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

Moaz Tajammal Hussain
School of Mathematics
Cardiff University
Cardiff, United Kingdom
HussainM26@cardiff.ac.uk

Charith Perera
School of Computer Science and Informtatics
Cardiff University
Cardiff, United Kingdom
PereraC@cardiff.ac.uk

***Abstract*—Deep learning or black-box models are widely used for anomaly detection in Internet of Things (IoT) data streams. We propose a technique to explain the output of a deep learning model used to detect anomalies in an IoT based industrial process. The proposed technique employs dual surrogate models to deliver black box model explanation. We have also developed an interactive dashboard to give further insights into the detected anomaly. The dashboard integrates our proposed deep learning explanation technique with historical logs to explain the detected anomaly for personas with different backgrounds.**

***Index Terms*—Explainable AI (XAI), Internet of Things, Long Short Term Memory Networks, Sensor Data, Anomaly Detection**

I. Introduction

Deep neural networks have shown robust anomaly detection capabilities. They are capable of capturing temporal and multimodal dependencies. Moreover, they allow for minimal manual feature engineering and domain knowledge independent data pre-processing [1]. Conversely, deep learning or 'black box' models [2], are difficult to *explain*.

This work presents a technique to explain the output of a unsupervised deep learning model. The well-known IoT dataset of Secure Water Treatment or SWaT [3] has been used for model training and anomaly detection. Anomalies are detected by monitoring reconstruction errors of LSTM Auto-encoder. LSTM Auto-encoder's (LSTM-AE) output for detected anomalies is then attempted to be explained by training a duo of Random Forest regression models. The surrogate models are trained to replicate the output of the LSTM-AE. We then use SHAP plots to explain the output of the surrogate models. Each surrogate model captures unique dependencies of the deep learning model for the probed output and is decrypted using *TreeSHAP* [4]. Finally, dashboard is designed to answer the questions (when, how, what, and why) associated with the detected anomaly for different personas.

II. Explainable Anomaly Detection

An Auto-encoder consisting of LSTM layers was trained using overlapping time sequenced training data. The breaking down of time series data into overlapping time sequences helps the deep learning model capture temporal dependencies. Model was trained to minimize the mean absolute reconstruction error of the target time sequence. Post model training, using the appropriate configurations, a reconstruction error threshold of 97$^{\text{th}}$ quantile was set after analysing the error distributions.

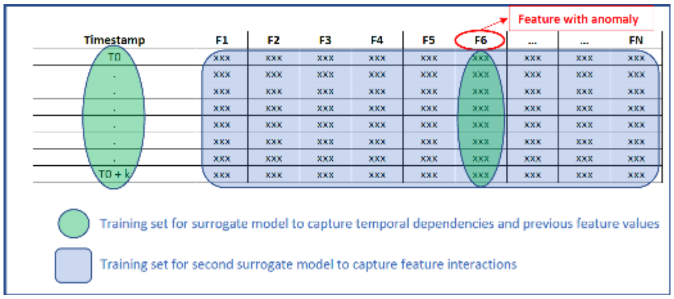

Feature with anomaly
Timestamp
F1
F2
F3
F4
F5
F6
...
...
FN
T0
.
.
.
.
.
T0 + k
xxx
Training set for surrogate model to capture temporal dependencies and previous feature values
Training set for second surrogate model to capture feature interactions

Fig. 1. Illustration of surrogate models input training data.

An anomaly score is then assigned, for readings whose reconstruction error exceeds the defined threshold, indicating the likelihood of the detected anomaly which can be used to appropriately alarm the user.

A. Deep Learning Model Local Explanation

Post-hoc model agnostic local explanation using *TreeSHAP* [4] is employed to generate a low time latency explanation of local surrogate models which replicate the deep learning model output being probed. Two surrogate models are trained using the probed feature output of the LSTM-AE as the common target for both the models but with a variation in the input features. Surrogate model A is trained using temporal components and past values of the feature being probed, whereas surrogate model B is trained using past values of all features. In summary (see Fig. 1 for an illustration),

1) Surrogate model A captures the temporal dependencies and contextual dependencies of the feature with respect to its previous values.
2) Surrogate model B captures the feature interactions.

Both surrogate models are based on random forest regression which is not a 'glass box' model by itself, however, it can easily be explained using *TreeSHAP* [4] and has the capability of learning complex non-linear relationships as opposed to simple regression models. Moreover, as the random forest regression model is composed of a forest of decision trees the

TABLE I
SUMMARY OF GENERATED EXPLANATIONS FOR EACH LEVEL

	Description
1A	When did the anomaly occur and How (much) anomalous was the event?
1B	What are the potential implications of the detected anomaly?
2A	How is the anomaly detected by the system?
2B	How well has the ML model learned the subject feature behaviour from the training?
3A	Why does the ML model predict this particular value for the subject feature in terms of influence of other feature values?
3B	Why does the ML model predict this particular value in terms of influence of the past feature values and temporal components?

