KAN vs KAN:
Examining Kolmogorov-Arnold Networks (KAN) Performance Under Adversarial Attacks

Nebojsa Djosic
nebojsa.djosic@torontomu.ca

Evgenii Ostanin
eostanin@torontomu.ca

Fatima Hussain
fatima.hussain@torontomu.ca

Salah Sharieh
salah.sharieh@torontomu.ca

Alexander Ferworn
aferworn@torontomu.ca

Toronto Metropolitan University, Toronto, Canada

Presenter
Nebojsa Djosic

SECURWARE 2024, November 3 - 7, Nice, France

IARIA

Toronto Metropolitan University

Nebojsa Djosic has extensive experience as a software developer, architect, and innovator, working for a few startups and major financial institutions in the Toronto area.

He made key contributions to several peer-reviewed publications and multiple patents in the cybersecurity domain.

Currently, he is pursuing a PhD in Computer Science at Toronto Metropolitan University (TMU).

Toronto Metropolitan University

The authors are members of the Computer Science Department at the Toronto Metropolitan University (formerly Ryerson).

They are actively involved in research and applied projects centered on leveraging Artificial Intelligence and Machine Learning for automation in key domains including cybersecurity, governance, and public safety.

Toronto
Metropolitan
University

Presentation Outline

- What are Kolmogorov-Arnold Networks (KANs)
- Good, Bad, and Different KAN Flavours
- Noise and Adversarial Attacks
- Methodology
- Results
- Conclusions and Future Work

Toronto Metropolitan University

Key Contributions

- Detailed evaluation of robustness
- Performance data on four different KANs
- Comparison of KAN models using different activation functions under Gaussian Noise, FGSM and PGD attacks

Toronto Metropolitan University

Kolmogorov-Arnold Representation Theorem

- Introduced by Andrey Kolmogorov in 1957
- Extended by Vladimir Arnold in 1963

Any multivariate continuous function $f(x_1,....,x_n)$ within a bounded domain can be represented as a superposition of continuous single-variable functions

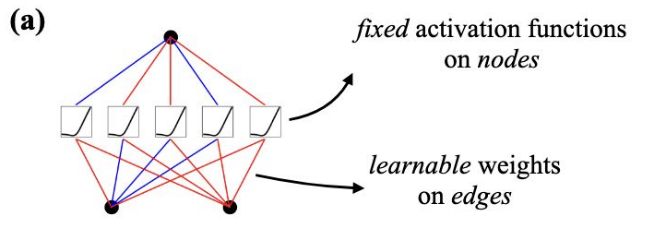
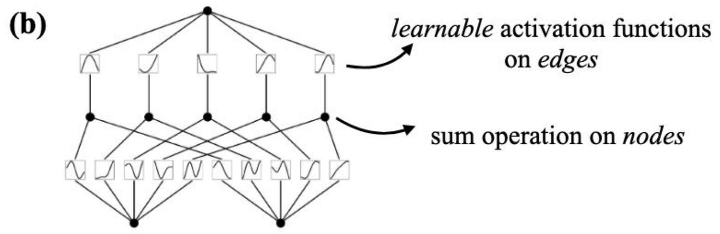
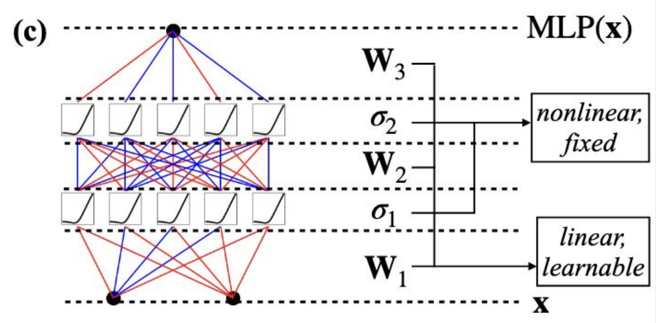
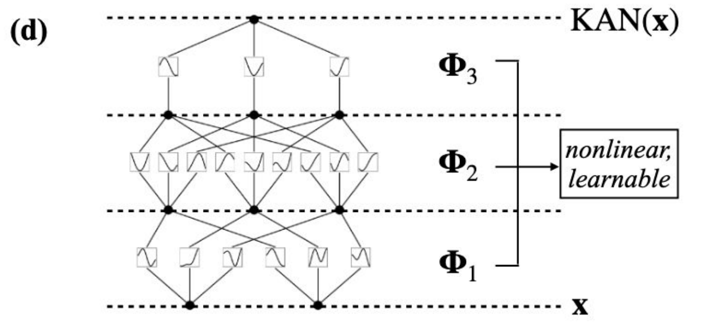
$$f(x) = f(x_1, ..., x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$$

where $\phi_{q,p} : [0,1] \to \mathbb{R}$ and $\Phi_q : \mathbb{R} \to \mathbb{R}$.

Key Insight:
One-dimensional functions can be represented by a B-spline curve and learned during training time!

Toronto Metropolitan University

Kolmogorov-Arnold Network Architecture

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	**Universal Approximation Theorem**	**Kolmogorov-Arnold Representation Theorem**
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$
Model (Shallow)		
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\mathbf{\Phi}_3 \circ \mathbf{\Phi}_2 \circ \mathbf{\Phi}_1)(\mathbf{x})$
Model (Deep)		

Z. Liu et al., Kan: Kolmogorov-arnold networks, 2024.

Good, Bad, Potential

😀 **Good** for computer vision vs MLP-Mixer, Convolutional Neural Networks, Vision Transformers. But not better than ResNet-18.

