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

Evgenii Ostanin
eostanin@torontomu.ca

Nebojsa Djosic
nebojsa.djosic@torontomu.ca

Fatima Hussain
fatima.hussain@torontomu.ca

Salah Sharieh
salah.sharieh@torontomu.ca

Alexander Ferworn
aferworn@torontomu.ca

Toronto Metropolitan University, Toronto, Canada

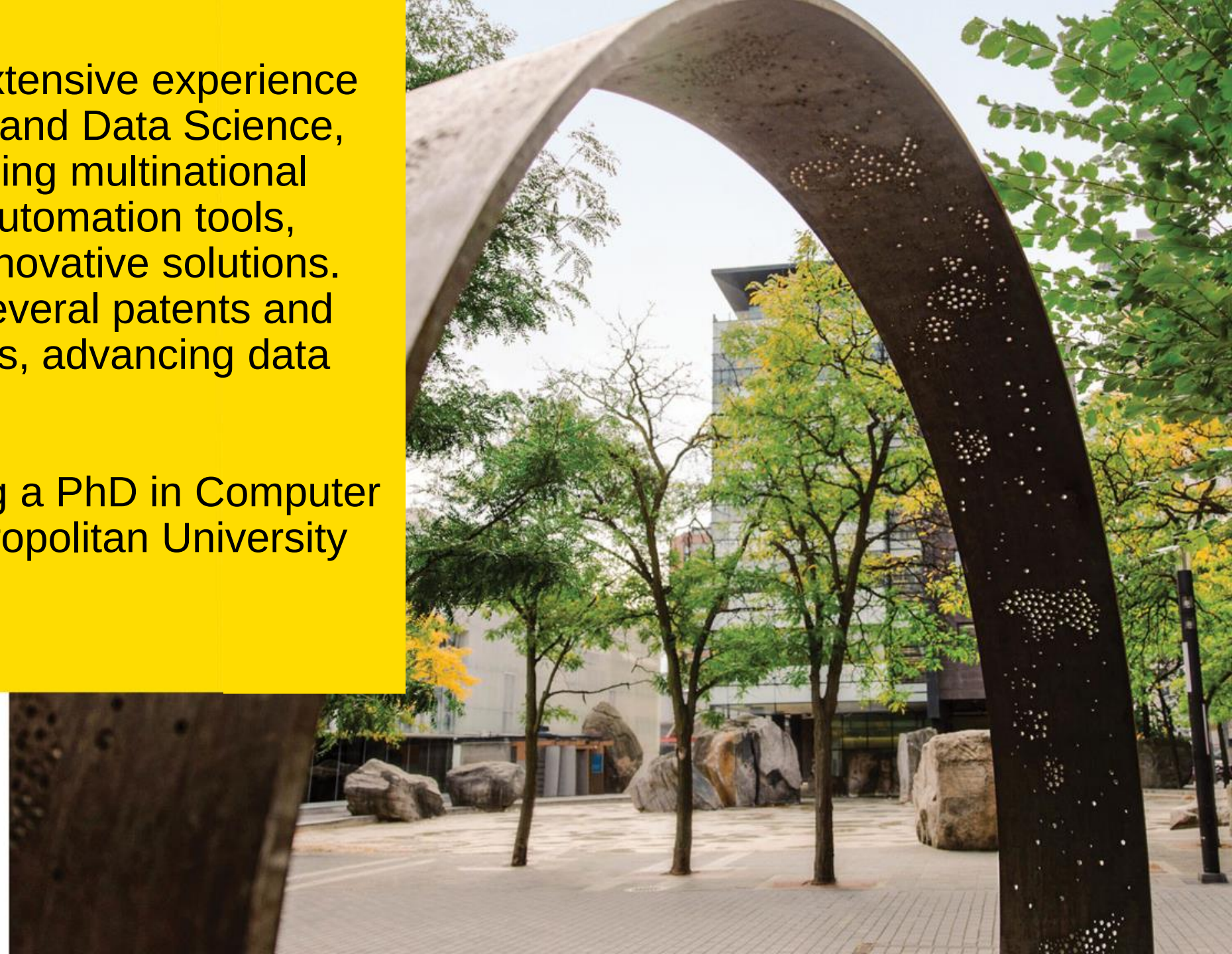
Presenter
Evgenii Ostanin

SECURWARE 2024, November 3 - 7, Nice, France

Evgenii Ostanin has extensive experience in Economics, Finance, and Data Science, having worked with leading multinational companies to develop automation tools, financial models, and innovative solutions. He has contributed to several patents and proof-of-concept projects, advancing data science and technology.

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

- Introduction and Problem Statement
- Kolmogorov-Arnold Networks (KANs) introduction
- Noise and Adversarial Attacks
- Methodology and Architecture
- Results
- Conclusions and Future Work

Toronto Metropolitan University

Key Contributions

- Comparative study evaluating the robustness of Kolmogorov-Arnold Networks (KANs) against adversarial attacks (FGSM, PGD) and Gaussian noise
- Demonstrates KANs' vulnerabilities under noisy and adversarial conditions, with significant accuracy drops compared to MLPs.
- Provides insights into KANs' sensitivity to perturbations, revealing areas for improvement in robustness and security.

Toronto Metropolitan University

Introduction

- **KANs**: Flexible, interpretable neural networks.
- **Research gap**: Robustness under adversarial attacks and noise not fully explored.
- **Goal**: Compare KANs vs. MLPs under adversarial attacks (FGSM, PGD) and Gaussian noise.

Problem Statement

- **MLPs:** Limited flexibility, vulnerable to noise/attacks.
- **KANs:** Learnable spline functions offer flexibility but may introduce sensitivity to perturbations.
- **Focus:** Testing robustness of KANs.

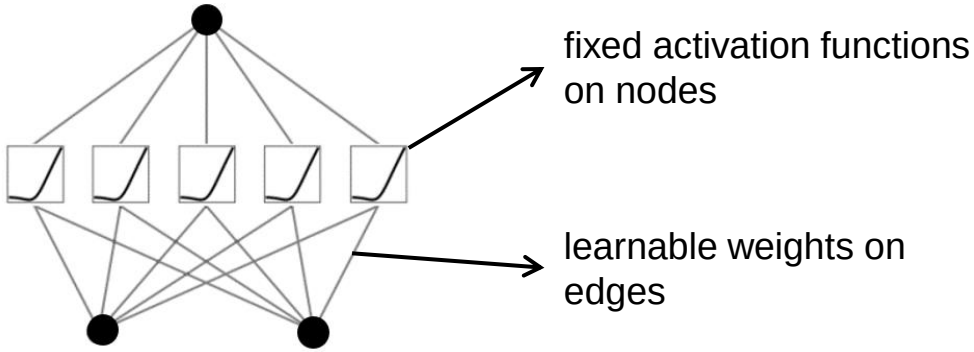
Toronto Metropolitan University

Kolmogorov-Arnold Network Architecture

*Based on **Kolmogorov-Arnold Representation Theorem:** Any multivariate continuous function $f(x_1,....,x_n)$ within a bounded domain can be represented as a superposition of continuous single-variable functions*

