IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 June 2022, accepted 21 June 2022, date of publication 27 June 2022, date of current version 1 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186328

RESEARCH ARTICLE

Divergence-Based Transferability Analysis for Self-Adaptive Smart Grid Intrusion Detection With Transfer Learning

PENGYI LIAO[1], (Graduate Student Member, IEEE), JUN YAN[2], (Member, IEEE), JEAN MICHEL SELLIER[3], AND YONGXUAN ZHANG[4]

[1]Department of Electrical and Computer Engineering (ECE), Concordia University, Montréal, QC H3G 1M8, Canada
[2]Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montréal, QC H3G 1M8, Canada
[3]Global AI Accelerator, Ericsson, Montréal, QC L4W 5E3, Canada
[4]Department of Computer Science and Software Engineering (CSSE), Concordia University, Montréal, QC H3G 1M8, Canada

Corresponding author: Jun Yan (jun.yan@concordia.ca)

This work was supported in part by the Ericsson—Global Artificial Intelligence Accelerator (GAIA) in Montréal and Mitacs Accelerate under Grant IT-15923, and in part by the Fonds de Recherche du Québec–Nature et Technologies (FRQNT) under Grant 2019-NC-254971.

ABSTRACT Machine learning is a popular approach to security monitoring and intrusion detection in cyber-physical systems (CPS) like the smart grid. However, these highly dynamic CPS operating in open environments can result in significant data distribution divergence, which may require the adaptation of a learned model. While transfer learning has been an effective approach to retain the performance against the divergence, there is still limited work on a more fundamental question that can be called *transferability*: when should one apply transfer learning? To address this challenge, this paper proposes a divergence-based transferability analysis to decide whether to apply transfer learning and autonomically adapt learning-based intrusion detectors. This work first identifies three metrics used to measure the divergence between data distributions, and then explores the relation between detector's accuracy drop and divergence in extensive temporal, spatial, and spatiotemporal experiments. Two regression models are trained to approximate the divergence-accuracy relation and then used to predict an accuracy drop which determines whether to apply transfer learning. Finally, a state-of-the-art domain adversarial neural network (DANN) classifier is adopted as the transfer learning model. Datasets from real normal operation profiles and simulated attacks are used to validate the effectiveness of the proposed transferability analysis against variations in attack timing, locations, and both. In all three scenarios, the proposed analysis demonstrated high accuracy in predicting accuracy drop from the divergence, with an RMSE lower than 4.20%, and the DANN can be timely triggered to achieve an accuracy improvement over 5.00%.

INDEX TERMS Transferability analysis, adversarial training, false data injection, intrusion detection, data distribution divergence, domain adaptation, smart grid.

I. INTRODUCTION

The smart grid is a trans-continental cyber-physical system (CPS) empowered by the advancement of declining costs in communications, advanced sensors, and distributed computing technology. The grid connects utilities and customers with two-way power and information flows to provide more efficiency, reliability, and safety of power delivery. However, the growing number of interconnections among billions of cyber-physical devices creates complex interdependence and vulnerabilities that will inevitably raise the occurrence of cyber attacks in power systems.

The impact of a cyber attack on CPS could be grievous and disastrous, as demonstrated by recent research efforts, business studies, and real-world incidences [1]–[3]. Machine learning (ML), which undergoes the rigorous process of

The associate editor coordinating the review of this manuscript and approving it for publication was Giambattista Gruosso.

designing and implementing algorithms with expected performance, has acclaimed significant attention in the smart grid security research community. A rich line of ML approaches have significantly enhanced the cyber-physical situational awareness and security monitoring [4], [5].

General ML approaches presume that the training and testing data are generated by identical or similar independent distribution. This assumption may not hold in many real-world systems and applications like the CPS, since the system dynamics may change the data distribution and thus fail the trained model. This poses a challenge to ML in the CPS scenarios. In the smart grid, as an example, the load demand is constantly changing, and there are variations from normal operations, grid topology and attack patterns [3], [6].

Moreover, despite the mounting risk, labeled attack data are still rare in the smart grid. The model trained on limited attack data can be brittle, and seemingly slight changes in the data distribution may lead to performance degradation [7]. Transfer learning (TL) is hence proposed to address this challenge by transferring learned knowledge from a labeled source domain to a related target domain. It has been extensively adopted and witnessed remarkable advances in natural language processing (NLP), image and video applications [8]. Lately, TL methods are applied to anomaly detection in Internet and cloud applications [9], which sheds new light on introducing TL to empower intrusion detection in highly dynamic cyber-physical power systems [6].

While most existing TL works focus on designing a sophisticated model to achieve state-of-the-art performance, little attention has been paid to answering a more fundamental question, especially in the field of CPS, that may be called transferability: when should one consider the performance of a trained model has degraded significantly enough to justify the need for TL, without having to retrain a new model from scratch? Studies have indicated that TL performance is related to the similarity between the source and target domains [10], and the effectiveness of TL may remain high in a certain range of distribution divergence, depending on how critical the application scenarios are.

We can use Fig. 1 as an illustrative example of the possible relation between the effectiveness of TL and distribution divergence. If the divergence between the source domain and target domain is within a small range, the ML model trained on the source domain can generally retain a good performance when applied to the target domain, so TL is not necessary as the performance boost would be trivial while the adaptation can be costly. Meanwhile, if the divergence is too large, even a TL model could suffer a severe accuracy drop on the target domain as the case is beyond transferable. In this case, one would better train a new model from scratch instead of applying TL. If the divergence is somewhere in between, it may be significant enough to degrade the performance of a trained ML model but not beyond what a TL model can handle. This will be the sweet spot where we can leverage TL to retain a good performance against the divergence by adapting - instead of re-applying or re-creating an ML model.

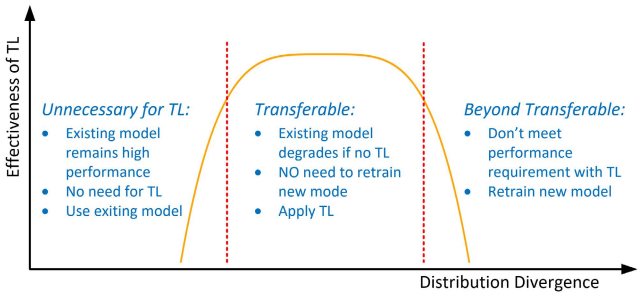


FIGURE 1. The relation between divergence and effectiveness of TL.

In this paper, we will mainly focus on the transferability between the first and the third situations, where one need to decide if the divergence is making it necessary to transfer an existing ML model with effective TL methods.

Based on the remaining gap and above analysis, this paper explores the landscape of TL by proposing a divergence-based transferability analysis in CPS. First, the proposed approach leverages three metrics chosen from different families to measure data distribution divergence between source and target domains. Then, two regression models are trained to approximate the relation and applied to predict the accuracy drop of the unlabeled target domain. Finally, the target domain which requires TL is identified, and we adopt domain adversarial neural networks (DANN) as the TL model to retain robust detection accuracy.

To validate the effectiveness of the proposed approach, this paper extracts one week of operation data from ISO New England as the normal data. The widely studied false data injection (FDI) attack is then used to generate the attack data. Considering attacks on different periods and/or locations in the smart grid, we synthesize the datasets with temporal, spatial, and spatiotemporal variations. The results demonstrate that the transferability analysis has high accuracy in predicting accuracy drop, and the intrusion detector can retain a robust accuracy after TL is timely applied.

The main contributions of this paper can be summarized as follows:

1) We propose a divergence-based transferability analysis to help evaluate the necessity of TL in security monitoring for CPS, such as intrusion detection in the smart grid.
2) We evaluate the accuracy drop and data distribution divergence with multiple metrics to reveal the relation between accuracy drop and divergence in dynamic CPS operations.
3) This paper considers the situation where attacks may happen at different time and/or locations during the power system operation, which may lead to temporal and/or spatial divergence that would result in an accuracy drop.
4) The results demonstrate three metrics (PAD, KL, and MMD) that are able to predict accuracy drop against both temporal and spatial divergence, which shows that

the transferability analysis can improve the adoption of TL in realistic CPS security monitoring.

