Zero-day Attack Detection in Digital Substations using In-Context Learning

Faizan Manzoor
Virginia Tech
Blacksburg, VA, USA

Vanshaj Khattar
Virginia Tech
Blacksburg, VA, USA

Chen-Ching Liu
Virginia Tech
Blacksburg, VA, USA

Ming Jin
Virgnia Tech
Blacksburg, VA, USA

***Abstract*—In this paper, we address the critical challenge of detecting zero-day attacks in digital substations that employ the IEC-61850 communication protocol to ensure the security and reliability of modern power systems. While many heuristic and machine learning (ML)-based methods have been proposed for attack detection in IEC-61850 digital substations, generalization to unknown or zero-day attacks remains a challenge. We propose an approach that leverages the in-context learning ability of transformer architecture, which enables the model to learn from a few examples of a new task without explicit retraining. Our experiments on the IEC-61850 dataset demonstrate that the proposed method achieves more than 87% detection accuracy on zero-day attacks while the existing baselines fail. We believe this work has the potential to enhance the security of digital substations by enabling the effective detection of zero-day attacks.**

***Index Terms*—In-context learning; IEC-61850; Intrusion Detection Systems; Zero-day attacks.**

I. Introduction

The IEC–61850 communication protocol is commonly used in digital substations to communicate between Intelligent Electronic Devices (IEDs) and Merging Units (MUs). Although the IEC-61850 allows for efficient connectivity and control in digital substations [1], there are numerous vulnerabilities that an attacker can exploit to disrupt the operation of these digital substations [2]. Recently, the occurrences of cyber-attacks on digital substations have increased. For example, in 2015, a coordinated cyber-attack was responsible for the mass-scale power outages in Ukraine [3]. In 2016, another cyber-attack in Ukraine also led to a mass power outage and affected the SCADA system at the transmission level [4]. Moreover, a recent study showed that millions of new cyber-attacks were detected annually worldwide from 2015 to 2020 [5]. As a consequence, cybersecurity of digital substations has recently been a focus for many researchers [6].

Intrusion Detection Systems (IDSs) play an important role in detecting potential attacks on digital substations so that timely action can be taken. Although IDS techniques are well-explored for traditional TCP/IP-based substation communication networks [7], the specific requirements and unique communication protocols of IEC–61850 substations have not been adequately addressed. Many existing IDS methods, whether heuristic-based [8] or machine learning (ML)-based [9], are designed for specific cases or trained on known attacks, which can limit their ability to generalize to zero-day attacks (unseen attacks).

In this paper, we propose a generalizable IDS framework for the IEC-61850 communication protocol that can detect zero-day attacks on digital substations. Our method leverages the ”in-context learning” (ICL) ability of transformer architectures [10] (not to be confused with power transformers in electrical substations), which have demonstrated potential in various domains, including natural language processing (NLP) and computer vision.

In-context learning is the ability to generalize rapidly from a few examples of a new task that have not previously been seen without any updates to the model, a key characteristic of many large language models (LLMs) [11]. For example, consider the following context examples of network packets provided to an LLM: $Packet1 = Normal; Packet2 = Normal; Packet3 = Attack$. Then, if the query sample to the LLM is $Packet4$, which shares similar characteristics with $Packet3$, the LLM may output ”Attack” as it is able to understand through the in-context examples that packets with certain features are classified as attacks. This ability of LLMs to understand the context and adapt their outputs accordingly without any additional training motivates the following question: **“How can this in-context learning ability be leveraged for zero-day attack detection?”** Insights from this investigation may guide the design and deployment of effective IDS in digital substations.

Main contributions. *1)* We propose an intrusion detection framework that leverages the ICL ability of transformer models to detect zero-day attacks in digital substations. *2)* We provide training and testing recommendations for the transformer architectures to be used in the IDS applications. *3)* Finally, we validate the effectiveness of our approach through extensive experiments on the IEC-61850 dataset and its ability to detect zero-day attacks.

A. Related Work

ML-based methods. In [12], the authors use a neural network-based approach to detect spoofed packets. Another work improved the previous approach using the decision trees and random forests [13]. However, many of these existing ML-based methods focus on specific attack cases and do not generalize to novel attacks. In contrast, our proposed ICL-based method focuses on zero-day attack detection.