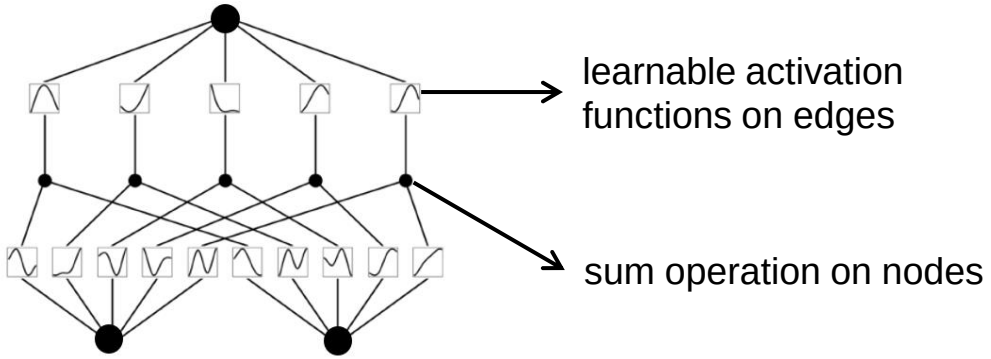
Traditional MLPs:

- Fixed activation functions like ReLU
- Linear weight matrices



KANs:

- Use *spline-based* activations on edges
- Greater **non-linearity** and **adaptivity**



Toronto Metropolitan University

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

Adversarial Attacks

Random Noise:

- Gaussian noise tests real-world resilience.

Fast Gradient Sign Method (FGSM):

- Introduces small, **calculated perturbations** to input data
- Simple yet powerful attack

Projected Gradient Descent (PGD):

- Iterative attack based on FGSM
- More effective and resilient

Impact on Neural Networks: Even minor perturbations can cause significant prediction shifts.

Toronto Metropolitan University

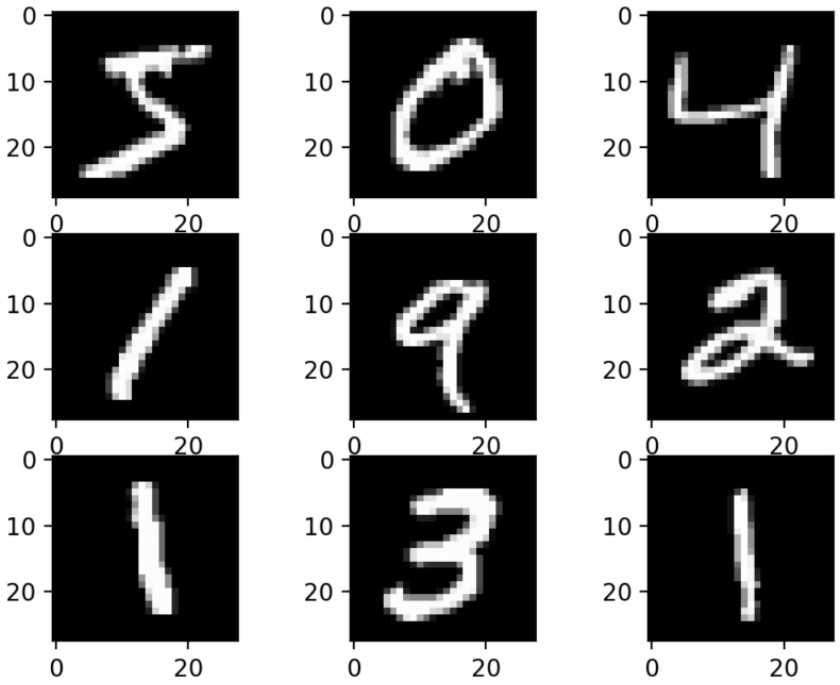
Methodology

Goal: Compare the robustness of KANs to MLPs under:

- **Clean conditions**
- **Noisy data** (Gaussian noise)
- **Adversarial attacks** (FGSM and PGD)

Dataset: MNIST (handwritten digits)

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score



KAN vs. MLP Architectures

MLP

Input Layer: Takes the raw input data (28x28 pixel image, or 784 features.).

Hidden Layers: Each hidden layer applies fixed activation functions (e.g., ReLU), and every neuron is fully connected to neurons in adjacent layers. (512, 256, 128, 64)

Output Layer: Produces the final classification or prediction (10 - one neuron per class for MNIST digits).

KAN

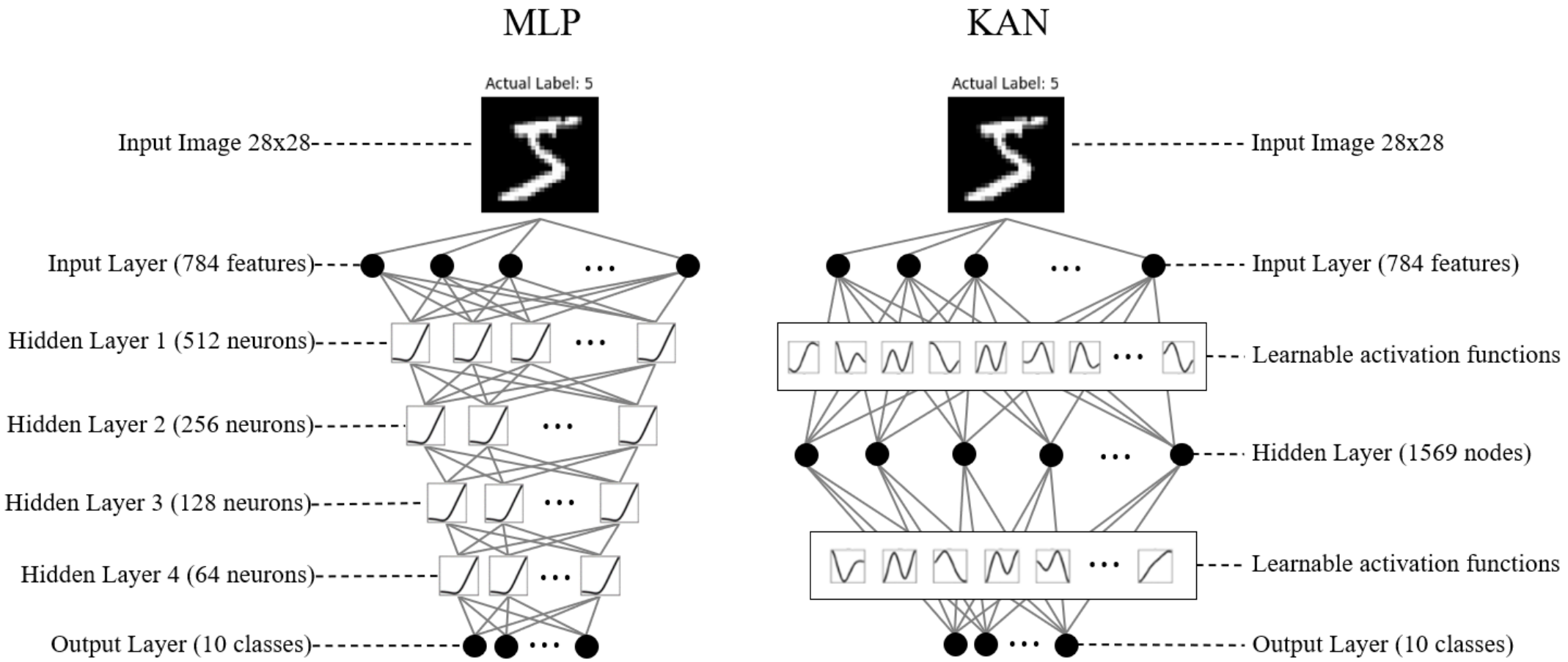
Input Layer: Similar to MLP, takes raw input data.

