# Cross-Protocol Domain Adaptation and Anomaly Generalization in Smart Grid Intrusion Detection Systems Using Tabular Transformers

Shivam Amrut Kale (SYMSc-CS)[1], Priya Savaleram Borhade (SYMSc-CS)[2],
Siddhi Ravindra Warade (SYMSc-CS)[3]
[1,2,3] *School of Information Technology (SOIT), Indira University, Pune, Maharashtra, India*

***Abstract*—Contemporary smart-grid infrastructures utilize a diverse range of industrial control and communication protocols, including Modbus, DNP3, and IEC 60870-5-104. Machine-learning-based Intrusion Detection Systems (IDSs) can operate effectively when the distributions of their training and deployment data are closely aligned. However, performance may degrade when a model trained for one protocol is applied to another, as differences in telemetry distributions, feature semantics, packet formats, and reporting behaviors can arise. This study presents a domain-adaptive tabular Transformer framework designed for anomaly detection across smart-grid protocols. The framework initially tokenizes both heterogeneous numerical and categorical attributes and subsequently employs a Transformer encoder to model the contextual dependencies among telemetry features. By leveraging labeled source data alongside unlabeled target data through a domain-adversarial component equipped with a Gradient Reversal Layer (GRL), the model fosters latent representations that are less reliant on the specific communication domain. The source domain anomaly classifier is trained using binary cross-entropy, while the domain discriminator is tasked with distinguishing between source and target representations. During cross-protocol transfer, the proposed framework is evaluated against XGBoost, a multilayer perceptron, and a non-adaptive FT-Transformer. During evaluation on the target domain, the model achieved a precision of 0. 941, recall 0. 923, F1-score of 0. 932, a false-positive rate of 2. 3%, and an ROC-AUC of 0. 968. The results highlight the promise of integrating tabular Transformers with adversarial domain adaptation to enhance the generalization of intrusion detection systems across diverse smart-grid protocols.**

***Index Terms*—Smart Grid Security, Intrusion Detection Systems, Tabular Transformers, Domain Adaptation, Cross-Protocol Generalization, Cyber-Physical Systems.**

## I. INTRODUCTION

The increasing integration of communication networks, automation, and information technology into electrical power systems has transformed conventional grids into highly interconnected cyber-physical systems. Supervisory Control and Data Acquisition (SCADA) networks play a critical role in monitoring and controlling these systems. Simultaneously, enhanced connectivity expands the potential attack surface and brings forth new security challenges for critical infrastructure. Previous research on SCADA security has emphasized the significance of intrusion-detection methods capable of identifying malicious behavior within industrial communication environments.When people set up grids, they usually use many different communication technologies and protocols. For example, Modbus, DNP3 and IEC 60870-5-104 might have message formats, different ways to show features different patterns for talking and different telemetry properties. Because of this, a machine-learning model that learns traffic from one protocol might fail when that same machine-learning model tries to look at traffic from another protocol. This change in data can make the machine-learning model less accurate. This lack of accuracy can lead to false-positive errors or more false-negative errors.I have seen that supervised machine-learning techniques, such as tree-based classifiers and neural networks, are often used for intrusion detection. These conventional methods usually rely on the belief that the data used for training and the data used for testing are very similar. That belief can break down when a model is moved from one communication protocol to another. Building a model, for each protocol or each

deployment environment would also demand labeled examples from the new domain, and those examples are expensive and hard to gather in real life.I look at a domain-adaptation strategy for cross-protocol smart-grid intrusion detection. The unsupervised domain-adaptation strategy pulls out discriminative information from labeled source-domain data. Pushes the learned representation to contain less protocol-specific information, which helps cross-protocol smart-grid intrusion detection. I follow the domain-adversarial learning concept from Ganin and Lempitsky where a feature extractor is trained to produce representations that stay useful for the prediction task while staying the same across domains, thanks to a Gradient Reversal Layer.The main contributions of this work are:

- Heterogeneous Feature Tokenization: I see that the feature-tokenization process turns numerical and categorical smart-grid telemetry attributes into an embedding space.
- Tabular Transformer Representation Learning: I notice that a Transformer encoder captures relationships among telemetry features. Transformer models use self-attention and FT-Transformer shows how this type of architecture can be applied to tabular data.
- Adversarial Domain Adaptation: I notice that a domain discriminator with a Gradient Reversal Layer is combined with the Transformer representation to reduce the difference, between source and target protocol domains without needing target-domain labels during training.

