GENERATIVE AI FOR CRITICAL INFRASTRUCTURE IN SMART GRIDS: A UNIFIED FRAMEWORK FOR SYNTHETIC DATA GENERATION AND ANOMALY DETECTION

Aydin Zaboli and Junho Hong
University of Michigan-Dearborn
Dearborn, MI United States
{azaboli, jhwr}@umich.edu

ABSTRACT

In digital substations, security events pose significant challenges to the sustained operation of power systems. To mitigate these challenges, the implementation of robust defense strategies is critically important. A thorough process of anomaly identification and detection in information and communication technology (ICT) frameworks is crucial to ensure secure and reliable communication and coordination between interconnected devices within digital substations. Hence, this paper addresses the critical cybersecurity challenges confronting IEC61850-based digital substations within modern smart grids, where the integration of advanced communication protocols, e.g., generic object-oriented substation event (GOOSE), has enhanced energy management and introduced significant vulnerabilities to cyberattacks. Focusing on the limitations of traditional anomaly detection systems (ADSs) in detecting threats, this research proposes a transformative approach by leveraging generative AI (GenAI) to develop robust ADSs. The primary contributions include the suggested advanced adversarial traffic mutation (AATM) technique to generate synthesized and balanced datasets for GOOSE messages, ensuring protocol compliance and enabling realistic zero-day attack pattern creation to address data scarcity. Then, the implementation of GenAI-based ADSs incorporating the task-oriented dialogue (ToD) processes has been explored for improved detection of attack patterns. Finally, a comparison of the GenAI-based ADS with machine learning (ML)-based ADSs has been implemented to showcase the outperformance of the GenAI-based frameworks considering the AATM-generated GOOSE datasets and standard/advanced performance evaluation metrics.

Keywords Anomaly detection · Data Generation · GOOSE · Generative AI · IEC61850 · Zero-day Attack

1 Introduction

Power grid blackouts considerably interrupt societal and economic operations, caused by factors such as human errors, technical malfunctions, and environmental incidents. Additionally, the risk of cyberattacks inducing these outages highlights an increasing susceptibility. Consequently, ensuring the cybersecurity of ICT systems integral to power grid functions has emerged as an imperative need [1]. Digital substations utilizing IEC61850 are pivotal within the power grid architecture, overseeing the allocation, conversion, and integration of energy streams. The advent of smart grids has amalgamated the power grid infrastructure with communication networks and computing functions, paving the way for numerous groundbreaking applications such as automated data acquisition and the remote management of electrical systems and elements [2, 3]. Nevertheless, integrating these systems introduces a range of security vulnerabilities to the smart grid. ADSs play a crucial role in identifying and mitigating malicious activities by adversaries. Historically, ADSs have proven effective in traditional ICT fields for this purpose. However, with the adoption of IEC61850 and the implementation of specific communication protocols, such as multicast messages (e.g., GOOSE), new avenues have emerged for customized malicious strategies that exhibit distinct traffic and attack patterns. These may comprise unauthorized data interceptions and denial-of-service (DoS) attacks. Thus, it is essential for ADSs to develop and refine new signatures for training, testing, validation, and assessment to address these emerging challenges effectively [4, 5].

1.1 Problem Statement

The criticality of substations is grounded in their function as nodes that handle multiple transmission lines, a feature that amplifies their impact on grid stability. Traditional contingency planning, such as $N-1$ analyses, assesses the system's resilience to the loss of a single component, but the compromise of a substation due to different types of cyberattack can lead to an $N-m$ scenario—simultaneous outages of multiple lines—overwhelming the grid's capacity and triggering catastrophic disruptions. This vulnerability is particularly heightened in the U.S., where the expansion of unmanned substations, dependent on remote access for maintenance, has created significant exposure points. These facilities are ideal targets for cyber intrusions, as both authorized engineers and malicious actors can exploit the same access mechanisms. The reliance on remote access in unmanned substations has advantages and limitations; while it facilitates operational efficiency, it also increases cybersecurity risks. The potential for unauthorized access by intruders, leveraging the same entry points as legitimate operators, underscores a critical gap in current security frameworks. This issue is compounded by the evolving nature of cyber threats, which can exploit vulnerabilities in communication protocols, leading to scenarios such as false data injection (FDI), DoS, and replay (RE) attacks. The need to reinforce these systems is evident, given their role in ensuring grid reliability and the potential for cascading failures that can disrupt entire regions.

Furthermore, ML techniques applied within ADSs play a crucial role in detecting and correcting inconsistencies in GOOSE multicast transmissions. These techniques are acknowledged for their accuracy and emphasis on data, offering a sophisticated foundation for cybersecurity measures. Nonetheless, they do present certain obstacles. A significant limitation is the need for ongoing model retraining whenever new attack vectors emerge. Upon detection of a novel attack pattern, the ML models require updating to integrate this novel data. This retraining procedure demands substantial time and resources, which results in a period of vulnerability during which the system is exposed to new threats (e.g., zero-day attacks) that have not yet been included in the model's intelligence framework [6]. Furthermore, the ability of these ML-driven ADSs to scale, alongside their efficiency in decision-making and data processing, is critically significant for the operational functionality. Scalability concerns focus on the model's proficiency in adapting and sustaining performance as the network expands or as data volume increases. The decision-making aspect relates to the model's capability for accurately identifying secure versus malicious actions, a challenge that grows more complicated with the advancement of sophisticated attack methods. Finally, the domain of data processing highlights the necessity for proficient management and analysis of extensive datasets that are frequently essential for system operations [7, 8]. The identified key deficiencies in existing ML-based ADSs pose substantial difficulties within industrial control system (ICS) networks, especially in settings where GOOSE messages are fundamental to time-sensitive communications.

1.2 Research Objectives

The successful fusion of traditional electrical grids with advanced communication networks and computational frameworks through smart grid technologies brings about significant security risks, particularly through the IEC61850 standard and its related protocols. This integration facilitates unique attack vectors marked by distinct traffic and attack patterns that pose considerable challenges to conventional ADSs. Addressing this evolving security framework requires the development of enhanced ADS capabilities featuring specialized signatures for robust training, testing, and validation against novel threats. This is especially crucial for identifying zero-day attacks in traffic patterns, which lack predetermined rules in practical scenarios, further complicated by the scarcity of realistic, balanced, and comprehensive IEC61850-based communication datasets. In light of these significant challenges, it is essential to develop adaptive, resilient, and scalable AD solutions that effectively reduce latency in integrating new threat intelligence while simultaneously improving decision-making processes and data management capabilities. Applications or systems known as GenAI tools employ large language models (LLMs) and sometimes other AI models to produce content customized to user inputs within distinct contexts, presenting a revolutionary approach via platforms such as OpenAI's ChatGPT [9], Anthropic Claude Pro [10] and Microsoft Copilot AI [11]. These GenAI solutions, crafted with precision for deep contextual understanding using advanced memory architectures and natural language processing (NLP) capabilities, demonstrate the ability to identify zero-day attacks through contextual evaluation, even in the absence of substantial prior knowledge of particular threat signatures [12], thus significantly decreasing the workload on human operators relative to traditional approaches that rely extensively on routine retraining procedures and fixed data models. Synthesizing pre-processed datasets and engineering ADSs that integrate past trends while independently analyzing changing network conditions and suggesting responses based on data insights, the goal of this research is to design security frameworks that possess the ability to autonomously respond, continuously learn, and process data at scale, thereby improving the robustness and dependability of digital substations. This is achieved by ensuring that innovative or complex threats are promptly detected, preventing significant disruptions to essential infrastructure systems [3, 13–16].

1.3 Literature Review

The evolution of digital substations has necessitated the development of sophisticated ADSs leveraging ML techniques. These systems analyze data patterns to identify cyber threats in real-time, ensuring power grid reliability [17–19]. Recent research has explored diverse ML approaches for AD. Alvee *et al.* [20] developed a convolutional neural network (CNN)-based methodology that converts binary files into images for ransomware detection, though limited by dataset constraints and scenario coverage. A real-time ADS using advanced ML was explored in [21], but faced challenges in processing large data volumes efficiently. Hybrid approaches combining CNN and long short-term memory (LSTM) networks [22] showed promise but struggled with generalization to new attack types. Several studies focused on IEC61850 protocol-specific solutions. Eynawi *et al.* [23] developed ML-based feature selection for GOOSE and Sampled Value (SV) messages, though computational overhead limited real-time deployment. Quincozes *et al.* [24] introduced innovative feature engineering for IEC61850, but relied on static datasets. A game theory integration by Jay [25] offered proactive defense mechanisms, though practical implementation remained computationally intensive. Bhattacharya *et al.* [26] achieved impressive results with k-Nearest Neighbors (KNN), but lacked deep learning (DL) integration and real-world deployment validation. Other approaches included gradient boosting [27], distributed security systems [28], and ML-driven GOOSE message analysis [29], each with specific limitations in scalability, computational demands, or protocol coverage. A GenAI-based ADS considering the human-in-the-loop (HITL) was proposed by Zaboli *et al.* [13] to implement the detection processes in IEC61850-based multicast messages, considering different GPT tools. However, there was no continuous learning and automated processing of actions in this research alongside the multicast messages extracted from the hardware-in-the-loop (HIL) testbed. Moreover, Zaboli *et al.* [5] suggested a novel GenAI-based ToD framework for the AD process to overcome the challenges given in the GenAI-based ADS using the HITL. It covered the learning and automated processes gaps and made a good comparison with the HITL technique considering different GPT tools. While this framework showcased the outperformance over the HITL process, the lack of a comparison of this framework with ML-based ADS was a gap of the research. Also, the balance and realistic issues of datasets considering the zero-day attack were missed as new threats emerged in real-world applications.

Recent data balancing methodologies build upon and extend established traditional approaches. Bhattacharya *et al.* [26] employed the RUS and SMOTE for imbalanced datasets in IEC61850-based messages. Although SMOTE can efficiently generate new samples for datasets with few dimensions, its performance drops markedly when dealing with high-dimensional data. Moreover, SMOTE's method of interpolation does not always guarantee that the synthetic data will be of high quality or that the interpolation is performed in the best possible way, which can lead to considerable noise in the dataset. Recent developments in GANs have achieved outstanding results in computer vision and image synthesis. As a result, researchers are now exploring the use of GANs to generate samples for minority classes, with the aim of addressing data imbalance challenges [30]. Some challenges in the data pre-processing part for the AD process can include data imbalance, absence of prior information, increased data complexity, data generation for anomalies with different patterns, resource extensiveness, and re-training process for zero-day attacks, which can be found in Table 1 with the relevant descriptions for each challenge [26, 31–36]. Dairi *et al.* introduced semi-supervised DL schemes that effectively learn temporal dependencies using only normal training data; nonetheless, their dependence on IEC60870-5-104 datasets (for SCADA control) creates a challenge when extending these methods to the semantics of GOOSE messages in digital substations [37]. López *et al.* [38] developed a substation-aware ADS that integrates substation topology to enhance detection accuracy, although its limited capability to infer the deep semantic content of GOOSE messages remains a critical gap. Moreover, Sahani *et al.* [39] provided a comprehensive survey of ML-based ADS in smart grids, offering valuable insights into ML applications while also revealing scalability and real-time processing challenges. Additionally, Anwar *et al.* [40] compared unsupervised learning algorithms for intrusion detection in the IEC60870-5-104 SCADA protocol, contributing to the understanding of detection performance while exposing low detection accuracies that limit practical deployment. An ADS along with network packet features for wide-area protection was proposed by Singh and Govindarasu [41], though the complexity in feature selection and high processing overhead hinders a smooth integration. Khaw *et al.* [42] presented a universal DL-based cyberattack detection system for transmission protective relays that simplifies model tuning across different fault types, but its reliance on static training datasets limits its adaptability to evolving zero-day attacks. Lim *et al.* [43] reviewed the application of GANs for AD in network security, demonstrating that these models can generate synthetic minority samples to improve the detection of rare attack patterns; however, they also highlighted challenges such as an over-reliance on predefined evaluation metrics and insufficient representation learning when dealing with extremely imbalanced datasets. Similarly, Yuan *et al.* [30] proposed a GAN framework, which leverages LSTM networks to capture temporal features and generate high-quality synthetic intrusion data for industrial control systems, enhancing the AD performance on imbalanced datasets; however, their approach is challenged by issues of mode collapse and limited diversity among the generated anomalies. In addition, Sauber-Cole *et al.* [44] surveyed GAN techniques for mitigating class imbalance in tabular data, emphasizing the promise of GAN architectures to produce representative minority instances; nevertheless, they noted persistent challenges, including architectural sensitivity and the lack of standardized evaluation metrics to reliably assess synthetic data quality. Manzoor *et al.* introduced a method that exploits the in-context learning (ICL) capability

Table 1: The challenges encountered in detecting anomalies within network systems in terms of data availability.

