2023 IEEE International Conference on Big Data and Smart Computing (BigComp) | 978-1-6654-7578-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigComp57234.2023.00049

MENDEL: Time series anomaly detection using transfer learning for industrial control systems

Jeongyong Park, Bedeuro Kim, and Hyoungshick Kim
Department of Computer Science and Engineering
Sungkyunkwan University
Suwon, Republic of Korea
{wjddyd, kimbdr, hyoung}@skku.edu

***Abstract*—Machine learning is commonly used to detect anomalies in industrial control systems (ICS). In general, building an anomaly detection model requires massive training data and computational resources. Therefore, an ideal solution is to use a pre-trained model instead of building each model completely from scratch. However, we cannot directly use a pre-trained model because each ICS dataset has its own unique features and characteristics. This paper proposes a practical transfer learning technique dubbed MENDEL (tiMe sEries aNomaly Detection using transfEr Learning) to efficiently build anomaly detection models, respectively, for different ICS domains. MENDEL first applies principal components analysis (PCA) to each model to obtain a fixed number of reduced features compatible with other models and then finds a reasonable mapping between different models' reduced features systemically for effective transfer learning. We evaluate the performance of MENDEL on two datasets (SWaT and WADI) with two models (InterFusion and USAD). Our evaluation results show that MENDEL can overall achieve high F1 scores even when a model is retrained with only a small proportion of the training dataset. For example, when we first train InterFusion with the SWaT train dataset and then retrain the trained model with only 10% of the entire WADI train dataset, the retrained InterFusion achieves an F1 score of 72%, which is better than an F1 score of 44% achieved by InterFusion with the entire SWAT training dataset.**

***Index Terms*—industrial control systems (ICS), anomaly detection, transfer learning, feature mapping**

I. Introduction

The term "Industrial Control Systems" (ICS) refers to a broad class of automation systems that provides control and monitoring functionality in manufacturing and industrial facilities. ICS is widely used in many facilities for supervisory control, data acquisition, and industrial automation. However, we must address security challenges to deploy and use ICS for critical infrastructures. Recently, several attack attempts (e.g., Stuxnet [1], Duqu [2], Flame [2], and Havex [3]) aimed to subvert ICS. Therefore, protecting ICS against cyber and physical security attacks is important.

A possible defense mechanism is to monitor sensors in an ICS and detect sudden changes in the monitored sensor values, which an attack attempt or system fault would cause. Many previous studies have proposed deep learning models [4]–[10] that could effectively detect abnormal behaviors of ICS. Overall, the effectiveness of deep learning models significantly decreases against new and unseen data. Therefore, we need to build a proprietary model for each ICS dataset. However, because training a deep learning model typically requires the collection of a massive dataset and using powerful computing resources, it would be expensive to train deep learning models completely from scratch for all different ICS datasets, respectively.

Many machine learning techniques have been attempting to reduce the training overhead caused by a large dataset. Transfer learning is a widely used technique to improve a model built from one domain by transferring to another model in a related domain. Transfer learning is commonly used in image and network domains, but there were only some cases (e.g., [11]) in the other domains using time-series data. This paper focuses on ICS data among time-series data because each ICS can produce a new and unique ICS dataset. Therefore, training a model for each ICS dataset is unavoidable to produce optimized results. To reduce the cost of training, we suggest using transfer learning techniques for ICS datasets to reduce the computational overhead for training.

To effectively perform transfer learning between two different ICS datasets, we need to transform the features in one dataset into the features in the other dataset [12]. Three feature transformation strategies (feature augmentation, feature reduction, and feature alignment) were presented in [12]. Among the three transform strategies, we choose the feature reduction method to find the

This work was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01343 and No.2022-0-00495).

978-1-6654-7578-5/23/$31.00 ©2023 IEEE
DOI 10.1109/BigComp57234.2023.00049

features for transfer learning because each ICS dataset has its own proprietary features. Next, we use a proper feature mapping method to map the features in one ICS dataset into the features in the other ICS dataset.

To show the feasibility of MENDEL, we evaluate the performance of MENDEL on two representative ICS datasets (SWaT [13] and WADI [14]) with two models (InterFusion [4] and USAD [5]). The experimental results show that MENDEL can achieve high F1 scores with only a small portion of the entire training dataset, reducing the training time dramatically. For the SWaT dataset, we first train a model with the WADI dataset and retrain the model with a portion of the SWaT training dataset; InterFusion achieves an F1 score of 87% with 1% of the SWAT training dataset; USAD achieves an F1 score of 86% with 5% of the SWAT training dataset. For the WADI dataset, we first train a model with the SWaT dataset and retrain the model with a portion of the WADI training dataset; InterFusion achieves an F1 score of 72% with 10% of the WADI training dataset; USAD achieves an F1 score of 48% with 5% of the WADI training dataset. From these experimental results, we confirm that MENDEL can reduce the cost of training to build an anomaly detection model with ICS datasets. Moreover, the retrained models through MENDEL produce F1 scores comparable with the F1 scores of 82% and 52%, respectively, achieved by the best models (InterFusion for SWaT and USAD for WADI) without using a pre-trained model.

