

# From Theory to Practice: Evaluating and Enhancing Kolmogorov-Arnold Networks (KAN) Robustness Under Adversarial Conditions

Evgenii Ostanin  
Toronto Metropolitan University  
Toronto, Canada  
eostanin@torontomu.ca

Nebojsa Djosic  
Toronto Metropolitan University  
Toronto, Canada  
nebojsa.djosic@torontomu.ca

Fatima Hussain  
Toronto Metropolitan University  
Toronto, Canada  
fatima.hussain@torontomu.ca

Salah Sharieh  
Toronto Metropolitan University  
Toronto, Canada  
salah.sharieh@torontomu.ca

Alexander Ferworn  
Toronto Metropolitan University  
Toronto, Canada  
aferworn@torontomu.ca

**Abstract**—Kolmogorov–Arnold Networks have emerged as promising architectures thanks to their adaptive activation functions and enhanced interpretability. However, their robustness under adversarial conditions remains underexplored. In this study, we evaluated four variants of Kolmogorov–Arnold Networks, Linear, Fourier, Jacobi, and Chebyshev against Gaussian noise and two gradient-based attacks (the Fast Gradient Sign Method and Projected Gradient Descent). Through detailed comparative analyses and adversarial training experiments with varying mixes of perturbed data, we reveal substantial differences in resilience across variants and relative to a multilayer perceptron baseline. Our results show that targeted adversarial training materially improves robustness under strong adversarial attacks. In particular, including only 5% Fast Gradient Sign Method examples and 5% Projected Gradient Descent examples in the training set restores between 60 and 90 percentage points of accuracy against these attacks. These findings clarify the factors influencing Kolmogorov–Arnold Network robustness and validate adversarial training as a practical hardening strategy for deployment in adversarially challenging environments.

**Keywords**—Kolmogorov–Arnold Networks; KAN; MNIST; FGSM; PGD; Classification; Adversarial Training.

## I. INTRODUCTION

The rapid advancement of Machine Learning (ML) has led to increasingly sophisticated models that perform well across a variety of tasks. Among these developments, Kolmogorov–Arnold Networks (KANs) represent a novel approach based on the Kolmogorov–Arnold representation theorem. KANs enhance interpretability and flexibility through learnable activation functions, dynamically adapting to data variations and potentially improving model robustness and generalization. Their robustness, however, especially under Adversarial Attacks (AA) and noisy data, remains an underexplored domain.

This paper extends our previous work [1], which analyzed the robustness of KAN architectures under AA. In that study, the focus was on evaluating the performance of different KAN implementations against Gaussian noise, Fast Gradient Sign Method (FGSM), and Projected Gradient Descent (PGD) attacks, comparing their vulnerabilities to a Multi-Layer Perceptron (MLP) classifier. Our findings showed that while KANs achieved higher accuracy than MLPs in clean

environments, they exhibited significant drops in accuracy when subjected to adversarial perturbations, with PGD having the most severe impact.

Traditional MLPs often struggle with capturing complex nonlinear relationships due to their reliance on fixed activation functions and linear weight matrices. This limitation can lead to suboptimal generalization in adversarial settings or when handling noisy data. To address these challenges, KANs introduce learnable activation functions on edges, allowing them to adapt dynamically to input variations, offering potential advantages in robustness and interpretability over traditional models [2].

The increasing sophistication of AA poses significant challenges for deep learning models, particularly in security-critical applications such as autonomous systems and cybersecurity. Attacks like the FGSM and PGD exploit weaknesses in models by introducing subtle alterations to input data. Additionally, the growing deployment of ML models in real-world applications exposes them to environmental noise, which can further degrade performance [3]–[5]. As a result, robustness against both AA and noise is an important requirement for deploying ML models in production and practical settings [6].

This extended paper expands our prior findings [1], [7] by systematically evaluating adversarial training as a novel approach to enhance the robustness of multiple KAN architectures. Specifically, we evaluate how different adversarial training compositions impact KAN resilience to AA. Our primary contributions include:

- A reassessment of the vulnerabilities of KAN architectures under adversarial conditions.
- Analyzing the impact of adversarial training with varying proportions of clean and adversarially perturbed samples.
- A comparative analysis of how different KAN models respond to adversarial training, highlighting the strengths and weaknesses of each approach.
- A discussion of the broader implications of KAN robustness and future research directions.

**Key Results:** Unprotected KAN models can lose up to 88% accuracy under strong PGD attacks. Injecting just 5%

adversarial samples per AA into the training set restores 60 to 90 percentage points of robustness against FGSM and PGD across all KAN variants. However, the Fourier KAN remains highly sensitive to Gaussian noise. Its noise accuracy stays below 20% even after adversarial training. These findings underscore the need for variant-specific hardening strategies.

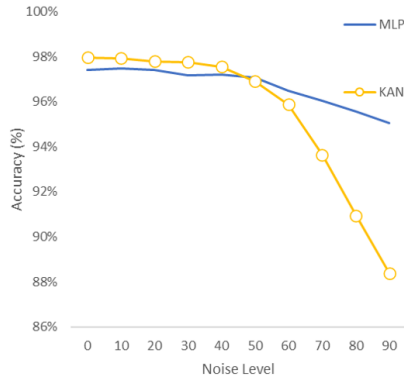


Figure 1. Model Accuracy Degradation After Noise Attack.

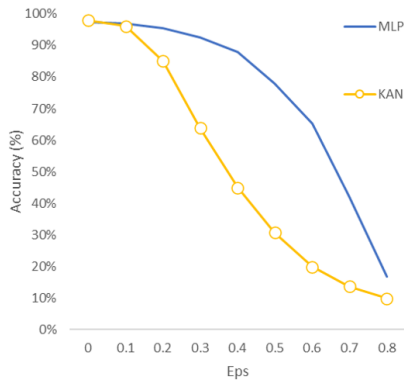


Figure 2. Model Accuracy Comparison After FGSM Attack.

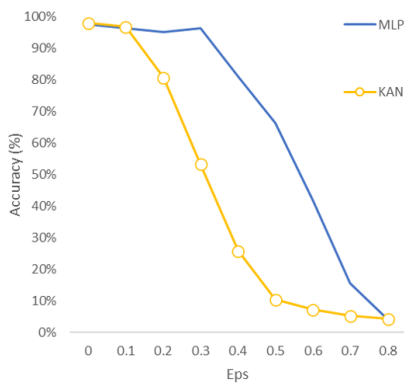


Figure 3. Model Accuracy Comparison After PGD Attack.

To support our initial findings, we include a series of visualizations. Figure 1 illustrates the accuracy degradation of MLPs and KANs under varying levels of Gaussian noise. Figure 2 shows accuracy degradation under increasing FGSM attack strength (eps.), highlighting KANs' greater sensitivity compared to MLPs. Finally, Figure 3 presents accuracy degradation under PGD attacks, where KANs demonstrate the most significant performance drop.

**Paper Structure:** The remainder of this paper is organized as follows: Section II reviews related work, including research on adversarial robustness and KAN applications. This section provides an overview of prior studies on KANs and AAs, positioning our work within the existing body of research. We discuss advancements in adversarial training techniques and their effectiveness in improving model resilience.

Section III details the methodology, including dataset preparation, attack methodologies, and adversarial training strategies. We describe the experimental setup, including the architecture of the tested KAN models, the parameters used for adversarial training, and the generation of adversarial examples using FGSM and PGD. This section also explains how different compositions of training data impact model robustness.

Section IV presents experimental results, evaluating the impact of adversarial training on model robustness. We provide a comparative analysis of the tested models under various adversarial conditions, supported by visualizations and performance metrics. This section highlights key trends observed across different KAN architectures and discusses the significance of adversarial training in mitigating accuracy degradation.

Sections V and VI conclude with a discussion of key findings and future research directions. We summarize the major contributions of this work, analyze the broader implications for secure ML applications, and propose areas for future exploration, including testing on more complex datasets and refining adversarial training techniques for enhanced KAN resilience.

## II. RELATED WORK

The robustness of ML models under adversarial conditions is critical for ensuring their reliability in real-world deployments, particularly in safety-critical applications. While traditional Neural Network (NN) architectures like MLPs have been extensively studied for their vulnerability to adversarial perturbations, KANs, with their unique architecture based on learnable activation functions, presents new opportunities and challenges in terms of robustness. This section provides an overview of foundational concepts and prior research related to KANs architectures, the underlying Kolmogorov-Arnold representation theorem, AAs, and adversarial training strategies, highlighting key insights and existing gaps in the literature.

### A. Kolmogorov-Arnold Representation Theorem

KANs represent a novel NN architecture derived from the Kolmogorov-Arnold representation theorem, providing a compelling alternative to traditional MLPs. Figure 4 from

[2] clearly illustrates the architectural differences between traditional MLP and KAN models. This innovative architecture fundamentally changes the traditional NN paradigm by introducing learnable activation functions along network edges, replacing the conventional fixed activation functions applied at nodes. The learnability of these functions allows for greater flexibility and interpretability, enabling the KAN models to dynamically adapt their internal transformations during training, potentially resulting in improved model generalizations, and adaptability to diverse and complex datasets.

