SAND2022-0699C

Deep Learning Architecture for Processing Cyber-Physical Data in the Electric Grid

Daniel Calzada
Sandia National Laboratories
Albuquerque, NM USA
dacalza@sandia.gov

Shamina Hossain-McKenzie
Sandia National Laboratories
Albuquerque, NM USA
shossai@sandia.gov

Zeyu Mao
Texas A&M University
College Station, TX USA
zeyumao2@tamu.edu

***Abstract*—Due to the increasing complexity of energy systems and consequent increase in attack vectors, protecting the power grid from unknown disturbances and attacks using special protection schemes is crucial. In this paper, we discuss the machine learning component of the HARMONIE special protection scheme which relies on a novel combination of graph neural networks and Transformer models to jointly process cyber (network) and physical data. Our approach shows promise in detecting cyber and physical disturbances and includes the capability to identify relevant portions of the input sequence that contribute to the model's prediction. With this in place, the end goal of developing automated mitigation strategies is within reach.**

***Index Terms*—special protection scheme, energy system, cybersecurity, cyber-physical data, network, rationale graph neural network, transformer models, machine learning**

I. Introduction

Cyber attacks targeting grid operations are increasing in frequency and intensity, as exemplified by the 2015 and 2016 cyber attacks to the Ukrainian grid [1]. Furthermore, with the increasing penetration of distributed energy resources (DER) such as solar photovoltaic (PV) systems and wind farms, new "smart" technologies are being integrated and connected to the bulk power system. These grid-edge devices, with novel communication and automation functionalities, are also becoming targets to cyber attacks and can cause detrimental impact propagation as DER penetration increases [2].

Special protection schemes (SPSs), also known as remedial action schemes (RASs), prioritize reliability and seek to maintain stability, acceptable voltages, and loading limits during disturbance. Unlike typical protection schemes, SPSs can take actions beyond the isolation of a fault and include changes to demand, generation, and system configuration [3]. However, it no longer suffices for SPSs to focus solely on predefined disturbances and reliability [4]. Resilience and unpredictable disturbances such as extreme weather and cyber attacks must be considered.

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. This material is based upon work supported by the Sandia Laboratory Directed Research and Development Project # 222444.

An SPS that can adapt to unpredictable events (without predefined conditions) and effectively respond to limit or eliminate the disruption quickly is needed. Furthermore, a SPS that is cyber-physical in analyzing collected data and taking response actions is required; it is no longer sufficient for a SPS to process only physical power system data and solely take physical-side actions. Cyber-side actions are necessary to eliminate malicious compromise [5].

To develop the capabilities needed for future SPSs, our project team is developing a defensive, wide-area SPS that learns system conditions, mitigates cyber-physical consequences, and preserves grid operation under diverse predictable and unpredictable disturbances. This harmonized automatic relay mitigation of nefarious intentional events (HARMONIE)-SPS will meet the needs stated above by processing both cyber and physical data from both relays and out-of-band (OOB) measurements, learning actual system conditions to adapt to both predictable and unpredictable disturbances, and take preemptive steps to prevent further cascading impact [6].

However, a key challenge in developing the HARMONIE-SPS methodology is the machine learning approach for using cyber-physical data to classify system conditions and deploy corrective actions. This approach necessitates 1) the linking of cyber and physical learning models to identify cyber and/or physical disturbances, 2) the isolation of information from the input to determine a cause of a disturbance, and 3) continuous online learning to adapt to new system conditions. In this paper, we explore deep learning architectures capable of moving toward a solution to the first two requirements. We present candidate model architectures, preliminary experimental results, and a path forward to further this line of research.

II. Background

As can be observed by a plethora of recent literature, the need for cross-domain analyses for cyber-physical systems is becoming more and more prominent. Rai et al. discuss this trend in a paper that reviews different modeling approaches for cyber-physical systems [7]. They identified two main directions, 1) model-based, physics informed and 2) machine learning, as well as the growing interest in the combination of both in a hybrid modeling approach. In future iterations on this work, we intend to incorporate the underlying physics

model, but in this work we focus on leveraging machine learning on the model-less communication network data and the underlying structured physical data.

In the work by Fink et al., the authors examined the use of machine learning for power system disturbance and cyber attack discrimination [8]. Their analysis focused on synchrophasor data, including power system quantities such as voltage and current as well as the status of system devices (e.g., relays, switches, transformers). However, they did not consider the communication traffic and deep-packet analysis within the cyber system. Wang et al. also develop a machine learning approach for detecting power system disturbances, including cyber attacks, in their paper [9]; they also focus only on synchrophasor measurements.

