Cyberattack Detection for Cyber Physical Systems Security – A Preliminary Study

Weizhong Yan[1], Lalit Mestha[2], Justin John[3], Daniel Holzhauer[4], Marc McKinley[5], and Masoud Abbaszadeh[6]

[1]*AI and Machine Learning, GE Global Research Center, Niskayuna, NY 12309*
yan@ge.com

[2,3,4,6]*Controls and Optimization, GE Global Research Center, Niskayuna, NY 12309*
{lalit.mestha, justin.john, daniel.f.holzhauer, abbaszadeh}@ge.com

[5]*GPS Technology, GE Power, Salem, VA 24153, USA*
marc.mckinley@ge.com

ABSTRACT

Cyber-physical systems (CPS) security has become an increasingly important research topic in recent years. Geared towards more advanced cyberattack detection techniques as part of strategies for enhancing the security of CPS, in this paper we propose a machine learning based cyber-attack detection scheme. The proposed scheme is a physical-domain technique; specifically, it assumes the physical measurements of the system carry sufficient information for capturing the system behavior, thus can be used for differentiating normal operation and attacks. CPS are complex in nature and the number of physical measurements available for CPS is often overwhelmingly high. Thus, accurately modeling CPS' dynamic behavior, more importantly, distinguishing normal and adversary activities based on the large number of physical measurements, can be challenging. To address the challenge, we have focused our research effort on feature engineering, that is, to intelligently derive a set of salient signatures or features from the noisy measurements. We make sure the derived features are more compact and, more importantly, have more discriminant power than the original physical measurements, thus enabling us to achieve more accurate and robust detection performance. To demonstrate the effectiveness of the proposed scheme, in our experimental study we consider gas turbines of combined cycle power plants as the cyber-physical system. Using the data from the high-fidelity simulation we show that our proposed cyberattack detection scheme is able to achieve high detection performance.

Weizhong Yan et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1. INTRODUCTION

Cyber-physical systems (CPS) are referred to an integral system of computation, networking and physical elements, having strong interaction between cyber and physical domains (Khaitan and McCalley, 2015). CPS provide the foundation of numerous critical infrastructure, such as, transportation networks, electric power distribution networks, and water and gas distribution networks. With the advent of Internet of things (IoT), more and more devices with security vulnerabilities are linked to CPS. Thus, ensuring CPS security (US DHS, 2018; Mitchell and Chen, 2014) becomes ever increasingly important, especially since the Stuxnet attack in 2010 (Falliere, et al., 2011).

Current intrusion detection systems (IDS), primarily designed for conventional Information Technology (IT) systems, are not effective for CPS; that is, cyber-threats can still penetrate through the IT protection layer and reach the physical "domain". Such attacks, if not detected and neutralized, can diminish the performance of the control system and may cause total shut down or catastrophic damage to the physical assets. For example, Stuxnet that penetrated the SCADA system of Iranian nuclear facilities and caused centrifuge motors to change their spinning frequency, which destroyed centrifuges (Falliere, et al., 2011).

Physical-domain cyberattack detection (pdCAD) techniques that perform detection by leveraging physical properties of the physical asset or process have proven to be more effective in securing CPS from malicious attacks (Urbina, et al., 2016).

A majority of the existing pdCAD methods involve explicitly modeling the "normal behavior" of the physical system based on the physics and declaring attack if the model behavior deviates from the normal behavior. For those pdCAD methods, building a high-fidelity model is the key to success.

State space models are predominantly used for building the physics-based models, especially when dealing with control systems, for example, in Ozay, et al. (2013), Kosut, et al. (2011), and Urbina, et al., (2016). The attack detection (AD) methods involving state space models are also called "state estimation-based detection".

More recently, machine learning technology has been adopted for pdCAD. Machine learning based attack detection methods generally do not need to explicitly model "normal behavior" of the system[1]. Rather the cyberattack detection problem is formulated as a classification problem, that is, to classify normal vs. attacks directly based on the available physical measurements. For example, Ozay et al. (2015) developed a physical layer attack detection framework for false data injection attack detection in smart grids. Others include Wallace et al. (2014), Wang et al (2017).

For real-world CPS applications, performing attack detection directly on physical measurements may be difficult in achieving the desired detection performance. With more and more devices connected to CPS, the CPS becomes more complex and the number of physical measurements can potentially become overwhelmingly large, which makes attack detection even more challenging.

