Evaluating the Robustness of Kolmogorov-Arnold Networks against Noise and Adversarial Attacks

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

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

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

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

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

***Abstract*—Kolmogorov-Arnold Networks (KANs) is a new perspective direction in Machine Learning (ML) domain. KANs use spline functions to enhance interpretability and adaptability of the ML models. However, their robustness against Adversarial Attacks (AAs) has not been fully researched. This paper aims to address this gap by evaluating KAN performance under Gaussian noise and AAs, by using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks. The objective of this paper is to assess the comparative robustness of KANs and Multi-Layer Perceptrons (MLPs) when exposed to Gaussian noise and adversarial attacks, aiming to identify areas of improvement for KANs and to provide insights into their performance under real-world, noisy conditions. The results show that KANs achieve higher accuracy than MLPs in a clean environment. At the same time, KANs demonstrate noticeable reduction in accuracy under conditions where increased noise and adversarial perturbations are present. KANs experience a more substantial accuracy drop under FGSM and PGD attacks compared to MLPs, which reveals critical areas for improvement and further research. The sensitivity of KANs to Gaussian noise further highlights their limitations in real-world scenarios. These findings underscore the need for further research to develop more resilient KAN architectures and better understand their role in secure ML systems.**

Keywords-Kolmogorov-Arnold Network, KAN, MLP, FGSM, PGD, MNIST, Classification.

I. Introduction

The rapid advancement of Machine Mearning (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 bring a promise to enhance models' interpretability and flexibility. However, the robustness of KANs, particularly to Adversarial Attacks (AAs) and noisy data, has not been thoroughly researched.

Traditional Multi-Layer Perceptrons (MLPs) often struggle with capturing complex nonlinear relationships due to their reliance on fixed activation functions and linear weight matrices. This can lead to limitations in model flexibility and interpretability, making them less effective in handling intricate patterns present in real-world data. Moreover, MLPs can be vulnerable to overfitting and may not generalize well to unseen data, especially under adversarial conditions or noise. To address these challenges, KANs introduce learnable activation functions on edges, replacing static weights with parameterized functions. This architectural shift enhances the model's ability to capture complex, nonlinear relationships, offering improved flexibility and interpretability over traditional MLPs [1].

Robustness of the ML models is an important quality for real-world applications, often characterized by suboptimal conditions. AAs, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), exploit the models' vulnerabilities by making small perturbations to input data, while noise can obscure key features and degrade performance.

This paper explores the robustness of KANs compared to MLPs, focusing on their resilience to Gaussian noise and AAs like FGSM and PGD. The goal of this paper is to assess the security and practical limitations of KANs by comparing their performance under various perturbations. All experiments are conducted using the Modified National Institute of Standards and Technology dataset (MNIST) [2], [3], a widely recognized benchmark for evaluating image classification models.

MLPs are selected as a benchmark for comparison because they represent one of the most widely used and established neural network architectures in machine learning. Their simplicity, effectiveness in various tasks, and resistance to adversarial conditions provide a useful baseline for evaluating the performance and robustness of newer, more complex architectures like KANs. Focusing on KANs provides an opportunity to assess a novel architecture that could potentially address some limitations of traditional models, thereby justifying its selection over other alternatives.

The primary objective of this paper is to systematically assess the robustness of KANs compared to traditional MLPs under adversarial conditions. Specifically, this study aims to evaluate how KANs and MLPs perform when exposed to Gaussian noise and AAs, such as the FGSM and PGD. By comparing their resilience across key metrics such as accuracy, precision, recall, and F1-score using the MNIST dataset, the paper seeks to identify the strengths and limitations

of KANs in real-world, noisy environments. The findings aim to inform further research and development of more robust KAN architectures for secure machine learning systems.

The remainder of this paper is organized as follows: Section II reviews related work and existing approaches in the field. Section III outlines the methodology, including the experimental setup and evaluation metrics. Section IV presents the results of the experiments, followed by a discussion in Section V. Finally, Section VI concludes the paper and suggests directions for future research.

