Robust Anomaly Detection in Cyber Physical System using Kullback-Leibler Divergence in Error Distributions

Simon Woo, Younggeun Kim, Jinwoo Cho
Sungkyunkwan University
Suwon, S. Korea
{swoo,dudrms33,caramel}@g.skku.edu

Shahroz Tariq, Sangyup Lee
Stony Brook University
Stony Brook, NY
{shahroz,sangyup}@stonybrook.edu

Jeong-Han Yun, Jonguk Kim, Hyoung Chun Kim
The Affiliated Institute of ETRI
Daejeon, S.Korea
{dolgam,jongukim,khche}@nsr.re.kr

ABSTRACT

We investigate anomaly detection in Cyber-Physical System (CPS), where anomalies are often attacks to CPS to disrupt the operations of critical infrastructures. We use secure water treatment (SWaT) systems dataset, where normal and attack states are simulated in the water tanks. Among different types of anomalies, we focus on detecting contextual anomaly, which can be difficult to detect with Out-Of-Limit threshold method. Recent research shows promising result in detecting anomalies from analyzing error distributions from the machine learning classifier. In a similar way, we statistically analyze prediction error patterns from RNN and MDN classifiers to detect anomalies. First, we generate anomaly scores with Local Outlier Factor (LOF) and remove point anomalies. With the fixed window size, empirical probability distribution is estimated, and we apply the sliding window to measure the difference of probability distributions between the other windows. To measure the difference efficiently between anomalies and normal data, we use Kullback-Leibler divergence. Our preliminary result shows that we can effectively detect contextual anomalies compared with Nearest Neighbor Distance (NND) approach.

KEYWORDS

Anomaly detection, Cyber-Physical System, Kullback-Leibler divergence

ACM Reference Format:
Simon Woo, Younggeun Kim, Jinwoo Cho, Shahroz Tariq, Sangyup Lee, and Jeong-Han Yun, Jonguk Kim, Hyoung Chun Kim. 2018. Robust Anomaly Detection in Cyber Physical System using Kullback-Leibler Divergence in Error Distributions. In *MileTS '19: 5th KDD Workshop on Mining and Learning from Time Series, August 5th, 2019, Anchorage, Alaska, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Anomaly detection in Cyber-Physical System (CPS) investigates the identification of unusual behaviors that are not exhibited under normal operating condition. In CPS, these anomalies may result from attacks on the control, network, or cyber-physical elements. At the same time, temporal high spike anomalies may be caused by hardware faults, operator errors, or even misconfigurations in the software. Therefore, detecting and separating anomalies from these temporal glitches or point anomalies are paramount and challenging tasks because false positives can be very expensive. For example, a false alarm can cause all running critical infrastructures to stop for inspection, or evacuate operators or people due to the possible dangers that can be caused by the attacks. Therefore, the capability to detect true anomalies apart from temporal operational glitches is critical. Typically, an anomaly detection technique can be a rule-based or statistical learning-based approach. The rule-based approach typically employs the Out-Of-Limit (OOL) threshold so that values measured above the OOL are considered as anomalies. Statistical machine learning based approaches typically fit a model to the data and detect anomalies, if prediction error is greater than the threshold. However, in many practical situations, there exist measurements that are greater than a threshold value but are normal. In addition, there will be contextual anomalies, which are less than the threshold but they are abnormal.

In this work, we hypothesize that these anomalies are originated from other probability distributions than the original nominal data distribution, because they are cyber attacks. Therefore, rather than using pre-defined threshold values, we aim to investigate patterns in time series and analyze its empirical probability distribution to detect contextual anomalies. In this work, we propose a statistical learning based approach to focus on detecting contextual anomalies. We take a similar approach to that of Machine Learning Anomaly Detection (MLAD) by Kaspersky Lab [6, 7], which exploits correlations in industrial traffic signals. For example, MLAD can train a recurrent neural network to recognize signal behavior under normal operating conditions and the data is presented as a multivariate time series. Next, MLAD predicts the values of all signals in real time for a certain future time interval and compares them with observed values. If the prediction error is greater than a statistically matched threshold defined at the training stage, MLAD can detect anomalies. We use dataset from the Secure Water Treatment (SWaT) testbed [12] collected by iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design. Then, we generate anomaly scores with the Local Outlier Factor (LOF) and calculate the probability distribution to detect the contextual anomalies. With the fixed window size, we estimate the empirical probability distribution and slide the window to measure the difference of probability distributions between previous and current time series. To measure the difference efficiently between time windows, we employ Kullback-Leibler (KL) divergence. Our preliminary result shows that we can effectively detect contextual

