Leveraging Conversational Generative AI for Anomaly Detection in Digital Substations

Aydin Zaboli, *Student Member, IEEE*, Seong Lok Choi, *Member, IEEE*, Junho Hong, *Senior Member, IEEE*

***Abstract*—This study addresses critical challenges of cybersecurity in digital substations by proposing an innovative task-oriented dialogue (ToD) system for anomaly detection (AD) in multicast messages, specifically, generic object oriented substation event (GOOSE) and sampled value (SV) datasets. Leveraging generative artificial intelligence (GenAI) technology, the proposed framework demonstrates superior error reduction, scalability, and adaptability compared with traditional human-in-the-loop (HITL) processes. Notably, this methodology offers significant advantages over machine learning (ML) techniques in terms of efficiency and implementation speed when confronting novel and/or unknown cyber threats, while also maintaining model complexity and precision. The research employs advanced performance metrics to conduct a comparative assessment between the proposed AD and HITL-based AD frameworks, utilizing a hardware-in-the-loop (HIL) testbed for generating and extracting features of IEC61850 communication messages. This approach presents a promising solution for enhancing the reliability of power system operations in the face of evolving cybersecurity challenges.**

***Index Terms*—Anomaly detection, GenAI, GOOSE, human-in-the-loop, SV, task-oriented dialogue.**

I. Introduction

The integration of IEC 61850-based digital substations into smart grids has revolutionized energy management, enabling automated data collection and remote control of electrical systems. However, the fusion of power infrastructure with communication networks has introduced various security vulnerabilities, requiring the implementation of robust anomaly detection systems (ADSs) to effectively address challenges concerning the preservation and defense of critical national assets [1]–[3].

While ML techniques have become instrumental in detecting anomalies within GOOSE and SV multicast messages, they face limitations in terms of scalability, decision-making efficacy, and data processing. The continuous need for model re-training to address new attack vectors creates temporal vulnerabilities and resource-intensive processes. Given these complexities and limitations, GenAI tools present a more flexible and adaptive approach to AD in digital substations [4]. GenAI, with its inherent capability to understand situational intricacies and subtleties, can potentially detect novel attacks without prior training, offering a more robust and efficient AD methodology. This innovative approach emphasizes the potential of GenAI tools to enhance cybersecurity by providing a tool that can dynamically evolve in response to emerging cyber threats, thereby addressing the limitations of traditional ML frameworks and HITL processes [5], [6]. Recent advancements in AD processes for securing digital substations have seen a significant shift toward the utilization of sophisticated ML models. These models have demonstrated considerable potential in enhancing AD capabilities by analyzing patterns and anomalies within datasets, thereby enabling real-time detection of cyberattacks and contributing to grid security [7]. Notable among these efforts is the AI-based ransomware detection approach proposed by Alvee *et al.* [8], that employs a convolutional neural network (CNN) and innovatively converts binary files into 2-D image files for detection, achieving a high accuracy rate of 96.22%. However, the approach's efficacy remains subject to debate due to the absence of HIL testbed data and the absence of various attack scenarios. Further contributions to this field include the research of Yang *et al.* [9] who introduced a novel methodology combining statistical analysis and ML models for AD processes, offering improved adaptability to evolving threats. Additionally, Zhu *et al.* [1] have developed a novel ADS specifically targeting manufacturing message specification (MMS)-based measurement attacks, incorporating advanced detection algorithms to enhance accuracy. These diverse approaches highlight the ongoing efforts to refine and optimize AD techniques in the context of digital substation security, although each methodology presents its own set of limitations and areas for further research and development. Current literature reviews indicate a notable absence of research exploring directed methodologies leveraging GenAI tools for attack and anomaly detection based on human recommendations. Existing research endeavors predominantly grapple with challenges in scalability, adaptability, robustness, and processing efficiency. Consequently, there is a pressing need for a framework that can address challenges with minimal effort and reduced reliance on continuous human expert intervention.