- M. Cheon, Demonstrating the efficacy of kolmogorov-arnold networks in vision tasks a preprint, 2024.
- B. Azam and N. Akhtar, Suitability of kans for computer vision: A preliminary investigation, 2024.

😞 **Bad** handling noise, hardware intensive applications.

- C. Zeng, J. Wang, H. Shen, and Q. Wang, Kan versus mlp on irregular or noisy functions, 2024
- V. D. Tran et al., Exploring the limitations of kolmogorov-arnold networks in classification: Insights to software training and hardware implementation, 2024

😉 **Show potential** for explainability and incremental, continuous learning, performance on mobile devices.

- Z. Liu et al., Kan: Kolmogorov-arnold networks, 2024.

Methodology: KAN Flavours

Linear KAN uses splines, based on the original pykan: (n * n) * 2 + 1; where n = 28 for MNIST dataset

- Z. Liu et al., Kan: Kolmogorov-arnold networks, 2024.

Naive Fourier KAN replaces splines with one-D Fourier coefficients, grid size 28 x 28 for MNIST dataset

- G. Noesis, “Pytorch layer for fourierkan”, https://github.com/GistNoesis/FourierKAN/tree/main, 2024

Chebyshev KAN replaces splines with Chebyshev polynomials.

- SynodicMonth, “Chebyshev polynomials kan”, https://github.com/SynodicMonth/ChebyKAN/, 2024

Jacobi KAN based on Chebyshev KAN but using orthogonal polynomials, we used special case, the Legendre where **α** and **β** are both 0.

- SpaceLearner, “Jacobi polynomials KAN”, https://github.com/SpaceLearner/JacobiKAN , 2024

Toronto Metropolitan University

Methodology: Adversarial Attacks

Gaussian Noise generation using high level 100.

Fast Gradient Sign Method (FGSM) generates adversarial data using random perturbations e.g. 0.1 to 0.8, higher values may become visible to a naked eye, we used 0.5, mid range, realistic, but still strong to break models. It’s fast, but not furious. Adversarial Robustness Toolbox ART.

- M.-I. Nicolae et al., Adversarial robustness toolbox v1.0.0, 2019.

Projected Gradient Descent (PGD) small, random perturbations progressively probe the model to maximize loss. Considered one of the strongest attacks, but it’s not fast. We used 0.5 level to maintain realistic, imperceptible level. Adversarial Robustness Toolbox ART.

- M.-I. Nicolae et al., Adversarial robustness toolbox v1.0.0, 2019.

Toronto Metropolitan University

Methodology

MNIST dataset with 33,600 training, 8,400 test samples of handwritten digits

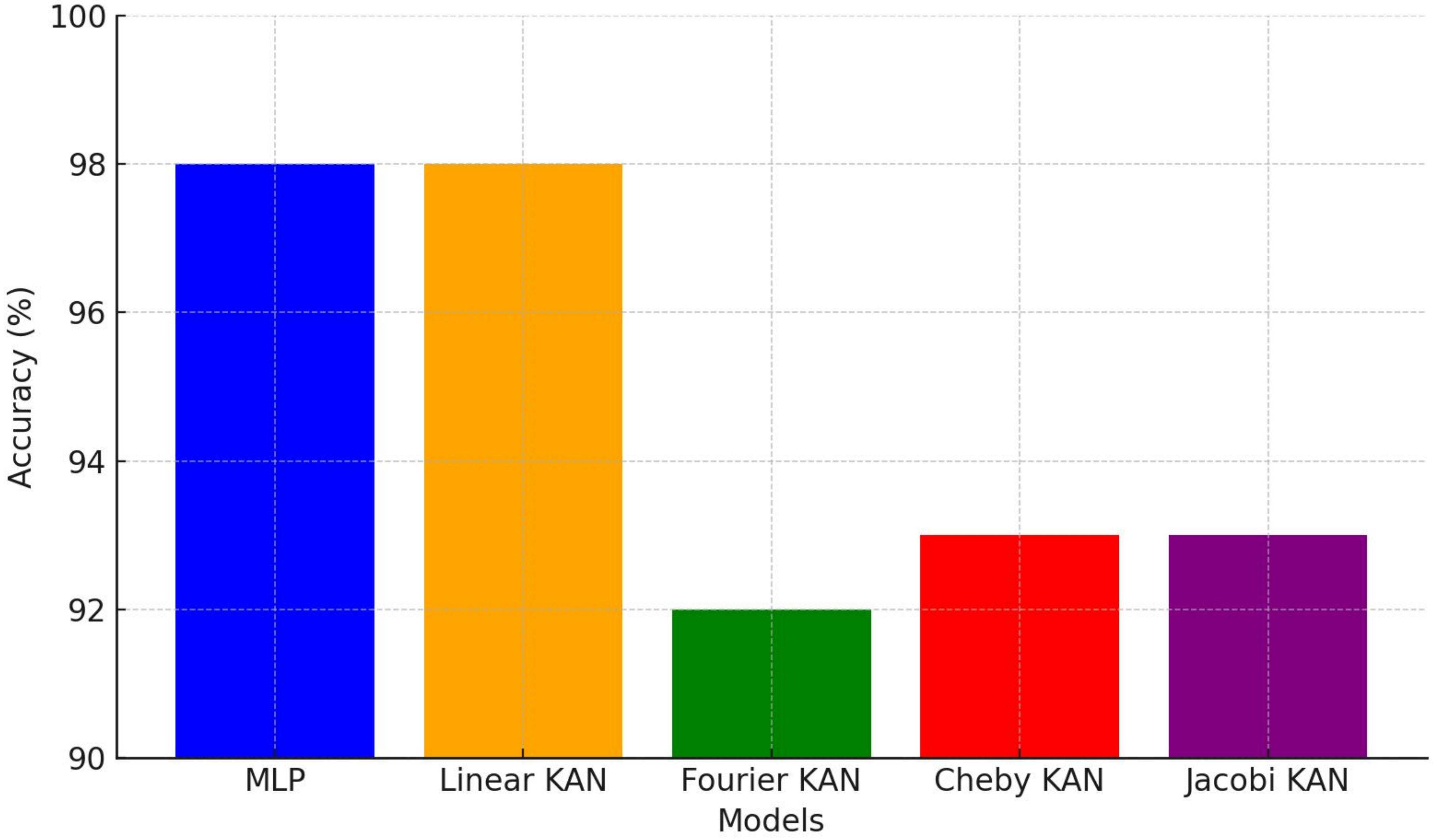
MLP Classifier used as a control: MNIST(28 x 28)→512→256→128→64→10

