GB–KAN: Gradient Boosting with Interpretable Kolmogorov–Arnold Networks

Janis Mohr[a] and Jörg Frochte[b]
Interdisciplinary Institute for Applied Artificial Intelligence and Data Science Ruhr, Bochum University of Applied Sciences, 42579 Heiligenhaus, Germany

Keywords: Gradient Boosting, Kolmogorov-Arnold Networks, Machine Learning, Interpretable Models.

Abstract: Gradient boosting remains a dominant approach for tabular data, but its widespread reliance on decision trees limits interpretability and the smoothness of learned functions. We propose Gradient-Boosted Kolmogorov-Arnold Networks (GB-KANs), a boosting framework that replaces trees with shallow Kolmogorov-Arnold networks as base learners. By fitting KAN shape functions to pseudo-residuals in an additive, stagewise manner, GB-KANs inherit the predictive strength of boosting while yielding per-feature functions that are easy to inspect. Across real-world datasets, GB-KANs achieve accuracy competitive with established baselines and produce shape functions that are faithful, sparse, and stable. They also yield well-calibrated probabilities without post-hoc correction. These properties position GB-KANs as a promising approach when interpretability is essential.

1 INTRODUCTION

Gradient boosting is a leading approach for tabular learning, with strong results across domains (Chen and Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018; McElfresh et al., 2023). In practice, modern libraries rely almost exclusively on decision trees as base learners. While trees are flexible and robust to heterogeneous features, their piecewise-constant structure hampers interpretability and gradient-based analysis and can behave unreliably outside the observed domain (Cai et al., 2023).

In parallel, interpretable modeling has advanced from classical generalized additive models (GAMs) to modern neural shape-function architectures. Explainable Boosting Machines (EBMs) (Nori et al., 2019) and Neural Additive Models (NAMs) (Agarwal et al., 2021) learn per-feature functional contributions, offering transparent insights while maintaining competitive accuracy. These methods show that structured decompositions can be both accurate and interpretable, but their additive structure limits higher-order interaction modeling compared to tree ensembles. Recently, Kolmogorov-Arnold Networks (KANs) (Liu et al., 2024) have been proposed as a flexible neural architecture that represents functions as compositions of one-dimensional splines. Their smooth structure and per-dimension functional decomposition make them appealing for interpretable learning, but they have not yet been explored in a boosting framework.

KANs offer several properties that make them promising as base learners in a boosting setup.Their spline-based parameterization induces smooth, low-complexity shape functions that can be plotted as univariate curves, offering direct and faithful visualizations of feature effects, in contrast to the discontinuous, high-dimensional decision boundaries of tree ensembles. The localized basis functions used by KANs can capture nonlinear effects while remaining sparse, potentially improving sample efficiency. Moreover, their differentiable structure permits gradient-based regularization such as total variation or group sparsity, further encouraging structured representations. These characteristics suggest that replacing trees with KANs could combine the interpretability of shape-based models with the predictive performance and robustness of gradient boosting.

This gap motivates our work. We propose *Gradient-Boosted Kolmogorov-Arnold Networks (GB-KANs)*, a new boosting framework that replaces decision trees with shallow KANs as base learners. This design combines the structured interpretability of KANs with the proven optimization dynamics of gradient boosting. We develop an efficient training procedure based on functional gradients with total variation and group sparsity regularization, enabling GB-KANs to learn sparse and smooth functional rep-

[a] https://orcid.org/0000-0001-6450-074X
[b] https://orcid.org/0000-0002-5908-5649

Mohr, J. and Frochte, J.
GB-KAN: Gradient Boosting with Interpretable Kolmogorov-Arnold Networks.
DOI: 10.5220/0014246800004052

In *Proceedings of the 18th International Conference on Agents and Artificial Intelligence (ICAART 2026) - Volume 3*, pages 2421-2432
ISBN: 978-989-758-796-2; ISSN: 2184-433X

resentations. To our knowledge, GB-KAN is the first boosting framework to combine competitive tabular performance with smooth, low-complexity, human-readable base learners.

Our main contributions are:

- **A Novel Gradient Boosting Framework Using KAN base learners.** This combines the interpretability of functional shape models with the performance of stage-wise boosting.
- **A Training Algorithm with Structural Regularization.** We integrate total variation and group sparsity penalties to enforce smoothness and sparsity in the learned functions.
- **An Empirical Evaluation across Real Datasets.** We show that GB-KANs achieve accuracy competitive with strong baselines while offering interpretable functional representations.

The remainder of the paper is organized as follows. Section 2 reviews related work on tree-based boosting and interpretable neural models. Section 3 introduces the GB-KAN framework and its regularized training procedure. Section 4 describes the datasets, experimental setup, and empirical results on accuracy, calibration, and interpretability. Section 5 concludes and outlines directions for future work.

2 RELATED WORK

To contextualize our contributions, we summarize advances in tree-based boosting, neural/differentiable additive models, and KANs, and identify unresolved limitations.

Tree-Based Gradient Boosting. Gradient boosting is a widely used and highly effective method for tabular data (Friedman, 2001). Modern implementations such as XGBOOST (Chen and Guestrin, 2016), LIGHTGBM (Ke et al., 2017), and CATBOOST (Prokhorenkova et al., 2018; Dorogush et al., 2018) refine the original framework with second-order optimization, shrinkage, feature subsampling, and extensive system-level engineering. These methods have become the de facto standard on many real-world tasks. However, their reliance on decision trees produces piecewise-constant predictions that are difficult to interpret and often unstable when extrapolating beyond the training domain. Post-hoc methods like SHAP (Lundberg and Lee, 2017) provide coarse summaries but do not fundamentally alter the tree ensemble's complexity.

Neural and Differentiable Boosting. Several recent approaches attempt to combine boosting with neural networks to improve smoothness and representation learning. DeepGBM (Ke et al., 2019) integrates gradient-boosted trees with deep neural networks via embedding layers, while NODE (Popov et al., 2019) implements fully differentiable oblique decision trees optimized end-to-end by gradient descent. TabNet (Arik and Pfister, 2021), DANet (Cheng et al., 2022), and SAINT (Somepalli et al., 2021) introduce attention-based architectures for tabular data that partially mimic boosting-like feature selection dynamics. (Mohr et al., 2023) propose Multiple Additive Neural Networks (MANNs), which use additive neural components in continual learning settings for regression and classification. Their architecture decomposes functions additively over features, similarly aiming for structure and additionally continuous learning. These approaches often achieve strong accuracy but remain black-box predictors, offering limited insight into feature effects.

