Cluster-based Deep One-Class Classification Model for Anomaly Detection

Younghwan Kim, Huy Kang Kim

School of Cybersecurity, Korea University, Korea
jamesck@korea.ac.kr, cenda@korea.ac.kr

Abstract

As cyber-attacks on Cyber-Physical System (CPS) become more diverse and sophisticated, it is important to quickly detect malicious behaviors occurring in CPS. Since CPS can collect sensor data in near real time throughout the process, there have been many attempts to detect anomaly behavior through normal behavior learning from the perspective of data-driven security. However, since the CPS datasets are big data and most of the data are normal data, it has always been a great challenge to analyze the data and implement the anomaly detection model. In this paper, we propose and evaluate the Clustered Deep One-Class Classification (CD-OCC) model that combines the clustering algorithm and deep learning (DL) model using only a normal dataset for anomaly detection. We use auto-encoder to reduce the dimensions of the dataset and the K-means clustering algorithm to classify the normal data into the optimal cluster size. The DL model trains to predict clusters of normal data, and we can obtain logit values as outputs. The derived logit values are datasets that can better represent normal data in terms of knowledge distillation and are used as inputs to the OCC model. As a result of the experiment, the F1 score of the proposed model shows 0.93 and 0.83 in the SWaT and HAI dataset, respectively, and shows a significant performance improvement over other recent detectors such as Com-AE and SVM-RBF.

Keywords: Anomaly detection, Knowledge distillation, Clustering, Deep learning

1 Introduction

Cyber-Physical System (CPS) is not only a smart factory system, but also a system integrating virtual space and real-world physical systems that control public infrastructure such as cars, airplanes, and railroads, which are closely related to our lives. Since the anomalies that occur in CPS can cause physical damage beyond the cyber domain, the detection of abnormal behavior is an important issue today. Actually, various cybersecurity issues have occurred in CPS, such as Stuxnet computer virus [1], SQL slammer worm attacks on the Nuclear Plant [2], a power blackout in Brazil [3], and Maroochy water breach [4]. However, there are several main issues concerned with abnormal behavior detection of CPS. Since the CPS datasets are big data and most of the data are normal, it is challenging to implement a good performance model that can detect small-scale anomalies. Also, since the data collected in CPS have high dimensionality and nonstationary characteristics [5], it is impossible to implement a good performance model without considering these factors. Many anomaly detection solutions have been developed for CPS. Representative methods of detecting anomaly behavior include rule-based and model-based. The rule-based method is a method of detecting anomalous patterns with predefined rules, and the model-based method is a method of detecting anomaly behavior with mathematical algorithms. For example, Wong *et al.* [6] proposed a rule-based anomaly detection model that characterizes each anomaly pattern with carefully evaluated rules using Fisher's exact and randomization tests. Also, Klerx *et al.* [7] identifies the types of anomalies occurring in individual event systems for anomaly detection and provides a model-based algorithm called Probabilistic Deterministic Time Transition Automatic Device (PDTTA). In this paper, we propose a novel solution that combines the clustering algorithm and deep learning (DL) model using only the normal data for anomaly detection. The autoencoder algorithm is used to perform dimensionality reduction to mitigate the high-dimensionality and non-stationary characteristics of the CPS dataset [8]. To reduce the dimension while keeping the original dataset's characteristics as much as possible, the optimized latent vector's dimension is determined by calculating the loss value according to the dimension size. We cluster the CPS dataset with the optimal cluster size using the K-means algorithm for the normal dataset and implement a DL model that predicts the cluster. The One-Class Classification model receives the logit value from the DL model and evaluates the anomaly detection performance of the test set. The contributions of our paper are summarized as follows.

*Corresponding Author: Huy Kang Kim; E-mail: cenda@korea.ac.kr
DOI: 10.53106/160792642021072204017

- To propose a novel solution that combines the clustering algorithm and DL models for anomaly detection;
- To implement an effective model that can distinguish between anomaly and normal using datasets labeled only as normal;
- To perform experimental analysis to evaluate the performance of the proposed model using public datasets.

The rest of the paper is organized as follows. Section 2 presents the related work on anomaly detection and knowledge distillation. Section 3 presents the overall process of evaluating the proposed models for anomaly detection. Section 4 presents the experimental results with various models and discusses the results. Section 5 discusses our findings and limitations, then finally, Section 6 concludes this paper.

2 Related Work

2.1 Anomaly Detection

An anomaly can be defined as "A data that does not follow the distribution of the rest of the data, as if it were generated by a different mechanism" [9]. Thus, anomaly detection means finding patterns in data that indicate unexpected behavior. However, models for anomaly detection are difficult to design because it is difficult to define normal areas including all possible normal behaviors, and often the data contain noise that tends to resemble real anomalies [10]. Furthermore, since CPS datasets are big data and most of the data are normal, the design of detection models is more difficult. Nevertheless, a lot of work has been performed in designing the anomaly detection model for CPS.