Adversarial Robustness Toolbox (ART) is a Python module with libraries for generating adversarial attacks

- M.-I. Nicolae et al., Adversarial robustness toolbox v1.0.0, 2019.

Google Colab environment, mostly CPU, and free GPU as much as allowed.

Results: Accuracy Before Attacks



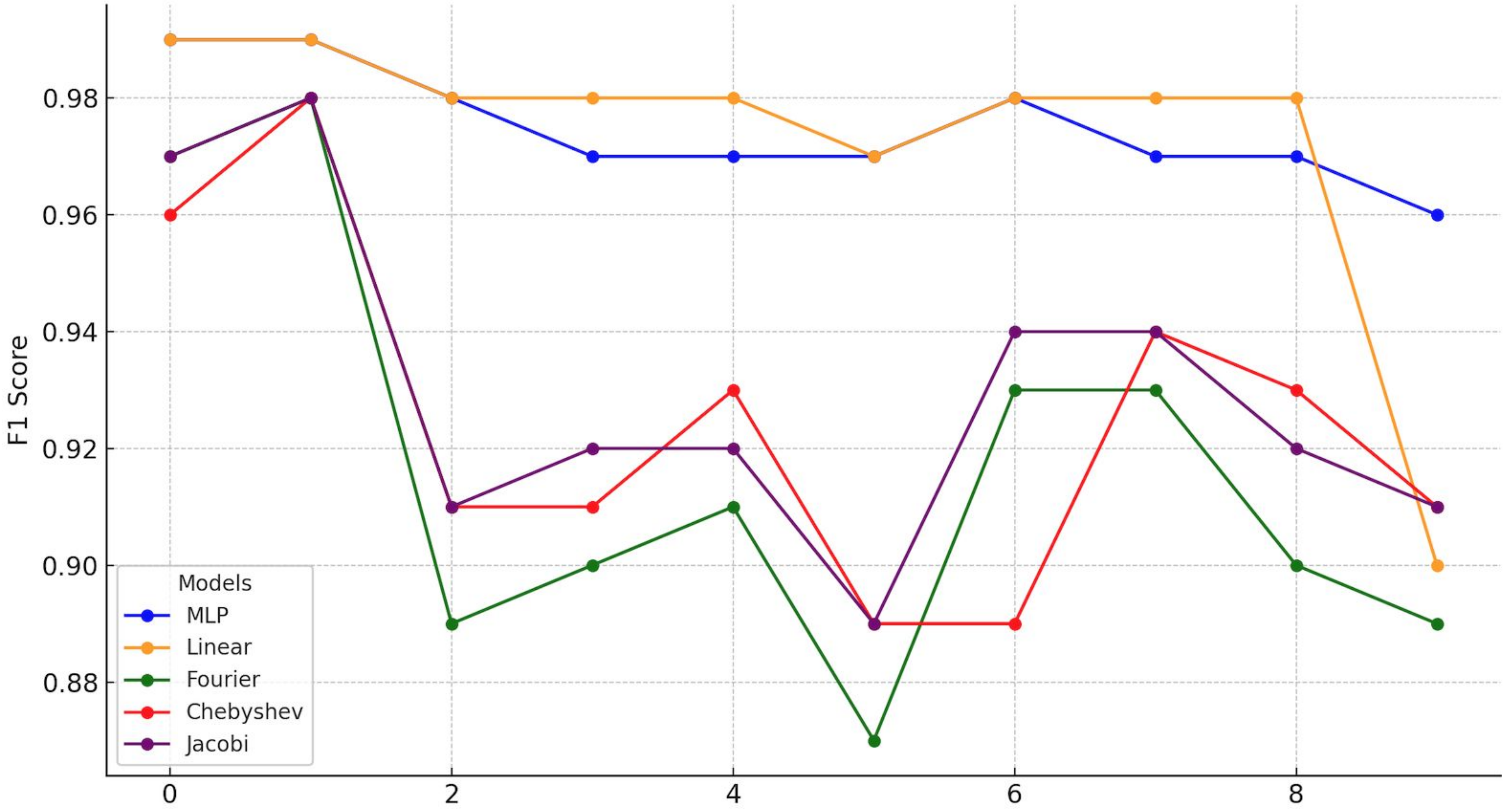
Results: Performance Before Attacks

TABLE I
PERFORMANCE METRICS BEFORE ATTACKS. (ACC. = ACCURACY)