MileTS '19, August 5th, 2019, Anchorage, Alaska, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

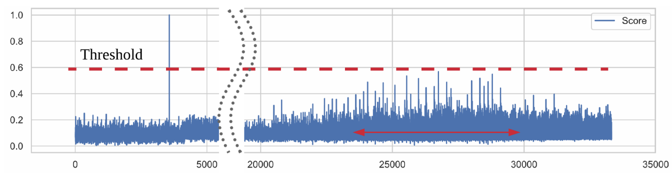

Figure 1: Example of contextual anomalies (red arrow) that are below the threshold (red dotted line), where the X-axis indicates time instances and the Y-axis is predicted error from neural networks. Thus, contextual anomalies are not detected by OOL method.

anomalies based on error distribution compared to the previous Nearest Neighbor Distances (NND) approach [20].

2 RELATED WORK

Anomaly detection on multivariate time series data: Anomaly detection in multivariate time series data is a challenging task and numerous approaches have been proposed in the past few years to tackle this problem. When identifying anomalies in Cyber-Physical Systems (CPS), the first-order approach can be implemented by building a knowledge base, when comprehensive and accurate domain knowledge are available [14, 15, 18, 19]. However, in modern CPS, developing a knowledge base from a large number of variables in a complex CPS is challenging. Recently, data-driven approaches such as deep-learning or unsupervised clustering-based methods, which do not require broad and specific knowledge of the domain, have been developed [5, 8–11]. These methods can learn and derive information and patterns from data, not using much expert domain knowledge.

Anomaly detection with Deep learning: As an extension of the data-driven approach, deep learning has been applied for detecting anomalies in time series data. Malhotra et al. [17] proposed a Long Short Term Memory (LSTM) networks based on Encoder Decoder scheme for anomaly detection. This model learns to reconstruct 'normal' time series behavior, and uses reconstruction error to detect anomalies. With this model, they were able to achieve better generalization capability than distance-based methods. In addition, Zhai et al. [21] developed Deep Structured Energy-Based Models (DSEBMs) which connects Energy-Based Models (EBMs) with a regularized autoencoder to eliminate the need for complicated sampling method. The model uses energy scores and reconstruction errors to decide anomalies. However, despite their effectiveness, these methods cannot jointly analyze temporal dependency, noise resistance, and the interpretation of the severity of anomalies [21]. On the other hand, Zhang et al. [22] made Multi-Scale Convolutional Recurrent Encoder-Decoder (MSCRED) to solve these problems, where it constructs multi-scale signature matrices to characterize multiple levels of the system statuses in time steps. However, most of the deep learning-based methods resort on some types of thresholding (Out-Of-Limit) to determine anomalies with the predicted error or reconstruction error.

Determining anomaly from error distributions: To detect contextual anomalies, Machine Learning Anomaly Detection (MLAD) by Kaspersky Lab [6, 7] was proposed to explore correlations in industrial traffic signals. MLAD trains the RNN to recognize signal behaviors under normal operating conditions and predicts the values of all signals in real time for a certain future time interval. If the prediction error is greater than a statistically matched threshold defined at the training stage, MLAD detects anomalies. In our work, we also explore the distributions of errors from RNN, because the probability density function (PDF) of errors can possibly express those anomalies as a value close to the mean. In other words, the PDF with anomalies in a given time window will have a high spike at the mean value, making it different from the PDF with normal data. Therefore, we expect that we can improve the detection of contextual anomalies by leveraging and comparing distance between PDFs from normal vs. abnormal data. In our study, we use Kullback-Leibler divergence, where KL divergence uses a distribution instead of a single data point. Therefore, it is useful to detect a continuous attack which is close to mean as shown by other research [1].

3 OUR APPROACH

The goal of our approach is to predict the values of signals from error distributions of neural networks we trained from CPS dataset. Specifically, we examine if the prediction error is greater than a statistically matched threshold defined at the training stage, we can detect anomalies. The main difference from MLAD [6, 7] is that we first remove point anomalies and then apply KL divergence to separate the anomalous PDF from the normal PDF.