Challenge	Description
Excessive dependence on prior knowledge [31]	Contemporary approaches to AD exhibit a pronounced reliance on established attack signatures and predefined network behavioral baselines. Such dependence on historical data and fixed analytical frameworks can restrict their effectiveness in detecting novel or rapidly evolving security threats (zero-day attacks).
Resource-demanding & time-consuming implementation [32]	Generating new signatures or updating profiles for current detection systems entails considerable time and resource investment, typically requiring the specialized expertise of network security professionals.
Unavailability of prior contextual attack data [33]	The scarcity of preceding intelligence, particularly regarding threat vectors such as zero-day attacks, represents a significant operational challenge. Moreover, unknown system vulnerabilities and evolving attack methodologies often further complicate the implementation of effective cybersecurity countermeasures.
Elevated data complexity and absence of real-time detection mechanisms [34]	Due to the exponential increase in data complexity and size, network traffic continues to intensify, making the real-time attack detection and the maintenance of consistent monitoring increasingly difficult.
Data augmentation aimed at anomalous trend emergence [35]	Conventional AD methods often encounter difficulties in producing unknown anomalous datasets, which are crucial for effectively training and evaluating detection algorithms.
Imbalanced class [26, 36]	The network AD inherently suffers from class imbalance, characterized by a privilege of normal instances relative to a scarcity of abnormal ones. Consequently, AD models may experience performance degradation when faced with extreme discrepancies in training samples, underscoring the criticality of addressing data imbalance.

inherent in transformer architectures to detect zero-day attacks in digital substations. Their technique allows the model to integrate new attack examples with minimal retraining, achieving detection accuracies exceeding 85% for zero-day scenarios where conventional state-of-the-art baselines fail. Nonetheless, the method is contingent on the diversity of training data, as its efficacy is highly dependent on the number and heterogeneity of attack classes included during training. Moreover, the overall performance of the system is significantly influenced by the quality of the weak classifiers, with notable performance degradation observed. Further, this research used the multi-mixing technique to generate synthetic datasets without good validation and meeting the IEC61850-based messages violation rules, which is a critical gap [45, 46]. Lin *et al.* [47] developed "CausalPrompt," a novel prompting strategy designed to adapt LLMs for classification and regression tasks via weakly supervised causal reasoning. By integrating domain-specific causal inferences during the fine-tuning phase, their approach enhanced the adaptability and resilience of energy systems against data distribution shifts. Their experimental results revealed improvements in predictive performance under feature changes. However, the method encounters challenges in safety-critical applications, as performance still degrades under significant feature shifts. Additionally, the approach's heavy reliance on domain expert reasoning, which is not always readily available, coupled with the high financial costs associated with fine-tuning commercial LLM APIs, presents further practical constraints. Also, the nature of the data generation process based on the different energy rules in terms of the unknown anomalies is not clear. Quincozes *et al.* [48] proposed the Efficacious Reproducer Engine for Network Operations (ERENO), a framework for generating realistic intrusion detection datasets specialized to IEC61850-based standards. Their system synthesized traffic features by integrating data from both network and physical domains, thereby facilitating cross-protocol detection between GOOSE and SV messages. The framework successfully generated datasets that model eight distinct use cases, including common attack types as well as normal network traffic and demonstrates that the integration of enriched substation features can enhance intrusion detection performance. Nevertheless, their implementation faces gaps in effectively detecting sophisticated masquerade attacks that are engineered to mimic legitimate behavior, and the current proof-of-concept primarily addresses attack scenarios based on illegitimate GOOSE messages, leaving other potential attack vectors and protocols less explored.

1.4 Contributions

According to the current research gaps, the enhancements of this research can be classified based on the data generation and efficiency of the proposed GenAI-based ADS based on the ToD framework [5]. Therefore, the main strengths of the proposed methodology using the previously proposed GenAI-based ADS can be summarized as follows:

- **An AATM technique for synthetic data generation:** A novel perturbation and mutation-based synthetic data generation methodology is proposed to address the critical challenge of insufficient and imbalanced IEC61850-based communication datasets in digital substations. The AATM technique employs protocol-aware transformation functions that generate realistic zero-day attack patterns while maintaining strict adherence to GOOSE protocol rules through gradient-guided perturbations and categorical feature mutations. This method can outperform existing approaches including conditional generative adversarial network (CGAN). This methodology establishes a robust foundation for generating protocol-compliant synthetic datasets essential for training advanced ADSs against evolving cyber threats in digital substations.
- **A validation of the GenAI-Based ADS considering the ToD framework, compared with ML-based models through standard and advanced performance metrics:** This method advances traditional ML-based ADSs by utilizing the contextual comprehension abilities of GenAI, along with integrating expertise from domain specialists through the continuous learning. Hence, a comparative analysis of the GenAI-based ADS with a ToD configuration is conducted to demonstrate the superiority of rule-based GenAI frameworks over traditional ML algorithms including Feedforward Neural Networks (FNN), Recurrent Neural Networks (RNN), and Support Vector Machines (SVM) for an AD process in IEC61850 communications. In this case, the AATM generated GOOSE datasets are considered for the comparative analysis which have better balance rate (BR) and realism rate (RR). The evaluation employs both standard and advanced metrics to provide comprehensive assessment of detection capabilities. This validation establishes that incorporating domain-specific rules and contextual understanding through GenAI significantly enhances AD performance beyond what traditional ML algorithms can achieve.

1.5 Paper Organization

The rest of this paper is organized as follows: Section 2 presents the IEC61850-based multicast data generation techniques, particularly GOOSE messages. Further, this section provides different steps of the generation process based on the proposed and current techniques, GOOSE rules, and validation framework. Section 3 demonstrates the description of the proposed GenAI-based ToD framework and modeling of this structure alongside ML-based ADSs. The results and discussion of the proposed AATM data generation techniques in terms of the balance and realistic aspects are given in Section 4. Also, a comparison of the proposed GenAI- and ML-based ADSs is implemented in this section considering the novel GOOSE message generation technique. Finally, the concluding remarks, along with potential directions for future research, are presented in Section 5.

2 Proposed IEC61850-based Multicast Data Generation Technique in Digital Substations

In a controlled and realistic environment, a HIL testbed is crucial for evaluating the interaction between cyber attacks and the robustness of power systems. This real-time HIL testbed comprises an integration of various elements, such as hardware, software, communication protocols, and simulation technologies, all integrated with GPS synchronization. The incorporation of these elements is essential for exploring the real-time dynamics inherent in communication and information processing. This understanding is critical for the analysis of cyber attacks, enhancing detection protocols, and developing robust strategies for effective mitigation [49]. The configuration of the HIL testbed comprises diverse elements such as protection IEDs, SDN switches, a GPS unit, a merging unit IED, a SCADA system, a real-time digital simulator, and an amplifier. The system utilizes a SCADA-based distribution management system (DMS) that gathers measurements and implements control commands utilizing DNP3 protocols. The designed IEDs possess the function of dispatching control signals to circuit breakers (CBs). A CB, in turn, is engineered to react to GOOSE messages by communicating its operational status—either open or closed—back to the protective IEDs. Moreover, the merging unit IED is responsible for transmitting digital current and voltage measurements from the digital real-time simulator to the protective IEDs by employing the amplifier [16, 17]. It is important to mention that specifics about the HIL testbed are outside the scope of this study, as this research primarily focuses on techniques for data pre-processing and generation. Within the HIL test environment, Wireshark (a tool for analyzing network packets) enables a comprehensive capture of communication packets. This procedure entails real-time observation and examination of network traffic in the HIL testbed, facilitating comprehensive tracking and recording of packet flows. Utilizing the functionalities of Wireshark, researchers can capture a snapshot of communication flows, facilitating a deeper comprehension of the interactions among various components within the test environment. This systematic approach guarantees precise extraction of essential packets, thereby enriching the research endeavor with significant insights into the operational behaviors of cyber-physical systems (CPSs) [13]. Detailed information on the datasets, the procedures for extraction, and the outlined features, along with the GOOSE rules for both normal and abnormal patterns, will be discussed in the forthcoming section. This methodology can similarly be applied to other multicast messages (e.g., SV messages) due to their unique rules and characteristics. Hence, this section introduces an innovative analytical framework for an analysis

of AD processes in IEC61850-based communication messages (specifically GOOSE messages). An in-depth analysis is outlined as follows:

2.1 Synthetic Balanced and Realistic Data Generation Process

This part shows the process for the generation of the GOOSE datasets in a way that can meet the requirements for the generation of realistic zero-day attacks and balancing issues. Hence, a comparison of the proposed technique known as the AATM with another methodology (i.e., CGAN [30, 43]) is carried out to show the better performance of this proposed method. Hence, a general framework of the data generation part considering the contributions and different steps is illustrated in Fig. 1. This pipeline begins with a raw packet capture step, proceeds through feature extraction and

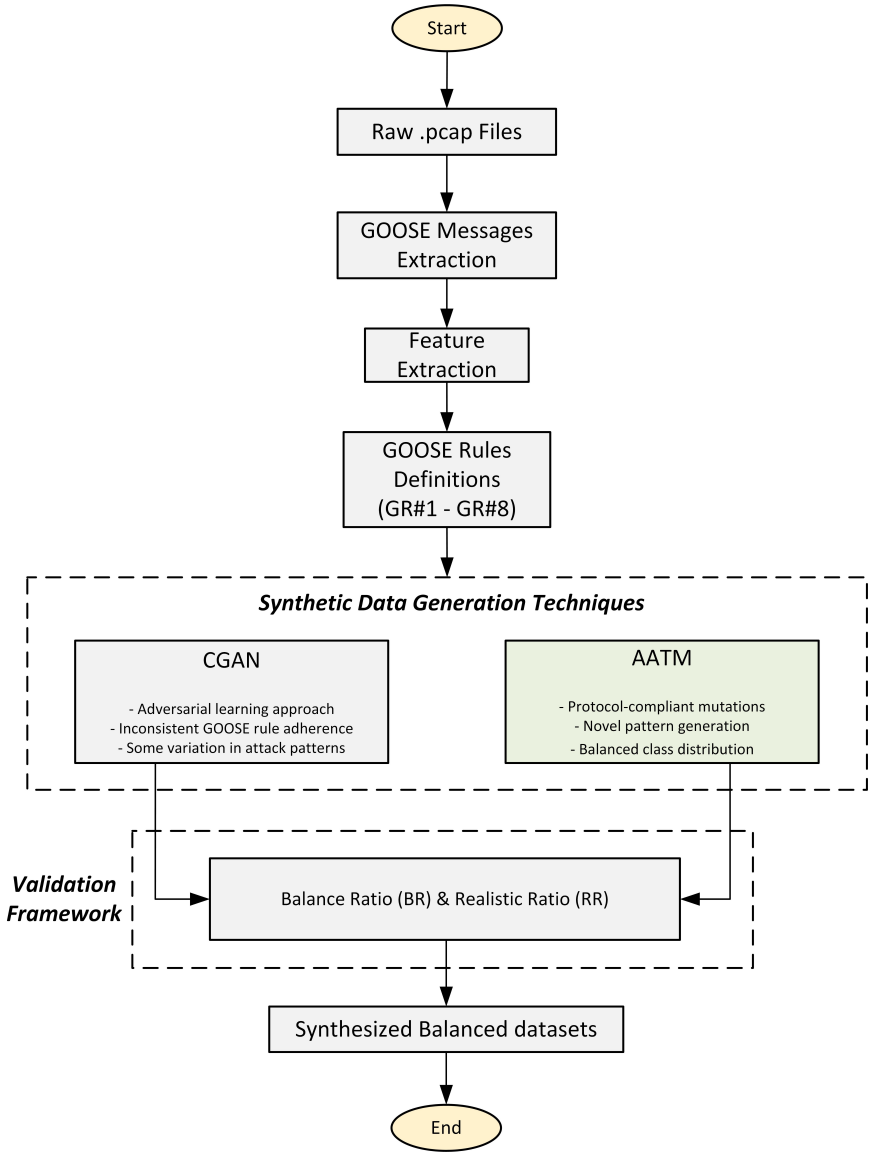


Figure 1: A systematic workflow of the proposed methodology for generating balanced and realistic GOOSE datasets.

GOOSE rules definitions. Before the GOOSE rules definitions, it is necessary to provide details of GOOSE features in a sample dataset. The GOOSE packet data encompasses 14 distinct data types, designated by the features extracted using tshark (a terminal-oriented version of Wireshark) which is shown in Fig. 2 [13]. The temporal attribute meticulously logs the exact transmission instance of a packet, detailing the time in hours, minutes, seconds, and milliseconds to ensure comprehensive precision. The abbreviations DM and SM stand for the destination and source media access control (MAC) addresses, respectively, acting as essential identifiers in the communication framework. In particular,

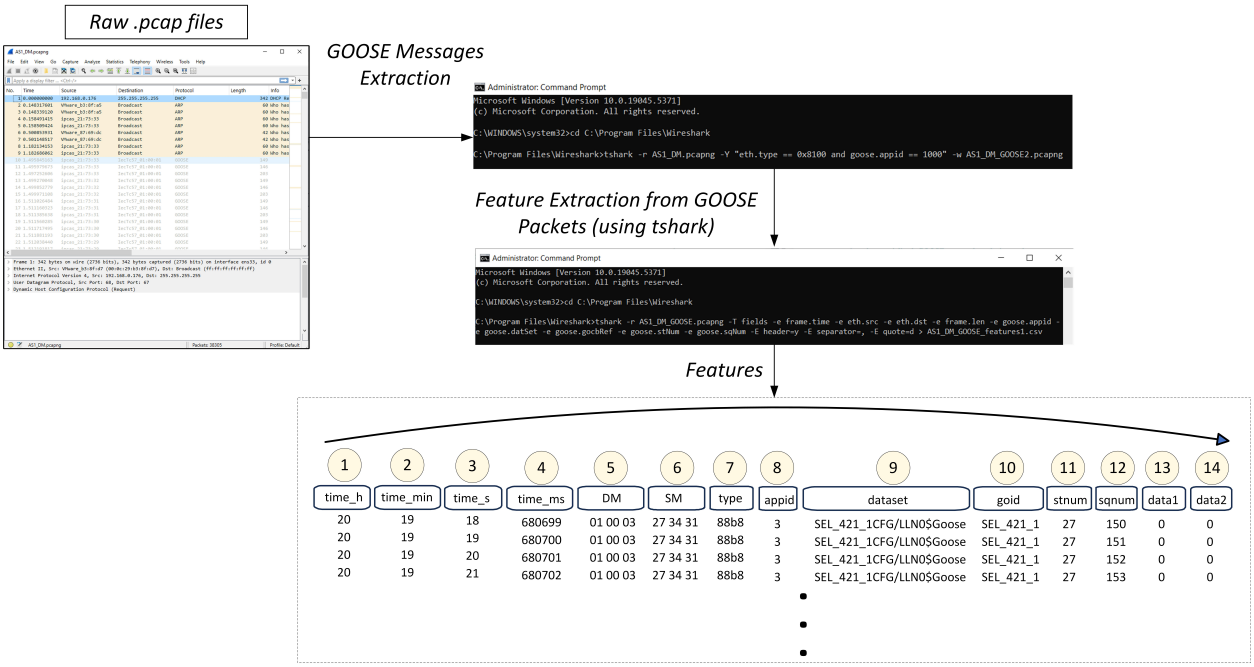


Figure 2: The process of GOOSE packets and features extraction using tshark.