Chen *et al.* [11] proposed an approach for learning invariants of CPS in machine learning models, such as SVM, to obtain an enhanced anomaly detection model. In this work, a preliminary investigation of the approach has been performed. Pasqualetti *et al.* [12] analyzed CPS-based monitoring limits for anomaly detection and proposed mathematical frameworks for CPS, attacks and monitors. Jones *et al.* [13] proposed an SVM-like algorithm that finds the signal time logic (STL) formula of the behavior domain from the dataset. However, this approach has a disadvantage that it is difficult to describe non-linear and high-dimensional datasets. Zhong *et al.* [14] performed anomaly detection using iForest, which is more scalable to high-dimensional data based on actual gas turbine data. Muralidhar *et al.* [15] developed a two-way model that can generate short-term and long-term forecasts of the system operating state using the seq2seq algorithm, reconstructed the missing data sequence from the log and evaluated the reconstruction performance. LSTM is one of the preferred models of anomaly detection using a CPS time-series dataset. Because RNN is unable to model long-term dependencies due to the vanishing gradients issues, LSTM is preferred over RNN for analysis of time-series datasets. Malhotra *et al.* [16] proposed a stacked LSTM model that would be able to learn higher-level temporal patterns without prior knowledge of the patterns. But the recall of the proposed model was very low, between 10% and 20%. Bontemps *et al.* [17] also applied a stacked LSTM model, which proposed a method of detecting collective anomalies that predict errors from a certain number of time steps instead of detecting anomaly from each time step. Ergen *et al.* [18] proposed a model to find the hyperplane that can separate anomalous data from normal data through the OCSVM model after extracting a fixed-length feature from LSTM. Guo *et al.* [19] proposed the Gaussian Mixture VAE (Variational Autoencoder) model. This is a generative model, which is effective for learning data in an unsupervised method. They trained the dependencies between time-series into Gated Recurrent Unit (GRU) cells to fit multidimensional time-series data and modeled them as Gaussian Mixture priors in the latent space. Wang *et al.* [20] proposed a composite auto encoder (Com-AE) model that learns a normal pattern. Common auto-encoder models are used to predict or reconstruct data separately, whereas the Com-AE model performs prediction and reconstruction on input simultaneously. Exponentially weighted moving average method (EWMA) was used to calculate the smoothing error for the normal data set and then used it as a threshold for detecting anomalies. Inoue *et al.* [21] proposed the SVM-RBF model for anomaly detection. In this research, the work in [20-21] is used to compare with our proposed method.

2.2 Knowledge Distillation

Knowledge distillation is a model that trains the Student Network from the Teacher Network. This is called Knowledge distillation because the knowledge of the Teacher Network, the more massive Neural Network, was distilled into the Student Network, the smaller Neural Network. Strategies such as knowledge distillation were first started in [22]. Bucilua˘ *et al.* [22] presented a method for compressing a large and complex ensemble model into a smaller and faster model. Hinton *et al.* [23] applied knowledge distillation by increasing the parameter called softmax temperature until the teacher network model suitably creates soft sets of targets and delivers them to the student network model.

3 Evaluation Framework of Proposed Model

Clustered Deep One-Class Classification (CD-OCC) solution has five main steps: (1) data collection, (2) data scaling and dimensionality reduction, (3)

clustering, (4) DL model and training, and (5) performance evaluation, which are described as follows. The overall process we proposed is shown in Figure 1.