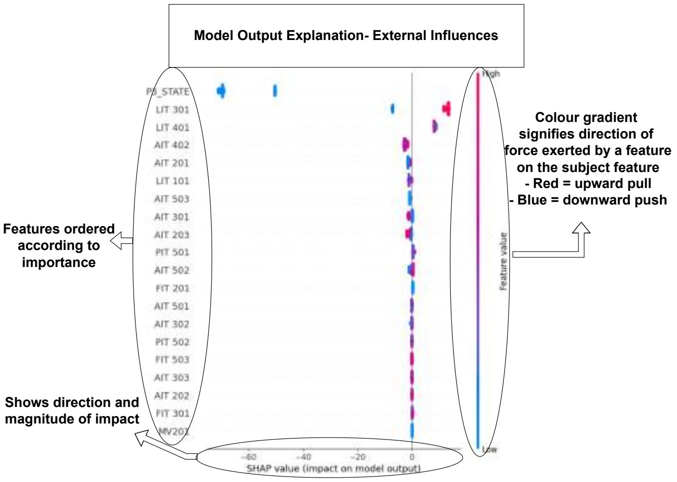
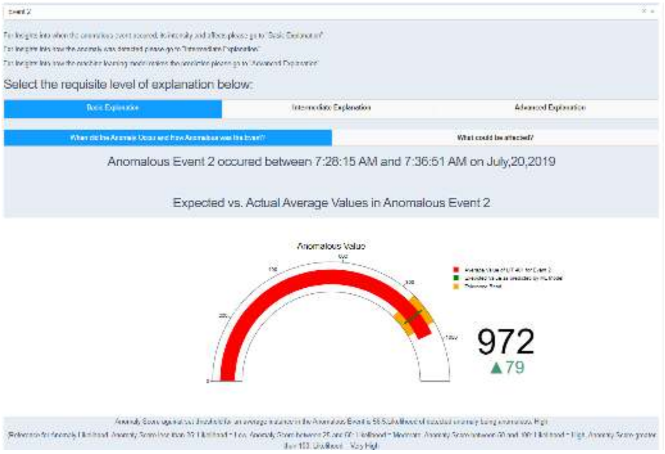


Fig. 2. SHAP Dependency Plot for Surrogate Models.



For Insights into when the anomalous event occured, its intensity and affects please go to 'Basic Explanation'

For Insights into how the anomaly was detected please go to 'Intermediate Explanation'

For Insights into how the machine learning model makes the prediction please go to 'Advanced Explanation'

Select the requisite level of explanation below:

Anomalous Event 2 occured between 7:28:15 AM and 7:36:51 AM on July,20,2019

Expected vs. Actual Average Values in Anomalous Event 2

Anomaly Score against set threshold for an average instance in the Anomalous Event is 56.5 Likelihood of detected anomaly being anomalous: High

Reference for Anomaly Likelihood: Anomaly Score less than 25: Likelihood = Low, Anomaly Score between 25 and 50: Likelihood = Moderate, Anomaly Score between 50 and 100: Likelihood = High, Anomaly Score greater than 100: Likelihood = Very High

Fig. 3. Level 1a Explanation Illustration.

input features do not have to be scaled which makes it simpler both for LSTM-AE model approximation and explanation generation. Lastly, the complexity of generating explanation for the surrogate models is reduced since only a single target feature is approximated by it as opposed to the original model.

SHAP force plot and SHAP dependency plot are used to deliver the local explanation of the black box model output. Force plot shows feature contributions in terms of magnitude and direction in making the model output go higher or lower. Whereas, the dependency plot outputs a summary of how each feature interacted with the feature being probed. Figure 2 illustrate how SHAP Dependency plot is used to explain output of a surrogate model.

The R^2 measure comparison against actual data for LSTM-AE and surrogate models is used to quantify how well the surrogate models replicate the LSTM-AE output.

III. EXPLANATION DASHBOARD

An explanation dashboard [6] is then produced in context of the following framework [5] of a good explanation:

1) Explanations should be contrastive.
2) Selective and focused.
3) Social context should be given major consideration.

Plotly's dash app was used to construct the explanation dashboard owing to the interactive features that it offers. Provenance logs were utilized to produce the explanations in a layered fashion progressing from basic to advanced.

The dashboard was constructed as a single page application divided in two halves. The top half contains important meta data related to the detected anomalies, some useful guidelines for the 'explainee' and some interactive features to let the user select which event or sensor they want the explanation for. The second half of the dashboard consists of three tabs with each tab representing one of the three levels of explanation. While each of the explanation level is further divided into two sub-levels making a total of 6 levels. Each highlighting different aspects of detected anomaly as required by the subject persona. Table I summarizes the explanations provided by each layer. Whereas Figure 3 provide an illustration of how level 1a explanation is delivered using the explanation dashboard.

A guide detailing the layout, flowchart of generating advanced explanation and how each explanation level delivers the explanation to the user can be accessed here: (Link).

REFERENCES

[1] Assendrop J.P., *Deep learning for anomaly detection in multivariate time series data*, 2007.

[2] Linardatos P., Papastefanopoulos V., Kotsiantis S., *Explainable AI: A Review of Machine Learning Interpretability Methods*, Entropy 2021, 2020.

[3] Goh J., Adepu S., Junejo K. N. Mathur A., *A Dataset to Support Research in the Design of Secure Water Treatment Systems*, 2016.

[4] Lundberg, Scott M., Lee S.-I., *A unified approach to interpreting model predictions*, Advances in Neural Information Processing Systems, 2017.

[5] Molnar C., *Interpretable Machine Learning- A Guide for Making Black Box Models Explainable*, Chapter: Model-Agnostic Methods, 2021.

[6] Liao Vera Q., Gruen Daniel, Miller Sarah, *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*, 2021.