III. Methodology

SPS efforts focus on automating triggering condition and corrective action parameters. HARMONIE-SPS augments these efforts and proposes novel real-time analysis of system conditions, using both cyber and physical data, to identify both triggering conditions and corrective actions. More details and discussion of the HARMONIE-SPS approach are given in [10]. For developing the machine learning component of HARMONIE-SPS, we will focus on a hybrid model-based and machine learning-based framework such that disturbances are not only classified correctly but also provide some insight to subsequently determine a suitable response for deployment.

One of our goals for HARMONIE-SPS is to jointly process both network data and physical data. This is a challenge given the different modalities of these disparate data types. Network traffic is typically modeled as a temporal graph with packets arriving at irregular intervals. Properties of network packets or flows are often discrete, such as port number and protocol. Physical data, however, is often expressed as a fixed-length vector of floating point values and is sampled at regular intervals.

To facilitate interleaving these two data streams, we elected to model the cyber-physical system as a graph (see Fig. 1), adding vertices for each node in the network and each synchrophasor and connecting these with edges that most closely mimic the cyber-physical system.

Using this data format, we can process data as either graphical, as messages being passed between vertices in a graph, or sequential, ordered by timestamp. For HARMONIE-SPS, a "timestep" would be either a message from a synchrophasor or a network packet or flow. In Section III-B we discuss the graphical processing of the data using a graph neural network (GNN) and in Section III-C we discuss processing the data as a stream using Transformer models. We can also chain these together, using the output of the GNN as input for the Transformer, creating a graphical-temporal deep learning model.

A. Data collection

To collect our data samples, we created various cyber and physical disturbances and contingencies on the Western Systems Coordinating Council (WSCC) 9-bus power system, consisting of 9 buses and 3 generators, with a representative communication network within a cyber-physical testbed [11]. The network diagram is outlined in Fig. 2 and the diagram of the physical system is given in Fig. 3. The disturbances were denial of service (DOS) attacks, false command injection (FCI) attacks, time delay (TD) attacks, and contingencies like single line-to-ground (SLG) faults.

Ultimately, we had 50 total scenarios, most of which contained cyber disturbances, physical disturbances, or both. Each scenario is roughly two minutes long with the disturbance (if present) happening at the one-minute mark. To allow our machine learning system to isolate rough temporal regions where disturbances happen, we treat each two minute capture as a training, validation, or test example and split it into overlapping 30-second time windows. (More details on splitting our dataset into folds are provided in Section IV.)

The models will be trained to identify whether or not a disturbance is present in a 30-second window of data. The advantage of splitting data like this is that it has low theoretical latency: The moment a disturbance happens, the machine learning model is capable of detecting it in the next data window it processes. The actual latency in a deployed system would depend primarily on system resources, configuration, and the latency of the underlying network. In our testbed environment, we could process an average of around one window per second, though we believe this can be optimized.

B. Graph neural network (GNN)

To capture and process the graphical component of the cyber-physical dataset, we employ a graph neural network (GNN) [12]. Simply put, we are using a graph neural network as a neural message passing algorithm: each vertex in the graph contains its own state vector and messages are passed along edges between neighbor nodes for a fixed number of iterations, updating the state vector for each vertex at each iteration as a function of its incoming messages. Since the edges (network flows or phasor readings) contain attributes, the messages being sent must be a function of the state vector of the source vertex and a vector representation of the edge attributes. At the end of this process, each vertex will have a state vector and each edge can also be represented as a function of the vector representation of the edge attributes and the state vectors of its two vertices.

C. Transformer model

To capture the sequential and temporal aspects of our cyber-physical data, we seek to employ a sequence processing architecture into our deep learning model. Traditionally, a recurrent neural network such as a Long Short Term Memory (LSTM) [13] or Gated Recurrent Unit (GRU) [14] would be employed for neural processing of sequential data. In recent years, however, the Transformer model [15] has been shown to yield superior performance on most tasks, especially in the natural language processing domain, which deals primarily

