A Review of An Interpretable Semi-Supervised System for Detecting Cyberattacks Using Anomaly Detection in Industrial Scenarios

Ángel Luis Perales Gómez*[1], Lorenzo Fernández Maimó[1], Alberto Huertas Celdrán[2] and Félix J. García Clemente[1]
[1] Faculty of Computer Science, University of Murcia, 30100 Murcia, Spain
angelluis.perales@um.es; lfmaimo@um.es; fgarcia@um.es
[2]Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, CH-8050, Switzerland
huertas@ifi.uzh.ch

***Resumen*—The Anomaly Detection systems based on Machine Learning and Deep Learning techniques showed great performance when detecting cyberattacks in industrial scenarios. However, two main limitations hinder using them in a real environment. Firstly, most solutions are trained using a supervised approach, which is impractical in the real industrial world. Secondly, the use of black-box Machine Learning and Deep Learning techniques makes it impossible to interpret the decision. This paper proposes an interpretable and semi-supervised system to detect cyberattacks in industrial settings. Besides, we validate our proposal using data collected from the Tennessee Eastman Process. To the best of our knowledge, our system is the only one that offers interpretability together with a semi-supervised approach in an industrial setting. Our system discriminates between causes and effects of anomalies and also achieved the best performance for 11 types of anomalies out of 20 with an overall recall of 0.9577, a precision of 0.9977, and a F1-score of 0.9711.**

***Index Terms*—anomaly detection, deep learning, explainable artificial intelligence, industry applications, root cause analysis.**

Tipo de contribución: Investigación ya publicada

I. Introduction

The industry is transitioning towards the Industry 4.0 paradigm, characterized by automating industrial processes through interconnected smart devices in factories. However, this connectivity to the internet has led to an increase in cyberattacks targeting industrial factories. To address this, the research community has embraced the Anomaly Detection (AD) paradigm to identify specialized cyberattacks affecting the industry. Although AD systems based on Machine Learning (ML) and Deep Learning (DL) techniques are effective, they face limitations in interpretability and the need for labeled data. To mitigate these challenges, a semi-supervised approach is proposed, enabling AD systems to be trained using only normal samples, thus reducing the resources required for training in industrial settings.

This paper reviews [1] and introduces three main contributions: 1) the relationship between the four desirable properties of interpretable ML/DL models and the industrial scenarios; 2) an interpretable and semi-supervised AD system specially designed for industrial scenarios, which use causal inference models to discriminate between causes and effects; and 3) the validation of the AD system in a realistic industrial scenario called Tennessee Eastman Process (TEP) where our solution not only discriminated between causes and effects but also achieved the best performance detection for 11 types of anomalies out of 20 with an overall recall of 0.9577.

II. Desirable properties of Interpretable ML/DL in Industrial Scenarios

This section describes the properties of interpretability methods and their relationship to industrial scenarios:

- Expressive power. Refers to the mechanism by which the decisions reached by the models are interpreted. In industrial environments, it is desirable for the selected mechanism to be easily and quickly understood.
- Translucency. Relationship between the interpretability mechanism and the model, including its parameters. The higher the translucency, the greater the dependency between the interpretability technique and the AD model. In industrial scenarios, the translucency needs to be low because the AD model can be changed by another.
- Portability. The range of ML/DL models applicable to the interpretability mechanism. In industrial settings, highly portable mechanisms are preferred because they can be used with a wide range of models.
- Algorithmic complexity. Time required to apply the interpretability mechanism. In industrial settings, the complexity needs to be low to facilitate quick detection.

III. A Semi-Supervised and Interpretable Approach to Anomaly Detection in Industrial Scenarios

The solution proposed in this work is based on the typical steps to training a semi-supervised model: data preprocessing, feature filtering, feature extraction, selecting the anomaly detection method, and validation. However, we introduce novel approaches in feature extraction and anomaly detection method steps as well as a new step for interpretability.

Feature extraction. In this step, we propose to extract features based on the dominant values from the autocorrelation and Discrete Fourier Transform (DFT).

Anomaly Detection Method. Our system is designed to detect anomalies when a specific threshold, determined by the largest z-score of the prediction error, is exceeded. To filter outliers from the z-score, we propose using the Inter-Quartile Range (IQR). Specifically, we filter out all values that exceed 1.5 times the IQR from the third quartile.

ISBN:978-84-09-62140-8

Tabla I
RECALL RATES FOR THE DIFFERENT WORKS (IN BOLDS THE SOLUTION THAT ACHIEVED THE BEST RESULT)