the designated DM address associated with GOOSE messages is referred to as (010003), and is directed at devices associated with this specific MAC address, whereas the SM address is depicted as 273431, which determines the transmitting IED. In GOOSE packet classification, the type indicator is specified by the value $88b8$. Furthermore, for GOOSE communications, the Application identifier (APPID) is set to 3. The dataset name and GOOSE identification are expressed by the dataset and goid attributes, which depend on the DM address. Additionally, state number (stNum) and sequence number (sqNum) are utilized in the context of GOOSE messages. In the analysis phase, data1 ($d1$) and data2 ($d2$) are considered, both of which are derived as binary features from GOOSE packets. Given a "*.pcap*" file, P containing all packets in which p represents only each individual packet. The GOOSE set is defined as G that only includes packets that meet specific criteria as described in Eq. (1) [50, 51]:

$$G = \{p \in P \mid p.\text{eth\_type} = 0 \times 8100 \wedge p.\text{appid} = 1000\} \tag{1}$$

The first condition shows the Ethernet type that must be 0×8100 for GOOSE messages, and the second condition is relevant to GOOSE APPID which should 1000 according to messages. By meeting these two conditions, the system guarantees extracting only GOOSE packets. This process can be carried out using tshark, enabling the capture of packet data from live networks as well as the reading of packets from pre-existing capture files. It provides the option to output a decoded representation of the data to the standard output or to write the packets to a file for subsequent analysis. Further, it can extract the features of data as well as a data conversion to other formats (e.g., .csv format). This extraction process ensures the capture of all relevant GOOSE messages while filtering out other traffic types. In the subsequent phase, GOOSE guidelines (*GR#1* through *GR#8*) are outlined to address a range of both normal and abnormal scenarios. Eqs. (2)–(9) demonstrate the GOOSE guidelines utilized in this study to examine the various anomalies within datasets as follows [5, 14]:

GR#1: In the event that sequential data packets possess the same *DM* and *SM* characteristics, the *sqNum* parameter must be incremented. If there are discrepancies, it is indicative of an abnormality.

$$GR\#1(G_i, G_{i-1}) = \begin{cases} 1, & \text{if } DM_i = DM_{i-1} \wedge SM_i = SM_{i-1} \wedge sqNum_i = sqNum_{i-1} + 1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

GR#2: In GOOSE communications, an anomaly that suggests a DI attack occurs when there is a transition in a data value, such as *data1* (*d1*) or *data2* (*d2*), from 0 to 1 or from 1 to 0, while the *stNum* remains unchanged and the *sqNum*

is consistently incremented in sequence.

$$GR\#2(G_i, G_{i-1}) = \begin{cases} 1, & \text{if } (d1_i \neq d1_{i-1} \vee d2_i \neq d2_{i-1}) \wedge (stNum_i = stNum_{i-1}) \wedge (sqNum_i = sqNum_{i-1} + 1) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

GR#3: Within GOOSE messages where the DM and SM are identical, the *stNum* is expected to either stay constant under normal scenarios or increment when data changes occur. A reduction in *stNum* is considered irregular unless it results from a system rollover—specifically, when *stNum* attains its maximum value of $2^{32} - 1$, it resets to 0 in the next GOOSE transmission—or from a valid system reboot.

$$GR\#3(G_i, G_{1:i-1}) = \begin{cases} 1, & \text{if } DM_i = DM_{i-1} = ... = DM_{i-n} \& SM_i = SM_{i-1} = SM_{i-n} \wedge \\ & (stNum_i = stNum_{i-1} \vee stNum_i = stNum_{i-1} + 1 \vee stNum_i = 0 \wedge \\ & stNum_{i-1} = 2^{32} - 1)) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

GR#4: Variations in the parameters *DM*, *SM*, *type*, *appid*, *dataset*, or *goid* imply the occurrence of abnormal conditions.

$$GR\#4(G_i, G_{i-1}) = \begin{cases} 1, & \text{if } DM_i = DM_{i-1} \wedge SM_i = SM_{i-1} \wedge type_i = type_{i-1} \wedge dataset_i = dataset_{i-1} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

GR#5: Although GOOSE messages transmit timestamps in a binary format, it is essential that the timestamps extracted and subsequently decoded within this dataset adhere to a uniform presentation format, specifically reflecting hours, minutes, seconds, and milliseconds (e.g., HH:MM:SS.mmm).

$$GR\#5(time_i) = \begin{cases} 1, & \text{if } time_i \text{ is in format} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

GR#6: Although it is recognized that the frequent occurrence of GOOSE messages is typical during protection operations—such as those involving busbar protection with several breaker activations lasting up to 500 ms—this framework analyzes the patterns in message frequency. The core of this monitoring focuses on the count of consecutive messages, identified by their timestamps at the millisecond level, and sets a standard threshold of 10 instances occurring within a 10 μs window.

$$GR\#6(G_{i-9:i}) = \begin{cases} 1, & \text{if } \forall j \in [i-9, i-1] : time_{j+1} - time_j \leq 10\mu s \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

GR#7: A data transmission interruption extending beyond 10 seconds serves as a sign of an anomalous condition.

$$GR\#7(G_i, G_{i-1}) = \begin{cases} 1, & \text{if } time_i - time_{i-1} \leq 10s \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

GR#8: In the context of GOOSE messages, a transition in a data value (e.g., *data1* or *data2*) from 1 to 0 or from 0 to 1, coupled with an unchanged *stNum* and a *sqNum* that remains at 0, signifies an irregularity suggestive of a possible RE attack.

$$GR\#8(G_i, G_{i-1}) = \begin{cases} 1, & \text{if } (d1_i \neq d1_{i-1} \vee d2_i \neq d2_{i-1}) \wedge (stNum_i = stNum_{i-1}) \wedge (sqNum_i = sqNum_{i-1}) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

According to Fig. 1, the next step presents two different techniques for synthetic data generation including CGAN and the proposed technique known as AATM. Compared with the CGAN technique, the AATM technique is characterized by its use of transformation functions instead of NNs, eliminating the need to train a model from scratch while operating through directed perturbations of existing samples under rule-based constraints and utilizing gradient information (i.e., derivatives or rates of change of a function) to guide these perturbations without building or training an NN architecture. Also, there are other techniques (e.g., multi-mixing [46]) which cannot meet requirements for balance and realistic concerns because the nature of this method is based on simple combinations with some weighting coefficients. While the multi-mixing technique merely interpolates between existing samples and cannot explore beyond known patterns, AATM applies protocol-aware transformations that can generate novel attack vectors while maintaining protocol compliance. Then, the generated datasets undergo evaluation for balance and realism checks before being used in the ADS frameworks. The following subsections represent the mathematical modeling of the CGAN and proposed AATM techniques along with GOOSE guidelines, considering the BR and RR functions to check the validity of synthesized datasets.

2.1.1 CGAN Technique

The CGAN formulation for GOOSE message generation employs an adversarial approach, as shown in Eq. (10) [30,44]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y,c)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y,c)))] \tag{10}$$

Where:

- $z \sim \mathcal{N}(0, I_d)$: A d-dimensional random noise vector sampled from a standard normal distribution, serving as the randomization source for generating diverse GOOSE messages while maintaining desired characteristics. Each dimension influencing different aspects of the synthetic message.
- $y \in \{0,1\}^m$ is a vector representing the attack class (e.g., DI, RE attack, and DoS) or normal traffic, allowing the CGAN to generate class-conditional samples. For GOOSE messages, m typically ranges from $5-8$ depending on how many attack types are modeled.
- $c \in \{0,1\}^p$ is a binary vector encoding the GOOSE context. These contextual factors ensure generated messages reflect realistic operational scenarios within IEC61850 environments.
- $G : \mathbb{R}^d \times \{0,1\}^m \times \{0,1\}^p \to \mathbb{R}^{14}$: The generator function that transforms the noise vector, conditioned on attack class and protocol context, into a synthetic 14-dimensional GOOSE message containing all required features.
- $D : \mathbb{R}^{14} \times \{0,1\}^m \times \{0,1\}^p \to [0,1] \times \{0,1\}^8$: The discriminator function that evaluates whether a given GOOSE message appears realistic given its claimed attack class and protocol context; outputting a probability between 0 and 1 considering the GOOSE rules where higher values indicate the message appears genuine rather than synthetic.

The adversarial loss for statistical realism and the rule compliance loss for protocol validity constitute the total loss function as Eq. (11):

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{rules}\mathcal{L}_{rules} \tag{11}$$

Where the rule compliance loss has different weights based on the security importance, as shown in Eq. (12):

$$\mathcal{L}_{rules} = \sum_{i=1}^{8} w_i(1 - GR_i(G(z|y,c)))^2 \tag{12}$$

The adversarial loss ensures generated messages match the statistical properties of real GOOSE traffic, while the rule compliance loss (weighted at λ_{rules} = 8.0) enforces adherence to the eight GOOSE rules. The generator is directly penalized for producing non-compliant GOOSE messages via the weighted sum of squared errors. Conversely, the discriminator learns about rule compliance indirectly by comparing real rule-following samples to generated ones. Within the rule compliance term, individual weights (i.e., $0 < \omega_i < 2$) reflect each rule's security importance: critical integrity rules (GR#3, GR#4) receive the highest weights (2.0), RE and data manipulation rules (GR#2, GR#8) receive moderate-high weights ($1.5 - 1.8$), and formatting/timing rules receive lower weights ($0.8 - 1.0$). This formulation creates a balanced objective that prioritizes both realistic message generation and protocol compliance, with emphasis on security-critical constraints.

2.1.2 Proposed AATM Technique

This proposed approach employs a protocol-aware transformation function which is particularly useful for communication messages. As communication messages have a specific pattern with some subtle changes, a proper definition of perturbations can help to distinguish these small changes. These subtle changes can even happen in some numbers and/or letters in the different features (both numerical and categorical) which the CGAN technique might fail to understand; these variations, specifically in cases of the generation of realistic zero-day attacks and a high similarity between the normal datasets and some types of attacks (e.g., RE attacks) [52, 53]. The following information shows the process of different steps in the proposed method.

Vector Representation Suppose a GOOSE dataset is represented as:

$$x_i = [x_i^{num}, x_i^{cat}] \tag{13}$$

Where x_i^{num} represents 9 numerical features (e.g., time, appid, and stNum) and x_i^{cat} represents 5 categorical features (e.g., DM, SM, and type). In the numerical part, three different functions including protocol compliance, balance, and novelty functions are presented in Eqs. (14)- (16) as follows:

$$f_{protocol}(x_i) = \sum_{j=1}^{m} w_j \cdot GR_j(x_i) \tag{14}$$

$$f_{balance}(x_i) = -\sum_{c \in C} \left| p_c - \frac{1}{|C|} \right| \tag{15}$$

$$f_{novel}(x_i) = \min_{x_j \in X} d(x_i, x_j) \tag{16}$$

The protocol compliance function, $f_{protocol}^{num}(x_i)$, uses weighted rule compliance metrics $GR_j(x_i)$ to guide perturbations toward protocol-valid regions, ensuring attacks remain credible by preserving necessary relationships between features. The balance function, $f_{balance}(x_i)$, calculates the negative sum of deviations between current class proportions, p_c, and ideal uniform distribution, $\frac{1}{|C|}$, directing mutations toward minority attack classes to create diverse datasets with equal representation across attack types; and the novelty function, $f_{novel}(x_i)$, calculates the minimum distance between a candidate sample and all training examples, encouraging exploration of unexplored feature space regions to discover attack vectors not present in existing datasets, thereby enhancing the model's ability to generate previously unknown attack patterns that could evade traditional detection systems while maintaining structural validity. The selection of weights, w_j, for the eight GOOSE protocol rules is crucial for effective attack generation and follows a priority-based approach: higher weights $(0.15 - 0.2)$ are assigned to critical sequence rules (i.e., stNum & sqNum) and timestamp validation as these are fundamental to the protocol's integrity and frequently monitored by ADSs. The medium weights (0.1) are given to APPID compliance and MAC address validity as they represent network-level identifiers that must appear legitimate. This weighted approach allows the proposed framework to generate attacks that target specific vulnerabilities while maintaining sufficient protocol compliance to avoid simple detection, creating attack effectiveness.

Gradient Computation The base gradient for perturbation in the numerical part can be mentioned as Eq. (17) which states a combination of presented functions along with different hyperparameters.

$$\delta_{base}^{num}(x_i) = \alpha \cdot \nabla_x f_{protocol}(x_i) + \beta \cdot \nabla_x f_{balance}(x_i) + \gamma \cdot \nabla_x f_{novel}(x_i) \tag{17}$$

The hyperparameters α, β, and γ in this equation are typically set to $\alpha = 0.4$, $\beta = 0.3$, and $\gamma = 0.3$, respectively. This weighting prioritizes protocol validity ($\alpha = 0.4$) to ensure attacks remain credible within the IEC61850 framework, while equally distributing the remaining influence between class balance and novelty exploration ($\beta = \gamma = 0.3$). These values can be adjusted based on specific attack objectives. For example, an increased β is required when greater attack diversity is needed, or raising γ is needed when novel attack discovery is prioritized over keeping the protocol compliance. The zero-day perturbation formulation represents the core innovation of the AATM approach, as stated in Eq. (18).

$$\delta_{zero}^{num}(x_i) = \delta_{base}^{num}(x_i) - \lambda \cdot \sum_{r \in R_{target}} \nabla_x GR_r(x_i) \tag{18}$$

According to this, the base perturbation is modified by subtracting the weighted gradient of targeted protocol rules, with λ controlling the violation strength and R_{target} specifying which rules to deliberately violate; this approach effectively pushes samples away from compliance with selected rules while maintaining overall protocol validity. The new numerical features are then generated using the projection function (Eq. (19)) which applies the zero-day perturbation to the original numerical features and projects the result into the valid numerical feature space through P^{num}.

$$x_{new}^{num} = P^{num}(x_i^{num} + \delta_{zero}^{num}(x_i)) \tag{19}$$

This ensures that perturbed values maintain protocol compliance except for the deliberately targeted violations, producing realistic attack vectors that specifically exploit the targeted protocol vulnerabilities.