The rest of this paper is organized as follows: Section II discusses the related work about transferability analysis. Section III illustrates the proposed transferability analysis for self-adaptive TL. Section IV introduces the experiments setup. Section V presents the simulation results and analysis. Section VI draws the conclusions and future works.

II. RELATED WORKS

In this paper, we are interested in the data distribution divergence that can lead to a significant accuracy drop and require TL. Divergence metrics like the maximum mean discrepancy (MMD) and Kullback–Leibler (KL) divergence are commonly used to measure distribution divergence in ML, including TL.

Recently, several studies outside the field of TL have started to establish a connection between the distribution divergence and accuracy. Among them, the most related work is from Elsahar and Galle [11], [12], who used various methods such as H-divergence, Fréchet distance and confidence-based metrics to predict the accuracy drop of modern NLP and computer version (CV) models under domain shifts. Both studies above used predicted accuracy drops to evaluate the robustness of trained models. However, these studies did not further explore the use of the predicted accuracy to determine whether/how the model shall be updated to retain the previous performance. Instead, in our work, we use divergence metrics as an indicator to determine whether one should apply TL, benchmarking three divergence metrics in predicting the accuracy drop to create a reliable predictor that will help operators decide whether to apply TL based on the predicted performance degradation of trained machine learning models.

There are also some studies that have used divergence metrics, such as MMD and cross-entropy, as the domain confusion loss in TL. Long *et al.* [13]–[15] leverage divergence metrics to measure the dissimilarity between the source and target domains as the domain confusion loss, then combine this loss with the classification loss as the total loss during training. Meanwhile, we introduce the divergence before the TL training process to decide whether one shall apply TL or not. Instead of using the divergence metrics as domain distribution dissimilarity in the training stage, we leverage the divergence metrics to predict the accuracy drop in the pre-training stage and trigger the TL process if the predicted accuracy drop is in a suitable range.

Divergence has also been introduced in power system studies. For example, Gupta *et al.* [16] use the relative entropy between normal and the perturbed power flow data, to predict the blackout risk. Tajdinian and Samet [17] propose a method based on Kullback–Leibler (KL) divergence for discriminating inrush and internal fault currents in power transformers. Compared to these works, our transferability analysis focuses on predicting the performance degradation based on the divergence metric to decide when TL should be triggered, instead of using the metric to measure the dissimilarity for a direct alert. We also consider different events of attacks, which may have more intentionally developed schemes than typical faults and thus can be more challenging to detect than the events in the aforementioned detectors using PMU-data.

It is notable that divergence itself can be used directly in the power system attack detection without TL. Chaojun *et al.* [18] use KL divergence to calculate the distance between normal and false data to identify the latter directly. Pal *et al.* [19] measure the Euclidean distance between real and tampered data to detect the data manipulation attacks directly. In both studies, the authors use divergence metrics to measure the dissimilarity between two data distributions and generate alerts for anomalies directly. However, their uses of the divergence have only focused on the dissimilarity between different events that typically have an inherent distinction in the data distribution at any moment. Meanwhile, our work considers the distribution divergence under the same event (normal and attack, respectively), which can change over time and space in a more intriguing way. In addition, these studies did not consider the following domain adaptation with TL, which can significantly retain or improve a trained model's robustness against the divergence. Our study uses divergence metrics not to generate alerts directly but to decide whether TL needs to be triggered. Meanwhile, as a pre-determination step of TL, our transferability analysis can be combined with TL to adapt the event detectors in the studies above to retain their performance under dynamic operating environments.

Inspired by the existing work, we investigate different divergence measurement metrics and choose three from them to measure the data distribution divergence between different domains. With these metrics, we want to identify the potential relation between attack detection accuracy drop in the smart grid and distribution divergence and approximate the relation through regression models.

III. TRANSFERABILITY ANALYSIS FOR SELF-ADAPTIVE TL

A. PROBLEM FORMULATION

CPS operating in open environments may have significant data distribution divergence, which may lead to accuracy degradation for a model trained on the source domain and tested on the target domain. Given source domain $\mathcal{D}_S$ and target domain $\mathcal{D}_T$, the distribution divergence may be caused by a variety of reasons. It can be due to covariate divergence, where only the feature distribution changes, i.e., $P_{D_S}(X) \neq P_{D_T}(X)$, but the conditional distribution remains the same, i.e., $P_{D_S}(Y|X) = P_{D_T}(Y|X)$. It may be caused by concept divergence, where $P_{D_S}(X) = P_{D_T}(X)$ and $P_{D_S}(Y|X) \neq P_{D_T}(Y|X)$, or label divergence, where $P_{D_S}(Y) \neq P_{D_T}(Y)$ and $P_{D_S}(X|Y) = P_{D_T}(X|Y)$, or a combination of the above divergence.

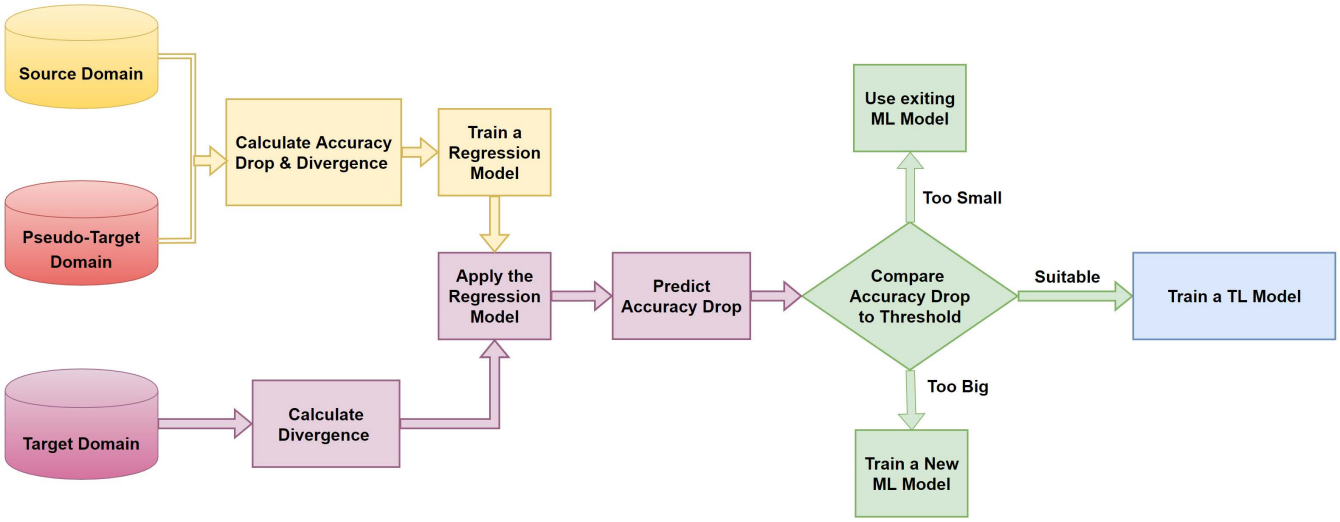


FIGURE 2. The proposed divergence-based transferability analysis for the smart grid intrusion detection.

This paper focuses on the covariate divergence, which often occurs in CPS intrusion detection scenarios because the system variations and attack variations will influence the normal and attack data distribution. In the power system, the system variations may be caused by the different load demands, normal operations, or topology changes. The attack variations could arise when the same scheme is launched again at different periods or locations in the grid. This paper considers the binary intrusion detection problem in the smart grid, which intends to classify the multivariate time series measurements data as attack events or normal operations. We focus on the scenarios where two consecutive attacks target on different time and/or different locations.