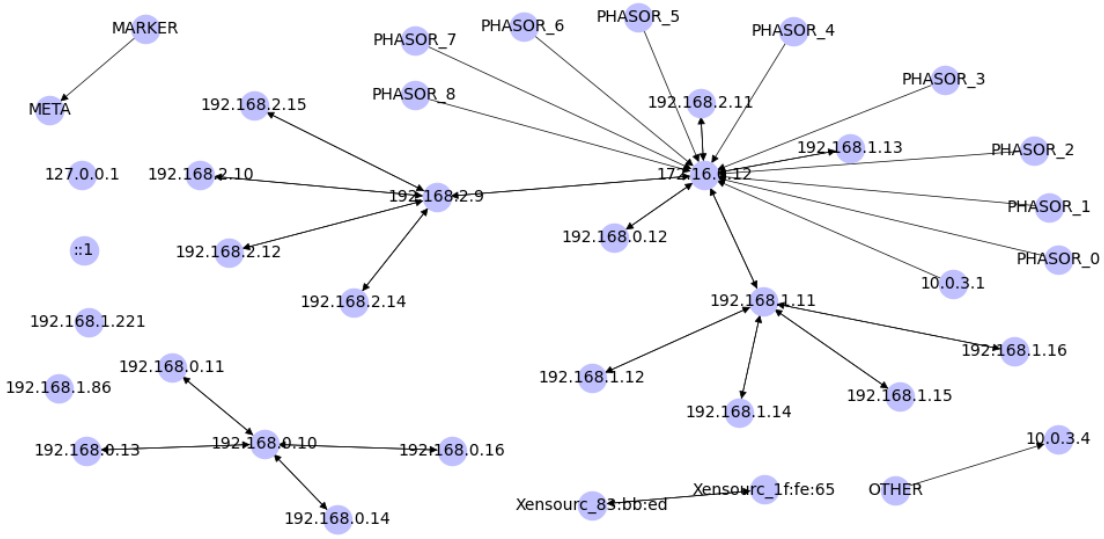

Fig. 1. The graph representation of the cyber-physical network as inputted to the neural network. Note the PHASOR_ nodes in the top center connecting the network and physical systems together. Some IPv6 nodes have been omitted for simplicity.

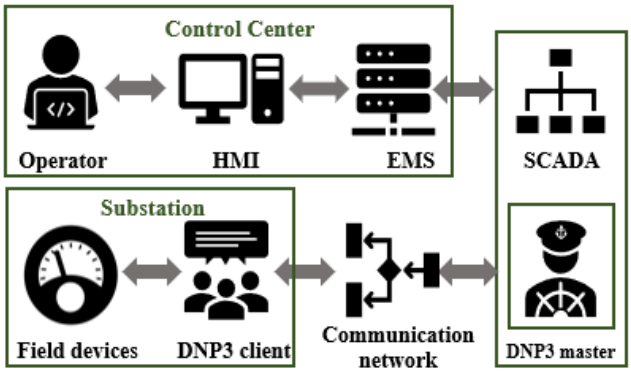

Fig. 2. The diagram of hierarchies for the simulated grid.

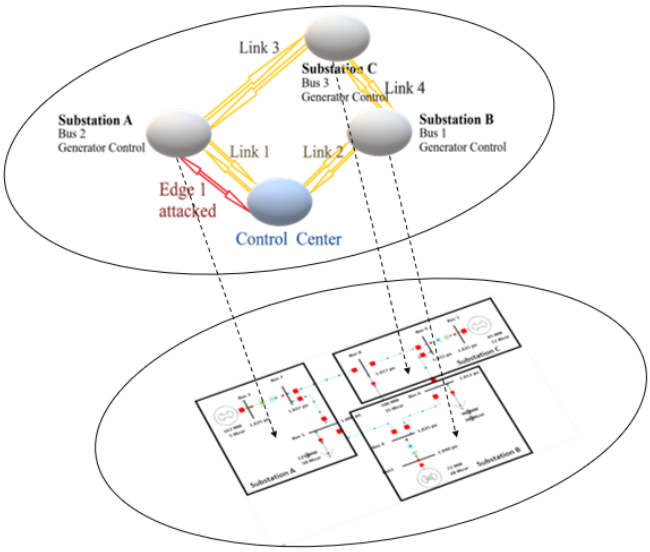

Fig. 3. The cyber-physical model for the WSCC 9-bus case.

with long sequences of input. The Transformer architecture has also been successfully applied to physical systems [16].

One limitation of Transformer models is their inability to efficiently process long sequences simultaneously. Due to the $\mathcal{O}(n^2)$ memory complexity of traditional Transformer models, modern hardware typically limits the size of a sequence to 512 timesteps. A naive solution to this would be to slice the sequence into 512-timestep windows and process each independently, but this loses the ability to model longer term dependencies, the very issue that Transformers were designed to address. To combat this, the Longformer [17] uses a combination of sliding window and global attention mechanisms to sparsify the attention matrix and the Reformer [18] uses Locality Sensitive Hashing [19] to more selectively compute attention scores.

Recently, the Big Bird Transformer architecture was proposed which used both a global and a randomized attention mechanism [20]. In our problem space, we expect each timestep to be related to temporally nearby timesteps. But, since the cyber and physical data streams are interleaved, related inputs may be separated by many timesteps. This would limit the usefulness of the sliding window component which only inspects a handful of adjacent inputs.