The foundational basis of KAN architectures lies in the Kolmogorov-Arnold Representation Theorem, first introduced by Andrey Kolmogorov in 1957 and later refined by Vladimir Arnold in 1963. Commonly referred to as the superposition theorem, it mathematically states that any continuous multivariate function  $f(x_1, \dots, x_n)$  defined within a bounded domain can be represented as a superposition of continuous univariate functions. Formally, the theorem is expressed as follows:

$$f(x) = f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left( \sum_{p=1}^n \phi_{q,p}(x_p) \right) \quad (1)$$

In (1)  $\phi_{q,p} : [0, 1] \rightarrow \mathbb{R}$  are continuous inner functions, and  $\Phi_q : \mathbb{R} \rightarrow \mathbb{R}$  represent continuous outer functions.

KAN models leverage this theorem by explicitly learning these univariate functions, typically using spline-based methods due to their computational efficiency, smoothness properties, and interpretability.

However, alternative activation functions beyond splines exist and may offer advantages depending on specific applications [8], [9]. Fourier-based activation functions, such as those employed in Naïve Fourier KAN [8], effectively handle periodic data and signals due to their inherent periodic properties. Polynomial-based activations, such as those used in Jacobi KAN and Chebyshev KAN [9], can provide computational simplicity while offering superior approximation capabilities in scenarios requiring less flexibility or complexity. Chebyshev polynomials, in particular, are notable for their numerical stability and efficient approximation characteristics for certain classes of functions [9].

Ultimately, identifying the optimal activation function involves balancing computational efficiency, robustness to adversarial perturbations, and task-specific performance requirements. This critical consideration, along with practical implications and empirical evaluations under adversarial scenarios, is addressed thoroughly in the experiments and results presented in later sections of this paper.

### B. Potentials and Limitations of KANs

KANs have been proposed as an innovative NN architecture offering unique advantages in interpretability and computational efficiency. Several studies have investigated their performance across various tasks, especially in computer vision. For instance, [10] evaluated KANs against established

architectures such as MLP-Mixer, Convolutional Neural Networks (CNNs), and Vision Transformers (ViTs) on widely-used benchmarks. The study highlighted that KAN models notably outperformed MLP-Mixer on datasets like CIFAR-10 and CIFAR-100, demonstrating the model's potential for achieving competitive accuracy. However, the same research observed that KAN architectures fell short when compared directly with deeper convolution-based models, specifically ResNet-18. Still, the computational efficiency advantage was evident, indicating that KANs could offer significant benefits in scenarios where resource constraints and computational efficiency are critical [11].

Further illustrating KANs' potentials, [10] also showed that KAN architectures achieve performance comparable to CNN and traditional MLP architectures on simpler image datasets, such as MNIST and CIFAR-10, with a considerably reduced number of parameters and lower computational requirements. This efficiency positions KANs as particularly suitable for deployment in resource-constrained environments, such as edge devices or embedded systems, where model size and computational efficiency are critical constraints.

Nevertheless, several studies have also highlighted notable limitations of KANs, particularly their sensitivity to noise. Research presented in [3] and [4] emphasizes that KANs exhibit significant performance degradation even when exposed to relatively small noise perturbations. These studies revealed that KANs can sometimes underperform compared to MLPs when the input data contains noise or irregularities, suggesting potential vulnerability in practical, real-world conditions. The spline-based activation functions used within KANs, while beneficial for smooth and continuous approximations, may contribute to increased sensitivity when encountering noisy inputs, as subtle perturbations can alter spline approximations disproportionately.

Moreover, the computational demands associated with spline optimization may exacerbate the sensitivity to noisy inputs, as these functions inherently attempt to closely fit the training data, increasing susceptibility to overfitting on noisy samples. These observations are further supported in [12], that highlight potential limitations of KANs in hardware and computational settings, particularly when working with complex datasets that demand higher computational resources. Their findings indicate that the increased complexity of learnable spline functions might lead to diminishing returns, where additional computational costs do not necessarily translate into proportional performance gains.

Similarly, [11] concludes that the practical advantages of KANs might not be evident for more challenging, complex datasets such as CIFAR-10, where traditional NN architectures like CNNs and ResNets typically dominate. They argue that despite their theoretical appeal and potential interpretability advantages, the practical benefits of employing KANs in more challenging or high-dimensional scenarios remain uncertain and require further validation.

Given these mixed findings, the robustness and practical efficiency of KANs need careful evaluation across diverse

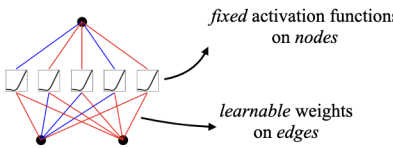
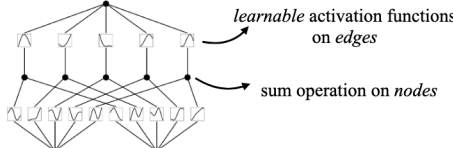
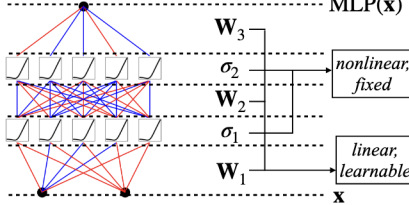
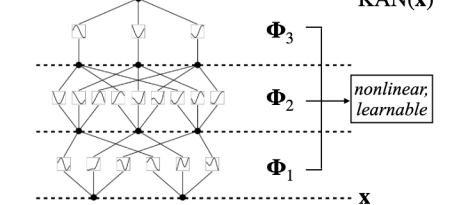
Model	<b>Multi-Layer Perceptron (MLP)</b>	<b>Kolmogorov-Arnold Network (KAN)</b>
Theorem	<b>Universal Approximation Theorem</b>	<b>Kolmogorov-Arnold Representation Theorem</b>
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(e)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left( \sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	(a)  fixed activation functions on nodes learnable weights on edges	(b)  learnable activation functions on edges sum operation on nodes
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	(c)  $\mathbf{W}_3$ $\sigma_2$ $\mathbf{W}_2$ $\sigma_1$ $\mathbf{W}_1$ $\mathbf{x}$ nonlinear, fixed linear, learnable	(d)  $\Phi_3$ $\Phi_2$ $\Phi_1$ $\mathbf{x}$ nonlinear, learnable

Figure 4. KAN vs MLP architectures compared, source: [2].

scenarios, datasets, and types of perturbations. While KANs clearly demonstrate potential for specific use-cases, particularly those prioritizing interpretability and computational efficiency, their sensitivity to adversarial and environmental noise requires comprehensive assessment and mitigation strategies, motivating further research into enhanced training techniques, such as adversarial training, which will be explored in subsequent sections of this paper.

### C. Adversarial Attacks

Recent advances in ML have sparked significant interest in understanding and mitigating vulnerabilities inherent to deep learning models. Central to this investigation are AAs, which strategically exploit vulnerabilities in models by introducing carefully crafted perturbations to the input data. The research into AAs has been especially vigorous in the computer vision domain, given the sensitivity of image-based models to subtle input changes that can drastically alter outputs [6], [13]. This has significant implications for applications relying heavily on image recognition, such as autonomous vehicles, security systems, and facial recognition technologies.

Among the numerous AA techniques developed, the FGSM and PGD have emerged as prominent benchmarks. FGSM, introduced by [14], crafts adversarial examples by leveraging gradients to add minimal perturbations that mislead the model's predictions. Due to its computational simplicity and effectiveness, FGSM remains widely used for initial robustness assessments. Conversely, PGD, introduced by [15], applies an iterative optimization-based procedure to find more potent perturbations, typically resulting in stronger attacks that are more challenging for models to withstand. Due to its iterative

nature, PGD has become the de facto standard for rigorous robustness evaluations, especially in the context of image classification tasks where even minor perturbations to input data can lead to substantial accuracy degradation [16].

Several defenses against these attacks have been proposed, ranging from detection and preprocessing approaches to robust training methodologies. Techniques such as adversarial example detection [17], diversity-enhancing strategies to mitigate attacks [18], and methods leveraging momentum to optimize the defense mechanism against PGD [19], have shown varying degrees of effectiveness. Despite these advancements, FGSM and PGD remain critically important for the systematic evaluation of model robustness due to their simplicity, efficiency, and established status in literature.

Tools like the Adversarial Robustness Toolbox (ART) [20] have been instrumental in facilitating systematic experimentation and reproducibility in adversarial research by providing standardized methods for generating adversarial examples and evaluating defenses. Likewise, benchmark datasets such as MNIST [21] (Modified National Institute of Standards and Technology handwritten-digit dataset) continue to serve as fundamental resources for comparative analyses due to their widespread acceptance, ease of use, and established benchmarks across a variety of ML models.

While KANs have begun to attract attention for their interpretability, adaptability, and computational advantages, their resilience to AA attacks remains significantly under-researched. Given the importance of robustness in safety-critical applications, understanding how various KAN architectures perform against established adversarial techniques like FGSM and PGD is crucial. In this extended study, we bridge

this critical research gap by systematically evaluating and comparing multiple KAN implementations under FGSM and PGD AAs. By doing so, we aim to identify the strengths and vulnerabilities inherent in these architectures, thereby laying the foundation for future research into targeted defense mechanisms specifically optimized for KAN-based models.