We are interested in the attack detection accuracy drop that requires the TL. We assume that the source domain consists of labeled normal data and attack data, where $\mathcal{D}_S = \{(x_{S_1}, y_{S_1}), \ldots, (x_{S_{n_1}}, y_{S_{n_1}})\}$. If we had a fully labeled target domain $\mathcal{D}_\mathcal{T}$, the accuracy drop could be measured by empirical data:

$$\Delta Pr = Pr_{D_S} - Pr_{D_T}, \tag{1}$$

where Pr_{D_S} is the accuracy of a model trained on a source domain $\mathcal{D}_\mathcal{S}$, and Pr_{D_T} is the accuracy of this model when applied to the target domain $\mathcal{D}_\mathcal{T}$. However, the detection system deployed in the smart grid detects attacks online, and the new generated target domain is unlabeled, i.e., $\mathcal{D}_T = \{(x_{T_1}), \ldots, (x_{T_{n_2}})\}$. So, the accuracy drop can not be calculated via (1) with the unlabeled target domain.

To solve the problem, we introduce the labeled pseudo target domain $\mathcal{D}_{T_s} = \{(x_{T_{s_1}}, y_{T_{s_1}}), \ldots, (x_{T_{sn_2}}, y_{T_{sn_2}})\}$. Since the pseudo target domain is labeled, we can measure the attack detection accuracy. We propose to explore the relation between accuracy drop and divergence between the source domain and the pseudo target domain, then use the relation to predict the accuracy drop of the unlabeled target domain. If we have the divergence-accuracy drop relation, the accuracy drop of the target domain can be predicted by:

$$\Delta Pr' = A(d), \tag{2}$$

where $\Delta Pr'$ is the predicted accuracy drop, A is the relation between accuracy drop and divergence, and d is the distribution divergence of the source domain and the target domain.

The proposed transferability analysis aims to predict accuracy drop and identify the unlabeled target datasets where a trained model will degrade significantly and call for TL. The challenges are how to measure the data distribution divergence in intrusion detection and how to approximate the relation between accuracy drop and divergence, which are tackled by the proposed framework in Fig. 2. The framework has two phases: 1) We measure the accuracy drop and divergence between each pair of source dataset and pseudo target dataset, then train regression models to approximate the divergence-accuracy drop relation. For an unlabeled target dataset, calculate the divergence and predict the accuracy drop with the relation model. 2) If the predicted accuracy drop is in a suitable range, trigger the TL.

B. DATA DISTRIBUTION DIVERGENCE METRICS

Considering the properties of different metrics, we select three widely used divergence metrics of different families from the literature:

- Classifier-based metric: depending on the capability of a basic classifier to discriminate between samples generated from source and target domains, like Proxy $\mathcal{A}$-Distance (PAD).
- Information theory-based metric: measuring the information gain required to code samples from one distribution using a code optimized for another distribution, like Kullback–Leibler (KL) divergence and Jensen-Shannon

(JS) divergence [20]. KL has shown effectiveness in predicting performance in sentiment analysis [21], so we pick KL in this paper.

- Higher-order moment-based metric: projecting higher-order moments of random variables to a new feature representation space, like Maximum Mean Discrepancy (MMD), Central Moment Discrepancy (CMD), and Correlation Alignment (CORAL) [22]. In this paper, we choose the MMD because MMD can make use of the kernel trick and is applicable to a wide range of data type [21].

1) PROXY $\mathcal{A}$-ISTANCE

PAD is a $\mathcal{H}$-divergence based metric proposed by Ben-David *et al.* [23]. Ben-David *et al.* have proven that the error of a trained model on a target domain is bounded by its error on the source domain and the $\mathcal{H}$-divergence between the source domain and the target domain. $\mathcal{H}$-divergence depends on the capability of a trained classifier to discriminate between samples generated from source and target domains. Although $\mathcal{H}$-divergence is hard to calculate, Ben-Davidr *et al.* approximates it by Proxy $\mathcal{A}$-Distance (PAD).

To calculate the PAD, source domain data and target domain data are mixed, and samples from source and target domains are labeled as 0 and 1, respectively. Then a classifier G_d is trained on the mixed dataset to distinguish between samples from source and target domains. Finally, the classifier is tested on the held-out test dataset. The PAD is defined as:

$$\epsilon\left(G_d\right) = \frac{1}{|D|} \sum_{x_i \in D'_s, D'_t} \left|G\left(x_i\right) - I\left(x_i\right)\right|, \tag{3}$$

$$PAD = 2(1 - 2\epsilon\left(G_d\right)), \tag{4}$$

where G_d is the trained classifier. $\epsilon\left(G_d\right)$ is the classifier's error on the held-out dataset D'_s and D'_t. I is the true domain label. In the experiments of this paper, following the approach of Ben-David *et al.* [23], we train a linear SVM as our classifier.

2) KULLBACK–LEIBLER (KL) DIVERGENCE

KL [17] is the relative entropy between two probability density functions $p(x)$ and $q(x)$:

$$D_{KL}(P||Q) = \int p(x) log \frac{p(x)}{q(x)} dx. \tag{5}$$

We adopt the work of Hershey and Olsen *et al.* [24] and consider the two datasets follow Gaussian Mixture Models (GMM). The marginal densities of $x \in \mathbb{R}^d$ under p and q are

$$p(x) = \sum_a \pi_a \mathcal{N}(x; \mu_a; \Sigma_a),$$

$$q(x) = \sum_b \pi_b \mathcal{N}(x; \mu_b; \Sigma_b). \tag{6}$$

To estimate $D(P||Q)$, we could conduct Monte Carlo simulation. Using n i.i.d. samples $\{x_i\}_n^{i=1}$, we have:

$$D_{MC}(P||Q) = \frac{1}{n} \sum_{i=1}^{n} log \frac{p(x_i)}{q(x_i)} \to D(P||Q). \tag{7}$$

The variance of the estimation error could be decreased when $n \to \infty$.

3) MAXIMUM MEAN DISCREPANCY (MMD)

MMD is a non-linear metric widely used in TL. MMD estimates divergence between two distributions based on the Reproducing Kernel Hilbert Space (RKHS) [25]. Given two datasets $X = \{x_1, x_2, \ldots, x_{n_1}\}$ and $Y = \{y_1, y_2, \ldots, y_{n_2}\}$ that come from two distribution P and Q, the empirical estimation of the distance is defined by:

$$D_{MMD}(X||Y) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi(x_i) - \frac{1}{n_2} \sum_{j=1}^{n_2} \varphi(y_j) \right\|_H. \tag{8}$$

where $\varphi(x)$: $\mathcal{X} \to \mathcal{H}$, is a kernel-based function mapping samples to a feature representation space in RKHS.

The feature representation varies with the different choices of kernels. In this paper, the radial basis function (RBF) kernel is adopted since the RBF kernel can take advantage of the Taylor expansion of the Gaussian function to map all the moments of two distributions [15].

C. REGRESSION MODEL

After measuring accuracy drop and data distribution divergence by the selected metrics, the potential relation between the distribution divergence and the detector accuracy drop are approximated through regression models, including linear regression and neural network regression.

1) LINEAR REGRESSION

A strong positive relation between detection accuracy drop and divergence can be observed in Fig. 5: Pearson correlation coefficient ρ is above 0.83 in all cases. According to Haldun [26], $0.8 < \rho <= 1$ shows a strong relation between two variables. Based on this observation, we first introduce a linear regression model.

$$\Delta Pr = w_1 d + w_0, \tag{9}$$

where w_0 and w_1 are parameters of the linear regression model, d is the distance between $\mathcal{D}_S$ to $\mathcal{D}_T$, and ΔPr is the accuracy drop of a model that is trained on $\mathcal{D}_S$ and applied to $\mathcal{D}_T$.

2) NEURAL NETWORK REGRESSION

Considering that the divergence-accuracy relation may not be linear, we also leverage a neural network (NN) regression model. The neural network regression model can learn a non-linear and complicated relation between accuracy drop and divergence.

$$\Delta Pr = f_{neural}(d), \tag{10}$$

where f_{neural} is a fully connected neural network, we adopt the same configuration as [11]. The input d is the distribution divergence of two domains. The output ΔPr is the accuracy drop.