This study presents an advanced GenAI-based ToD system, offering advantages over HITL processes and ML algorithms for AD in multicast messages. Built on historical human recommendations, it automates decision-making by emulating patterns, potentially reducing errors over time. The learning capabilities and data processing proficiency enable scalability, though challenges remain in building user trust. The following points encapsulate contributions of the proposed approach:

- This paper introduces a groundbreaking implementation of a GenAI-based ToD system for the efficient and reliable detection of anomalies in multicast messages. This innovative approach presents a method for bolstering the security and operational stability of smart grid infrastructures.
- An analysis of the proposed methodology was conducted employing a diverse array of advanced metrics. This rigorous evaluation process not only rectifies the limitations identified in prior research but also institutes a new paradigm for assessing ADS efficacy with improved efficiency, flexibility, and expandability.

The subsequent sections are organized thusly: An assessment

A. Zaboli and J. Hong are with the Department of Electrical and Computer Engineering, University of Michigan – Dearborn, Dearborn, MI 48128, USA.

S. L. Choi is with the Power Systems Engineering Center, National Renewable Energy Laboratory (NREL), Golden, CO 80401, USA.

of IEC 61850-based protocols, and human-derived rules is presented in Section II. Section III elucidates the proposed AD methodology in GOOSE/SV datasets. Section IV encompasses a detailed discussion of results, conducting a comparative analysis of the efficacy of the HITL approach with that of the proposed framework. Finally, this research concludes in Section V.

II. Substation Cyber Imperatives

The integration of advanced communication technologies into digital substations necessitates robust cybersecurity measures, encompassing multi-layered protective strategies and regular assessments to address evolving cyber threats. An HIL testbed provides a controlled environment for investigating the interplay between cyber breaches and power system resilience. GOOSE and SV packet extraction from the HIL testbed is executed using Wireshark, enabling comprehensive network traffic analysis. This process facilitates detailed observation of communication patterns within the testbed environment. This methodical approach ensures accurate data collection and provides critical insights into cyber-physical system (CPS) dynamics [6]. The next section describes datasets, their feature extraction process and human recommendations integral to the proposed framework.

A. GOOSE and SV Dataset Features & Rules

The most important features of GOOSE messages can be considered as time, destination MAC address (*DM*) (01 00 03), source MAC address (*SM*) (27 34 31), *type* (88*b*8), application identifier (*appid*) (3), *dataset* ($SEL\_421\_1CFG/LLN0\$Goose$), GOOSE identifier (*goid*) (SEL_421_1), state number (*stnum*) (27), sequence number (*sqnum*), and *data1*/*data2* values that are binary. Let G = (*time*, *DM*, *SM*, *type*, *appid*, *dataset*, *goid*, *stnum*, *sqnum*, *data1*/*data2*) represent the features of a GOOSE message. Eqs. (1)–(8) illustrate the GOOSE rules (i.e., *GR#1* to *GR#8*) employed in this paper to check the different abnormalities of datasets. This paper presents different levels of considering rules as without training, WT (i.e, no rules), partial training, PT (i.e., *GR#1* to *GR#5*), and full training, FT (i.e., *GR#1* to *GR#8*). *GR#1*: When consecutive data packets exhibit identical *DM* and *SM* attributes, the *sqnum* parameter should incrementally advance. Any inconsistency reveals an abnormality.

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

GR#2: Modifications in *data1* (*d1*) or *data2* (*d2*) necessitate a unitary increment in *stnum* and an *sqnum* reset to 0. Non-compliance with this protocol indicates an anomaly.

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

GR#3: For data with identical *DM* and *SM*, the *stnum* must maintain a monotonically increasing sequence. Any regression in *stnum* value constitutes an anomaly.

$$GR\#3(G_i, G_{1:i-1}) = \begin{cases} 1, & \text{if } DM_i = SM_i \wedge stnum_i > \\ & \max_{j<i}(stnum_j) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

GR#4: Any alteration in *DM*, *SM*, *type*, *appid*, *dataset*, or *goid* parameters is indicative of an anomalous condition.

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

GR#5: The *time* column must adhere to a format delineating hour, minute, and second with microsecond precision. Any deviation from this temporal feature constitutes an anomaly.

$$GR\#5(time_i) = \begin{cases} 1, & \text{if } time_i \text{ is in format HH:MM:SS.mmmmmm} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

GR#6: The occurrence of data frequency exceeding 10 instances within a 10 μs interval is classified as an anomalous event.

