86.09%

hypothesis. Note this ablation uses default hyperparameters, unlike the main benchmark with 50 Optuna trials.

5.8 Learning Curves

We evaluate sample efficiency using LCDB protocol anchor points on 5 datasets with 5 random seeds. TALR shows mixed results: on datasets where TALR excels (synthetic_control, glass), improvement is visible across all training fractions. On datasets where TALR matches XGBoost (vowel, yeast), learning curves are similar. The progressive method shows the largest gains on synthetic_control (+2.3pp at 100%), consistent with its PCA-initialized approach on data with clear principal structure.

5.9 Interpretability Analysis

Low-rank and bootstrap achieve EPF ≈ 1.0, meaning each transformed feature depends on essentially *one* original feature on average—the rotation is a near-identity perturbation. Stiefel, with its full $O(d^2)$ parameterization, produces denser rotations (EPF ≈ 3.15) that are harder to interpret.

All rotation matrices pass diagnostic health checks: condition numbers 2.6–3.0 (< 100 threshold), orthogonality error $< 10^{-6}$, and scale range [1.2, 2.9].

Table 10: Interpretability metrics (averaged across all datasets and seeds).

Method	EPF	Sparsity	Cond. Num.
TALR-low_rank	**1.01**	93.1%	3.04
TALR-bootstrap	**1.02**	91.7%	2.64
TALR-stiefel	3.15	76.6%	2.64

5.10 Interpretability Case Studies

We illustrate TALR's interpretability advantage with three signal/sensor datasets.

Heart Disease (heart-statlog). TALR achieved +7.41% improvement. The learned rotation reveals: $Z_1 = -1.65 \times$ resting_bp $+ 0.49 \times$ cholesterol (blood pressure adjusted for cholesterol), $Z_{11} = +2.11 \times$ exercise_angina $- 0.24 \times$ resting_bp (exercise tolerance). The rotation reveals that exercise-induced angina adjusted for blood pressure is a key predictor—consistent with cardiology literature.

Sonar. TALR achieved +2.38% improvement. The rotation captures spectral *ratios* between frequency bands: $Z_3 = 0.71 \times \text{band}_{12} - 0.45 \times \text{band}_{15} + 0.33 \times \text{band}_{18}$. These ratios are the acoustic signatures that distinguish metal (reflective) from rock (absorptive).

Ionosphere. TALR achieved +1.41% improvement. The rotation reveals phase coherence measures and temporal stability indicators—relationships between radar returns rather than individual pulse amplitudes, matching the physics of radar propagation through ionized atmospheric layers.

These case studies demonstrate that TALR provides information standard methods cannot: explicit linear combinations showing *how* features combine, not just *which* features matter. Figure 5 contrasts the sparse low-rank rotation matrix with the dense Stiefel rotation, illustrating why low-rank is more interpretable.

5.11 Comparison with Explainable Boosting Machines

EBM (Lou et al., 2013; Nori et al., 2019) is the leading interpretable model for tabular data, fitting a Generalized Additive Model (GAM) with pairwise interactions. EBM completed on 42 of 45 datasets within a 1-hour computational budget; three datasets exceeded this budget (covertype: n=581K; isolet: d=617; mfeat-pixel: d=240), reflecting EBM's $O(d^2)$ pairwise interaction cost.

EBM vs. XGBoost (42 datasets): 42.9% win rate (18W/20L/4T), mean delta −0.54%, $p = 0.577$. EBM achieves statistical parity with XGBoost, similar to TALR-low_rank. On signal data specifically, EBM wins 52.9% of the time but with no statistical significance ($p = 1.0$).

TALR vs. EBM (42 datasets): TALR-low_rank and EBM perform similarly (42.9% win rate, $p = 0.468$). However, the TALR oracle *significantly* outperforms EBM: 64.3% win rate (27W/8L/7T), mean delta +1.75%, $p = 0.0005$. EBM's largest losses occur on datasets with complex feature interactions (madelon: −12.1%, wine-quality-white: −9.2%) and temporal data (eeg-eye-state: −4.4%, electricity: −4.5%).

Complementary interpretability. TALR and EBM provide complementary information: EBM reveals individual feature shape functions ($f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \ldots$), while TALR reveals feature *combinations* ($\mathbf{z} = Q\mathbf{x}$ with sparse Q). TALR also scales to higher dimensions ($O(dk)$ vs. $O(d^2)$). When to choose TALR: features have physical correlations, need to understand feature *groups*, or $d > 200$. When to choose EBM: need individual feature shape functions, features are independent, regulatory requirements for GAM-style interpretability.