Dataset: We use data collected from the Secure Water Treatment (SWaT) [12] collected by iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design. The dataset has 11 continuous water purification operations of the plant control network in SWaT. Out of 11 days, the amount of normal operation data is 7 days, while data collected from attack scenarios consist of 4 days. All network traffic, sensor and actuator data in the control network were collected during this period. The properties of this SWaT dataset are summarized as follows:

- Network Traffic and Cyber-Physical Properties: Not only the dataset contains all the network traffic captured through 11 days, but it also includes all the values gathered from all the 51 sensors and actuators available in SWaT.
- Data Labels: Class label indicating either normal or abnormal behavior.
- Attack Scenarios: SUTD has developed an attack model, which can generate attacks in the dataset. The attack model considers the intent space of a CPS as an attack model, where 36 attacks were launched throughout 4 days.

In particular, the water purification process in SWaT system is composed of 6 sub-stations with more than 1 variable each, and it is indicated as P1 through P6 as follows:

- P1: Raw water supply and storage
- P2: Pre-treatment
- P3: Ultrafiltration and backwash
- P4: De-Chlorination System
- P5: Reverse Osmosis (RO)
- P6: RO permeate transfer, UF backwash and cleaning

Error dataset generation from neural networks: We used a Recurrent Neural Network (RNN) with Mixture Density Network (MDN) [2] to obtain the prediction errors. The main benefit of using MDN is that the model can learn the distribution of the data, where

Table 1: P1 substation's error dataset after applying LOF on the extracted features, where the column time series represents the original features' name.

Window #	Sum	Mean	...	Max	Time Series	Outlier Score
1	-0.03	-0.03	...	0.01	MV101	1.12
2	-0.06	-0.06	...	-0.14	P101	1.00
3	-0.08	-0.08	...	-0.08	P102	1.06

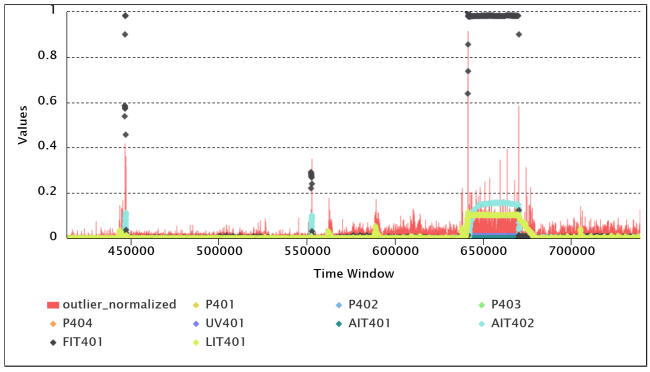

Figure 2: Local Outlier Factor score vs the values of each data feature in the same time window of P4 sub-station.

a sequence of inputs may lead to several distinct future possibilities. In order to predict normal sequences, the model is trained with normal operation data in the SWaT dataset. Past 90 seconds were used as an input to the model to predict the future 10 seconds as an output. The prediction error is calculated by subtracting the actual value with the predicted sequence value. The prediction error of this model is then collected and saved as an error dataset.

Time window conversion: We created the equal size time windows and aggregated these values to extract features for each time window such as sum, mean, median, min, max, kurtosis and skewness.

Detecting and removing point anomalies: We used Local Outlier Factor (LOF) [3] to generate an outlier score for each time window in the dataset as shown in Fig. 2 and then grouped the outlier data separately from the nominal data using K-means clustering [16]. The outlier group is then removed from the dataset so that there are only the contextual anomalies left in the dataset.

Kullback-Leibler divergence: We assume that if an attack is performed during normal operating condition, its distribution of time series will be different from the distribution of times series in normal operation conditions. Therefore, we use *Kullback-Leibler (KL) divergence* [13] to measure the difference in distribution between normal vs. attack instance. Let $X_1 : (\Omega, S) \rightarrow (\mathbb{R}, \mathcal{B})$ and $X_2 : (\Omega, S) \rightarrow (\mathbb{R}, \mathcal{B})$ be discrete random variables with mass f_1 and f_2, and let $\mathcal{X}$ be a support of X_1. KL divergence D_{KL} between two different probability distributions f_1 and f_2 is defined as follows:

$$D_{KL}(f_1|f_2) := E_{X_1}\left[\log \frac{f_1(X)}{f_2(x)}\right] = \sum_{x\in\mathcal{X}} f_1(x) \log \frac{f_1(x)}{f_2(x)}. \quad (1)$$

KL divergence has an important property: $D_{KL}(f_1|f_2) = 0$, if and only if two distributions are same almost everywhere. However, it