Categorical Feature Processing A categorical transition matrix, $T_{r,j}$ to show how feature j should change to affect rule r as Eq. (20):

$$T_{r,j} = \begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,5} \\ T_{2,1} & T_{2,2} & \cdots & T_{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ T_{m,1} & T_{m,2} & \cdots & T_{m,5} \end{bmatrix} \tag{20}$$

Next, the categorical mutations can be defined as Eq. (21) to generate zero-day attacks in terms of non-numerical parts.

$$M_{zero}(x_i) = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_5 \end{bmatrix} \tag{21}$$

Where each m_j is calculated based on Eq. (22), and the base mutation combines protocol, balance, and novelty objectives:

$$m_j = \text{base\_mutation}_j - \lambda \cdot \sum_{r \in R_{target}} T_{r,j} \tag{22}$$

$$\text{base\_mutation}_j = \alpha \cdot \text{protocol\_mut}_j + \beta \cdot \text{balance\_mut}_j + \gamma \cdot \text{novelty\_mut}_j \tag{23}$$

Now, a new categorical feature vector can be given according to Eq. (24) that describes how categorical features in GOOSE messages are mutated during zero-day attack generation. It works by first finding the index position of the current categorical values, applying the calculated mutation value, and then selecting the corresponding categorical value at the new index position while ensuring it remains within the valid set of possible values through the modulo operation.

$$x_{new,j}^{cat} = C_j[(I_j(x_{i,j}^{cat}) + \text{Round}(m_j)) \mod |C_j|]_{1\times 2} \tag{24}$$

Where C_j is the set of valid values for the categorical feature j, and I_j maps the categorical value to its index. Further, "Round" converts the mutation to an integer value. Also, the modulo operation helps to ensure that the resulting index does not exceed the bounds of the array. Suppose that multiple values in the DM column are as follows:

C_{DM} = ["01 00 03," "01 00 04," "01 00 05"] with $|C_{DM}| = 3$.

- For a row with DM = "01 00 03" and $m_{DM} = 1.5$:

- I_{DM}("010003") $= 0$

- $Round(m_{DM}) = 2$

- New index $= 0 + 2 = 2$

- $x_{new,DM} = C_{DM}[2]$ = "010005"

However, if $m_{DM} = 2.5$:

- Round(m_{DM}) $= 3$

- New index $= 0 + 3 = 3$

This would be an error since the valid indices are only 0, 1, and 2. This is why the modulo operation is important to bound the indices to the available size. Hence, the complete process of the proposed technique can be summarized as follows to clearly show different steps in plain language.

1. Split the original GOOSE message into numerical and categorical components
2. Compute numerical perturbation $\delta_{zero}^{num}(x_i)$
3. Compute categorical mutations $M_{zero}(x_i)$
4. Apply perturbation to numerical features
5. Apply mutations to categorical features
6. Combine into final vector $x_{new} = [x_{new}^{num}, x_{new}^{cat}]$

This process demonstrates the AATM's ability to precisely target specific protocol rules for violation while maintaining overall message validity, creating attacks that are challenging to detect. The next part shows the validation process including the BR and RR for further processing in the first layer of this framework.

2.1.3 Validation Framework

A validation framework focusing exclusively on two critical quality metrics, including BR and RR, is presented to evaluate the quality of generated synthesized datasets.

- Balance Rate Assessment of Generated Datasets

The BR quantifies dataset balance using Eq. (25) that the first segment focuses on the relationship between the most extreme classes in the dataset, and the second segment shows the normalized Shannon entropy [54, 55].

$$BR(X) = \frac{1}{2}\left(\frac{\min_i(n_i)}{\max_i(n_i)} + E\right) \tag{25}$$

Where n_i is the number of samples in class i, $\frac{\min_i(n_i)}{\max_i(n_i)}$ is the inverse of the imbalance ratio (ranges from 0 to 1), $E = -\frac{\sum_{i=1}^{K} p_i \log_2(p_i)}{\log_2(K)}$ is the normalized Shannon entropy, $p_i = \frac{n_i}{\sum_{j=1}^{K} n_j}$ is the proportion of samples in class i and K is the total number of classes. The first part directly measures how much smaller your least represented class is compared to your most common class. This is particularly important for GOOSE message security applications because it highlights if any attack class is significantly infrequent compared to normal traffic. While the first part only looks at the extremes, Shannon entropy examines the distribution across all classes simultaneously. It measures how evenly distributed the samples are across every class, not just the smallest and largest ones. Also, $BR(X)$ has a range of $0 < BR(X) \leq 1$ which 0 and 1 show the completely imbalanced and perfectly balanced datasets, respectively. Hence, this combination shows two key aspects of class balance in terms of the ratio between least and most frequent classes, addressing extreme imbalances as well as the overall diversity of the distribution, capturing how evenly samples are distributed across all classes.

- Realism Rate Assessment of Generated Datasets

For synthesized GOOSE datasets, an RR can be presented that evaluates protocol compliance through essential rule verification as shown in Eq. (26) where $GR_i(x)$ evaluates compliance with each essential GOOSE protocol rule.

$$RR(x) = \prod_{i=1}^{8} GR_i(x) \tag{26}$$

Moreover, Eq. (27) expresses the scoring mechanism for each protocol rule using function $GR_i(x)$, which assigns a value of 1 when rule i is completely satisfied, thus signifying perfect compliance, while an exponential penalty is applied in cases of rule violation.

$$GR_i(x) = \begin{cases} 1, & \text{if rule } i \text{ is fully satisfied} \\ e^{-\lambda_i \cdot V_i(x)}, & \text{Otherwise} \end{cases} \tag{27}$$

$V_i(x)$ represents the normalized severity of the violation (scaled between 0 and 1) and λ_i is the importance weight assigned to that specific rule. This formulation effectively creates a continuous scoring mechanism that severely

penalizes violations of critical rules (those with higher λ_i values) while allowing minor deviations in less critical aspects, making it particularly suitable for evaluating synthesized GOOSE messages [56].

$$\lambda_i = \begin{cases} 5, & \text{for critical integrity rules: } GR_3, GR_8 \\ 3, & \text{for important rules: } GR_1, GR_2, GR_4 \\ 2, & \text{for structural rules: } GR_5, GR_7 \\ 1, & \text{for timing pattern rules: } GR_6 \end{cases}$$

The violation severity, $V_i(x)$, provides a standardized illustration for quantifying deviations from GOOSE protocol rules, with each measure ($V_1(x) - V_8(x)$) targeting a specific aspect of message integrity as shown in Table 2. $V_1(x)$ assesses sqNum correctness, $V_2(x)$ detects DI patterns, $V_3(x)$ monitors stNum consistency, $V_4(x)$ measures unexpected field changes as a proportion, $V_5(x)$ validates timestamp formatting, $V_6(x)$ evaluates message frequency anomalies on a continuous scale, $V_7(x)$ quantifies communication gaps relative to acceptable thresholds, and $V_8(x)$ identifies RE attack signatures. In this context, X represents a single dataset (a group of messages), and $|X|$ is the number of messages

Table 2: The violation severity measures for GOOSE protocol rules.

Measure	Description	Value
$V_1(x)$	sqNum increment violation	Binary: 1 if sqNum does not increment correctly, 0 otherwise.
$V_2(x)$	Data change with unchanged stNum violation	Binary: 1 if data change logic is violated, 0 otherwise.
$V_3(x)$	stNum monotonicity violation	Binary: 1 if stNum decreases inappropriately, 0 otherwise.
$V_4(x)$	Field integrity violation	$\frac{\text{(Number of unexpectedly changed critical features)}}{\text{(Total number of critical features)}}$
$V_5(x)$	Timestamp format violation	Binary: 1 if timestamp format is invalid, 0 otherwise.
$V_6(x)$	Message frequency violation	$\min(1, \frac{\text{Observed frequency}}{\text{Maximum threshold}} - 1)$
$V_7(x)$	Temporal gap violation	$\min(1, \frac{\text{Gap duration} - 10s}{30s})$ for gaps > 10s
$V_8(x)$	RE attack indicator	Binary: 1 if data changes without stNum increment and sqNum reset, 0 otherwise

within that specific dataset. Hence, the aggregate realism can be given as Eq. (28):

$$RR(X) = \frac{1}{|X|} \sum_{x \in X} RR(x) \tag{28}$$

In which $RR(X) \geq 0.95$ shows the excellent realism of the generated synthesized dataset. This methodology offers a robust and compelling strategy for assessing the authenticity of generated GOOSE datasets, all while ensuring adherence to fundamental protocol compliance standards is not compromised. To recap, this section presented the data pre-processing, a generation of normal and zero-day attacks, and data post-processing according to the balance and realism assessments. Different methods, including the proposed AATM technique, are represented according to the application for GOOSE messages in digital substations considering the rules and data features. The purpose of the upcoming sections is to use these pre-processed datasets for an AD based on the proposed frameworks.

3 GenAI-based Anomaly Detection Systems in Digital Substations

ML-based ADSs have served as the basis for identifying anomalies within IEC61850 message frameworks. Although these approaches are known for their accuracy and reliance on data, they face a substantial limitation. Specifically, with the emergence of novel attack patterns (i.e., zero-day attacks), the models necessitate re-training. This requirement for model re-training is resource-intensive and time-consuming, introducing periods of vulnerability during which the system may not be equipped to cope with these unforeseen threats until they are integrated into the model's knowledge repository [3]. Conversely, GenAI tools present a more flexible and versatile strategy. Differing from traditional ML models, GenAI systems possess the ability to comprehend context, thereby enabling them to identify and address emerging threats without the need for prior explicit training. This capability of contextual comprehension reduces the necessity for constant updates and retraining in the rapidly changing landscape of cyber threats. By interpreting and accommodating new data, GenAI tools offer a robust and effective approach to AD in digital substations, leveraging

NLP capabilities [15]. According to the challenges mentioned and the comprehensive literature surveys, a GenAI-based ADS framework considering the ToD system is presented in [5] to show its performance evaluations.

Moreover, this section identifies critical sources for the dataset used in training the GenAI-enhanced ADS, along with the criteria applied for selecting these datasets. Compared to publicly accessible datasets, those sourced from the HIL testbed deliver high-resolution, authentic data that accurately represents actual substation operations [13]. However, it is challenging to have balanced and realistic datasets, without the existence of zero-day attacks. Hence, this paper evaluates the performance of the proposed ADS with 5,000 GOOSE datasets generated using the proposed AATM technique, which cover a wide range of classes including the normal and abnormal scenarios, and the BR and RR of the generated datasets are better than the CGAN approach.

3.1 GenAI-based Task-Oriented Dialogue ADS

The GenAI-powered ToD system's architectural framework incorporates advanced computational strategies through its hierarchical processing structure as proposed in our previous research [5]. The full training (FT) configuration of this framework is demonstrated in Fig. 3 that considers all GOOSE rules. It is designed to optimally leverage the structured communication protocols prevalent in digital substation environments. This system's belief state is mathematically

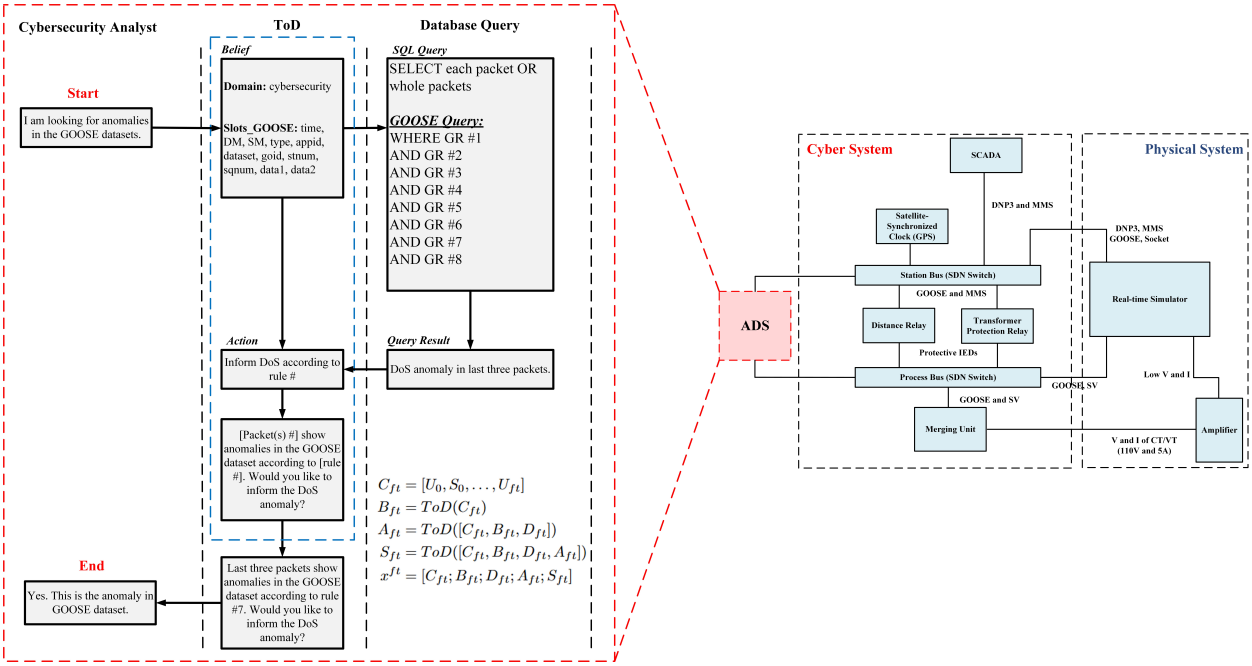


Figure 3: The proposed GenAI-based ToD ADS framework in the full training levels considering all GOOSE rules.

represented by aggregated sequences of packets and system states and is iteratively refined through rule-based SQL query executions that assess compliance to predefined operational constraints, including temporal synchronization, sqNum progression, and data integrity verification protocols [57, 58]. By integrating adaptive validation tools characterized by a hybrid function that includes belief states, dynamic conditions, and evolving rule sets, the system effectively distinguishes between legitimate operational behaviors and anomalies within the IEC61850 communication network.