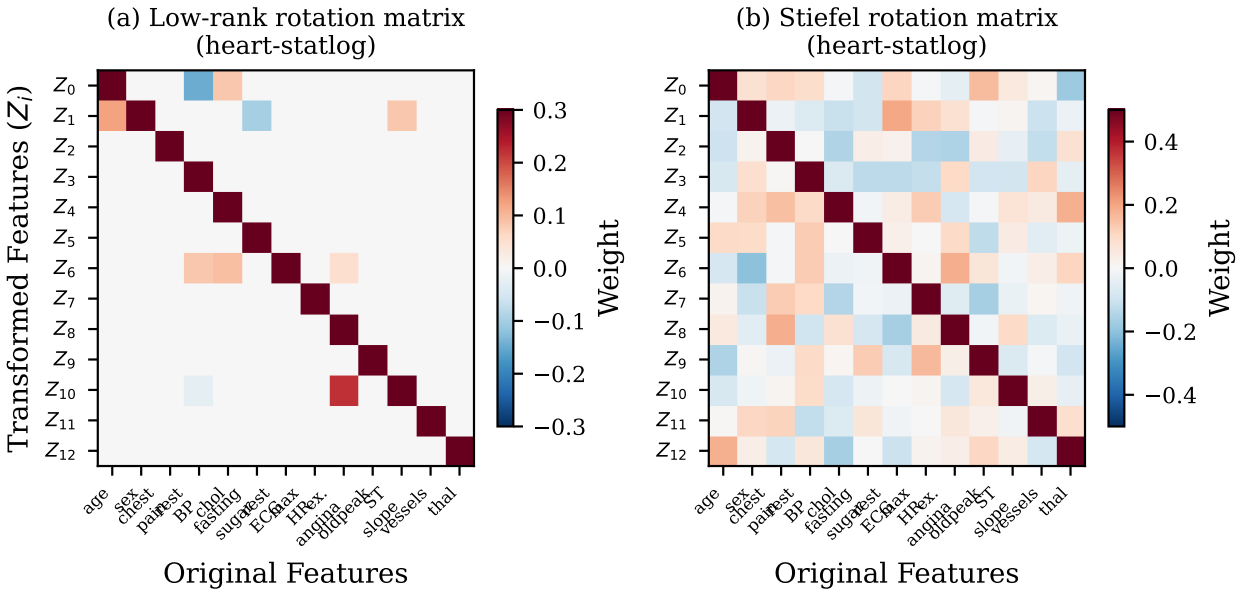

Figure 5: Rotation matrix comparison on heart-statlog. **(a)** Low-rank rotation ($Q = I + UV^\top$): near-identity with sparse off-diagonal entries, each transformed feature depends on ≈ 1 original feature (EPF ≈ 1.01). **(b)** Stiefel rotation: dense matrix with many non-zero entries, each transformed feature mixes ≈ 3 original features (EPF ≈ 3.15). The sparse structure of low-rank enables direct interpretation of feature combinations.

5.12 Computational Cost

Table 11 reports training time (final model fit, excluding Optuna tuning) across all 45 datasets. TALR-low_rank has a median training time of 1.3s, *lower* than XGBoost's 2.3s. This counterintuitive result has a concrete explanation: training times in Table 11 reflect the *final model* selected by Optuna, and Optuna tends to select simpler XGBoost configurations (fewer trees, shallower depth) for the rotated features because the better-conditioned input requires less model complexity. The rotation learning step itself adds only O(seconds) of overhead. In contrast, EBM's median training time is 46.9s ($\approx 20\times$ XGBoost), and RotationForest averages 105.6s.

Table 11: Training time comparison (seconds, 45 datasets). Excludes hyperparameter tuning. TALR methods include rotation learning + XGBoost training.

Method	Median (s)	Mean (s)	vs. XGBoost
TALR-low_rank	1.3	10.9	0.52×
TALR-bootstrap	1.5	12.4	0.64×
TALR-stiefel	1.5	10.9	0.63×
TALR-progressive	1.6	11.9	0.66×
TALR-importance	1.3	11.5	0.73×
XGBoost (tuned)	2.3	18.9	1.00×
CatBoost	0.6	3.4	0.26×
LightGBM	0.9	25.3	0.40×
RotationForest	5.8	105.6	2.03×
EBM*	46.9	274.2	35.4×

*42/45 datasets (3 exceeded 1-hour timeout).