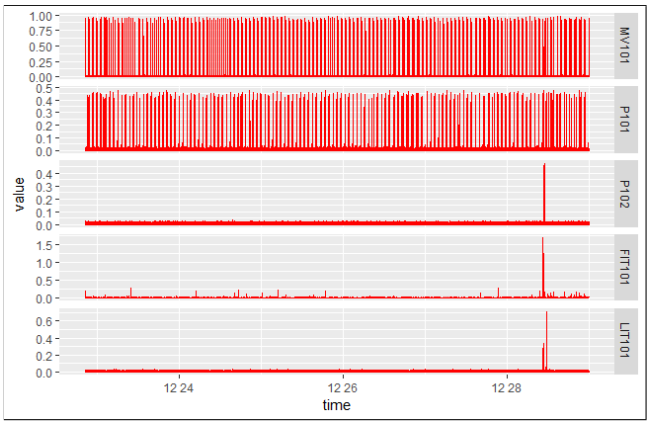

Figure 3: Prediction error produced from neural networks measured at $P1$ (raw water supply and storage).

is different with the distance in that KL divergence is not *symmetric*, i.e. $D_{KL}(f_1|f_2) \neq D_{KL}(f_2|f_1)$. Although KL divergence is not a distance, its simplicity makes it useful for anomaly detection, as shown in Afgani et al. [1]. In this work, we also employ KL divergence to detect attack distributions and compared to the baseline threshold method.

4 EXPERIMENT

Local outlier detection and removal: In order to estimate the manifest PDF of each time window, local outliers were removed, which are typically high spike point anomalies. We used LOF with lower bound (K-minimum) of 10 and upper bound (K-maximum) of 20. Using LOF, we computed outlier scores for each time window. The K-means clustering algorithm with $K = 2$, max runs = 10 and *Bregman divergence* is used as a distance metric to compute the outlier scores between the cluster the dataset into outliers and nominal clusters. These nominal clusters are further investigated to detect the contextual anomalies. After deleting local outliers, time series within the sub-stations were grouped together. For example, 5 time series from $P1$, including $MV101$, $P101$, $P102$, $FIT101$ and $LIT101$ in SWAT data, were aggregated. We believe it will be more reliable to utilize all aggregated sensor data than to consider only a single time series, when assuming sensors in the same station are correlated with one another. Next, we present the resulting prediction errors from neural networks in Fig. 3. Although point anomalies can be observed as local spikes, it is difficult to distinguish contextual anomalies visually from Fig. 3. For example, $MV101$ and $P101$ have numerous local spikes, which are not real anomalies. To eliminate those spikes, the summation of absolute distances between these time series is calculated as follows:

$$\sum_{i\neq j}^{5} |X_{it} - X_{jt}|, \quad (2)$$

where t indicate time variable. Given that each sensor can catches signals simultaneously, Eq. 2 can capture the correlation between the time series. Moreover, the moving average method is used to produce a smoother time series which can generate a more meticulous PDF. In addition, time series seasonal differencing [4] was applied to offset seasonality. And data is transformed using

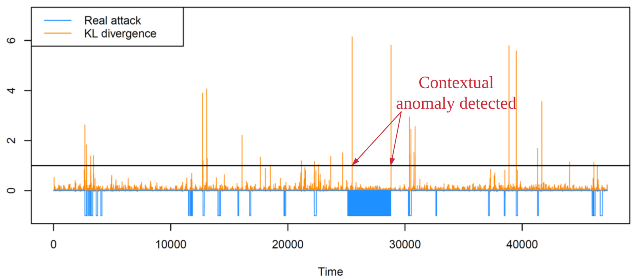

Figure 4: KL divergence with size of window 180 and interval 10 measured at $P1$ (raw water supply and storage).

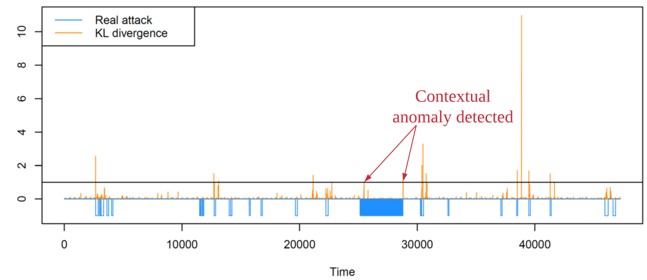

Figure 5: KL divergence with size of window 360 and interval 10 measured at $P1$ (raw water supply and storage).

calculation, $Y_t = X_t - X_{t-l}$, where t is a time index and l is time lag. Then, Y_t has no seasonal effect with l lag size. Since one hour effect was strong, 3600 lag differencing was used. Lastly, we varied window sizes (interval) of PDFs to detect contextual anomalies.