Our key contributions are summarized as follows:

- We present a transfer learning method dubbed MENDEL to build a highly accurate anomaly detection model for ICS, significantly reducing the training cost of the model (see Section III);
- We demonstrate via experiments using publicly accessible two ICS datasets (SWaT [13] and WADI [14]) and two anomaly detection models (InterFusion [4] and USAD [5]), the efficiency and feasibility of MENDEL under various conditions (see Section IV);
- We release the source code of MENDEL (https://github.com/SKKU-SecLab/MENDEL) for the research community interested in developing anomaly detection models for ICS.

II. Background

A. Industrial control system (ICS)

Industrial control systems (ICS) refer to various control systems used to control and monitor the operating processes of industrial infrastructures, such as gas pipelines, water treatment, and transportation networks [15]. ICS has several components: (1) the field site responsible for the physical process in infrastructure, (2) the fieldbus network that transmits data input from physical sensors and actuators, and (3) the control center that can manipulate and monitor the state of the field site through collected data. Recently, Internet communication protocols such as Ethernet and TCP/IP have been used for the fieldbus to improve network efficiency and accuracy. However, those network interfaces can be the main targets of attackers who want to exploit ICS. In this paper, we focus on detecting anomalous operations that can occur in the physical processes of the field site due to cyber attacks or system faults.

B. Anomaly detection in multivariate time series

In ICS, the collected data from sensors have temporal (continuous) properties because sensors measure some physical status and report the measured values to controllers periodically. For example, in a water treatment system, a controller receives information about the current height of a water tank from a sensor and transmits the command to an actuator to perform an algorithm to keep the height of the water tank below a specific water height. During this process, we need to collect the data measured by the sensor over time. Therefore, the collected data have temporal (continuous) properties. We call such data with temporal features *time series data*. In practice, most time series data of ICS have high dimensionality called *multivariate time series* data in which logs are collected at every instant from interconnected field devices [6]. Many machine learning models have been proposed to build accurate anomaly detection models on multivariate time series data [4]–[6], [16], [17]. The anomaly detection approaches for multivariate time series data are divided into supervised learning [18] and unsupervised learning [7], [8], [17], [19], [20], depending on whether label data is required for model training. In recent studies, the importance of unsupervised learning is growing to overcome the disadvantage of identifying only the anomaly type seen in the training data of supervised learning and the limitation that it is difficult to obtain label data in the real world.

C. Feature transformation for transfer learning

Feature transformation refers to processes that transform a given source task model's features into a target task model's features for transfer learning [12]. Feature transformation can be divided into feature mapping, feature selection, and feature clustering. Among them, feature mapping aims to increase the performance of transfer learning by matching similar features and measuring the similarity between the features of the original data and the target data applied feature reduction.

III. Overview of MENDEL

We propose a transfer learning technique dubbed MENDEL to construct a model for anomaly detection

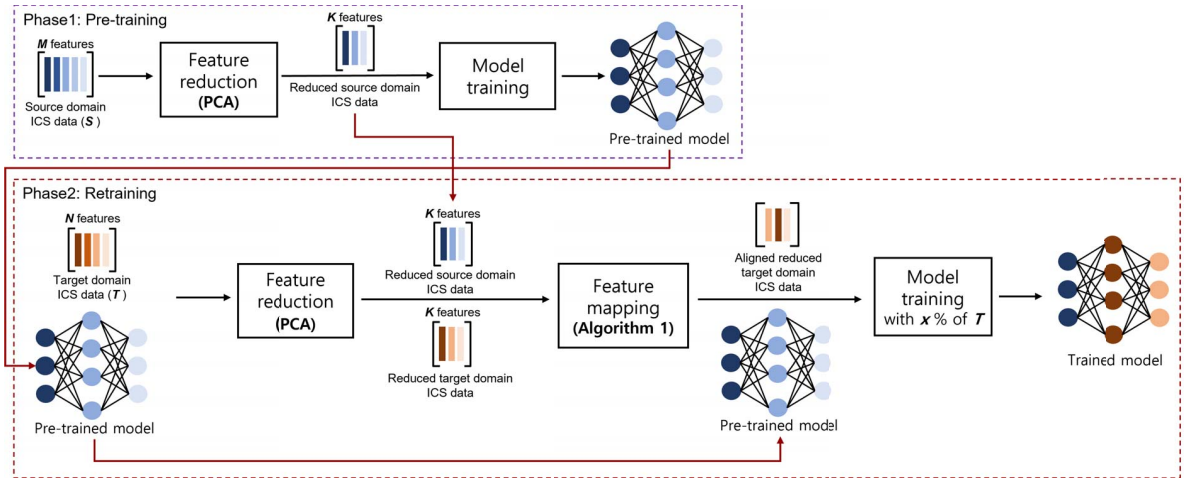

Fig. 1. Overview of MENDEL.