With the regression models, we can measure the divergence between the source domain and the unlabeled target domain, and predict the accuracy drop according to the divergence. If the divergence is greater than the accuracy drop threshold Π, we will trigger TL for the target domain. The entire proposed transferability analysis process is summarized in Algorithm 1.

Algorithm 1 Transferability Analysis.

Input: The set $\mathcal{S}$ of labeled source dataset $\mathcal{D}_S$; The set $\mathcal{T}_f$ of labeled pseudo target dataset $\mathcal{D}_{T_s}$; The set $\mathcal{T}$ of unlabeled target dataset $\mathcal{D}_T$; Accuracy drop upper bound Π and lower bound π

Output: TL decision

for $\mathcal{D}_S, \mathcal{D}_{T_s}$ in $\mathcal{S}, \mathcal{T}_f$ **do**
  # Measure divergence
  $d \leftarrow D(D_S||D_{T_s})$
  # Measure accuracy drop
  Train a classifier on $\mathcal{D}_S$, calculate accuracy Pr_{D_S}
  Apply classifier on $\mathcal{D}_{T_s}$, calculate accuracy $Pr_{D_{T_s}}$
  $\Delta Pr \leftarrow Pr_{D_S} - Pr_{D_{T_s}}$
end for
# Train a regression model
$\Delta Pr = A(d)$
# Predict accuracy drop for target domain
for $\mathcal{D}_S, \mathcal{D}_T$ in $\mathcal{S}, \mathcal{T}$ **do**
  $d' \leftarrow D(D_S||D_T)$
  $\Delta Pr' = A(d')$
  # Make TL decision
  if $\Delta Pr' <= \pi$ **then**
    Use exiting ML model
  else if $\pi < \Delta Pr' < \Pi$ **then**
    Train a TL model
  else
    Train a new ML model
  end if
end for

IV. EXPERIMENTS SETUP

This section will introduce our experiments setup to validate the transferability analysis for intrusion detection.

A. NORMAL DATA

To establish experiments based on realistic scenarios, we obtain public load demand from ISO New England [27] from August 24th to 30th, 2019, as shown in Fig. 3. In ISO New England, the demand was reported every 5 minutes. To increase the sampling rate and maintain the trend of the demand curve, the demand data is interpolated with a 1-second interval by the Spline method in MATLAB.

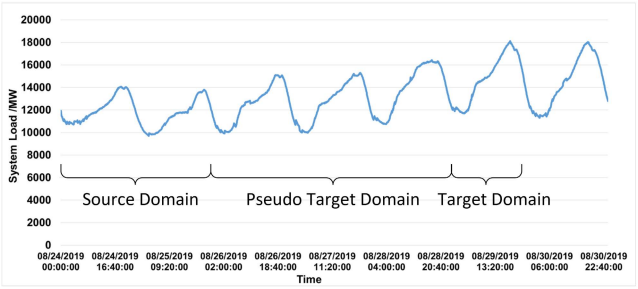


FIGURE 3. One week load demand of ISO New England [27].

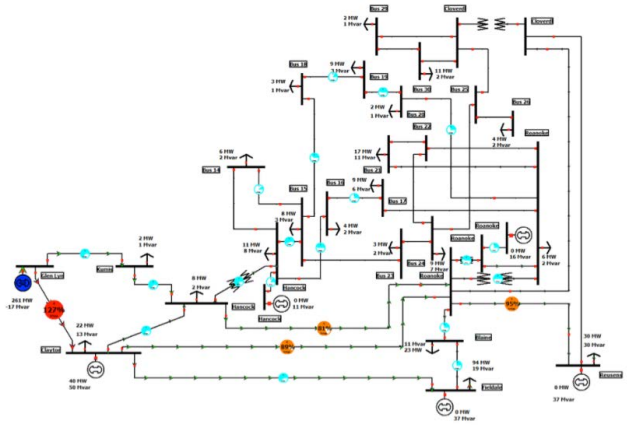


FIGURE 4. The IEEE 30-bus system by the Illinois Center for a Smarter Electric Grid (ICSEG) [28].

The IEEE 30-bus system [28] is selected as the simulation scenario, and MATLAB toolbox MATPOWER is leveraged to generate and synthesize the above load demand. As illustrated in Fig. 4, the system consists of 30 buses and 41 branches with a total load demand of 189.2 MW. We first assume that the default operating point in the 30-bus system is at its peak (100%) and match the total load demand to the peak load of the data we obtained from ISO New England. Then we assume that the total demand of the IEEE 30-bus system follows the same changes as that of the ISO New England grid (in terms of percentage w.r.t. the peak load). For example, if the total demand of the ISO New England drops from 100% at the peak to 80% after 2 hours, the total demand of the IEEE 30-bus system will also decrease to 80% of its own peak after 2 hours. This matching will allow us to apply the same aggregated load profile of the ISO New England to that of the IEEE 30-bus system.

Meanwhile, we follow [29] and introduce variations into load for each node over time. Based on [29], we assume that at a given period, if the load of the entire grid is changed by $x\%$, the corresponding individual load change across 30 buses follows a normal distribution with a mean of $x\%$ and a variance of $y\%$. For example, from t_k to t_{k+1}, if the total load of the grid is increased by 3%, the load of each node may increase similarly but with potential random variations, such as 3.2% or 2.6%, and the average will be 3%. In our

experiments, x is the change obtained from ISO New England load profile, $y = x/100$. 142 measurements over a 1-second interval are calculated and collected through the DC optimal power flow (DC-OPF) solver in MATPOWER as normal data.

B. ATTACK DATA

Distinct attack models have been proposed and developed to analyze and enhance the security of the smart grid in the past two decades [3]. The false data injection (FDI), first proposed by Liu *et al.* [30], is sophisticatedly designed to exploit a mathematical vulnerability in the residual-based bad data detector (BDD) and stealthily compromise measurements from electricity grid sensors in a coordinated fashion [31], [32]. A successful FDI attack will evade the detection and pose a severe threat to power system state estimators (PSSE) in the supervisory control and data acquisition (SCADA) systems, possibly inflicting severe impacts like power outages, physical damages, and monetary losses [33]. The FDI attack has successfully attracted the attention of lots of researchers. Therefore, in this paper, we choose the FDI attack as the attack model.

We present the formulation of FDI attack under the power flow model as [30]:

$$\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}}, \tag{11}$$

where $\mathbf{z} = z_1, z_2, \ldots, z_n$ represents the physical measurements, $\hat{\mathbf{x}} = \hat{x_1}, \hat{x_2}, \ldots, \hat{x_m}$ is the estimated states, $\mathbf{H}$ is an $n \times m$ Jacobian matrix of power grid topology, and $\mathbf{r}$ is the residual. The traditional BDD calculates the $L_2 - norm$ of residual between observed measurements $\mathbf{z}$ and estimated measurements $\mathbf{H}\hat{\mathbf{x}}$, and adopts the statistical residual tests to detect the presence of bad data via comparing the residual with a threshold τ. $\mathbf{r} > \tau$ indicates the presence of an attack.

If the attacker chooses the attack vector $\mathbf{a} = \mathbf{Hc}$, where $\mathbf{c} \sim N(0, \sigma_c^2)$ is the false state error injected into the system, and injects it into the measurements $\mathbf{z}$ by $\mathbf{z_a} = \mathbf{z} + \mathbf{a}$, the new residual will be:

$$\mathbf{r_a} = \mathbf{z_a} - \mathbf{H}\hat{\mathbf{x}_a} = (\mathbf{z} + \mathbf{a}) - \mathbf{H}(\hat{\mathbf{x}}+) = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}}. \tag{12}$$

The new residual remains the same, allowing the FDI attack to bypass the residual-based BDD. This paper will use $\mathbf{z_a}$ as the attack data, generated from the false state $\mathbf{c}$ with a mean of zero and a variance of $\sigma_c^2 = 0.1$.