KL divergence calculation: To compute KL divergence, we set the size of time series windows equally. Since the error data was generated by a neural network model with the size 90, multiples of 90 were used as the window size. With the data points within the window, the empirical PDF was estimated using normal kernel smoothing. Also, KL divergence was computed with the neighboring PDF by Eq. 1, where f_1 is the previous PDF which represents a prior distribution, and f_2 is the current PDF which is a posterior distribution. As a result, KL divergence can measure correlations between PDF's within 1 time lag.

5 RESULT

We present anomaly detection performance using KL divergence values for $P1$ (raw water supply and storage as defined in Section 3) with the window sizes of 180 and 360 in Fig. 4 and 5, respectively. We used the fixed time interval of 10, where the X-axis is the time. In the Y-axis, orange lines indicate KL divergence values and blue lines are actual attacks instances. We also plot the classical threshold values (black lines) as a baseline method for comparisons in each figure.

In our approach, we determine the specific instance is an anomaly, when KL divergence value is relatively high. As shown in Figs. 4 and 5, there exists one contextual anomaly (shown as solid blue rectangle), starting from 25, 000 to 29, 000. As shown in each figure, our approach detects the contextual anomaly effectively, showing the high KL divergence value at the start and the end of contextual anomaly interval. However, we can observe that a classical baseline threshold method shown in a black line fails to detect all the contextual anomalies. We also present a more detailed zoomed region of contextual anomalies that occurred between 25, 000 and 29, 000 in Fig. 6. As shown in Fig. 6, KL divergence can detect the beginning and the end of the contextual anomaly effectively.

Comparison to NND: Additionally, we simulated and compared our KL divergence to Nearest Neighbor Distances (NND) based method by Yun et al. [20], whose research aims to identify anomalies using Euclidean distance. Figures 7 and 8 show anomaly detection performance using NND for the same time series measured at $P1$. The calculated NNDs are shown in orange and actual attacks are shown in blue, and they are presented in the Y-axis. Similarly, the X-axis represents the time. As shown in Figs. 7 and 8,

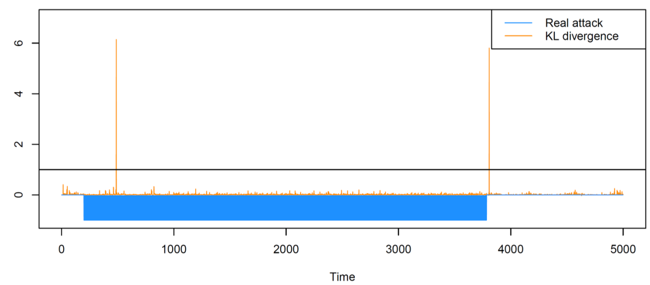

Figure 6: Contextual anomaly detection using KL divergence with size of window 180 and interval 10 at $P1$ (raw water supply and storage).

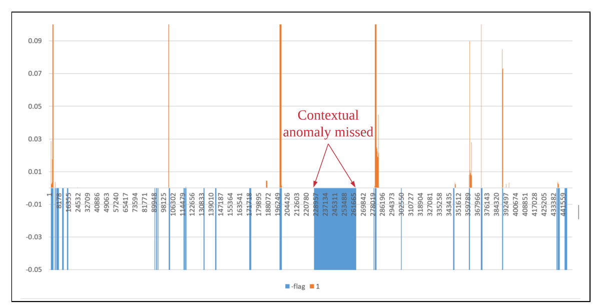

Figure 7: NND with window size 1 measured at $P1$ (raw water supply and storage).

a blue rectangular region is a contextual anomaly and it cannot be detected by NND based approach. We believe that the main reason of NND's poor performance is due to the fact that Euclidean distance can capture and produces a large distance for the point anomalies. However, NND fails to produce the large distance for contextual anomalies. Further research is needed to compare a side-by-side anomaly detection performance between different distance measures using more dataset.

Therefore, our study need to develop a method to classify contextual attack. Moreover, it is difficult to compare performance of KL divergence with classical method, since it converts original data into shrink data such as PDF. In this sense, delicate measure to show performance should be recommended. Nonetheless, KL divergence method can detect contextual anomalies quickly with a simple criterion.