II. Related Work

KANs are a new neural architecture based on the Kolmogorov-Arnold representation theorem, offering an alternative to traditional MLPs. Instead of fixed activation functions on nodes and linear weight matrices, KANs use learnable activation functions on edges, with each weight replaced by a 1D learnable function parameterized as a spline. This new architecture promises improved accuracy and interpretability compared to traditional MLPs and can find application in various domains [1]. Figure 1 provides a comparative visualization of KANs and MLPs. On the left, the KAN architecture is shown, where the edges represent learnable activation functions, unlike traditional networks. In KANs, nodes perform sum operations across these learned functions, enabling greater flexibility and non-linearity. On the right, the MLP architecture is depicted, where the edges correspond to learnable weights and fixed activation functions are applied at each node. This distinction highlights the novel design of KANs, which replace static activations with adaptive spline functions, allowing for potentially better handling of complex relationships in data compared to MLPs.

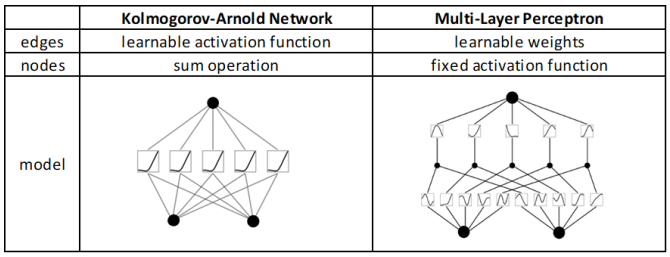

Figure 1. KAN and MLP Architecture, derived from [1]

KANs have been explored in computer vision, where they have been compared to architectures like MLP-Mixer, CNNs, and Vision Transformers. Studies [4]–[6] demonstrate that KANs can achieve competitive accuracy on datasets like CIFAR10 and MNIST while offering benefits in computational efficiency and parameter reduction. However, they sometimes fall short compared to models like ResNet-18, indicating both their potential and limitations.

In time series analysis, KANs have been applied to capture complex temporal patterns and enhance model interpretability. Models like Temporal Kolmogorov-Arnold Transformer (TKAT) [7], Temporal Kolmogorov-Arnold Networks (T-KAN), and Multivariate Temporal Kolmogorov-Arnold Networks (MT-KAN) [8] have shown improved performance in handling multivariate data streams and detecting concept drift. These studies highlight KANs' adaptability and efficiency, particularly in forecasting tasks [9].

However, KANs have limitations that were discussed in several studies. For instance, [10] compares KANs with MLPs and finds that KANs do not always outperform MLPs. KANs can fall behind when dealing with irregular or noisy functions. Both models struggle with noise, and while increasing training data helps, KANs often match, rather than surpass, MLPs in such noisy conditions. [11] further emphasizes KANs' sensitivity to noise, showing that even small amounts can significantly degrade performance. Although oversampling and denoising techniques can mitigate these issues the increased computational cost can become a limiting factor for the practical applications of KANs. Additionally, [12] explores KANs in hardware applications, finding that they fall short of MLPs in complex datasets and require more hardware resources. [5] also concludes that benefits of KANS for more complex datasets like CIFAR-10 are not evident. While KANs excel in capturing complex patterns and promise improvements in interpretability, their vulnerability to noise and ability to handle more complex tasks raises concerns about potential susceptibility to AAs, an area yet to be thoroughly explored.

Recent developments in AAs have focused on refining techniques that exploit vulnerabilities in machine learning models [13]–[16], particularly in computer vision [17], [18]. Two well-researched methods for evaluating model robustness are the FGSM and PGD. FGSM, introduced by [19], generates adversarial examples through small perturbations to input data, which can cause models to make incorrect predictions. PGD, a more iterative and sophisticated method, was introduced by [20] and is now a benchmark for testing resilience against stronger attacks. These methods are particularly impactful in computer vision, where minor input changes can lead to significant shifts in model outputs [21]. Defenses against FGSM and PGD have been explored [22]–[24]. Despite these advancements, FGSM and PGD remain the standard for assessing robustness of ML models.

Tools like the Adversarial Robustness Toolbox (ART) provide methods for crafting adversarial examples and defenses [25], and datasets such as MNIST [2] are frequently used for benchmark adversarial vulnerability across studies. While KANs have been applied to various domains, their resistance to AAs, particularly FGSM and PGD, is largely unexplored. This paper aims to fill that gap, enhancing the understanding of KANs' robustness in adversarial settings.