As such, we propose keeping only the randomization component by randomly assigning each timestep into one of N windows, each of a fixed size, and using full attention matrices within those windows. This allows information to spread across all regions of the sequence. The randomness innate with this approach would also serve as regulation for the complex neural network. A visual example of these attention weights is shown in Fig. 4. In this paradigm, each subwindow would be a sequence of 512 timesteps that has the potential to process

pieces of data across all 30 seconds of its sliding window.

D. Rationales

Given our stated goal of isolating inputs to identify a cause of a disturbance, we seek a solution to interpret our network's prediction to understand which timesteps (packets or physical data) are most relevant in making an assessment of a time window. Though deep learning models are powerful, they are notoriously difficult to interpret [21]. One advance in this area is Rationale Neural Networks [22]. In this architecture, the model is trained with a masking layer as the first layer of the model, and it learns to identify which timesteps are necessary for subsequent layers of the network and which can be masked or hidden. For each input, a mask probability is learned and timesteps are kept or removed by sampling using this probability. By regularizing the number of unmasked input timesteps, we encourage the model to present only the most relevant input features to the GNN and/or the Transformer. The rationales for our model's prediction, then, are simply the timesteps that remained unmasked.

The Rationale Neural Network in our case varied greatly across training runs, with some models masking around 95% of the edges to some masking almost none.

IV. Experimental Results

To test the efficacy of our method, we trained 20 versions of the model on various slices of data. Only 50 scenarios were available to us, so to make the most of this small dataset we report all our results using cross validation. We split our data into five random folds, each with ten scenarios. Since scenarios are further broken down into overlapping sliding windows, each window will not be independent from some others within the same scenario, so all sliding windows from the same scenario were placed into the same fold.

After this, we train independent models for each fold, withholding that ten-scenario test fold for evaluation. Within the four remaining training folds, we reserved one as the validation set for model selection, choosing the model which performed best on this validation fold. In summary, of the five folds, we assigned one as the test fold, one as the validation fold, and the remaining three as the training folds. In all of our experiments, one model was trained for each of these settings for a total of 20 distinct models.

To further reduce the high variance of our models incurred by training on such a small dataset, we elect to use bagging to combine multiple models into one. Specifically, we average the output of each of the four models trained on each test fold. The result is five aggregate models, one per test fold, each consisting of four models, one per validation fold. To reduce variance further, we ran each sliding window through the model four times, each time resampling which edges are kept in the Rationale Neural Network layer. The outputs predictions of all four runs are averaged to create the overall prediction for that sliding window. (Due to this innate randomness, these metrics are approximate and running the evaluation again on the same models would produce results differing by around 0.02.)

For each model architecture, we present Receiver Operator Curve (ROC) plots and Area Under the Curve (AUC) scores for detecting cyber disturbances and physical disturbances. Additionally, we include a confusion matrix for a decision threshold of 0.5 and its corresponding Matthew's Correlation Coefficient (MCC) scores. We also analyze the average percentage of edges that are selected by the Rationale Neural Network for propagation to the GNN and/or RNN. A low percentage indicates that the Rationale Neural Network significantly downsampled the edges being used. A summary of the experimental results is included in Table I. Detailed explanations and analyses of the experiments are provided in the subsequent subsections.

A. Traditional Transformer

To understand the effects of our random-windowed Transformer variant, we will first replace it with a sliding window Transformer operating on 512-timestep chunks, illustrated in the center of Fig. 4. This is, in essence, the same as the random-windowed Transformer without shuffling across timesteps. In Fig. 5, we present the confusion matrices and receiver operator curve for this experiment.

B. Random-windowed Transformer

In this experiment, we analyze the performance of our model using only the random-windowed Transformer described in Section III-C. In Fig. 6, we present the results for the random-windowed Transformer.

We see that the ROC and MCC scores for detecting cyber disturbances are better than those for physical disturbances. We suspect this is because many of the cyber disturbances are caused by single packets which are much simpler for a statistical model to identify. Physical disturbances, on the other hand, may require inferring patterns from long sequences of incoming data before arriving at a conclusion.

Surprisingly, we observe that the traditional non-randomized Transformer performs on par with the random-windowed Transformer. Due to this and the fact that the traditional approach to Transformers is more studied in literature, we intend to use a traditional Transformer as opposed to the random-windowed Transformer in future iterations of this project work.

C. GNN only

We will also examine the effects of removing the Transformer entirely and only using the graph neural network. The results of this experiment will help give a rough estimate of the relative contribution of the GNN. In Fig. 7, we present ROCs and confusion matrices, along with AUC and MCC scores for detecting cyber disturbances and physical disturbances.