(Hybrid) Interpretable Additive Models. A complementary line of work has focused on interpretable predictive modeling by constraining model structure. Classical generalized additive models (GAMs) (Hastie and Tibshirani, 1986) decompose predictions into per-feature contributions using smooth basis functions. Explainable Boosting Machines (EBMs) (Nori et al., 2019) extend GAMs with gradient boosting to learn flexible shape functions and sparse pairwise interactions, while Neural Additive Models (NAMs) (Agarwal et al., 2021) employ neural networks to learn smooth per-feature functions end-to-end. These approaches show that enforcing structured decompositions can yield models that are both accurate and interpretable, yet their restricted additive structure can limit performance, especially on tasks with complex higher-order feature interactions. Several works explore hybrids between neural architectures and additive or symbolic structure. GA^{2}Ms (Caruana et al., 2015) and Neural GAMs (Yang et al., 2021) aim to combine the interpretability of GAM-like decompositions with the flexibility of neural networks. Symbolic regression ensembles (Kamienny et al., 2022) and sparse concept models (Ross et al., 2022) attempt to learn human-readable functions from data while preserving accuracy. While promising, these models often rely on global optimization or discrete search and do not benefit from the regularizing effect of stage-wise boosting. They also rarely offer mechanisms to control capacity through shrinkage or to incorporate gradient-based smoothness penalties.

Kolmogorov-Arnold Networks. Kolmogorov-Arnold Networks (Liu et al., 2024) are a recent neural architecture inspired by the Kolmogorov-Arnold representation theorem (Kolmogorov, 1957; Arnold, 1957). They represent functions as compositions of one-dimensional spline-based transformations, yielding interpretable and smooth shape functions. Early work has shown their potential as interpretable alternatives to multilayer perceptrons, but they have not yet been explored in ensemble or boosting frameworks (Barašin et al., 2024). Moreover, prior studies have not examined how to regularize KANs for sparsity or stability in high-dimensional settings (Cruz et al., 2025).

Our work bridges these lines of research by combining the structured interpretability of KANs with the optimization dynamics of gradient boosting. To our knowledge, GB-KANs are the first boosting framework to employ KANs as base learners. This design enables stage-wise additive optimization over smooth, low-complexity functional components, offering a new accuracy-interpretability trade-off not explored in prior work.

3 METHODOLOGY: GRADIENT-BOOSTED KOLMOGOROV-ARNOLD NETWORKS (GB - KAN)

Shallow Kolmogorov–Arnold Networks (KANs) are integrated as weak learners in a standard functional gradient-boosting loop. Boosting provides robust stage-wise additive optimization on tabular data; KANs contribute smooth, spline-based per-feature shape functions that are directly inspectable. Using shallow KANs preserves the weak-learner regime, while optional total-variation and group-sparsity penalties encourage smoothness and compact feature use. This section first recalls the general formulation of functional gradient boosting, then introduces our KAN base learners, and finally describes the regularized training procedure that yields stable and interpretable models.

3.1 Gradient Boosting Framework

We formulate our approach within the functional gradient boosting framework (Friedman, 2001). Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and a differentiable loss $\ell(\hat{y}, y)$, boosting constructs an additive model

$$F^{(T)}(\mathbf{x}) = \sum_{t=1}^{T} \nu\, f_t(\mathbf{x}; \theta_t) \tag{1}$$

where f_t are base learners with parameters θ_t, and $\nu \in (0,1]$ is a learning rate. At each stage t, the model fits a new base learner to the negative gradient, i.e. the pseudo-residuals:

$$r_i^{(t)} = -\left.\frac{\partial \ell(\hat{y}, y_i)}{\partial \hat{y}}\right|_{\hat{y}=F^{(t-1)}(\mathbf{x}_i)} \tag{2}$$

The new learner f_t is trained to minimize the squared error to these residuals, followed by a line search to find the optimal step size:

$$\gamma_t = \arg\min_{\gamma} \sum_{i=1}^{n} \ell\big(F^{(t-1)}(\mathbf{x}_i) + \gamma f_t(\mathbf{x}_i), y_i\big) \tag{3}$$

and the model is updated as

$$F^{(t)} \leftarrow F^{(t-1)} + \nu\gamma_t f_t.$$

3.2 Kolmogorov-Arnold Base Learners

Each base learner f_t is a shallow Kolmogorov-Arnold Network (KAN) (Liu et al., 2024), which represents multivariate functions as compositions of one-dimensional spline functions. Concretely, given d input features, a KAN computes

$$f_t(\mathbf{x}) = \sum_{j=1}^{m} w_j \phi_j\Big(\sum_{k=1}^{d} u_{jk}\, s_{jk}(x_k)\Big) \tag{4}$$

where s_{jk} are trainable spline basis functions, u_{jk} are linear weights, ϕ_j are activation functions (e.g., identity or ReLU), and w_j are output weights. This construction induces smooth, low-complexity shape functions that can be visualized and interpreted per feature. A *shallow Kolmogorov–Arnold Network (KAN)* is a single-layer KAN where each input feature is mapped through a spline-based univariate function $S_j(x_j)$, and the outputs are linearly combined. This additive form preserves interpretability via per-feature shape functions while providing smooth, differentiable approximations suitable as weak learners in boosting.

3.3 Regularized Training Procedure

To encourage sparse and structured representations, we add total variation and group sparsity penalties to the base learner objective:

$$\begin{aligned}\mathcal{L}_t = \frac{1}{n}\sum_{i=1}^{n}\big(f_t(\mathbf{x}_i) - r_i^{(t)}\big)^2 \\ +\lambda_{\text{tv}}\sum_{j,k}\text{TV}(s_{jk}) \\ +\lambda_{\text{grp}}\sum_{k}\|u_{\cdot k}\|_2\end{aligned} \tag{5}$$

Here $\mathrm{TV}(s_{jk})$ denotes the discrete total variation of the spline coefficients for s_{jk}, i.e., if s_{jk} is parameterized by coefficients $(\alpha_m)_m$, then

$$\mathrm{TV}(s_{jk}) = \sum_m |\alpha_{m+1} - \alpha_m| .$$

We optimize $\mathcal{L}_t$ using gradient-based methods for a fixed number of epochs. After training f_t, we perform a line search to obtain γ_t and update the ensemble as described above. Here $u_{\cdot k}$ denotes the vector of all incoming weights from feature k across all hidden units.