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

GR#7: A temporal gap in data transmission exceeding 10 seconds is indicative of an anomalous condition.

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

GR#8: Upon detection of alterations in *d1* or *d2*, the *stnum* should remain constant while the *sqnum* undergoes an increment. Any deviation depicts an anomaly.

$$GR\#8(G_i, G_{i-1}) = \begin{cases} 1, & \text{if } (d1_i \neq d1_{i-1} \vee d2_i \neq d2_{i-1}) \wedge stnum_i = \\ & stnum_{i-1} \wedge sqnum_i > sqnum_{i-1} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

In the case of SV datasets, the most important features can be denoted as *time*, *DM* (04 00 01), *SM* (27 22 13), *type* (88*ba*), *appid* (40), sampled value identifier (*svid*) (4000), and sample count (*smpcnt*). Let S = (*time*, *DM*, *SM*, *type*, *appid*, *svid*, *smpcnt*) represent the features of an SV message. Eqs. (9)–(16) illustrate the SV rules (i.e., *SR#1* to *SR#8*) utilized to check the anomalies of SV datasets. A similar procedure of rules for different level of training in GOOSE messages can be applied to SV messages. *SR#1*: The *smpcnt* parameter is constrained to the integer interval $[0, 4799]$ for 60 Hz systems (i.e., $80 \times 60 = 4800$) or $[0, 3999]$ for 50 Hz systems (i.e., $80 \times 50 = 4000$). Any value outside this prescribed range is classified as an anomalous condition.

$$SR\#1(S_i) = \begin{cases} 1, & \text{if } 0 \leq Smpcnt_i \leq 4799 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

SR#2: The *smpcnt* parameter should exhibit a uniformly increasing series from 0 to 4799 for 60 Hz systems, followed by a reset to 0. Any deviation from this progression is indicative of an anomaly.

$$SR\#2(S_i, S_{i-1}) = \begin{cases} 1, & \text{if } (Smpcnt_i > Smpcnt_{i-1} \wedge \\ & Smpcnt_i \leq 4799) \vee \\ & \quad (Smpcnt_i = 0 \wedge Smpcnt_{i-1} = 4799) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

SR#3: The *smpcnt* parameter must maintain a non-decreasing sequence until it attains the value 4799, whereupon it resets to 0. Any deviation from this pattern denotes an anomaly.

$$SR\#3(S_i, S_{i-1}) = \begin{cases} 1, & \text{if } Smpcnt_i \geq Smpcnt_{i-1} \vee \\ & (Smpcnt_i = 0 \wedge Smpcnt_{i-1} = 4799) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

SR#4: The parameters *DM*, *SM*, *type*, *appid*, and *svid* must maintain invariance across all conditions. Any changes in these values signifies an anomalous state.

$$SR\#4(S_i, S_{i-1}) = \begin{cases} 1, & \text{if } DM_i = DM_{i-1} \wedge SM_i = SM_{i-1} \wedge \\ & type_i = type_{i-1} \wedge apid_i = apid_{i-1} \wedge \\ & svid_i = svid_{i-1} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

SR#5: The temporal data field must conform to a hierarchical structure of hour, minute, second, and microsecond. A deviation from this chronological format indicates an anomaly.

$$SR\#5(time_i) = \begin{cases} 1, & \text{if } time_i \text{in format HH:MM:SS.mmm} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

SR#6: The standard temporal interval is constrained to the range of 200 to 215 μs. Any deviation from this prescribed timeframe indicates an anomalous condition.

$$SR\#6(S_i, S_{i-1}) = \begin{cases} 1, & \text{if } 200\mu s \leq time_i - time_{i-1} \leq 215\mu s \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

SR#7: The occurrence of data frequency exceeding 12 instances within a 2.083-ms interval is an anomalous event.

$$SR\#7(S_{i-11:i}) = \begin{cases} 1, & \text{if } time_i - time_{i-11} \leq 2.083ms \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

SR#8: The *smpcnt* parameter should exhibit a unitary increment with each successive instance. Any irregularity in this sequential progression implies an anomalous condition.