#### *D. Adversarial Training in ML*

Adversarial training has emerged as one of the most prominent and effective strategies for improving the robustness of ML models against AAs. Initially introduced by [14] as a defense against the FGSM, adversarial training involves the augmentation of training datasets with adversarially perturbed samples. This augmentation forces the model to encounter and learn from specifically crafted examples during training, thereby facilitating the development of more robust decision boundaries and improving model generalization to unseen adversarial inputs.

Subsequently, [15] significantly enhanced adversarial training by employing PGD as the adversarial example generator. PGD-based adversarial training iteratively applies small perturbations to input data, guiding the model toward learning highly robust and generalizable features. Due to its iterative nature, this method has been established as the state-of-the-art approach for benchmarking robustness in deep learning models. Empirical results consistently confirm that PGD-trained models exhibit significantly improved resilience compared to models trained using traditional or non-adversarial methods.

Building on these seminal studies, [22] proposed the TRADES method, introducing a theoretically-principled framework that explicitly balances the trade-off between adversarial robustness and natural accuracy. The TRADES framework introduces a regularization term that penalizes deviations from robust behavior while maintaining model performance on clean data. This approach has demonstrated notable improvements in robustness compared to standard adversarial training techniques, especially in image classification benchmarks.

Furthermore, [23] proposed integrating feature denoising techniques within adversarial training frameworks, enhancing the resilience of models against AAs by explicitly denoising intermediate feature representations during training. By embedding feature denoising mechanisms directly into adversarial training procedures, their method not only mitigates adversarial perturbations but also reduces the model's vulnerability to natural variations in data. These advancements underscore adversarial training as a continually evolving field, with methods becoming progressively sophisticated to counter increasingly powerful AAs.

However, despite the proven efficacy of adversarial training in enhancing model robustness, it introduces significant computational overhead and complexity [15]. Training models using adversarial techniques typically require extended computational resources and time due to the iterative generation of adversarial examples. Moreover, selecting suitable parameters, such as perturbation magnitude, training composition,

and learning rates, becomes critical to achieving optimal performance without compromising model accuracy on clean data. Careful dataset preparation, hyperparameter tuning, and rigorous empirical validation remain essential to leveraging the full benefits of adversarial training methodologies. Addressing these computational challenges and identifying efficient adversarial training strategies tailored to specific NN architectures, including KANs, remain vital areas for ongoing research and development.

#### *E. Adversarial Training Applied to KANs*

At the time of the publication of our original paper, the robustness of KAN architectures under adversarial conditions had begun receiving increased attention. Recent studies have expanded on the initial exploration of KAN vulnerabilities, systematically evaluating their performance under various adversarial perturbations and comparing them against traditional NN architectures. For instance, [24] investigated the application of KANs in Wi-Fi-based positioning systems, examining their response to adversarial manipulations in wireless signal inputs. Similarly, [25] assessed robustness aspects of KANs across a range of image classification benchmarks, providing valuable comparative analyses that underscore both strengths and limitations of KAN models in adversarial conditions. Another recent study by [26] evaluated the resilience of KAN architectures to AAs within broader applied ML contexts, highlighting the nuanced sensitivity of spline-based activation functions used within KAN models.

Despite the increasing focus on evaluating KAN robustness, the specific application of adversarial training methodologies to KAN architectures remains notably underexplored. To date, adversarial training has predominantly been applied to well-established models such as CNNs and transformers, whereas its impact on KAN models has yet to be rigorously investigated. Although the inherent flexibility and adaptivity of KANs suggest that adversarial training could significantly enhance their robustness, systematic empirical studies in this area are scarce. Consequently, many aspects remain unexplored, including how different compositions and intensities of adversarially perturbed data influence the training process, as well as the specific interactions between spline-based activation functions and adversarial samples.

Given this substantial gap, there is an important opportunity for research that specifically investigates adversarial training tailored to the unique properties of KAN architectures. Detailed analyses examining the relationship between adversarial perturbation strategies (such as FGSM and PGD) and the adaptability of KAN activation functions could provide essential insights for designing more robust models. Additionally, exploring computationally efficient adversarial training methodologies suitable for the unique structural properties of KANs could further unlock their potential for secure, real-world deployment. Addressing these open questions will be critical for future research, ultimately informing best practices for integrating adversarial training strategies into the design and deployment of KAN models.

### III. METHODOLOGY

The primary objective of our methodology is to assess how different KAN architectures respond relative to each other and the baseline MLP classifier under adversarial perturbations, placing emphasis on comparative robustness rather than absolute performance optimization. While we acknowledge that each evaluated model could potentially benefit from further tuning through parameter optimization, architectural adjustments, or advanced regularization methods, we operate under the assumption that the relative effects of AAs will remain consistent regardless of these enhancements.

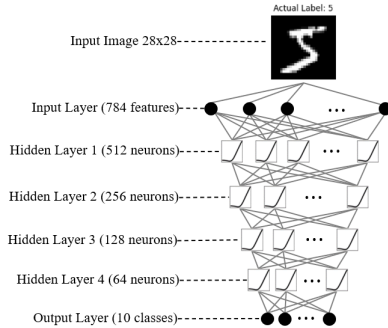


Figure 5. MLP Architecture, source: [7].

This assumption provides a clear foundation for comparing the intrinsic robustness characteristics of various KANs implementation. Nevertheless, future research should rigorously investigate the validity and generalizability of this assumption by exploring the impact of advanced training techniques on robustness outcomes.

The general structure of the KAN networks architecture used in our experiments is illustrated in Figures 5 and 6, which highlight the key differences between traditional MLPs and KAN models. All evaluated KAN models follow this fundamental architectural concept, where traditional node-based activation functions are replaced with edge-based learnable activation functions. The adversarial robustness of four distinct KAN implementations is systematically examined: *Linear (Efficient) KAN* [27], *Naïve Fourier KAN* [28], *Jacobi KAN* [29], and *Chebyshev KAN* [30].

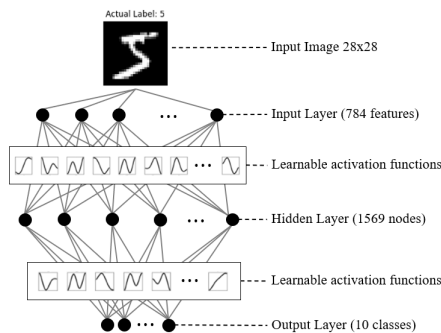


Figure 6. KAN Architecture, source: [7].

The robustness of each model is evaluated under controlled adversarial conditions, involving Gaussian noise as well as two widely recognized AAs technique: the FGSM, and PGD. These adversarial perturbations are generated and administered using the ART [20]. Performance robustness metrics such as accuracy, precision, recall, and F1-scores are utilized to provide a comprehensive understanding of model sensitivity to adversarial perturbations. The well-known MNIST dataset [21], consisting of 33,600 training samples and 8,400 test samples of handwritten digits, provides a standard benchmark that ensures consistency and comparability of results across models.

Furthermore, to extend our previous findings and explore potential improvements in model robustness, we introduce adversarial training by augmenting the original training dataset with adversarially perturbed examples. Specifically, we construct three training sets with varying proportions of clean MNIST samples combined with adversarial samples generated by FGSM, PGD, and Gaussian noise. The training dataset compositions are (i) 85% clean data and 5% of each perturbation type, (ii) 70% clean data and 10% each of noise, FGSM, and PGD, and (iii) 55% clean data and 15% each of noise, FGSM, and PGD. Through this systematic approach, we aim to evaluate how the inclusion of adversarial examples during training influences the robustness and generalizability of different KAN architectures.

In the subsequent sections, detailed results from these experiments will be analyzed, highlighting insights into the relative effectiveness of adversarial training strategies across diverse KAN implementations. Metrics including accuracy, precision, recall, and F1-scores provide a comprehensive understanding of robustness gains and vulnerabilities under adversarial conditions, guiding future research directions toward optimized KAN training strategies.

#### A. Model Architectures

In this research, we evaluate the robustness of different KANs architecture against AA and compare their performance with a traditional MLP baseline. All architectures use the MNIST dataset [21] and share common parameters for training, such as an AdamW optimizer with a learning rate of 0.001, weight decay for regularization, and an exponential learning rate scheduler to adjust the learning rate dynamically throughout training.

**MLP Classifier** is utilized as a baseline reference. The model comprises five fully-connected layers, progressively decreasing in size:  $784 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 10$ . ReLU activation and dropout layers with probability 0.2 are employed after each layer, ensuring model regularization and reduced overfitting. The final output layer utilizes a softmax activation function, providing classification probabilities for each of the ten MNIST digit classes.

The primary models under investigation are four distinct implementations of KANs, each employing unique activation functions along their edges. All KAN implementations share a similar basic architecture, as depicted in Figure 6, but differ substantially in their choice of edge-based activation functions.



**Linear KAN (Efficient KAN)** [27] utilizes spline-based activation functions parameterized by spline order and grid size. Specifically, the implementation uses spline order 3 and grid size 5, corresponding to a computationally efficient parameterization recommended by the authors. The architecture employs the standard configuration derived from the Kolmogorov-Arnold theorem, where the input dimension of 784 (the MNIST image size of  $28 \times 28$ ) is decomposed into one-dimensional spline functions along the network edges. This configuration results in  $(28 \times 28) \times 2 + 1 = 1569$  spline parameters, providing the model with substantial flexibility for capturing MNIST data patterns efficiently.