This procedure yields an additive model whose components are smooth and interpretable KANs while retaining the optimization dynamics of gradient boosting.

While the general gradient boosting formulation (Section 3.1) includes a per-stage line search to determine γ_t, our implementation uses a fixed shrinkage parameter ν for efficiency. Conceptually, ν can be interpreted as absorbing γ_t into a constant step size, i.e. using $\eta_t = \nu\gamma_t$ with $\gamma_t \equiv 1$. This simplification preserves the boosting dynamics while reducing computational overhead.

3.4 Interpretability in GB - KANs

A key advantage of GB–KANs is that they expose *native* explanatory objects: each base learner is a composition of one-dimensional spline functions, and the boosted ensemble remains an additive sum of such components. Let $S_j(x_j)$ denote the per-feature effect, centered to zero mean for identifiability. Because the overall predictor is

$$F^{(T)}(\mathbf{x}) = \sum_{t=1}^{T} \nu f_t(\mathbf{x}),$$

global and local explanations can be derived directly from $\{S_j\}$, without post-hoc surrogates. Here $f_{t,j}$ denotes the feature-wise marginal contribution extracted from the stage-t KAN (via its edge-wise spline decomposition; other inputs are held fixed at a reference and edge contributions are aggregated). We use the term interpretability for model-intrinsic structures that are directly understandable by humans, and explainability for artifacts that justify specific predictions. Models with intrinsic interpretability typically provide explanations ante hoc via their own decomposition. Therefore, GB–KAN, unlike most models, does not rely on post-hoc explainability techniques (Ribeiro et al., 2016), but constitutes an interpretable model class by design (Doshi-Velez and Kim, 2017). Its shallow KAN weak learners expose human-readable spline functions, making the model intrinsically interpretable in the sense of (Rudin, 2018). At the same time, the additive decomposition allows faithful, prediction-level explanations, thus bridging interpretability and explainability.

Algorithm 1: GB-KAN training.

Require: Data $\{(x_i, y_i)\}_{i=1}^n$, loss $\ell \in \{\mathrm{MSE}, \mathrm{logistic}\}$, rounds T, learning rate ν, KAN hyperparams (degree/knots/hidden units/layers; optional skip), row subsampling rate π_{row}, validation set $\mathcal{V}$ with patience P.

Init:
if $\ell = \mathrm{MSE}$ **then** $F^{(0)}(x) \leftarrow \frac{1}{n}\sum_{i=1}^n y_i$
else $F^{(0)}(x) \leftarrow \log\frac{\pi+\varepsilon}{1-\pi+\varepsilon}$, with $\pi = \frac{1}{n}\sum_i \mathbb{1}\{y_i = 1\}$
best $\leftarrow +\infty$, $t_\star \leftarrow 0$, no_imp $\leftarrow 0$.
for $t = 1, 2, \ldots, T$ **do**
  Draw a row subset $\mathcal{S}_t$ of size $\lfloor \pi_{\mathrm{row}} n \rfloor$ *without replacement*.
  Residuals:
  if $\ell = \mathrm{MSE}$ **then** $r_i^{(t)} \leftarrow y_i - F^{(t-1)}(x_i)$
  else $p_i \leftarrow \sigma(F^{(t-1)}(x_i))$, $r_i^{(t)} \leftarrow y_i - p_i$ (*binary*)
  Stage fit: train a shallow KAN $f_t(\cdot;\theta_t)$ on $\{(x_i, r_i^{(t)}) : i \in \mathcal{S}_t\}$ by least squares (mini-batch AdamW; optional smoothness penalty set to 0 unless stated).
  Update: $F^{(t)}(x) \leftarrow F^{(t-1)}(x) + \nu f_t(x;\theta_t)$.
  Validation:
  if $\ell = \mathrm{MSE}$ **then** val $\leftarrow \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(y_i - F^{(t)}(x_i))^2$
  else val $\leftarrow \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \mathrm{logloss}(y_i, \sigma(F^{(t)}(x_i)))$
  Early stop:
  if val $<$ best **then** best $\leftarrow$ val; $t_\star \leftarrow t$; no_imp $\leftarrow 0$
  else no_imp $\leftarrow$ no_imp+1;
    if no_imp $> P$ **then break**
return $F^{(t_\star)}$.

Global Shape Functions and Sensitivities. For each feature x_j, we obtain a smooth effect curve

$$S_j(x_j) = \sum_{t=1}^{T} \nu f_{t,j}(x_j), \qquad S_j'(x_j) = \tfrac{d}{dx_j} S_j(x_j),$$

where $f_{t,j}$ is the stage-t contribution of x_j. The shape S_j summarizes directionality and saturation, while its derivative S_j' highlights regions of high sensitivity. From these we can identify *regimes* (intervals that are approximately monotone or flat) and *hot zones* (regions with large $|S_j'|$). These regimes form the basis for simple human-readable rules.

Stage-Wise Provenance. Because boosting is stage-wise, the evolution of feature effects can be visualized through *storyboards*, plotting the incremental contributions $\Delta F_j^{(t)}(x_j) = \nu f_{t,j}(x_j)$ alongside the cumulative shape S_j. This reveals when residual structures are corrected and whether interactions emerge progressively through repeated co-updates of different features.

Sparsity and Structure. The edge weights u_{jk} in the base learners induce a sparse feature×unit connectivity pattern when aggregated across stages. We summarize this with edge-sparsity heatmaps and with an *active feature count* based on $\|S_j\|_{2,\mathcal{G}}$ over a grid $\mathcal{G}$. When monotonicity constraints are imposed, we further report the fraction of grid points where empirical derivatives S_j' violate the expected sign.

Local-to-Global Explanations. For individual predictions, per-feature contributions $S_j(x_j)$ naturally form waterfall plots that sum to $F^{(T)}(\mathbf{x})$. Each local contribution can be overlaid on the corresponding global curve $S_j(\cdot)$ to check for consistency. Because $F^{(T)}$ is additive, attribution methods such as SHAP or integrated gradients reduce to simple per-stage decompositions, which can even be refined to edge-level components within a KAN.