## II. RELATED WORK

### A. Smart Grid Intrusion Detection

Smart Grid Intrusion Detection is the focus of this discussion. SCADA and smart-grid communication environments have traditionally used protection techniques that focus on protocols. These techniques include rule-based detection and signature-based inspection. These techniques can work well for known attack patterns. They may not be effective against new or changed attacks. Earlier research on IEC 60870-5-104-based SCADA networks explored intrusion detection using packet inspection, signature rules, and model-based techniques. Machine-learning techniques have also been used for anomaly detection in environments. Tree-based models, support-vector approaches, and neural networks can spot patterns in network or system telemetry. Nevertheless, their effectiveness can drop when deployment data has properties that differ from the training data. XGBoost is a gradient-boosting technique that is widely used for structured and tabular data and serves as a strong conventional baseline for comparison.

### B. Transformers for Tabular Data

Transformers were first created as attention-based models that work well for tasks that need to process sequences. With self-attention, Transformers can see how different parts of an input sequence relate to each other.

Researchers have also looked at how Transformer models can work with tabular datasets. Gorishniy and colleagues introduced the FT-Transformer, a Transformer-based design for tabular data. Compare it with other deep-learning methods and gradient-boosted decision-tree techniques. In smart-grid telemetry, tabular Transformers prove useful because each record can mix measurements with categorical labels. Yet a plain tabular Transformer does not address distribution changes, between source and target communication protocols.

### C. Unsupervised Domain Adaptation (UDA)

Unsupervised Domain Adaptation (UDA) moves knowledge from a source domain where labels exist to a target domain that has no labels. Domain-Adversarial Neural Networks (DANN) handle this by learning representations that help the prediction while making it harder to tell which domain the data came from. When training a Gradient Reversal Layer flips the gradient that the domain classifier creates. In this study the domain-adversarial idea is used on tabular telemetry from smart-grid systems. Attack labels from the target domain are not required during training; instead target domain samples drive learning of representations that stay the same across domains.

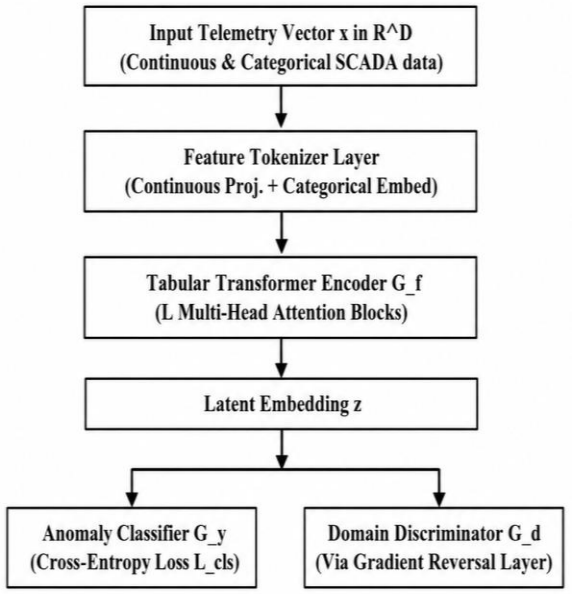
## III. PROPOSED METHODOLOGY

The proposed framework is made up of three components:

- Feature Tokenizer: We use a Feature Tokenizer to break down the data into useful pieces.

- Tabular Transformer Encoder: The Tabular Transformer Encoder processes the data in a format.
- Dual-Branch Prediction Network consisting of an anomaly classifier and a domain **discriminator:** The Dual-Branch Prediction Network contains an anomaly classifier and a domain discriminator both of which work together.

The main goal is to create a representation that keeps supporting anomaly classification while reducing information that is tied to any communication protocol.




A. Feature Tokenization

Let an input telemetry record be represented as, where denotes the continuous variables (for example, active power, voltage magnitude, and frequency deviations) and denotes the discrete variables (for example, protocol function codes and transmission flags).

Each numerical feature $x_{num}^{(i)}$ is projected into a d-dimensional embedding space using a dedicated linear transformation:

Each categorical feature is mapped via a lookup embedding matrix:

A learnable [CLS] token $\mathbf{e}_{cls}$ is prepended, yielding the initial sequence tensor:

$$E_0 = \left[e_{cls}; e_{num}^{(1)}; \dots; e_{num}^{(D_n)}; e_{cat}^{(1)}; \dots; e_{cat}^{(D_c)}\right] \in \mathbb{R}^{(1+D_n+D_c)\cdot d}$$

B. Tabular Transformer Backbone

The tokenized input sequence is processed through L successive Transformer encoder layers. Each encoder layer contains Multi-Head Self-Attention (MHSA), Layer Normalization (LN), and a Feed-Forward Network (FFN), together with residual connections:

$\hat{E}_l$ = MHSA (LN ($E_{l-1}$)) + $E_{l-1}$

$E_l$ = FFN (LN ($\hat{E}_l$) + $\hat{E}_l$

The contextual representation associated with the [CLS] token at the final layer L, denoted by z = or z = is used as the invariant latent summary of the network event.