KANLinear Layers: Replaces standard hidden layers with layers using learnable spline-based activation functions, allowing flexibility in how activations adapt during training.

Output Layer: Similar to MLP, outputs predictions (e.g., digits 0–9 for classification).

Toronto Metropolitan University

KAN vs. MLP Architectures



MLP: Fixed weights and ReLU activations.

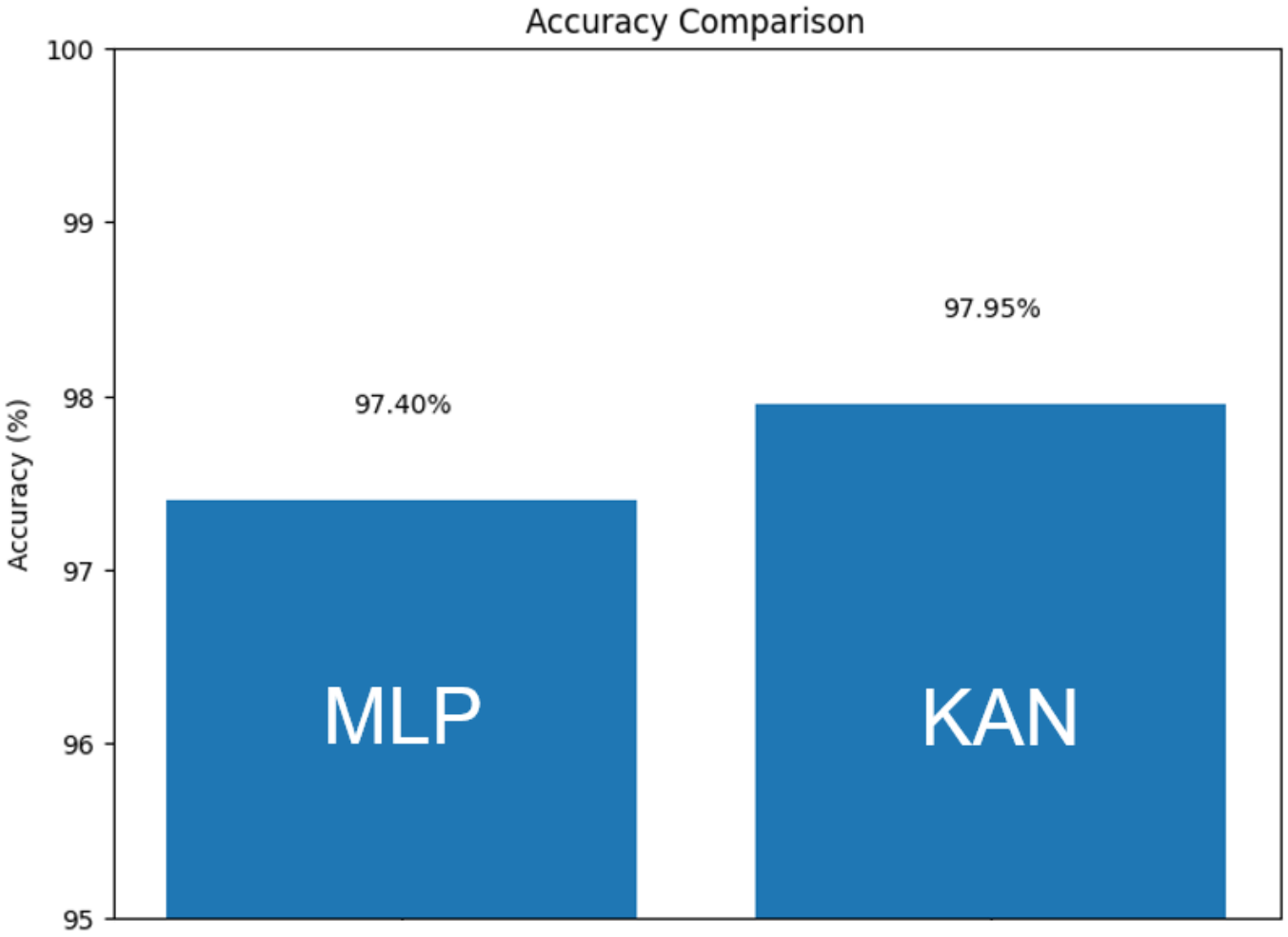
KAN: Spline-based activations on edges.

Toronto Metropolitan University

Results – Clean Data

- In clean conditions, KANs performed slightly better than MLPs.
- Accuracy: KAN - 97.95%, MLP - 97.40%
- KAN’s flexible architecture leads to better handling of data variability.