Our approach for zero-day attack detection also shares some similarities with other advanced ML paradigms, such

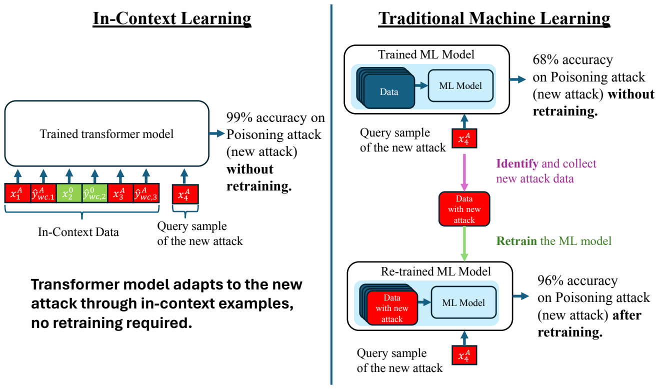
In-Context Learning
Trained transformer model
99% accuracy on Poisoning attack (new attack) without retraining.
x_1^A $\hat{y}_{wc.1}^A$ x_2^0 $\hat{y}_{wc,2}^0$ x_3^A $\hat{y}_{wc,3}^A$ x_4^A
In-Context Data
Query sample of the new attack
Transformer model adapts to the new attack through in-context examples, no retraining required.
Traditional Machine Learning
Trained ML Model
Data
ML Model
68% accuracy on Poisoning attack (new attack) without retraining.
Query sample of the new attack
x_4^A
Identify and collect new attack data
Data with new attack
Retrain the ML model
Re-trained ML Model
Data with new attack
ML Model
96% accuracy on Poisoning attack (new attack) after retraining.
Query sample of the new attack
x_4^A

Fig. 1. Traditional ML models, such as Random Forest, Decision Trees, and Convolutional Neural Networks trained on a specific dataset, often fail to identify novel attacks during the deployment phase. To adapt to the new threats, they require retraining with datasets that include these novel attacks. In contrast, the proposed ICL-based approach can detect novel attacks even if they were not included in the training dataset. ICL allows it to use the in-context data and weak labels (WLs) to better generalize and recognize new attacks without the need for retraining or parameter updates.

as transfer learning [14], out-of-distribution (OOD) detection [15], meta-learning [16], and multi-task learning [17]. These techniques also aim to improve the generalization capability of ML models and enable them to adapt to new tasks or unseen data. However, our approach specifically leverages the ICL ability of transformer models, which allows them to adapt to new tasks based on input context without additional training.

Heuristic-based methods. In [18], the authors use timestamp and sequence numbers to detect a replay attack. In [19], authors investigated the injected spoofing attack on the IEC-61850-based standard. However, the proposed model was specifically designed for spoofing attacks and may have limited applicability to other types of attacks without further adaptations. The work in [20] develops an IDS for the IEC-61850 protocol, but they only consider information carried within sampled value messages, which restricts its application to other message-sharing protocols (such as GOOSE).

The rest of the paper is organized as follows. In Section II, we provide the preliminaries on the transformer architecture and the ICL. Section III provides our proposed IDS methodology, and Section IV includes experiments and validates our zero-day attack detection approach.

II. Preliminaries

A. Transformer Architecture and In-context Learning (ICL)

The GPT-2 (Generative Pre-trained Transformer 2) is a large-scale language model that has shown remarkable performance in various NLP tasks [21]. GPT-2 employs a transformer architecture [10], which relies on a self-attention mechanism to model long-range dependencies in sequential data. It consists of multiple layers of multi-head self-attention and feed-forward neural networks, enabling the model to capture complex patterns and relationships within the input sequences.

One of GPT -2's key strengths is its ability to perform in-context learning. The model can adapt its predictions based on the provided examples, allowing it to generalize to new tasks without retraining. This capability is particularly relevant in intrusion detection, as it enables the detection of novel zero-day attacks in digital substations without requiring the model to be retrained on the unseen attack.

We chose GPT-2 as the backbone of our proposed framework as it has a more manageable model size and computational requirements, making it more practical for deployment in resource-constrained environments such as digital substations. Moreover, the availability of pre-trained GPT-2 models and the associated training code facilitates the implementation and reproducibility of our approach.

In-context learning (ICL). ICL is an attractive property of transformer models that allows them to adapt to new tasks given as input context without updating the model parameters [11]. Consider a transformer model M_θ, where θ are the model parameters that take sequence length of N as input, where $N-1$ samples are input-label pairs which we call in-context data $D_{N-1} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{N-1}, y_{N-1})\}$. The N^{th} sample, denoted as x^q, is the query point for which we want to predict the label. The model predicts the label for x^q as follows:

$$y^q = M_\theta(x^q; D_{N-1}).$$

The transformer achieves the above output by computing the following conditional probability: $P(y^q | x^q, D_{N-1})$.