III. Methodology, Tools and Environment

Methodology: The experiments compare two machine learning models: a KAN and a traditional MLP-based feedforward classifier. Both are trained on the MNIST dataset to maintain consistency. Their robustness is evaluated through various performance metrics under different conditions, including noise and AAs. Model architectures and pre-trained weights remain

unchanged throughout the experiments, with only the test data being manipulated for assessment.

Both models are trained using the MNIST dataset [2], [3], which contains samples of handwritten digits. Initial evaluations are conducted on unaltered test data to set a performance baseline. Metrics like accuracy, confusion matrices, precision, recall, and F1-scores are used to assess both models. The same evaluation approach is maintained across all experiments to ensuring uniformity.

Experiments: To test noise sensitivity, Gaussian noise with mean zero and varying standard deviations is added to the test data, with noise levels increasing from 10 to 90 in increments of 10. These standard deviation values represent the intensity of the noise added, simulating conditions from mild to severe distortion. The noise is added to each pixel in the images, introducing variability that can blur edges and obscure important features necessary for accurate classification. This method measures the models' ability to maintain accuracy as the noise level increases, providing insights into each model's robustness in noisy environments To test robustness against FGSM attack, adversarial examples are created by introducing perturbations in input data and applied to both models. The attack's strength is controlled by the epsilon parameter, ranging from 0.1 to 0.8. Models' accuracy degradation is tracked as epsilon increases, showing each model's vulnerability to FGSM attacks. Similarly, the PGD attack, a more iterative adversarial method, is tested with epsilon values between 0.1 and 0.8. This reveals how both models handle stronger attacks. Once the experiments on noise, FGSM, and PGD attacks are completed the results are aggregated to compare the robustness of KAN and the MLP. Table I summarizes a performance comparison of MLP and KAN models under different scenarios. A discussion follows, highlighting key performance strengths and weaknesses under different conditions and further research directions are suggested.

Tools and environment: The MNIST dataset is used throughout the experiments [2], [3], while adversarial examples are generated using the Adversarial Robustness Toolbox (ART) [25]. The KAN implementation is sourced from GitHub repositories [26], [27] and the MLP is implemented using PyTorch and Scikit-learn python libraries. These standardized tools ensure the experiments' reproducibility and reliability.

Models Architectures

The MLP implementation is a feedforward neural network with five hidden layers, each followed by ReLU activation and dropout for regularization. The input is a flattened 28x28 pixel image, resulting in 784 features. The layers progressively reduce in size from 512 to 64 neurons, and the output layer contains 10 neurons for the digit classes. The model uses the AdamW optimizer with a learning rate of 0.001 and weight decay to prevent overfitting. This regularization penalizes larger weight values and encourages the model to maintain smaller weights, which helps prevent overfitting by reducing model complexity and improving generalization to unseen data. An Exponential Learning Rate Scheduler adjusts the learning rate during training, and the CrossEntropyLoss function is used for classification.

The KAN implementation [27] leverages spline-parametrized univariate functions instead of traditional activations, based on the Kolmogorov-Arnold theorem. It begins with a 784-feature input layer, followed by two KANLinear layers, which transform the features using spline-based activations. The first KANLinear layer outputs 1569 units, while the second produces 10 units corresponding to the digit classes. The model includes customizable spline parameters like grid size and spline order, which allow it to learn complex functions. Regularization techniques, including activation and entropy penalties, are applied to maintain model stability. Like the MLP, KAN is optimized with AdamW and trained with the CrossEntropyLoss function, with the learning rate dynamically adjusted by an Exponential Scheduler. Figure 2 provides architectural diagrams for MPL and KAN implementations.

IV. Results

Default models: In the initial experiment on the clean MNIST test dataset, both the MLP and KAN performed similarly well. The MLP achieved an accuracy of 97.40%, while KAN outperformed it with 97.95%. Both models showed comparable precision, recall, and f1-scores, with minor differences in misclassifications (Table II). KAN's slight edge in accuracy and recall suggests better handling of data variability.