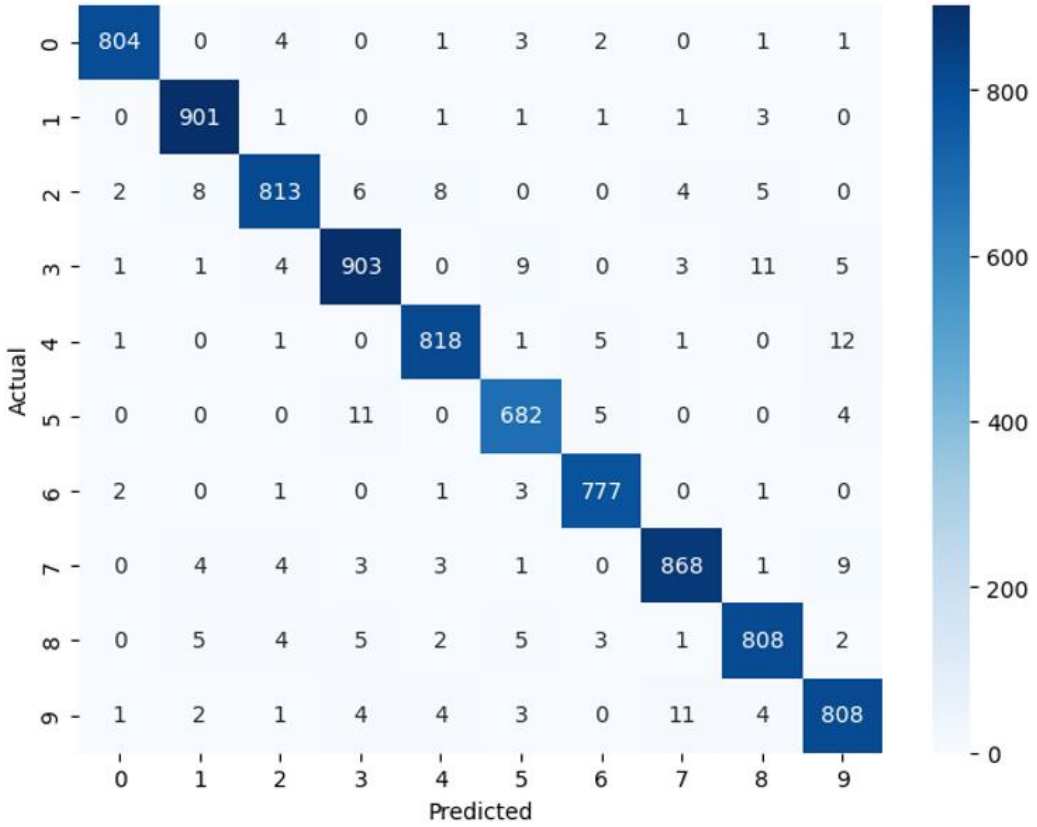
Results – Clean Data



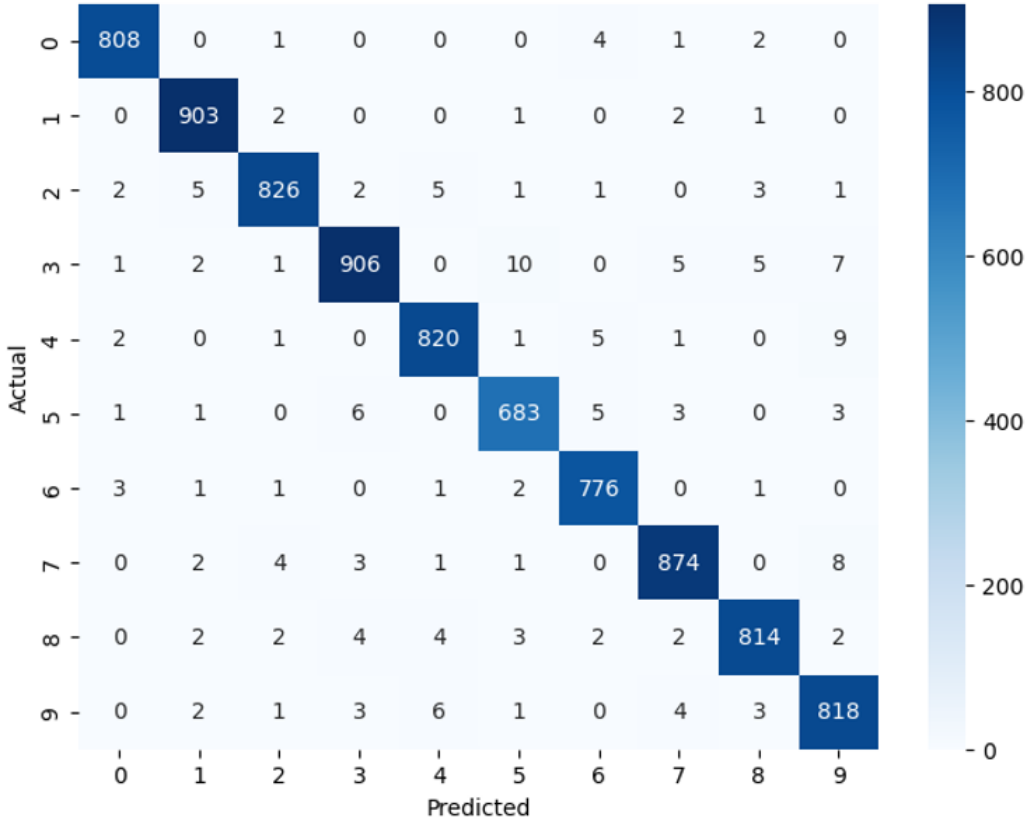
Toronto Metropolitan University

Results – Clean Data

MLP



KAN

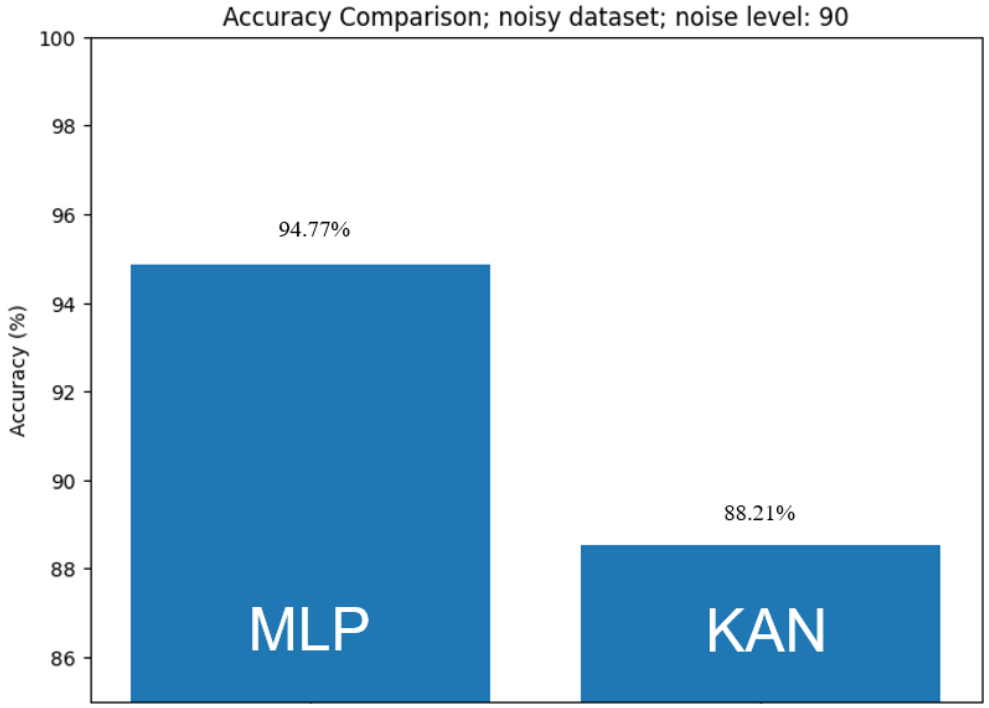


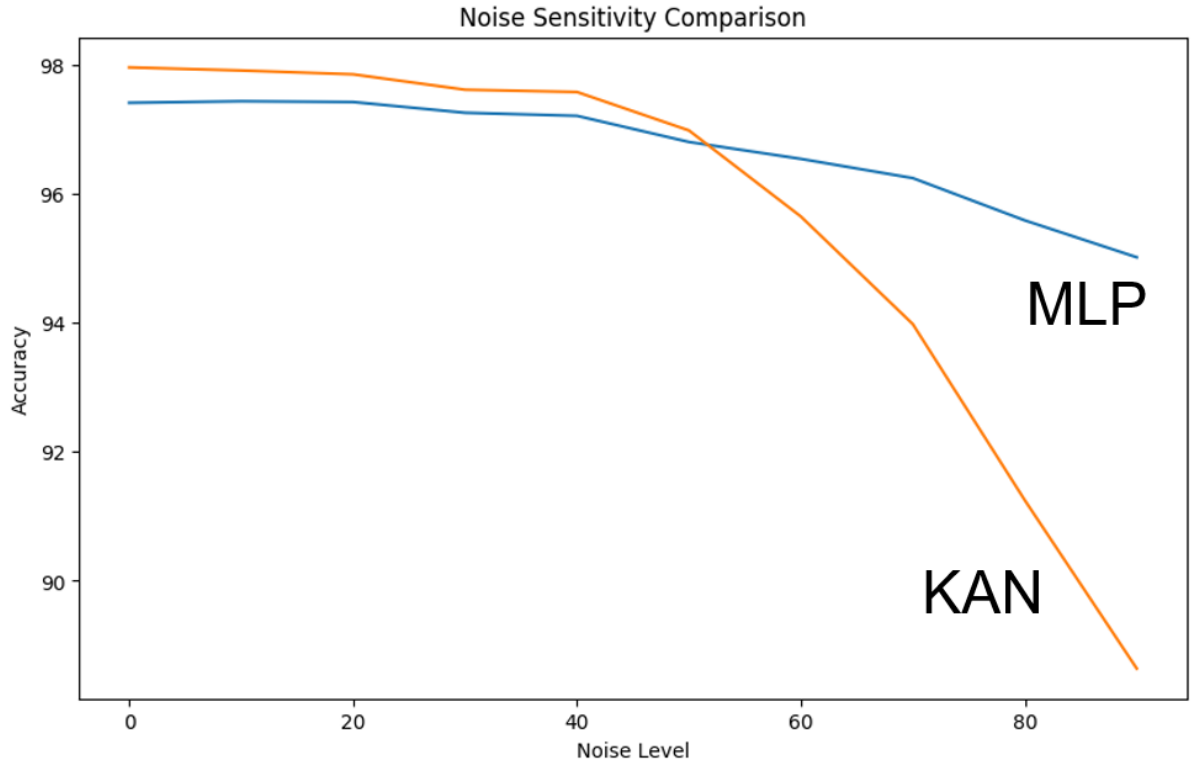
Toronto Metropolitan University

Results – Gaussian Noise

- KAN’s accuracy degraded faster as noise increased compared to MLPs.
- Noise level 90: KAN - 88.21%, MLP - 94.77%
- KANs struggle more under noisy conditions, especially for digits like 1 and 8.