(see Figure 1). Our challenge is how to use transfer learning when a source domain ICS dataset and a target domain ICS dataset have different features. MENDEL first uses a feature reduction technique so that each ICS dataset should have the same number of features. In our implementation, we use principal component analysis (PCA) [21] to determine k non-correlated features for each ICS dataset. Next, MENDEL finds the mapping between the features in a source domain ICS dataset and the features in a target domain ICS dataset by comparing their statistical distributions, respectively. MENDEL constructs a pre-trained model for a source domain ICS dataset as follows:

1) MENDEL uses PCA to reduce the dimension of a source domain ICS dataset to k features.
2) MENDEL uses the reduced set of k features to train a model for the source domain ICS dataset.

For a target domain ICS dataset, MENDEL then retrains the pre-trained model with some samples of the target domain ICS dataset as follows:

1) MENDEL uses PCA to reduce the dimension of a target domain ICS dataset to k features.
2) MENDEL uses Algorithm 1 to map the k features of the source domain ICS dataset to the k features of the target domain ICS dataset.
3) MENDEL retrains the pre-trained model with $x\%$ of the reduced target domain ICS dataset.

To use a pre-trained model (built with a source domain ICS dataset), we need to map the k reduced features of the source domain ICS dataset to the k reduced features of the target domain ICS dataset. We propose a greedy algorithm to find a reasonable mapping between two ICS datasets (see Algorithm 1).

Our key idea is simple: each feature in the source domain ICS dataset should be mapped to the feature

Algorithm 1 Feature mapping algorithm.

```
S: a set of features in the source domain dataset.
T: a set of features in the target domain dataset.
Initialize a priority queue Q ← ∅.
for all t ∈ T do
    Compute t's feature importance score s_t.
    priority[t] ← s_t.
    Insert t into Q.
end for
Compute the JSD of all pairs between S and T.
while Q is not empty do
    t_max ← Q.dequeueMax().
    s_i = argmin_{s_p ∈ S} JSD(t_max, s_p).
    Output (t_max, s_i).
    Delete s_i in S.
end while
```

in the target domain ICS dataset with the most similar statistical distribution. Therefore, MENDEL computes the similarity scores of all pairs of source and target domain features using Jensen Shannon Distance (JSD) [22], a metric to assess the difference between two distributions, and chooses the most similar k pairs of source and target domain features in a greedy fashion. MENDEL first sorts the features in the target domain ICS dataset according to their feature importance scores in descending order. We use the extreme gradient boosting (XGBoost) algorithm [23] to compute feature importance scores. Before computing the feature importance scores, MENDEL applies the min-max normalization technique, which confines the scale of each feature from 0 to 1. Next, in each loop iteration, MENDEL picks the most important feature, finds its corresponding feature with the minimum JSD in the source domain ICS

dataset, and removes those features. This is repeated until there are no more features to be mapped. Algorithm 1 describes this procedure in detail.

IV. Experiments and analysis

In this section, we introduce our experimental setup and results in detail. First, we will explain the two datasets used in the experiments. Second, we will explain the evaluation metrics. Third, we will explain the hyperparameters used in each model and the number of features used in the experiments. Finally, the experimental results will be presented.

A. Datasets

We use two representative ICS datasets collected from a water treatment testbed for research: Secure Water Treatment (SWaT) [13] and Water Distribution (WADI) [14]. The SWaT dataset was collected for normal and cyber attack operating conditions for 11 days in an environment designed to resemble the existing water treatment plant's physical processes and control systems. The WADI is an extension of the SWaT, which has 75 more features than SWaT and was collected under normal and cyber attack operating conditions for 14 days. Table I summarizes the basic properties of the two datasets.

TABLE I
Description of the datasets.

Datasets	Train	Test	Features	Version
SWaT	495,000	449,919	51	A2
WADI	1,209,601	172,801	123	A1

B. Metrics

We use the F1 score, Precision, Recall, and Accuracy as anomaly detection performance metrics to evaluate the performance of each model for two datasets. Accuracy is an indicator of the model's prediction accuracy, but it does not represent the model performance for incorrect predictions such as FP and FN. In addition, since both WADI and SWAT are imbalanced (e.g., high normal class and low anomaly class), performance indicators are selected, considering that the performance of the evaluation indicators may be biased toward a specific class. The training and test times are used to measure the models' efficiency.

