Shift-Invariant Attribute Scoring for Kolmogorov-Arnold Networks via Shapley Value

ICML 2026 Mechanistic Interpretability Workshop *Virtual Poster*

Wangxuan Fan*[1], Ching Wang*[1], Siqi Li[2], Nan Liu[2]

[1]National University of Singapore [2]Duke-NUS Medical School, Singapore

Equal contribution

Take-home Message

ShapKAN uses *Shapley-value* neuron attribution to prune Kolmogorov-Arnold Networks (*KANs*) in a shift-invariant way, producing more stable importance scores, better symbolic recovery, and stronger generalization than vanilla KAN pruning.

1 Motivation

- KANs are promising because spline-based edge activations support compact symbolic representations.
- However, pruning KANs is challenging because importance is attached to functions and neurons, not simple scalarweights.
- Vanilla magnitude-based pruning is sensitive to input coordinate shifts, causing unstable neuron rankings.
- We need a pruning score that reflects actual contribution and remains stable under domain shift.

2 Core Idea

- Model each KAN layer as a cooperative game: neurons are players and subsets of neurons form coalitions.
- Use Shapley value to quantify each neuron's contribution fairly and robustly.
- Compute exact Shapley values for small layers and permutation-sampling estimates for wider layers.
- Use a bottom-up multi-layer pruning strategy to remove low-importance neurons layer by layer.

$$x_{l+1,j} = \sum_{i \in S_l} \phi_{l,j,i}(x_{l,i})$$

$$\mathrm{Shap}(x_{l,i}) \approx \frac{1}{m} \sum_{t=1}^{m} \left(v(S_{L_l}^{(t)} \cup \{x_{l,i}\}) - v(S_{L_l}^{(t)}) \right),$$

Flexible pruning criteria

ratio pruning • number pruning • threshold pruning

3 Why ShapKAN?

- **Shift-invariant attribution under covariate shift;**
- **Principled importance scoring via Shapley axioms;**
- **Preserves interpretability while enabling effective compression.**

5 Contributions

1. A Shapley-based explainable pruning framework for KANs.
2. Efficient approximation strategies for scalable multi-layer pruning.
3. Extensive evidence on synthetic and real-world datasets showing robustness, generalization, and interpretability gains.

6 Conclusion

- ShapKAN provides stable, shift-robust neuron attribution for KAN pruning.
- It improves both prediction-related generalization and non-prediction interpretability tasks.
- The framework is practical, scalable, and readily applicable to future KAN variants.