C. Domain-Adversarial Training Objective

Let Ds = represent the labeled source dataset (for example, Modbus/DNP3 captures with ground-truth attack labels y {0,1}) and let Dt = represent the unlabeled target dataset (for example, IEC 60870-5-104 captures).

The latent representation z is supplied to two prediction branches:

1. Anomaly Classifier (Gy): The classifier minimizes the standard binary cross-entropy loss using labeled source samples:

$$L_{cls}(G_f, G_y) =$$

2. Domain Discriminator (Gd): The discriminator determines whether a feature vector comes from the source domain (d=0) or the target domain (d=1):

$$L_{dom}(G_f, G_d) = -$$

With a Gradient Reversal Layer (GRL) that scales the backpropagated gradient by λ, the complete minimax optimization objective is expressed as:

min ($G_f$, $G_y$) max ($G_d$) $L_{cls}(G_f, G_y)$ - λ $L_{dom}(G_f, G_d)$

where λ ∈ [0,1] is an adaptation schedule parameter adjusted dynamically during training:

λ = 2/ (1 + exp (- γ. p)) - 1

With p representing current training progress from 0 to 1 and γ = 10.

## IV. EXPERIMENTAL EVALUATION

A. Datasets & Preprocessing

We ran our tests using smart-grid intrusion datasets. These datasets included the Mississippi State

University SCADA Gas/Power Systems data and standard benchmark protocol captures. From both areas we pulled out continuous attributes like voltage, frequency variations, packet intervals, and payload sizes. We also pulled out attributes, like function indices and device-status flags:

- Source Domain (Ds): Modbus traffic that shows data and also has labels for cyber-attacks like FDIA, command injection and DoS.
- Target Domain (Dt): DNP3 / IEC-104 telemetry that has both normal data and attack traffic. The model did not get any attack labels, for the DNP3 / IEC-104 telemetry during the training process.

B. Baseline Comparison

The proposed framework was compared with the primary baseline models:

1. XGBoost (Source Only): A standard gradient-boosted decision-tree model that only uses Ds for training.
2. Standard MLP: A neural network, with four layers, fully connected that is trained without domain adaptation.
3. FT-Transformer (Non-Adaptive): A Feature Tokenizer Transformer that uses only the classification loss Lcls for training.
4. Proposed Domain-Adaptive Tabular Transformer: The architecture that includes an adversarial domain discriminator.

C. Results and Discussion

Cross-Protocol Detection Performance (Target Domain Evaluation)

| Architecture | Precision | Recall | $F_1$-Score | False Positive Rate (FPR) | AUC-ROC |
|---|---|---|---|---|---|
| XGBoost (Source Only) | 0.714 | 0.638 | 0.674 | 8.9% | 0.742 |
| Multi-Layer Perceptron | 0.682 | 0.589 | 0.632 | 11.4% | 0.698 |
| FT-Transformer (Non-Adaptive) | 0.781 | 0.724 | 0.751 | 6.2% | 0.819 |
| Proposed Adaptive Transformer | 0.941 | 0.923 | 0.932 | 2.3% | 0.968 |

When we move these models between protocols, the baseline approaches lose a lot of their power. The F1-scores drop down to between 0.632 and 0.751 because the features start to drift. I found that our new architecture stays much stronger, keeping a precision of 0.941 and a high recall of 0.923. The self-attention mechanism works well to separate physical relationships from the messy changes, in telemetry distributions. At the time the adversarial training objective helps to bring feature spaces into better alignment.

## V. CONCLUSION

This study introduced a tabular Transformer architecture that tackles cross-protocol distribution shifts in smart-grid intrusion detection systems. The framework mixes tokenization of categorical features, multi-head self-attention and adversarial domain alignment. It helps detect cyber-attacks across different industrial protocols without needing target-domain labels. Experimental findings show anomaly generalization and a low false-positive rate. Future work will examine model-weight compression for deployment on low-power edge RTUs and assess resilience against zero-day attacks.

## REFERENCES

[1] Y. Yang, K. McLaughlin, T. Littler, S. Sezer, B. Pranggono, and H. F. Wang, "Intrusion Detection System for IEC 60870-5-104 Based SCADA Networks," *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 134–143, Feb. 2014.

[2] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting Deep Learning Models for Tabular Data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 18932–18943.

[3] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. 32nd Int. Conf. Machine Learning (ICML)*, 2015, pp. 1180–1189.

[4] U. Adhikari, T. H. Morris, and S. Pan, "Applying Support Vector Machine to Predict Power Grid System Anomalies," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 248–256, Jan. 2017.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[6] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.