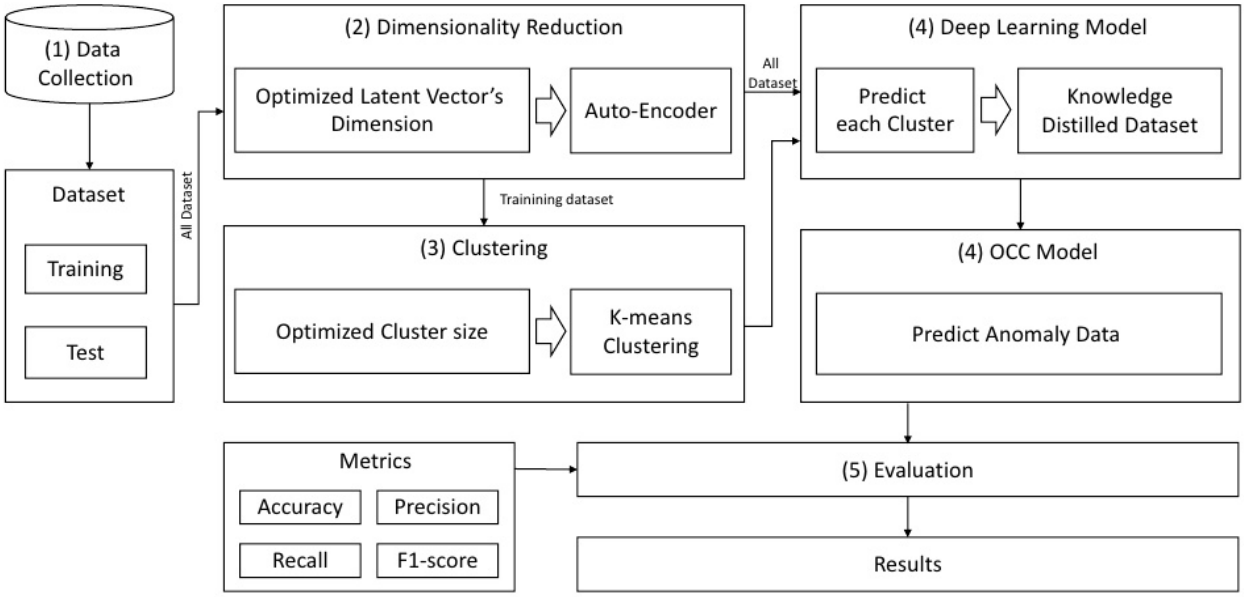


Figure 1. Overall process of the proposed model

Step 1. Data collection: Raw data collection is not easy because data cannot be obtained from the actual control system operating environment. Most datasets used for anomaly detection are generated through testbeds. In this paper, the data collection step is simplified using a publicly available dataset named HIL based Augmented ICS testbed (HAI) dataset [24] and Secure Water Treatment (SWaT) dataset [25]. HAI dataset is a dataset collected through a testbed built from 2017 using industrial controls, sensors, and actuators from GE, Emerson, and Siemens. The HAI testbed consists of three control systems (boiler, turbine, and water treatment). The control loops of the three control systems form a thermal power plant in a hardware in the loop (HIL) simulator. The HIL simulator controls each control system's field devices based on the signal values of these control systems. The dataset was developed for the study of detecting anomalies at CPS, such as railways, water treatments, and power plants, and contains normal and anomalous data corresponding to 34 attack scenarios. SWaT dataset is a dataset collected through a testbed created by SUTD University in 2015. The testbed collects data through a 6-step filtration process that can generate 5 gallons/minute of filtration water per minute. Each stage of SWaT communicates with PLCs via connected sensors and actuators based on the Ethernet ring topology. The PLCs interact with each other via a separate network based on the Ethernet star topology. Each PLC reads the sensor's latest data and calculates the appropriate signal to send to the actuator. The dataset was developed for the study of anomaly detection in water treatment plants and contained normal and anomalous data for a total of 36 attack scenarios. The details of each dataset are described in Table 1.

Table 1. The details of the SWaT and HAI dataset

Dataset	Normal data	Anomalous data	Total
SwaT	1,387,098	54,621	1,441,719
HAI	812,585	96,415	909,000

Step 2. Data scaling and dimensionality reduction: Since raw data has data-specific characteristics and distribution for each feature, performance will deteriorate if you use the data as it is. We use the MinMax scaler to scale the features of the data to the same extent. The use of high dimensional data can cause the performance problem because of the curse of dimensionality. To solve this problem, we apply the auto-encoder to reduce the dimensionality while preserving the unique properties of the data. To determine the latent vector's optimal size in the auto-encoder, it calculates the mean square error (MSE) loss value according to the latent vector's size. The latent vector size with minimal MSE loss is determined as the optimal latent vector.

Step 3. Clustering: We define a dataset consisting of normal data only as a train set, and cluster the train set based on the following assumptions.

- We can subdivide the normal data based on features, and the subdivided normal groups are relatively stable and representative;
- The probability distribution of classification into subdivided normal groups will differ between normal and anomalous data.

Note that we use a train set consisting of only

normal data for clustering because we want to obtain a probability distribution that characterizes the normal data. We use the K-means cluster algorithm to cluster train set only. Since the K-means clustering algorithm has the advantage of less computing without requiring prior information about the data in the analysis target, it is suitable for clustering large-capacity CPS datasets. However, when using the K-means clustering, we have to determine the initial cluster size. In our study, we use DBI (Davis-Bouldin Index) to find the optimal cluster size. DBI is useful for finding the optimal cluster size among various cluster sizes because it has the advantage of fast, easy, and providing consistent values. Since DBI has a high value for the cluster size that has a close distance between clusters and a long distance between all elements in the cluster, we define the cluster size with the lowest DBI value as the optimal cluster size.

Step 4. Training with DL and OCC model: From the perspective of knowledge distillation, the DL model is a teacher network that delivers the learning results for the dataset to the student networks, OCC models. When the amount of data is sufficient, the DL model generally outperforms the traditional machine learning models. Therefore, we tested the datasets with DL models (e.g., DNN, CNN and RNN). As a result of experimenting with the DNN, CNN, and RNN model in [26], we select the DNN model since the DNN model showed the highest performance. Detailed parameter values of the DL model are written in Section 4. DL model derives the logit values by predicting the clusters. Logit is the input to softmax in deep learning. The final layer in deep learning has logit values which are raw values for prediction by softmax. When obtaining the probability value for the multi classification problem, the softmax value is mainly used. We use logit because softmax has a much higher probability for a high score and a much lower probability for a low score. In our experiment, the logit presents the probability distribution value that predicts normal data clusters.