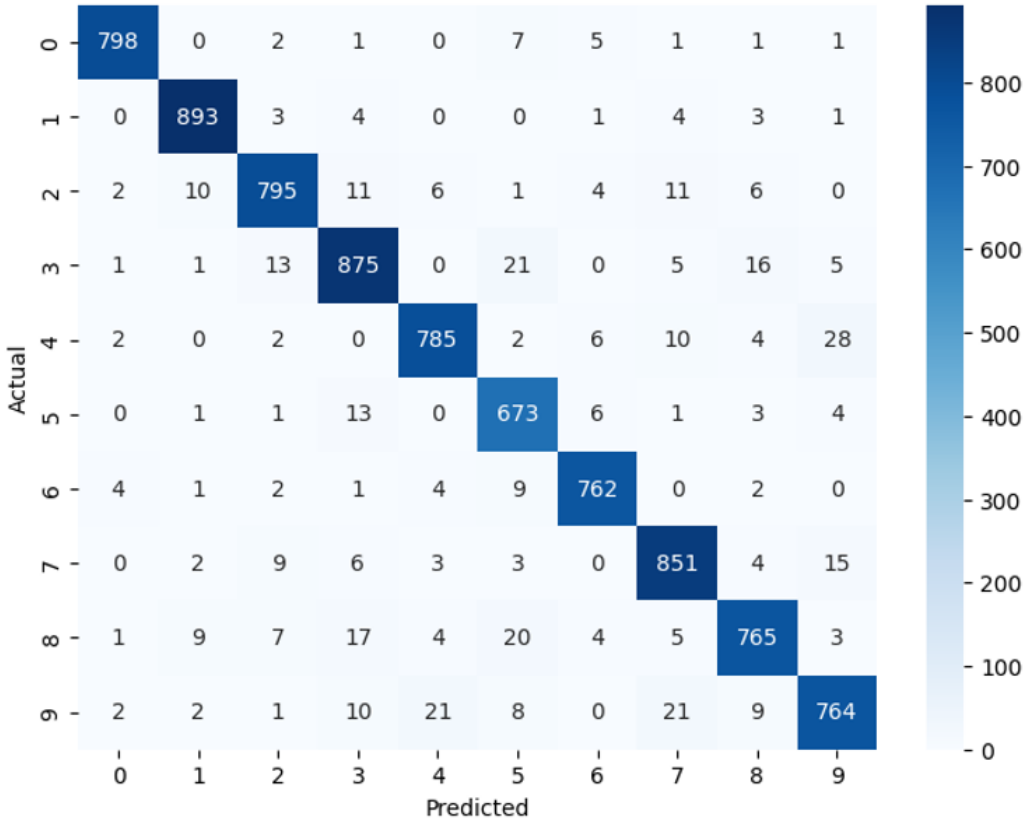
Results – Gaussian Noise



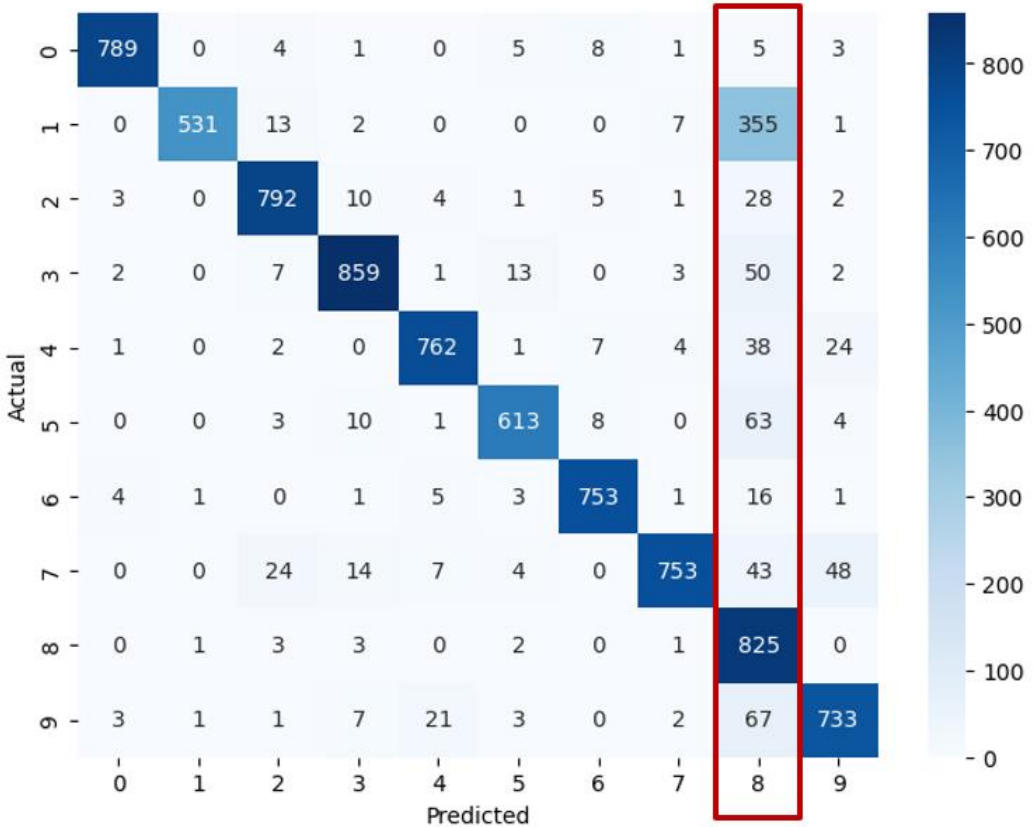


Results – Gaussian Noise

MLP



KAN



Toronto Metropolitan University

Results – Adversarial Attacks

FGSM Attack (ε = 0.3):