B. Intrusion Detection in Digital Substations

In the digital substations that use the IEC-61850 protocol, there are 3 main network protocols: Sampled Values (SV), Generic Object-Oriented Substation Events (GOOSE), and MMS [6]. In this paper, we only consider the SV and GOOSE, as they are the ones involved in substation protection functions. SV packets transmit sampled voltage and current values from MUs to IEDs, while GOOSE packets enable fast and reliable communication between IEDs for exchanging control commands. Despite the benefits offered by the IEC-61850, it also introduces vulnerabilities that can be exploited by attackers. For example, the lack of authentication and encryption in GOOSE and SV protocols can allow attackers to inject false data, manipulate control commands, or launch denial-of-service attacks [22]. These vulnerabilities highlight the need for developing effective IDS for IEC-61850-based digital substations.

III. Methodology

In this section, we describe our overall ICL-based IDS framework and the training and testing procedures in detail. Figure 2 and 3 show our overall framework.

A. Training Data Generation

Many recent works have highlighted the importance of training data diversity to foster ICL capabilities in transformer models [23]. Specifically, within the cybersecurity context, this translates to the inclusion of various types of attack scenarios. However, there are not many existing attack datasets for the IEC-61850 protocol. To increase the diversity of the

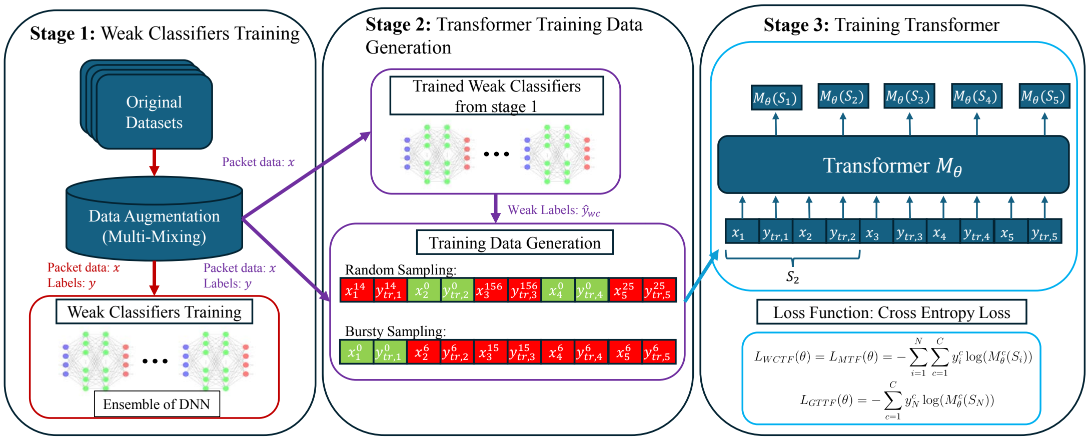
Stage 1: Weak Classifiers Training
Original Datasets
Packet data: x
Data Augmentation (Multi-Mixing)
Packet data: x Labels: y
Packet data: x Labels: y
Weak Classifiers Training
Ensemble of DNN
Stage 2: Transformer Training Data Generation
Trained Weak Classifiers from stage 1
Weak Labels: $\hat{y}_{wc}$
Training Data Generation
Random Sampling:
x_1^{14} $y_{tr,1}^{14}$ x_2^0 $y_{tr,2}^0$ x_3^{156} $y_{tr,3}^{156}$ x_4^0 $y_{tr,4}^0$ x_5^{25} $y_{tr,5}^{25}$
Bursty Sampling:
x_1^0 $y_{tr,1}^0$ x_2^6 $y_{tr,2}^6$ x_3^{15} $y_{tr,3}^{15}$ x_4^6 $y_{tr,4}^6$ x_5^6 $y_{tr,5}^6$
Stage 3: Training Transformer
$M_\theta(S_1)$ $M_\theta(S_2)$ $M_\theta(S_3)$ $M_\theta(S_4)$ $M_\theta(S_5)$
Transformer M_θ
x_1 $y_{tr,1}$ x_2 $y_{tr,2}$ x_3 $y_{tr,3}$ x_4 $y_{tr,4}$ x_5 $y_{tr,5}$
S_2
Loss Function: Cross Entropy Loss
$L_{WCTF}(\theta) = L_{MTF}(\theta) = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log(M_\theta^c(S_i))$
$L_{GTTF}(\theta) = -\sum_{c=1}^{C} y_N^c \log(M_\theta^c(S_N))$

Fig. 2. Transformer training as explained in Section III-C. In stage 1, synthetic data is generated using multi-mixing. Using this data, weak classifiers are trained. In stage 2, data samples $(x_1, y_{tr,1}, \ldots, x_5, y_{tr,5})$ are generated for training, where x_i represents packets from synthetic data and $y_{tr,i}$ represents their labels that can be ground-truth (y_i), weak classifier $(\hat{y}_{wc,i})$ or mixture of both $(y_{mix,i})$. Finally, in stage 3, the transformer is trained with cross-entropy loss, where the loss function depends on the training strategy: ground-truth Trained Transformer (GTTF), Weak Classifier Trained Transformer (WCTF), or Mixture of weak classifiers and ground-truth Transformer (MTF).