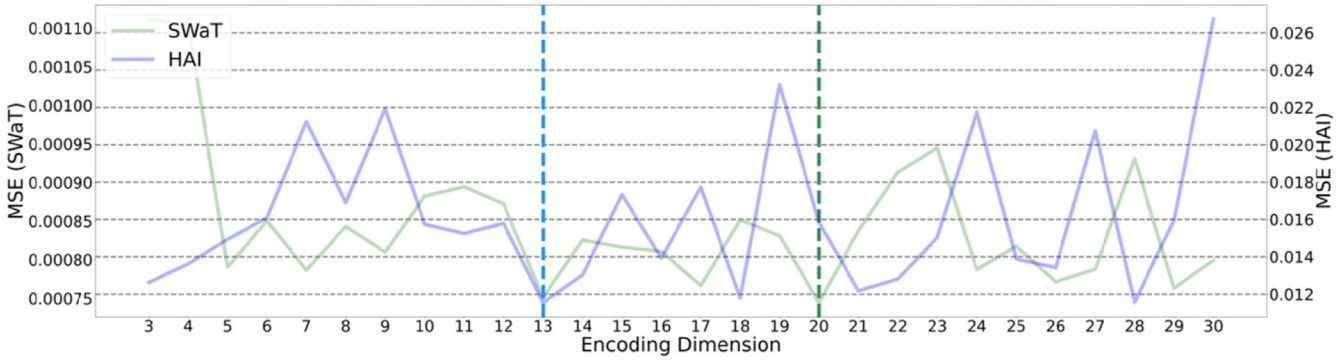


Figure 2. MSE for encoding dimension

Figure 3. DBI score of each dataset

Step 5. Performance evaluation: In the proposed CD-OCC model, DL models derive a knowledge distillation dataset and deliver it to OCC models to detect the anomalies. We evaluate the CD-OCC model using the knowledge distillation dataset and show that they outperform other recent detectors, such as Com-AE [20], SVM-RBF [21]. In the dataset, each record is labeled as "anomaly" or "normal", and the detection result is evaluated through the records. As the evaluation metric, accuracy, recall, precision, and f1-score are used.

4 Experimental Analysis

4.1 Optimized Latent Vector for Auto-encoder

We adjust the features of raw dataset through a min-max scaler and reduce dimensionality through auto-

encoder. The high dimensionality and non-stationary characteristics of the target dataset increase the complexity of detecting anomalies. Therefore, we mitigate high dimensionality and non-stationary characteristics by reducing the dimension using auto-encoder. In the experiment, two layers are used in the encoder part, and the number of neurons in each is 64 and 32. The decoder part is assigned the number of neurons of 32 and 64 symmetrical to the encoder part, respectively. The latent vector used as the reduced dimension of the dataset is located between the encoder and the decoder and is set to a value between 0 and 30 to find the optimized size. Adam was selected as the optimizer, and training epochs were set to 100. And to reduce the training time, the earlystop is used. we derive an optimized dimension value by calculating the MSE according to the dimension of the latent vector, as shown in Figure 2. Experimental results show that the optimal latent vector dimensions for SWaT and HAI are 20 and 13, respectively.

4.2 Optimized Cluster Size

We perform k-means clustering by constructing a train set with only normal data from a reduced-dimensional dataset. To find the optimal cluster size, we compute the DBI score for various cluster sizes. The cluster size is equal to the input data's dimension for the next step, deep learning. Since a cluster size that is too small limits the normal data's characterization, we set the cluster size from 5 to 30 to determine the optimal cluster size. The DBI score by cluster size is shown in Figure 3. As a result of the experiment, the cluster size representing the smallest DBI score is 24 for SWaT and 7 for HAI.

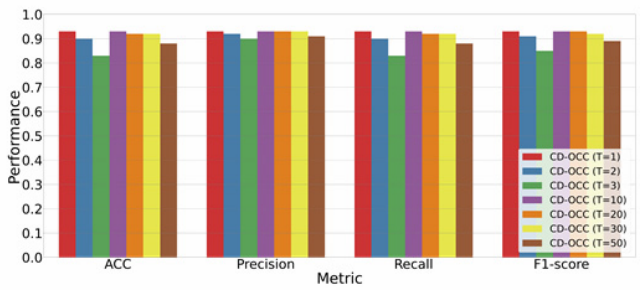
4.3 Derive Knowledge Distillation Dataset

The DL model trains to predict which cluster each record in the train set belongs to. As a result of experimenting with the DNN, CNN, and RNN model in [26], the DNN model showed the highest performance, so we select the DNN model. Our DL model consists of three dense layers. We use ReLU for the activation function in the first two layers. The number of neurons is 64 and 32, respectively. The activation function in the last layer is softmax, and the number of neurons is the optimized cluster size of each dataset. We set an optimizer as Adam, and the training epoch number as 100. Like the auto- encoder, earlystop is used to reduce training time. We perform 10-fold cross validation to prevent overfitting in the learning process, and derive logit values as the output of the DL model. The logit derived from the DL model's training result is a knowledge distillation dataset that more clearly represents the distribution of normal data. It is transferred to the OCC model in the next step. In terms of knowledge distillation, the concept of temperature is used to prevent the probability value from being biased to a specific class. We obtain logit values according to temperature and compare the anomaly detection performance.