**Naïve Fourier KAN** [28] modifies the standard spline-based KAN by employing Fourier series coefficients to parameterize the learnable activation functions. Fourier-based activation functions provide smooth and periodic approximations, which inherently bound the activation functions numerically and avoid the common issues associated with spline parameterizations going out of grid bounds. Specifically, the Fourier KAN configuration used in our experiments employs grid size 56, corresponding to twice the dimension of input features, along with initialization parameters that ensure numerical stability and smoothness of learned functions.

**Chebyshev KAN (ChebyKAN)** [30] substitutes spline functions with Chebyshev polynomials. Chebyshev polynomials, due to their orthogonality and numerical stability, provide efficient approximations suitable for polynomial interpolations over bounded intervals. In our experiments, we employed Chebyshev polynomials of degree 7, aiming to balance approximation accuracy and computational efficiency. ChebyKAN requires fewer parameters to achieve comparable performance relative to spline-based KANs, making it appealing for scenarios where computational resources are constrained.

**Jacobi KAN (JacobiKAN)** [29], derived from the ChebyKAN framework, uses Jacobi polynomials, a broader family of orthogonal polynomials parameterized by two additional parameters ( $a$ ,  $b$ ) controlling polynomial shape. In our experiments, we selected a polynomial degree of 7 with default parameters  $a = 0.0$  and  $b = 0.0$  - a special case of Jacobi, the Legendre polynomials. This is typically used for MNIST classifications. JacobiKAN provides an adaptive and flexible framework capable of adjusting polynomial forms according to task-specific data characteristics. However, this flexibility introduces additional complexity, requiring careful parameter tuning during training.

All four KAN implementations share a fundamental architectural structure illustrated in Figure 6, differing primarily in the form of their learnable activation functions. By evaluating these architectures systematically, our study seeks to quantify and understand the impact of different parameterizations on model robustness against adversarial perturbations and noise.

## B. Attack Architecture

**Noise Attack:** We conducted Gaussian noise attacks at a noise level of 100 to evaluate the robustness of the models

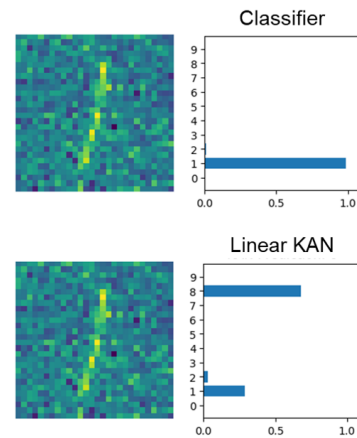


Figure 7. The Gaussian Noise Attack Example.

under extreme conditions. This high noise level was deliberately selected to amplify performance degradation, facilitating a clear comparison across the different KAN architectures and the baseline MLP model. Our prior research [7] valued the noise sensitivity of a single KAN model by incrementally increasing noise levels to determine its robustness relative to the MLP. In the current study, we shift our focus to systematically comparing multiple KAN variants, maintaining the MLP as a consistent baseline for robustness benchmarking. An MNIST digit example (digit '1') corrupted by Gaussian noise at the level of 100 is shown in Figure 7, illustrating the extreme noise conditions used in our robustness assessments.

**FGSM Attack:** The ART [20] was utilized to generate adversarial examples and implement the FGSM attack across all models. Perturbations were introduced into the MNIST test dataset to create adversarial samples, with the epsilon parameter typically ranging from 0.1 to 0.8. A higher epsilon increases perturbation visibility in images. For this research, an epsilon value of 0.5 was selected, sufficient to significantly degrade model performance without introducing visually noticeable distortions, thus preserving realism in the adversarial scenario. An example of an MNIST digit (digit '1') subjected to the FGSM attack is shown in Figure 8, highlighting how subtle perturbations can drastically alter model predictions.

**PGD Attack:** We also employed ART [20] to facilitate the PGD AAs. PGD iteratively generates small random perturbations to the input data, progressively maximizing the loss function. Each iteration incrementally adjusts perturbation magnitude, while carefully controlling the maximum perturbation size to maintain imperceptibility to human observers. This iterative approach positions PGD as one of the strongest first-order AAs methods available, significantly more potent than FGSM. Consistent with the FGSM setup, a perturbation level of 0.5 was adopted to simulate realistic adversarial conditions. Figure 9 presents an MNIST digit example (digit '1') after a PGD attack, demonstrating the iterative nature of this strong adversarial perturbation and its effect on model classification.

**Tools and environment:** All KAN implementations are

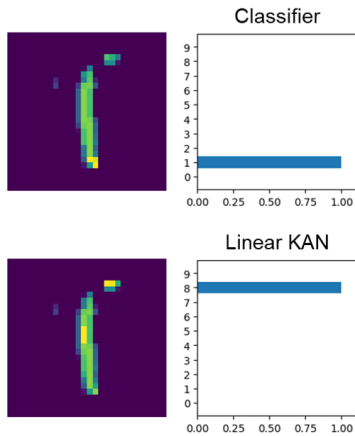


Figure 8. The FGSM Attack Example.

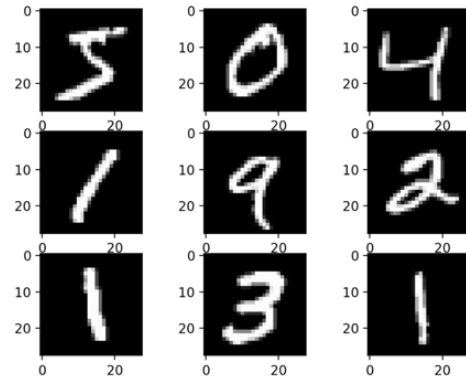


Figure 10. MNIST Dataset Example.

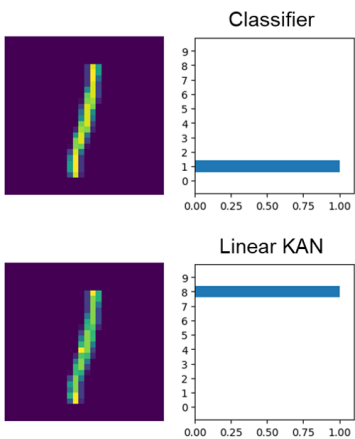


Figure 9. PGD Attack Example.

obtained from publicly available GitHub repositories [28]–[31], alongside ART [20]. The baseline MLP classifier was implemented independently using PyTorch and Scikit-learn Python libraries. The Google Colab cloud environment was utilized to conduct all experiments, ensuring consistency in hardware and software configurations. This standardized experimental environment is intended to facilitate reproducibility and validation of our results.

### C. Experiments

All models, including the four KAN architectures: Linear (Efficient) KAN, Naïve Fourier KAN, Jacobi KAN, Chebyshev KAN, and the MLP baseline, were initially trained and evaluated in a controlled, non-adversarial setting using the MNIST dataset [21]. Figure 10 illustrates example digits from the MNIST dataset used in all experiments. Performance metrics, including accuracy, precision, recall, and F1 scores, were recorded for each architecture to establish a robust baseline for subsequent adversarial analyses.

We then assessed each model's robustness under adversarial conditions by subjecting them individually to Gaussian

noise, FGSM, and PGD attacks. For each attack scenario, we computed the relative change in performance metrics compared to their baseline values. These results provided insights into the vulnerabilities of each KAN variant relative to the MLP classifier, allowing for a systematic analysis of model-specific weaknesses and strengths under adversarially perturbed conditions.

Expanding on this initial analysis, adversarial training experiments were conducted to explore strategies for enhancing model robustness. Specifically, models were retrained with adversarially augmented datasets composed of varying proportions of clean and perturbed data, as follows:

- 85% clean MNIST data combined with 5% each of Gaussian noise, FGSM, and PGD perturbed samples.
- 70% clean MNIST data combined with 10% each of Gaussian noise, FGSM, and PGD perturbed samples.
- 55% clean MNIST data combined with 15% each of Gaussian noise, FGSM, and PGD perturbed samples.

This adversarial training strategy aimed to quantify how incorporating a controlled proportion of adversarially generated data into the training set affects model performance and robustness. Each model was retrained separately under these three training set compositions, and performance metrics were reevaluated on clean as well as adversarially perturbed test sets (noise, FGSM, PGD). The goal was to identify optimal training compositions capable of significantly enhancing robustness without severely compromising accuracy on clean data.

To ensure consistency and reproducibility, all training sessions employed identical hyperparameters, including the AdamW optimizer with a learning rate of 0.001, weight decay for regularization, and an exponential learning rate scheduler. Each training scenario was repeated multiple times to ensure the reliability of observed improvements in robustness metrics.

In the results section that follows, detailed analyses will be presented, comparing performance outcomes from standard training versus adversarial training across all tested models. This comprehensive experimental approach provides critical insights into the efficacy of adversarial training for improving



KAN robustness, contributing valuable guidance for future research into secure and robust NN architectures.