$$SR\#8(S_i, S_{i-1}) = \begin{cases} 1, & \text{if } Smpcnt_i = (Smpcnt_{i-1} + 1) \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

The proposed framework is crafted in which GOOSE (*GR#1* to *GR#8*) and SV (*SR#1* to *SR#8*) rules guide SQL queries that filter messages to detect anomalies. In general, this framework uses partial training to help analysts apply rules for identifying abnormal patterns, suggesting additional rules if no anomalies are found. In full training, this method tries to confirm anomalies, by analyzing recent message packets and refining detection through belief-action updates. This adaptive approach enhances AD accuracy by dynamically incorporating rule adjustments which will be elaborated in the next section.

III. A Novel Anomaly Detection System Paradigm: Integrating GenAI with Task-Oriented Dialogue

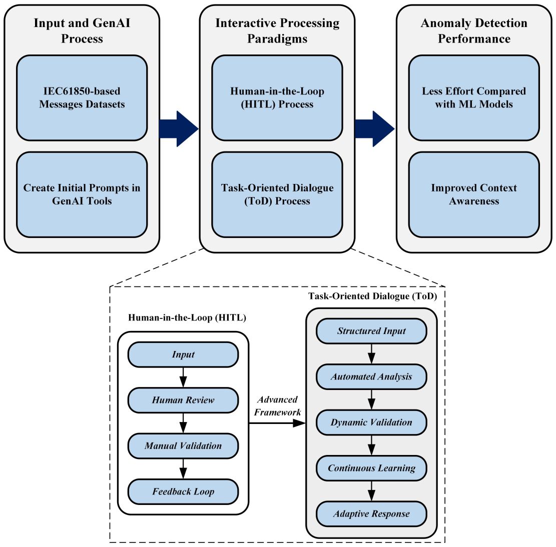
Input and GenAI Process
IEC61850-based Messages Datasets
Create Initial Prompts in GenAI Tools
Interactive Processing Paradigms
Human-in-the-Loop (HITL) Process
Task-Oriented Dialogue (ToD) Process
Anomaly Detection Performance
Less Effort Compared with ML Models
Improved Context Awareness
Human-in-the-Loop (HITL)
Input
Human Review
Manual Validation
Feedback Loop
Advanced Framework
Task-Oriented Dialogue (ToD)
Structured Input
Automated Analysis
Dynamic Validation
Continuous Learning
Adaptive Response

Fig. 1. A general framework of interactive GenAI-based AD process.

A GenAI-driven approach to AD in GOOSE and SV datasets can enhance the detection process, using interactive processing paradigms such as HITL and ToD as demonstrated in Fig. 1. The framework begins with gathering IEC61850-based message datasets. Initial prompts are crafted in GenAI tools to help the system identify anomalies by establishing normal operational patterns and highlighting potential threats. The interactive processing then enhances this process through HITL and ToD which the HITL process integrates human feedback to iteratively refine the model's detection accuracy, while ToD allows the GenAI system to engage in dialogue, gathering context and adjusting its response. These steps create a flexible, adaptable detection system that can efficiently identify anomalies. Compared to the ToD approach, the HITL process exhibits several limitations. While HITL relies heavily on constant human review and manual validation, requiring time and effort for each feedback loop, ToD offers a more efficient automated analysis system with dynamic validation capabilities. The HITL process's dependence on human intervention creates scalability constraints and slower processing cycles. Different steps of the ToD system are organized below to show its outperformance on HITL processes and ML models. ToD systems' efficacy is traditionally evaluated through their proficiency in distinct subtasks, encompassing dialogue state tracking (also known as belief state management), dialogue management (which includes action and decision prediction), and the generation of responses utilizing an SQL query database. In this GenAI-based framework, SQL query block acts as a filtering tool that efficiently identifies anomalies in data by checking GOOSE and SV messages against predefined rules. This approach simplifies the AD by processing only relevant data and maintaining consistent rule application to detect issues. This partitioning of tasks has facilitated the development of specialized models for each subtask, a methodology that has gained considerable adoption within the field [10]. The present research endeavors to investigate the efficacy of a unified and end-to-end model in managing these multilayered functions, as illustrated in Fig. 2 along with CPS. This model incorporates a cybersecurity analyst component, implemented as a GenAI tool, which processes GOOSE and SV data to detect anomalies, leveraging packets, ToD labels, and anomaly scores as inputs to foster an understanding of anomalous characteristics. A separation of the proposed ToD framework's steps illustrated in Fig. 1 and elaborated below.