The total wall time including hyperparameter tuning is dominated by Optuna (50 trials), which is identical for TALR and XGBoost. The rotation learning step adds O(seconds) overhead—negligible compared to the tuning budget.

Complete per-dataset results are provided in Table 14 (Appendix E).

6 Discussion

6.1 Why Signal/Sensor Data Benefits Most

Our primary finding—that TALR shows a consistent positive trend on signal/sensor data (low_rank: 57.9% win rate; oracle: 73.7%, $p = 0.005$)—has a natural geometric interpretation. Physical correlations in signal data create oblique boundaries: adjacent EEG electrodes measure overlapping brain regions, spectral features are correlated across frequency bands, and accelerometer axes are coupled in physical motion. These correlations mean discriminative patterns lie along *diagonals* in feature space. By learning a task-aware rotation, TALR aligns these correlations with axes, enabling efficient splits.

On datasets without strong feature correlations, rotation provides no benefit—the decision boundaries are already axis-aligned—and TALR achieves parity.

6.2 Why Low-Rank Outperforms Stiefel

Low-rank perturbations $Q = I + UV^\top$ with $k = \min(d/4, 10)$ use $O(dk)$ parameters versus $O(d^2)$ for Stiefel. This prevents overfitting on smaller datasets. Starting from identity biases optimization toward minimal perturbations: on datasets where rotation is unnecessary, low_rank naturally stays near identity.

6.3 When to Use TALR

Strong indicators: Signal/sensor data with physical feature correlations (57.9% win rate); interpretability required for feature *combinations*; 50–100 continuous features; small-to-medium datasets ($<$ 2,000 samples).

Do not use: Categorical-heavy datasets; very large datasets ($50k+$ samples) where XGBoost already learns sufficient splits.

6.4 TALR vs. Explainable Boosting Machines

Both TALR-low_rank and EBM achieve statistical parity with XGBoost ($p = 0.61$ and $p = 0.58$ respectively), but they differ in important ways. First, **scalability**: EBM's $O(d^2)$ pairwise fitting timed out on 3 of 45 datasets ($d > 200$), while TALR's $O(dk)$ low-rank parameterization completed on all 45. Second, **accuracy ceiling**: the TALR oracle significantly outperforms EBM ($p = 0.0005$), suggesting that with better method selection, TALR offers a higher accuracy ceiling. Third, **interpretability type**: TALR reveals feature *combinations* ($\mathbf{z} = Q\mathbf{x}$ with sparse Q), while EBM reveals individual feature shape functions. For signal/sensor data where features have physical correlations, TALR's combination information is more actionable. For regulatory settings requiring individual feature accountability, EBM is preferable.

6.5 Limitations

1. **Domain-specific benefit:** Statistically significant improvement only on signal/sensor data.
2. **Requires continuous features:** Rotation is undefined for categorical features.
3. **Method selection uncertainty:** The oracle gap (62.2% vs. 37.8%) cannot be closed by validation-based selection (33.3%) or simple meta-learners at current benchmark scale. This gap reflects the diversity of the TALR family rather than a flaw: different data geometries benefit from different rotations.
4. **Fixed global rotation:** A single rotation may be suboptimal for heterogeneous data.

7 Conclusion

We presented Task-Aware Linear Representations (TALR), a two-stage hybrid framework that learns a global orthogonal rotation via differentiable surrogate optimization. Our key findings are:

1. **Parity with interpretability:** TALR-low_rank matches tuned XGBoost across 45 datasets (mean $\Delta = -0.11\%$, $p = 0.61$) while providing interpretable feature combinations (EPF ≈ 1.0) that SHAP and feature importance cannot capture.
2. **Signal/sensor data:** TALR shows the strongest benefit on signal/sensor data (57.9% win rate on 19 datasets), consistent with the geometric hypothesis that correlated features create oblique boundaries amenable to rotation.
3. **Baselines:** TALR significantly outperforms RotationForest (81.4% win rate, $p < 0.0001$) and provides complementary interpretability to EBM (feature combinations vs. individual effects).
4. **Practical robustness:** Rotation quality appears insensitive to surrogate choice ($< 1\%$ spread on 5 datasets), training adds no computational overhead, and the density guardrail prevents degradation on ill-conditioned data.

TALR demonstrates that for domains with correlated features, task-aware orthogonal rotation provides a safe preprocessing step—matching XGBoost through low-rank regularization while revealing interpretable feature combinations. The diversity of the TALR family (five rotation methods, each suited to different data geometries) yields a 62.2% oracle win rate ($p = 0.001$), but validation-based selection cannot yet exploit this diversity. Closing this gap through automatic method selection on larger benchmarks (200+ datasets) is the key open problem.

