Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Hybrid Cyber Threats Detection Using Explainable AI in Industrial IoT

The proposed approach to detect and classify the sophisticated cyber threats is named as hybrid cyber threats abbreviated as “HCTs” in Industrial Internet of Things (IIoT). They are known as complex cyber-attacks that are combination of various different technologies which are used to exploit the vulnerabilities that are present in IIoT systems. These attacks can result in severe consequences like data breaches, facility damage and even other casualties. They want to create a robust intrusion detection system which is capable of recognizing and also classify the HCTs in IIoT environments while also providing the interpretability for the security analysts.

The proposed methodology for HCTs is as followed:

1. Modelling: This is the first step that utilize’s the attention mechanism to make the neural network focus on parts which are to the input data. This attention mechanism we used is proposed by Google which is based on the transformer model [1]. To encode the positional information into the input sequences positional embedding is used along with the attention mechanism. The model used is single-head attention model which is known to have for the encoder-decoder architecture with multiple encoder and decoder layers as shown in the below figure [Fig \_]

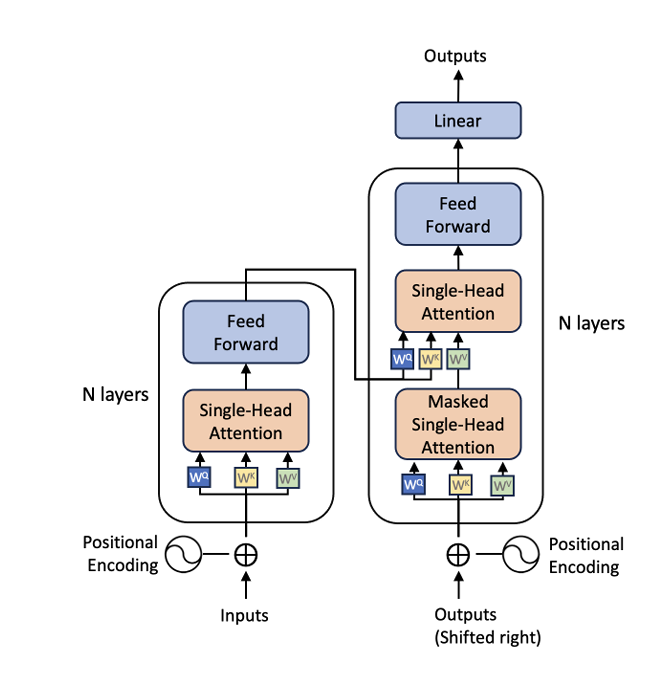


Fig \_ : Attention model [1]

1. Data Processing: This is the second step in this we sort the dataset based on the timestamp and then are divided into two different groups namely:
   1. D1 – Normal and regular which are continuous samples.
   2. D2 – remaining data which may contain both normal and attack samples

Then the data is standarized using z-score normalization.

1. Anomaly Detection: This is the third step, in this the data from D1 is pre-trained using the single-head attention model. Once the model is trained is model it is used to predict the anomaly scores on the data from D2 and the scores is added as new column in the add for D2 this can be seen in algorithm in the below figure [Fig \_]