C. CASE SETUP FOR TRANSFERABILITY ANALYSIS

Three scenarios are considered to validate the effectiveness of the proposed approach, including temporal, spatial, and spatiotemporal cases. As illustrated in Fig. 3, if the source domain is the data from t_1 to t_2, and the target domain is the data from t_5 to t_6, the power grid operators can choose the data from t_3 to t_4 as the pseudo target domain, where $t_3 > t_2$ and $t_4 < t_5$. So, the pseudo target domain data is the historical data and do not need to be labeled in real-time. Since it is historical data, if both normal and attack data are available, the power grid operators can choose a certain part of data from the historical data and label them as the pseudo target domain. Considering the labeled attack data could be extremely rare in the smart grid compared with the labeled normal data, if there is no attack data available, the power grid operators can review the attack models [3] in the literature and synthesize the most prominent attacks. This can still be helpful in defense planning and operations against the most prominent subset of attacks.

1) TEMPORAL CASES

First, we consider a known attack returning at different times. We assume that attackers launch the attack vector at the same locations but across different periods in temporal cases. Since the load demand and its patterns vary significantly throughout the day, we select the 4-hour time window data as the source domain and target domain to best capture the characteristics of data distributions.

As illustrated in Table 1, the source data consists of normal data launched on Day 1 and attack data with the same hours collected on Day 2. Considering the load patterns distinct in different periods of a day, we define 4 cases based on our previous work [6] according to the variation of load demand: the valley, the ascending slope, the peak, and the descending slope. For the pseudo target domain, considering attack data are rare in the smart grid, we use a 4-hour time window to collect normal data from Day 3 to Day 5. The normal data in the pseudo target domain will be used to approximate the divergence-accuracy drop relation. We also use the 4-hour time window but divide Day 6 into six intervals as the target domain for testing.

TABLE 1. Setup of cases in the temporal scenario.

#	Source Domain on Day 1 (Normal) and Day 2 (Attack)		Pseudo Target Domain from Day 3 to Day 5	Target Domain on Day 6 for Testing
	Cases	Hours		
1	Valley	2–5	4-hour time window of normal data.	2 hours of normal followed by 2 hours of attack.
2	Ascending	11–14		
3	Peak	17–20		
4	Descending	21–24		

2) SPATIAL CASES

For spatial cases, we consider attacks returning at different locations. We assume the load demand will be similar in source and target domains. We select the same 4-hour time window of the different days for source and target data while choosing different attack locations for the target data.

Since we use the IEEE 30-bus system, there are a total of 30 potential buses to be attacked. However, some buses carry zero loads and are non-attackable. We follow the reference [34] and notice that attackers will not choose to attack 15 specific buses. Hence for the IEEE 30-bus system, we have 15 attackable source domain datasets and target domian datasets. We also assume that the attackers only inject one

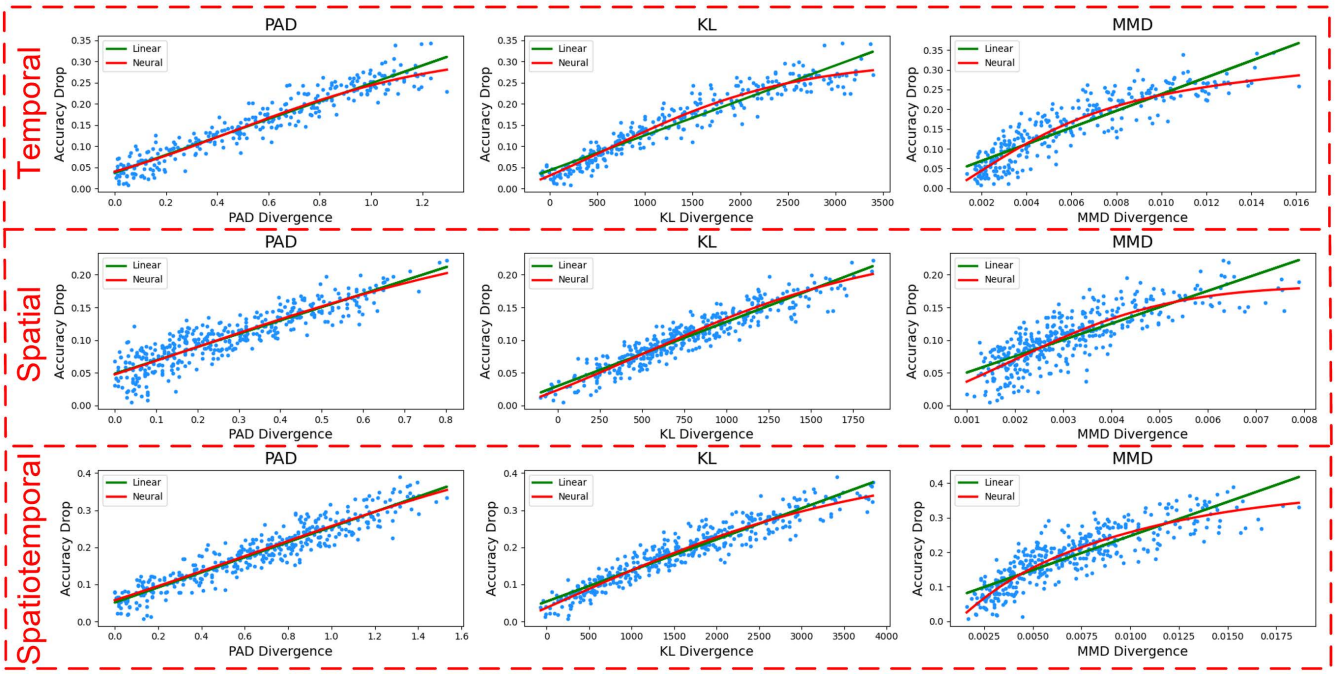


FIGURE 5. Relation between actual detection accuracy drop and divergence measured by selected metrics in temporal, spatial, and spatiotemporal experiments. Each dot corresponds to a pair of source dataset and pseudo target dataset. We also plot the linear regression line and neural network regression line in green and red.

bus when launching the attack. By conducting training and testing on 15×15 pairs experiment, the non-transfer methods perform worst when target buses are 14, 16, and 19. Thus, we select 15 buses as the source domain datasets separately and Buses 14, 16, 19 as target domain datasets.

3) SPATIOTEMPORAL CASES

For spatiotemporal cases, we consider attacks happen at different time and locations. We assume that the time and locations of attack in the target domain vary from the source domain. To this end, we select 4 hours (''valley'') as the source load demand pattern and another 4 hours (''Peak'') as the target load demand pattern. And we inject different buses for source domain datasets and target domain datasets.

D. CLASSIFIER ARCHITECTURE

We use the DANN [35] as our benchmark TL model, which aims to learn domain-invariant features by maximizing the domain discriminator loss and minimizing the label predictor loss. Zhang and Yan [6] propose a DANN-based framework in the smart grid and show their frame is sufficiently powerful to perform well on intrusion detection in the smart grid. For the basic classification model used to calculate the classification accuracy drop in the transferability analysis, we extract the Feature Extractor and the Label Predictor from DANN and combine them into a Multi-layer Perceptron (MLP) [36], which contains 5 layers and 592 neurons in total.

We train the MLP on the source domain and test it on the target domain to acquire the classification accuracy drop. A threshold Π is set to indicate whether the data distribution divergence could have a significantly negative effect on the trained ML model. Considering FDI is a severe threat, we use 10% of accuracy drop as the threshold in triggering TL in experiments.

To evaluate TL performance after identifying the tasks, we compare the detection accuracy of DANN and a non-transfer ML method. Since MLP has demonstrated superior accuracy and computation efficiency in intrusion detection [37], we choose MLP as the non-transfer ML method and follow the same configuration as that in transferability analysis. All classifiers are implemented in Scikit-learn and Keras with manually optimized parameters releasable upon request.