4 Key Experimental Findings

(a) Shifted Domain Sensitivity: Vanilla KAN vs. ShapKAN

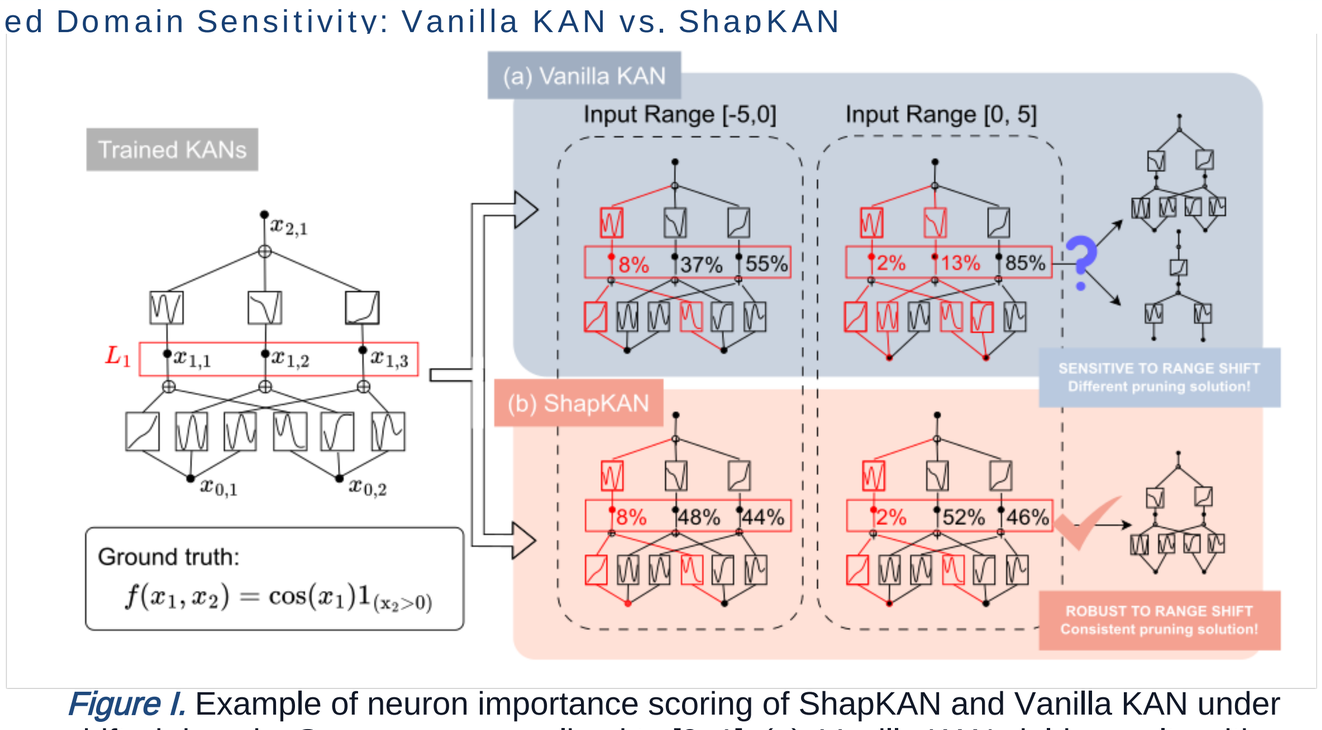

Figure I. Example of neuron importance scoring of ShapKAN and Vanilla KAN under shifted domain. Scores are normalized to [0, 1]. (a): Vanilla KAN yields results with significant fluctuation. (b):ShapKAN provides more robust scoring consequence and consistent pruning solution.

(b) Cross-domain Robustness of Neuron Rankings

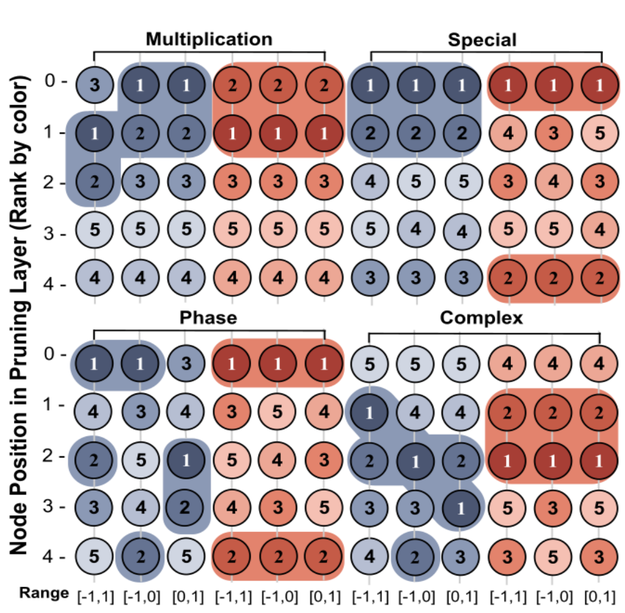

Figure 2. Cross-domain robustness comparison of neuron importance ranking betwcen Vanilla KAN and ShapKAN. Each circle represents a node's importance rank evaluated by scoring methods. Shadow indicates the most substantial nodes. Numeric record of mean, standard deviations and pruning set are reported in the paper.