Class	MLP			Linear KAN			Fourier KAN			Cheby KAN			Jacobi KAN		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0	0.99	0.99	0.99	0.99	0.99	0.99	0.96	0.97	0.97	0.96	0.96	0.96	0.96	0.97	0.97
1	0.98	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.98	0.97	0.99	0.98	0.97	0.99	0.98
2	0.98	0.97	0.98	0.99	0.97	0.98	0.91	0.87	0.89	0.92	0.90	0.91	0.91	0.90	0.91
3	0.98	0.96	0.97	0.98	0.97	0.98	0.90	0.91	0.90	0.93	0.90	0.91	0.92	0.91	0.92
4	0.98	0.97	0.97	0.98	0.97	0.98	0.91	0.92	0.91	0.91	0.94	0.93	0.91	0.94	0.92
5	0.97	0.97	0.97	0.97	0.97	0.97	0.88	0.86	0.87	0.89	0.90	0.89	0.90	0.88	0.89
6	0.98	0.99	0.98	0.97	0.99	0.98	0.92	0.95	0.93	0.93	0.95	0.89	0.93	0.96	0.94
7	0.97	0.98	0.97	0.98	0.98	0.98	0.94	0.92	0.93	0.95	0.93	0.94	0.94	0.93	0.94
8	0.97	0.97	0.97	0.98	0.98	0.98	0.90	0.90	0.90	0.92	0.93	0.93	0.92	0.92	0.92
9	0.95	0.97	0.96	0.97	0.97	0.97	0.89	0.90	0.89	0.91	0.91	0.91	0.92	0.91	0.91
Acc.	-	-	0.98	-	-	0.98	-	-	0.92	-	-	0.93	-	-	0.93