Gaussian Noise: When exposed to noisy data, both models showed accuracy degradation as noise levels increased (Figure 3). At a noise level of 90, the MLP showed 94.77% accuracy, while KAN's performance dropped to 88.21%. The MLP showed stronger resistance to noise, achieving higher precision, recall, and f1-scores across all digits. KAN particularly struggled with digits 1 and 8, where performance dropped drastically (Table III). At lower noise levels, KAN outperformed the MLP, but its accuracy deteriorated more quickly at higher noise levels.

Fast Gradient Sign Method: Under the FGSM AA, both models experienced accuracy declines as epsilon increased (Figure 4). At epsilon 0.3, the MLP showed 92.48% accuracy, while KAN's dropped to 63.87%. The MLP retained more consistent precision, recall, and f1-scores, while KAN saw sharp declines, particularly with digits 1 and 8 (Table IV). KAN's accuracy fell rapidly as epsilon increased, indicating a greater vulnerability to adversarial attacks compared to the MLP, which maintained resilience until epsilon values grew larger.

Projected Gradient Descent: During the PGD attack, a more iterative adversarial method, both models again showed performance declines (Figure 5). At epsilon 0.3, the MLP showed 96.29% accuracy, while KAN dropped significantly to 53.12%. The MLP exhibited strong overall precision, recall, and f1-scores, while KAN struggled significantly, particularly with digits 1 and 8, where precision and recall dropped sharply (Table V). KAN's performance deteriorated more rapidly than the MLP as the epsilon value increased emphasizing its greater vulnerability to stronger adversarial attacks.

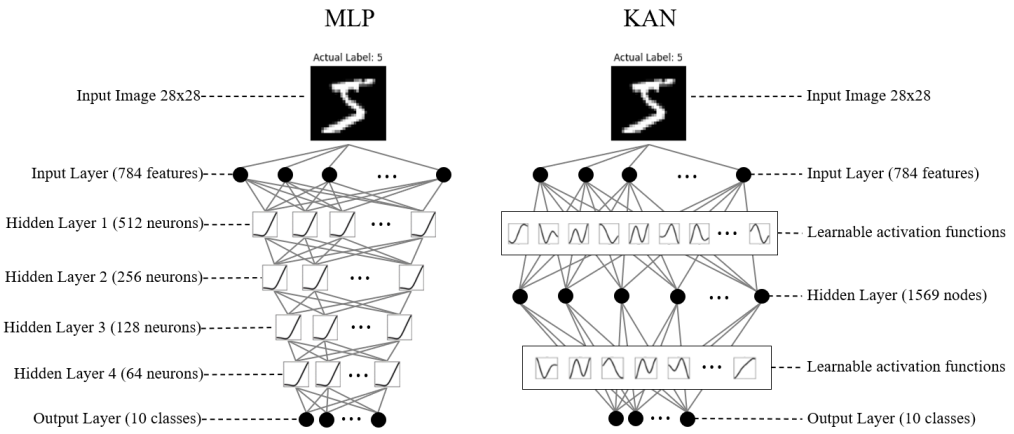

Figure 2. MPL and KAN Implementations

TABLE I
PERFORMANCE COMPARISON OF MLP AND KAN MODELS UNDER DIFFERENT SCENARIOS

Scenario	MLP				KAN			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Default models	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
Gaussian Noise (Level 30)	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Gaussian Noise (Level 60)	0.97	0.96	0.97	0.96	0.96	0.96	0.96	0.96
Gaussian Noise (Level 90)	0.95	0.95	0.95	0.95	0.89	0.92	0.89	0.89
FGSM (eps. 0.1)	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96
FGSM (eps. 0.3)	0.92	0.93	0.92	0.92	0.64	0.77	0.64	0.64
FGSM (eps. 0.6)	0.65	0.67	0.65	0.66	0.20	0.43	0.20	0.19
PGD (eps. 0.1)	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97
PGD (eps. 0.3)	0.96	0.96	0.96	0.96	0.53	0.68	0.53	0.55
PGD (eps. 0.6)	0.42	0.46	0.42	0.42	0.07	0.08	0.07	0.02