To address the challenge, in this paper, we propose a behavior-based machine learning attack detection scheme. The key component of the proposed scheme is its advanced feature engineering module, which is to intelligently derive a set of salient signatures or features from the noisy measurements. Such derived features not only are more compact and less noisy, but also have more discriminant power than the original physical measurements, thus enabling us to achieve more accurate and robust detection performance. Also, our proposed scheme differs from other existing machine learning based detection methods in that we innovatively adopt the extreme learning machines (ELM), a recently developed machine learning algorithm (Huang, et al., 2012), as our attack detection algorithm.

Power plants are an important CPS application. It is our national interest to have an enhanced resilience of power plants by reducing the risk of operational disruptions due to cyber-attacks, which is exactly what the Cybersecurity for Energy Delivery Systems (CEDS) programs of the Department of Energy (DOE) are aiming for. In this paper we will use the gas turbines of combined cycle power plant as the CPS for demonstrating our proposed attack detection scheme.

The reminder of the paper is organized as follows. The proposed methodology is described in detail in Section 2. Section 3 presents our experimental study and its results, while Section 4 concludes the paper.

[1]Some machine learning-based attack detection models, e.g., one-class classifier-based attack detection, do model the "normal behavior" of the system.

2. PROPOSED CYBERATTACK DETECTION SCHEME

As a physical-domain approach, our proposed behavior-based machine learning attack detection scheme has a key capability of *learning* the dynamic behavior of the physical systems, more precisely the behavior difference between normal operations and adversary attacks, from the physical (sensors, actuators and controllers) measurements. With the increasing complexity of physical systems, the number of physical measurements can be overwhelmingly large, which makes learning from physical measurements very challenging. To address the challenge, as the key technical component of the proposed scheme we turn the raw physical measurements into more salient features and build our attack detection algorithms on the feature space, which results in more accurate detection performance. The overall structure of the proposed scheme is shown in Figure 1. Its two critical components, *feature engineering* and *attack detection modeling*, are described in detail in the following two subsections.

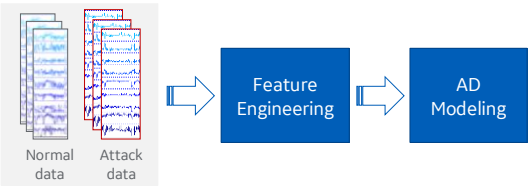

Figure 1. Overall block diagram of the proposed cyberattack detection scheme.

2.1. Multi-Modal Feature Engineering

The primary goal of our feature engineering is to derive a set of signatures or features from the raw physical measurements, which not only have more expressive power for capturing physical behavior, but also have more discriminant power distinguishing normal and malicious activities, thus improving the attack detection performance. Feature engineering is an important process in the pipeline of developing predictive analytical solutions. In literature, there are numerous feature extraction methods available, ranging from traditional statistics based to modern deep representation learning (Yan, 2015). In this paper, considering the unique properties associated with CPS security, we propose using statistics-based and physics-based features. To capture the temporal effects or dynamics of the underlying system, we use a sliding window sliding over time and all features are calculated over the sliding windows. Spatially, we calculate features on individual (univariate) and

multiple (multivariate) measurements. Let's assume we have *n* physical measurements, $s^{(1)}, s^{(2)}, \dots s^{(n)}$, covering sensor, actuator, and control measurements, and the window width for sliding windows is *w*.

2.1.1. Univariate-based features

For each individual measurement, $s^{(i)}$, its windowed segment of measurements at time t is $m_t^{(i)} = s_{t-w}^{(i)}, s_{t-w+1}^{(i)}, \dots, s_t^{(i)}$. Several statistical descriptors can be calculated for this segment of measurements, $m_t^{(i)}$, for example,

$$f_1^{(i)} = median(m_t^{(i)}) \quad (1)$$

$$f_2^{(i)} = std(m_t^{(i)}) \quad (2)$$

$$f_3^{(i)} = max(m_t^{(i)}) \quad (3)$$

$$f_4^{(i)} = max(m_t^{(i)}) - \min(m_t^{(i)}) \quad (4)$$

$$f_5^{(i)} = s_t^{(i)} \quad (5)$$

In addition, for each measurement we also calculate statistics that captures the rate-of-change of the measurements within the window concerned, for example, the maximum absolute rate-of-change, the average of absolute rate-of-change, etc.