Toronto Metropolitan University

Results: F1 Score Before Attacks



Toronto Metropolitan University

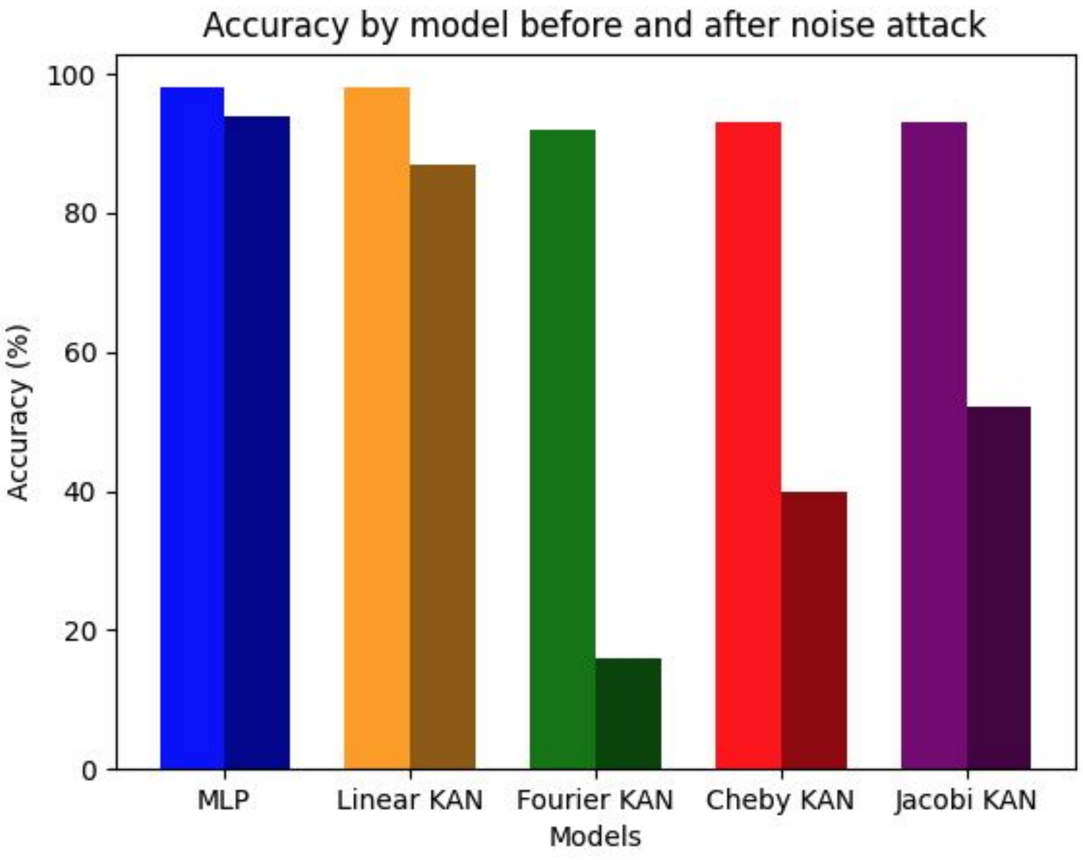
Results: After Noise Attacks Table

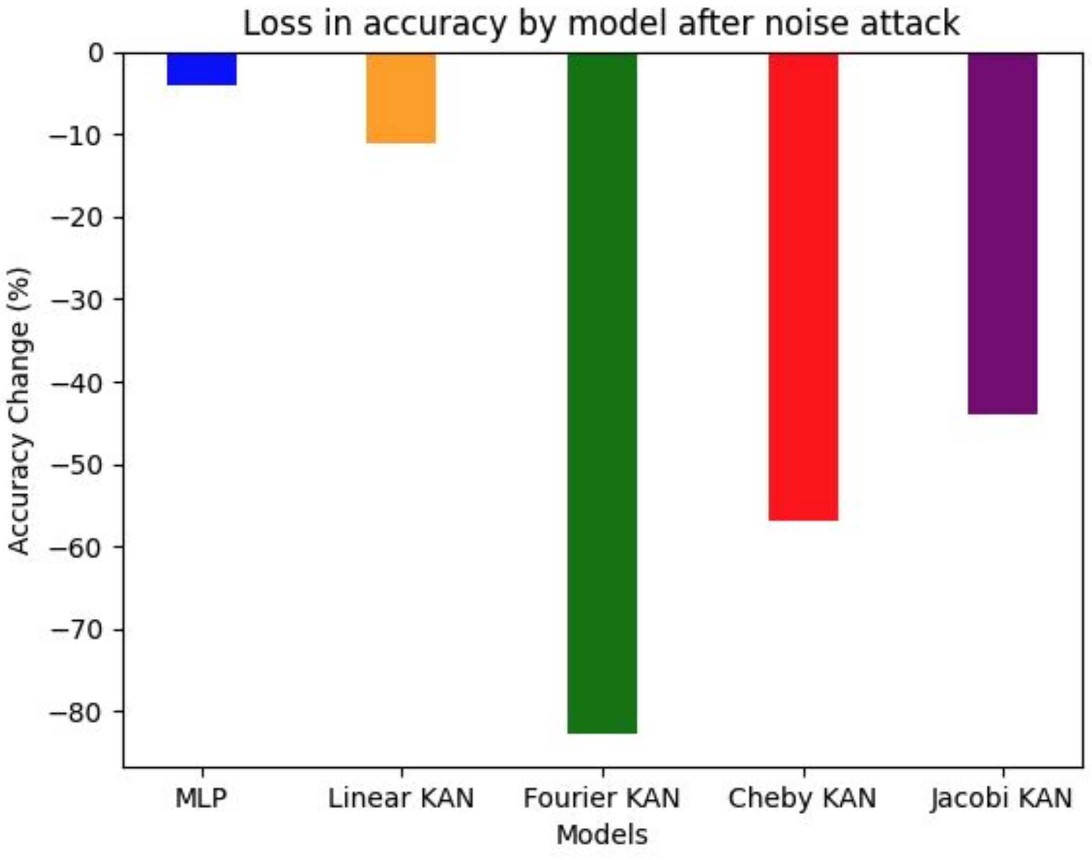
TABLE II
PERFORMANCE METRICS AFTER NOISE ATTACK. (ACC. = ACCURACY)

Class	MLP			Linear KAN			Fourier KAN			Cheby KAN			Jacobi KAN		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
0	0.98	0.98	0.98	0.98	0.97	0.97	0.22	0.11	0.14	0.84	0.44	0.58	0.91	0.55	0.69
1	0.97	0.97	0.97	0.99	0.55	0.71	0.25	0.08	0.12	0.98	0.09	0.16	0.97	0.40	0.57
2	0.94	0.94	0.94	0.90	0.92	0.91	0.14	0.22	0.17	0.28	0.68	0.39	0.38	0.82	0.52
3	0.91	0.93	0.92	0.93	0.91	0.92	0.16	0.21	0.18	0.36	0.57	0.44	0.51	0.57	0.54
4	0.95	0.94	0.95	0.95	0.86	0.90	0.18	0.11	0.14	0.74	0.19	0.30	0.88	0.29	0.44
5	0.91	0.93	0.92	0.95	0.86	0.90	0.11	0.11	0.11	0.30	0.45	0.36	0.26	0.80	0.39
6	0.95	0.97	0.96	0.96	0.96	0.96	0.18	0.15	0.16	0.69	0.38	0.49	0.90	0.44	0.59
7	0.94	0.95	0.94	0.98	0.82	0.89	0.21	0.17	0.19	0.69	0.24	0.35	0.79	0.51	0.62
8	0.94	0.91	0.93	0.53	0.98	0.69	0.14	0.32	0.19	0.30	0.63	0.40	0.53	0.42	0.47
9	0.92	0.90	0.91	0.87	0.88	0.88	0.17	0.11	0.14	0.49	0.32	0.39	0.63	0.44	0.52
Acc.	-	-	0.94	-	-	0.87	-	-	0.16	-	-	0.40	-	-	0.52