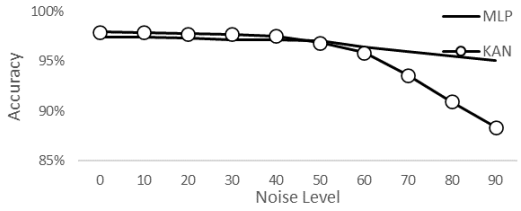

Figure 3. Gaussian Noise: Accuracy by Noise Level

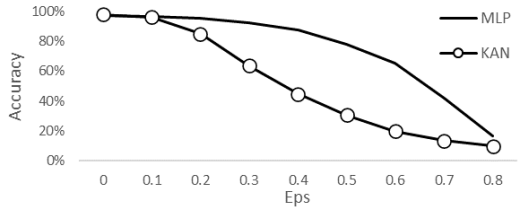

Figure 4. FGSM: Accuracy by Eps Level

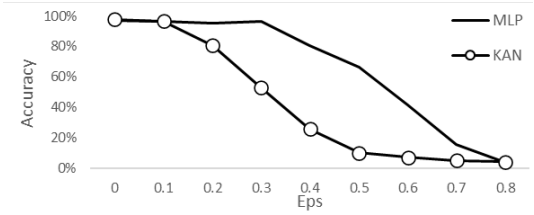

Figure 5. PGD: Accuracy by Eps Level

Discussion: Across all experiments, KAN displayed greater sensitivity to noise and adversarial attacks especially in more challenging conditions, while the MLP showed more stable performance and resilience. The vulnerabilities of KANs to noise and adversarial attacks could be linked to their reliance on spline-based transformations, which may be more sensitive to perturbations compared to the simpler linear activations used in MLPs. KANs' flexibility in modeling complex functions might lead to overfitting, making them less robust when faced with data that deviates from the training distribution, such as noisy inputs or adversarial perturbations. The spline functions used in KANs may also be more prone to distortions from small input changes, explaining their susceptibility to adversarial attacks. Additionally, KANs' complexity might hinder their ability to generalize well in adversarial scenarios, where simpler MLP structures could offer more stability.

V. CONCLUSION AND FUTURE WORK

The results of the experiments show clear differences in how KANs and classic MLPs handle AAs and noise. In clean conditions, both models perform similarly, with KAN slightly outperforming the MLP in accuracy (97.95% vs. 97.40%) and

TABLE II
PERFORMANCE COMPARISON: DEFAULT MODELS.

Class	MLP			KAN		
	Precision	Recall	F1-score	Precision	Recall	F1-score
0	0.99	0.99	0.99	0.99	0.99	0.99
1	0.98	0.99	0.98	0.98	0.99	0.99
2	0.98	0.96	0.97	0.98	0.98	0.98
3	0.97	0.96	0.97	0.98	0.97	0.97
4	0.98	0.97	0.98	0.98	0.98	0.98
5	0.96	0.97	0.97	0.97	0.97	0.97
6	0.98	0.99	0.98	0.98	0.99	0.98
7	0.98	0.97	0.97	0.98	0.98	0.98
8	0.97	0.97	0.97	0.98	0.97	0.98
9	0.96	0.96	0.96	0.96	0.98	0.97

TABLE III
PERFORMANCE COMPARISON: MODELS EXPOSED TO NOISE LEVEL 90.

Class	MLP			KAN		
	Precision	Recall	F1-score	Precision	Recall	F1-score
0	0.98	0.98	0.98	0.98	0.98	0.98
1	0.98	0.97	0.97	1.00	0.58	0.73
2	0.95	0.95	0.95	0.92	0.95	0.94
3	0.93	0.95	0.94	0.94	0.92	0.93
4	0.95	0.94	0.94	0.95	0.91	0.93
5	0.91	0.94	0.93	0.96	0.88	0.92
6	0.97	0.97	0.97	0.97	0.94	0.96
7	0.95	0.96	0.95	0.98	0.85	0.91
8	0.94	0.92	0.93	0.56	0.99	0.72
9	0.92	0.91	0.92	0.90	0.88	0.89

TABLE IV
PERFORMANCE COMPARISON: MODELS EXPOSED TO FGSM, EPS0.3.