2.1.2. Multivariate-based features

To capture the relationships among many variables, we calculate two groups of features using multiple variables. The first group of features are to capture the relations between pairs of measurements, which are either defined by domain experts or learned from the data. The relations can be, for example, the difference or the ratio of two measurements, and can also be covariance of two measurements. The number of features resulted in this group is equal to the number of relations defined.

The second group of features are the residuals of some physical models, that is, the differences between the true measurements and the model predictions. The number of physical models as well as the input and output of each the physical models are application-dependent and determined based on domain knowledge. As a result, we often call this group of features as "physics-based". The physical models can be the first principal models or data-driven models. Assume i^{th} model has *n* inputs, $x^{(i)} \in \Re^n$, and *m* outputs, $y^{(i)} \in \Re^m$. At each time stamp (sample), this model will give us *m* residuals, $R_j^{(i)} = \left|y_j^{(i)} - \bar{y}_j^{(i)}\right|, j = 1,2, \dots, m$. We can directly use these m residuals as the features. Alternatively, for each of the residuals, we can calculate statistics (e.g., mean, standard deviation) of the residuals over the sliding window, *w*, and use the calculated statistics as features.

Concatenating all the calculated features described above, that is, the univariate-based and the multivariate-based features, gives us the final feature set, which is then used as the input to our attack detection model.

2.2. Attack Detection Modeling

Attack detection is a binary classification problem where inputs to the classifier are the features extracted from the physical measurements and output is either normal or attack status. In literature there are numerous types of classifiers available. In this paper we adopt the extreme learning machine (ELM) as the classifier, considering the unique characteristics associated with the ELM. ELM is a special type of feed-forward neural networks introduced by Huang, et al. (2006). ELM was originally developed for the single hidden layer feedforward neural networks (SLFNs) and was later extended to the generalized SLFNs where the hidden layer needs not be neuron alike (Huang, et al., 2012). Unlike in traditional feed-forward neural networks where training the network involves finding all connection weights and bias, in ELM, connections between input and hidden neurons are randomly generated and fixed, that is, they do not need to be trained. Thus, training an ELM becomes finding connections between hidden and output neurons only, which is simply a linear least squares problem whose solution can be directly generated by the generalized inverse of the hidden layer output matrix (Huang, et al., 2012). Because of such special design of the network, ELM training becomes very fast. Numerous empirical studies and recently some analytical studies as well have shown that ELM has better generalization performance than other machine learning algorithms including SVMs and is more efficient and effective for both classification and regression tasks (Huang, et al., 2012).

3. EXPERIMENTS

To validate the proposed scheme, in this section we conduct experimental study by using simulated data.

3.1. Data

In this paper, the high-fidelity hardware-in-the-loop (HWIL) threat simulator is used for generating multiple data sets, i.e., normal, attack and evaluation data sets, for developing our gas turbine AD algorithms.

Normal operational data are simulated to cover a wide range of turbine operation conditions. Specifically, ambient temperature, pressure and humidity are varied to capture environmental variations. Fuel composition, compressor flow variation and turbine efficiency are varied to capture the expected operating conditions for the gas turbine. A design of experiments was performed to blend these factors into a reduced set of simulation runs. It was determined that steady state operating points were required as well as dynamic load conditions. Load level and load rate of change were variations that factored into the DOE runs. The DOE runs contained three different types of variations Plackett-Burman

(PB) full factorial; PB 11 factorial; & Pseudorandom binary signal (PRBS). Table 1 summarizes the DOE runs under normal operations.

For attack data, we simulate 11 different attacks, each with 3 attack levels. Table 2 shows the summary of the Plackett-Burman DOE runs.

Note that those simulation runs have different lengths in time. For DOE runs, the lengths vary from 200 seconds to 300 seconds; for PRBS runs, the lengths vary from 2000 seconds to 25000 seconds.

Table 1. Summary of Simulation Runs for Normal Operations.

