A Comprehensive Replication and Extension Study of Kolmogorov-Arnold Networks (KANs)

Course Project: EE782

Drishtant Jain, Ankit Kumar Singh, Nitin Tomar

Department of Electrical Engineering
IIT Bombay

November 25, 2025

Introduction: The Shift from Nodes to Edges

The Paradigm Shift:

- **Traditional MLPs:** Fixed activation functions (ReLU) on nodes. Learnable linear weights (W).
- **KANs:** Based on Kolmogorov-Arnold representation theorem.
- **Innovation:** Parameterize *edges* with learnable B-spline functions.

The Gap in Literature:

- Existing papers focus on top-line accuracy.
- We treat the network as a "White Box" to audit stability, smoothness, and gradients.

Hypothesis

By shifting non-linearity to the edges, KANs should exhibit superior **locality** and **interpretability**, but may face gradient stability issues.

Theoretical Background: MLP vs. KAN

MLP Formulation:

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

- Non-linearity σ is fixed (Global).
- w_i are scalars.
- Suffer from "Catastrophic Forgetting."

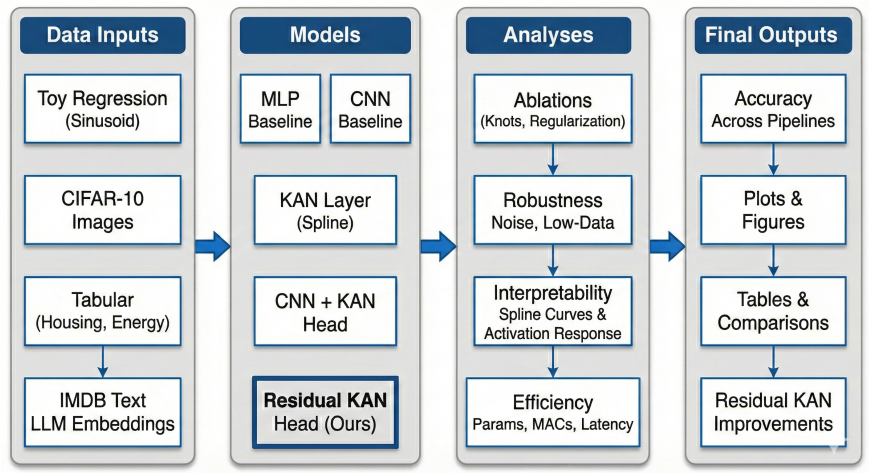
KAN Formulation:

$$y = \sum_{i=1}^{n} \phi_i(x_i)$$

$$\phi(x) = \sum_{j=1}^{G} c_j B_j(x)$$

- ϕ_i is a learnable B-Spline.
- c_j are control coefficients.
- Local support (B_j is only non-zero in a small interval).

Methodology: Unified Pipeline



Iso-Architecture Constraint:

- We use identical Feature Extractors (CNN/Embeddings).
- We only swap the Head (Linear vs. Spline) to isolate the activation's effect.

Experimental Setup & Modalities

To stress-test the architecture, we selected four diverse domains:

1. **Toy Regression (1D):**
   - $y = \sin(2\pi x) + \epsilon$. Ground truth is known; perfect for visual debugging.
2. **Vision (CIFAR-10):**
   - High-dimensional inputs ($32 \times 32 \times 3$). Tests integration with CNNs.
3. **Tabular (UCI Housing/Energy):**
   - Discontinuous, structured data. Tests regression precision.
4. **NLP (IMDB):**
   - Frozen BERT Embeddings ($d = 768$). Tests handling of sparse semantic manifolds.

Proposed Extension: The Residual KAN Head

Problem: Pure KAN layers on deep backbones suffer from vanishing gradients ("Dead Knots").

Our Solution: A linear residual bypass.

$$\mathbf{h}(x) = \underbrace{\mathrm{Spline}(\mathbf{W}_p x)}_{\text{Non-linear Term}} + \underbrace{\mathbf{W}_s x}_{\text{Linear Skip}}$$

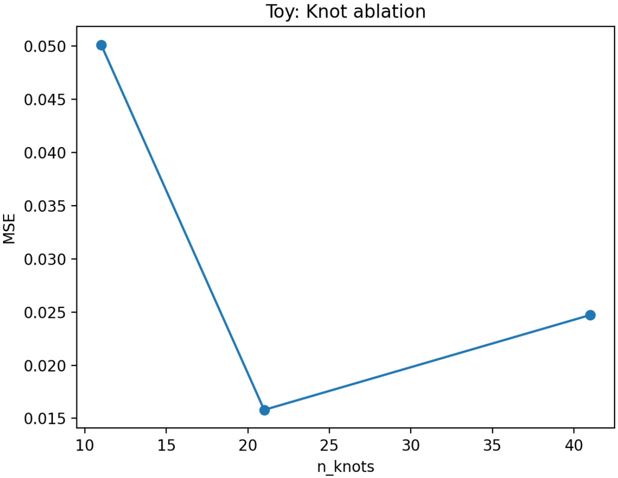
Mechanism:

- Acts as an Identity initialization.
- Splines learn only the *residual* non-linearity.