V. RESULTS AND DISCUSSIONS

We first evaluate the performance of the selected metrics in predicting accuracy drop.

A. EVALUATION OF TRANSFERABILITY ANALYSIS

Fig. 5 shows the relation between actual detection accuracy drop (y-axis) and the divergence (x-axis) measured by selected metrics in temporal, spatial, and spatiotemporal scenarios. Note that the pseudo target domain datasets in the temporal scenario are attack-free as illustrated in Table 1, while the pseudo target domain datasets in the spatial and spatiotemporal scenarios contain attack data.

TABLE 2. Error of accuracy drop prediction in the target domain.

Scenarios	Metrics	Linear Regression		NN Regression*	
		RMSE	MaxAE	RMSE	MaxAE
Temporal	PAD	2.38	8.62	2.34	7.72
	KL	2.52	7.88	2.22	7.43
	MMD	3.65	10.94	3.26	9.14
Spatial	PAD	1.88	6.35	1.87	6.32
	KL	1.43	4.48	1.41	4.25
	MMD	2.45	7.51	2.36	7.79
Spatiotemporal	PAD	2.77	8.28	2.75	8.74
	KL	2.60	7.46	2.46	7.53
	MMD	4.20	12.74	3.84	12.72

* The NN Regression refers to the neural network regression.

1) COMPARISON BETWEEN THREE SCENARIOS

We can find a strong positive relation between accuracy drop and distribution divergence with a high Pearson correlation coefficient: ρ is above 0.83 in all experiments. This observation also indicates that it is feasible to predict the accuracy drop by distribution divergence. Among the three scenarios, spatial cases have the lowest divergence and accuracy drop. This is because the normal data of source and target domains are from the same load demand pattern and share similar distributions. Meanwhile, spatiotemporal cases have the most significant divergence and accuracy drop, since the source domain and target domain vary in both temporal and spatial variables.

2) COMPARISON OF DIVERGENCE METRICS AND REGRESSION MODELS

To show the relation between accuracy drop and divergence measured by the selected metrics in the target domain, following Elsahar and Galle [11], we first make a comparison in predicting classification accuracy between the selected metrics and baseline. The baseline directly measures the mean of actual accuracy drop in each scenario and takes the mean as its prediction. Table 2 shows the root mean squared error (RMSE) and maximum absolute error (MaxAE) of different metrics with the two regression models. The baseline RMSE, i.e., the standard deviation of the actual accuracy drop, are 7.09%, 4.18%, 8.19% in temporal, spatial, and spatiotemporal scenarios, respectively. The baseline MaxAE are 18.91%, 13.87%, 20.28% in each scenario.

Overall, all our selected metrics improve significantly over the baseline in both RMSE and MaxAE. For instance, in temporal cases, PAD and KL with either linear regression or neural network regression both decrease RMSE to below 2.38% and MaxAE to under 8.62%. MMD performs slightly worse than the first two metrics but still achieves high performance compared to the baseline. MMD has an RMSE of 3.65% and MaxAE of 10.94% with linear regression, and an RMSE of 3.26% and MaxAE of 9.14% with neural network regression. Among three metrics, PAD and KL have comparable performance and show robust prediction power in all three scenarios. In addition, PAD and KL are more accurate than MMD in RMSE and MaxAE.

TABLE 3. Comparison of DANN and MLP against returning attacks at different hours.

Cases	Source Hours	Source Accuracy	Target Hours	Target Accuracy	Predicted Drop	Actual Drop	DANN Accuracy	DANN Improvement
1	2-5 (Valley)	95.72	1-4	86.54	8.65	9.18	95.56	+9.02
			5-8	85.25	12.40	10.47	93.81	+8.57
			9-12	76.97	16.44	18.75	87.82	+10.85
			13-16	66.01	33.07	29.71	90.29	+24.29
			17-20	71.81	25.15	23.91	92.96	+21.15
			21-24	83.90	10.51	11.82	95.33	+11.43
2	11-14 (Ascending)	96.15	1-4	95.30	2.40	0.85	94.81	**-0.48**
			5-8	90.98	4.60	5.17	94.88	+3.90
			9-12	79.77	12.98	16.38	93.73	+13.96
			13-16	76.75	21.67	19.40	88.93	+12.17
			17-20	73.01	26.49	23.14	89.82	+16.81
			21-24	84.42	13.22	11.73	91.16	**+6.74**
3	17-20 (Peak)	95.93	1-4	80.51	11.85	15.42	94.65	+14.14
			5-8	86.40	8.74	9.53	95.46	+9.06
			9-12	80.39	15.84	15.54	87.34	+6.95
			13-16	67.69	29.70	28.24	93.29	**+25.60**
			17-20	74.58	19.33	21.35	89.90	+15.32
			21-24	84.10	8.03	11.83	93.97	+9.87
4	21-24 (Descending)	97.18	1-4	93.63	5.60	3.55	94.63	+1.00
			5-8	91.08	5.92	6.10	92.44	+1.36
			9-12	79.58	19.46	17.60	95.40	+15.82
			13-16	78.38	18.00	18.80	88.47	+10.10
			17-20	70.96	27.40	26.22	95.75	+24.79
			21-24	88.17	9.54	9.01	95.42	+7.24

The underscored hours refer to the cases where the predicted accuracy drop is lower than the threshold Π, so no TL is required.

Neural network regression has a slightly smaller RMSE than linear regression in all scenarios, but their general performance is close. Overall, the RMSE of two regression models with any metrics is lower than 4.20% in all cases, implying the predicted accuracy drop is close to the ground accuracy drop. In addition to using attack-free pseudo target datasets, we also evaluate extra cases where there might be attacks in the pseudo target dataset and it is shown that the regression based method still remained robust. This indicates a strong relation between accuracy drop and divergence, and we can use this relation to predict accuracy drop.

B. FDI DETECTION ACCURACY

Based on the above observation, we can measure the divergence and leverage the regression relation to predict the accuracy drop of an unlabeled target domain dataset, and determine whether to trigger TL accordingly.

1) FDI DETECTION ON TEMPORAL SCENARIO

The detection accuracy of DANN and MLP is illustrated in Table 3. We underscore the target domain hours where the predicted accuracy drop is below the threshold of 10% and TL is not required. We find that all the underscored target hours are during Hours 1-4, 5-8, and 21-24. The reason is the load demand of the aforementioned hours on Day 6 is overlapped with most of the source domain load demand on Day 1 and 2. The overlapping implies less distribution divergence, so the MLP has a smaller accuracy drop than other hours.

Compared with the TL-required cases, DANN gives less improvement in the underscored cases. Furthermore, DANN may not provide improvement when divergence is relatively small. For instance, during target hours 1-4 in Case 2, the predicted accuracy drop is only 2.40%, and there is no need for TL. If we apply TL, DANN degrades the accuracy by 0.48% compared to MLP. In all TL-required cases, however, DANN shows better performance over MLP. The greatest and lowest improvements are +25.60% during Hours 13-16 in Case 3 and +6.74% during target hours 21-24 in Case 2. Overall, the results suggest that DANN can retain the FDI detection accuracy while the non-transfer classifier fails to adapt when there is a significant data distribution divergence caused by variations in load demand.

2) FDI DETECTION ON SPATIAL SCENARIO

The classification accuracy of MLP and DANN on spatial cases is reported in Fig. 6, and the white triangle shows the average accuracy of each method. For the spatial cases, since the data distributions of normal data are similar between source and target domains, the non-transfer method is able to identify most of the attack samples and achieve higher accuracy than temporal cases. However, DANN still demonstrates improvements in all cases. Comparing DANN and MLP, it can be found that the greatest and lowest average improvements reach 7.91% and 5.83% when the target buses are 19 and 16. The results show that DANN can achieve high detection accuracy against variations in attack locations.