C. Experiment setup

We used two models (InterFusion [4] and USAD [5]) for experiments. InterFusion is a state-of-the-art model that learns both time and inter-metric features using Hierarchical Variational Autoencoder (HAVE) [24], [25]. InterFusion consists of three structures: offline training, online detection, and anomaly interpretation. USAD is an unsupervised-based anomaly detection method based on an autoencoder architecture whose learning is inspired by GAN. USAD provides a new architecture design using two-phase with adversarial training. Training in USAD is done with two autoencoders. The first autoencoder generates the reconstructed data for the original input dataset using a pair of encoders and decoders. The second autoencoder intentionally generates anomaly data by adapting adversarial training to the output of the decoder and then giving feedback to the encoder again. This two-phase training process allows USAD to detect anomalies similar to the original samples.

To use the optimized models, we used the same hyperparameters as in their original papers [4], [5]. Table II presents the hyperparameters for InterFusion and USAD. However, we note that InterFusion produced a significantly lower estimated F1 score (0.4392) for the WADI dataset than the results presented in the original paper [4].

TABLE II
Hyperparameters for InterFusion and USAD.

Hyperparameters	InterFusion	USAD
Window size	60	60
Epoch	15	70
Batch size	50	200

We first determine the optimal number of reduced features (i.e., k) for PCA in MENDEL and evaluate each model's performance with varying the proportion of the samples for retraining (i.e., $x\%$).

D. Performance without retraining

As baselines, we built InterFusion and USAD with the SWaT and WADI datasets, respectively, without retraining. Table III shows the evaluation results. For the SWaT dataset, InterFusion achieved an F1 score of 0.8261; USAD achieved an F1 score of 0.8226. For the WADI dataset, InterFusion achieved an F1 score of 0.4440; USAD achieved an F1 score of 0.5280.

TABLE III
Performance of two models without retraining.

Measure	SWaT		WADI	
	InterFusion	USAD	InterFusion	USAD
F1 score	0.8261	0.8226	0.4440	0.5280
Precision	0.8948	0.9297	0.8026	0.6671
Recall	0.7213	0.6951	0.3038	0.4112
Accuracy	0.9298	0.9284	0.9234	0.9206
Training time (s)	9250.3	2099.4	17288.2	3029.5
Test time (s)	13483.8	8.6	1430.5	9.2

E. Optimization of the number of reduced features

To determine the optimal number of reduced features for PCA, we first evaluated the performance of two

anomaly detection models, InterFusion and USAD, with varying the number of reduced features (i.e., k).

Table IV presents the performance of InterFusion built by MENDEL for the SWaT dataset with the number of reduced features. We first trained InterFusion with 100% of the WADI train dataset, then retrained InterFusion with 3% of the SWaT train dataset, and evaluated the retrained InterFusion with the SWaT test dataset. The retrained InterFusion produced the best F1 score (0.8772) when $k = 15$.

TABLE IV
PERFORMANCE OF INTERFUSION VIA MENDEL FOR SWAT WITH THE NUMBER OF REDUCED FEATURES.

Measure	InterFusion					
	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$	$k = 35$
F1 score	0.8400	**0.8772**	0.8586	0.8432	0.8559	0.8496
Precision	0.8941	0.9139	0.9098	0.9142	**0.9414**	0.8427
Recall	0.7920	0.8434	0.8129	0.7825	0.7846	**0.8566**
Accuracy	0.9633	**0.9713**	0.9675	0.9646	0.9679	0.9631
Training time (s)	**257.6**	257.7	277.6	278.3	262.7	271.9
Test time (s)	7541.7	**7198.8**	7670.3	7473.8	7221.5	7381.8

Table V presents the performance of USAD built by MENDEL for the SWaT dataset with the number of reduced features. We first trained USAD with 100% of the WADI train dataset, then retrained USAD with 3% of the SWaT train dataset, and evaluated the retrained USAD with the SWaT test dataset. The retrained USAD produced the best F1 score (0.8617) when $k = 25$.

TABLE V
PERFORMANCE OF USAD VIA MENDEL FOR SWAT WITH THE NUMBER OF REDUCED FEATURES.

Measure	USAD					
	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$	$k = 35$
F1 score	0.8286	0.8471	0.8418	**0.8617**	0.8587	0.8259
Precision	0.9879	0.9193	0.9521	0.9545	0.9603	**0.9918**
Recall	0.7135	**0.7853**	0.7544	**0.7853**	0.7765	0.7075
Accuracy	0.9641	0.9655	0.9655	**0.9693**	0.9689	0.9637
Training time (s)	58.8	61.1	**57.2**	63.2	72.6	63.8
Test time (s)	**7.7**	8.2	8.5	7.8	8.1	7.9

Table VI presents the performance of InterFusion built by MENDEL for the WADI dataset with the number of reduced features. We first trained InterFusion with 100% of the SWaT train dataset, then retrained InterFusion with 3% of the WADI train dataset, and evaluated the retrained InterFusion with the WADI test dataset. The retrained InterFusion produced the best F1 score (0.6541) when $k = 30$.

TABLE VI
PERFORMANCE OF INTERFUSION VIA MENDEL FOR WADI WITH THE NUMBER OF REDUCED FEATURES.