Domain- Specific Uses. While the same interpretability tools apply across application types, their role differs. In semantically meaningful tabular data (e.g., medical or socio-economic domains), shapes, storyboards, and rules yield concise and accessible summaries for human decision-makers. In scientific or physical systems, the same smooth effects can be projected onto physics-informed bases, approximated with symbolic surrogates, or tested for invariances; providing insight into underlying laws. Thus GB–KANs offer a unified framework whose interpretability adapts naturally to different contexts.

3.5 Complexity & Regularization

One GB–KAN stage trains a shallow KAN with m units for E epochs on n samples and d features; the per-epoch cost is $O(n \cdot d \cdot m)$, yielding $O(T \cdot E \cdot n \cdot d \cdot m)$ over T stages. Convergence follows the standard functional boosting view for convex, differentiable losses (Friedman, 2001); in practice we rely on fixed shrinkage and early stopping. Generalization is controlled by the usual path-length perspective; small learning rate ν, moderate T, and weak learners and by KAN-native structure: capped spline resolution/width and optional total-variation and group-sparsity penalties. We leave monotonicity constraints disabled unless stated, to isolate the effect of boosting with KANs. Our empirical results (Section 4) confirm stable convergence across datasets and folds.

4 EXPERIMENTS

Gradient-Boosted Kolmogorov-Arnold Networks (GB-KANs) are evaluated on real-world tabular datasets. Our experiments address three aspects: (i) predictive accuracy, (ii) calibration, and (iii) interpretability and stability of the learned representations. The key question is whether GB-KANs can match state-of-the-art baselines in accuracy while offering smoother, more interpretable shape functions and better-calibrated confidence estimates.

4.1 Datasets

We use five established tabular benchmarks, selected to cover both classical ML benchmarks and high-stakes real-world applications:

Diabetes progression **(Efron et al., 2004)**
Regression dataset with n=442, d=10 covariates (age, sex, BMI, blood pressure, six serum markers). Target: one-year progression.

Concrete Strength **(Yeh, 1998)**
Regression dataset with n=1030, d=8 mixture features; physically interpretable rules (e.g., water–cement ratios).

Breast Cancer **(Dua and Graff, 2017)**
Binary classification, n=569, d=30 image-derived features; canonical medical benchmark.

California Housing **(Kelley Pace and Barry, 1997)**
Regression dataset with n=20,640, d=8; California house value prediction, large-scale regression.

FICO Credit Risk **(Chen et al., 2018)**
Binary classification, n=10,459, d=23; credit bureau features, high-stakes financial domain.

For classification datasets we use stratified 5-fold cross-validation; for regression datasets we use standard 5-fold cross-validation with shuffling and fixed random seeds. Continuous features are standardized; categorical features are one-hot encoded. "For categorical variables, we add an explicit 'Missing' category for absent values within each training fold and group levels with fewer than 5 samples (or $< 0.5\%$ frequency) into a single 'Other' category. These encodings are fitted on the training fold and applied to

the corresponding validation/test splits to avoid leakage. Concrete Strength is included as an interpretability case study rather than in the main accuracy table, due to its small size and strong collinearities.

4.2 Baselines and Training Setup

We compare GB-KANs against strong tree-based, neural, and non-boosting baselines: XGBOOST (Chen and Guestrin, 2016), GRADIENTBOOSTINGREGRESSOR from SCIKIT-LEARN, LIGHTGBM (Ke et al., 2017), CATBOOST (Prokhorenkova et al., 2018), RANDOMFOREST, KNN, SVR-RBF, RIDGE, LASSO (Efron et al., 2004), ELASTICNET, TABTRANSFORMER (Huang et al., 2020), and shallow MLPS. Hyperparameters are tuned by grid search on the training folds. The search ranges are capacity-matched: for example, boosting baselines explore $\nu \in \{0.05, 0.1\}$ and $T \in \{100, 200, 300\}$; GB-KANs explore the same ranges plus spline grid resolution (4-10 knots), polynomial degree (1-3), and hidden width (10-12). This alignment avoids conferring method-specific advantages. All methods use the same splits, seeds, and evaluation protocol.

4.3 Evaluation Protocol

Unless otherwise stated, we use 5-fold cross-validation (stratified for classification) and report fold means. For brevity, we omit confidence intervals in the main tables.

For regression we report root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). For classification we measure probabilistic and ranking quality: cross-entropy (log-loss), Brier score (Brier, 1950) and ROC-AUC (Bradley, 1997). We quantify calibration using expected calibration error (ECE) (Nixon et al., 2019; Posocco and Bonnefoy, 2021) with both 10/20 equal-width bins and 10 quantile bins:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left| \widehat{\mathrm{acc}}(B_b) - \widehat{\mathrm{conf}}(B_b) \right| \qquad (6)$$

where B_b are bins, $\widehat{\mathrm{acc}}(B_b)$ is empirical accuracy, and $\widehat{\mathrm{conf}}(B_b)$ is mean predicted confidence. No post-hoc calibration is applied. Results can be seen in table 1. The original Diabetes dataset is a regression task. For our *calibration* analysis we use a binary variant: on each training fold we compute a threshold τ (the median of the target on that fold) and define labels as $y_i^{\mathrm{bin}} = \mathbb{1}\{y_i \geq \tau\}$. The same τ is then applied to the corresponding validation/test split to avoid leakage. All preprocessing steps and the threshold are fitted strictly within the training fold. Overall, GB–KAN achieves the lowest ECE under most binning schemes, indicating better calibrated probabilities than tree-based boosting baselines. While CatBoost is competitive under 20-bin uniform discretization, its performance is less consistent across other binning strategies. This suggests that GB–KAN not only produces accurate predictions, but also well-calibrated confidence estimates without requiring post-hoc calibration.

Table 1: Calibration summary (ECE with 10/20 uniform bins and 10 quantile bins) on Diabetes (binary variant). Lower is better; values are 5-fold means.

Model	ECE@10 (uniform)	ECE@20 (uniform)	ECE@10 (quantile)
GB-KAN	0.034	0.115	0.067
XGBoost	0.065	0.111	0.081
LightGBM	0.057	0.124	0.078
CatBoost	0.075	0.086	0.088

Table 2: Breast cancer classification benchmark. Lower log-loss and Brier indicate better calibration; higher ROC–AUC indicates stronger discrimination. Best results in **bold** and mean across 5 folds.