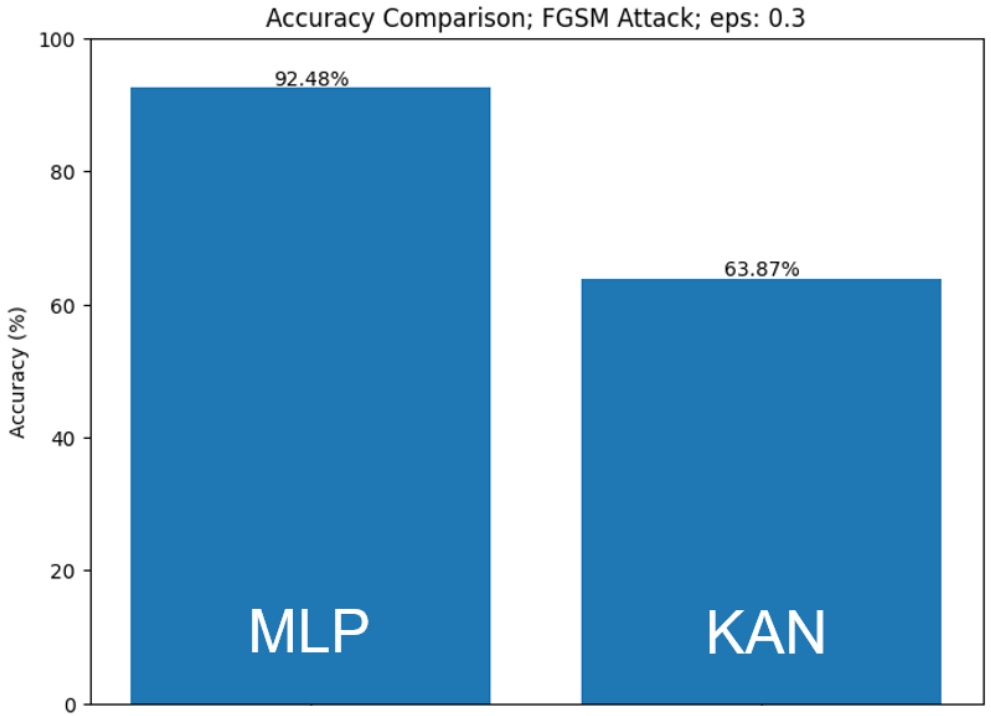
- KAN's accuracy drops to 63.87% compared to MLP's 92.48%.
- KAN is more sensitive to small perturbations.

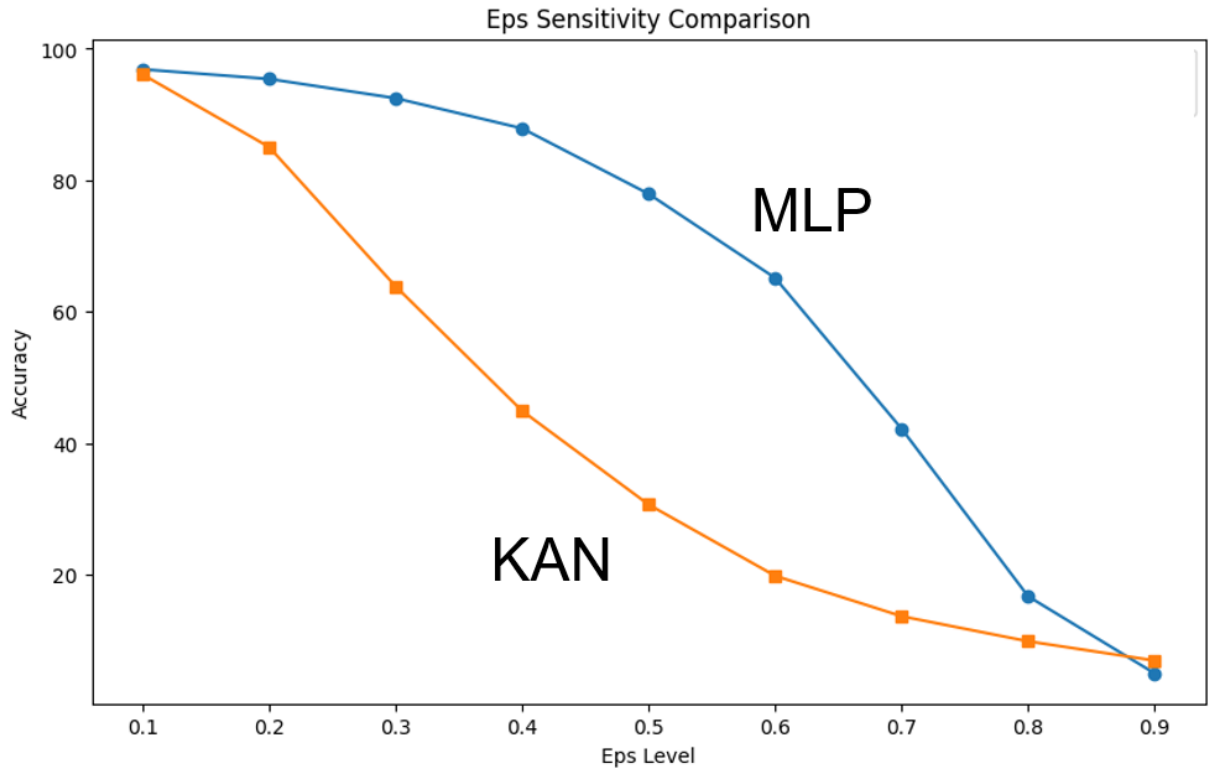
PGD Attack (ε = 0.3):

- KAN drops further to 53.12%, while MLP remains relatively robust at 96.29%.