Class	MLP			KAN		
	Precision	Recall	F1-score	Precision	Recall	F1-score
0	0.98	0.99	0.98	0.97	0.91	0.94
1	0.98	0.98	0.98	0.67	0.02	0.03
2	0.97	0.94	0.96	0.91	0.78	0.84
3	0.91	0.92	0.91	0.81	0.79	0.80
4	0.86	0.89	0.88	0.77	0.64	0.70
5	0.85	0.95	0.90	0.77	0.56	0.65
6	0.96	0.96	0.96	0.92	0.82	0.87
7	0.92	0.91	0.92	0.96	0.41	0.58
8	0.88	0.88	0.88	0.24	0.96	0.38
9	0.90	0.80	0.85	0.58	0.61	0.60

TABLE V
PERFORMANCE COMPARISON: MODELS EXPOSED TO PGD, EPS0.3.

Class	MLP			KAN		
	Precision	Recall	F1-score	Precision	Recall	F1-score
0	0.98	1.00	0.99	0.96	0.94	0.95
1	1.00	0.98	0.99	0.00	0.00	0.00
2	1.00	1.00	1.00	0.88	0.64	0.74
3	0.93	0.98	0.96	0.90	0.67	0.77
4	0.90	0.97	0.94	0.68	0.34	0.46
5	0.94	0.91	0.93	0.78	0.41	0.54
6	0.99	1.00	0.99	0.90	0.71	0.80
7	0.97	0.93	0.95	1.00	0.33	0.49
8	0.93	0.98	0.95	0.17	0.98	0.29
9	0.96	0.84	0.90	0.52	0.43	0.47

showing marginally better metrics overall. However, KANs struggle when noise is introduced. As noise levels increase, KANs experience a sharper drop in accuracy compared to the MLP. This indicates that while KANs perform better in clean environments, they suffer accuracy degradation under noisy conditions, revealing a weakness in robustness in real-life scenarios. Under FGSM and PGD AAs, KANs demonstrate even greater vulnerability. Their accuracy declines much faster than that of the MLP as the epsilon value rises. For example, at epsilon 0.3, KANs' accuracy falls to 63.87%, while the MLP still shows 92.48%. This trend continues with increasing perturbations, showing that KANs are more vulnerable to AAs than the MLP. Although KANs show high performance in optimal conditions, they face challenges in robustness and security. Their rapid decline in accuracy under noise and adversarial conditions suggests they are more vulnerable than traditional models. This poses risks in security-sensitive applications, where resilience against such attacks is crucial.

Future Research Directions

Improving KANs' robustness could involve exploring advanced regularization methods, adversarial training, or defense mechanisms tailored for KANs. Additionally, designing architectures that better handle noisy inputs and conducting more comprehensive security analyses across diverse attacks and datasets would further enhance KANs' resilience and security.

The future research directions include:

- Investigating and developing advanced robustness techniques tailored for KANs. This may include exploring novel regularization methods, adversarial training, or defensive strategies specifically designed to improve KANs' resilience to noise and AAs. Trying different types of activation functions could provide new insights into improving model performance and robustness. Activation functions based on Fourier transforms, for instance, can capture periodic patterns in data, while Chebyshev and Jacobi polynomials might offer superior approximation capabilities for certain types of functions. Investigating these alternatives could lead to the development of KANs that are more resilient to noise and adversarial attacks by leveraging the mathematical properties of these functions.
- Designing KAN architectures that inherently handle noisy inputs better. This might involve incorporating noise-robust activation functions or more sophisticated noise-handling mechanisms within the network. One approach could be to employ regularization methods that penalize overly sensitive spline functions, making the network less reactive to small perturbations in the input. Another strategy is to integrate preprocessing steps or layers within the KAN that filter out noise before it propagates through the network.
- Conducting a thorough security analysis of KANs across a broader range of AA methods and datasets is imperative, as it can provide deeper insights into KANs vulnerabilities and help in devising more effective defense strategies. Future work will involve testing KANs against more

sophisticated attacks like the Carlini & Wagner attack, DeepFool, and black-box attacks to evaluate their robustness comprehensively. Additionally, experimenting with diverse datasets such as CIFAR-10, ImageNet, or domain-specific datasets will help assess the generalizability of KANs' resilience across different types of data.