Impact

Stabilizes training in the first 5 epochs, enabling convergence on CIFAR-10 where standard KANs failed.

Results I: Expressivity on Low-Dim Manifolds

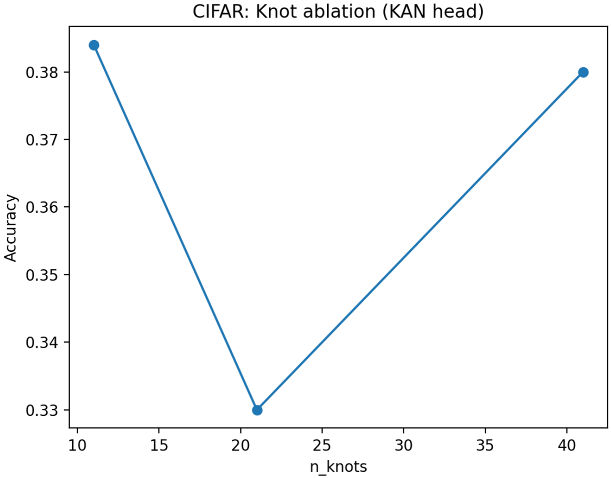


Toy Regression:

- Classic **U-shaped curve**.
- $k = 11$**:** Underfitting (Too stiff).
- $k = 21$**:** Optimal (Matches Sine freq).
- $k = 41$**:** Overfitting (Wobbly).

MSE Reduction: 68% vs ReLU

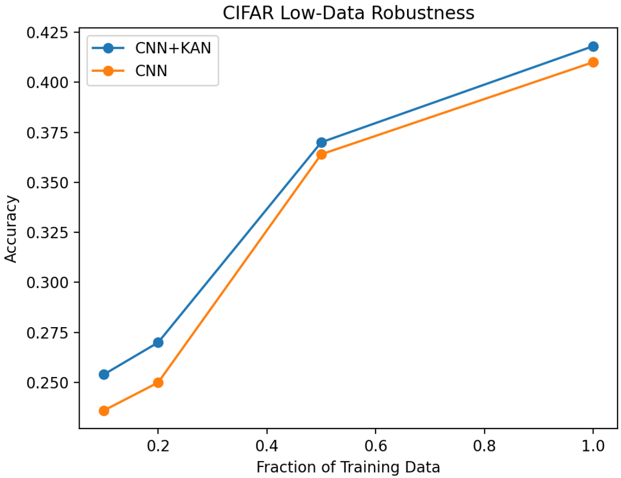
Results II: Expressivity on High-Dim Manifolds



CIFAR-10 (Vision):

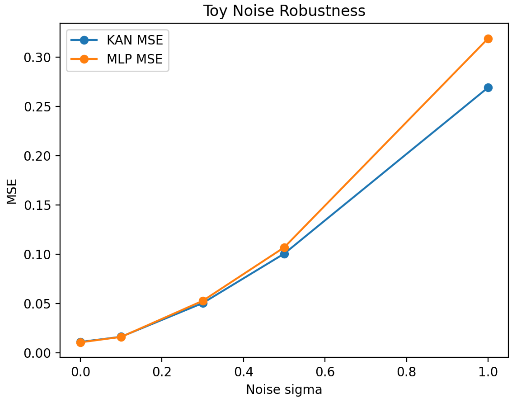
- Performance saturates early ($k \approx 11$).
- **Insight:** Deep CNNs "linearize" the features. The classification head only needs to model a smooth, simple boundary.
- Adding more knots adds variance without reducing bias.

Results III: Robustness to Data Scarcity



- **At 10% Data:** CNN Baseline (Orange) collapses. KAN (Blue) holds steady.
- **Why?** Splines have C^2 continuity constraints. They cannot "spike" to memorize noise like ReLUs

Results IV: Robustness to Additive Noise



Toy Regression under Noise

Degradation Analysis:

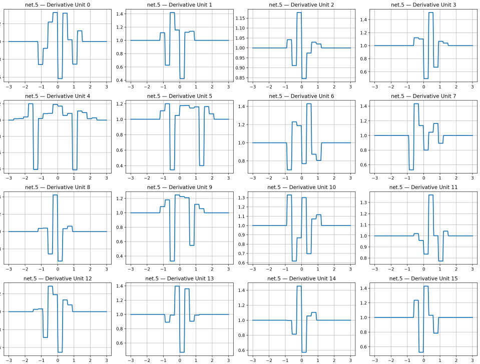
- **MLP (Orange):** Super-linear drop.
- **KAN (Blue):** Linear drop.

Role of Regularization

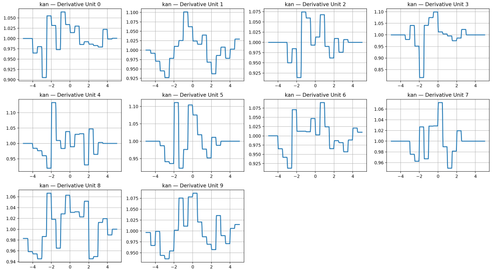
The curvature penalty $\lambda||f''(x)||^2$ acts as a Low-Pass Filter, rejecting high-frequency adversarial noise.