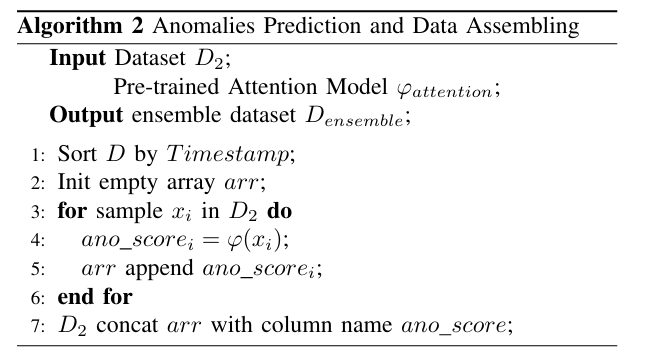
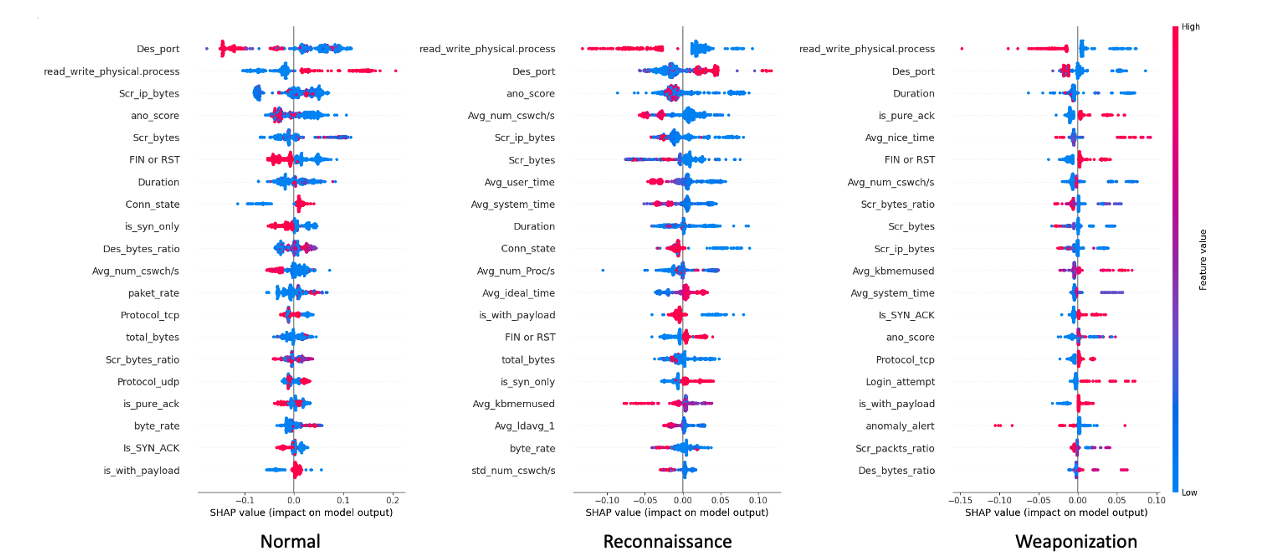


Fig \_: Algorithm2 pseudo code [1]

Attack Classification and Model: In this attack classification phase first we split the assembled dataset (D\_ensemble) which contains the original features along with the predicted anomaly scores which we split into training and testing data sets. The training set is then used to train the Random Forest classifier as an ensemble learning algorithm that can combine multiple decision trees together to improve the performance and reduce overfitting issue. The Random Forest model uses the Classification and Regression Trees for the base learners which is used as base for building decision trees on the Gini impurity criterion by using the process known as bagging which is also known as “bootstrap aggregating”. Random Forest model creates the ensemble of decision trees for each one of the trained data in a random subset of the training data. The final prediction is made as the combination of all the predictions from all trees which are present in the ensemble. The Random Forest model is then trained on the training set of D\_ensemble which now includes the anomaly scores and other required features which are selected from network traffic and the host logs.

The trained Random Forest model can classify attacks effectively but it lacks interpretability of the attacks which is very important to the security analysts for performing the investigation on cyber threats.

To provide explanations SHAP stands for “SHapley Additive exPlanations” , it is a technique based on game theory and Shapley values. A SHAP explainer is created on the trained Random Forest model by computing SHAP values that gives credit or importance to each of the feature for the particular prediction. Then it visualizes the feature importance using the heatmap’s or bar plots depending on the need of the user, this can be seen in the below figure [Fig:\_]. Where the x-axis represents the SHAP value Additionally, It can provide the local explanation’s for each individual sample in the dataset by highlighting the specific features that contributed to the result by model's prediction along with the corresponding SHAP value. These interpretable explanations are then used by security analysts in understanding the causes and the list of the features which contributed to the detected threats which can help in facilitating the investigation and also help in decision-making processes after an IIoT system has been attacked.



Fig\_ : RF Explanation using SHAP

Reference:

1. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.