Domain-Adversarial Transfer Learning for Robust Intrusion Detection in the Smart Grid

Yongxuan Zhang and Jun Yan

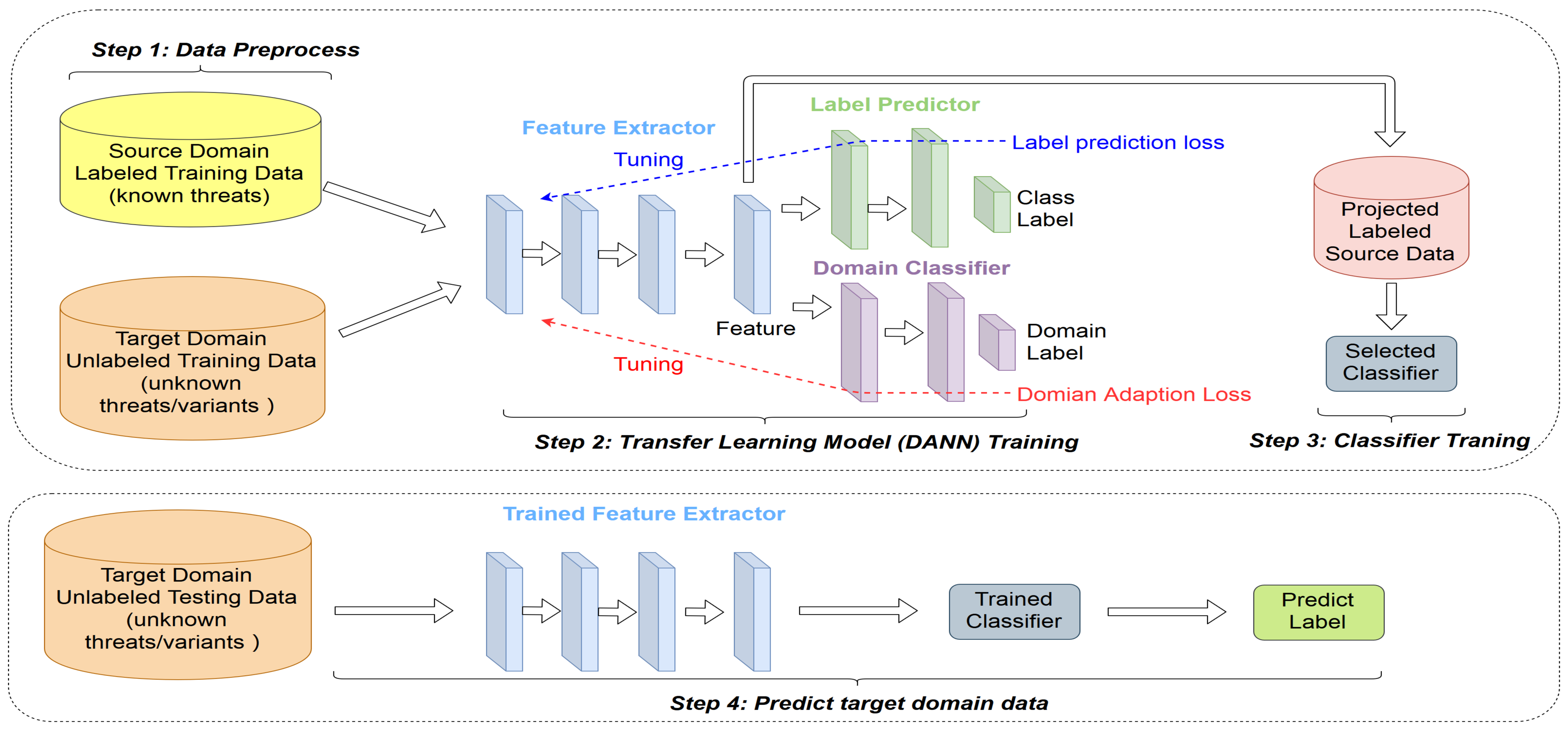
Concordia University, Montréal, Québec, Canada

Cyber-Physical Security and Intrusion Detection for the Smart Grid

- Cyber-physical smart grid mainly consists of generation, transmission, distribution systems communicating over a vast network in the cyberspace.
- The smart grid faces growing cyber-physical attack threats aimed at the critical systems and processes communicating over the complex cyber-infrastructure.
- Machine learning (ML)-based detection and classification have been increasingly effective and adopted against sophisticated attacks [1]-[3].
- Limitation of ML-based methods: Classic machine learning methods may not perform well once the data distribution has changed after training.