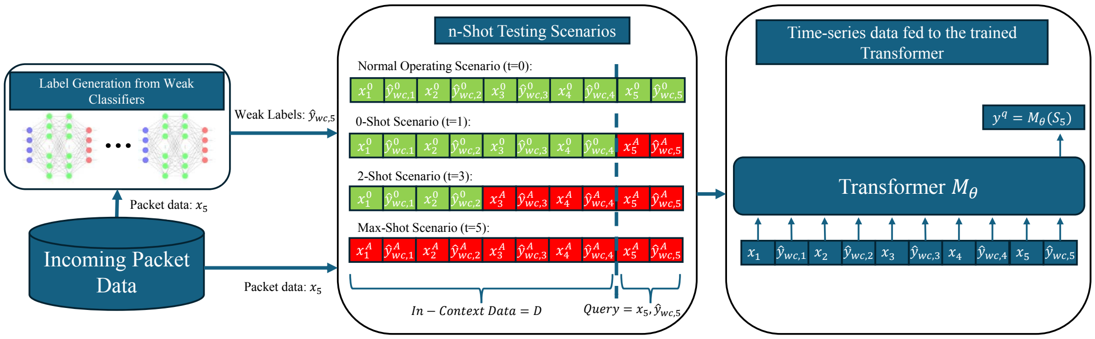
Label Generation from Weak Classifiers
Weak Labels: $\hat{y}_{wc,5}$
Packet data: x_5
Incoming Packet Data
Packet data: x_5
n-Shot Testing Scenarios
Normal Operating Scenario (t=0):
x_1^0 $\hat{y}_{wc,1}^0$ x_2^0 $\hat{y}_{wc,2}^0$ x_3^0 $\hat{y}_{wc,3}^0$ x_4^0 $\hat{y}_{wc,4}^0$ x_5^0 $\hat{y}_{wc,5}^0$
0-Shot Scenario (t=1):
x_1^0 $\hat{y}_{wc.1}^0$ x_2^0 $\hat{y}_{wc,2}^0$ x_3^0 $\hat{y}_{wc,3}^0$ x_4^0 $\hat{y}_{wc,4}^0$ x_5^A $\hat{y}_{wc,5}^A$
2-Shot Scenario (t=3):
x_1^0 $\hat{y}_{wc,1}^0$ x_2^0 $\hat{y}_{wc,2}^0$ x_3^A $\hat{y}_{wc,3}^A$ x_4^A $\hat{y}_{wc,4}^A$ x_5^A $\hat{y}_{wc,5}^A$
Max-Shot Scenario (t=5):
x_1^A $\hat{y}_{wc,1}^A$ x_2^A $\hat{y}_{wc,2}^A$ x_3^A $\hat{y}_{wc,3}^A$ x_4^A $\hat{y}_{wc,4}^A$ x_5^A $\hat{y}_{wc,5}^A$
$In - Context\ Data = D$
$Query = x_5, \hat{y}_{wc,5}$
Time-series data fed to the trained Transformer
$y^q = M_\theta(S_5)$
Transformer M_θ
x_1 $\hat{y}_{wc,1}$ x_2 $\hat{y}_{wc,2}$ x_3 $\hat{y}_{wc,3}$ x_4 $\hat{y}_{wc,4}$ x_5 $\hat{y}_{wc,5}$

Fig. 3. Transformer testing phase as explained in Section III-D. Incoming packet (x_5) is received and passed to pre-trained weak classifiers to generate weak labels $(\hat{y}_{wc,5})$. This data-label pair query $(x^{qs} = (x_5, \hat{y}_{wc,5}))$ is then appended at the end of the sequence. The trained transformer model sees different scenarios depending on the incoming packets in the in-context data and then predicts the label y^q for x_5 using ICL.

training data, we introduce an approach called ”multi-mixing”, where the key idea is to generate synthetic attacks without collecting additional data by linearly combining features from different attack classes. By creating these synthetic examples, we aim to expose the transformer model to a wider range of attack patterns, potentially improving its ability to detect unseen attacks. To formalize this, consider a dataset with C distinct classes. Multi-mixing generates a new synthetic class as follows:

$$A_{\text{new}} = \sum_{i=1}^{C} \alpha_i A_i, \tag{1}$$

where A_i represents all data points (shuffled) belonging to the i^{th} class, $\alpha_i \in [0, 1]$ denotes the weight of the i^{th} class incorporated into the new class A_{new}. Each of these new classes is assigned a new label, which allows the model to learn more latent features and generalize better. We will denote the number of attack classes generated using the multi-mixing by T, which represents the training attack diversity.