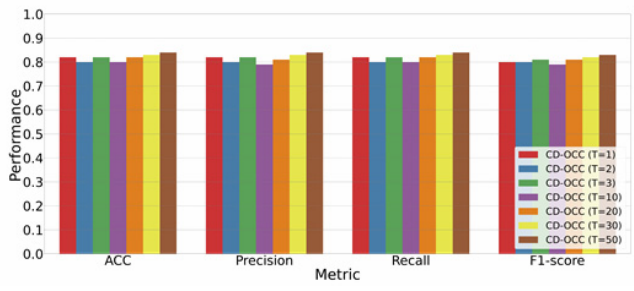
4.4 Evaluate Performance for Anomaly Detection

We use the OCC model for anomaly detection. According to Kim et al. [26], the iForest model has the best performance among various OCC models. Thus, we select iForest for our CD-OCC model. To prove the CD-OCC model's superiority, We compare the performance with the latest anomaly detectors such as com-AE and SVM-RBF. We also compare the anomaly detection performance when the dataset derived from the DL model is not only logit but also softmax. Figure 4 shows the anomaly detection performance when using logit and softmax in the CD-OCC model. Figure 4(a) and Figure 4(b) show the result of detecting anomaly using Logit in SWaT and HAI dataset, and compare the performance of models according to temperature values. In SWaT and HAI dataset, the performances change according to temperature, but the fluctuations are not large, and overall detection performances are high. It shows the highest detection performance when the temperature is 1 and 50 in SWaT and HAI, respectively. Figure 4(c) and Figure 4(d) show anomaly detection performance for softmax values in SWaT and HAI dataset.

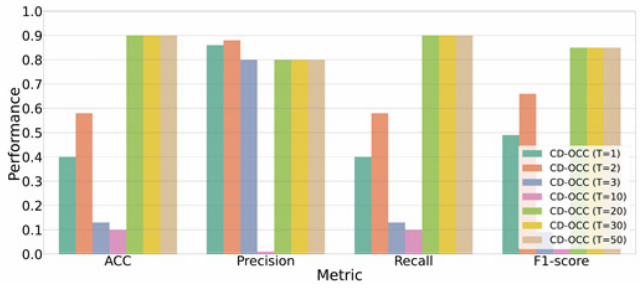
It shows the highest detection performance when the temperature is 20 and 1 in SWaT and HAI, respectively. Unlike logit, softmax's performances are highly dependent on temperature. Figure 4(e) and Figure 4(f) show the anomaly detection performance of logit and softmax in SWaT and HAI dataset as a box plot. Overall, the anomaly detection performances using the logit are higher than the softmax, and the fluctuation according to temperature is much less. This is because when the DL model predicts the cluster, softmax makes it have a much higher probability for high scores and a much lower probability for low scores. Figure 5 compares the performance of the CD-OCC model with other latest models, such as Com-AE and SVM-RBF. Figure 5(a) and Figure 5(b) show the experimental results for SWaT and HAI, respectively. The F1-score of our CD-OCC model is 0.93 and 0.83 in SWaT and HAI, respectively, showing higher detection performance than other models. Table 2 shows detailed numerical values indicating the experimental results.



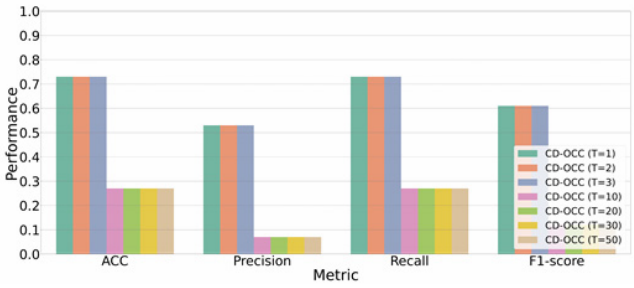
(a) SWaT (logit)



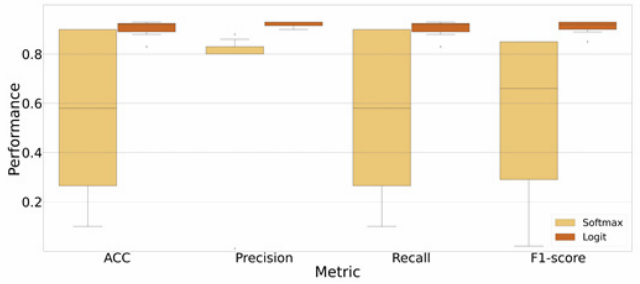
(b) HAI (logit)



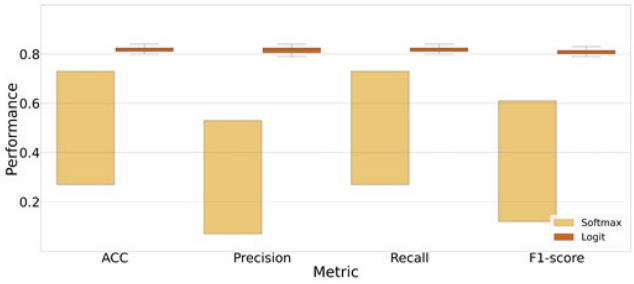
(c) SWaT (softmax)