As in Section IV-B, the ROC and MCC scores for detecting cyber disturbances are better than those for physical disturbances, but here the difference is even more pronounced. Because a vanilla GNN does not effectively process temporal

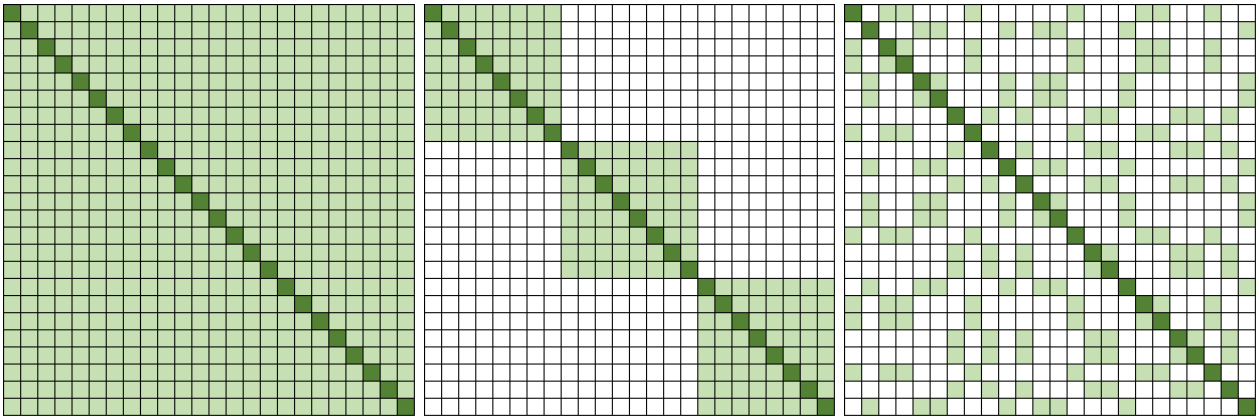

Fig. 4. Various Transformer attention mechanisms. The green squares are timesteps with nonzero attention weights. Left: Full attention matrices where memory usage is $\mathcal{O}(n^2)$. Center: Long sequence split into fixed-sized windows with constant memory usage but no long term dependency modeling. Right: Randomized attention windows with constant memory usage and information spreading out across the entire sequence.

TABLE I
EXPERIMENTAL RESULTS SUMMARIZED

Architecture	Rationale %	Cyber Disturbance Detection		Physical Disturbance Detection	
		MCC	AUC	MCC	AUC
Traditional Transformer	39.3%	0.77	**0.98**	0.57	0.85
Random-windowed Transformer	46.1%	0.70	0.95	**0.63**	**0.87**
GNN	48.0%	**0.85**	0.96	0.18	0.68
GNN + Transformer	N/A	0.74	0.97	0.30	0.77

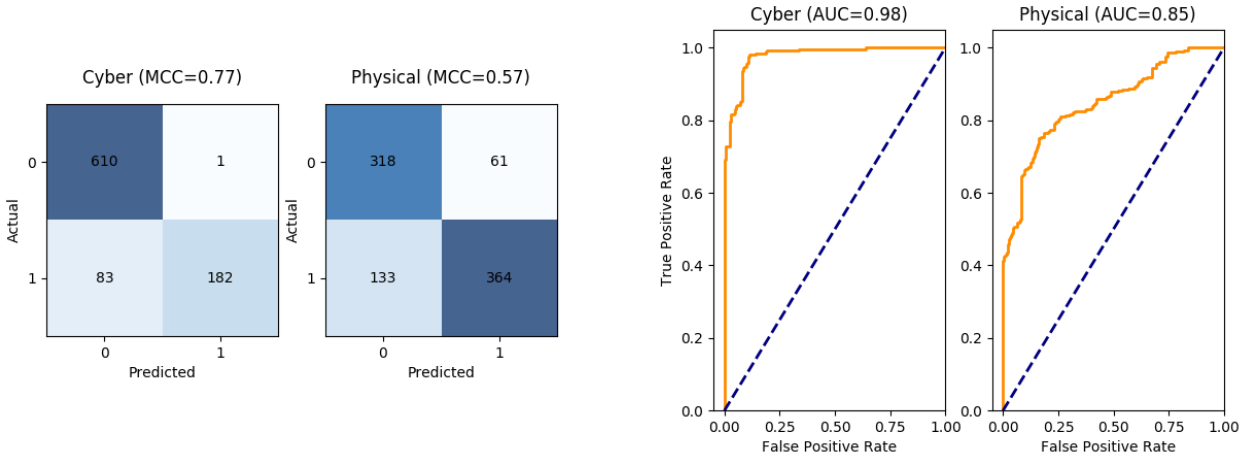

Fig. 5. The confusion matrices, MCC scores, receiver operator curves (ROCs), and AUC scores for the traditional Transformer model detecting cyber and physical disturbances.

data like sensor measurements, it is unsurprising that the GNN alone is unable to identify physical disturbances consistently. Our GNN includes an edge attribute for the timestamp of each measurement in addition to the measurement values themselves, but as seen in these results, that is insufficient for the graph neural network to make accurate assessments of the physical system.