Measure	InterFusion					
	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$	$k = 35$
F1 score	0.4362	0.4917	0.5888	0.6360	**0.6541**	0.6417
Precision	0.2939	0.4015	0.6840	0.8263	**0.8905**	0.8458
Recall	**0.8448**	0.6343	0.5169	0.5169	0.5169	0.5169
Accuracy	0.8742	0.9245	0.9584	0.9659	**0.9685**	0.9667
Training time (s)	445.3	438.8	437	**425.6**	442.8	439.5
Test time (s)	**1195.2**	1269.3	1222.7	1226.1	1393.6	1319.9

Table VII presents the performance of USAD built by MENDEL for the WADI dataset with the number of reduced features. We first trained USAD with 100% of the SWaT train dataset, then retrained USAD with 3% of the WADI train dataset, and evaluated the retrained USAD with the WADI test dataset. The retrained USAD produced the best F1 score (0.4589) when $k = 35$.

TABLE VII
PERFORMANCE OF USAD VIA MENDEL FOR WADI WITH THE NUMBER OF REDUCED FEATURES.

Measure	USAD					
	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$	$k = 35$
F1 score	0.3827	0.2892	0.3633	0.3664	0.4207	**0.4589**
Precision	0.3288	0.2636	0.3030	0.3510	0.4662	**0.5718**
Recall	**0.4576**	0.3199	0.4535	0.3832	0.3832	0.3832
Accuracy	0.9149	0.9094	0.9084	0.9236	0.9392	**0.9479**
Training time (s)	93.8	94.1	110.9	108.8	**91.8**	109.1
Test time (s)	**3.0**	3.8	**3.0**	3.1	3.2	3.4

Based on these results, we recommend using at least $k = 15$. In addition, the percentage of explained variance provided by PCA was used to choose the number of nonlinear principal components. The explained variance ratio is the percentage of variance that is attributed to each of the selected features. Table VIII presents the explained variance ratio for k features in the SWaT and WADI datasets, respectively. For SWaT, we obtained 99.9% of the best explained variance ratio when $k = 30$ while for WADI, we obtained 97.9% of the best explained variance ratio when $k = 35$. Therefore, we finally recommend using $k = 30$. All subsequent experiments were conducted using $k = 30$.

TABLE VIII
EXPLAINED VARIANCE RATIO FOR k FEATURES.

k	Dataset	
	SWaT	WADI
10	96.0%	82.0%
15	98.0%	89.0%
20	99.4%	93.0%
25	99.8%	95.3%
30	**99.9**%	96.9%
35	**99.9**%	**97.9**%

F. Effects of the train data size for retraining

MENDEL aims to build a retrained model with a small proportion of a target domain ICS dataset comparable with the model trained with the entire target domain ICS dataset. To find the optimal size of the samples for retraining, we evaluated the performance of two anomaly detection models, InterFusion and USAD, with x% of a target domain ICS dataset.

Table IX presents the performance of InterFusion built by MENDEL for the SWaT dataset with varying the size of retraining samples. We first trained InterFusion with 100% of the WADI train dataset, then retrained

InterFusion with $x\%$ of the SWaT train dataset, and evaluated the retrained InterFusion with the SWaT test dataset. Surprisingly, the retrained InterFusion produced the best F1 score (0.8757) when $x = 1\%$.

TABLE IX
PERFORMANCE OF INTERFUSION VIA MENDEL WITH $x\%$ OF THE SWAT TRAIN DATASET.

Measure	InterFusion				
	$x = 1\%$	$x = 3\%$	$x = 5\%$	$x = 7\%$	$x = 10\%$
F1 score	**0.8757**	0.8559	0.8659	0.8586	0.8584
precision	**0.9775**	0.9414	0.9660	0.9481	0.9378
Recall	**0.7932**	0.7846	0.7846	0.7846	0.7913
Accuracy	**0.9726**	0.9679	0.9705	0.9686	0.9683
Training time (s)	**145.6**	428.4	738.2	1007.1	1241.6
Test time (s)	7292.7	**7221.5**	7850.4	8050.6	8169.4

Table X presents the performance of USAD built by MENDEL for the SWaT dataset with varying the size of retraining samples. We first trained USAD with 100% of the WADI train dataset, then retrained USAD with $x\%$ of the SWaT train dataset, and evaluated the retrained USAD with the SWaT test dataset. The retrained USAD produced the best F1 score (0.8757) when $x = 5\%$.

TABLE X
PERFORMANCE OF USAD VIA MENDEL WITH $x\%$ OF THE SWAT TRAIN DATASET.