B. In-context Data from Weak Classifiers

In our proposed approach, the transformer model relies on ICL to adapt to new attack scenarios. However, during testing, the true labels of the incoming data packets are not available. Therefore, we use weak classifiers, which are pre-trained models (e.g., neural networks), for the initial predictions for the incoming data packets. These are then used as pseudo-labels in the in-context data for the transformer model. We refer to these classifiers as ’weak’ because they might not have perfect detection accuracy but rather provide initial predictions that can guide the transformer model’s ICL process. To mitigate the impact of individual weak classifier errors, we concatenate the outputs from multiple such “weak classifiers”, creating a collective output $\hat{y}_{wc} \in \mathbb{R}^d$, where d denotes the number of weak classifiers. We use deep neural networks as weak classifiers tailored for multi-class classification.

C. Training the Transformer Model

To train the transformer model for intrusion detection, there are 3 possible label choices that can be used for each in-context sample: *1)* the ground-truth labels; *2)* the weak classifier labels; *3)* a mixture of weak classifier and the ground-truth labels. In the experiments section, we show that using a mixture of weak classifiers and ground-truth labels gives the best zero-day attack detection accuracy.

We train the transformer model with N input-label pairs, denoted as $\{(x_1, y_{tr,1}), (x_2, y_{tr,2}), \ldots, (x_{N,y_{tr,N}})\}$, where tr denotes the training sample. Each x_i is sampled randomly and burstily [24] from the T attack classes and the normal data. The label $y_{tr,i}$ for each $i \in \{1, 2, \ldots, N\}$ depends on the chosen training approach (*1*), *2*), or *3*)). When we only use the weak classifier labels, the label is denoted as $\hat{y}_{wc,i}$. For the mixture approach, as $y_{mix,i}$. When we only use ground-truth labels, the label is $y_{tr,i} = y_i$, with the exception for the N^{th} case, where the label is $y_{tr,N} = \hat{y}_{wc,N}$. This adjustment sets $y_{tr,N}$ as $\hat{y}_{wc,N}$ rather than the actual y_N to discourage the model from relying exclusively on y_N when predicting for x_N. This approach helps ensure the model's generalization ability during testing.

For each input $i \in \{1, 2, \ldots, N\}$, transformer model considers the context data $S_i = (x_1, y_{tr,1}, \ldots, x_i, y_{tr,i})$ -where $D_{i-1} = \{(x_1, y_{tr,1}), (x_2, y_{tr,2}), \ldots, (x_{i-1}, y_{tr,i-1})\}$ is the in-context data and i^{th} set is the query set $x^{qs} = (x_i, y_{tr,i})$- to make its prediction $M_\theta(S_i)$ for the target y_i. Note, instead of using just the query point x_i for which we want to predict the true label, we append $y_{tr,i}$ as well, which provides extra information to the transformer model about what the true label could be.

We design two different cross-entropy loss functions for training: *(i)* WCTF (Weak Classifier Trained Transformer)/ MTF (Mixed Trained Transformer) loss and *(ii)* GTTF (ground-truth Trained Transformer) loss:

$$L_{WCTF}(\theta) = L_{MTF}(\theta) = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log(M_\theta^c(S_i)), \quad (2)$$

$$L_{GTTF}(\theta) = -\sum_{c=1}^{C} y_N^c \log(M_\theta^c(S_N)), \quad (3)$$

where $M_\theta^c(\cdot)$ gives the probability of $x_i \in c$, C is the total number of classes, and y_i^c is an indicator showing whether class label c is the correct label for x_i. The reason for using a different loss function for GTTF is to avoid calculating the loss for $i < N$, which prevents the model from developing a bias towards learning from $y_{tr,i}$ for $i < N$ that represents only the ground-truth labels. Moreover, despite initially training for multi-class classification, our study focuses on binary classification as we are only interested in inferring if the incoming packet is anomalous or not.

D. Testing

Once trained, we can deploy the transformer model to detect anomalies in real time. The most recent packet received and its weak classifier labels are treated as the query set, $x^{qs} = (x_N, \hat{y}_{wc,N})$, while the preceding $N-1$ packets and their weak labels serve as the in-context data D_{N-1}. Under standard substation operations, we would anticipate that all packets within the sequence are normal. However, in the event of an attack, the x_N becomes anomalous, while the earlier in-context packets would likely still reflect normal conditions. We refer to the transformer model's ability to detect anomalies where the context remains normal while only the query point is anomalous—as its **zero-shot** performance. This scenario assesses the model's ability to detect completely novel attacks based solely on its learned representations.