Structured Input: This block captures structured inputs essential for the ToD system to function effectively which

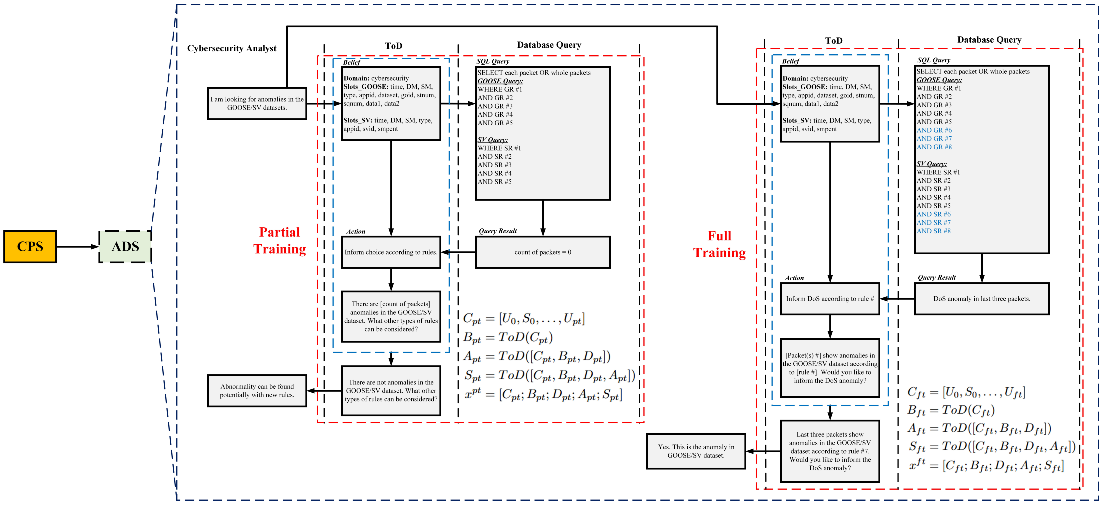
CPS
ADS
Cybersecurity Analyst
I am looking for anomalies in the GOOSE/SV datasets.
Partial Training
ToD
Database Query
Belief
Domain: cybersecurity
Slots_GOOSE: time, DM, SM, type, appid, dataset, goid, stnum, sqnum, data1, data2
Slots_SV: time, DM, SM, type, appid, svid, smpcnt
SQL Query
SELECT each packet OR whole packets
GOOSE Query:
WHERE GR #1
AND GR #2
AND GR #3
AND GR #4
AND GR #5
SV Query:
WHERE SR #1
AND SR #2
AND SR #3
AND SR #4
AND SR #5
Action
Inform choice according to rules.
Query Result
count of packets = 0
There are [count of packets] anomalies in the GOOSE/SV dataset. What other types of rules can be considered?
There are not anomalies in the GOOSE/SV dataset. What other types of rules can be considered?
Abnormality can be found potentially with new rules.
$C_{pt} = [U_0, S_0, \dots, U_{pt}]$
$B_{pt} = ToD(C_{pt})$
$A_{pt} = ToD([C_{pt}, B_{pt}, D_{pt}])$
$S_{pt} = ToD([C_{pt}, B_{pt}, D_{pt}, A_{pt}])$
$x^{pt} = [C_{pt}; B_{pt}; D_{pt}; A_{pt}; S_{pt}]$
Full Training
ToD
Database Query
Belief
Domain: cybersecurity
Slots_GOOSE: time, DM, SM, type, appid, dataset, goid, stnum, sqnum, data1, data2
Slots_SV: time, DM, SM, type, appid, svid, smpcnt
SQL Query
SELECT each packet OR whole packets
GOOSE Query:
WHERE GR #1
AND GR #2
AND GR #3
AND GR #4
AND GR #5
AND GR #6
AND GR #7
AND GR #8
SV Query:
WHERE SR #1
AND SR #2
AND SR #3
AND SR #4
AND SR #5
AND SR #6
AND SR #7
AND SR #8
Action
Inform DoS according to rule #
Query Result
DoS anomaly in last three packets.
[Packet(s) #] show anomalies in the GOOSE/SV dataset according to [rule #]. Would you like to inform the DoS anomaly?
Last three packets show anomalies in the GOOSE/SV dataset according to rule #7. Would you like to inform the DoS anomaly?
Yes. This is the anomaly in GOOSE/SV dataset.
$C_{ft} = [U_0, S_0, \dots, U_{ft}]$
$B_{ft} = ToD(C_{ft})$
$A_{ft} = ToD([C_{ft}, B_{ft}, D_{ft}])$
$S_{ft} = ToD([C_{ft}, B_{ft}, D_{ft}, A_{ft}])$
$x^{ft} = [C_{ft}; B_{ft}; D_{ft}; A_{ft}; S_{ft}]$