Measure	USAD				
	$x = 1\%$	$x = 3\%$	$x = 5\%$	$x = 7\%$	$x = 10\%$
F1 score	0.8330	0.8587	**0.8619**	0.8617	0.8579
precision	0.9505	0.9603	0.9637	**0.9651**	0.9584
Recall	0.7413	0.7765	**0.7783**	**0.7783**	0.7765
Accuracy	0.9639	**0.9689**	0.9695	0.9696	0.9687
Training time (s)	**24.0**	72.0	89.0	162.5	196.5
Test time (s)	8.6	**8.1**	8.2	10.2	8.5

Table XI presents the performance of InterFusion built by MENDEL for the WADI dataset with varying the size of retraining samples. We first trained InterFusion with 100% of the SWaT train dataset, then retrained InterFusion with $x\%$ of the WADI train dataset, and evaluated the retrained InterFusion with the WADI test dataset. The retrained InterFusion produced the best F1 score (0.7215) when $x = 10\%$.

TABLE XI
PERFORMANCE OF INTERFUSION VIA MENDEL WITH $x\%$ OF THE WADI TRAIN DATASET.

Measure	InterFusion				
	$x = 1\%$	$x = 3\%$	$x = 5\%$	$x = 7\%$	$x = 10\%$
F1 score	0.5353	0.6541	0.6549	0.6529	**0.7215**
precision	0.6528	0.8905	**0.8935**	0.6780	0.6197
Recall	0.4536	0.5169	0.5169	0.6295	**0.8631**
Accuracy	0.9546	0.9685	**0.9686**	0.9614	0.9616
Training time (s)	**145.6**	428.4	738.2	1007.1	1421.6
Test time (s)	1407.1	1393.6	**1388.3**	1406.2	1407.6

Table XII presents the performance of USAD built by MENDEL for the WADI dataset with varying the size of retraining samples. We first trained USAD with 100% of the SWaT train dataset, then retrained USAD with $x\%$ of the WADI train dataset, and evaluated the retrained USAD with the WADI test dataset. The retrained USAD produced the best F1 score (0.4825) when $x = 5\%$.

TABLE XII
PERFORMANCE OF USAD VIA MENDEL WITH $x\%$ OF THE WADI TRAIN DATASET.

Measure	USAD				
	$x = 1\%$	$x = 3\%$	$x = 5\%$	$x = 7\%$	$x = 10\%$
F1 score	0.3057	0.4207	**0.4825**	0.4630	0.4448
precision	0.2927	0.4662	**0.6513**	0.5846	0.5299
Recall	0.3199	**0.3832**	**0.3832**	**0.3832**	**0.3832**
Accuracy	0.9163	0.9392	**0.9526**	0.9488	0.9449
Training time (s)	**36.6**	91.8	161.3	249.9	347.3
Test time (s)	**3.1**	3.2	3.2	3.4	3.2

These results demonstrate that MENDEL can be used to build an effective anomaly detection model by retraining a small proportion of a target domain ICS dataset. For the SWaT dataset, MENDEL constructed an InterFusion model achieving an F1 score of 0.8757, which is better than 0.8261 achieved by InterFusion without retraining. Similarly, for the WADI dataset, MENDEL constructed an InterFusion model achieving an F1 score of 0.7251, which is significantly better than 0.4440 achieved by InterFusion without retraining. For the SWaT dataset, MENDEL constructed a USAD model achieving an F1 score of 0.8619, which is better than 0.8226 achieved by USAD without retraining. However, for the WADI dataset, MENDEL failed to produce a better USAD model than USAD without retraining. The best retrained USAD model achieved an F1 score of 0.4825, which is worse than 0.5280 achieved by USAD without retraining.

V. LIMITATIONS

We evaluated the performance of MENDEL with only two ICS datasets, SWaT and WADI. Therefore, we need to consider additional ICS datasets such as HIL-based Augmented ICS (HAI) [26] for generalization. We surmise that the performance of MENDEL would be downgraded with an ICS dataset having features which are significantly different from the features of a pre-trained model.

VI. RELATED WORK

A. Unsupervised anomaly detection

Unsupervised anomaly detection models are broadly categorized into three types: Autoencoder (AE)-based, Long Short-Term Memory (LSTM)-based, and Generative adversarial networks (GAN)-based.

The AE-based approaches [4], [5], [7], [8], [18] detect anomalies by measuring the differences between the output samples through the encoding/decoding processes and the original data samples. Audibert et al. [5] presented an effective anomaly detection model dubbed

USAD that considers adversarial examples in the AE training process.

The LSTM-based approaches [19], [20], [27] extends LSTM models to overcome the limitations of conventional LSTM. Ergen et al. [19] proposed a joint optimization model using an LSTM-based anomaly detection algorithm to resolve the shortcomings of existing studies in which performance varies greatly depending on the window size. The joint optimization is composed of stacked LSTM and One-class SVM (OC-SVM). The joint optimization comprises stacked LSTM and OC-SVM. Stacked LSTM converts all input window sizes into a fixed size and detects anomalies through the OC-SVM classifier. According to the results of OC-SVM, joint optimization is performed while updating the parameters of stacked LSTM and OC-SVM. Li et al. [20] proposed an LSTM-based AE model with OC-SVM to solve several problems of unbalanced and high-dimensional datasets. This model consists of two main parts (AE and OC-SVM). The model reconstructs input data through the AE part and classifies the reconstructed data as normal or anomalous in the OC-SVM part.