#### IV. RESULTS

This section presents a comprehensive analysis of our experimental results, structured systematically into five subsections. We begin by establishing baseline performance metrics for all evaluated models in the absence of adversarial conditions. Subsequent subsections report detailed findings on model robustness under Gaussian noise, FGSM, and PGD AAs. Finally, we present a thorough evaluation of the impact of adversarial training on model resilience, comparing performance across varying proportions of adversarially perturbed training data. The analyses provided herein offer valuable insights into the relative strengths and vulnerabilities of different KAN architectures compared to the baseline MLP classifier, highlighting critical considerations for enhancing model robustness.

##### A. Before Attacks

Initially, we evaluated all models under clean (non-adversarial) conditions using the MNIST dataset, as detailed in Table I. This baseline evaluation provides an essential reference point for assessing subsequent robustness to adversarial perturbations.

TABLE I  
ACCURACY BY MODEL. TRAIN SET: 100% MNIST.

Model	Clean	Noise100	FGSM 0.5	PGD 0.5
Classifier	0.98	0.94	0.79	0.66
KAN Linear	0.98	0.86	0.29	0.11
Naïve Fourier	0.92	0.16	0.11	0.22
Jacobi	0.93	0.51	0.08	0.05
Cheby	0.92	0.39	0.05	0.04

Accuracy results before AAs are visualized in Figure 11, clearly indicating that the MLP Classifier and the Linear KAN both achieve nearly identical accuracy (98%), establishing a strong performance baseline. Conversely, the other three KAN variants: Naïve Fourier, Jacobi, and Chebyshev exhibit somewhat lower accuracy scores (92-93%). Although the primary objective of this study focuses on evaluating relative robustness under adversarial conditions rather than absolute accuracy, these performance discrepancies warrant further exploration. Future research may investigate whether model-specific architectural differences, parameter settings, or alternative optimization strategies might account for these performance gaps and potentially improve the absolute accuracy of the affected KAN architectures.

Another notable observation relates to computational complexity and training duration. Despite improvements from the use of Google Colab's free-tier T4 GPU, training times for KAN models remained substantially longer compared to the simpler MLP architecture. Specifically, KAN architectures typically required roughly ten times longer to train than

the baseline MLP. This discrepancy, attributed primarily to the computational overhead associated with spline-based and polynomial-based activation functions, highlights a significant practical consideration for real-world deployment and iterative training workflows.

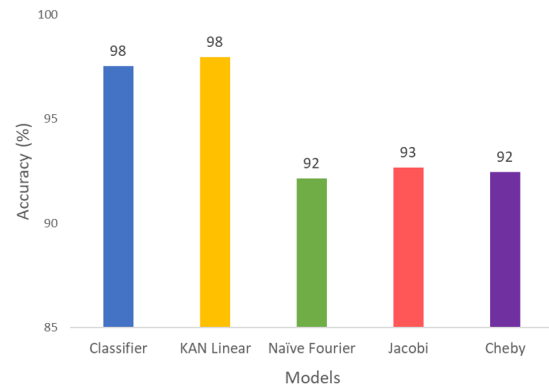


Figure 11. Model Accuracy Comparison Before Attacks.

An additional intriguing finding involves the class-wise balance of model performance, particularly evident from examining the F1 scores for individual digit classes, as illustrated in Figure 12. The Linear KAN model demonstrates generally balanced F1 scores across most digit classes but exhibits a pronounced drop in performance on digit 9. Other KAN models similarly reflect class imbalance patterns, suggesting inherent limitations or biases within their activation function parameterizations, and suggesting that certain activation functions or training methodologies may disproportionately impact specific digit classes. Investigating the causes of these class-specific discrepancies may offer valuable insights into further optimizing KAN architectures or identifying data-specific challenges. Such analyses remain outside the scope of this current study but represent promising avenues for future research.

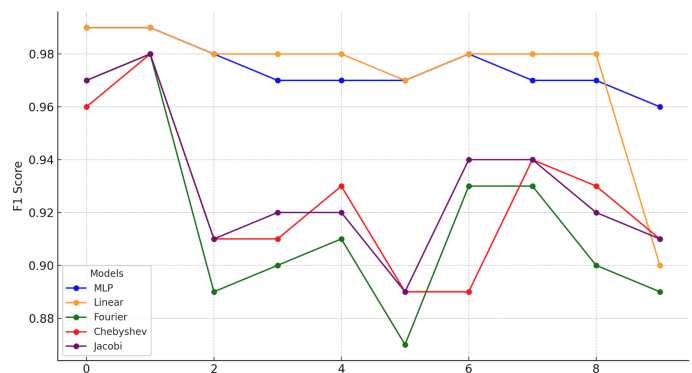


Figure 12. Model F1 Score Comparison Before Attacks, source: [1].

Overall, these baseline performance evaluations provide essential context for subsequent analyses of model robust-

ness under various adversarial perturbations, enabling precise quantification of robustness changes attributed explicitly to adversarial conditions.

### B. Gaussian Noise Attack Results

All evaluated models exhibited reduced accuracy when exposed to Gaussian noise at the extreme level of 100, as detailed in Table I. Figure 13 clearly illustrates the drop in accuracy for each model before and after the noise attack.

The MLP Classifier demonstrated robust performance, maintaining high accuracy at 94%, reflecting only a modest reduction of approximately 4%. The Linear KAN model also performed relatively well under noisy conditions, achieving an accuracy of 86%, though this still represents a notable accuracy drop of about 12%. In contrast, the other evaluated KAN architectures: Naïve Fourier, Jacobi, and Chebyshev, experienced severe degradation in performance, with accuracy declining dramatically to 16%, 51%, and 39%, respectively.

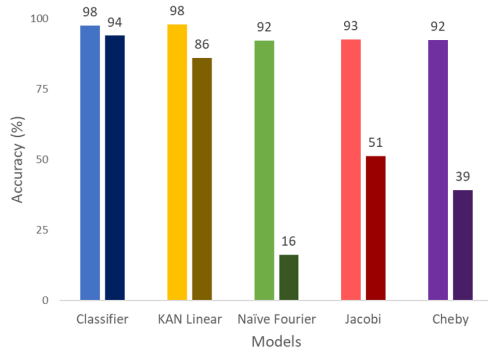


Figure 13. Model Accuracy Comparison After Noise Attack.

Figure 14 provides a visualization of the percentage accuracy losses, further underscoring the pronounced vulnerability of polynomial-based KAN models to Gaussian noise attacks. While Linear KAN demonstrates comparatively moderate sensitivity to noise, its accuracy loss is still substantially higher than the baseline MLP, suggesting inherent architectural vulnerabilities of KAN models under noisy conditions. These observations emphasize the necessity of further investigation into mechanisms underlying KAN models' sensitivity to noise, guiding future enhancements in model robustness.

TABLE II  
ACCURACY REDUCTION, (%).

Model	Noise100	FGSM 0.5	PGD 0.5
Classifier	4	18	31
KAN Linear	12	69	87
Naïve Fourier	76	81	70
Jacobi	41	84	88
Cheby	53	88	88

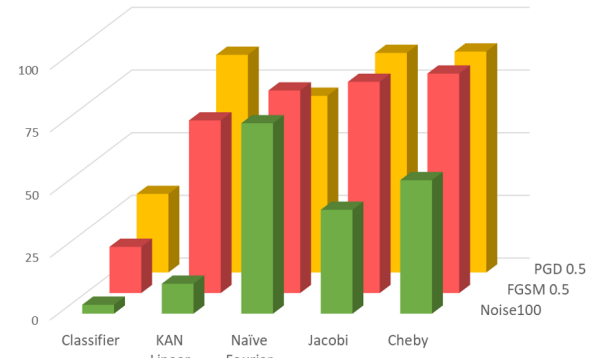


Figure 14. Accuracy Loss Comparison by Attack.

### C. FGSM Attack Results

Under the FGSM attack with a perturbation parameter ( $\epsilon = 0.5$ ), all evaluated models experienced significantly greater accuracy losses compared to the Gaussian noise attack. Figure 15 clearly illustrates this reduction in accuracy scores for each model when subjected to FGSM-generated adversarial examples.

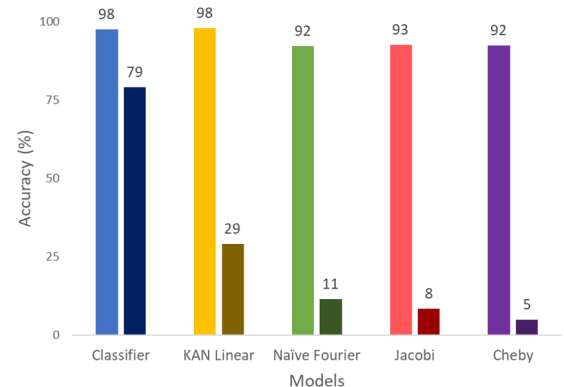


Figure 15. Model Accuracy Comparison After FGSM Attack.

Consistent with previous attack outcomes, the baseline MLP classifier demonstrated the strongest resilience among all models, yet it still experienced a substantial accuracy drop from 98% to 79%, representing a relative accuracy loss of approximately 18%. The Linear KAN model, while maintaining higher absolute accuracy compared to other KAN architectures, displayed a considerable accuracy reduction, falling from 98% to just 29%. Notably, this corresponds to a relative loss of approximately 69% in accuracy, highlighting Linear KAN's vulnerability to the FGSM attack.