As the attack persists, the attacker continues to send anomalous packets, which begin to appear in the in-context data. The model's performance in this setting evaluates its ability to adapt and detect the new attack based on these few new unlabeled examples. This gradual transition from normal to anomalous in the in-context data allows us to evaluate the model's **n-shot** performance, where n denotes the number of anomalous packets present within the in-context data D_{N-1}.

Note that our approach differs from traditional ICL. Instead of having access to true labels, we rely solely on pseudo-labels generated by weak classifiers. From a deployment perspective, this methodology is especially beneficial for real-time intrusion detection in digital substations. It enables the model to quickly adapt to new types of attacks based on just a few observed instances.

IV. Experiments

In this section, we validate the proposed ICL-based IDS framework on a real-world attack dataset called ERENO–IEC–61850 [6]. Specifically, the dataset contains 7 types of GOOSE-based attacks: random and inverse replay, masquerade fake normal, masquerade fake, message injection, high-status number, and high-rate flooding. Additionally, it includes two types of SV-based attacks: message injection and inverse replay. Through the experiments, we aim to answer the following questions: **Q1:** How does increasing training attack diversity improve zero-day attack detection in digital substations? **Q2:** What is the most effective way to leverage weak classifiers and ground-truth labels during training to enhance the model's performance? **Q3:** How well does the proposed approach is able to detect zero-day or unseen attacks?

We utilize GPT-2 transformer architecture for our experiments [21]. For training, we specifically choose 5 attacks: 3 from the GOOSE and 2 from the SV, along with normal data. We treat the 5 selected attacks and the normal data as in-distribution (ID) data, while all the other data from the dataset as out-of-distribution (OOD), which we use to test the zero-day attack detection accuracy. We train the transformer with an in-context sample size, $N = 11$.

A. The Impact of Training Data Diversity

We answer Q1 by examining the influence of training data diversity on ICL and zero-day attack detection by training the GPT-2 transformer on different numbers of attack classes generated during training, i.e., $T \in \{100, 300, 500, 700, 900\}$. These attacks are generated using the multi-mixing approach discussed in Section III-A. Firstly, from Figure 4, we see that increasing the number of shots (attack instances in the in-context data) improves the accuracy on all four types of attacks not seen during the training.

Second, we also see that as the training data diversity increases, the ICL performance also improves. This is indi-

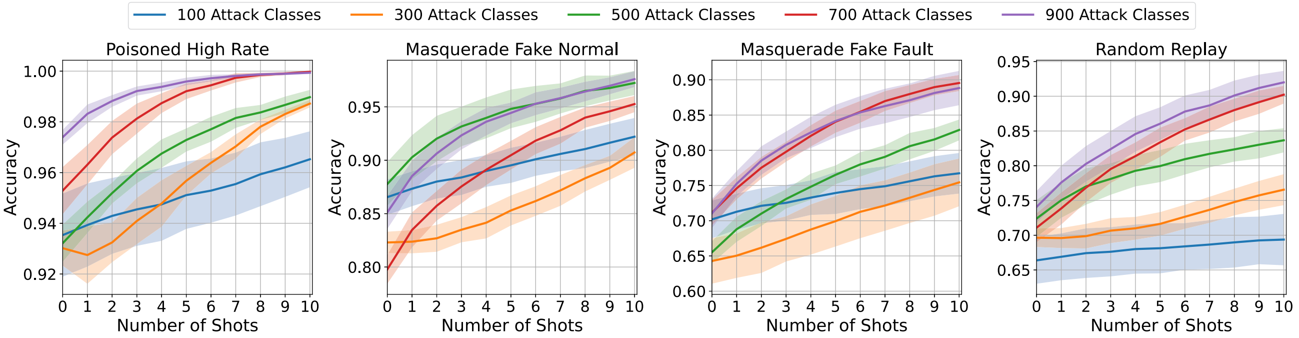

Fig. 4. Performance of WCTF against the number of shots for out-of-distribution (OOD) attacks across different numbers of training attack diversity.

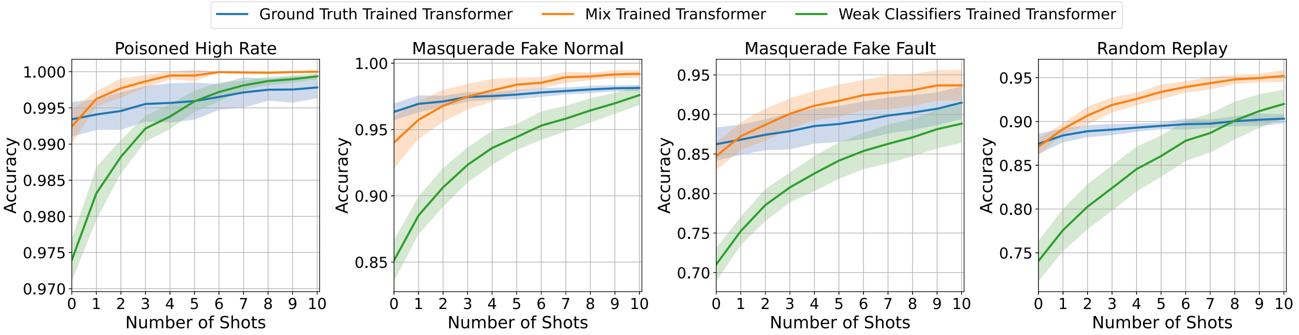

Fig. 5. Performance of transformer for different OOD attacks with an increasing number of shots under different training strategies: GTTF, WCTF, and MTF.