Domain Adversarial Neural Network for Robust Intrusion Detection

- Domain Adversarial Neural Network [4] is one of the state-of-the-art transfer learning methods, which leverages the adversarial training to help neural network find a mapping of data from two domains into same space with similar distribution. Then we can apply the trained source domain classifier on target domain.
- Overview of the DANN based framework:



- Objective Function:

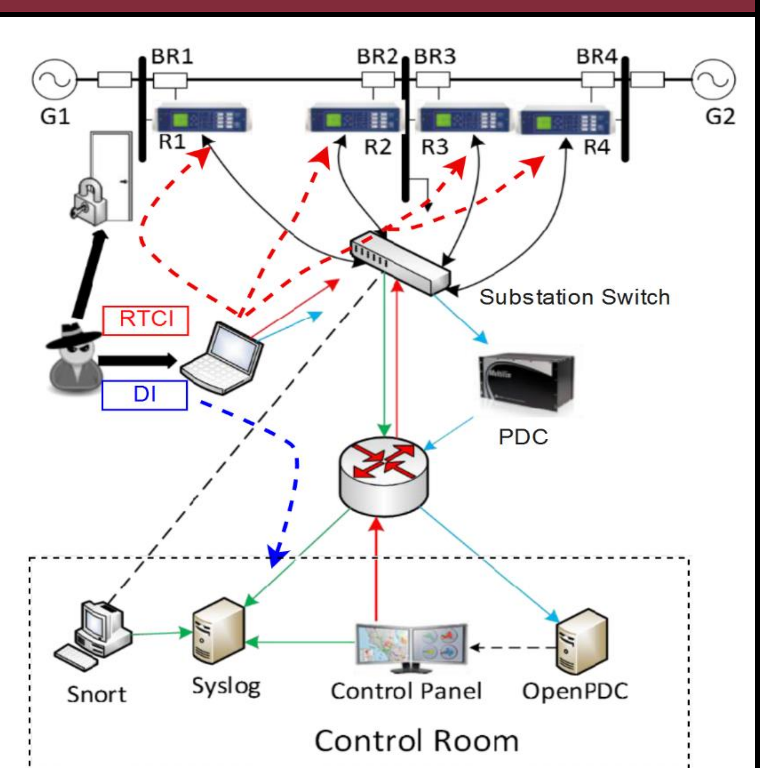
$$\min_{\mathbf{W},\mathbf{b},\mathbf{V},\mathbf{c}} \left[\frac{1}{n_S} \sum_{\mathbf{x} \in \mathcal{D}_S} L_y(\mathbf{x}, y) + \lambda \cdot R(\mathbf{W}, \mathbf{b}) \right]$$

Combining the losses of event misclassification and domain separation.

Dataset

- The Dataset is from a hardware-in-the-loop testbed by University of Alabama in Huntsville and the Oak Ridge National Lab (ORNL) [5], [6].

Classes	Scenarios	Descriptions
Normal	No Events	Normal operation with load variation
Attack	Data Injection (DI)	Attacker manipulates current, voltage, etc. to mislead controllers and/or operators into mal-operations.
	Remote Tripping Command Injection (RTCI)	Attacker sends a command to a relay and open a circuit breaker, directly causing a line outage.



Experiments Description

- We create several cases where there is unseen attack in testing set or same attack with different locations.

Cases	Threat in Source Domain	Threat(s) in Target Domain	Transfer Between
1	DI	RTCI	Different types of attack (measurement to command)
2	DI	DI and RTCI	
3	RTCI	DI	Different types of attack (command to measurement)
4	RTCI	DI and RTCI	
5	RTCI-15 (RelayR1)	RTCI-16 (RelayR2)	Different location of same attack
6	RTCI-17 (Relay R3)	RTCI-18 (Relay R4)	

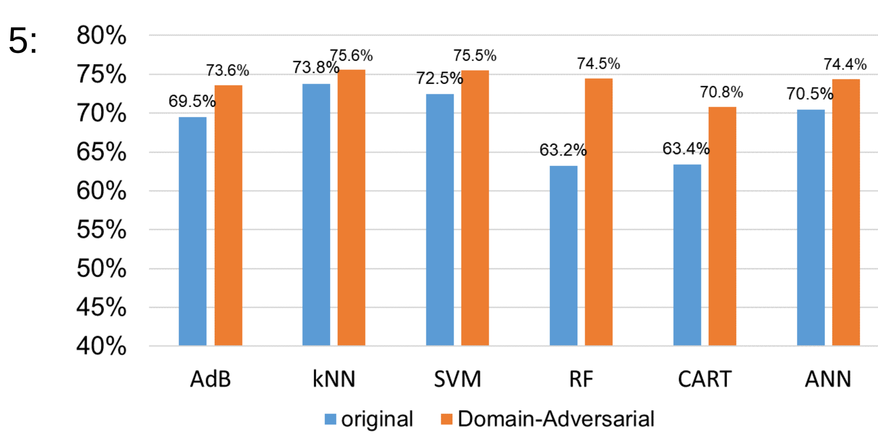
Classification Performance

- Transfer to new attacks:

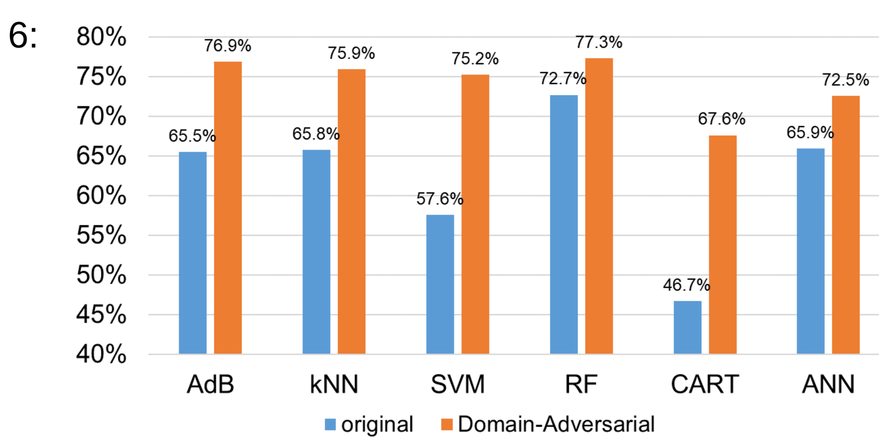
Cases	Methods	AdaBoost	kNN	SVM	Random Forest	CART	ANN
1	Original	72.0%	77.6%	57.9%	51.4%	53.2%	83.0%
	Domain-Adversarial	86.8%	86.0%	82.9%	88.2%	76.3%	84.5%
	Improvement	**+ 14.8%**	**+ 8.5%**	**+ 25.0%**	**+ 36.8%**	**+ 23.1%**	**+ 1.5%**
2	Original	77.3%	82.7%	71.0%	84.5%	76.7%	86.5%
	Domain-Adversarial	94.2%	90.4%	85.5%	95.2%	79.2%	87.8%
	Improvement	**+ 16.9%**	**+ 7.8%**	**+ 14.5%**	**+ 10.7%**	**+ 2.5%**	**+ 1.3%**
3	Original	73.0%	75.1%	64.8%	76.0%	61.3%	81.7%
	Domain-Adversarial	83.6%	82.2%	80.9%	84.5%	75.7%	83.6%
	Improvement	**+ 10.6%**	**+ 7.1%**	**+ 16.1%**	**+ 8.5%**	**+ 14.4%**	**+ 1.9%**
4	Original	71.2%	80.5%	66.7%	83.0%	69.5%	85.9%
	Domain-Adversarial	89.3%	88.2%	85.3%	90.0%	79.6%	87.6%
	Improvement	**+ 18.1%**	**+ 7.7%**	**+ 18.6%**	**+ 7.0%**	**+ 10.1%**	**+ 1.6%**

- Transfer to different locations:

Case 5:



Case 6:



Conclusions

- All baseline classifiers can benefit significantly from the domain-adversarial training and demonstrate robust performance against unseen types and different locations of threats.
- For future work we will extend the knowledge transfer ability among:
  - *Events*, e.g. normal operations, planned maintenance, system faults, extreme weather damage, and intentional attacks;
  - *Scenarios*, e.g., heterogeneous manufacturers, protocols, standards, topologies, and wired/ wireless configurations.
- References:

[1] H. He and J. Yan, “Cyber-physical attacks and defences in the smart grid: a survey,” *IET Cyber-Physical Systems: Theory & Applications*, vol. 1, no. 1, pp. 13–27, 2016.
[2] Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection, ” IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.
[3] M. Ozay, I. Esnaola, F. Yarman Vural, S. Kulkarni, and H. Poor, “Machine learning methods for attack detection in the smart grid,” IEEE Transactions on Neural Networks and Learning Systems, vol. 27,no. 8, pp. 1773–1786, Aug. 2016.
[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Lavi-olette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” Journal of Machine Learning Research, vol. 17,no.1, pp. 2096–2030, Jan. 2016.
[5] R. Borges Hink, J. Beaver, M. Buckner, T. Morris, U. Adhikari, and S. Pan, “Machine learning for power system disturbance and cyber-attack discrimination,” in 2014 7th International Symposium on Resilient Control Systems (ISRCS), Aug. 2014, pp. 1–8.
[6] Industrial control system (ICS) cyber attack datasets. https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets.

NSERC CRSNG
Fonds de recherche sur la nature et les technologies Québec
Mitacs
The work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de Recherche du Québec - Nature et Technologies (FRQNT) and Mitacs.
Yongxuan Zhang: z_yongxu@encs.concordia.ca Jun Yan: jun.yan@concordia.ca
Université Concordia University
GINA CODY SCHOOL OF ENGINEERING AND COMPUTER SCIENCE