The GAN-based approaches [6], [10] exploit GAN models by artificially generating samples that cannot be seen in the training dataset to improve the model's robustness against new and unseen samples. Li et al. [6] proposed an LSTM-based GAN model that can detect various attacks due to real-world dynamic complexities of Cyber-Physical Systems (CPS). This model consists of a generator and discriminator made of LSTM-RNN and suggests a new anomaly score named the Discrimination and Reconstruction score (DR-score). Singla et al. [10] proposed a GAN model to solve the problem of lack of labeled data in network intrusion detection (NID). This model uses PCA to apply domain adaptation and uses three loss functions (discriminator's domain loss, generator's domain loss, and generator's class loss) in the discriminator.

There were several attempts to apply transfer learning to improve performance [28]. Maschler et al. [28] proposed a model that can be applied to other fields where training data is lacking in the field with prior knowledge. This model provides pre-training and interpretation using LSTM, CNN, and fully connected models.

However, existing approaches mainly focused on developing a model with a given ICS dataset. Therefore, a time consuming training process is needed for each ICS dataset. In this paper, we propose a new technique dubbed MENDEL to develop an effective anomaly detection model without retraining the entire ICS dataset.

B. Transfer learning

Transfer learning [29] is a machine learning technique in which learning of a target task is carried out with given data and knowledge transferred from a source task. Transfer learning techniques have been mainly used in studies using images and network traffic data.

Abbas et al. used a CNN model for transfer learning [30] consisting of three phases (decompose, transfer, and compose). In the decompose phase, an ImageNet model is used as a pre-trained model to perform feature extraction. Fine-tune and optimization are performed in the transfer phase, and classification is performed in the composing phase. Wang et al. [31] proposed a scheme to detect unknown attacks by converting time series data into image data using Mahalanobis Distance (MD) matrix. This method can achieve high efficiency in processing a large-scale dataset. Borgli et al. [32] focused on improving the detection accuracy of anomaly detection models using a hyperparameter optimization technique for transfer learning.

There are several studies [33], [34] using network traffic data. Mahdavi et al. [33] used incremental learning to solve the problem of lack of labeled data in NID and overcame the challenges that new technologies accompany new vulnerabilities. This study used incremental learning in transfer learning to decrease training time. Zhang et al. [34] proposed an anomaly detection model using transfer learning techniques. Their experimental results demonstrated that transfer learning could effectively reduce the training overhead required to learn the characteristics of new samples, which is consistent with our experimental results.

Previous studies showed that transfer learning techniques would be useful for image and network domains. However, these approaches cannot directly be applied to multivariate time series data. To address this issue, Wang et al. [31] proposed a transfer learning technique using a CNN by converting ICS data into images. We extend previous transfer learning approaches into a more practical one for ICS data to directly process multivariate time series data with different feature sets.

VII. Conclusion

In this study, we presented a transfer learning technique dubbed MENDEL to efficiently train anomaly detection models for ICS using a pre-trained model. MENDEL uses PCA to reduce the features of a given dataset to a specified number of features so that each model can be constructed with the same number of reduced features. Consequently, a source domain model with reduced features can be used for a target domain model through retraining. For retraining, MENDEL uses a proper mapping to efficiently map the source domain model's features into the target domain model's features. Our evaluation results demonstrated that MENDEL could be used to construct an effective anomaly detection for ICS in multivariate time series data. Overall, the

models built by MENDEL achieved high F1 scores even when a model is retrained with only a small proportion of the target domain dataset, which can significantly reduce the training overhead to build an anomaly detection model for a new ICS dataset.

REFERENCES

[1] M. Faisal and M. Ibrahim, "Stuxnet, Duqu and beyond," *International Journal of Science and Engineering Investigations*, vol. 1, no. 2, pp. 75–78, 2012.

[2] B. Bencsáth, G. Pék, L. Buttyán, and M. Felegyhazi, "The Cousins of Stuxnet: Duqu, Flame, and Gauss," *Future Internet*, vol. 4, no. 4, pp. 971–1003, 2012.

[3] J. Rrushi, H. Farhangi, C. Howey, K. Carmichael, and J. Dabell, "A quantitative evaluation of the target selection of Havex ICS malware plugin," in *Industrial control system security (ICSS) workshop*, 2015.

[4] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate Time Series Anomaly Detection and Interpretation Using Hierarchical Inter-Metric and Temporal Embedding," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3220–3230.

[5] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised Anomaly Detection on Multivariate Time Series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3395–3404.