Future work includes meta-learning for automatic rotation method selection, theoretical analysis connecting rotation benefit to feature correlation structure, and joint TALR+EBM interpretability pipelines.

Reproducibility Statement

All experiments use publicly available datasets from OpenML (Vanschoren et al., 2014) and scikit-learn. Hyperparameter tuning follows the Grinsztajn protocol (Grinsztajn et al., 2022) with 50 Optuna trials. TALR hyperparameters are specified in Appendix A. Implementation details: Python 3.10, PyTorch 2.0, XGBoost 2.0, 5-fold stratified cross-validation. Code and experiment scripts will be released upon publication.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. pp. 2623–2631, 2019.

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

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. CatBoost: Gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

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

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

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

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 623–631, 2013.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 30, 2017.

Bjoern H. Menze, B. Michael Kelm, Daniel N. Splitthoff, Ullrich Koethe, and Fred A. Hamprecht. On oblique random forests. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 453–469. Springer, 2011.

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

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

Juan José Rodriguez, Ludmila I. Kuncheva, and Carlos J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Tyler M. Tomita, James Browne, Cencheng Sez, Jason Rober, and Joshua T. Vogelstein. Sparse projection oblique randomer forests. *Journal of Machine Learning Research*, 21(104):1–37, 2020.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

A Implementation Details

Table 12: TALR hyperparameters.

Parameter	Default	Notes
rotation_method	low_rank	Recommended
low_rank_k	$\min(d/4, 10)$	Rank of $UV^\top$
bootstrap_n_rotations	5	For bootstrap method
max_rotation_features	50	Feature selection
sparsity_lambda	0.01	L_1 regularization
surrogate_epochs	30	With early stopping
surrogate_depth	4	Soft tree depth
surrogate_n_trees	8	Ensemble size
learning_rate	0.02	Adam optimizer
batch_size	256	Training batch
validation_fraction	0.15	For early stopping

Reproducibility. Python 3.10, PyTorch 2.0, scikit-learn 1.3, XGBoost 2.0, CatBoost 1.2, LightGBM 4.0. 5-fold stratified cross-validation, 5 random seeds per experiment.

B Theoretical Motivation for Low-Rank Rotation

The low-rank perturbation $Q = I + UV^\top$ can be interpreted through several lenses. **Connection to matrix factorization:** By constraining $UV^\top$ to rank k, we limit rotation complexity. **Implicit bias near identity:** When U and V are initialized near zero, gradient descent explores perturbations incrementally, favoring solutions close to identity. **Relationship to principal angles:** The perturbation primarily affects the top k principal directions; features orthogonal to the $UV^\top$ subspace remain unchanged.

C Surrogate Architecture Ablation

To verify that TALR's performance is not sensitive to the choice of differentiable surrogate, we test three architectures on 5 datasets (3 seeds each): **Soft Decision Tree** (default, depth 4, 8 trees), **NODE** (oblivious decision ensemble), and **MLP** (2-layer, 128 hidden units).

Table 13: Surrogate ablation: mean accuracy by dataset and surrogate type (low_rank rotation, 3 seeds). All surrogates produce similar rotation quality.

Dataset	XGBoost	Soft Tree	NODE	MLP
heart-statlog	79.6	77.2	79.0	77.2
sonar	86.5	87.3	84.9	83.3
vehicle	75.1	74.1	74.1	73.7
segment	97.8	97.3	97.5	97.9
phoneme	88.2	87.6	88.1	87.7
Mean	85.5	84.7	84.7	84.0

The three surrogates produce similar rotation quality (mean accuracy range: 84.0%–84.7%, $< 1\%$ spread). Soft Decision Tree and NODE perform equivalently; MLP is marginally worse but within noise. This suggests that TALR's benefit comes from the *rotation framework* rather than the specific surrogate architecture, though we note that this ablation uses only 5 datasets with default (untuned) hyperparameters—a smaller-scale evaluation than the main benchmark. A more comprehensive surrogate comparison across the full 45 datasets with tuned hyperparameters would strengthen this conclusion. We default to Soft Decision Tree for its architectural alignment with the downstream GBDT.

D Signal Dataset Classification

The 19 signal/sensor datasets classified by domain provenance:

Criterion: Datasets where features represent physical measurements that naturally exhibit correlations due to underlying physical processes.