Toronto Metropolitan University

Results – FGSM

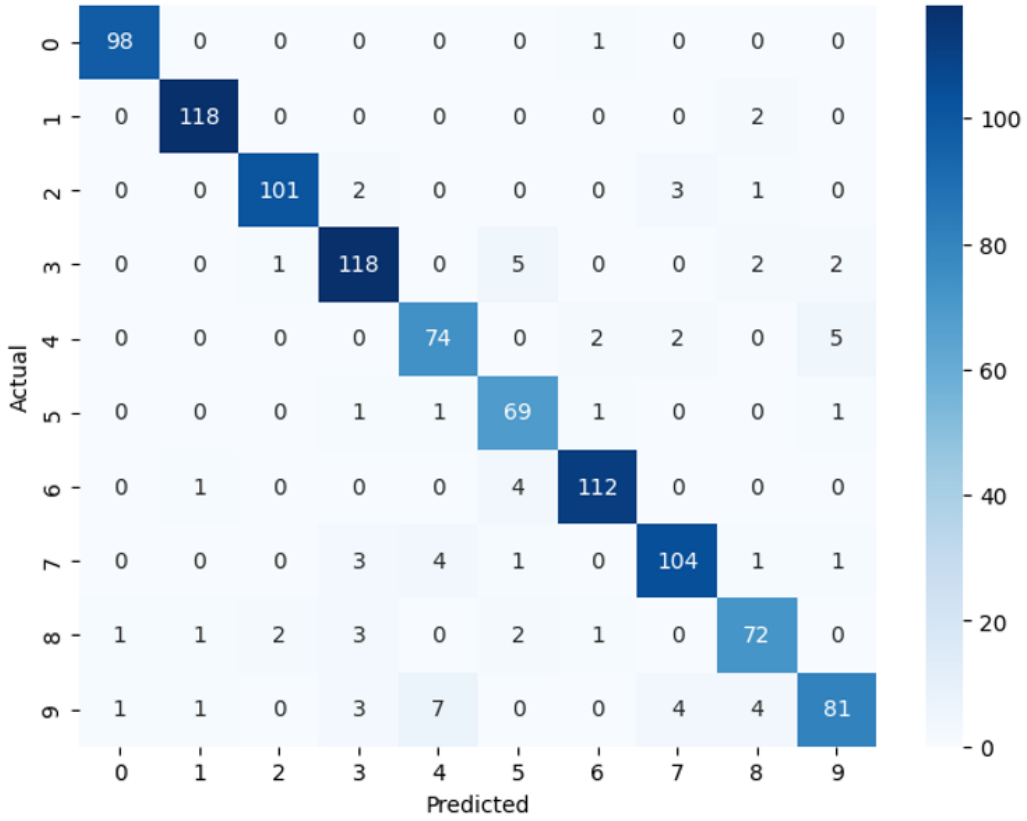




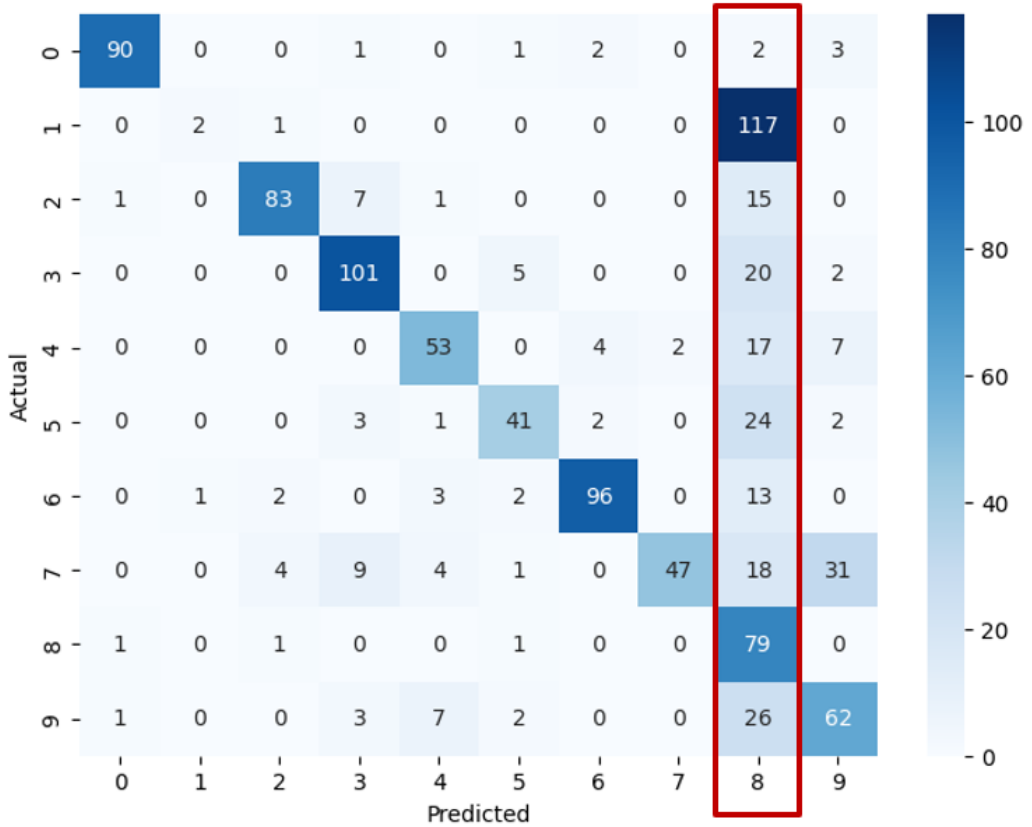
Toronto Metropolitan University

Results – FGSM

MLP

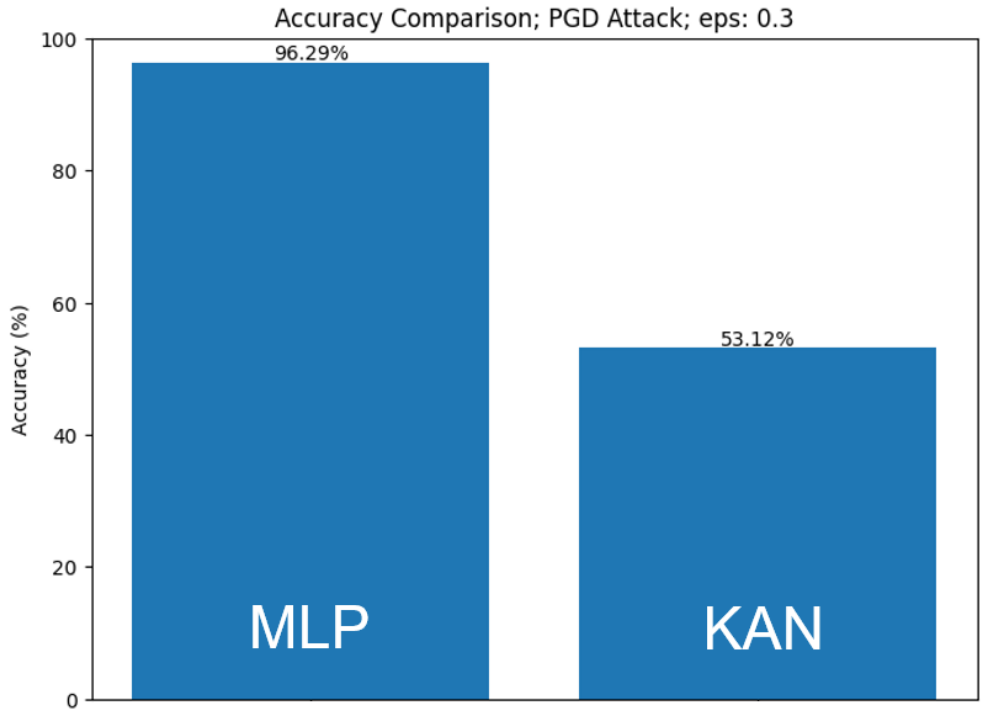


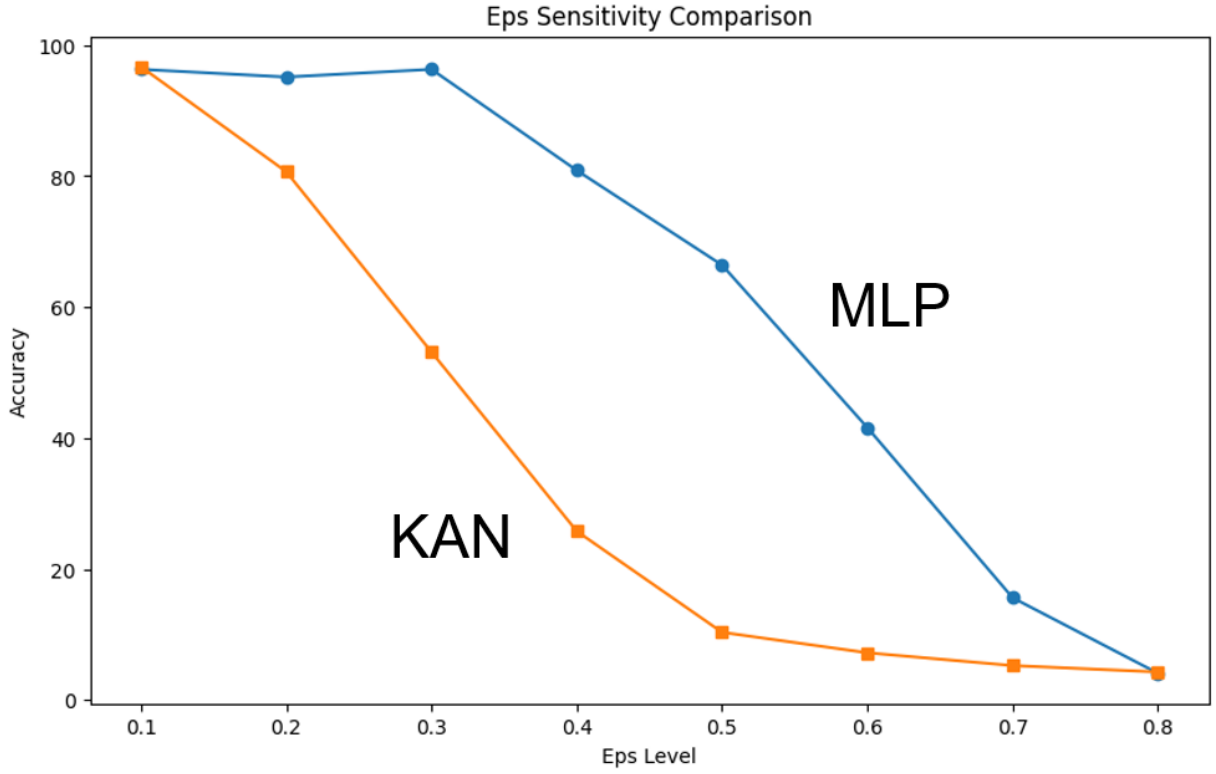
KAN



Toronto Metropolitan University

Results – PGD

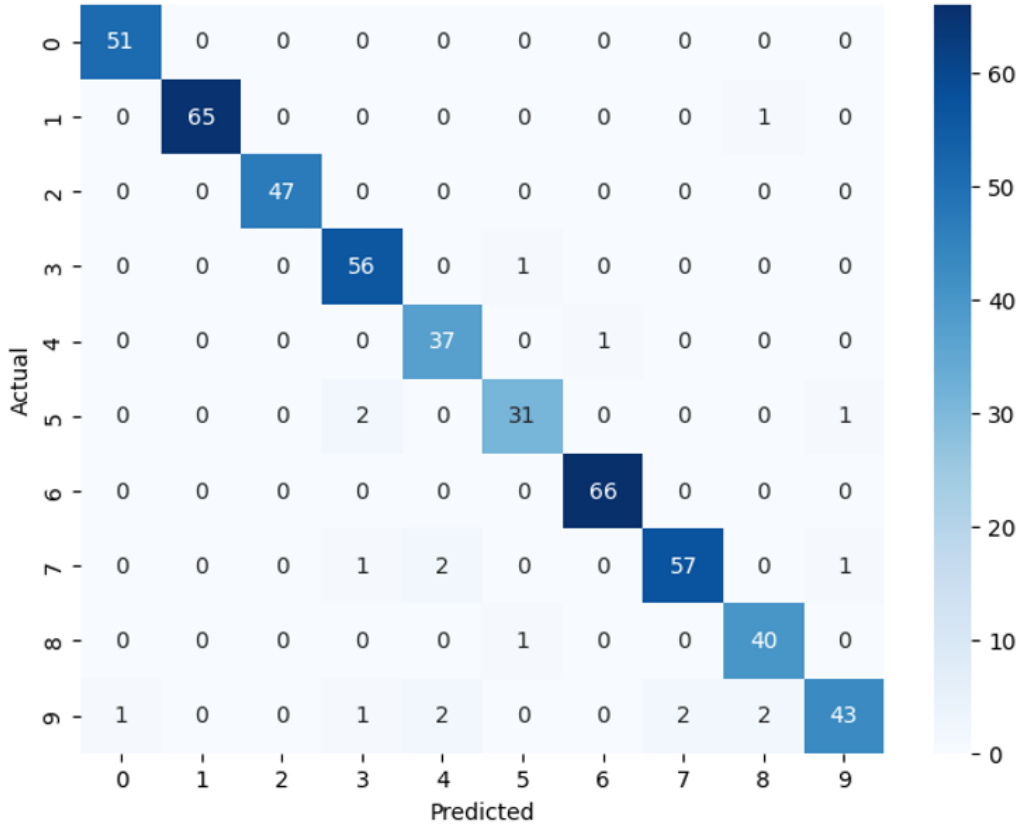




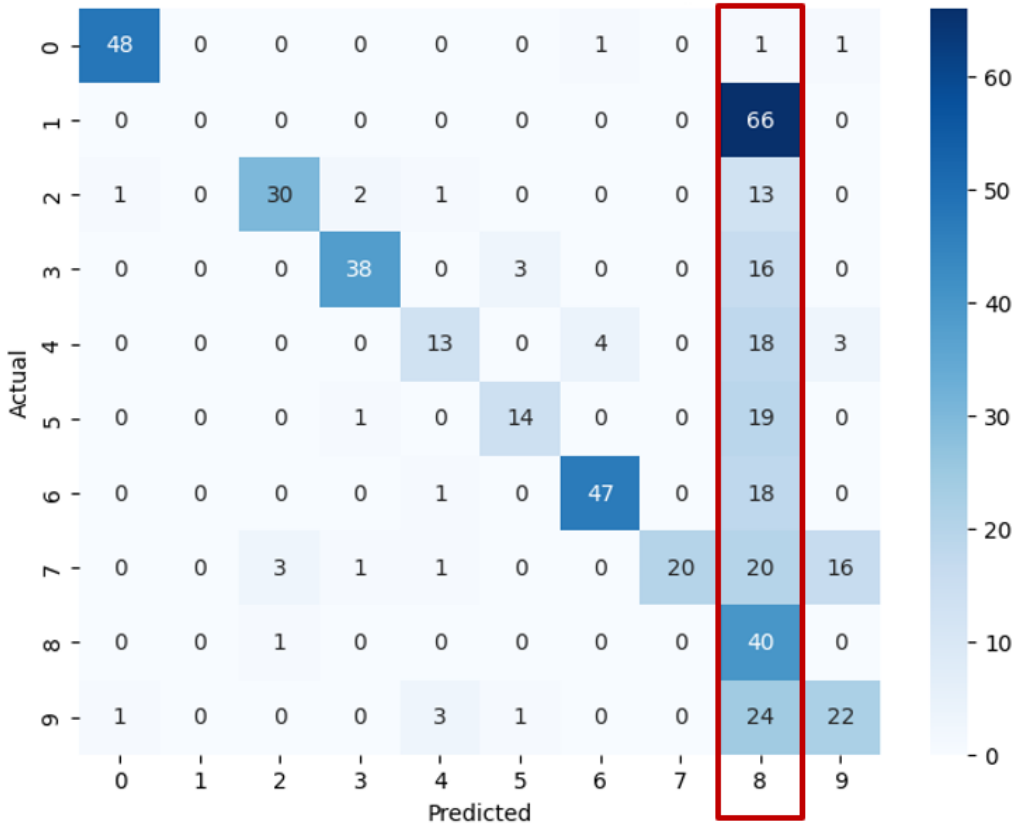
Toronto Metropolitan University

Results – PGD

MLP



KAN



Toronto Metropolitan University

Results Summary

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

Toronto Metropolitan University

Conclusions

- KAN excels in clean environments.
- KAN suffers more from noise and adversarial attacks than MLP.
- Possible causes: Spline function sensitivity and overfitting.
- Next steps: Enhancing resilience for secure ML applications.

Toronto Metropolitan University

Future Work

- Investigating and developing advanced robustness techniques tailored for KANs.
- Designing KAN architectures that inherently handle noisy inputs better.
- Conducting a thorough security analysis of KANs across a broader range of AA methods.

SECURWARE 2024, November 3 - 7, Nice, France

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

Q & A

eostanin@torontomu.ca

Toronto Metropolitan University

IARIA