Interestingly, the Naïve Fourier KAN model, which exhibited poor performance under Gaussian noise conditions, showed a relatively stronger resilience compared to other

polynomial-based KAN variants under FGSM perturbations, achieving an accuracy of 11%. While still significantly affected, this result contrasts sharply with its extreme vulnerability under noise attacks. Jacobi and Chebyshev KAN models suffered the most severe accuracy losses, dropping from initial accuracies around 92-93% to below 10% accuracy post-FGSM attack, underscoring their heightened sensitivity to adversarially generated perturbations.

The relative accuracy losses across models under different attack conditions are summarized in Table II and visually depicted in Figure 14. This comprehensive visualization emphasizes the particularly devastating impact of the FGSM attack on the polynomial-based KAN models.

An intriguing observation from these results is the apparent inverse performance relationship between the polynomial-based KAN models' responses to Gaussian noise and FGSM attacks. This phenomenon, visually apparent in the comparison of Figures 14 and 15, suggests distinct underlying vulnerabilities to different perturbation types. This finding provides a compelling direction for future research, potentially exploring the underlying mechanisms driving these divergent responses, and informing more targeted strategies for robustness enhancement.

#### D. PGD Attack Results

Under the PGD attack at an intensity level of 0.5, all tested models suffered severe accuracy degradation. Figure 16 illustrates a significant decline in accuracy for each model when subjected to the PGD adversarial perturbations.

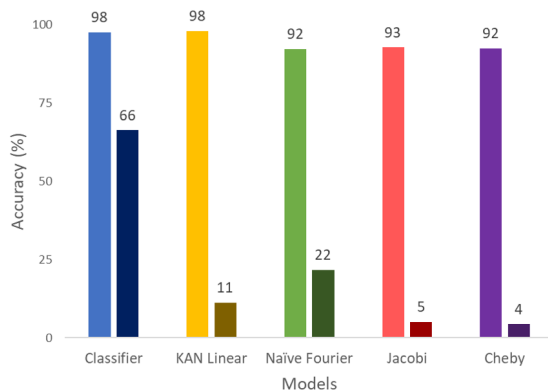


Figure 16. Model Accuracy Comparison After PGD Attack.

Interestingly, while the MLP classifier demonstrated the highest absolute accuracy (66%) following the PGD attack, it experienced a considerable relative accuracy loss of about 31%, highlighting significant vulnerability despite its robustness under other conditions. The Linear KAN, which performed well under noise attacks, showed an exceptionally high sensitivity to PGD attacks, with accuracy plunging drastically to 11%, reflecting an 87% relative loss.

On the contrary, other KAN architectures, particularly Naïve Fourier KAN showed slightly better resilience in relative terms compared to Linear KAN, albeit their absolute accuracy scores remained critically low (22%). Notably, Chebyshev and Jacobi KAN displayed minimal accuracy retention at around 5%.

It is important to emphasize that the overall degradation was catastrophic for all models. The accuracy for the majority of digit classes dropped dramatically to zero or near-zero for all models, indicating extensive vulnerability across all tested NN variants.

Nevertheless, from a relative accuracy retention standpoint, Naïve Fourier showed marginally better resilience than other KAN variants, making it the least affected architecture under the PGD attack scenario. This nuanced difference, although minor in absolute terms, presents an intriguing avenue for future investigation into what specific properties of Naïve Fourier activation functions might contribute to improved resilience against iterative adversarial perturbations like PGD.

These observations are visually summarized in Figure 16, clearly highlighting the extensive accuracy loss across all models, reinforcing the potent effectiveness of PGD attacks on current NN architectures.

#### E. Adversarial Training Results

To further examine the robustness of our models, we conducted adversarial training experiments by progressively reducing the proportion of clean MNIST data in the training set (85%, 70%, and 55%) and simultaneously increasing the adversarially perturbed examples. Tables III, IV, and V summarize the performance of each model under these conditions.

TABLE III  
ACCURACY BY MODEL. TRAIN SET: 85% MNIST.

Model	Clean	Noise100	FGSM 0.5	PGD 0.5
Classifier	0.97	0.93	0.96	0.88
KAN Linear	0.98	0.80	0.98	0.99
Naïve Fourier	0.92	0.18	0.97	0.98
Jacobi	0.92	0.82	0.69	0.74
Cheby	0.92	0.50	0.88	0.90

TABLE IV  
ACCURACY BY MODEL. TRAIN SET: 70% MNIST.

Model	Clean	Noise100	FGSM 0.5	PGD 0.5
Classifier	0.97	0.92	0.96	0.89
KAN Linear	0.97	0.76	0.98	0.98
Naïve Fourier	0.92	0.18	0.94	0.97
Jacobi	0.92	0.80	0.66	0.72
Cheby	0.91	0.54	0.91	0.93

Our adversarial training experiments revealed significant robustness gains across all evaluated models, demonstrating substantial resilience improvements against FGSM and PGD attacks, even when training data contained high proportions

TABLE V  
ACCURACY BY MODEL. TRAIN SET: 55% MNIST.

Model	Clean	Noise100	FGSM 0.5	PGD 0.5
Classifier	0.97	0.92	0.96	0.89
KAN Linear	0.97	0.71	0.94	0.96
Naïve Fourier	0.92	0.19	0.95	0.98
Jacobi	0.91	0.80	0.65	0.71
Cheby	0.91	0.60	0.92	0.96

of adversarial samples. Linear KAN exhibited remarkable improvement, achieving 98% accuracy under FGSM and 99% accuracy under PGD with 85% clean data. Even at the lowest clean data level (55%), Linear KAN maintained 94% and 96% accuracy for FGSM and PGD respectively, though accuracy dropped significantly to 71% under high-level noise attacks. Figure 17 visually highlights the robustness improvement across adversarial scenarios, and Table VI provides detailed information.

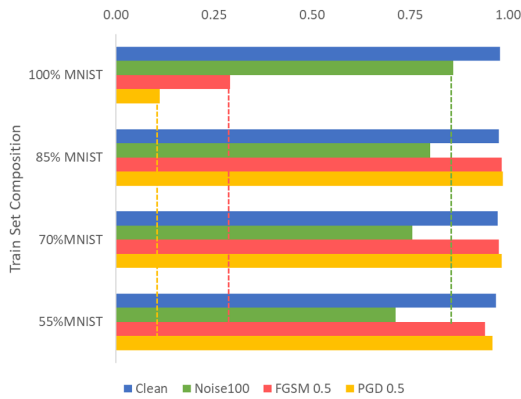


Figure 17. Model Accuracy by Train set Linear KAN.

The Naïve Fourier KAN demonstrated a dramatic transformation, jumping from poor performance (11% FGSM and 22% PGD accuracy at 100% MNIST clean data) to 97% and 98% accuracy respectively at 85% clean data. Even further reduction to 55% clean data sustained high performance, achieving 95% for FGSM and 98% for PGD (Figure 18, Table VII). However, Naïve Fourier continued to underperform in noise attacks across all data compositions, never exceeding 20% accuracy.

Jacobi and Cheby KAN models also improved significantly, albeit with more moderate gains. Jacobi KAN, which initially had catastrophic performance under FGSM (8%) and PGD (5%) at 100% MNIST, improved considerably to 69% and 74% respectively at 85% MNIST clean data. However, further reduction in clean data slightly diminished robustness, settling at 65% FGSM and 71% PGD at 55% MNIST (Figure 19,

TABLE VI  
PERFORMANCE METRICS BY ATTACK. KAN LINEAR.

	TestSet	Precision	Recall	F1-score	Accuracy
100% MNIST	Clean	0.98	0.98	0.98	0.98
	Noise	0.90	0.86	0.86	0.86
	FGSM	0.50	0.29	0.30	0.29
	PGD	0.20	0.11	0.07	0.11
85% MNIST	Clean	0.98	0.98	0.98	0.98
	Noise	0.86	0.80	0.81	0.80
	FGSM	0.98	0.98	0.98	0.98
	PGD	0.99	0.99	0.99	0.99
70% MNIST	Clean	0.97	0.97	0.97	0.97
	Noise	0.83	0.76	0.75	0.76
	FGSM	0.98	0.98	0.98	0.98
	PGD	0.99	0.98	0.98	0.98
55% MNIST	Clean	0.97	0.97	0.97	0.97
	Noise	0.82	0.71	0.71	0.71
	FGSM	0.94	0.94	0.94	0.94
	PGD	0.96	0.96	0.96	0.96

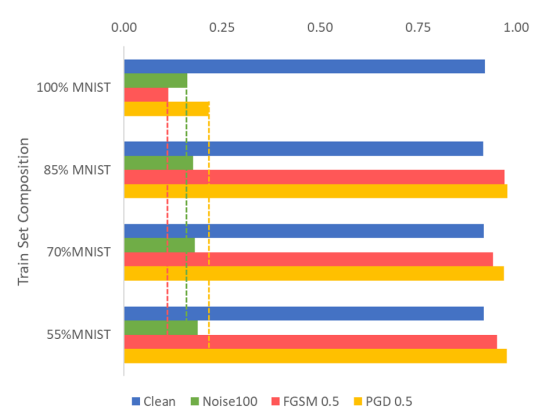


Figure 18. Model Accuracy by Train set Naive Fourier.