Fig. 2. Advancing multicast messages security: A GenAI-based ToD System for the AD process.

corresponds to data parameters (e.g., *time*, *DM*). For PT and FT levels, input packets are represented as $C_{pt} = [U_0, S_0, \dots, U_{pt}]$ and $C_{ft} = [U_0, S_0, \dots, U_{ft}]$, respectively, which denotes a collection of structured data extracted from GOOSE/SV packets.

Automated Analysis: This step involves analyzing data automatically, applying AD rules based on prior rules. It includes SQL queries for messages to extract anomaly-related data. During the PT level, automated analysis is stated as $B_{pt} = \text{ToD}(C_{pt})$ which indicates the transformation of the structured input to identify preliminary anomalies based on rules. Then, the analysis is refined in the FT level as $B_{ft} = \text{ToD}(C_{ft})$, providing a deeper analysis for complete training.

Dynamic Validation: This step includes verifying identified anomalies dynamically, using a set of conditions. It includes queries to count specific packet conditions. In the PT phase, dynamic validation refines the analysis with additional packet transformations as $S_{pt} = \text{ToD}([C_{pt}, B_{pt}, D_{pt}, A_{pt}])$ where D_{pt} includes additional conditions. Further, the validation includes more advanced criteria in the FT phase as $S_{ft} = \text{ToD}([C_{ft}, B_{ft}, D_{ft}, A_{ft}])$, increasing the complexity and robustness of validation checks with further transformations.

Continuous Learning: The framework adapts over time, incorporating feedback to refine the AD process. This aligns with the transition from partial to full training levels in the proposed ToD framework. In this case, the learning happens iteratively in PT with A_{pt}, where rules are revised to enhance AD as $x^{pt} = [C_{pt}; B_{pt}; D_{pt}; A_{pt}; S_{pt}]$. Moreover, the learning process is formalized with a higher count of structured feedback and adaptive rules in FT level as $x^{ft} = [C_{ft}; B_{ft}; D_{ft}; A_{ft}; S_{ft}]$.

Adaptive Response: Based on AD outcomes, the system generates adaptive responses to inform the cybersecurity analyst of potential issues in the GOOSE/SV data. The responses align with the system's blocks for communicating anomaly findings. The adaptive responses for PT and FT levels involve generating responses based on the findings in x^{pt} and x^{ft}, respectively.

An evaluation of contemporary AD paradigms, including HITL processes, ML architectures, and the novel GenAI-based ToD framework, illuminates the efficacy and versatility of the latter approach. While HITL methodologies are inherently limited by the availability and expertise of human operators, and ML models, despite their computational capability, are constrained by input/output structures and reduced adaptability, this system emerges as a more comprehensive and efficient solution. This innovative framework harnesses the power of natural language dialogue to accommodate complex queries and responses, demonstrating remarkable adaptability to emergent scenarios and anomalies through advanced prompt engineering techniques, while concurrently offering interpretable insights.