Toronto Metropolitan University

Results: After Noise Attacks





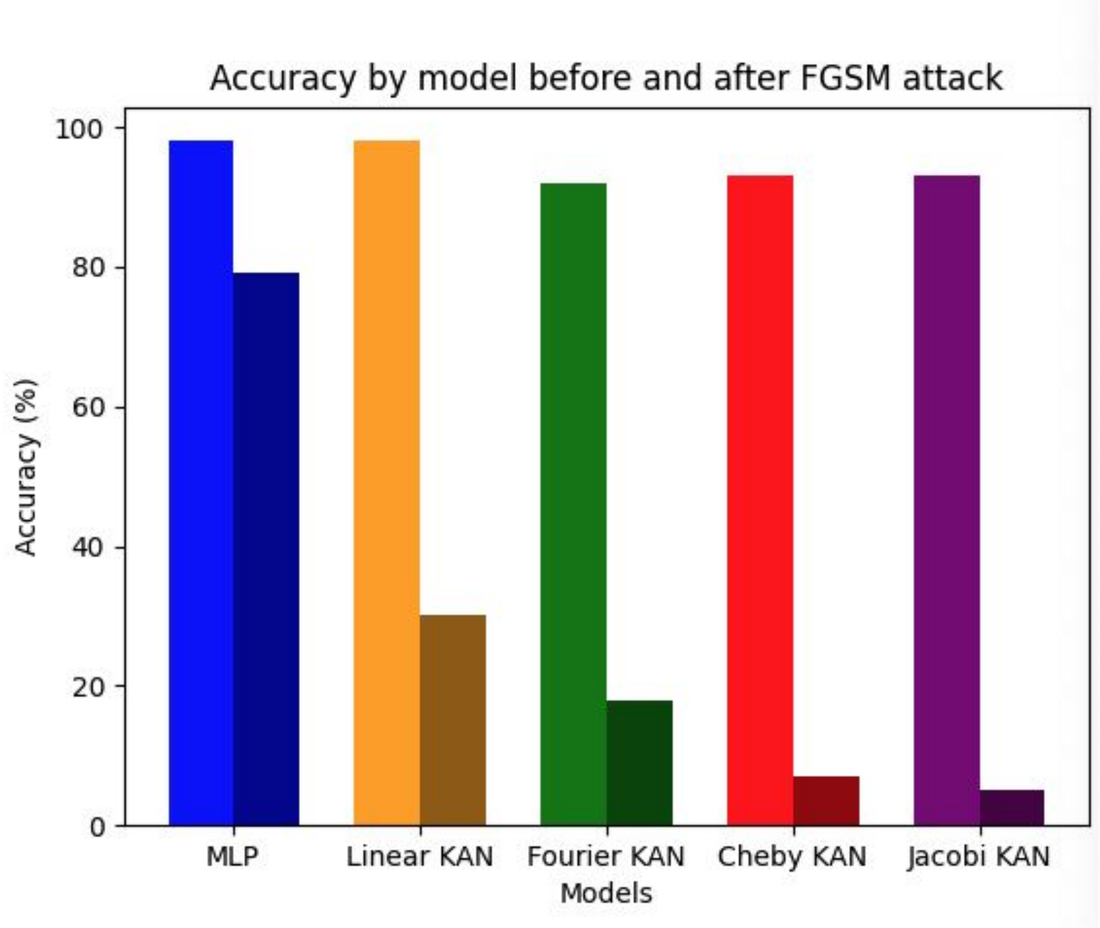
Results: After FGSM Attacks Table

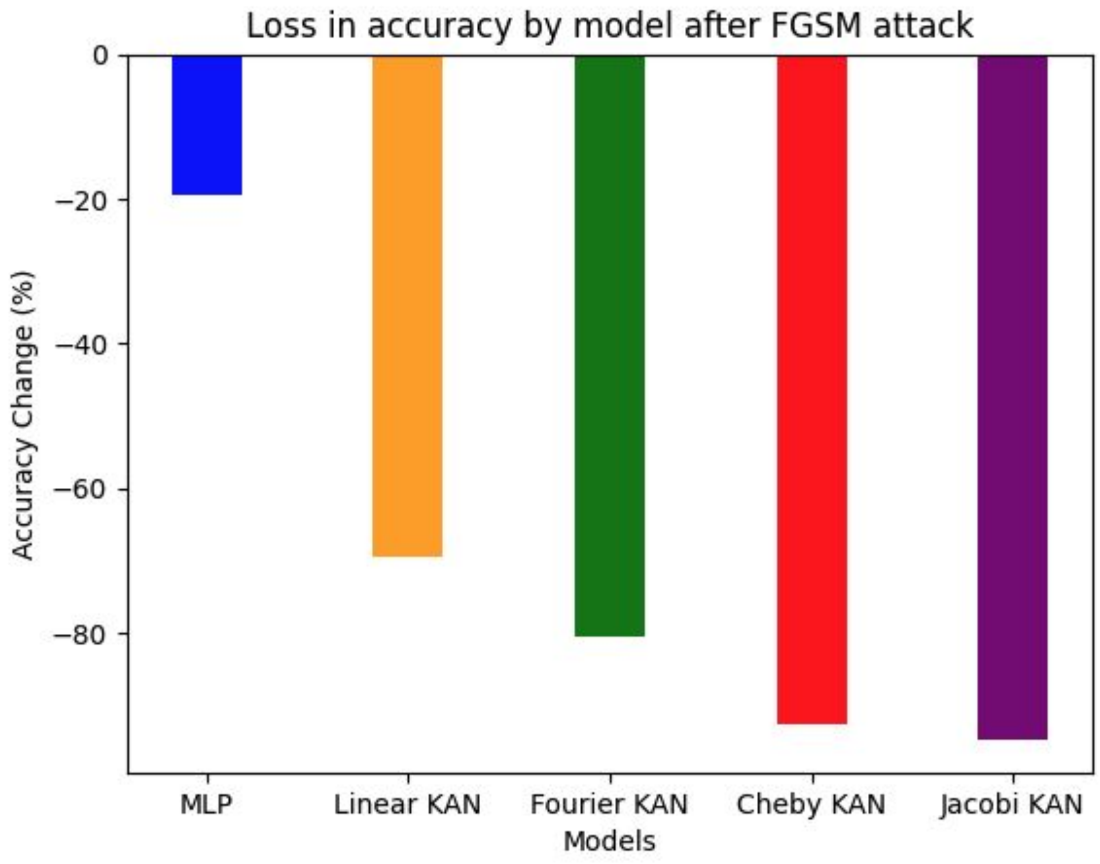
TABLE III
PERFORMANCE METRICS AFTER FGSM ATTACK. (ACC. = ACCURACY)