Descriptions	# of simulation runs
Full factorial DOE	
4 factors (ambient temp., ambient pressure, relative humidity, and compressor flow) at 3 levels; and load at 8 levels (from 16.8MW to 200MW)	648
Placket-Burman DOE	
11 factors at 2 levels; and load at 8 levels (from 16.8MW to 200 MW)	96
Pseudo-random binary signal (PRBS) runs	
Fixed ambient conditions, while varying load from 1MW to 200MW with ramp rate varying from 10 MW/min to 18MW/min	7
Total =	**751**

Table 2. Summary of Simulation Runs for Attacks.

Descriptions	# of simulation runs
Placket-Burman DOE	
11 factors Placket Burman DOE at 3 attack levels, 3 ambient temperatures, and 8 levels of load (from 16.8MW to 200 MW)	936
Total =	**936**

3.2. AD Modeling and Performance Evaluation

3.2.1. AD model details

A typical gas turbine of power plants has a large number of physical measurements available. For AD modeling in this paper, we down-select a total of 24 measurements, out of which 15 are sensor measurements, 5 are actuator measurements, and 4 are control measurements. Out of the 24 selected physical measurements, we take 20 measurements for univariate feature calculation and other 4 for multivariate feature calculation.

For each of the 20 measurements, we calculate 5 statistical features, that is, median, variance, kurtosis, range, current value, of the signals within the sliding window. For the 4 measurements, we form 2 pairs and for each pair, we calculate mean difference between the two measurements.

We identified 3 physics models. Model 1 has 7 inputs and 4 outputs, model 2 has 3 inputs and 1 output, and model 3 has 2 inputs and 1 output. The three models are all data-driven models (ELMs) and are trained with normal data only. We calculate 5 residual statistics for each of the 6 model outputs, which gives us 30 physics-based features.

So overall, we have 132 (100+2+30) features calculated for each sample or each sliding window.

For ELM model design, we set the number of hidden neurons to be the default value of 1000, as suggested in Huang et al. (2012). The activation function for the hidden neurons is the sigmoid function, $G(w,b,x) = 1/(1+\exp(-(W^T x+b))$. The model parameter, C, is empirically determined via cross-validation by trying 20 different values, i.e., $C = [2^{-9}, 2^{-8}, \dots 2^{10}]$.

3.2.2. Performance evaluation method and metrics

To assess the performance of the proposed attack detection scheme, we use the Receiver Operating Characteristic (ROC) curves and the area-under-curve (AUC) as the classification performance measures. We employ 10-fold cross-validation for model training and validation. To obtain more robust comparison we run the 10-fold cross-validation 10 times, each time with different randomly splitting of 10 folds of the data. All experiments conducted in this paper are performed in Matlab® environment.

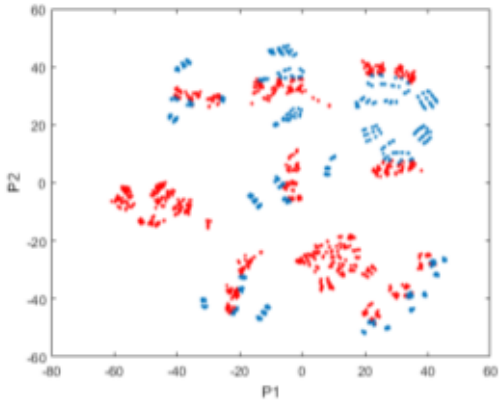

Figure 2. t-SNE projection of the extracted features.

3.3. Results

To help understand how well the 130 calculated features are in distinguishing normal and attack cases, we project the

features from original 130 dimensions to 2 dimensions using the t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008), as shown in Figure 2. We clearly see a good separation between normal and attack cases, granted the decision boundary is highly nonlinear, which gives us a confidence that we should have a good detection performance.

Figure 3 shows ROC curves for 10 different 10-fold cross-validation runs. To quantify the ROCs, we calculate the true positive rate (sensitivity) for the false positive rate set to be 1.0% and show them in a confusion matrix as shown in Table 3 below. Note that the numbers in Table 3 are the averages of the results of the 10 runs. As one can see from the table, our proposed attack detection scheme can achieve a true positive rate of 99.04% at the false positive rate of 1%, which is an excellent detection performance.