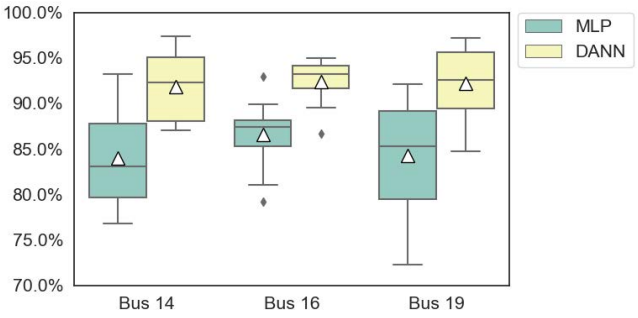


FIGURE 6. Box plots of accuracy with source attack launched on 15 attackable buses and target attack on Buses 14, 16, 19.

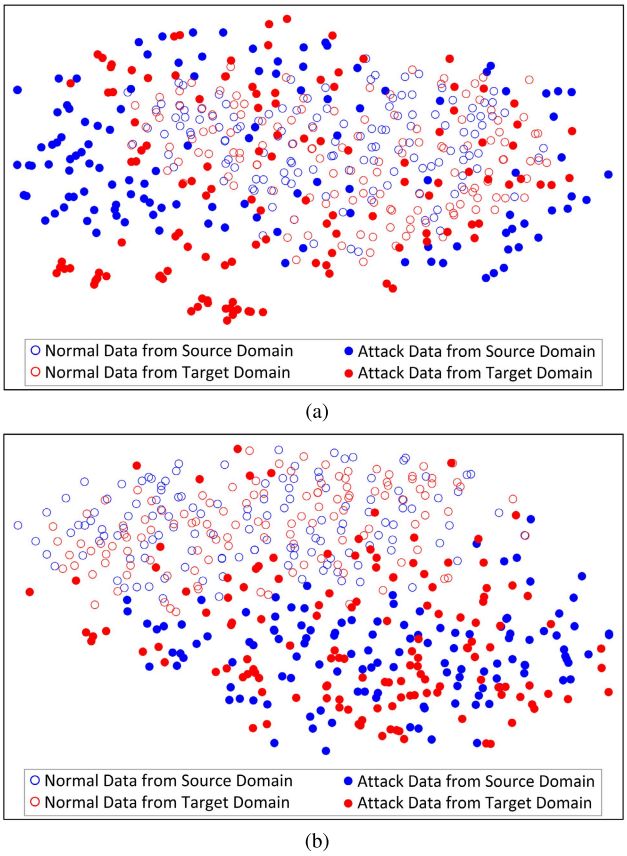


FIGURE 7. The t-SNE visualization of DANN effectiveness: distribution of extracted features (a) without transfer; and (b) with transfer.

We also use t-SNE to visualize the distribution of extracted features without and with DANN in Fig. 7. In both subfigures, the normal data in the source and target domains have a close distribution since they are from similar load demand. Without DANN, attack data from the source domain and target domain are not mixed because they target different locations. After applying DANN, however, the attack data of two domains are mixed well. Overall, the distribution of the two datasets becomes similar after the domain adversarial training. Moreover, normal data are distributed on the upper left, and attack data are distributed on the lower right, making it easier for DANN to distinguish them.

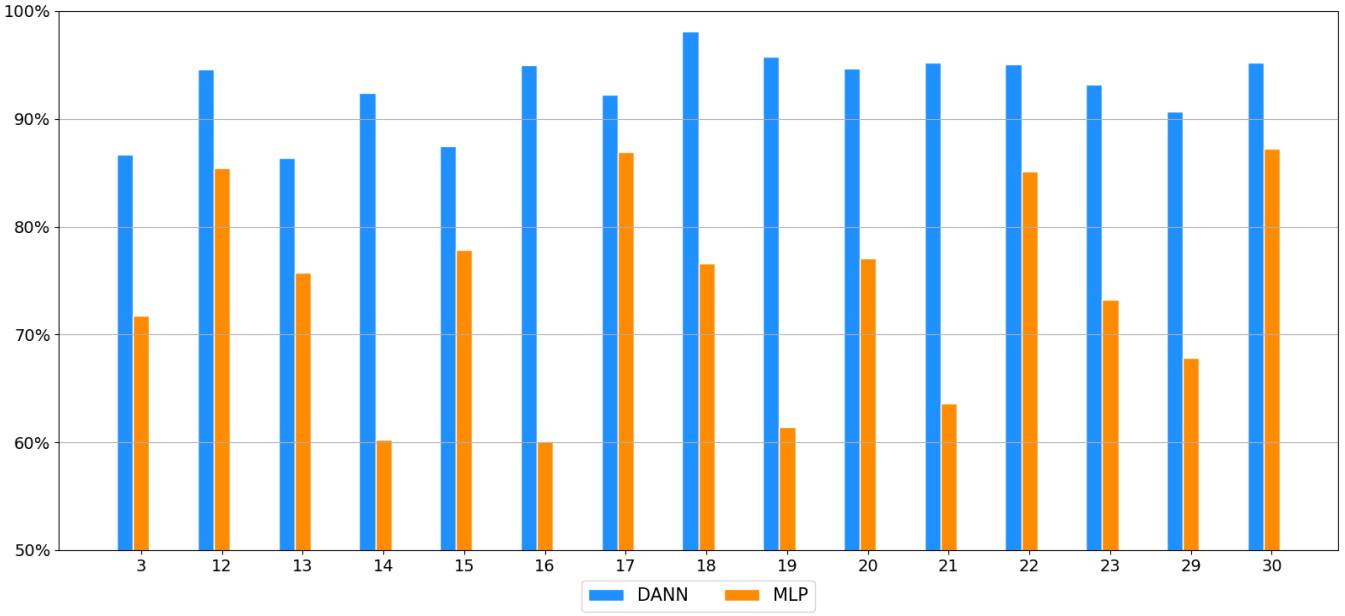


FIGURE 8. Spatiotemporal cases accuracy with one bus as target domain and other 14 buses as source domain separately.

3) FDI DETECTION ON SPATIOTEMPORAL SCENARIO

We further test out the classifiers on spatiotemporal cases and demonstrate the average accuracy of each method in Fig. 8 with one bus as the target domain and the other 14 buses as the source domain separately. For the spatiotemporal cases, the data distribution divergence on spatial and temporal dimensions further degrades the baseline methods, but DANN still demonstrates a significant improvement. The average improvement is 18.84% compared to MLP. The greatest and lowest average improvements reach 34.37% when the target bus is 19 and 5.34% when the target bus is 17. The results suggest that, once timely triggered, DANN can enable robust performance in detecting FDI attacks with spatial and temporal variants.

VI. CONCLUSION

This paper studies the problem of when one should apply TL for intrusion detection in the smart grid. We propose a divergence-based transferability analysis to justify the necessity of TL. First, we leverage three metrics of different properties to evaluate the distribution divergence and use two regression models to approximate the relation between accuracy drop and divergence. The result shows that the selected metrics are capable of predicting the accuracy drop of a trained model on an unseen dataset with distribution divergence. Furthermore, we consider attacks may happen at different times, locations, and both in dynamic cyber-physical systems, and adopt domain adaptive training as our TL model on realistic datasets. The TL results show that DANN can retain high detection accuracy against temporal, spatial, and spatial-temporal divergence, implying our approach is promising for future applications in real-world systems.

There are more sophisticated scenarios for the FDI and other attacks in the studies, e.g., coordinated cyber-physical attacks (CCPAs) [38] and coordinated topology attacks [39], among others. In the future, we will study more advanced coordinated attack scenarios to verify the performance and improve the understanding of transferability analysis.

REFERENCES

[1] C. Tu, X. He, X. Liu, and P. Li, "Cyber-attacks in PMU-based power network and countermeasures," *IEEE Access*, vol. 6, pp. 65594–65603, 2018.