Table VIII).

Cheby KAN showed an impressive recovery from initial single-digit accuracy figures to consistently high performances (88% FGSM and 90% PGD at 85% MNIST), improving further as clean data proportion decreased, reaching 92% FGSM and 96% PGD at 55% MNIST (Figure 20, Table IX).

The MLP classifier displayed robust and consistent improvement, maintaining high performance with minor fluctuations. With 85% clean data, the MLP reached 96% FGSM and 88% PGD accuracy, and notably, further reductions of clean data to 55% sustained performance, yielding 96% FGSM and 89% PGD accuracy (Figure 21, Table X).

TABLE VII  
PERFORMANCE METRICS BY ATTACK. NAÏVE FOURIER.

	TestSet	Precision	Recall	F1-score	Accuracy
100% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.17	0.16	0.16	0.16
	FGSM	0.22	0.11	0.07	0.11
	PGD	0.53	0.22	0.20	0.22
85% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.19	0.18	0.17	0.18
	FGSM	0.97	0.97	0.97	0.97
	PGD	0.98	0.98	0.98	0.98
70% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.20	0.18	0.17	0.18
	FGSM	0.94	0.94	0.94	0.94
	PGD	0.97	0.97	0.97	0.97
55% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.20	0.19	0.19	0.19
	FGSM	0.95	0.95	0.95	0.95
	PGD	0.98	0.98	0.98	0.98

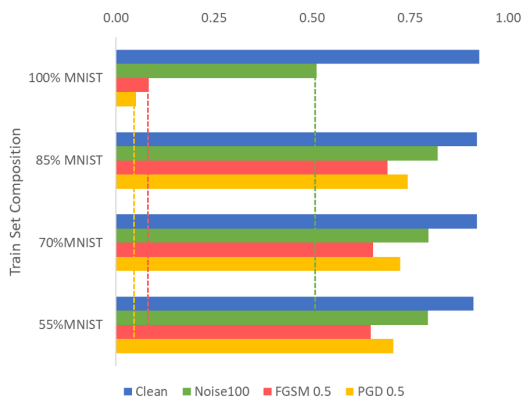


Figure 19. Model Accuracy by Train set Jacobi.

TABLE VIII  
PERFORMANCE METRICS BY ATTACK. JACOBI.

	TestSet	Precision	Recall	F1-score	Accuracy
100% MNIST	Clean	0.93	0.93	0.93	0.93
	Noise	0.68	0.51	0.52	0.51
	FGSM	0.07	0.08	0.05	0.08
	PGD	0.05	0.05	0.02	0.05
85% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.83	0.82	0.82	0.82
	FGSM	0.69	0.69	0.69	0.69
	PGD	0.74	0.74	0.73	0.74
70% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.82	0.80	0.80	0.80
	FGSM	0.66	0.66	0.65	0.66
	PGD	0.72	0.72	0.72	0.72
55% MNIST	Clean	0.91	0.91	0.91	0.91
	Noise	0.82	0.80	0.80	0.80
	FGSM	0.66	0.65	0.65	0.65
	PGD	0.71	0.71	0.70	0.71

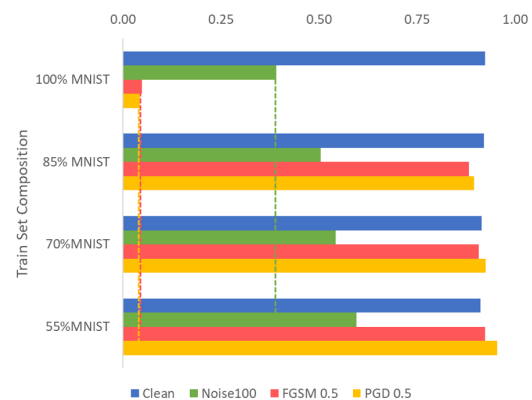


Figure 20. Model Accuracy by Train set Cheby.

## V. EVALUATION AND DISCUSSION

Our experiments show that the same spline flexibility that gives KANs their predictive power also makes them prone to overfitting. Under adversarial perturbations, KANs can lose far more accuracy than a standard MLP, creating a serious risk in security-sensitive contexts.

Adversarial training is an effective countermeasure. Injecting a small fraction (5%) of adversarial samples per AA type into the training set restores robustness across all variants, improving accuracy by more than 60 points under strong PGD attacks. Although generating those perturbed samples adds computational cost, the resulting resilience gains justify this overhead for any real-world KAN deployments.

Not all KANs respond equally. Linear and orthogonal-spline (Jacobi, Chebyshev) variants recover most of their robustness with modest adversarial mixing. The Fourier KAN, while nearly perfect under adversarial attacks after training, remains

highly vulnerable to Gaussian noise. Its noise accuracy never exceeds 20 % even at high perturbation ratios. This tells us that adversarial training alone cannot address stochastic-noise weaknesses; techniques such as input denoising is required. The Jacobi KAN shows the smallest net gain overall and may benefit from hybrid hardening tailored to its spline structure.

The next section outlines future directions, including systematic tuning of adversarial and noise ratios for each KAN type, theoretical analysis of spline susceptibility, and more efficient adversarial-sample generation methods.

## VI. CONCLUSION AND FUTURE WORK

In this work, we have: 1) Quantified the vulnerability of four KAN architectures, revealing up to 88% under adversarial attacks (and up to 76% under noise conditions). 2) Shown that modest adversarial training (5% perturbed samples per



TABLE IX  
PERFORMANCE METRICS BY ATTACK. CHEBY.

	TestSet	Precision	Recall	F1-score	Accuracy
100% MNIST	Clean	0.93	0.92	0.92	0.92
	Noise	0.56	0.39	0.38	0.39
	FGSM	0.08	0.05	0.03	0.05
	PGD	0.01	0.04	0.01	0.04
85% MNIST	Clean	0.92	0.92	0.92	0.92
	Noise	0.65	0.50	0.51	0.50
	FGSM	0.88	0.88	0.88	0.88
	PGD	0.91	0.90	0.89	0.90
70% MNIST	Clean	0.92	0.91	0.91	0.91
	Noise	0.68	0.54	0.55	0.54
	FGSM	0.91	0.91	0.91	0.91
	PGD	0.93	0.93	0.92	0.93
55% MNIST	Clean	0.91	0.91	0.91	0.91
	Noise	0.70	0.60	0.61	0.60
	FGSM	0.92	0.92	0.92	0.92
	PGD	0.96	0.96	0.95	0.96

TABLE X  
PERFORMANCE METRICS BY ATTACK. CLASSIFIER.

	TestSet	Precision	Recall	F1-score	Accuracy
100% MNIST	Clean	0.98	0.98	0.98	0.98
	Noise	0.94	0.94	0.94	0.94
	FGSM	0.79	0.79	0.79	0.79
	PGD	0.66	0.66	0.65	0.66
85% MNIST	Clean	0.97	0.97	0.97	0.97
	Noise	0.93	0.93	0.93	0.93
	FGSM	0.96	0.96	0.96	0.96
	PGD	0.89	0.88	0.88	0.88
70% MNIST	Clean	0.97	0.97	0.97	0.97
	Noise	0.92	0.92	0.92	0.92
	FGSM	0.96	0.96	0.96	0.96
	PGD	0.89	0.89	0.89	0.89
55% MNIST	Clean	0.97	0.97	0.97	0.97
	Noise	0.92	0.92	0.92	0.92
	FGSM	0.96	0.96	0.96	0.96
	PGD	0.89	0.89	0.89	0.89

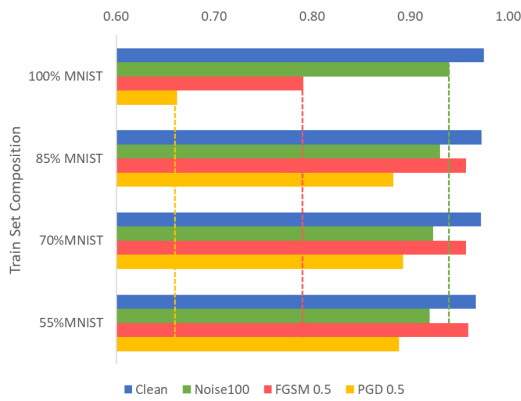


Figure 21. Model Accuracy by Train set Classifier.

AA type) recovers more than 60 points of robustness. 3) Identified that Fourier KANs remain noise-sensitive even after adversarial training, highlighting the need for future research and hybrid noise–adversarial defenses.

Our results demonstrate significant variation among KAN models in their response to AA and provide a comparative analysis against traditional MLP classifiers. Consistent with previous findings [1], [7], in the unprotected setting, the MLP baseline outperforms all KAN variants under FGSM and PGD attacks. However, after adversarial training, Linear and Fourier KANs exceed the MLP’s robustness, reaching nearly 99% accuracy against PGD, highlighting the effectiveness of targeted hardening for these architectures.

