**Doctoral School of Information and Biomedical Technologies**
**Polish Academy of Sciences (SD TIB PAN)**

## SUBJECT:

**Distinguishing component faults from deliberate attacks: interpretable machine learning for anomaly detection in IoT and engineering systems**

## SUPERVISOR:

dr hab. Joanna Kołodziej

joanna.kolodziej@nask.pl

Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB), ul. Kolska 12, Warszawa

## DESCRIPTION:

In the era of advanced IoT system development and the increasing role of systems in securing technology infrastructure, data processing, and transmission, modern solutions for detecting anomalies and external attacks are critical to the reliable operation of IoT systems. The primary challenge of anomaly detection, particularly in critical environments of infrastructure and complex large-scale IT systems, is often the absence or only residual characteristics of these anomalies – it is unclear whether the cause is a failure of system components or an external attack. In such cases, it is essential to react quickly and reconfigure the system in a way that does not compromise its operation, even in emergency mode.

This uncertainty about the origin of an anomaly is the central research problem of the thesis. Existing anomaly and intrusion detection systems typically either reduce detection to a binary decision (normal vs. anomalous) or classify anomalies into known attack families. The question of whether an observed deviation is caused by a benign component fault (e.g., sensor degradation or drift) or by a deliberate external attack (e.g., data injection or jamming) has received little attention in the literature, although it is the one the operator most needs answered: a fault calls for maintenance and reconfiguration, an attack for security countermeasures. The problem is hardest in the early stage of detection, when no labelled examples of the ongoing incident are available.

The lack of complete information about the nature of anomalies in the initial stages of detection makes it challenging to train machine learning (ML) models and deep learning (DL) models, which are commonly used in so-called Anomaly (Intrusion) Detection Systems (ADS/IDS). These methods are used to identify suspicious patterns in monitoring data from IoT systems, which requires large training datasets. Unsupervised learning methods solve the problem of costly and time-consuming data annotation (using unlabelled data), but the efficiency and accuracy of such systems are generally not high. Another proposal is federated learning methods (FL) without the need for data centralisation.

In recent years, so-called “oversampling” methods have often been used in anomaly detection, which allow the training dataset to be enlarged in case of a shortage of labelled data or handling data imbalances. Generative adversarial networks (GANs) are used to detect generative “patterns” of anomalies. However, these methods are generally computationally expensive and negatively impact ITS performance, failing in cases where minor differences exist between data patterns depicting normal system fluctuations and disturbed system fluctuations.

Given the limitations of the ML and DL models and their variants, there is a need to build and analyse models that can better generalise, capture complex nonlinear relationships in data, and offer computational performance suitable for real-time IoT deployments. A promising alternative is the Kolmogorov-Arnold network (KAN). Unlike traditional DL and ML models, which rely on fixed activation functions, KANs use learnable one-dimensional functions at their edges, allowing them to approximate complex, multidimensional functions with fewer parameters and greater interpretability. This makes KANs particularly suited to the unpredictable and evolving nature of attack and anomaly detection, where patterns of these attacks can change rapidly and feature relationships that are often non-linear, as is often the practice case.

For anomaly-source identification, the interpretability of KANs is treated as a hypothesis to be tested. The learned one-dimensional edge functions can be inspected directly, and the work will examine whether their shape separates fault signatures (such as a smooth, gradual sensor drift) from attack signatures (such as a localised injected perturbation). A positive result would give operators an explainable indication of the probable cause of an anomaly. A negative result, obtained under a fair comparison protocol with parameter- and compute-matched ML and DL baselines, is also a valid scientific contribution of the thesis.

The purpose of this work is to develop KAN-based models for the detection of anomalies and the identification of their probable source (component failure vs. external attack) in IoT systems, along with a comprehensive comparative analysis with existing ML, DL, and FL models. The work involves designing a reliable case study that utilises an implementable system for speech analysis, image analysis, or wireless network operation (e.g., LoRa WAN) in the monitoring and management of safe intelligent transportation in cities. LoRaWAN-based sensor networks are the primary candidate for this case study: their strict device constraints and distributed gateway architecture match the resource and deployment assumptions of the proposed models.

**REQUIREMENTS:**

- MSc degree in computer science, mathematics, AI, or related field
- High programming skills in Python
- Experience with ML, DL and FL tools and model evaluation frameworks
- Knowledge of current research on anomaly and intrusion detection methods
- Advanced level of English (spoken and written);

**BIBLIOGRAPHY:**


- M. Frustaci, P. Pace, G. Aloi, and G. Fortino, “Evaluating critical security issues of the IoT world: Present and future challenges,” *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2483–2495, Aug. 2018.
- SonicWall. (Mar. 27, 2023). *Annual Number of Internet of Things (IoT) Malware Attacks Worldwide From 2018 to 2022 (in Millions)*. [Online]. Available: https://www.statista.com/statistics/1377569/worldwide-annual-internet-of-things-attacks/
- G. Kocher and G. Kumar, “Machine learning and deep learning methods for intrusion detection systems: Recent developments and challenges,” *Soft Comput.*, vol. 25, no. 15, pp. 9731–9763, Aug. 2021.
- Z. Azam, M. M. Islam, and M. N. Huda, “Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree,” *IEEE Access*, vol. 11, pp. 80348–80391, 2023.
- Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, “KAN: Kolmogorov–Arnold networks,” 2024, arXiv:2404.19756.
- G. De Carlo, A. Mastropietro, and A. Anagnostopoulos, “Kolmogorov–Arnold graph neural networks,” 2024, arXiv:2406.18354.
- Z. Li, “Kolmogorov–Arnold networks are radial basis function networks,” 2024, arXiv:2405.06721.
- A. Aghaei, “FKAN: Fractional Kolmogorov–Arnold networks with trainable Jacobi basis functions,” 2024, arXiv:2406.07456.
- S. Zinage, S. Mondal, and S. Sarkar, “DKL-KAN: Scalable deep kernel learning using Kolmogorov–Arnold networks,” 2024, arXiv:2407.21176.
- R. Yu, W. Yu, and X. Wang, “KAN or MLP: A fairer comparison,” 2024, arXiv:2407.16674.

- I. Butun, N. Pereira, and M. Gidlund, “Security risk analysis of LoRaWAN and future directions,” *Future Internet*, vol. 11, no. 1, art. 3, 2019.
- E. Aras, G. S. Ramachandran, P. Lawrence, and D. Hughes, “Exploring the security vulnerabilities of LoRa,” in *Proc. IEEE Int. Conf. Cybernetics (CYBCONF)*, 2017, pp. 1–6.