	OOD				ID					
Models	Poisoned High Rate	Masquerade Fake Fault	Masquerade Fake Normal	Random Replay	Normal	High Status Number	Inverse Replay	Injection	SV High Status Number	SV Injection
Logistic Regression	0.311	0.173	0.387	0.027	0.884	0.164	0.323	0.165	0.087	0.298
Decision Tree	1.000	0.992	0.029	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SVM	0.898	0.756	0.645	0.995	0.821	0.999	0.995	0.995	0.995	0.995
Naive Bayes	1.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DNN	0.544	0.532	0.648	0.087	0.476	0.578	0.728	0.419	0.478	0.476
CNN	0.978	0.854	0.717	0.939	1.000	1.000	0.998	1.000	1.000	1.000
RNN	0.684	0.839	0.608	0.996	0.996	0.999	0.968	0.999	1.000	1.000
LSTM	0.584	0.784	0.886	0.996	0.998	1.000	0.728	0.995	1.000	1.000
Hard Voting WC	0.992	0.729	0.597	0.827	1.000	0.999	0.997	1.000	0.978	1.000
Soft Voting WC	0.994	0.754	0.627	0.873	1.000	0.999	0.997	1.000	0.978	1.000
MTF 0-Shot **(Ours)**	0.992	0.871	0.847	0.940	0.999	0.982	0.987	1.000	0.937	1.000
MTF Max-Shot **(Ours)**	1.000	0.952	0.936	0.992	0.999	0.989	0.998	1.000	1.000	1.000

TABLE I

PERFOMANCE OF DIFFERENT MODELS ON OUT-OF-DISTRIBUTION (OOD) AND IN-DISTRIBUTION (ID) DATA.

cated by improved performance on both zero-shot and n-shot scenarios, when n varies from 1 to 10.

We repeated these experiments for WCTF and saw similar trends of improved performance with the number of shots and training data diversity. These results suggest that the transformer model trained on a higher diversity of attacks will be more effective at predicting OOD attacks in a zero-shot scenario and will also generalize more rapidly in an n-shot setting for intrusion detection in digital substations. Given that the 900 attack classes in the training dataset performed the best, we report the next experiments only for this case.

B. Comparative Analysis of Labels During Training

We answer Q2 by investigating the impact of 3 possible label choices during training: *1)* ground-truth labels; *2)* weak classifier labels; *3)* a mixture of weak classifier and the ground-truth labels. First, we experimented with different mixing ratios. Our experiments showed that the attack detection accuracy for the ratio of 60% of the weak classifier labels and 40% of ground-truth labels during training provided the best attack detection accuracy during testing. Moreover, Figure 5 shows the comparison of attack detection accuracy during testing when the transformer model is trained on ground-truth only, weak classifier only, and mixed labels (60% ratio) during training. It clearly shows the advantage of mixing the ground-truth and weak classifier labels during training, as it leads to the best zero-shot and n-shot performance.

The GTTF model shows high zero-shot performance but low ICL, while WCTF demonstrates the opposite. This suggests an inverse relation between zero-shot and ICL, possibly due to the bias introduced by the normal samples in the in-context dataset D_{N-1}. However, the MTF exhibits both high zero-shot performance and ICL, defying expectations. Under the zero-shot setting, the MTF distinguishes whether the query sample belongs to the D_{N-1} distribution. As the scenario transitions to n-shot, the MTF leverages its ICL, combining the zero-shot capabilities of GTTF and the ICL of WCTF. Overall, our analysis suggests that we can enhance the effectiveness of transformer-based IDS for digital substations by incorporating a combination of weak classifiers and ground-truth labels during the training process.

C. Zero-day Attack Detection

Finally, we answer Q3 by validating the trained model on the attacks not seen during the training (OOD attacks) and on the attacks that are seen during the training (ID attacks). We compare our results against widely known ML-based classification methods. The GPT-2 transformer is trained on 900 attack classes and a mix of 60-40 ratio of weak classifiers and ground-truth labels.