Model	LogLoss ↓	Brier ↓	ROC–AUC ↑
GB–KAN	**0.074**	**0.024**	**0.997**
CatBoost	0.092	0.025	0.995
XGBoost	0.093	0.031	0.994
MLP	0.102	0.029	0.996
LightGBM	0.103	0.028	0.991
Random Forest	0.131	0.038	0.993
Gradient Boosting	0.173	0.041	0.990

4.4 Predictive Performance on Regression and Classification

We evaluate GB–KAN on predictive accuracy. On the binary classification breast cancer dataset, GB–KAN achieves the strongest overall results (Tab. 2). It delivers the lowest log-loss (0.074) and Brier score (0.024), while also reaching the highest ROC–AUC (0.997). Tree-based ensembles (CatBoost, XGBoost, LightGBM) and Random Forest remain competitive in discriminative power, but underperform in calibration. Compared to the neural MLP baseline, GB–KAN attains superior calibration and slightly higher ROC–AUC. These findings confirm that GB–KAN combines state-of-the-art predictive performance with reliable probability estimates.

Table 3 reports regression results. On the **Diabetes** dataset, GB–KAN achieves the lowest error (RMSE 53.84, MAE 42.87), while reaching an R^2 of 0.48. Linear baselines such as Ridge, ElasticNet,

Table 3: Regression benchmarks on Diabetes and California Housing. Best per dataset in **bold** and mean across 5 folds.

Model	RMSE ↓	MAE ↓	R^2 ↑
Diabetes			
GB-KAN	***53.84***	***42.87***	***0.48***
Ridge	54.83	44.24	0.48
ElasticNet	54.84	44.26	0.48
Lasso	54.85	44.27	0.48
CatBoost	56.38	45.57	0.45
Random Forest	57.59	46.77	0.43
LightGBM	57.84	46.29	0.42
GB (sklearn)	58.53	47.49	0.41
MLP	59.40	46.74	0.39
XGBoost	60.52	48.54	0.37
TabTransformer	67.68	53.50	0.21
California Housing			
XGBoost	**0.45**	**0.30**	**0.85**
LightGBM	0.48	0.32	0.83
CatBoost	0.49	0.33	0.82
GB-KAN	0.49	0.33	0.82
Random Forest	0.50	0.33	0.81
GB (sklearn)	0.52	0.36	0.80
MLP	0.53	0.35	0.79
TabTransformer	0.56	0.39	0.76
ElasticNet	0.73	0.53	0.60
Lasso	0.73	0.53	0.60
Ridge	0.73	0.53	0.60

and Lasso attain similar R^2 values (0.48), but with higher prediction error. Tree ensembles achieve R^2 values between 0.45 and 0.37, while TabTransformer lags behind at 0.21. This highlights GB–KAN as a strong contender that balances predictive accuracy with interpretability, though it trades off some explained variance compared to linear models.

On the large-scale **California Housing** dataset, gradient-boosted trees dominate. XGBoost achieves the best results (R^2=0.85, RMSE 0.45, MAE 0.30), followed closely by LightGBM (0.83). GB–KAN and CatBoost are statistically tied (R^2=0.82, RMSE 0.49, MAE 0.33), placing them just behind the top tree learners and ahead of Random Forest (0.81) and the standard GradientBoostingRegressor (0.80). Neural baselines perform somewhat worse, with MLP at R^2=0.79, while the TabTransformer reaches 0.76. Linear methods (Ridge, Lasso, ElasticNet) trail far behind at R^2=0.60, confirming the advantage of nonlinear models in this high-dimensional setting. Overall, GB–KAN scales competitively to large datasets, nearly matching CatBoost and narrowing the gap to XGBoost, while continuing to offer its interpretability advantages. On all datasets, GB-KAN training time is within a factor of 2.5–3× of XGBoost. This overhead is mainly due to spline evaluations but remains practical for the tabular scales considered.

4.5 Interpretability Evaluation

Beyond accuracy and calibration, we evaluate GB-KAN explanations along both *qualitative* and *quantitative* dimensions.

Qualitative Inspection. Stage-wise refinement plots ("boosting storyboards") visualize how successive stages refine feature effects and highlight emergent interactions. For tabular datasets (e.g., DIABETES), we extract per-feature *shape cards* $S_j(x_j)$ with 95% cross-validation bands. These reveal clinically interpretable regimes, e.g., flat risk below BMI 25, sharp increase between 30-35, and saturation thereafter. Local-to-global plots confirm that individual predictions decompose faithfully into these global curves.

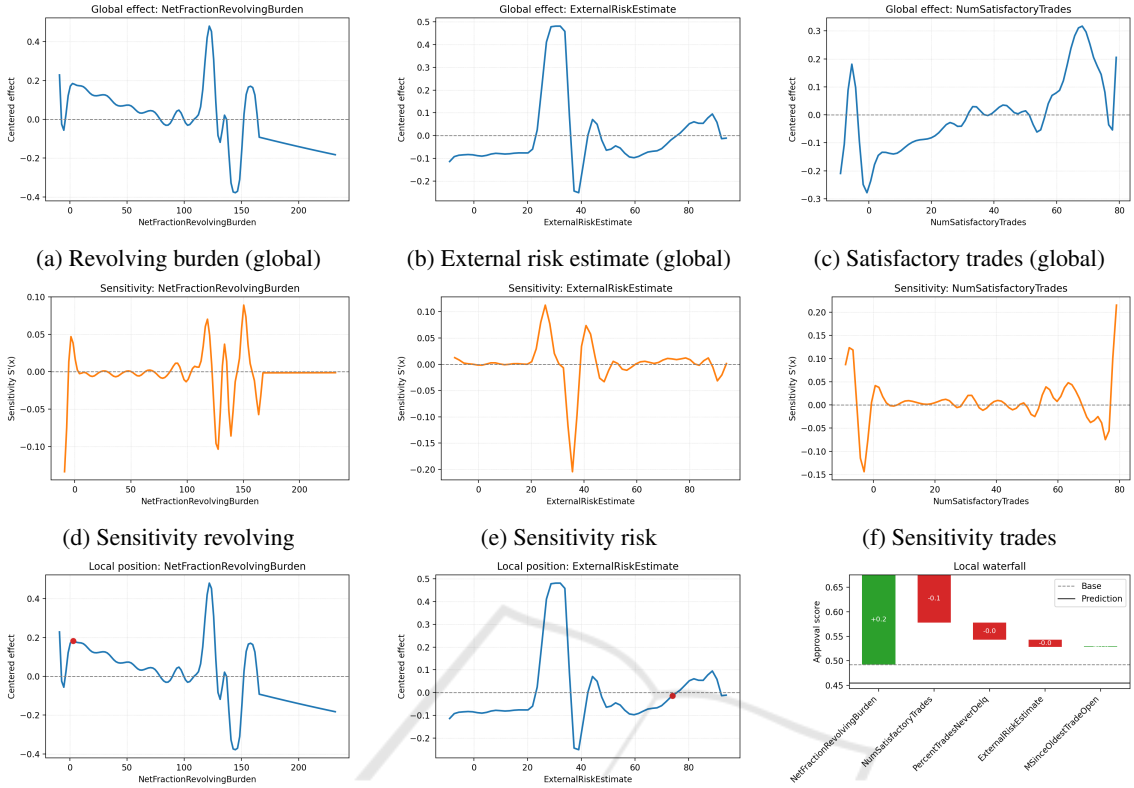
Quantitative Diagnostics. We report three families of metrics:

- **Complexity:** total variation of S'_j (lower = smoother).
- **Sparsity:** number of active features with $\|S_j\|_{2,\mathcal{G}}$ exceeding a 1% threshold of the maximum.
- **Stability:** mean Spearman correlation of discretized S_j curves across folds for the top-k features (k=10)..

Table 4 summarizes these metrics for GB-KAN compared with XGBoost+SHAP and MLP+Integrated Gradients. Across datasets, we find that GB-KAN achieves substantially lower sparsity than baselines (using only ∼7 out of 10 features on average for Breast Cancer) while maintaining high stability ($\tau \approx 0.92$). By contrast, XGBoost and MLP typically use all features (sparsity = 10) but show reduced stability across folds ($\tau < 0.8$). Complexity values are higher for GB-KAN, reflecting the expressive spline parameterization of KAN shape functions; however, this complexity remains structured and interpretable, unlike the opaque latent complexity of MLPs.

Table 4: Interpretability metrics. Lower complexity and higher stability are better.

Model	Complexity ↓ (TV of S'_j)	Sparsity ↓ (active feats)	Stability ↑ (cross-fold)
GB-KAN	*2018*	*7.2*	*0.92*
XGBoost+SHAP	*1466*	*10*	*0.75*
MLP+IG	*762*	*10*	*0.68*



(a) Revolving burden (global) (b) External risk estimate (global) (c) Satisfactory trades (global)

(d) Sensitivity revolving (e) Sensitivity risk (f) Sensitivity trades

(g) Local2global revolving (h) Local2global risk (i) Local waterfall

Figure 1: **Interpretability in the FICO dataset.** Row 1: global effects reveal domain-plausible patterns. Row 2: derivative sensitivities highlight unstable regions and monotonic segments. Row 3: local2global and waterfall plots provide applicant-level auditability and contextualization of predictions.

4.6 Interpretability Case Study: FICO Credit Risk

To further illustrate the interpretability advantages of GB–KAN, we analyze the FICO credit risk dataset, a benchmark widely studied in fairness-aware machine learning. Unlike our study on Concrete Strength, which emphasizes physically grounded rules in engineering, this example highlights interpretability in a high-stakes financial application.

Global effects. We focus on three representative features with the highest edge weight magnitudes: revolving burden, external risk estimate, and number of satisfactory trades Figure 1 (top row) shows the global shape functions for key features. We observe clear and domain-plausible patterns:

- **NetFractionRevolvingBurden (credit utilization):** low utilization yields a strong positive effect on approval, while high utilization ($>150\%$) is sharply penalized.
- **ExternalRiskEstimate (risk score proxy):** values around 20–30 lead to strongly positive contributions, while mid-range values ($\sim$40) are penalized.
- **NumSatisfactoryTrades:** additional satisfactory trades improve creditworthiness, but the effect saturates beyond $\sim$60 trades.

Sensitivity Analysis. The derivative plots (middle row of Figure 1) capture local sensitivities $S'_j(x)$, revealing where predictions are unstable. For example, around mid-range values of *ExternalRiskEstimate*, small changes can strongly alter predicted approval, reflecting regions of higher uncertainty. This motivates regularization or monotonicity constraints in sensitive domains.

Local-to-Global Explanations. The bottom row of Figure 1 shows local2global plots, situating an individual applicant (red dot) within the global effect curve. This highlights, for instance, how a low revolving burden can compensate for fewer satisfactory

trades. Such explanations allow credit officers or applicants to understand why a prediction was made in context of the global model behavior.

Local Decomposition. Finally, the waterfall plot (bottom right of Figure 1) provides a full decomposition of a single decision: low revolving burden contributes $+0.2$ to approval, while a limited number of satisfactory trades and moderate delinquency history subtract ~ 0.1 each. Such additive explanations make the decision process directly auditable.

Derived Decision Rules. From these interpretable objects, GB–KAN allows practitioners to extract transparent rules that are both human-readable and predictive:

> *Applicants with low revolving utilization (<50%) and more than ∼20 satisfactory trades are consistently favored. High utilization (>150%) or mid-range external risk scores (∼40) strongly reduce approval likelihood. Longer credit history (>600 months) adds further positive contributions, but with weaker magnitude.*

Such rules are not explicitly encoded in the data but emerge naturally from the smooth, spline-based structure of GB–KAN. Importantly, these statements are both *global* (model-wide trends) and *local* (applicant-specific), which is essential for high-stakes decisions in regulated domains.

4.7 Interpretability on the Cement Dataset

We next turn to the concrete strength dataset, where the same tools provide scientifically coherent explanations in an engineering domain. We examine three complementary families of interpretability plots. Together, they provide a multi-perspective account of how features contribute to predictions.

Global Shape Functions. On Concrete Strength, shapes for cement, water, and age (Figure 2a–c) align with engineering knowledge: strength increases with cement, decreases with excessive water, and saturates with curing age. Simple parametric surrogates (e.g., saturating exponentials for age, concave quadratics for water) approximate these shapes and match known empirical laws like Abrams' law. show representative feature effects for `cement`, `water`, and `age`. These curves are smooth and semantically meaningful. Cement has a strong monotone positive effect, consistent with the known role of cement content in increasing compressive strength. These approximations confirm that GB–KAN not only recovers domain-consistent trends (Abrams' law) (Yeh, 2006; Benaicha et al., 2019) but also aligns with well-established empirical formulas from material science.