The framework efficiently interacts with CPS components through advanced interfacing techniques, which facilitate its integration with supervisory control systems. This enables real-time coordination of identified anomalies with responses from protective relays within the power grid setup. A continuous learning approach, utilizing iterative feedback loops that incorporate both detection outcomes and scenario simulations, progressively refines classification boundaries while preserving the computational efficiency necessary for urgent protection strategies. This detailed training approach guarantees that the system's query processing mechanism, which evaluates multi-conditional rules, attains an ideal balance between detection sensitivity and specificity. Therefore, it reinforces a strong security barrier for the protection of critical infrastructure, as indicated by performance metrics that surpass those of conventional detection methods. More information regarding the parameters and detailed analysis of different blocks is expanded in [5]. The intention of this paper is to evaluate the FT level of the GenAI-based ToD for the AD process, trained by the proposed AATM-generated GOOSE datasets, which are more realistic and balanced. Also, the previous research

assessed the performance of different HITL and ToD in various GPT tools to show the better performance, in which the ToD framework implemented in Anthropic Claude Pro had the best performance in terms of the evaluation metrics as well as the accuracy. Hence, this paper uses the GenAI-based ToD framework implemented in Anthropic Claude Pro to make a comparison with other ML-based ADSs.

3.2 ML-based ADSs vs. GenAI-based ADSs

Traditional ML approaches, including FNNs, RNNs, and SVM, have shown promising results in detecting anomalies in IEC61850-based multicast messages. Hence, this part shows the modeling of these ML algorithms, particularly for GOOSE messages, as well as the general GenAI-based AD model. However, the emergence of GenAI technologies, particularly transformer-based architectures, has opened new possibilities for the AD concept. Hence, a general mathematical modeling of these ML algorithms, in addition to the GenAI-based system, is given below. More information and discussion about the results of the AD process considering different classes and their performance metrics are provided in the next section [59].

3.2.1 Feedforward Neural Network (FNN)

represents the fundamental architecture of DL, consisting of multiple layers of interconnected neurons where information flows one way from input to output without cycles or loops. In the context of the GOOSE AD process, FNNs are configured with an input layer that accepts normalized feature vectors extracted from GOOSE messages (e.g., timing parameters, sqNum, and data values), followed by multiple hidden layers with non-linear activation functions (typically ReLU), and an output layer with softmax activation for multi-class classification [60]. According to the application of the AD process in GOOSE messages, given an input vector $\mathbf{x} \in \mathbb{R}^n$, the FNN computes the output through successive layer transformations as Eq. (29):

$$\mathbf{h}^{(l)} = f^{(l)}(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}) \tag{29}$$

where $\mathbf{h}^{(l)}$ represents the activation of layer l, $\mathbf{W}^{(l)}$ is the weight matrix, $\mathbf{b}^{(l)}$ is the bias vector, and $f^{(l)}$ is the activation function. For the AD process, the network is trained to minimize the reconstruction error through the loss function in Eq. (30):

$$\mathcal{L}_{FNN} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 + \lambda \sum_{l=1}^{L} \|\mathbf{W}^{(l)}\|_F^2 \tag{30}$$

where $\hat{\mathbf{x}}_i$ is the reconstructed input and λ controls regularization, thus,the anomaly score can be computed as Eq. (31):

$$A_{FNN}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \tag{31}$$

3.2.2 Recurrent Neural Network (RNN)

is designed to process sequential data by forming directed graphs between nodes across time steps, enabling them to capture temporal patterns and maintain internal states. These networks consist of three types of nodes—input nodes that receive external data, output nodes that produce results, and hidden nodes that transform information. RNNs have proven particularly effective for applications requiring temporal understanding. According to the application of this research, RNNs process GOOSE messages as sequences, maintaining a "memory" of previous messages [19, 61]. For each message at time t, the hidden state update can be given as Eq. (32):

$$h_t = \tanh(W_h \cdot h_{t-1} + W_x \cdot x_t + b) \tag{32}$$

where h_t is the current hidden state, h_{t-1} is the previous state, and x_t is the current input. To handle long sequences, a long-short term memory (LSTM) with three gates can be employed as Eq. (33):

$$\begin{aligned} \text{Forget gate: } f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ \text{Input gate: } i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \text{Output gate: } o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \end{aligned} \tag{33}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. Then, the final hidden state is used for classification as the AD process, shown in Eq. (34):

$$P(\text{anomaly}) = \sigma(W_{out} \cdot h_T + b_{out}) \tag{34}$$

3.2.3 Support Vector Machine (SVM)

represents a supervised learning algorithm employed primarily for data classification tasks. The fundamental objective of SVM implementation involves achieving precise categorization of previously unseen data through the optimization of a decision boundary that reduces misclassification rates. This methodology operates through a two-phase process; initially, the algorithm undergoes training using labeled datasets to establish optimal parameters, followed by the application of the trained model to generate class predictions for new, unlabeled instances [62, 63]. GOOSE messages in IEC61850 networks can be monitored for anomalies using One-Class SVM. Given GOOSE message features $\mathbf{x}$, this algorithm learns a decision boundary around normal GOOSE behavior such that the optimization problem is as Eq. (35):

$$\min_{\mathbf{w},\xi,\rho} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \tag{35}$$

subject to:

$$\begin{gathered}\mathbf{w}^T\phi(\mathbf{x}_i) \geq \rho - \xi_i \\ \xi_i \geq 0\end{gathered} \tag{36}$$

The weight vector $\mathbf{w} \in \mathbb{R}^d$ defines the orientation of the separating hyperplane in the feature space, while the offset parameter ρ determines the hyperplane's distance from the origin. The slack variables $\xi_i \geq 0$ enable soft-margin classification by allowing some training points to lie within the margin or on the wrong side of the decision boundary, providing robustness against outliers. The parameter $\nu \in (0, 1]$ serves as a user-defined regularization constant that controls the trade-off between maximizing the margin and minimizing the fraction of outliers, effectively determining the upper bound on the fraction of training errors and the lower bound on the fraction of support vectors. The kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ computes the similarity between data points in the transformed feature space, with the RBF kernel parameter $\gamma > 0$ controlling the influence radius of each support vector; smaller values create smoother decision boundaries and larger values allow more complex, localized boundaries. Finally, the Lagrange multipliers $\alpha_i \geq 0$ determine the contribution of each training sample to the final decision function, with non-zero values identifying the support vectors that define the decision boundary. Using an RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, the decision function becomes as Eq. (37). For a new GOOSE message, if $f(\mathbf{x}) < 0$, it is classified as an anomaly.

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i, \mathbf{x}) - \rho\right) \tag{37}$$

3.2.4 GenAI-based ADSs

leverage self-attention mechanisms to capture complex temporal dependencies in GOOSE message sequences. Given a sequence of GOOSE messages $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the model learns normal communication patterns through attention-based reconstruction [15, 64]. The architecture consists of an encoder-decoder transformer with positional encoding. The multi-head self-attention mechanism computes as Eq. (38):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{38}$$

where queries Q, keys K, and values V are linear projections of the GOOSE features. For the AD process, the reconstruction loss can be expressed as Eq. (39):

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 + \lambda\sum_{h=1}^{H}\text{Entropy}(A_h) \tag{39}$$

where $\hat{\mathbf{x}}_i$ is the reconstructed GOOSE message and A_h represents attention weights from head, h. The entropy term encourages focused attention patterns. The anomaly score combines reconstruction error and attention anomaly that is represented in Eq. (40):

$$A(\mathbf{x}) = \alpha\|\mathbf{x} - \hat{\mathbf{x}}\|^2 + (1-\alpha)\max_h \mathrm{KL}(A_h \| A_h^{normal}) \tag{40}$$

This approach excels at detecting temporal anomalies (i.e., unusual message sequences), contextual anomalies (messages inconsistent with historical patterns), and attention-based anomalies where the model's attention deviates from learned normal patterns, providing interpretable detection through attention visualization. The next section shows the results and discussion according to the proposed framework.

4 Results and Discussion

This section presents the results and discussion considering the different layers of the proposed framework, along with the performance evaluation of the ADSs based on a comparison of the proposed methodology with other ML-based methods. Further, three advanced performance evaluation metrics (i.e., informedness, Markedness, and Matthews correlation coefficient - MCC) are provided for a comparison of ADSs in power system applications in addition to standard metrics.

4.1 Validation of Proposed AATM Methodology for GOOSE Data Generation

The empirical assessment of the proposed AATM methodology demonstrates significant advancements in the capabilities for generating synthetic data when contrasted with the CGAN method. Figure 4 provides an extensive visualization of class distributions spanning three scenarios: the original GOOSE dataset, samples synthesized via the CGAN method, and data generated by the AATM technique. As can be seen, the original dataset exhibits significant class imbalance,

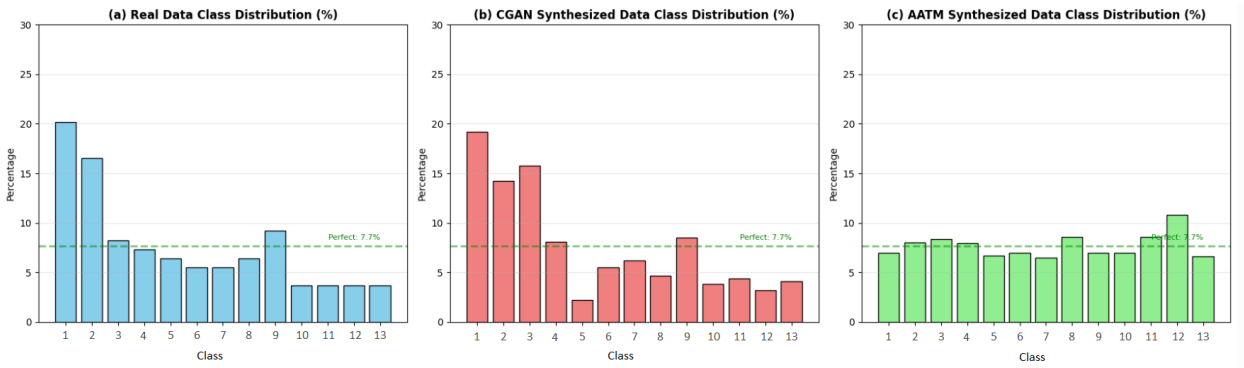


Figure 4: A representation of classes' distributions of GOOSE datasets considering different classes, (a) real data, (b) CGAN synthesized data, and (c) AATM synthesized data.

where prevalent categories such as "Normal" traffic (approximately 20%) and data injection ("DI") attacks (17%) significantly outweigh minority classes, notably "SP-dataset" errors which constitute merely 4% of the total samples. Such distributional disparities present fundamental challenges for developing robust ADSs capable of recognizing diverse attack/error patterns with equal effectiveness. The horizontal axis of the figure shows different classes including the attacks/errors which these classes are enumerated in Table 3. An examination of the synthetic data produced by the CGAN indicates an unforeseen intensification of the existing class imbalance within the original dataset. As demonstrated in Table 3, the CGAN model generates 19.2% of "Normal" class instances, whereas it only produces 2.2% of "SP-time" samples, thereby intensifying the distributional imbalance instead of alleviating it. This occurrence can be attributed to the inherent bias in CGAN models, where the generator network tends to replicate patterns that are dominant in majority classes due to their statistical prominence during the training phase. The "DOS" and "DI" attack classes also exhibit disproportionate representation at 15.8% and 14.2%, respectively, while essential minority classes such as "SP-dataset" and "Zero-day" attacks remain markedly underrepresented at 3.9% and 3.2%, respectively.

On the other hand, the AATM technique demonstrates exceptional capacity for generating synthetically balanced data distributions. The proposed method achieves remarkable uniformity across all 13 classes, with representation percentages restricted within a range of 6.5% to 10.8%. This balanced generation paradigm is particularly evident

Table 3: A distribution of different classes for CGAN- and AATM-generated GOOSE datasets.

	Class	CGAN Count	CGAN %	AATM Count	AATM %
1	Normal	961	19.2%	**350**	**7.0%**
2	DI	712	14.2%	**401**	**8.0%**
3	DOS	789	15.8%	**419**	**8.4%**
4	RE	403	8.1%	**396**	**7.9%**
5	SP-time	111	2.2%	**334**	**6.7%**
6	SP-DM	277	5.5%	**348**	**7.0%**
7	SP-SM	311	6.2%	**325**	**6.5%**
8	SP-type	233	4.7%	**430**	**8.6%**
9	SP-appid	424	8.5%	**349**	**7.0%**
10	SP-dataset	193	3.9%	**348**	**7.0%**
11	SP-goid	219	4.4%	**428**	**8.6%**
12	Packet Loss	206	4.1%	**331**	**6.6%**
13	Zero-day	161	3.2%	**541**	**10.8%**
	Total	5000	100.0%	5000	100.0%

in the transformation of traditionally infrequent categories: "SP-time" classes increase from 2.2% under CGAN to 6.7% with the AATM method, while "Zero-day" attacks experience a substantial enhancement from 3.2% to 10.8%. Simultaneously, overly dominant classes undergo appropriate reduction, with "Normal" traffic decreasing from 19.2% to 7.0%, thereby contributing to overall distributional balance. These distribution percentages show that the CGAN technique generated more "Normal" classes as it could not get all the correct patterns for this class. Also, some of the attacks/errors were mistakenly generated as they could be the "Zero-day" attacks. The quantitative assessment presented in Fig. 5 further validates the superiority of the AATM methodology through two critical metrics. The BR, an index that

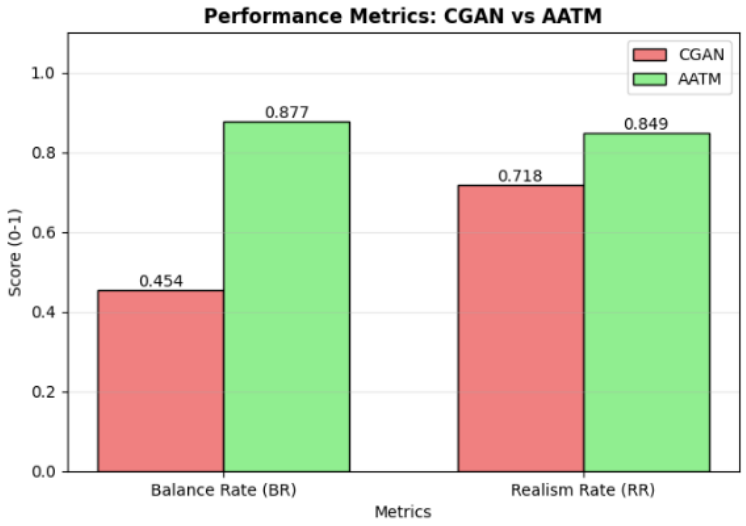


Figure 5: The BR and RR of CGAN- and proposed AATM-generated datasets.

quantifies the distributional uniformity across generated samples, demonstrates a notable increase from 0.454 in CGAN to 0.877 in AATM, reflecting a 93% enhancement in class balance. This substantial improvement underscores the AATM's effectiveness in addressing the core challenge of generating representative samples across all classes, including attacks/errors and normal data, regardless of their frequency in the training dataset. Furthermore, RR, which evaluates the credibility and quality of synthesized samples, shows an advancement from 0.718 to 0.849, signifying an 18% enhancement in the fidelity of generated data. This simultaneous progress in both BR and RR underscores the AATM's capability in producing high-quality synthetic samples while preserving class balance. The implications of these findings reach far beyond simple enhancements in statistical metrics, providing substantial advantages for real-world cybersecurity implementations. ADSs that are trained on datasets with unbalanced distributions often demonstrate a reduced ability to accurately identify rare yet possibly disastrous attack paths. By generating balanced synthetic datasets, the AATM technique facilitates the creation of detection models that maintain an equitable performance across a wide array of threat signatures, achieving similar levels of detection accuracy. This characteristic proves

particularly valuable for zero-day attack detection, where AATM's generation of 10.8% samples compared to CGAN's 3.2% provides sufficient training instances for models to develop robust recognition capabilities for these critical yet infrequent threats. Despite these findings, it is important to evaluate certain limitations in the context of interpreting the results. The observed residual variation in class percentages produced by the AATM algorithm, which varies between 6.5% and 10.8%, illustrates that achieving an entirely uniform distribution remains an ongoing challenge yet to be completely resolved. Future research could investigate the scalability attributes of the AATM technique when applied to datasets with an expanded number of classes. Additionally, these studies might assess the feasibility of integrating constraints specific to particular domains to not only enhance the realism of samples but also to ensure a balanced distribution across the dataset.