IV. Results and Discussion

A comparative analysis is carried out to evaluate the performance of the proposed framework against the HITL process. This evaluation aims to illuminate the potential advantages of the proposed approach, offering insights into its scalability and adaptability in complex scenarios. This study utilizes advanced evaluation metrics including informedness, markedness, Matthews correlation coefficient (MCC), and geometric mean (GM), each ranging from -1 to 1, except for the GM, which is between 0 and 1, to assess the consistency, decision-making, quality, and balance between normal and abnormal datasets, respectively. In the context of GOOSE/SV datasets, these metrics serve distinct purposes: Informedness measures the model's ability to detect anomaly-indicative patterns, Markedness assesses its proficiency in reducing false positives (FPs) and negatives (FNs), MCC provides a balanced performance measure in scenarios with rare anomalies, and the GM evaluates the model's precision in AD while maintaining low FP rates [11]. In this application, a value of 1 for informedness, markedness, MCC, or GM indicates optimal performance, reflecting perfect anomaly detector, crucial for maintaining the accuracy and reliability of digital substation communications. This section endeavors to elucidate the results obtained through the application of advanced metrics across diverse methodologies, classified by their respective training levels (i.e., WT, PT, and FT). To simplify this process and minimize human effort, an image-based framework based on rules was conceptualized and implemented. The empirical evidence presented in Table I demonstrates the superiority of the proposed framework across all training levels. By comparison, the HITL process exhibits suboptimal performance, with the first three metrics yielding

TABLE I
A COMPARATIVE ANALYSIS OF AD REGARDING WT, PT, FT LEVELS.

GOOSE						
Method	**HITL**			**ToD**		
Metrics	WT	PT	FT	WT	PT	FT
Informedness	0.22	0.3964	0.5709	0.825	0.8998	0.9492
Markedness	0.233	0.416	0.599	0.8296	0.8998	0.9492
MCC	0.0247	0.2054	0.4142	0.822	0.8997	0.9491
Geometric Mean	0.5865	0.6844	0.7784	0.9105	0.9512	0.9746
SV						
Method	**HITL**			**ToD**		
Metrics	WT	PT	FT	WT	PT	FT
Informedness	**0**	0.5	0.8833	0.8296	0.8968	0.9467
Markedness	**0**	0.3836	0.7407	0.825	0.8968	0.9467
MCC	**0**	0.3713	0.8432	0.823	0.8966	0.9466
Geometric Mean	0.5	0.7483	0.9397	0.9143	0.9484	0.9733

values of 0, indicating that its predictive capabilities for positive and negative data rarely surpass random chance. This deficiency manifests in elevated rates of FPs and FNs, substantially compromising the prediction accuracy. An in-depth analysis of Fig. 3 reveals the proposed method's unquestionable superiority across all metrics, thereby affirming its reliability in the context of AD for GOOSE/SV datasets. Moreover, the model's decision-making efficacy when applied to SV dataset is comparable to that of a coin toss, rendering these models unreliable due to the absence of meaningful correlations and the inability to consistently identify correct predictions. The advanced metrics visualized in Fig. 3 further confirm the exceptional performance of the GenAI-based framework, particularly at the FT level, where values approaching 1 underscore its good efficiency, scalability, adaptability, and reliability. Additionally, Fig. 4 presents a novel comparative framework, investigating various models across different training levels based on accuracy and improvement differentials. This visualization serves to articulate the incremental percentage gains at each level. Notably, this model exhibits minimal increases when implemented within the proposed framework, suggesting a robust correlation across training levels and demonstrating exceptional AD capabilities even in the absence of training, as evidenced by the negligible performance disparities across levels. Conversely, the HITL process displays the highest incremental percentage, indicating enhanced adaptability in synthesizing new rules. Nevertheless, the integration of HITL methodology further enhances the ToD framework by incorporating human expertise. This combination allows for a refinement of the model's accuracy and responsiveness through continuous feedback mechanisms. As a result, the AD process becomes more precise and dependable, effectively leveraging both automated systems and human insight to identify and respond to anomalies more efficiently.