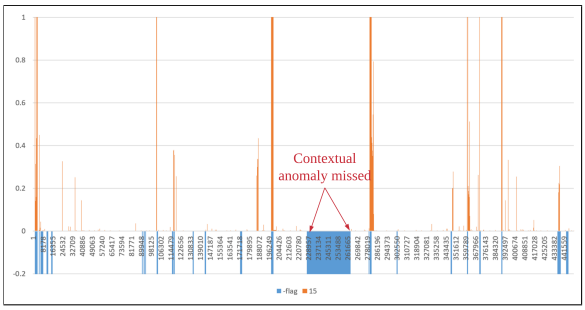

Figure 8: NND with window size 15 measured at $P1$ (raw water supply and storage).

6 CONCLUSION AND FUTURE WORK

We focus on detecting the contextual anomaly, using error patterns produced from a Recurrent Neural Network (RNN) with Mixture Density Network (MDN). Our work shows promising results in detecting a contextual anomaly by measuring differences in the distribution in a fixed time window, using KL divergence. We show that measuring difference in distributions is more effective way to detect anomalies than that of actual distances. However, further research is needed to more effectively differentiate the cases, where contextual and point anomalies coexist.

REFERENCES

[1] Mostafa Afgani, Sinan Sinanovic, and Harald Haas. 2008. Anomaly detection using the Kullback-Leibler divergence metric. In *2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies*. IEEE, 1–5.

[2] Christopher M Bishop. 1994. *Mixture density networks*. Technical Report. Citeseer.

[3] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *ACM sigmod record*, Vol. 29. ACM, 93–104.

[4] Peter J Brockwell, Richard A Davis, and Matthew V Calder. 2002. *Introduction to time series and forecasting*. Vol. 2. Springer.

[5] Ricardo JGB Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10, 1 (2015), 5.

[6] Pavel Filonov, Fedor Kitashov, and Andrey Lavrentyev. 2017. Rnn-based early cyber-attack detection for the tennessee eastman process. *arXiv preprint arXiv:1709.02232* (2017).

[7] Pavel Filonov, Andrey Lavrentyev, and Artem Vorontsov. 2016. Multivariate industrial time series with cyber-attack simulation: Fault detection using an lstm-based predictive data model. *arXiv preprint arXiv:1612.06676* (2016).

[8] Sylvain Fuertes, Gilles Picart, Jean-Yves Tourneret, Lotfi Chaari, André Ferrari, and Cédric Richard. 2016. Improving Spacecraft Health Monitoring with Automatic Anomaly Detection Techniques. In *14th International Conference on Space Operations*. 2430.

[9] Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. 2004. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3. IEEE, 430–433.

[10] Zengyou He, Xiaofei Xu, and Shengchun Deng. 2003. Discovering cluster-based local outliers. *Pattern Recognition Letters* 24, 9-10 (2003), 1641–1650.

[11] Tsuyoshi Idé, Spiros Papadimitriou, and Michail Vlachos. 2007. Computing correlation anomaly scores using stochastic nearest neighbors. In *Seventh IEEE international conference on data mining (ICDM 2007)*. IEEE, 523–528.

[12] Sridhar Adepu (iTrust Centre for Research in Cyber Security Singapore University of Technology and Design). 2019. SWaT Secure Water Treatment dataset description. https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/#swat

[13] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[14] Teresa F Lunt, R Jagannathan, Rosanna Lee, Sherry Listgarten, David L Edwards, Peter G Neumann, Harold S Javitz, and Al Valdes. 1988. Ides: The enhanced prototype-a real-time intrusion-detection expert system. In *SRI International, 333 Ravenswood Avenue, Menlo Park*. Citeseer.

[15] Teresa F Lunt, Ann Tamaru, and F Gillham. 1992. *A real-time intrusion-detection expert system (IDES)*. SRI International. Computer Science Laboratory.

[16] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.

[17] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. 2016. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148* (2016).

[18] Mark Rolincik, Michael Lauriente, Harry C Koons, and David Gorney. 1992. An expert system for diagnosing environmentally induced spacecraft anomalies. (1992).

[19] Michael M Sebring. 1988. Expert systems in intrusion detection: A case study. In *Proc. 11th National Computer Security Conference, Baltimore, Maryland, Oct. 1988.* 74–81.

[20] Jeong-Han Yun, Yoonho Hwang, Woomyo Lee, Hee-Kap Ahn, and Sin-Kyu Kim. 2018. Statistical Similarity of Critical Infrastructure Network Traffic Based on Nearest Neighbor Distances. In *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 577–599.

[21] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. 2016. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717* (2016).

[22] Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V Chawla. 2018. A Deep Neural Network for Unsupervised Anomaly Detection and Diagnosis in Multivariate Time Series Data. *arXiv preprint arXiv:1811.08055* (2018).