#	FCNN	DBN	GAN-DBN	GAN-SRCC-DBN	GAN-PCC-DBN	GAN-MI-DBN	FCNN	IPCA	Ours
0	0.21	0.941	0.989	**0.997**	0.955	0.994	-	-	0.9575
1	0.98	0.978	0.997	0.997	0.996	0.993	**1**	**1**	0.9917
2	0.99	0.951	0.981	0.998	0.984	0.985	0.9951	0.9954	**1**
3	0.25	0.208	0.241	0.330	0.402	0.283	-	-	-
4	0.98	0.972	0.983	0.992	0.988	0.977	**1**	**1**	**1**
5	0.98	0.936	0.938	0.967	0.974	0.971	**1**	**1**	**1**
6	**1**	0.992	0.999	0.999	0.997	0.991	**1**	**1**	**1**
7	**1**	0.957	0.971	0.973	0.981	0.971	**1**	**1**	**1**
8	0.98	0.934	0.943	0.950	0.954	0.935	0.9806	0.9846	**0.9990**
9	0.27	0.191	0.201	**0.422**	0.410	0.280	-	-	-
10	0.89	0.946	0.962	0.965	**0.974**	0.961	0.9396	0.9157	0.9542
11	0.94	0.948	0.948	0.971	0.955	0.970	0.9720	0.9567	**0.9838**
12	0.99	0.965	0.971	0.993	0.990	0.994	0.9869	0.9978	**0.9997**
13	0.96	0.917	0.931	0.957	0.940	0.955	0.9578	**0.9623**	0.8982
14	0.98	0.927	0.934	0.944	0.948	0.920	0.9997	**1**	**1**
15	0.31	0.471	0.510	**0.633**	0.552	0.603	-	-	-
16	0.92	0.775	0.907	0.924	0.919	0.902	0.9541	0.9570	**0.9926**
17	**0.98**	0.780	0.822	0.935	0.901	0.926	0.9593	0.9684	0.9032
18	**0.98**	0.914	0.943	0.955	0.961	0.949	0.9415	0.9515	0.8107
19	0.90	0.945	0.959	0.982	0.970	0.981	**0.9918**	0.9917	0.8669
20	0.91	0.881	0.900	0.952	0.932	0.9600	0.9362	**0.9650**	0.8804
Overall	0.8285	0.8347	0.8586	0.8970	0.8915	0.8810	0.9773	**0.9792**	0.9577

Interpretability. We propose grouping features in different sets depending on sensor/actuator they come from. Then, to discover the relationship between features, we build a graph using the Distance Correlation (dCor) to compute the adjacency matrix. Finally, to create a directed graph, we used the Information-Geometric Causal Inference (IGCI) technique that indicates the cause and the effect between features.

IV. EXPERIMENTAL RESULTS

This section details the steps to validate the AD system proposed in this work using TEP testbed.

Data preprocessing. First, we split the data into training, validation and test sets, excluding anomalies 3, 9, and 15 because the introduced perturbations were not sufficient to produce anomalies and, therefore, they are wrong labeled.

Feature Filtering. In this step, using the Pearson's correlation, we did not find any data leakage between features and the label. Subsequently, we did not remove any feature since we did not find any features with zero variance. Finally, we performed the Kolmogorov-Smirnov test and concluded that the distribution is preserved in training, validation, and test.

Feature Extraction. In this step, we selected the three dominant frequencies obtained from the DFT, along with the three dominant coefficients returned by the autocorrelation technique, to be added to the dataset for each feature. After applying this step, the dataset contained 364 features.

Anomaly Detection Method. To select the best model with the best hyper-parameters combination, we used a random search. The best model was a 1-dimensional Convolutional Neural Network (CNN) with two CNN and two fully connected layers, utilizing the LeakyReLU as activation function with an alpha of 0.5 and an output layer of 21 neurons. Finally, we computed the threshold to detect anomalies using the z-score and the validation set.

Interpretability. In this step, we grouped the original, autocorrelation, and frequency features by the sensor/actuator from which they were obtained. However, in industrial scenario, the values of all features are correlated in a normal behavior. Therefore, we only consider a pairs of features when the dCor in validation exceed the dCor in normal behavior. Finally, we used IGCI to determine the cause and the effect.

Validation. We used Precision, Recall, and F1-score, which are the common metrics to measure the detection performance in AD. Table I shows the recall rates per anomaly type achieved by different works. Regarding the global performance, our solution achieved an excellent precision rate with a result of 0.9974, a recall of 0.9577, and an F1-score of 0.9771.

V. CONCLUSIONS

This work presented an interpretable and semi-supervised system to detect cyberattacks on industrial control systems. The core of our proposal is the introduction of a new step focused on interpretability that uses dCor and IGCI to determine the effect of each anomaly. Additionally, we provide specific steps for both feature extraction and AD model selection. The solution was validated using TEP and achieved state-of-the-art in most anomaly types, obtaining the third-best overall recall.

ACKNOWLEDGEMENTS

This work has been funded under Grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033, NextGenerationEU/PRTR, UE, Grant PID2021-122466OB-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE, by the strategic project CDL-TALENTUM/DEFENDER from the Spanish National Institute of Cybersecurity (INCIBE), by the Recovery, Transformation and Resilience Plan, Next Generation EU, by the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberForce (CYD-C-2020003), and by the University of Zurich (UZH).

REFERENCIAS

[1] Perales Gómez, Á. L., Fernández Maimó, L., Huertas Celdrán, A., & García Clemente, F. J. (2023). An interpretable semi-supervised system for detecting cyberattacks using anomaly detection in industrial scenarios. *IET Information Security*.