D. GNN and Transformer

Finally, to quantify the advantage of using the GNN and random-windowed Transformer jointly, we train and evaluate models where the output edge vectors of the GNN are fed as input to the Transformer. As described in Section III, this architecture is expected to model both graphical and temporal aspects of our data and thus outperform the others. The results of this experiment are shown in Fig. 8.

Unsurprisingly, the GNN with the Transformer outperformed the GNN alone in terms of AUC in detecting a

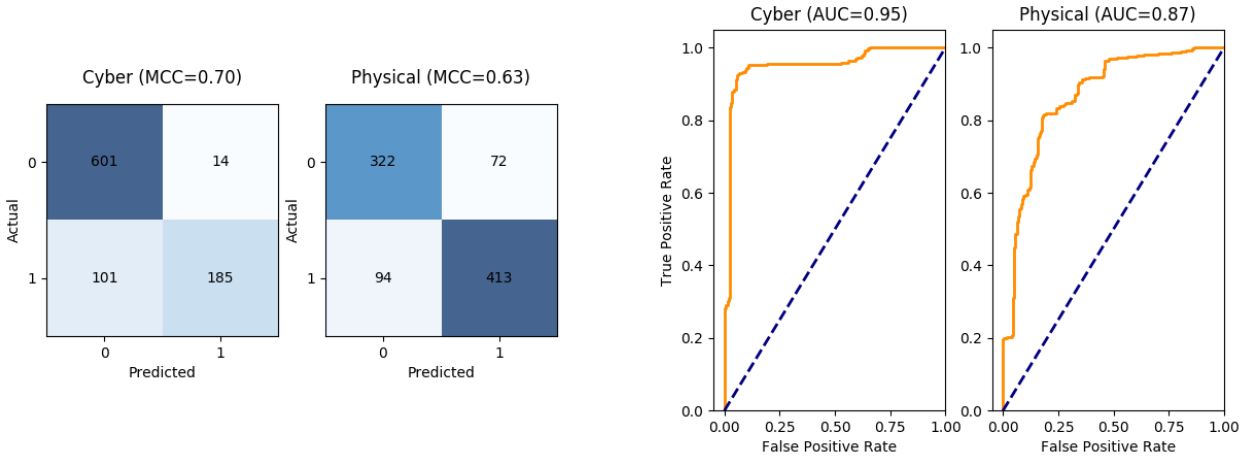

Fig. 6. The confusion matrices, MCC scores, receiver operator curves (ROCs), and AUC scores for the random-windowed Transformer model detecting cyber and physical disturbances.

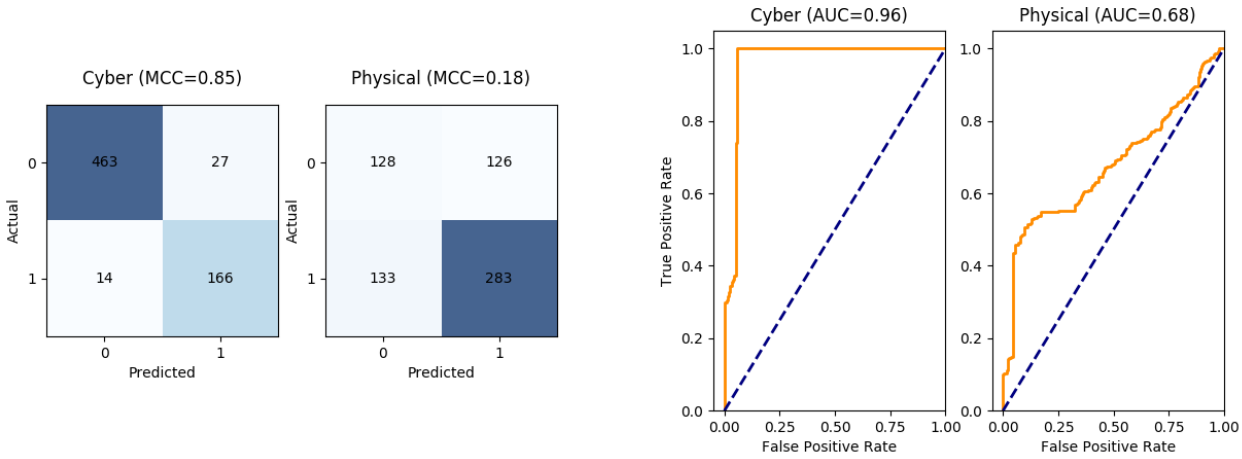

Fig. 7. The confusion matrices, MCC scores, receiver operator curves (ROCs), and AUC scores for the graph neural network model detecting cyber and physical disturbances.

physical disturbance, though the difference in detecting a cyber disturbance is negligible.