(d) HAI (softmax)

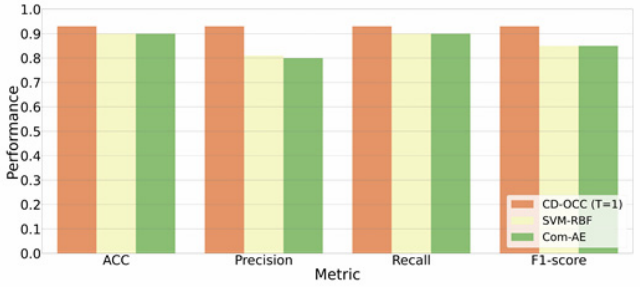


(e) SWaT (logit vs. softmas)

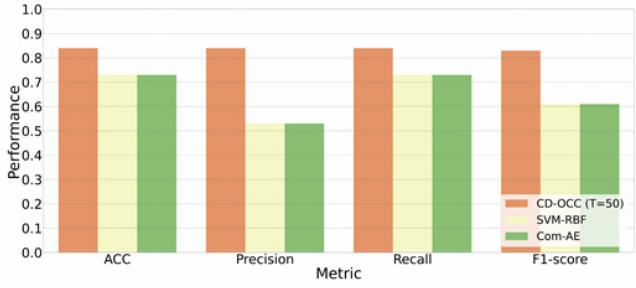


(f) HAI (logit vs. softmas)

Figure 4. Performance comparison by logit, softmax and temperature



(a) SWaT



(b) HAI

Figure 5. Performance comparison with other latest models

5 Discussion

5.1 Dimensionality Reduction

As the amount of information collected in the CPS increased, the size and dimension of the data became larger. Xu *et al.* [27] pointed out that such a high-dimensional characteristic of data makes the objects of data almost equal by making the distance between objects of data very close. To address this problem, dimensionality reduction techniques such as PCA or auto-encoder have been used in several studies [5, 28-31]. In particular, auto-encoder is a more effective nonlinear technology than PCA when reducing the dimensionality of high-dimensional data. Sakurada *et al.* [30] found that auto-encoder can detect subtle anomalies that linear PCA fails and increase its accuracy by denoising. Therefore, we reduced dimensionality through auto-encoding and then performed training.

Table 2. CD-OCC vs. Other models

Models	SWaT				HAI			
	ACC	Precision	Recall	F1-score	ACC	Precision	Recall	F1-score
CD-OCC (logit)								
Temp = 1	**0.93**	**0.93**	**0.93**	**0.93**	0.82	0.82	0.82	0.80
Temp = 2	0.90	0.92	0.90	0.91	0.80	0.80	0.80	0.80
Temp = 3	0.83	0.90	0.83	0.85	0.82	0.82	0.82	0.81
Temp = 10	0.93	0.93	0.93	0.93	0.80	0.79	0.80	0.79
Temp = 20	0.92	0.93	0.92	0.93	0.82	0.81	0.82	0.81
Temp = 30	0.92	0.93	0.92	0.92	0.83	0.83	0.83	0.82
Temp = 50	0.88	0.91	0.88	0.89	**0.84**	**0.84**	**0.84**	**0.83**
CD-OCC (softmax)								
Temp = 1	0.40	0.86	0.40	0.49	0.73	0.53	0.73	0.61
Temp = 2	0.58	0.88	0.58	0.66	0.73	0.53	0.73	0.61
Temp = 3	0.13	0.80	0.13	0.09	0.73	0.53	0.73	0.61
Temp = 10	0.10	0.01	0.10	0.02	0.27	0.07	0.27	0.12
Temp = 20	0.90	0.80	0.90	0.85	0.27	0.07	0.27	0.12
Temp = 30	0.90	0.80	0.90	0.85	0.27	0.07	0.27	0.12
Temp = 50	0.90	0.80	0.90	0.85	0.27	0.07	0.27	0.12
Other Models								
Com-AE	0.90	0.80	0.90	0.85	0.73	0.53	0.73	0.61
SVM-RBF	0.90	0.81	0.90	0.85	0.73	0.53	0.73	0.61

5.2 Dataset Reliability Issues

Due to the nature of the OCC model, which trains with only normal data, it is necessary to use purely normal data in the training phase. When anomalies that appear to be normal input the training phase, excellent performance cannot be achieved. Besides, due to the regular tests such as the overall test, although it is not anomalies, different normal data than usual may be included in the training phase. In this regard, data labeled normal can be grouped in detail according to characteristics through clustering. If the data labeled with normal contains anomalies or a rare category of normal data, it can be classified in a sophisticated clustering process. In this way, we try to solve the reliability issue of normal data through a K-means clustering algorithm, while detecting anomalies outside the normal clusters.

Since the K-means clustering algorithm has the advantage of not requiring prior information about the data to be analyzed and has a small amount of computation, it is suitable for clustering large CPS datasets. In the future, we will find out the most suitable clustering algorithm for the CD-OCC model by comparing and analyzing the performance with various clustering algorithms such as OPTICS and DBSCAN in addition to the K-means clustering algorithm.