Table III-D shows the final results where the red cells represent 'failure cases,' which are defined as follows: *1)* inability to achieve an accuracy threshold of 80% for a specific attack, and *2)* failure to exceed a 99.5% accuracy for normal. The table clearly shows that the widely used ML-based methods experienced at least one failure case, indicating a shortfall in detecting certain types of zero-day attacks. Moreover, traditional ensembling techniques, such as hard voting and soft voting, when applied to weak classifiers also experience failure cases. In contrast, our proposed ICL-based method, when assessed under the 0-shot and max-shot conditions, did not show such failure cases. This observation suggests an intrinsic capacity of our model for recognizing unseen attacks.

As adversarial strategies become more sophisticated alongside advancements in AI, it is plausible that attackers will devise maneuvers that elude the detection capabilities of conventional ML and DL models, as well as rule-based systems. However our model will detect such maneuvers and generalize better using ICL. It's important to note that the optimal ratio of weak classifier labels to ground-truth labels may vary depending on the specific characteristics of the power system and the available resources for labeling.

V. Conclusion and Future Work

In this paper, we proposed an ICL-based approach for detecting zero-day attacks in IEC-61850-based digital substations. Our approach leverages the ICL capabilities of transformer models to adapt to unseen attack scenarios, enabling the detection of zero-day attacks. While our ICL-based approach has shown promising results, there are limitations to be addressed in future work. One important direction is to extend the approach to handle more complex attack scenarios, such as multi-stage attacks and coordinated attacks on multiple substations. Additionally, investigating the integration of our approach with other security measures, such as anomaly detection and threat intelligence sharing, could provide a more comprehensive defense strategy.

References

[1] International Electrotechnical Commission et al. Communication networks and systems for power utility automation. *IEC Std*, 61850, 2013.

[2] Abdulrahaman Okino et.al. Otuoze. Smart grids security challenges: Classification by sources of threats. *Journal of Electrical Systems and Information Technology*, 5(3):468–483, 2018.

[3] Michael J Assante. Confirmation of a coordinated attack on the ukrainian power grid. *SANS Industrial Control Systems Security Blog*, 207, 2016.

[4] Defense Use Case. Analysis of the cyber attack on the ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)*, 388(1-29):3, 2016.

[5] Global new malware volume (2020) statista. http://www.statista.com/statistics/680953/global-malwarevolume/.

[6] Silvio E Quincozes, Célio Albuquerque, Diego Passos, and Daniel Mossé. Ereno: A framework for generating realistic iec–61850 intrusion detection datasets for smart grids. *IEEE Transactions on Dependable and Secure Computing*, 2023.

[7] Alexandru Stefanov and Chen-Ching Liu. Cyber-power system security in a smart grid environment. In *2012 IEEE PES Innovative Smart Grid Technologies (ISGT)*, pages 1–3. IEEE, 2012.

[8] Junho Hong, Chen-Ching Liu, and Manimaran Govindarasu. Integrated anomaly detection for cyber security of the substations. *IEEE Transactions on Smart Grid*, 5(4):1643–1653, 2014.

[9] Nitasha Sahani, Ruoxi Zhu, Jin-Hee Cho, and Chen-Ching Liu. Machine learning-based intrusion detection for smart grid computing: A survey. *ACM Transactions on Cyber-Physical Systems*, 7(2):1–31, 2023.

[10] Vaswani et.al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Tom Brown et.al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[12] Mohamad et.al. El Hariri. The iec 61850 sampled measured values protocol: Analysis, threat identification, and feasibility of using nn forecasters to detect spoofed packets. *Energies*, 12(19):3731, 2019.

[13] Ustun et.al. Artificial intelligence based intrusion detection system for iec 61850 sampled values under symmetric and asymmetric faults. *Ieee Access*, 9:56486–56495, 2021.

[14] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[15] Yang et.al. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[16] Vanshaj Khattar, Yuhao Ding, Bilgehan Sel, Javad Lavaei, and Ming Jin. A cmdp-within-online framework for meta-safe reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[17] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

[18] Junho Hong, Chen-Ching Liu, and Manimaran Govindarasu. Detection of cyber intrusions using network-based multicast messages for substation automation. In *ISGT 2014*, pages 1–5. IEEE, 2014.

[19] Vetrivel Subramaniam et.al. Rajkumar. Cyber attacks on power system automation and protection and impact analysis. In *2020 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. IEEE, 2020.

[20] Yi et.al. Yang. Intrusion detection system for iec 61850 based smart substations. In *2016 IEEE power and energy society general meeting (PESGM)*, pages 1–5. IEEE, 2016.

[21] Alec Radford et.al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[22] Reda et.al. Vulnerability and impact analysis of the iec 61850 goose protocol in the smart grid. *Sensors*, 21(4):1554, 2021.

[23] Allan et.al. Raventós. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Stephanie et.al. Chan. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.