(c) Symbolic Recovery on Multiplication Task

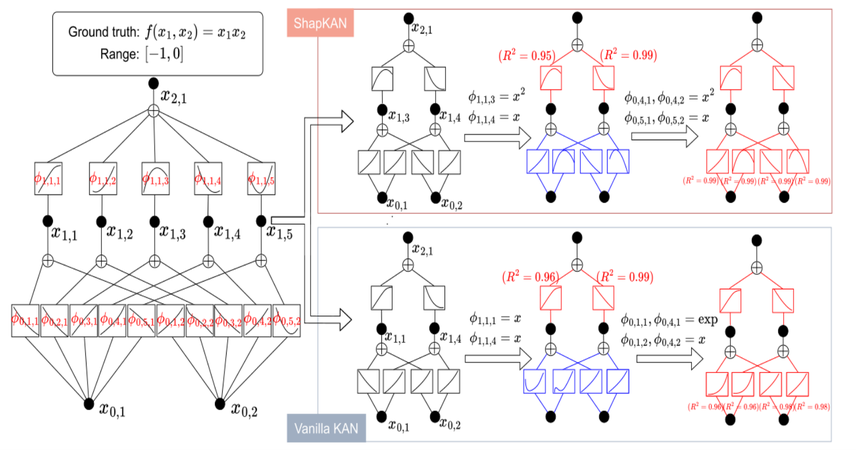

Figure 3. Symbolic recovery on multiplication task (f(x_1, x _2) = x_1 · x_2. ShapKAN accurately recovers the ground -truth symbolic structure,while Vanilla KAN recovers suboptimal forms.

(d) Convergence of Permutation Sampling

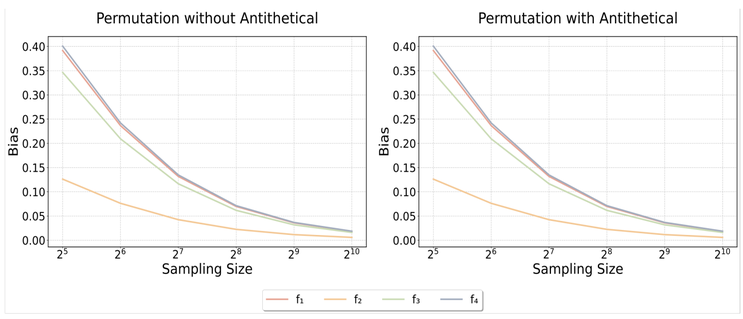

Figure 4. Convergence analysis of permutation sampling methods on simulated datasets.
Left: Standard permutation sampling.
Right: Antithetical permutation sampling.

(e) Generalization on Real-world Datasets

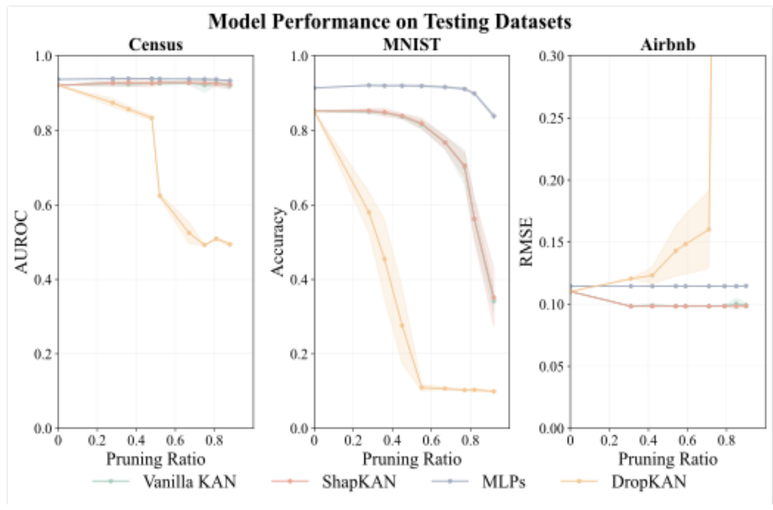

Figure 5. Comparison of generalization capacity. Mean and standard deviation are reported based on 10 times independent experiments. For accuracy and Area under the Receiver Operating Characteristic Curve (AUROC), higher are better; for RMSE, lower is better. In Airbnb dataset, DropKAN's RMSE significantly exceeds 0.3 when the ratio is above 0.6.

NUS National University of Singapore · DukeNUS Medical School · Digital Medicine Lab, Duke-NUS Medical School, National University of Singapore

Code & Data

github.com/YuanJrShiuan/KAN_Pruning_Shapley