5.3 Applying New Attack Scenarios in CPS

In recent years, research on the types of attacks on CPS has been actively conducted. Amin *et al.* [32] reported denial-of-service attacks on network control systems, Liu *et al.* [33] showed the possibility of a false data injection attack, and Teixeira *et al.* [34] analyzed the effect of replay attack in CPS. Apart from this, research into advanced and intelligently evolving CPS attack techniques is ongoing. For the anomaly detection model to learn the attack patterns of the evolved CPS, datasets should be implemented, including attack scenarios based on the latest attack patterns. However, since it is difficult to obtain data from the actual control system operating environment, creating a high-quality dataset that goes beyond a simple attack level is an ongoing challenge.

6 Conclusion

A Cluster based Deep One-Class Classification (CD-OCC) model is proposed in this paper. This model is trained with only a normal dataset for use in a real CPS environment where anomalous data is sparse. Our model clusters the normal dataset and then trains to predict the cluster through the DL model. To ensure the practicality of the proposed model, we conducted comprehensive experiments. We use the softmax and logit values derived from the training process of the DL, and use the temperature values to mitigate the bias of the probability values. The probability values derived from the DL are used as inputs to the OCC model to evaluate anomaly detection performance. As a result of the experiment, the F1 score is 0.93 and 0.83 in the SWaT and HAI dataset, respectively, higher than other recent detectors such as Com-AE and SVM-RBF. In addition to the CD-OCC model we propose, there are many existing detection anomaly methods. It is meaningful to propose a novel approach to detecting anomalies by using the clustering algorithm and the DL model from the knowledge distillation perspective. Efforts to Minimize the normal detected as anomalies

or the anomalies detected as normal are important tasks in the CPS environment. In future work, we will expand research to implement high-performance models while minimizing the error rate. We will also apply our approach to other datasets and analyze the impact on performance according to different cluster algorithms.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)(No. 2018-0-00232, Cloud-based IoT Threat Autonomic Analysis and Response Technology).

References

[1] J. P. Farwell, R. Rohozinski, Stuxnet and the future of cyber war, *Survival*, Vol. 53, No. 1, pp. 23-40, February, 2011.

[2] S. Kuvshinkova, SQL slammer worm lessons learned for consideration by the electricity sector, *North American Electric Reliability Council*, Vol. 1, No. 2, p. 5, June, 2003.

[3] J. P. Conti, The day the samba stopped, *Engineering & Technology*, Vol. 5, No. 4, pp. 46-47, March, 2010.

[4] J. Slay, M. Miller, Lessons learned from the maroochy water breach, *International conference on critical infrastructure protection*, Hanover, NH, USA, 2007, pp. 73-82.

[5] J. Camacho, A. P. Villegas, P. G. Teodoro, G. M. Fernandez, PCA-based multivariate statistical network monitoring for anomaly detection, *Computers & Security*, Vol. 59, pp. 118-137, June, 2016.

[6] W. Wong, A. Moore, G. Cooper, M. Wagner, Rule-based anomaly pattern detection for detecting disease outbreaks, *18th National Conference on Artificial Intelligence*, Edmonton, Alberta, Canada, 2002, pp. 217-223.

[7] T. Klerx, M. Anderka, H. K. Buning, S. Priesterjahn, Model-based anomaly detection for discrete event systems, *the IEEE 26th International Conference on Tools with Artificial Intelligence*, Limassol, Cyprus, 2014, pp. 665-672.

[8] Y. Yao, J. Kang, J. Lee, A survey on deep learning-based anomaly detection models for time series data, *Journal of the Korea Information Science Society*, Vol. 46, No. 1, pp. 919-921, June, 2019.

[9] J. Han, M. Kamber, J. Pei, *Data mining: concepts and techniques*, Elsevier, 2011.

[10] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM computing surveys (CSUR)*, Vol. 41, No. 3, pp. 1-58, July, 2009.

[11] Y. Chen, C. M. Poskitt, J. Sun, Towards learning and verifying invariants of cyber-physical systems by code mutation, *International Symposium on Formal Methods*, Limassol, Cyprus, 2016, pp. 155-163.

[12] F. Pasqualetti, F. Dorfler, F. Bullo, Attack detection and identification in cyber-physical systems, *IEEE transactions on automatic control*, Vol. 58, No. 11, pp. 2715-2729, November, 2013.

[13] A. Jones, Z. Kong, C. Belta, Anomaly detection in cyber-physical systems: A formal methods approach, the *53rd IEEE Conference on Decision and Control*, Los Angeles, CA, USA, 2014, pp. 848-853.

[14] S. Zhong, S. Fu, L. Lin, X. Fu, Z. Cui, R. Wang, A novel unsupervised anomaly detection for gas turbine using isolation forest, *IEEE International Conference on Prognostics and Health Management (ICPHM)*, San Francisco, CA, USA, 2019, pp. 1-6.