4.2 Performance Evaluation Metrics in an AD Process

In this section, a comparative analysis is conducted to assess the performance and effectiveness of the proposed GenAI-based ToD framework over ML-based ADSs. Fig. 6 shows the different descriptions and formulations for standard and advanced evaluation metrics to make a comparison between these frameworks. These metrics represent

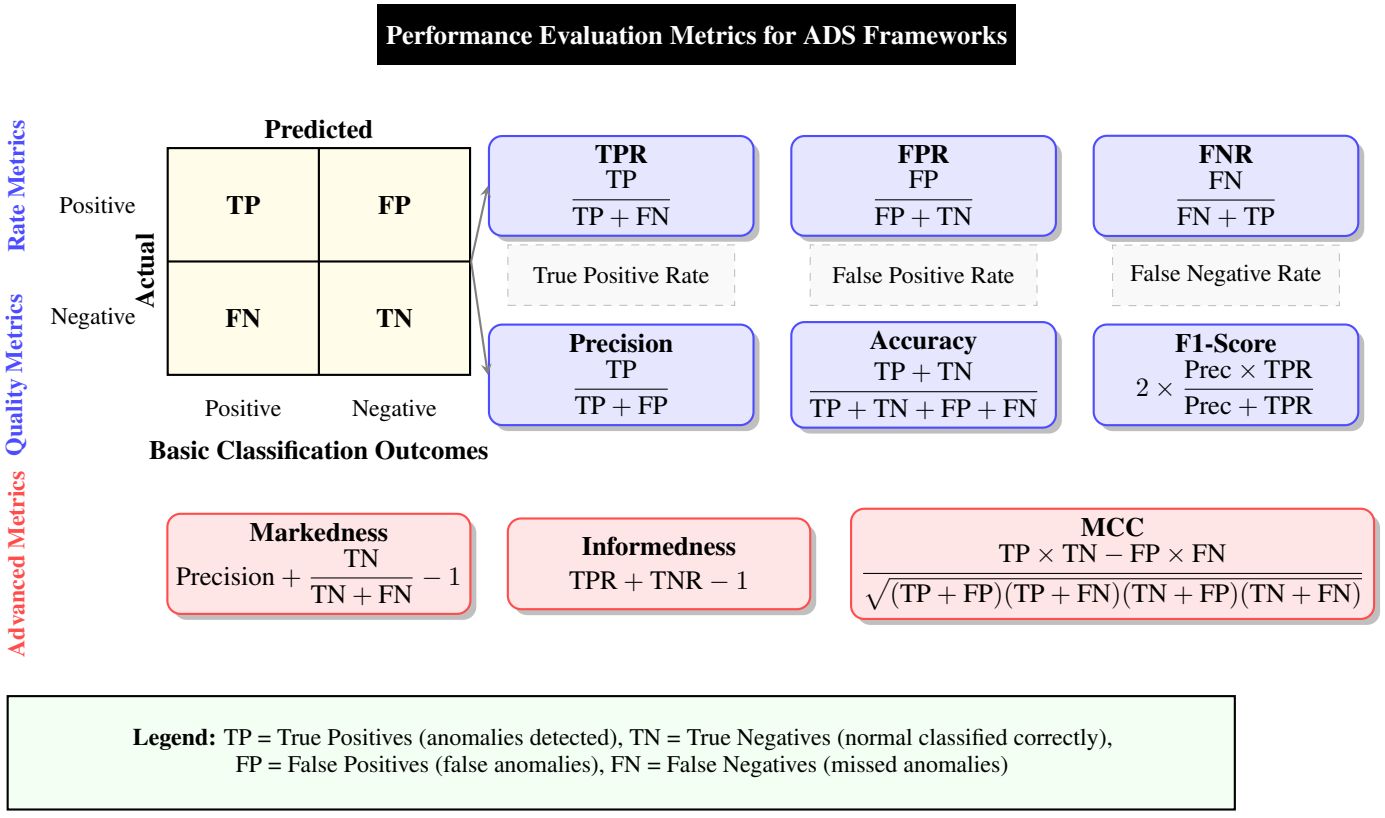


Figure 6: Performance evaluation metrics and their mathematical formulations for assessing ADS frameworks. The confusion matrix (left) shows basic classification outcomes, which are used to derive rate metrics (blue boxes) and advanced evaluation metrics (red boxes).

essential evaluation criteria for ADSs implementing security monitoring of GOOSE messages. The fundamental classification metrics include the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) indicators. In this context, these primary metrics quantify correctly identified anomalies, properly classified normal traffic, incorrectly flagged normal communications, and undetected anomalous events, respectively. The true positive rate (TPR), calculated as $\frac{TP}{TP+FN}$, quantifies the system's sensitivity in detecting actual anomalies. Conversely, the false positive rate (FPR), expressed as $\frac{FP}{FP+TN}$, measures the likelihood of false alarms, while the false negative rate (FNR), determined by $\frac{FN}{TP+FN}$, evaluates detection failures, a particularly critical metric for security applications in industrial control systems. The Precision, formulated as $\frac{TP}{TP+FP}$, indicates the reliability of positive predictions, while accuracy, calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, assesses overall classification correctness across the entire messages collection. The

F1-Score, expressed as $2 \times \frac{Precision \times TPR}{Precision + TPR}$, provides a harmonic mean of precision and TPR, offering a balanced assessment when both false alarms and missed detections possess significant operational implications in substations.

Within this work, advanced metrics such as Markedness, Informedness, and the MCC are utilized to evaluate the consistency, decision-making precision, and classification quality, ranging from -1 to 1. These metrics prove beneficial in the AD process to assess the efficiency and applicability of frameworks. Specifically, in the context of AD applied to GOOSE datasets, Markedness is indispensable for measuring the model's capability to reduce both false positives (FPs) and false negatives (FNs). A high Markedness value signifies a robust AD framework that minimizes erroneous alerts, thereby ensuring stability and optimal performance at substations by decreasing unnecessary disruptions. Meanwhile, Informedness measures the model's proficiency in recognizing dataset pattern variations that indicate anomalies. Particularly for AD scenarios using GenAI models — where real anomalies may occur infrequently — MCC is advantageous as it ensures the model's performance reflects its true efficacy and is not excessively affected by a larger class size [65].

4.3 A Comparative Analysis of the GenAI-based and ML-based ADSs

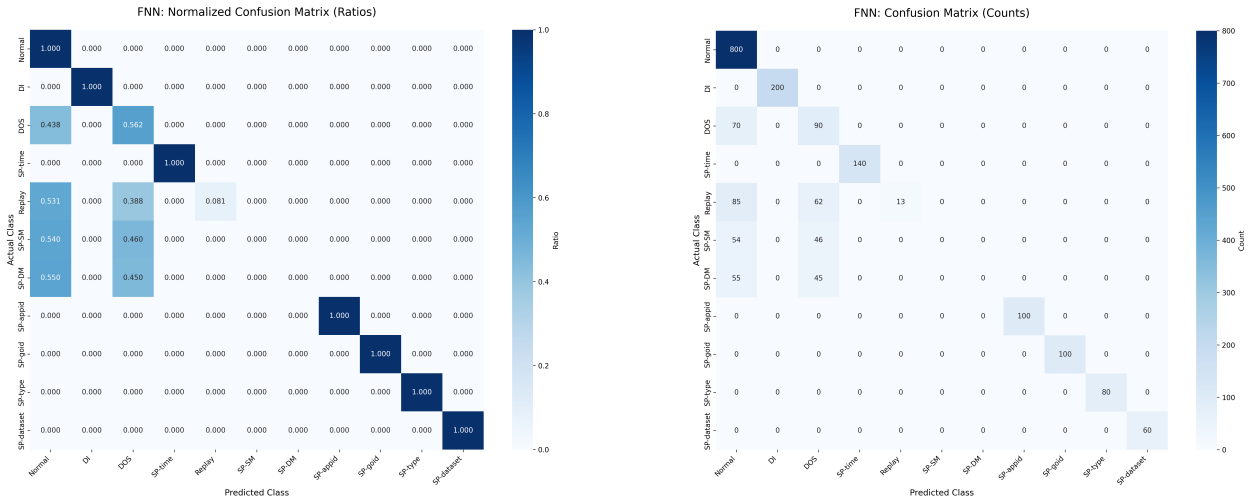
This section presents a comparative analysis of four ADSs considering GOOSE datasets generated using the proposed AATM technique. The comparison includes three ML algorithms — FNN, RNN, SVM, in addition to a GenAI-based ADS. The empirical results provide robust evidence of the superior effectiveness exhibited by the GenAI-based ADS, especially when evaluated against ML models. It is presented in Table 4 that the proposed GenAI approach achieves an outstanding classification accuracy rate of 97.9%. This table demonstrates a significant enhancement in performance

Table 4: A comparison of GenAI- and ML-based ADSs using AATM-generated GOOSE datasets.

Algorithms	FNN	RNN	SVM	Anthropic Claude Pro (GenAI-based ADS)
Standard Metrics				
TPR	79%	87.9%	79.1%	**97.9%**
FPR	**0%**	10.6%	**0%**	3.2%
FNR	21%	12.08%	20.9%	**2.1%**
Precision	**100%**	92.5%	**100%**	97.9%
Accuracy	87.4%	88.5%	87.4%	**97.5%**
F1-Score	88.3%	90.2%	88.3%	**97.9%**
Advanced Metrics				
Markedness	0.76	0.756	0.761	**0.947**
Informedness	0.79	0.773	0.791	**0.947**
MCC	0.775	0.764	0.776	**0.945**

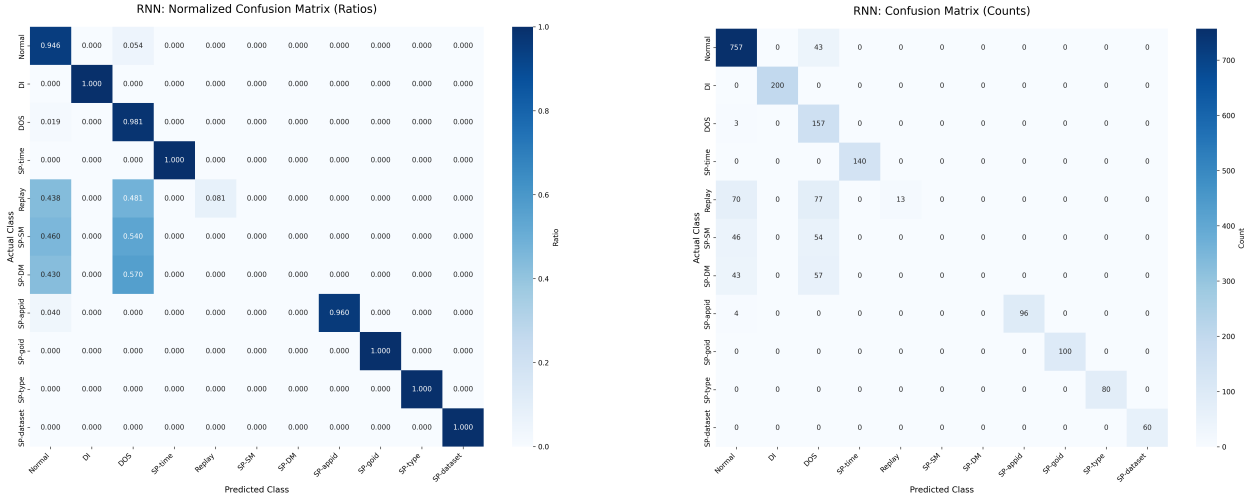
metrics compared to ML models, such as FNN, which achieves an accuracy rate of 87.4%, RNN achieving 88.5%, and SVM, also at 87.4%. The observed improvement (approximately a 10 percentage point increase in the accuracy metric) represents a significant advancement in the domain of GOOSE messages monitoring. Further, this proposed framework showed better performance in almost all other metrics. A thorough analysis of the confusion matrices demonstrated in Figs. 7 – 12 offers a comprehensive understanding of the classification features displayed by different approaches. The GenAI framework's confusion matrices, i.e., Figs. 10 – 12 demonstrate prominent diagonal concentration with negligible inter-class confusion, reflecting robust differentiating capabilities. Notably remarkable is the normalized confusion matrix of the GenAI implementation (Fig. 11), which showcases near-unity values along the diagonal for the majority of classes, contrasting markedly with the substantial class overlap observed in ML approaches, especially when distinguishing between structurally similar class signatures.