Class	MLP			Linear KAN			Fourier KAN			Cheby KAN			Jacobi KAN		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0	0.96	0.92	0.94	0.97	0.62	0.75	0.76	0.19	0.31	0.0	0.0	0.0	0.0	0.0	0.0
1	0.93	0.95	0.94	0.0	0.0	0.0	0.32	0.3	0.31	0.0	0.0	0.0	0.0	0.0	0.0
2	0.88	0.77	0.82	0.65	0.38	0.48	0.11	0.73	0.2	0.05	0.04	0.04	0.1	0.13	0.11
3	0.74	0.82	0.78	0.54	0.28	0.37	0.25	0.02	0.03	0.17	0.02	0.04	0.33	0.08	0.13
4	0.67	0.72	0.69	0.35	0.18	0.24	0.26	0.06	0.1	0.0	0.0	0.0	0.0	0.0	0.0
5	0.75	0.81	0.78	0.22	0.11	0.15	0.13	0.05	0.08	0.0	0.0	0.0	0.0	0.0	0.0
6	0.9	0.84	0.87	0.94	0.42	0.58	0.37	0.06	0.1	0.0	0.0	0.0	0.0	0.0	0.0
7	0.78	0.85	0.81	0.5	0.01	0.02	0.37	0.27	0.31	0.0	0.0	0.0	0.0	0.0	0.0
8	0.59	0.6	0.59	0.12	0.94	0.22	0.0	0.0	0.0	0.06	0.73	0.12	0.03	0.32	0.06
9	0.6	0.5	0.55	0.31	0.23	0.26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Acc.	-	-	0.79	-	-	0.3	-	-	0.18	-	-	0.07	-	-	0.05