V. Conclusion

These experimental results provide a number of insights that will guide the HARMONIE-SPS machine learning framework and can guide other SPSs in the future.

First, we see that in our scenarios and contingencies, cyber disturbances are generally easier to detect than physical disturbances. As noted above, we suspect that modeling the long-term trajectory of the physical data is a more difficult task for our deep learning model than searching for a small number of malicious or problematic network packets. Along these lines, as expected, we observe that the GNN has the most difficulty identifying a physical disturbance.

Second, we see that using the GNN and Transformer together in this way does not yield the performance increase we expected. While the GNN and Transformers perform well on detecting cyber disturbances and physical disturbances respectively, the GNN + Transformer model underperforms the best model in each of those categories. This suggests that while the GNN and Transformer model each contribute valuable information to the process, there is room to improve the way we link them together into a cohesive model.

Third, we see that the Rationale Neural Network kept 40-50% of the edges (packets or phasor datapoints). Upon closer inspection, the edges most often kept are TCP packets, and all phasor measurement edges seem to have been assigned approximately the same probabilities. While removing 50-60% of edges is a good start, we see value in exploring techniques to reduce the number of edges retained and thus further isolate the source of the disturbance.

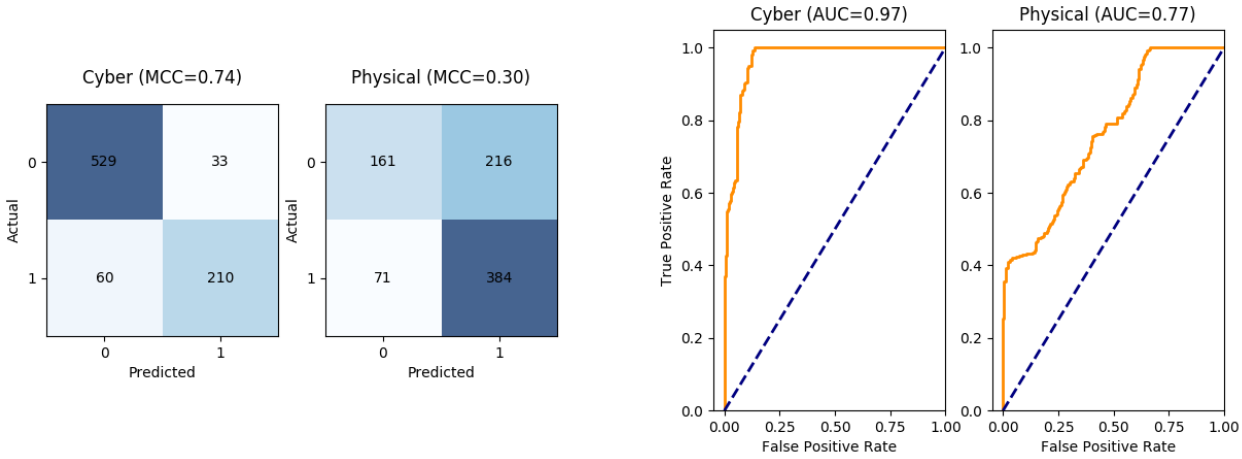

Fig. 8. The confusion matrices, MCC scores, receiver operator curves (ROCs), and AUC scores for the GNN and random-windowed Transformer operating in series to detect cyber and physical disturbances.

A. *Future work*

In this paper, we have shown a deep learning framework for analyzing the cyber and physical states of the WSCC 9-bus system and related communication network. Next steps include deploying this framework in the real-time HARMONIE-SPS testbed defined in [23]. On the machine learning side, future work on this effort may include scaling beyond the WSCC 9-bus system into a more complex power system. Since the model proposed in this framework is generally applicable to any cyber-physical system structured as a graph of information flows, we do not anticipate any structural changes necessary to the underlying model architecture. Instead, we anticipate requiring more training and validation data consisting of more instances of the same contingencies discussed in Section III-A.