Local Sensitivity. First-derivative plots (Figure 2d–f) highlight where predictions are most sensitive to small changes in input. For instance, the sharp initial sensitivity to curing age underscores the importance of early hydration processes, while cement remains consistently influential across its range. These derivative views add nuance by revealing where in feature space small perturbations matter most. Note that the apparent differences in the derivative plots across features (e.g., water, age, cement) are primarily due to their different scales. The vertical axes are not normalized.

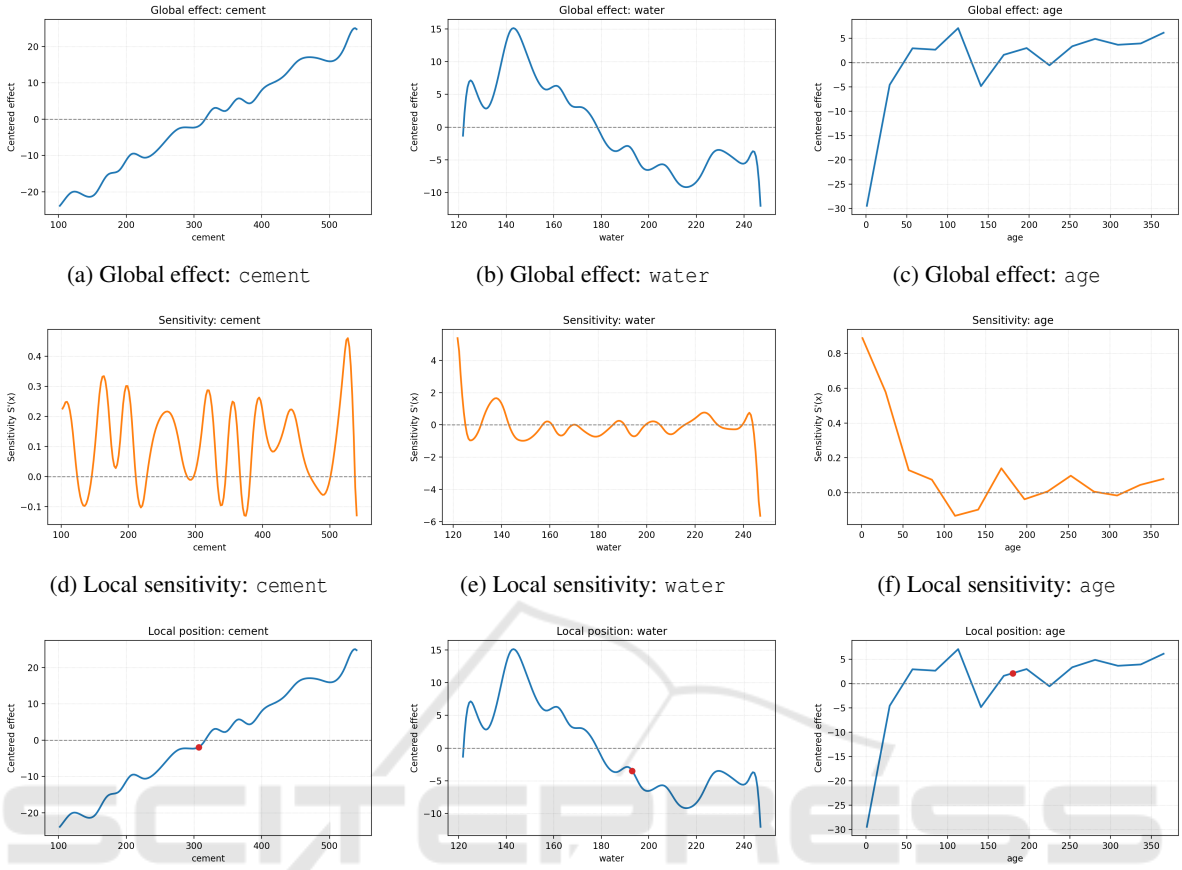
Local-to-Global Alignment. To test coherence across scales, we visualize local positions within global curves (Figure 2g–i). For an individual instance, the red markers illustrate how its prediction decomposes along the same trends observed globally. For example, a sample with high water content is placed on the steeply negative region of the water effect curve, clearly explaining the lowered prediction. This consistency across local and global explanations is a key advantage over post-hoc methods, which often lack such alignment.

Finally, we analyze model internals. The waterfall plot in Figure 3 decomposes a single prediction into signed feature contributions, illustrating how positive contributions from `age` and `superplasticizer` partially offset negative effects of `slag` and `water` and `cement` leading to a prediction lower then the base.

Rule Extraction. From regime structure in S_J we can extract simple rules (e.g., cement up $\rightarrow$ strength up; water $> 125 \rightarrow$ strength down; age up to 100 days $\rightarrow$ strength up then plateau), providing concise summaries for domain experts. These rules correspond closely to engineering intuition: strength increases steadily with higher cement content, decreases with excess water, and rises sharply with curing age before plateauing. Such extracted rules provide an immediately usable description of model behavior that is accessible to domain experts.

4.8 Ablation Studies

On Breast Cancer, GB–KAN is sensitive to learning rate and number of boosting rounds, which substantially reduce log-loss, whereas varying hidden



(a) Global effect: `cement` (b) Global effect: `water` (c) Global effect: `age`

(d) Local sensitivity: `cement` (e) Local sensitivity: `water` (f) Local sensitivity: `age`

(g) Local-to-global: `cement` (h) Local-to-global: `water` (i) Local-to-global: `age`

Figure 2: **Interpretability for the concrete strength dataset.** Columns correspond to features (`cement`, `water`, `age`); rows correspond to interpretability views. (a-c) Global effect functions show domain-consistent nonlinearities. (d-f) Local sensitivities (first derivatives) highlight where predictions are most reactive. (g-i) Local-to-global alignment plots place individual instances (red markers) on the global curves.