The performance metric analysis indicates that although the SVM algorithm achieves a comparable FPR of 3.2% to the GenAI model, the latter distinguishes itself by simultaneously maintaining the minimal FNR of 2.1%, substantially lower than the 21%, 12.08%, and 20.9% observed for FNN, RNN, and SVM approaches, respectively. Such balanced error characteristics prove essential for practical deployment scenarios where both false alarms and undetected threats must be minimized. Additional performance metrics meticulously confirm the superiority of the GenAI framework. This system achieves Markedness and Informedness values reaching 0.947, a notable advancement well beyond the ML benchmarks typically observed between 0.76 and 0.791. Furthermore, it performs well in accurately classifying classes, effectively distinguishing between situations indicative of attack and those that are non-malicious, thereby demonstrating a high degree of precision in its operation. Furthermore, the MCC, recording at 0.945, provides additional validity to the comprehensive effectiveness of the model. This coefficient reflects an enhancement of approximately 20% compared to standard baseline methodologies, confirming GenAI's advanced capabilities and distinguished performance.



(a) (b)

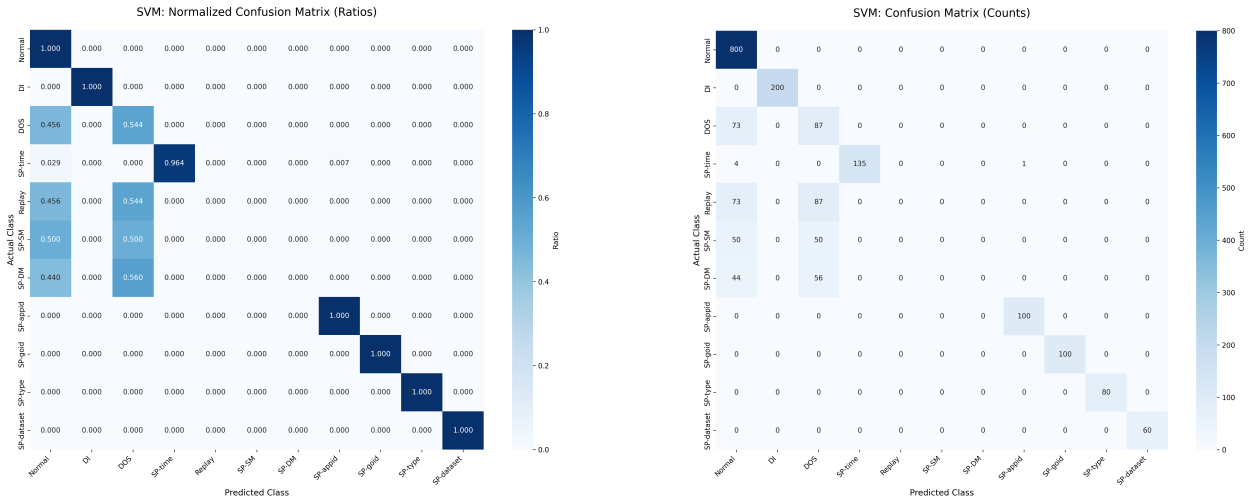
Figure 7: Confusion matrices of an FNN-based ADS, trained by the proposed AATM-generated GOOSE datasets, (a) normalized ratios (b) counts.



(a) (b)

Figure 8: Confusion matrices of an RNN-based ADS, trained by the proposed AATM-generated GOOSE datasets, (a) normalized ratios (b) counts.

Analysis of the multi-class confusion matrix (i.e., Fig. 12) clarifies the GenAI framework's classification capabilities across diverse attack/error categories. The system demonstrates remarkable competence in distinguishing subtle attack variations including DI, DOS, temporal anomalies (i.e., SP-time), feature-based manipulations (SP-[feature]), and RE attacks, with negligible inter-category confusion. This refined detection resolution originates from the framework's inherent capacity for semantic interpretation of message dynamics, as illustrated in the diagnostic output where specific message deviations and contextual anomalies are clearly identified. The GenAI model possesses an advanced capacity for semantic understanding, which facilitates its detection of complex attack/error patterns that are often missed by traditional statistical approaches. This performance is quantitatively supported by its Markedness and Informedness scores, both achieving a value of 0.947. The model's advantage in this domain stems from its context-based reasoning abilities concerning message violations, differing markedly from methods that depend solely on numerical pattern recognition. According to this concept, a textual portion of the "Response" for GenAI-based ADS in Anthropic Claude Pro is presented in the following box for more clarification.



(a) (b)

Figure 9: Confusion matrices of an SVM-based ADS, trained by the proposed AATM-generated GOOSE datasets, (a) normalized ratios (b) counts.

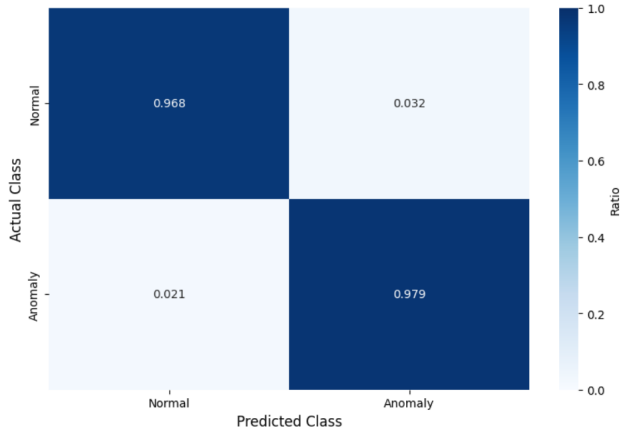


Figure 10: A normalized confusion matrix of the GenAI-based ADS, trained by the proposed AATM-generated GOOSE datasets considering the normal and anomalous classes.

Response

1. **Semantic Pattern Understanding Phase**
   I would analyze each dataset by:
   - **Reading the temporal flow** of messages within each dataset
   - **Understanding the protocol semantics** - recognizing that GOOSE messages follow specific IEC61850 patterns
   - **Contextual interpretation** of field relationships (DM/SM pairs, stNum/sqNum)
2. **Rule-Based Reasoning with Contextual Intelligence**
   Unlike rigid rule engines, I would apply the 8 rules with contextual understanding:
   Rule Application Strategy:
   - **Rule 1 & 3:** Track sqNum progression and stNum monotonicity within DM/SM contexts
   - **Rule 2 & 8:** Understand the semantic relationship between data changes and sqNum

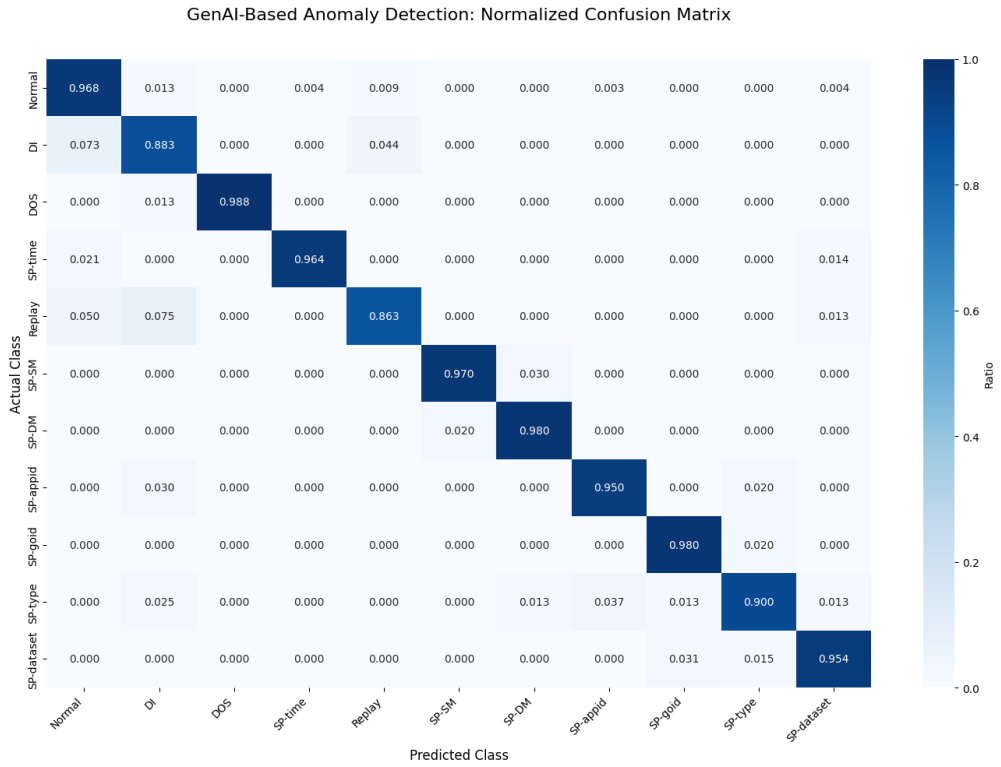


Figure 11: A normalized confusion matrix of the GenAI-based ADS, trained by the proposed AATM-generated GOOSE datasets for all classes.

- **Rule 4:** Detect configuration changes in categorical fields
- **Rule 5-7:** Analyze temporal patterns for DOS and timing anomalies

3. **Multi-Level Anomaly Classification Process**

   Level 1 - Pattern Recognition:

   ```
   For each dataset:
   1.  Identify message sequence patterns
   2.  Detect temporal anomalies (ms-level for DOS, second-level for SP-time)
   3.  Recognize protocol violations
   ```

   Level 2 - Semantic Classification:

   - **Normal:** All rules satisfied, expected GOOSE behavior
   - **DI:** StNum changes without proper data correlation
   - **DOS:** $>$ 10 messages within 10 microseconds
   - **SP-time:** $>$10 second gaps between messages
   - **SP-[feature]:** Unexpected changes in specific fields
   - **RE:** Repeated sequence patterns

   Level 3 - Contextual Validation:

   - Cross-reference patterns with known GOOSE protocol behavior
   - Apply semantic understanding of industrial control system operations

   Example Analysis Output:

   ```
   Dataset #1:  ANOMALY (DI Class)
   ```

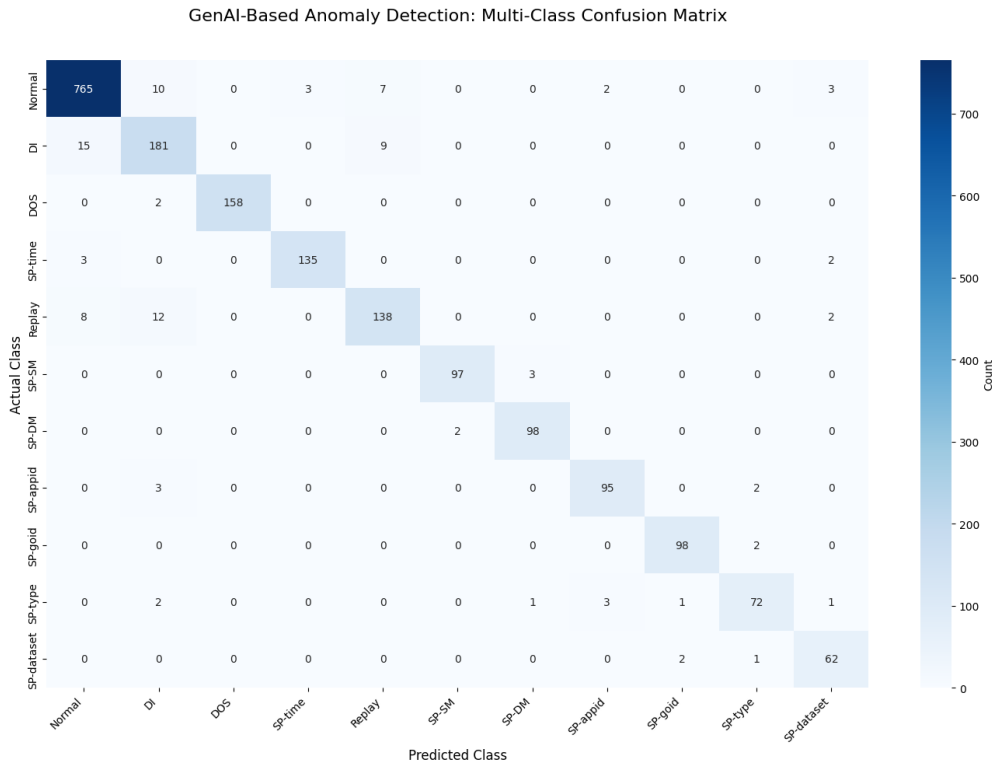


Figure 12: A confusion matrix (counts) of the GenAI-based ADS, trained by the proposed AATM-generated GOOSE datasets for all classes.

```
Reasoning: Detected stNum change from 27 to 28 at row 5 without corresponding
data1/data2 change. This violates the expected GOOSE protocol behavior where
stNum changes should correlate with data modifications.

Dataset #2: NORMAL
Reasoning: All sequences follow expected patterns. sqNum increments properly
(150→151→152), stNum remains constant at 27, no timing anomalies detected.

Dataset #3: ANOMALY (DOS Class)
Reasoning: Identified 12 messages within 8 microseconds (rows 3-14),
indicating a DOS pattern with abnormally rapid message transmission.
```

As can be observed, it can follow the provided rules in addition to the patterns in the datasets to provide a suitable reason according to each class. Some examples of outputs based on three datasets are represented at the end of the box.

