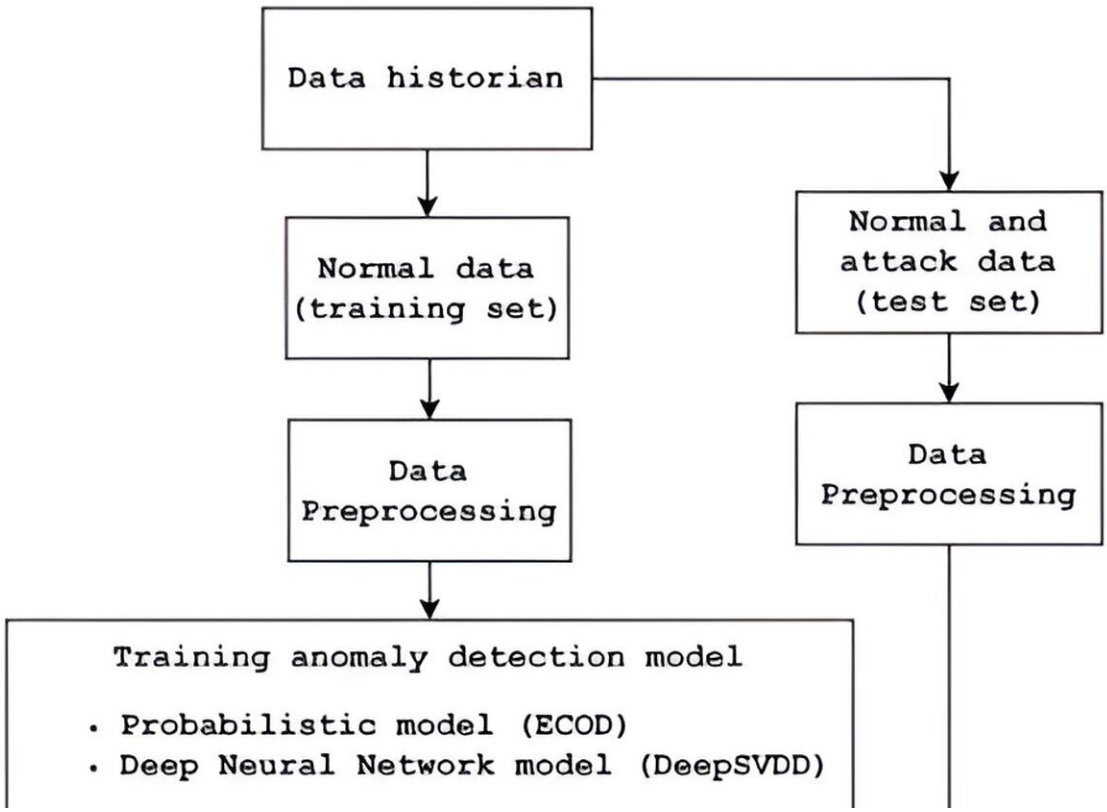
Addressing AI 'hallucinations' and improving anomaly detection algorithms in industrial systems

July 31 2024



Overall methodology of the study. Credit: *Proceedings of the 10th ACM Cyber-Physical System Security Workshop* (2024). DOI: 10.1145/3626205.3659147

Significant strides in addressing the issue of AI "hallucinations" and

improving the reliability of anomaly detection algorithms in critical national infrastructures (CNIs) have been made by scientists based in Bristol's School of Computer Science.

Recent advances in artificial intelligence have highlighted the technology's potential in anomaly detection, particularly within sensor and actuator data for CNIs. However, these AI algorithms often require extensive training times and struggle to pinpoint specific components in an anomalous state. Furthermore, AI's decision-making processes are frequently opaque, leading to concerns about trust and accountability.

To help combat this, the team brought in a number of measures to boost efficiency including:

1. Enhanced anomaly detection: Researchers employed two cutting-edge anomaly detection algorithms with significantly shorter training times and faster detection capabilities, while maintaining comparable efficiency rates. These algorithms were tested using a dataset from the operational water treatment testbed, SWaT, at the Singapore University of Technology and Design.
2. Explainable AI integration: To enhance transparency and trust, the team integrated eXplainable AI models with the anomaly detectors. This approach allows for better interpretation of AI decisions, enabling human operators to understand and verify AI recommendations before making critical decisions. The effectiveness of various XAI models was also evaluated, providing insights into which models best aid human understanding.
3. Human-centric decision making: The research emphasizes the importance of human oversight in AI-driven decision-making processes. By explaining AI recommendations to human operators, the team aims to ensure that AI acts as a decision-support tool rather than an unquestioned oracle. This

methodology introduces accountability, as human operators make the final decisions based on AI insights, policy, rules, and regulations.

4. Scoring system development: A meaningful scoring system is being developed to measure the perceived correctness and confidence of the AI's explanations. This score aims to assist human operators in gauging the reliability of AI-driven insights.

These advancements not only improve the efficiency and reliability of AI systems in CNIs but also ensure that human operators remain integral to the decision-making process, enhancing overall accountability and trust.

Dr. Sarad Venugopalan, co-author of the study, explained, "Humans learn by repetition over a longer period of time and work for shorter hours without being error prone. This is why, in some cases, we use machines that can carry out the same tasks in a fraction of the time and at a reduced error rate.

"However, this automation, involving cyber and physical components, and subsequent use of AI to solve some of the issues brought by the automation, is treated as a black box. This is detrimental because it is the personnel using the AI recommendation that is held accountable for the decisions made by them, and not the AI itself.

"In our work, we use explainable AI, to increase transparency and trust, so the personnel using the AI is informed why the AI made the recommendation (for our domain use case) before a decision is made."

This research is part of the MSc thesis of Mathuros Kornkamon, under the supervision of Dr. Sridhar Adepu. The paper is published in *Proceedings of the 10th ACM Cyber-Physical System Security Workshop*.

Dr. Adepu added, "This work discovers how WaXAI is revolutionizing anomaly detection in industrial systems with explainable AI. By integrating XAI, human operators gain clear insights and enhanced confidence to handle security incidents in critical infrastructure."

More information: Kornkamon Mathuros et al, WaXAI: Explainable Anomaly Detection in Industrial Control Systems and Water Systems, *Proceedings of the 10th ACM Cyber-Physical System Security Workshop* (2024). DOI: 10.1145/3626205.3659147

Provided by University of Bristol