REFERENCES

[1] Z. Liu *et al.*, *Kan: Kolmogorov-arnold networks*, retrieved: September 2024, Apr. 2024.

[2] L. Deng, "The mnist database of handwritten digit images for machine learning research", *IEEE Signal Processing Magazine*, vol. 29, pp. 141–142, 6 2012, ISSN: 10535888. DOI: 10.1109/MSP.2012.2211477.

[3] OpenML, "Mnist handwritten digit dataset", retrieved: September 2024, [Online]. Available: https://www.openml.org/d/554.

[4] M. Cheon, *Demonstrating the efficacy of kolmogorov-arnold networks in vision tasks a preprint*, retrieved: September 2024, 2024.

[5] B. Azam and N. Akhtar, *Suitability of kans for computer vision: A preliminary investigation*, retrieved: September 2024, Jun. 2024.

[6] A. D. Bodner, J. N. Spolski, A. S. Tepsich, and S. Pourteau, *Convolutional kolmogorov-arnold networks*, retrieved: September 2024, 2024.

[7] R. Genet and H. Inzirillo, *A temporal kolmogorov-arnold transformer for time series forecasting*, retrieved: September 2024, Jun. 2024.

[8] K. Xu, L. Chen, and S. Wang, *Kolmogorov-arnold networks for time series: Bridging predictive power and interpretability*, retrieved: September 2024, Jun. 2024.

[9] C. J. Vaca-Rubio, L. Blanco, R. Pereira, and M. Caus, *Kolmogorov-arnold networks (kans) for time series analysis*, retrieved: September 2024, May 2024.

[10] C. Zeng, J. Wang, H. Shen, and Q. Wang, *Kan versus mlp on irregular or noisy functions*, retrieved: September 2024, 2024.

[11] H. Shen, C. Zeng, J. Wang, and Q. Wang, *Reduced effectiveness of kolmogorov-arnold networks on functions with noise*, retrieved: September 2024, Jul. 2024.

[12] V. D. Tran *et al.*, *Exploring the limitations of kolmogorov-arnold networks in classification: Insights to software training and hardware implementation*, retrieved: September 2024, Jul. 2024.

[16] K. Sadeghi, A. Banerjee, and S. K. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning", *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, pp. 450–467, 4 Aug. 2020, ISSN: 2471285X. DOI: 10.1109/TETCI.2020.2968933.

[17] B. Xi, "Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 12, p. 1511, 5 Sep. 2020, ISSN: 19390068. DOI: 10.1002/wics.1511.

[13] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning", in *Proceedings of the ACM Conference on Computer and Communications Security*, Oct. 2011, pp. 43–58. DOI: 10.1145/2046684.2046692.

[14] 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, 1 Jan. 2022, ISSN: 1557380X. DOI: 10.1177/1548512920951275.

[15] G. Apruzzese, L. Ferretti, M. Colajanni, and M. Marchetti, "Addressing adversarial attacks against security systems based on machine learning", in *2019 11th international conference on cyber conflict (CyCon)*, vol. 900, 2019, pp. 1–18.

[18] G. R. Machado, E. Silva, and R. R. Goldschmidt, *Adversarial machine learning in image classification: A survey towards the defender's perspective*, retrieved: September 2024, Sep. 2020. DOI: 10.1145/3485133.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and harnessing adversarial examples*, retrieved: September 2024, Mar. 2015.

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, *Towards deep learning models resistant to adversarial attacks*, retrieved: September 2024, Jun. 2017.

[21] 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. DOI: 10.3390/bdcc8010008.

[22] 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.

[23] 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.

[24] 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. DOI: 10.5220/0010938400003116.

[25] M.-I. Nicolae *et al.*, *Adversarial robustness toolbox v1.0.0*, retrieved: September 2024, 2019.

[26] Z. Liu, "Kolmogorov-arnold networks (kans)", retrieved: September 2024, [Online]. Available: https://github.com/KindXiaoming/pykan.

[27] H. Cao, "An efficient implementation of kolmogorov-arnold network (kan)", retrieved: September 2024, [Online]. Available: https://github.com/Blealtan/efficient-kan.