One critical observation in our study was the significant imbalance across classes within KAN models, particularly evident in Figure 12. Further investigation into the underlying causes of this imbalance could provide valuable insights into

improving the robustness and general performance of KANs. Understanding these mechanisms might not only enhance our theoretical understanding of KAN architectures but also guide practical improvements for diverse applications.

Adversarial training experiments provided substantial new insights. Introducing progressively greater proportions of adversarial data into the training sets notably improved resilience across all models. This approach significantly enhanced KAN models’ robustness, especially Linear and Naïve Fourier KANs, which achieved near-perfect accuracy (98%–99%) under both FGSM and PGD attacks with 85% clean data. Remarkably, even reducing clean training data to as low as 55%, these models maintained high accuracy (above 94%), demonstrating their considerable potential for adversarial robustness. In contrast, Jacobi and Cheby KANs showed substantial, though less pronounced, improvements, indicating that different activation functions significantly influence adversarial training outcomes.

Our study did not specifically address training efficiency, but the substantial training time observed for KAN models highlights a potential area for future research. Understanding and optimizing the trade-off between training efficiency and adversarial robustness, especially for novel architectures like KANs, is critical for broader adoption and practical applications.

#### Future Research Directions

Building on our results, we identify several promising areas for future investigation:

- Deepening theoretical understanding of why certain KAN models (e.g., Fourier) exhibit greater resistance to PGD

attacks, potentially guiding new architectural designs or activation function choices.

- Developing specialized adversarial robustness training strategies tailored explicitly for different KAN architectures to further leverage their inherent strengths.
- Exploring additional AA methodologies and evaluating KAN robustness on more diverse datasets. Future work should rigorously test KAN robustness using datasets beyond MNIST, such as CIFAR-10 or ImageNet, to validate the generalizability of our findings and their practical implications.
- Investigating and addressing the observed class imbalance issue within KAN models to improve both robustness and general classification performance.
- Assessing the balance between computational efficiency, training time, and model robustness to enhance the practical deployment of KAN models in real-world applications.

Pursuing these identified research directions will significantly deepen our theoretical understanding of KAN robustness, fostering advancements toward practically deployable, secure, and interpretable ML models.

#### ACKNOWLEDGMENT

We acknowledge the use of various general-purpose online and cloud-based tools, including those with AI-driven features, during the preparation of this work.

#### REFERENCES

- [1] N. Djosic, E. Ostanin, F. Hussain, S. Sharieh, and A. Ferworn, "KAN vs KAN: Examining Kolmogorov-Arnold networks (KAN) performance under adversarial attacks", in Proceedings of the SECURWARE 2024, The Eighteenth International Conference on Emerging Security Information, Systems and Technologies, Nov. 2024, pp. 17–22.
- [2] Z. Liu et al., "KAN: Kolmogorov-Arnold networks", Apr. 2024, [Online]. Available: <http://arxiv.org/abs/2404.19756>. Accessed: 14 May 2025.
- [3] C. Zeng, J. Wang, H. Shen, and Q. Wang, "KAN versus MLP on irregular or noisy functions", 2024, [Online]. Available: <https://arxiv.org/abs/2408.07906>. Accessed: 14 May 2025.
- [4] H. Shen, C. Zeng, J. Wang, and Q. Wang, "Reduced effective-ness of Kolmogorov-Arnold networks on functions with noise", Jul. 2024, [Online]. Available: <http://arxiv.org/abs/2407.14882>. Accessed: 14 May 2025.
- [5] D. Dasgupta, Z. Akhtar, and S. Sen, "Machine learning in cybersecurity: A comprehensive survey", Journal of Defense Modeling and Simulation, vol. 19, pp. 57–106, Jan. 2022.
- [6] B. Xi, "Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges", Wiley Interdisciplinary Reviews: Computational Statistics, vol. 12, p. 1511, Sep. 2020.
- [7] E. Ostanin, N. Djosic, F. Hussain, S. Sharieh, and A. Ferworn, "Evaluating the robustness of Kolmogorov-Arnold networks against noise and adversarial attacks", in Proceedings of the SECURWARE 2024, The Eighteenth International Conference on Emerging Security Information, Systems and Technologies, Nov. 2024, pp. 11–16.
- [8] J. Xu et al., "Fourierkan-gcf: Fourier Kolmogorov-Arnold network – an effective and efficient feature transformation for graph collaborative filtering", 2024, [Online]. Available: <https://arxiv.org/abs/2406.01034>. Accessed: 14 May 2025.
- [9] S. Sidhartha, A. Keerthana, R. Gokul, and K. Anas, "Chebyshev polynomial-based Kolmogorov-Arnold networks: An efficient architecture for nonlinear function approximation", 2024, [Online]. Available: <https://arxiv.org/abs/2405.07200>. Accessed: 14 May 2025.
- [10] M. Cheon, "Demonstrating the efficacy of Kolmogorov-Arnold networks in vision tasks a preprint", 2024, [Online]. Available: <https://arxiv.org/abs/2406.14916>. Accessed: 14 May 2025.
- [11] B. Azam and N. Akhtar, "Suitability of KANs for computer vision: A preliminary investigation", Jun. 2024, [Online]. Available: <http://arxiv.org/abs/2406.09087>. Accessed: 14 May 2025.
- [12] V. D. Tran et al., "Exploring the limitations of Kolmogorov-Arnold networks in classification: Insights to software training and hardware implementation", Jul. 2024, [Online]. Available: <http://arxiv.org/abs/2407.17790>. Accessed: 14 May 2025.
- [13] G. R. Machado, E. Silva, and R. R. Goldschmidt, "Adversarial machine learning in image classification: A survey towards the defender's perspective", Sep. 2020, [Online]. Available: <http://arxiv.org/abs/2009.03728>. Accessed: 14 May 2025.
- [14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples", Mar. 2015, [Online]. Available: <https://arxiv.org/abs/1412.6572>. Accessed: 14 May 2025.
- [15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks", Sep. 2019, [Online]. Available: <http://arxiv.org/abs/1706.06083>. Accessed: 14 May 2025.
- [16] W. Villegas, A. Jaramillo-Alcázar, and S. Luján-Mora, "Evaluating the robustness of deep learning models against adversarial attacks: An analysis with FGSM, PGD and CW", Big Data and Cognitive Computing, vol. 8, p. 8, Jan. 2024.
- [17] Y. Jang, T. Zhao, S. Hong, and H. Lee, "Adversarial defense via learning to generate diverse attacks", in In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2740–2749.
- [18] G. Sriramanan, S. Addepalli, A. Baburaj, and R. V. Babu, "Guided adversarial attack for evaluating and enhancing adversarial defenses", in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 20 297–20 308.
- [19] S. Mohandas, N. Manwani, and D. P. Dhulipudi, "Momentum iterative gradient sign method outperforms PGD attacks", in International Conference on Agents and Artificial Intelligence, vol. 3, Science and Technology Publications, Lda, 2022, pp. 913–916.
- [20] M.-I. Nicolae et al., "Adversarial robustness toolbox v1.0.0", 2019, [Online]. Available: <https://arxiv.org/abs/1807.01069>. Accessed: May 2025.
- [21] L. Deng, "The MNIST database of handwritten digit images for machine learning research", IEEE Signal Processing Magazine, vol. 29, pp. 141–142, Jun. 2012.
- [22] H. Zhang et al., "Theoretically principled trade-off between robustness and accuracy", in International conference on machine learning, PMLR, 2019, pp. 7472–7482.
- [23] C. Xie, Y. Wu, L. van der Maaten, A. Yuille, and K. He, "Feature denoising for improving adversarial robustness", 2019.
- [24] M. Goswami, R. Chatterjee, S. Mahato, and P. K. Pattnaik, "Adversarial-ensemble Kolmogorov Arnold networks for enhancing indoor wi-fi positioning: A defensive approach against spoofing and signal manipulation attacks", 2025.
- [25] T. Alter, R. Lapid, and M. Sipper, "On the robustness of Kolmogorov-Arnold networks: An adversarial perspective", 2024.
- [26] A. D. M. Ibrahim, Z. Shang, and J.-E. Hong, "How resilient are Kolmogorov-Arnold networks in classification tasks? A robustness investigation", Applied Sciences, vol. 14, no. 22, 2024.
- [27] H. Cao, "An efficient implementation of Kolmogorov-Arnold network (KAN)", 2024, [Online]. Available: <https://github.com/Blealtan/efficient-kan>. Accessed: 14 May 2025.
- [28] G. Noesis, "Pytorch layer for FourierKAN", 2024, [Online]. Available: <https://github.com/GistNoesis/FourierKAN/tree/main>. Accessed: 14 May 2025.
- [29] SpaceLearner, "Jacobi polynomials KAN", 2024, [Online]. Available: <https://github.com/SpaceLearner/JacobiKAN>. Accessed: 14 May 2025.
- [30] SynodicMonth, "Chebyshev polynomials KAN", 2024, [Online]. Available: <https://github.com/SynodicMonth/ChebyKAN/>. Accessed: 14 May 2025.
- [31] Z. Liu, "Python Kolmogorov-Arnold networks (KANs)", 2024, [Online]. Available: <https://github.com/KindXiaoming/pykan>. Accessed: 14 May 2025.