[6] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate Anomaly Detection For Time Series Data With Generative Adversarial Networks," in *Proceeding of the 29th International Conference on Artificial Neural Networks*, 2019, pp. 703–716.

[7] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini, "Anomaly detection with conditional variational autoencoders," in *Proceeding of the 18th IEEE International Conference on Machine Learning and Applications*, 2019, pp. 1651–1657.

[8] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng *et al.*, "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," in *Proceedings of the 27th International Conference World Wide Web Conference*, 2018, pp. 187–196.

[9] L. Chen, Y. Li, X. Deng, Z. Liu, M. Lv, and H. Zhang, "Dual Auto-Encoder GAN-Based Anomaly Detection for Industrial Control System," *Applied Sciences*, vol. 12, no. 10, p. 4986, 2022.

[10] A. Singla, E. Bertino, and D. Verma, "Preparing network intrusion detection deep learning models with minimal data using adversarial domain adaptation," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, 2020, pp. 127–140.

[11] E. Otović, M. Njirjak, D. Jozinović, G. Mauša, A. Michelini, and I. Stajduhar, "Intra-domain and cross-domain transfer learning for time series data—how transferable are the features?" *Knowledge-Based Systems*, vol. 239, p. 107976, 2022.

[12] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[13] A. P. Mathur and N. O. Tippenhauer, "SWaT: a water treatment testbed for research and training on ICS security," in *"Proceedings of the 2nd International Workshop on Cyber-physical Systems for Smart Water Networks"*, 2016, pp. 31–36.

[14] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: a water distribution testbed for research in the design of secure cyber physical systems," in *Proceedings of the 3rd International Workshop on Cyber-physical Systems for Smart Water Networks*, 2017, p. 25–28.

[15] K. Stouffer, J. Falco, K. Scarfone *et al.*, "Guide to industrial control systems (ICS) security," *NIST special publication*, vol. 800, no. 82, pp. 16–16, 2011.

[16] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2828–2837.

[17] X. Xie, B. Wang, T. Wan, and W. Tang, "Multivariate abnormal detection for industrial control systems using 1D CNN and GRU," *IEEE Access*, vol. 8, pp. 88 348–88 359, 2020.

[18] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proceeding of the 18th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 2366–2370.

[19] T. Ergen and S. S. Kozat, "Unsupervised anomaly detection with LSTM neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 8, pp. 3127–3141, 2019.

[20] M. Said Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "Network anomaly detection using LSTM based autoencoder," in *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, 2020, pp. 37–45.

[21] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[22] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Transactions on Information theory*, vol. 49, no. 7, pp. 1858–1860, 2003.

[23] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[24] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," *Advances in neural information processing systems*, vol. 29, 2016.

[25] J. Tomczak and M. Welling, "VAE with a VampPrior," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1214–1223.

[26] H.-K. Shin, W. Lee, J.-H. Yun, and B.-G. Min, ""Two ICS Security Datasets and Anomaly Detection Contest on the HIL-based Augmented ICS Testbed"," in *Proc. of the Cyber Security Experimentation and Test Workshop*, 2021.

[27] X. Zhou, Y. Hu, W. Liang, J. Ma, and Q. Jin, "Variational LSTM enhanced anomaly detection for industrial big data," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3469–3477, 2020.

[28] B. Maschler, T. Knodel, and M. Weyrich, "Towards deep industrial transfer learning for anomaly detection on time series data," in *Proceeding of the 26th IEEE International Conference on Emerging Technologies and Factory Automation*, 2021, pp. 1–8.

[29] L. Torrey and J. Shavlik, "Transfer Learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[30] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Detrac: transfer learning of class decomposed medical images in convolutional neural networks," *IEEE Access*, vol. 8, pp. 74 901–74 913, 2020.

[31] W. Wang, Z. Wang, Z. Zhou, H. Deng, W. Zhao, C. Wang, and Y. Guo, "Anomaly detection of industrial control systems based on transfer learning," *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 821–832, 2021.

[32] R. J. Borgli, H. K. Stensland, M. A. Riegler, and P. Halvorsen, "Automatic hyperparameter optimization for transfer learning on medical image datasets using bayesian optimization," in *Proceeding of the 13th International Symposium on Medical Information and Communication Technology*, 2019, pp. 1–6.

[33] E. Mahdavi, A. Fanian, A. Mirzaei, and Z. Taghiyarrenani, "ITL-IDS: Incremental Transfer Learning For Intrusion Detection Systems," *Knowledge-Based Systems*, vol. 253, p. 109542, 2022.

[34] S. Zhang, Z. Zhong, D. Li, Q. Fan, Y. Sun, M. Zhu, Y. Zhang, D. Pei, J. Sun, Y. Liu *et al.*, "Efficient kpi anomaly detection through transfer learning for large-scale web services," *IEEE Journal on Selected Areas in Communications*, 2022.