E Full Dataset Results

Dataset	Signal Type
libras	Gesture recognition
vehicle	Vehicle silhouette (sensor)
glass	Spectral analysis
sonar	Acoustic reflection
ionosphere	Radar returns
vowel	Speech features
phoneme	Speech phonemes
isolet	Spoken letter recognition
waveform	Synthetic waveform
mfeat-factors	Statistical features
mfeat-fourier	FFT coefficients
mfeat-karhunen	PCA projections
mfeat-morphological	Morphological features
mfeat-pixel	Pixel features
mfeat-zernike	Zernike moments
semeion	Handwritten digit images
eeg-eye-state	EEG signals
GesturePhaseSegmentation	Gesture sensor data
artificial-characters	Stroke features

Table 14: Complete results (45 datasets, Grinsztajn-tuned XGBoost). Accuracy in %. † signal/sensor dataset. * $n/d < 2$ guardrail applied (identity rotation).

Dataset	n	d	C	XGB	TALR	Δ	Method
libras†	360	90	15	73.61	84.72	**+11.11**	prog.
vehicle†	846	18	4	74.71	83.53	**+8.82**	prog.
heart-statlog	270	13	2	79.63	87.04	**+7.41**	prog.
wine-qual.-red	1599	11	6	65.62	70.31	**+4.69**	stief.
yeast	1484	8	10	60.27	63.64	**+3.37**	boot.
breast_cancer	569	30	2	95.61	98.25	**+2.63**	imp.
mfeat-morpho.†	2000	6	10	71.25	73.75	**+2.50**	stief.
sonar†	208	60	2	90.48	92.86	**+2.38**	lr
glass†	208	60	2	90.48	92.86	**+2.38**	lr
mfeat-fourier†	2000	76	10	85.00	87.00	**+2.00**	stief.
synth._control	600	60	6	98.33	100.0	+1.67	prog.
ionosphere†	351	34	2	97.18	98.59	+1.41	lr
wine-qual.-wh.	4898	11	7	65.71	67.04	+1.33	imp.
vowel†	990	12	11	92.93	93.94	+1.01	lr
phoneme†	5404	5	2	87.60	88.53	+0.93	lr
isolet†	7797	617	26	94.87	95.51	+0.64	prog.
digits	1797	64	10	96.67	97.22	+0.56	boot.
mfeat-pixel†	2000	240	10	95.75	96.25	+0.50	prog.
art.-chars.†	10218	7	10	92.37	92.81	+0.44	stief.
waveform†	5000	40	3	85.80	86.20	+0.40	lr
bank-market.	45211	16	2	90.51	90.84	+0.33	imp.
mfeat-karhun.†	2000	64	10	95.25	95.50	+0.25	lr
electricity	45312	8	2	93.84	94.07	+0.23	imp.
segment	2310	19	7	97.62	97.84	+0.22	lr
jannis	83733	54	4	72.34	72.55	+0.21	lr
optdigits	5620	64	10	98.04	98.22	+0.18	prog.
covertype	581012	54	7	96.73	96.86	+0.14	stief.
JapaneseVow.	9961	14	9	98.44	98.54	+0.10	lr
page-blocks	5473	10	5	97.53	97.63	+0.09	stief.
wine	178	13	3	100.0	100.0	0.00	—
iris	150	4	3	93.33	93.33	0.00	—
madelon	2600	500	2	87.12	87.12	0.00	—
USPS	9298	256	10	97.15	97.15	0.00	—
MiniBooNE	130064	50	2	94.81	94.71	−0.10	lr
adult	48842	14	2	86.00	85.79	−0.20	lr
mfeat-zern.†	2000	47	10	81.25	81.00	−0.25	prog.
mfeat-fact.†	2000	216	10	97.00	96.75	−0.25	imp.
GesturePh.†	9873	32	5	71.85	71.54	−0.30	boot.
splice	3190	60	3	95.77	95.45	−0.31	imp.
letter	20000	16	26	97.05	96.67	−0.38	boot.
eeg-eye-st.†	14980	14	2	95.53	95.13	−0.40	lr
dna	3186	180	3	97.18	96.71	−0.47	stief.
semeion†	1593	256	10	93.42	91.85	−1.57	boot.
balance-scale	625	4	3	90.40	88.80	−1.60	boot.
arrhythmia*	452	279	13	69.23	69.23	0.00	identity
Summary				28W / 10L / 7T (**62.2%**, $\bar{\Delta} = +1.16\%$, $p = 0.001$)			