[15] N. Muralidhar, S. Muthiah, K. Nakayama, R. Sharma, N. Ramakrishnan, Multivariate long-term state forecasting in cyber-physical systems: A sequence to sequence approach, *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 543-552.

[16] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, Long short term memory networks for anomaly detection in time series, *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium, 2015, pp. 89-94.

[17] L. Bontemps, V. L. Cao, J. Mcdermott, N. A. Le-Khac, Collective anomaly detection based on long short-term memory recurrent neural networks, *International Conference on Future Data and Security Engineering*, Can Tho City, Vietnam, 2016, pp. 141-152.

[18] T. Ergen, A. H. Mirza, S. S. Kozat, Unsupervised and semi-supervised Anomaly Detection with LSTM Neural Networks, *arXiv preprint arXiv:1710.09207*, October, 2017.

[19] Y. Guo, W. Liao, Q. Wang, L. Yu, T. Ji, P. Li, Multidimensional time series anomaly detection: A gru-based gaussian mixture variational autoencoder approach, *Asian Conference on Machine Learning*, Beijing, China, 2018, pp. 97-112.

[20] C. Wang, B. Wang, H. Liu, H. Qu, Anomaly detection for industrial control system based on autoencoder neural network, *Wireless Communications and Mobile Computing*, Vol. 2020, Article ID 8897926, August, 2020.

[21] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, J. Sun, Anomaly detection for a water treatment system using unsupervised machine learning, *IEEE international conference on data mining workshops (ICDMW)*, New Orleans, LA, USA, 2017, pp. 1058-1065.

[22] C. Bucilua, R. Caruana, A. Niculescu-Mizil, Model compression, *12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA, 2006, pp. 535-541.

[23] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531*, March, 2015.

[24] H. K. Shin, W. Lee, J. H. Yun, H. C. Kim, Implementation of programmable CPS testbed for anomaly detection, *12th USENIX Conference on Cyber Security Experimentation and Test*, Santa Clara, CA, USA, 2019, pp. 1-9.

[25] J. Goh, S. Adepu, K. N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, *11th International Conference on Critical Information*

Infrastructures Security, Paris, France, 2016, pp. 88-99.

[26] Y. Kim, H. K. Kim, Anomaly detection using clustered deep one-class classification, *15th Asia Joint Conference on Information Security (AsiaJCIS)*, Taipei, Taiwan, 2020, pp. 151-157.

[27] X. Xu, H. Liu, M. Yao, Recent progress of anomaly detection, *Complexity*, Vol. 2019, pp. 1-11, January, 2019.

[28] H. Ringberg, A. Soule, J. Rexford, C. Diot, Sensitivity of PCA for traffic anomaly detection, *ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, San Diego, California, USA, 2007, pp. 109-120.

[29] D. Brauckhoff, K. Salamatian, M. May, Applying PCA for traffic anomaly detection: Problems and solutions, *IEEE INFOCOM*, Rio de Janeiro, Brazil, 2009, pp. 2866-2870.

[30] M. Sakurada, T. Yairi, Anomaly detection using autoencoders with nonlinear dimensionality reduction, *2nd Workshop on Machine Learning for Sensory Data Analysis*, Gold Coast, QLD, Australia, 2014, pp. 4-11.

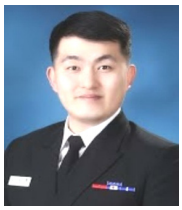
[31] Y. Wang, H. Yao, S. Zhao, Auto-encoder based dimensionality reduction, *Neurocomputing*, Vol. 184, pp. 232-242, April, 2016.

[32] S. Amin, A. A. Cardenas, S. S. Sastry, Safe and secure networked control systems under denial-of-service attacks, *International Workshop on Hybrid Systems: Computation and Control*, San Francisco, CA, USA, 2009, pp. 31-45.

[33] Y. Liu, P. Ning, M. K. Reiter, False data injection attacks against state estimation in electric power grids, *ACM Transactions on Information and System Security (TISSEC)*, Vol. 14, No. 1, pp. 1-33, May, 2011.

[34] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, S. S. Sastry, Cyber security analysis of state estimators in electric power systems, *49th IEEE conference on decision and control (CDC)*, Atlanta, GA, USA, 2010, pp. 5991-5998.

Biographies



Younghwan Kim received a B.S. degree from the Department of Mechanical Engineering, Naval Academy and an M.S. degree from the Department of Computer Science, Korea National Defense University. He is currently pursuing a Ph.D. degree with School of Cybersecurity, Korea University, under the supervision of H.K. Kim. His research interests include data mining, rumor detection, and machine learning.

Huy Kang Kim is a professor in School of Cybersecurity, Korea University. He received his Ph.D. in industrial and systems engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2009. He received an M.S. degree in industrial engineering from KAIST in 2000. He received a B.S. degree in industrial management from KAIST in 1998. He founded A3 Security Consulting, the first information security consulting company in South Korea in 1999. Before joining Korea University, he was a technical director (TD) and a head of information security department of NCSOFT (2004-2010), one of the most famous MMORPG companies in the world. His recent research is focused on solving many security problems in online games based on the user behavior analysis.