Toronto Metropolitan University

Results: After FGSM Attacks





Toronto Metropolitan University

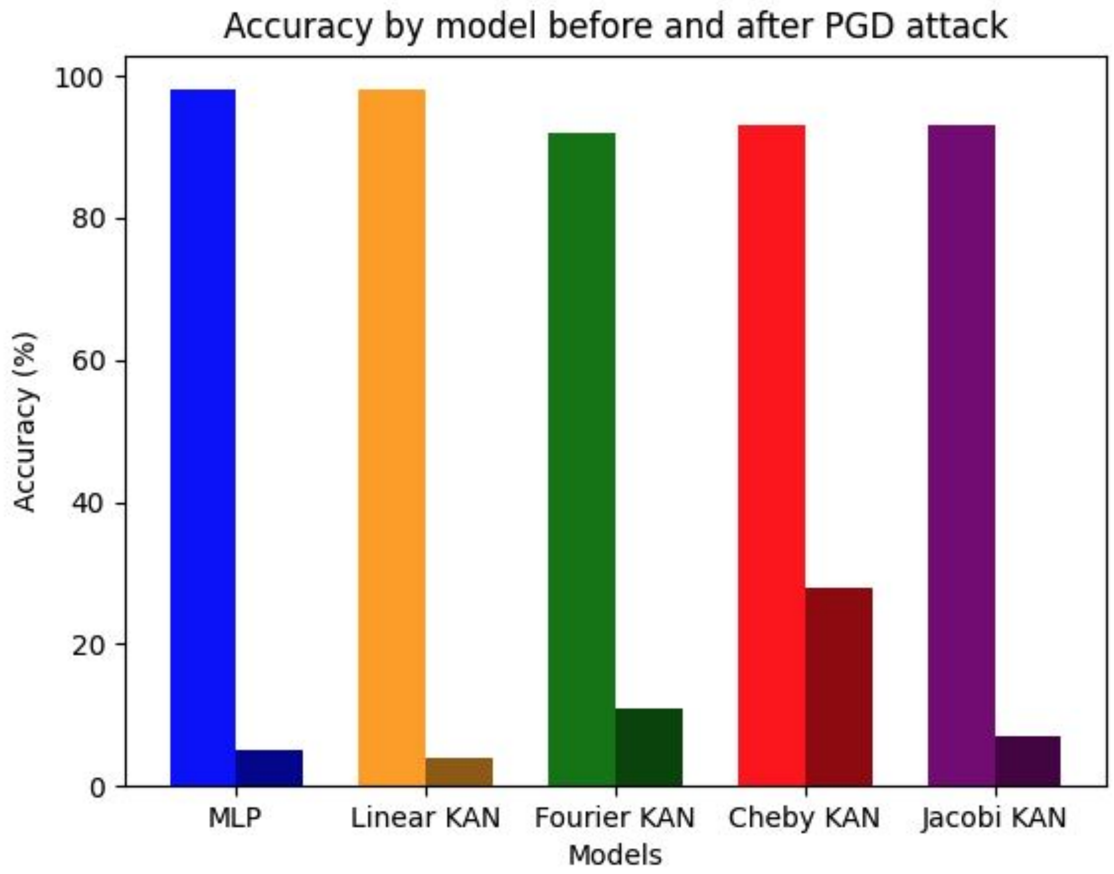
Results: After PGD Attacks Table

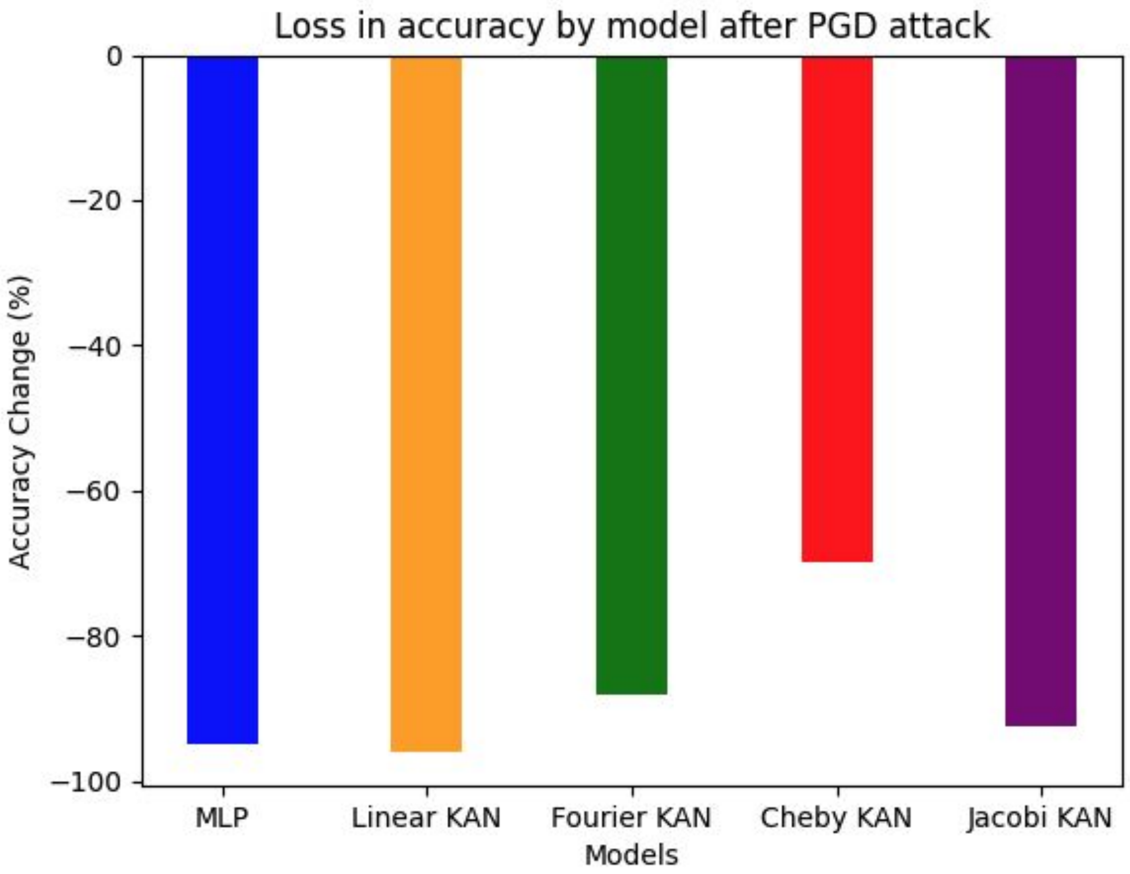
TABLE IV
PERFORMANCE METRICS AFTER PGD ATTACK. (ACC. = ACCURACY)

Class	MLP			Linear KAN			Fourier KAN			Cheby KAN			Jacobi KAN		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0	0.0	0.0	0.0	0.0	0.0	0.0	0.71	0.1	0.17	0.79	0.22	0.34	0.96	0.88	0.92
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.47	0.36	0.41	0.91	0.94	0.93
2	0.04	0.02	0.03	0.12	0.15	0.13	0.23	0.06	0.1	0.13	0.94	0.24	0.7	0.66	0.68
3	0.33	0.02	0.03	0.6	0.05	0.1	0.17	0.02	0.03	1.0	0.02	0.03	0.64	0.86	0.73
4	0.0	0.0	0.0	0.0	0.0	0.0	0.06	0.03	0.04	0.57	0.21	0.31	0.53	0.5	0.51
5	0.0	0.0	0.0	0.0	0.0	0.0	0.12	0.03	0.05	0.67	0.24	0.35	0.6	0.82	0.69
6	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.03	0.06	0.58	0.17	0.26	0.82	0.76	0.79
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.52	0.41	0.46	0.71	0.85	0.78
8	0.05	0.63	0.1	0.02	0.27	0.05	0.09	0.98	0.16	0.38	0.22	0.28	0.41	0.37	0.38
9	0.0	0.0	0.0	0.0	0.0	0.0	0.09	0.02	0.03	0.0	0.0	0.0	0.27	0.12	0.16
Acc.	-	-	0.05	-	-	0.04	-	-	0.11	-	-	0.28	-	-	0.7

Toronto Metropolitan University

Results: After PGD Attacks





Toronto Metropolitan University

Conclusions

- All KANs except Linear show class imbalance
- Significant variations between different KAN models
- MLP more robust and resilient except under PGD attacks
- All KANs except Linear performed better under PGD attacks
- Chebyshev KAN resisted PGD with accuracy around 0.3 vs mostly 0 for the rest

Future Work

- Investigating the observed differences.
- Training KAN models specifically for handling AA.
- Improving AA methods (less successful attacking KANs)
- Looking into some resistance of KAN models to PGD
- Using different datasets

Toronto Metropolitan University

SECURWARE 2024, November 3 - 7, Nice, France

KAN vs KAN: Examining Kolmogorov-Arnold Networks (KAN) Performance Under Adversarial Attacks

Q & A

nebojsa.djosic@torontomu.ca

Toronto Metropolitan University

IARIA