Explainable Sensor Data-Driven Anomaly Detection in Internet of Things Systems

Moaz Tajammal Hussain, Charith Perera
Cardiff University

Motivation & Aim

- Explainability is a key hurdle in the wider adoption of deep neural networks for various tasks including anomaly detection.
- A low-time latency and computationally efficient method is proposed to generate explanations for deep learning output.
- Additionally, an explanation dashboard is created to explain the detected anomaly for different personas using deep learning explanation and provenance logs.

Methodologies

Anomaly Detection — **LSTM Auto-Encoder**

Unsupervised deep learning prediction-based technique is used to detect anomalies in data stream.

Explanation Generation — **Root Method: Shapley Additive explanation (SHAP)**

A derivative of game theory based SHAP, TreeSHAP, is used to generate post-hoc model agnostic local explanations for the deep learning model output.

Dual surrogate models are then employed to explain the model output with respect to feature interactions, previous feature values and temporal dependencies

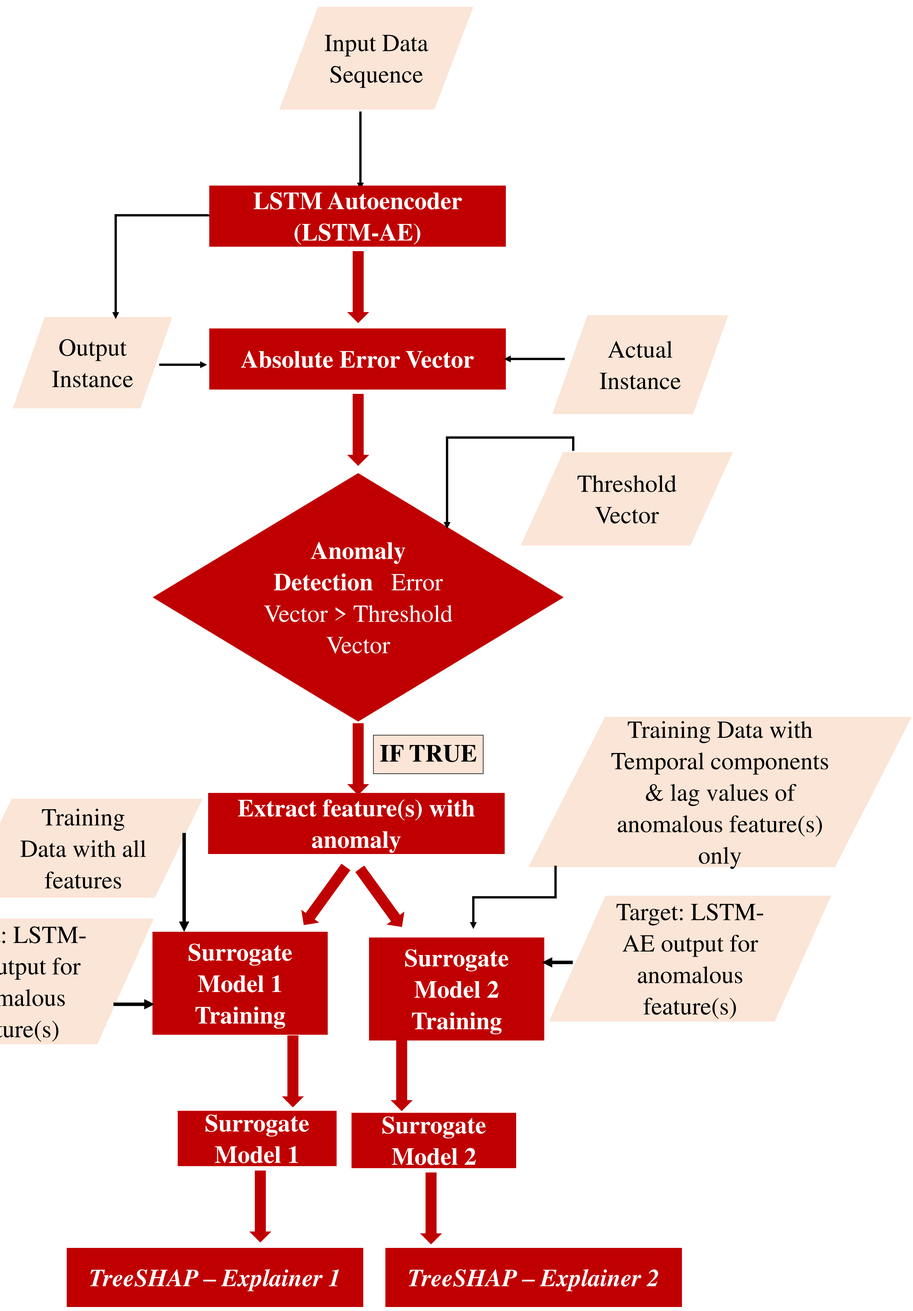
Dataset

SWaT (Secure Water Treatment)

Six-stage water treatment testbed

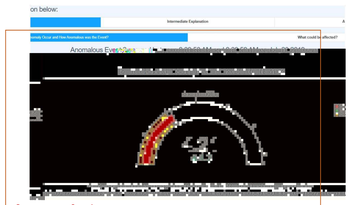
14,998 observations from **78** IoT devices

Anomaly Detection and Black Box Model Explanation

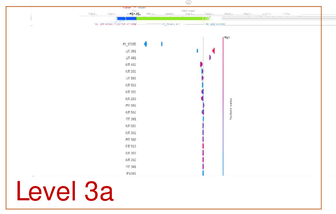


Explanation Dashboard

- **Interactive Dashboard**
- **Layered Explanation**
- **Provenance Logs**


Level 1a


Level 2a


Level 3a

Level	Explanation
1a	When and How Much?
1b	Potential Implications
2a	How is the Anomaly Detected?
2b	Model Training
3a	Model Output (feature interactions)
3c	Model Output (temporal interactions)