Final Report

Project Title

- A Kolmogorov-Arnold Network for Interpretable Cyberattack Detection in AGC Systems.

Group Members and Their Contributions

- Member 1: Jehad Adel Jilan. AI specialization

Project Mentor

Dr. Ahmad Saber: supervisor

Project Objectives and Goals

- Develop an interpretable method to detect stealthy FDIA attacks on AGC systems.
- Address the limitations of black-box ML models.

Current Progress and Milestones Achieved

- Developed and trained the KAN model for FDIA detection in AGC systems.
- Generated symbolic expression to explain model behavior.
- Evaluated model performance on simulated AGC datasets.

Methodology and Approach

- Data Preparation: Simulate AGC measurements under normal and attack scenarios.
- Feature Engineering: Extract statistical and nonlinear features from time-series data.
- Model Training: Use Kolmogorov-Arnold Networks (KAN) to learn complex nonlinear relationships.
- Symbolic Regression: Generate clear mathematical formulas that explain the model's decisions.

- Pruning and Fine-Tuning: Simplify the model while recovering accuracy to balance interpretability and performance.
- Evaluation: Measure binary classification accuracy and formula precision on test data.

Preliminary Results or Findings

- KAN model achieved up to 95.90% detection accuracy for FDIA attacks.
- Generated symbolic formulas maintained ~95.97% accuracy, supporting interpretability.
- Insights confirm that interpretable models can approach black-box accuracy while improving operator trust.

Timeline and Schedule

Task	Start Date	End Date	Status
Domain knowledge about the problem	06/06/2025	15/06/2025	Completed
Complete the model	16/06/2025	01/07/2025	Completed
Present the experiment	02/07/2025	07/07/2025	Completed
Complete the conference paper	10/07/2025	15/07/2025	Completed
Set up the real-world data using Opal-RT system	16/07/2025	01/10/2025	In progress

Challenges and Obstacles Encountered

- Balancing the trade-off between the KAN model's accuracy and the symbolic formula.

Next Steps and Future Work

- improve the interpretability of attack detection and identification
- Support real-time processing capabilities.
- Identify the type of attack and determine the supposed signals

Resources and Tools Used

- https://arxiv.org/pdf/2503.02281
- https://arxiv.org/pdf/2504.09363

Resources and Tools Used

- Approved paper: https://drive.google.com/file/d/1mCjtcYW3u74B3_oKS4oal9TEHS5I33OH/view?usp=sharing
- Symbolic expression: https://docs.google.com/document/d/15XFchnXIS-95QvIlu5wOv_YIKUY3XSkeL7HjnB4s4OM/edit?usp=sharing
- Code: https://drive.google.com/file/d/1OiFPIRD6bgzRhcUDQWXPxSi_H-qz_Fhv/view?usp=drive_link