Furthermore, we see potential for our model to identify and isolate more complex cyber disturbances. The cyber disturbances used in this paper are relatively simple and are able to be detected even without the full utilization of the graph neural network module. We propose developing more complex cyber contingencies to test the bounds of the graph neural network's capabilities. One such contingency could be a malicious insider progressively gaining access to various nodes on the network before deploying an attack.

Finally, there is value in a future rigorous exploration of the model architecture and hyperparameter space. In particular, this work assumes that the best way to combine the GNN with the Transformer is to link them in series, with the output of the GNN being fed to the input of the Transformer. It is not clear that this would outperform a parallel implementation with the GNN and Transformers each running in parallel on the original data and having their results concatenated together when making the final predictions.

ACKNOWLEDGMENT

The authors would like to thank Adam Summers, Nick Jacobs, Chris Goes, and Treya Houston from the Sandia National Laboratories team for their contributions on this and other aspects of the HARMONIE-SPS effort, as well as consultants to the project, Jason Stamp and Matthew Reno. In additon, the authors wish to thank the other members of the Texas A&M University team consisting of Leen Al Homoud and Komal Shetye, overseen by professor Katherine Davis, for their indispensable contributions to the project.

REFERENCES

[1] "Analysis of the cyber attack on the Ukrainian power grid," *Electricity Information Sharing and Analysis Center (E-ISAC)*, vol. 388, 2016.

[2] C. Lai, N. Jacobs, S. Hossain-McKenzie, P. Cordeiro, O. Onunkwo, J. Johnson, "Cyber Security Primer for DER Vendors, Aggregators, and Grid Operators," Sandia National Laboratories, Sandia Report SAND2017-13113, Dec. 2017.

[3] Western Electricity Coordinating Council (WECC), "Remedial Action Scheme Design Guide," 2011.

[4] H. Li, K. Shetye, S. Hossain-McKenzie, K. Davis, and T. Overbye, "Investigation of Automated Corrective Actions for Special Protection Schemes," SAND2020-9602, Sandia National Laboratories, Tech. Rep., 2020.

[5] Z. Mao, A. Sahu, P. Wlazlo, Y. Liu, A. Goulart, K. Davis, and T. J. Overbye, "Mitigating tcp congestion: A coordinated cyber and physical approach," in *2021 North American Power Symposium (NAPS)*, 2021, pp. 1–6.

[6] N. Jacobs, A. Summers, S. Hossain-McKenzie, D. Calzada, H. Li, Z. Mao, C. Goes, K. Davis, and K. Shetye, "Next-generation relay voting scheme design leveraging consensus algorithms," in *2021 IEEE Power and Energy Conference at Illinois (PECI)*, 2021, pp. 1–6.

[7] R. Rai and C. K. Sahu, "Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus," *IEEE Access*, vol. 8, pp. 71 050–71 073, 2020.

[8] R. C. Borges Hink, J. M. Beaver, M. A. 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)*, 2014, pp. 1–8.

[9] D. Wang, X. Wang, Y. Zhang, and L. Jin, "Detection of power grid disturbances and cyber-attacks based on machine learning," *Journal of Information Security and Applications*, vol. 46, pp. 42–52, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214212618305866

[10] S. Hossain-McKenzie, D. Calzada, N. Jacobs, C. Goes, A. Summers, K. Davis, H. Li, Z. Mao, T. Overbye, and K. Shetye, "Adaptive, cyber-physical special protection schemes to defend the electric grid against

predictable and unpredictable disturbances," in *2021 Resilience Week (RWS)*. IEEE, 2021, pp. 1–9.

[11] A. Sahu, P. Wlazlo, Z. Mao, H. Huang, A. Goulart, K. Davis, and S. Zonouz, "Design and evaluation of a cyber-physical testbed for improving attack resilience of power systems," *IET Cyber-Physical Systems: Theory & Applications*, vol. n/a, no. n/a. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cps2.12018

[12] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] N. Geneva and N. Zabaras, "Transformers for modeling physical systems," *arXiv preprint arXiv:2010.03957*, 2020.

[17] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[18] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.

[19] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.

[20] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences." in *NeurIPS*, 2020.

[21] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao *et al.*, "Interpretability of deep learning models: A survey of results," in *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE, 2017, pp. 1–6.

[22] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," *arXiv preprint arXiv:1606.04155*, 2016.

[23] A. Summers, C. Goes, D. Calzada, N. Jacobs, S. Hossain-McKenzie, and Z. Mao, "Towards cyber-physical special protection schemes: Design and development of a co-simulation testbed leveraging sceptre™," in *2022 Power and Energy Conference at Illinois (PECI)*. IEEE, 2022, pp. 1–7.