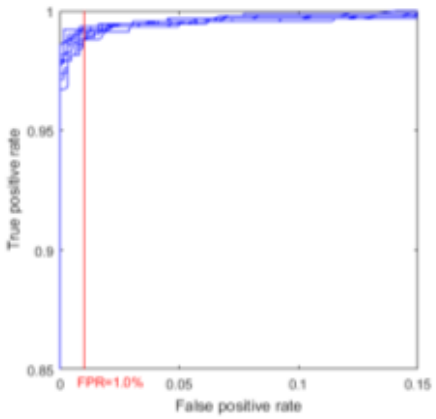

Figure 3. ROC curves.

Table 3. The confusion matrix.

		Predicted	
		Normal	Attack
True	Normal	**99.00**	1.00
	Attack	0.96	**99.04**

4. CONCLUSIONS

Securing CPS against cyberattacks is of great importance. Traditional intrusion detection systems (IDS) designed for IT systems are not effective in detecting CPS attacks. Detection methods that leverage physical properties of physical systems have proved to be more effective in CPS attack detection.

In this paper we propose a strategy that maximally leverage physical properties into our attack detection. The core of our proposed strategy is to accurately capture the physical behavior difference of the physical system between normal operation and adversary activities (attacks). To better capture the physical behavior using the physical measurements (sensors, actuators, and controllers), we intelligently map the raw physical measurements to a salient feature set that signifies the difference between normal operation and attack. Using the salient feature set as the input, our attack detection algorithm, an ELM based binary classification model, achieves excellent detection performance based on simulated gas turbine data.

ACKNOWLEDGEMENT

This research work is partially supported by contract number DEOE0000833 awarded in 2016 by the United States Department of Energy (DOE)'s Cyber-security for Energy Delivery Systems (CEDS) R&D Program.

REFERENCES

Falliere, N., Murchu, L.O. and Chien, E. (2011). W32. stuxnet dossier. *White Paper, Symantec Corp., Security Response*, 2011.

Humayed, A., Lin, J. and Li, F. (2017). Cyber-Physical Systems Security – A Survey. IEEE Internet of Things Journal, 4/6, Dec. 2017, pp. 1802 – 1831.

Huang, G.B., Zhou, H.M., Ding, X.J., and Zhang R. (2012). Extreme learning machine for regression and multiclass classification, *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, Vol. 42, No. 2, April 2012, pp. 513 – 529.

Huang, G.B., Zhu, Q.Y., and Siew, C.K. (2006). Extreme learning machine: Theory and applications, *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, Dec. 2006.

Khaitan, S.K. and McCalley, J.D. (2015). Design Techniques and Applications of Cyber Physical Systems: A Survey. *IEEE Systems Journal*, Vol.9, Issue 2, June 2015.

Kosut, O., Jia, L., Thomas, R.J. and Tong, L. (2011). Malicious data attacks on the smart grid. *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.

Mitchell R. and Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Computing Surveys (CSUR)*, 46(4):55, 2014.

Ozay, M., Esnaola, I. and Vural, F.T.Y. (2016). Machine Learning Methods for Attack Detection in the Smart Grid. *IEEE Transactions on Neural Networks and Learning Systems*, 27/8, Aug. 2016.

Ozay, M., Esnaola, I., Vural, F.T.Y, Kulkarni, S.R. and Poor, H.V. (2013). Sparse attack construction and state estimation in the smart grid: Centralized and distributed models. *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1306–1318, Jul. 2013.

Urbina, J. Giraldo, J., Cardenas, A.A., Valente, J., Faisal, M., Tippenhauer, N.O., Ruths, J. and Sandberg, H. (2016). Survey and new directions for physics-based attack detection in control systems. *NIST GCR 16-010, National Institute of Standards and Technology*, Gaithersburg MD, 20899, November 2016.

US Department of Homeland Security, Cyber Physical Systems Security. https://www.dhs.gov/science-and-technology/csd-cpssec.

van der Maaten, L.J.P. and G.E. Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579-2605, 2008.

Wallace, N., Ponomarev, S. and Atkison, T. A dimensional transformation scheme for power grid cyber event detection. *in Proc. Cyber Inf. Security Res. Conf.*, 2014, pp. 1–12.

Wang, Y., Amin, M.M., Fu, J. and Moussa, H.B. (2017). A novel data analytical approach for false data injection cyber-physical attack mitigation in smart grids. *IEEE Access*, Vol. 5, 2017.

Yan, W. (2015). Feature Engineering for PHM Applications. Tutorial at the Annual Conference of the Prognostics and Health Management Society 2015, Coronado, California 92118 USA.