[2] L. F. F. De Almeida, J. R. D. Santos, L. A. M. Pereira, A. C. Sodre, L. L. Mendes, J. J. P. C. Rodrigues, R. A. L. Rabelo, and A. M. Alberti, "Control networks and smart grid teleprotection: Key aspects, technologies, protocols, and case-studies," *IEEE Access*, vol. 8, pp. 174049–174079, 2020.

[3] H. He and J. Yan, "Cyber-physical attacks and defences in the smart grid: A survey," *IET Cyber Phys. Syst., Theory Appl.*, vol. 1, no. 1, pp. 13–27, Dec. 2016.

[4] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: A review," *IEEE Access*, vol. 7, pp. 13960–13988, 2019.

[5] P. I. Radoglou-Grammatikis and P. G. Sarigiannidis, "Securing the smart grid: A comprehensive compilation of intrusion detection and prevention systems," *IEEE Access*, vol. 7, pp. 46595–46620, 2019.

[6] Y. Zhang and J. Yan, "Semi-supervised domain-adversarial training for intrusion detection against false data injection in the smart grid," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[7] J. Q. Nonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Theoretical Views on Dataset and Covariate Shift*. Cambridge, MA, USA: MIT Press, 2009, pp. 39–43.

[8] Y. Abdulazeem, H. M. Balaha, W. M. Bahgat, and M. Badawy, "Human action recognition based on transfer learning approach," *IEEE Access*, vol. 9, pp. 82058–82069, 2021.

[9] Z. Taghiyarrenani, A. Fanian, E. Mahdavi, A. Mirzaei, and H. Farsi, "Transfer learning based intrusion detection," in *Proc. 8th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2018, pp. 92–97.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 1–9.

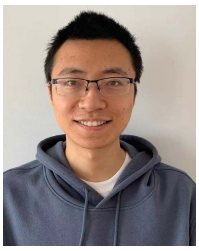
[11] H. Elsahar and M. Gallé, "To annotate or not? Predicting performance drop under domain shift," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019, pp. 2163–2173.
[12] W. Deng and L. Zheng, "Are labels always necessary for classifier accuracy evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 1, pp. 15069–15078, Dec. 2020.
[13] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 97–105.
[14] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4068–4076.
[15] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1–9.
[16] S. Gupta, S. Waghmare, F. Kazi, S. Wagh, and N. Singh, "Blackout risk analysis in smart grid WAMPAC system using KL divergence approach," in *Proc. IEEE 6th Int. Conf. Power Syst. (ICPS)*, Mar. 2016, pp. 1–6.
[17] M. Tajdinian and H. Samet, "Divergence distance based index for discriminating inrush and internal fault currents in power transformers," *IEEE Trans. Ind. Electron.*, vol. 69, no. 5, pp. 5287–5294, May 2022.
[18] G. Chaojun, P. Jirutitijaroen, and M. Motani, "Detecting false data injection attacks in AC state estimation," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2476–2483, Sep. 2015.
[19] S. Pal, B. Sikdar, and J. Chow, "Detecting data integrity attacks on SCADA systems using limited PMUs," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Nov. 2016, pp. 545–550.
[20] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Jun. 2004, pp. 31–40.
[21] A. R. Kashyap, D. Hazarika, M.-Y. Kan, and R. Zimmermann, "Domain divergences: A survey and empirical analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 1830–1849, doi: 10.18653/v1/2021.naacl-main.147.
[22] B. Sun, J. Feng, and K. Saenko, *Correlation Alignment for Unsupervised Domain Adaptation*. Cham, Switzerland: Springer, 2017, pp. 153–171.
[23] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.
[24] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 317–320.
[25] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. 49–57, Jul. 2006.
[26] H. Akoglu, "User's guide to correlation coefficients," *Turkish J. Emergency Med.*, vol. 18, no. 3, pp. 91–93, Sep. 2018.
[27] (2019). *ISO New England—Energy, Load, and Demand Reports*. [Online]. Available: https://www.iso-ne.com/isoexpress/web/reports/load-and-demand
[28] I. C. for a Smarter Electric Grid (ICSEG). *IEEE 30-Bus System*. [Online]. Available: https://icseg.iti.illinois.edu/ieee-30-bus-system/
[29] M. H. Hassan, S. Kamel, M. A. El-Dabah, T. Khurshaid, and J. L. Dominguez-Garcia, "Optimal reactive power dispatch with time-varying demand and renewable energy uncertainty using Rao-3 algorithm," *IEEE Access*, vol. 9, pp. 23264–23283, 2021.
[30] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inform. Syst. Secur.*, vol. 14, no. 1, p. 13, Jun. 2011.
[31] R. Deng, G. Xiao, R. Lu, H. Liang, and A. V. Vasilakos, "False data injection on state estimation in power systems—Attacks, impacts, and defense: A survey," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 411–423, Apr. 2017.
[32] A. S. Musleh, G. Chen, and Z. Y. Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2218–2234, May 2020.
[33] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 Ukraine blackout: Implications for false data injection attacks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3317–3318, Jul. 2016.
[34] M. Rahman, Y. Li, and J. Yan, "Multi-objective evolutionary optimization for worst-case analysis of false data injection attacks in the smart grid," *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2020, pp. 1–8.
[35] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Lavi-olette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Jan. 2016.
[36] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.
[37] S. Park and H. Park, "Ann based intrusion detection model," *Proc. Workshops Int. Conf. Adv. Inf. Netw. Appl.* Matsue, Japan: Springer, 2019, pp. 433–437.
[38] R. Deng, P. Zhuang, and H. Liang, "CCPA: Coordinated cyber-physical attacks and countermeasures in smart grid," *IEEE Trans. Smart Grid.*, vol. 8, no. 5, pp. 2420–2430, Sep. 2017.
[39] S. Liu, B. Chen, T. Zourntos, D. Kundur, and K. Butler-Purry, "A coordinated multi-switch attack for cascading failures in smart grid," *IEEE Trans. Smart Grid*, vol. 5, no. 3, pp. 1183–1195, May 2014.

PENGYI LIAO (Graduate Student Member, IEEE) received the B.S. degree in electrical and electronic engineering from the Huazhong University of Science and Technology, China, in 2017, and the M.Sc. degree in electrical engineering from Wuhan University, China, in 2021. He is currently pursuing the M.Sc. degree in electrical and computer engineering with Concordia University, Canada. His current research interests include machine learning, transfer learning, smart grid, and cyber-physical systems security.

JUN YAN (Member, IEEE) received the B.E. degree from Zhejiang University, Hangzhou, China, in 2011, and the M.Sc. and Ph.D. degrees (with Excellence in Doctoral Research) in electrical engineering from the University of Rhode Island, Kingston, RI, USA, in 2013 and 2017, respectively.

He joined the Concordia Institute for Information Systems Engineering, Concordia University, Montréal, QC, Canada, in December 2017, where he is currently an Associate Professor (with early tenure) and a Founding Member of the Security Research Center (SRC) and the Applied Artificial Intelligence Institute (AI)^2. His research interests include computational intelligence and cyber-physical security, with applications in smart grids, smart cities, and other safety-critical smart infrastructures. He was a recipient of the Best Paper Award of IEEE ICC 2014, the Best Student Paper Award of IEEE WCCI 2016, and the Best Readings of IEEE ComSoc (2013), among others.

JEAN MICHEL SELLIER studied mathematical physics at the University of Catania, Italy, and gained experience during his various postdoctoral positions at Imperial College London (U.K.) and at INRIA (France). He has also been a Research Associate and an Assistant Professor with Purdue University, IN, USA, and an Associate Professor with the Bulgarian Academy of Sciences. He is currently a Senior Data Scientist with Global AI Accelerator (GAIA), Ericsson, Canada.



YONGXUAN ZHANG received the B.S. degree from the Department of Information Engineering, Xi'an Jiaotong University, China, in June 2016, and the M.Sc. degree in computer science from Concordia University, in 2020, where he contributed to this study. During his master's degree, his research interests include machine learning, transfer learning, and smart grid security.