5 Conclusions and Future Work

This section generally presents the AATM technique for balanced and realistic data generation, novel GenAI-based ADSs, and an enhanced performance of GenAI-based ADSs over ML-based ADSs. Initially, because there are insufficient datasets of IEC61850-based communications as well as the generation of realistic zero-day attacks, this research proposes a novel pre-processing technique known as AATM for data generation which is perturbation- and mutation-based to enhance the RR and BR of the generated synthesized datasets, which show enhancements over another pre-processing technique. Secondly, a GenAI-based ToD ADS is presented in IEC61850-based communication messages in digital substations which rather outperforms the traditional ML-based ADSs in terms of no necessity for the re-training process, less effort, and the ability of analysis of categorical features in multi-cast messages. Then, considering the generated datasets and suggested GenAI-based ADS, the performance of this ADS is assessed based on standard and advanced performance metrics. According to the results, it demonstrates that the GenAI-based ToD framework

implemented by Anthropic Claude Pro has superior performance compared with other ML models. Finally, the system's capacity for semantic comprehension, demonstrated through its proficiency in contextualizing message/pattern anomalies and detecting sophisticated attack/error signatures undetected by solely quantitative approaches, represents a qualitative evolution beyond traditional pattern recognition models. The framework's consistent high performance across diverse normal/threat vectors, including data manipulation, operational issues, temporal attacks, and message RE scenarios, in combination with its inherent scalability and sustainability features, validates its suitability for production deployment. These findings position GenAI as a crucial technology in securing critical infrastructure, providing a resilient, explainable, and evolutionary security solution capable of addressing emergent threats within digital substation environments while preserving the rigorous reliability requirements fundamental to electrical grid operations.

The development of security frameworks for substations opens up numerous promising future pathways. Present detection techniques must evolve to include the entire suite of IEC61850 communication protocols. This extension effort results in a comprehensive monitoring ecosystem, surpassing the constraints associated with single-protocol analysis. Tailored intelligent systems can be deployed within utility-managed infrastructure using community-developed language models. These deployments establish self-contained operational environments that satisfy rigorous regulatory compliance requirements. Power system physics-aware computational models facilitate instantaneous threat identification within these frameworks. This architectural approach naturally extends toward interconnected substation networks. Mechanisms for cryptographically protected information exchange facilitate joint threat evaluation among installations spread over various geographic locations while maintaining the self-governance of each facility. The system's intelligent pattern recognition abilities allow for the dynamic fine-tuning of detection criteria, accommodating evolving threat scenarios. Thus, these adaptive frameworks significantly reduce the need for human intervention while preserving visibility, which is crucial for securing critical infrastructure.

References

[1] 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.

[2] Chih-Che Sun, Adam Hahn, and Chen-Ching Liu. Cyber security of a power grid: State-of-the-art. *International Journal of Electrical Power & Energy Systems*, 99:45–56, 2018.

[3] Junho Hong, Tai-Jin Song, Hyojong Lee, and Aydin Zaboli. Automated cybersecurity tester for IEC 61850-based digital substations. *Energies*, 15(21):7833, 2022.

[4] Silvio E. Quincozes et al. Ereno: A framework for generating realistic IEC–61850 intrusion detection datasets for smart grids. *IEEE Transactions on Dependable and Secure Computing*, pages 1–15, 2023.

[5] Aydin Zaboli, Yong-Hwa Kim, and Junho Hong. An advanced generative AI-based anomaly detection in iec61850-based communication messages in smart grids. *IEEE Access*, 2025.

[6] Omar A Beg et al. A review of AI-based cyber-attack detection and mitigation in microgrids. *Energies*, 16(22):7644, 2023.

[7] Junho Hong and Chen-Ching Liu. Intelligent electronic devices with collaborative intrusion detection systems. *IEEE Transactions on Smart Grid*, 10(1):271–281, 2017.

[8] Ying Chen, Junho Hong, and Chen-Ching Liu. Modeling of intrusion and defense for assessment of cyber security at power substations. *IEEE Transactions on Smart Grid*, 9(4):2541–2552, 2016.

[9] OpenAI. Chatgpt. `https://openai.com/chatgpt`. Accessed: Feb. 2024.

[10] Anthropic. Anthropic Claude Pro. `https://www.anthropic.com`, 2023. Accessed: 2024-11-10.

[11] Microsoft Corporation. Microsoft Copilot AI. `https://copilot.microsoft.com/`, 2023. Accessed: Feb. 1, 2024.

[12] Mikhail Smolin. Gencoder: A generative AI-based adaptive intra-vehicle intrusion detection system. *IEEE Access*, 2024.

[13] Aydin Zaboli, Seong Lok Choi, Tai-Jin Song, and Junho Hong. Chatgpt and other large language models for cybersecurity of smart grid applications. In *2024 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2024.

[14] Aydin Zaboli, Seong Lok Choi, and Junho Hong. Leveraging conversational generative AI for anomaly detection in digital substations. *arXiv preprint arXiv:2411.16692*, 2024.

[15] Sukhpal Singh Gill and Rupinder Kaur. ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3:262–271, 2023.

[16] Chee-Wooi Ten, Junho Hong, and Chen-Ching Liu. Anomaly detection for cybersecurity of the substations. *IEEE Transactions on Smart Grid*, 2(4):865–873, 2011.

[17] In-Sun Choi, Junho Hong, and Tae-Wan Kim. Multi-agent based cyber attack detection and mitigation for distribution automation system. *IEEE Access*, 8:183495–183504, 2020.

[18] Philipp Kreimel et al. Anomaly detection in substation networks. *Journal of Information Security and Applications*, 54:102527, 2020.

[19] Xuelei Wang et al. Anomaly detection for insider attacks from untrusted intelligent electronic devices in substation automation systems. *IEEE Access*, 10:6629–6649, 2022.

[20] Syed RB Alvee, Bohyun Ahn, Taesic Kim, Ying Su, Young-Woo Youn, and Myung-Hyo Ryu. Ransomware attack modeling and artificial intelligence-based ransomware detection for digital substations. In *2021 6th IEEE Workshop on the Electronic Grid (eGRID)*, pages 01–05. IEEE, 2021.

[21] Manikant Panthi. Anomaly detection in smart grids using machine learning techniques. In *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 220–222. IEEE, 2020.

[22] Ankitdeshpandey and R Karthi. Development of intrusion detection system using deep learning for classifying attacks in power systems. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2019*, pages 755–766. Springer, 2020.

[23] Ahmad Eynawi, Aneeqa Mumrez, Ghada Elbez, and Veit Hagenmeyer. Machine learning-based feature selection for intrusion detection systems in IEC 61850-based digital substations. In *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7. IEEE, 2024.

[24] Vagner E Quincozes, Silvio E Quincozes, Célio Albuquerque, Diego Passos, and Daniel Mossé. Feature extraction for intrusion detection in IEC-61850 communication networks. In *2022 6th Cyber security in networking conference (CSNet)*, pages 1–7. IEEE, 2022.

[25] Devika Jay. Deception technology based intrusion protection and detection mechanism for digital substations: a game theoretical approach. *IEEE Access*, 11:53301–53314, 2023.

[26] Souradeep Bhattacharya, Nazmus Saqib, and Manimaran Govindarasu. ML-based anomaly detection system for IEC 61850 communication in substations. In *2024 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2024.

[27] Darshana Upadhyay, Jaume Manero, Marzia Zaman, and Srinivas Sampalli. Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids. *IEEE Transactions on Network and Service Management*, 18(1):1104–1116, 2020.

[28] Ruoxi Zhu, Chen-Ching Liu, Junho Hong, and Jiankang Wang. Intrusion detection against mms-based measurement attacks at digital substations. *IEEE Access*, 9:1240–1249, 2020.

[29] Taha Selim Ustun et al. Machine learning-based intrusion detection for achieving cybersecurity in smart grids using IEC 61850 GOOSE messages. *Symmetry*, 13(5):826, 2021.

[30] Lixiang Yuan, Siyang Yu, Zhibang Yang, Mingxing Duan, and Kenli Li. A data balancing approach based on generative adversarial network. *Future Generation Computer Systems*, 141:768–776, 2023.

[31] Juliette Dromard and Philippe Owezarski. Study and evaluation of unsupervised algorithms used in network anomaly detection. In *Proceedings of the Future Technologies Conference (FTC) 2019: Volume 2*, pages 397–416. Springer, 2020.

[32] Peng Lin, Kejiang Ye, and Cheng-Zhong Xu. Dynamic network anomaly detection system by using deep learning techniques. In *Cloud Computing–CLOUD 2019: 12th International Conference, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25–30, 2019, Proceedings 12*, pages 161–176. Springer, 2019.

[33] Innocent Mbona and Jan HP Eloff. Detecting zero-day intrusion attacks using semi-supervised machine learning approaches. *IEEE Access*, 10:69822–69838, 2022.

[34] Jean-Paul A Yaacoub, Ola Salman, Hassan N Noura, Nesrine Kaaniche, Ali Chehab, and Mohamad Malli. Cyber-physical systems security: Limitations, issues and future trends. *Microprocessors and microsystems*, 77:103201, 2020.

[35] Yanfang Fu, Yishuai Du, Zijian Cao, Qiang Li, and Wei Xiang. A deep learning model for network intrusion detection with imbalanced data. *Electronics*, 11(6):898, 2022.

[36] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, 53(3):1–37, 2020.

[37] Abdelkader Dairi, Fouzi Harrou, Benamar Bouyeddou, Sidi-Mohammed Senouci, and Ying Sun. Semi-supervised deep learning-driven anomaly detection schemes for cyber-attack detection in smart grids. In *Power systems cybersecurity: Methods, concepts, and best practices*, pages 265–295. Springer, 2023.

[38] Jose Antonio Lopez, Iñaki Angulo, and Saturnino Martinez. Substation-aware. an intrusion detection system for the iec 61850 protocol. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*, pages 1–7, 2022.

[39] 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.

[40] Mahwish Anwar, Anton Borg, and Lars Lundberg. A comparison of unsupervised learning algorithms for intrusion detection in iec 104 scada protocol. In *2021 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–8. IEEE, 2021.

[41] Vivek Kumar Singh and Manimaran Govindarasu. A cyber-physical anomaly detection for wide-area protection using machine learning. *IEEE Transactions on Smart Grid*, 12(4):3514–3526, 2021.

[42] Yew Meng Khaw, Amir Abiri Jahromi, Mohammadreza FM Arani, Scott Sanner, Deepa Kundur, and Marthe Kassouf. A deep learning-based cyberattack detection system for transmission protective relays. *IEEE Transactions on Smart Grid*, 12(3):2554–2565, 2020.

[43] Willone Lim, Kelvin Sheng Chek Yong, Bee Theng Lau, and Colin Choon Lin Tan. Future of generative adversarial networks (gan) for anomaly detection in network security: A review. *Computers & Security*, 139:103733, 2024.

[44] Rick Sauber-Cole and Taghi M Khoshgoftaar. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *Journal of Big Data*, 9(1):98, 2022.

[45] Faizan Manzoor, Vanshaj Khattar, Akila Herath, Clifton Black, Matthew C Nielsen, Junho Hong, Chen-Ching Liu, and Ming Jin. Detecting zero-day attacks in digital substations via in-context learning. *arXiv preprint arXiv:2501.16453*, 2025.

[46] Faizan Manzoor, Vanshaj Khattar, Chen-Ching Liu, and Ming Jin. Zero-day attack detection in digital substations using in-context learning. In *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 220–225. IEEE, 2024.

[47] Tung-Wei Lin, Vanshaj Khattar, Yuxuan Huang, Junho Hong, Ruoxi Jia, Chen-Ching Liu, Alberto Sangiovanni-Vincentelli, and Ming Jin. CAUSALPROMPT: Enhancing LLMs with weakly supervised causal reasoning for robust performance in non-language tasks. In *ICLR Workshop: Tackling Climate Change with Machine Learning*, 2024.

[48] Silvio Ereno 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*, 21(4):3851–3865, 2023.

[49] Junho Hong et al. Implementation of secure sampled value (SeSV) messages in substation automation system. *IEEE Transactions on Power Delivery*, 37(1):405–414, 2021.

[50] Partha P Biswas, Heng Chuan Tan, Qingbo Zhu, Yuan Li, Daisuke Mashima, and Binbin Chen. A synthesized dataset for cybersecurity study of iec 61850 based substation. In *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–7. IEEE, 2019.

[51] 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.

[52] Monowar H Bhuyan, Dhruba Kumar Bhattacharyya, and Jugal K Kalita. Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials*, 16(1):303–336, 2013.

[53] Sebastian Garcia, Martin Grill, Jan Stiborek, and Alejandro Zunino. An empirical comparison of botnet detection methods. *computers & security*, 45:100–123, 2014.

[54] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[55] Mingkai Yu, Jianxiao Wang, Jie Yan, Lin Chen, Yang Yu, Gengyin Li, and Ming Zhou. Pricing information in smart grids: A quality-based data valuation paradigm. *IEEE Transactions on Smart Grid*, 13(5):3735–3747, 2022.

[56] Yi Yang, HT Jiang, Kieran McLaughlin, L Gao, YB Yuan, W Huang, and Sakir Sezer. Cybersecurity test-bed for iec 61850 based smart substations. In *2015 IEEE Power & Energy Society General Meeting*, pages 1–5. IEEE, 2015.

[57] Weihao Zeng, Dayuan Fu, Keqing He, Yejie Wang, Yukai Xu, and Weiran Xu. DivTOD: Unleashing the power of LLMs for diversifying task-oriented dialogue representations. *arXiv preprint arXiv:2404.00557*, 2024.

[58] Ellie S Paek, Talyn Fan, James D Finch, and Jinho D Choi. Enhancing task-oriented dialogue systems through synchronous multi-party interaction and multi-group virtual simulation. *Information*, 15(9):580, 2024.

[59] Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*, 2024.

[60] Yu Huang and Liangyuan Su. Design of intrusion detection and response mechanism for power grid scada based on improved lstm and fnn. *IEEE Access*, 2024.

[61] Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. A review of unsupervised machine learning frameworks for anomaly detection in industrial applications. In *Science and Information Conference*, pages 158–189. Springer, 2022.

[62] Chao Meng, Xue Song Jiang, Xiu Mei Wei, and Tao Wei. A time convolutional network based outlier detection for multidimensional time series in cyber-physical-social systems. *IEEE Access*, 8:74933–74942, 2020.

[63] Anish Jindal, Amit Dua, Kuljeet Kaur, Mukesh Singh, Neeraj Kumar, and Sukumar Mishra. Decision tree and svm-based data analytics for theft detection in smart grid. *IEEE Transactions on Industrial Informatics*, 12(3):1005–1016, 2016.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[65] Isaac Martín De Diego et al. General performance score for classification problems. *Applied Intelligence*, 52(10):12049–12063, 2022.