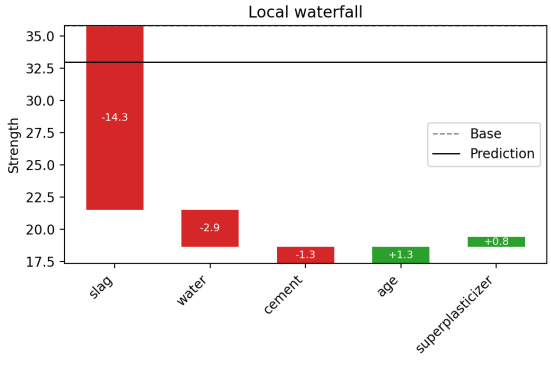


Figure 3: Local waterfall decomposition of a single prediction for the cement dataset.

width and spline knots has negligible effect (Fig. 4). ROC–AUC remains saturated across all settings, suggesting that modest-capacity base learners suffice when boosting is well tuned.

5 CONCLUSION

We introduced GB–KANs, which combine gradient boosting with interpretable KAN base learners to produce smooth per-feature shape functions and competitive, well-calibrated performance on tabular benchmarks. Across real-world tabular datasets, GB-KANs achieve predictive performance competitive with state-of-the-art boosting libraries while providing better-calibrated probability estimates. Importantly, the model exposes human-readable shape functions at each stage, enabling one to trace how feature effects evolve through boosting rounds and to detect emergent interactions.

Limitations and Future Work. Future work includes multi-output extensions, deeper KANs, second-order boosting, and explicit interaction dis-

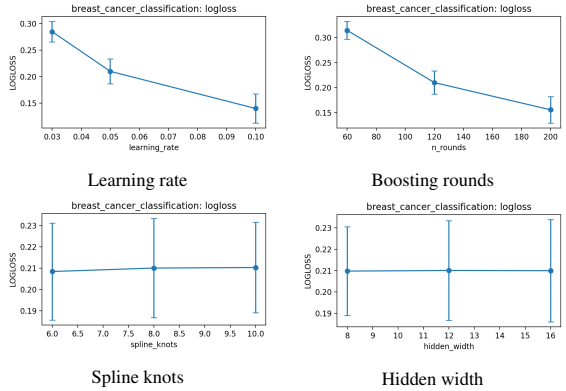


Figure 4: Ablation study on *Breast Cancer*: effect of key factors on log-loss (lower is better). Learning rate and boosting rounds strongly reduce loss, while architectural parameters (knots, width) have little effect. ROC–AUC remains saturated at ≈ 0.997 across all settings (not shown).

covery to further improve high-dimensional performance while preserving interpretability. Overall, GB-KANs provide a new accuracy-interpretability trade-off for tabular learning: they retain the reliability of boosting while exposing the internal structure of the learned function in a human-readable form. In practice, GB-KANs are particularly attractive when tree-based boosting is deemed too opaque, but plain additive or linear models underfit. In those cases, GB-KAN offers a middle ground: performance close to XGBoost/LightGBM on large tabular tasks, interpretability on par with GAM-style methods, and a computational footprint comparable to other boosting libraries.

ACKNOWLEDGMENTS

This work was funded by the Federal Ministry of Research, Technology and Space under 01IS24034B.

REFERENCES

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. (2021). Neural additive models: interpretable machine learning with neural nets. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Arik, S. O. and Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 6679–6687.

Arnold, V. I. (1957). On functions of three variables. *Doklady Akademii Nauk SSSR*, 114(4):679–681.

Barašin, I., Bertalanič, B., Mohorčič, M., and Fortuna, C. (2024). Exploring kolmogorov-arnold networks for interpretable time series classification.

Benaicha, M., Hafidi Alaoui, A., Jalbaud, O., and Burtschell, Y. (2019). Dosage effect of superplasticizer on self-compacting concrete: correlation between rheology and strength. *Journal of Materials Research and Technology*, 8(2):2063–2069.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Cai, Y., Ma, Y., Dong, Y., and Yang, H. (2023). Extrapolated random tree for regression. In *Proc. 40th Int'l Conf. Machine Learning (ICML)*, volume 202.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730.

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., and Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk.

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

Cheng, Z., Li, Y., Wang, Y., Wang, M., and Ji, R. (2022). Danet: Deep attentive network for tabular data. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 6437–6444.

Cruz, G. G., Renczes, B., Runacres, M. C., and Decuyper, J. (2025). State-space kolmogorov-arnold networks for interpretable nonlinear system identification. *IEEE Control Systems Letters*, pages 1–6.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. S. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. abs/2012.06678.

Kamienny, P., d'Ascoli, S., Lample, G., and Charton, F. (2022). End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Ke, G., Xu, Z., Zhang, J., Bian, J., and Liu, T.-Y. (2019). Deepgbm: A deep learning framework distilled by gbdt for online prediction tasks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 384–394.

Kelley Pace, R. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Kolmogorov, A. N. (1957). On the representation of continuous functions of several variables as superpositions of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, 114:953–956.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. (2024). Kan: Kolmogorov–arnold networks.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4765–4774.

McElfresh, D., Khandagale, S., Valverde, J., C., V. P., Ramakrishnan, G., Goldblum, M., and White, C. (2023). When do neural nets outperform boosted trees on tabular data? In *Proc. of the 37th International Conference on Neural Information Processing Systems*.

Mohr, J., Tousside, B., Schmidt, M., and Frochte, J. (2023). Multiple additive neural networks: A novel approach to continuous learning in regression and classification. In *15th International Conference on Neural Computation Theory and Applications*.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Nori, H., Jenkins, S., Koch, P., and Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability.

Popov, S., Morozov, S., and Babenko, A. (2019). Neural oblivious decision ensembles for deep learning on tabular data. In *International Conference on Learning Representations (ICLR)*.

Posocco, N. and Bonnefoy, A. (2021). Estimating expected calibration errors. In *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part IV*, page 139–150, Berlin, Heidelberg. Springer-Verlag.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ross, A., Marasović, A., Bhagavatula, C., and Choi, Y. (2022). Discovering latent concepts learned in bert. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13243–13257.

Rudin, C. (2018). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206 – 215.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bansal, Y., and Goldstein, T. (2021). Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. In *Neural Information Processing Systems*.

Yang, H., Zhang, X., and Sudjianto, A. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4699–4711.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28:1797–1808.

Yeh, I.-C. (2006). Generalization of strength versus water–cementitious ratio relationship to age. *Cement and Concrete Research*, 36(10):1865–1873.