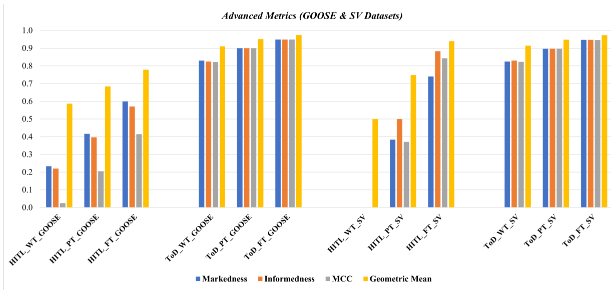

Fig. 3. A comparative assessment of GenAI-based HITL and the proposed framework: Insights from advanced metrics applied to GOOSE and SV datasets.

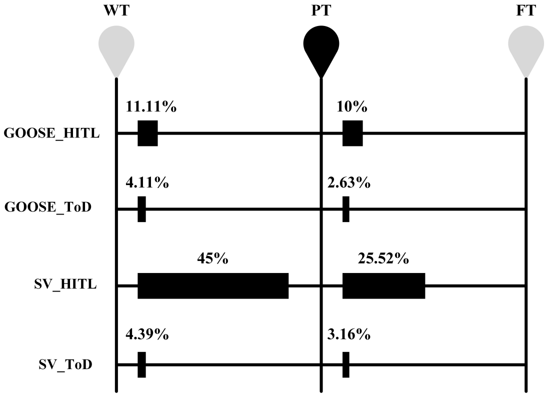

Fig. 4. Accuracy metrics for various models based on training levels.

V. CONCLUSION AND FUTURE DIRECTIONS

This study introduces a GenAI-based ToD framework for an efficient and reliable AD in IEC61850-based multicast messages. The framework's scalability and adaptability are validated through comparative analysis with HITL process, employing advanced metrics to assess reliability and correlation capabilities—aspects previously overlooked in smart grids. Also, this GenAI-based methodology demonstrates acceptable performance across all metrics. Future research directions will include integrating a self-learning component to expand the applicability to other messages, such as MMS, and incorporating natural language processing (NLP) metrics to enhance the quality of GenAI-generated outputs.

REFERENCES

[1] R. Zhu, C.-C. Liu, J. Hong, and J. Wang, "Intrusion detection against mms-based measurement attacks at digital substations," *IEEE Access*, vol. 9, pp. 1240–1249, 2020.

[2] J. Hong and C.-C. Liu, "Intelligent electronic devices with collaborative intrusion detection systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 271–281, 2017.

[3] J. Hong *et al.*, "Automated cybersecurity tester for IEC61850-based digital substations," *Energies*, vol. 15, no. 21, p. 7833, 2022.

[4] S. M. S. Mohammadabadi *et al.*, "Generative artificial intelligence for distributed learning to enhance smart grid communication," *International Journal of Intelligent Networks*, 2024.

[5] H. S. Mavikumbure *et al.*, "Generative AI in cyber security of cyber physical systems: Benefits and threats," in *2024 16th International Conference on Human System Interaction (HSI)*. IEEE, 2024, pp. 1–8.

[6] A. Zaboli, S. L. Choi, T.-J. Song, and J. Hong, "Chatgpt and other large language models for cybersecurity of smart grid applications," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, 2024.

[7] X. Wang *et al.*, "Anomaly detection for insider attacks from untrusted intelligent electronic devices in substation automation systems," *IEEE Access*, vol. 10, pp. 6629–6649, 2022.

[8] S. R. Alvee *et al.*, "Ransomware attack modeling and artificial intelligence-based ransomware detection for digital substations," in *2021 6th IEEE Workshop on the Electronic Grid (eGRID)*. IEEE, 2021, pp. 01–05.

[9] L. Yang *et al.*, "A new methodology for anomaly detection of attacks in IEC 61850-based substation system," *Journal of Information Security and Applications*, vol. 68, p. 103262, 2022.

[10] S. Hu *et al.*, "Dialight: Lightweight multilingual development and evaluation of task-oriented dialogue systems with large language models," *arXiv preprint arXiv:2401.02208*, 2024.

[11] I. M. De Diego *et al.*, "General performance score for classification problems," *Applied Intelligence*, vol. 52, no. 10, pp. 12 049–12 063, 2022.