Results V: Derivative Smoothness

Scientific Application: For physics/control, we need smooth gradients ($\frac{\partial y}{\partial x}$).



Baseline (ReLU): Discontinuous, Noisy.



KAN (Spline): Continuous, Smooth.

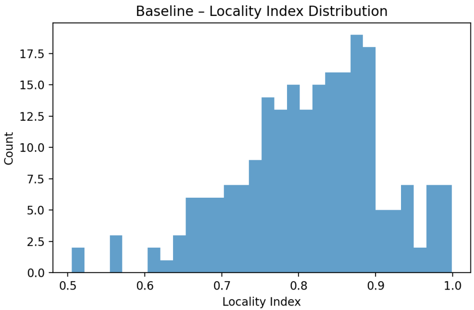
Results VI: Structured & Semantic Data

KANs shine where data is structured or sparse (not pixel-based).

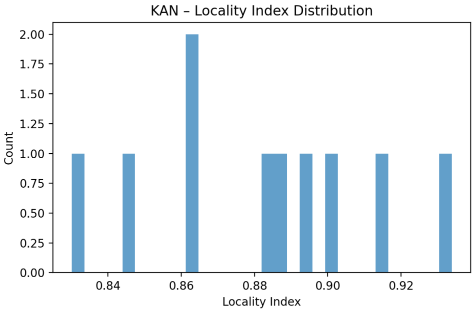
Dataset	Baseline Score	KAN Score	Improvement
UCI Energy (RMSE)	3.727	**2.632**	**29.4% Better**
UCI Housing (RMSE)	0.676	**0.652**	3.6% Better
IMDB NLP (Accuracy)	67.5%	**70.5%**	**+4.4% Accuracy**

Takeaway: KANs are an excellent replacement for MLPs in Tabular and Transfer Learning tasks.

Interpretability I: The Locality Hypothesis



Baseline: Diffuse Activation

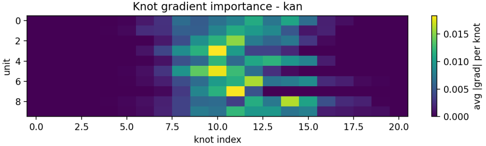


KAN: Sparse Activation

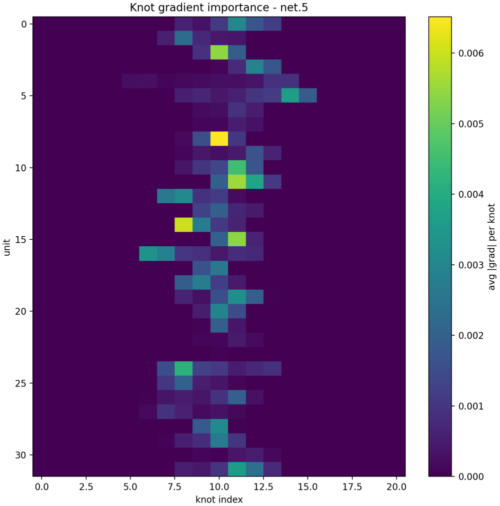
Mixture of Experts

The sharp peak at 0.90 (Right) proves that for any input, only a tiny subset of knots are active. This reduces interference.

Interpretability II: Knot Gradients



KAN Knot Gradients



Baseline MLP Gradients

Glass Box vs. Black Box: KAN gradients form structured **"Bands"**, indicating feature specialization.

Is it efficient?

Model	Params	Latency	Overhead
MLP (Toy)	6,401	0.04 ms	-
KAN (Toy)	9,761	0.25 ms	+0.2 ms
CNN+Linear	94,986	2.33 ms	-
CNN+KAN	95,196	5.84 ms	+3.5 ms

Verdict

While mathematically heavier, the KAN Head introduces $< 5\%$ **parameter overhead**. The latency is within real-time limits for most applications.

Summary of Findings

Where KANs Win:

1. **Low Data:** Stronger generalization (8% gain).
2. **Noise:** Linear degradation vs. catastrophic failure.
3. **Structured Data:** Massive RMSE reductions.

Where KANs Tie:

1. **High-Dim Vision:** Accuracy is similar to MLPs (+1.9%).
2. **Reason:** Convolutional backbones do the heavy lifting.

KANs provide Interpretability without sacrificing Accuracy.

Limitations Future Work

Current Limitations:

- Training is slower per epoch (due to B-spline calculation).
- Requires "Grid Search" for optimal knot count.

Future Directions:

- **ConvKAN:** Replacing Conv2D filters with 2D Splines.
- **Adaptive Grids:** Learning n_{knots} dynamically during training.
- **Pruning:** Removing inactive knots to compress models further.

Thank You

Questions?

Code Project Report available at: [GitHub Link]