International Journal Of
Industrial Electronics Control and Optimization
Print ISSN: 3517-2645
Online ISSN: 3568-2645
Homepage: https://ieco.usb.ac.ir/

SIP Signaling Anomaly Detection in IMS/VoLTE Networks: A Survey, Taxonomy, and Conceptual Multi-Level Machine Learning Framework

Ali Kuhestani[1] | Mojtaba Karimi[1] | Mohammad Reza Keshavarzi[2]

Faculty of Electrical and Computer Engineering, Qom University of Technology, Qom, Iran[1]
ICT Research Institute, Iran Telecommunication Research Center (ITRC), Tehran, Iran[2]

Corresponding author's email: kuhestani@qut.ac.ir

Article Info

Article type:
Review Article

Article history:
Received: 27-June-2026
Received in revised form: 09-October-2026
Accepted: 06-November-2026
Published online: **********

Keywords:
SIP signaling,
IMS/VoLTE,
Anomaly detection,
Machine learning,
Telecom security.

ABSTRACT

Session Initiation Protocol (SIP) is a core signaling protocol in IMS and VoLTE networks, supporting user registration, authentication, call setup, session management, and call termination. Abnormal SIP signaling behavior may affect service availability, call setup performance, troubleshooting activities, and the operational stability of IMS components. Existing studies on SIP anomaly detection have addressed flooding, malformed messages, registration abuse, dialog misuse, and machine learning-based detection; however, many approaches remain limited in their ability to jointly consider message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level behaviors.

This paper presents a survey, taxonomy, and conceptual multi-level machine learning framework for SIP signaling anomaly detection in IMS/VoLTE environments. First, SIP-related anomalies are categorized according to their operational behavior and protocol impact. Then, existing rule-based, statistical, machine learning, deep learning, and federated learning approaches are reviewed and compared in terms of applicability, explainability, data requirements, and operational limitations. Based on the identified gaps, a conceptual six-level feature framework is proposed, covering message, temporal, transaction, dialog, error, and entity features. The framework is intended to support the future development of explainable and data-driven SIP anomaly detection systems.

Since the scope of this work is conceptual rather than experimental, the paper also outlines key research directions for future validation, including controlled IMS/VoLTE testbed implementation, feature extraction from labeled SIP traces, model comparison under class imbalance, and integration of explainability mechanisms for operational use.

I. Introduction

IP Multimedia Subsystem (IMS) and Voice over LTE (VoLTE) networks provide important infrastructure for delivering voice and multimedia services in modern mobile communication systems. In these networks, the Session Initiation Protocol (SIP) is used to support essential signaling procedures, including user registration, authentication, session establishment, call control, and session termination. The reliability of IMS/VoLTE services therefore depends not only on general network availability, but also on the correct exchange, timing, and ordering of SIP messages.

SIP-based systems have long been recognized as vulnerable to signaling-related threats such as impersonation, message manipulation, flooding, malformed signaling, and denial-of-service attacks [1], [6]. With the deployment of IMS and VoLTE in mobile operator environments, these issues have become more critical. Unlike generic VoIP systems, IMS/VoLTE networks involve telecom-specific entities and service procedures, meaning

Copyright © Author(s).
Publisher: University of Sistan and Baluchestan

that abnormal SIP behavior can directly affect call availability, signaling load, and user quality of experience. Previous studies have also shown that VoLTE and IMS systems may suffer from implementation weaknesses, signaling abuse, amplification attacks, call denial-of-service, and privacy leakage [11]–[16], [21].

The main problem addressed in this paper is the detection and classification of SIP signaling anomalies in IMS/VoLTE networks. SIP anomalies may appear in several forms, including REGISTER flooding, INVITE flooding, registration abuse, malformed messages, abnormal transaction behavior, invalid dialog sequences, excessive error responses, and unusual activity from users or network entities. These anomalies can disrupt normal IMS/VoLTE signaling procedures and may lead to service degradation, failed call setup, increased signaling overhead, or denial of service.

A key challenge is that SIP anomalies are not always visible through simple traffic-volume indicators. Some attacks are high-rate and can be detected based on message frequency, while others affect the logical structure of SIP procedures. For example, an anomaly may appear as an incomplete transaction, an invalid dialog sequence, an unusual response-code distribution, or abnormal behavior from a specific user entity. Therefore, SIP anomaly detection in IMS/VoLTE networks requires a procedure-aware perspective that considers message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level behavior.

Several approaches have been proposed for SIP, VoIP, IMS, and VoLTE security. Early studies mainly focused on rule-based, threshold-based, and statistical techniques for detecting SIP flooding and denial-of-service attacks. Detection of flooding attacks against IMS networks, adaptive z-score CUSUM-based REGISTER flooding detection, and SIP message flooding detection have been investigated in previous studies [5], [8], [9]. Stateful approaches have also been introduced. For example, SIPAD models SIP behavior using a stateful rule tree, showing the importance of considering SIP procedure states rather than only message volume [10].

Machine learning-based methods have also been applied to SIP anomaly detection. Previous studies used classifiers such as k-nearest neighbors and support vector machines for IMS intrusion detection and SIP traffic monitoring [2]–[4]. Other studies considered labeled VoIP datasets, recurrent neural networks for SIP-based DDoS detection, and machine learning-based intrusion detection in enterprise VoIP environments [7], [17], [22]. More recent research has examined SIP dialog prediction, abnormal SIP dialog classification, abnormal signaling pattern detection, and federated learning for SIP-DDoS anomaly detection [18]–[20], [23].

Machine learning and deep learning methods are particularly promising because they can model complex and nonlinear relationships in signaling data. Compared with static thresholds and manually defined rules, learning-based approaches may be more adaptable to different traffic conditions and more suitable for detecting anomalies that are difficult to express through fixed signatures.

Despite these advances, existing studies still have several limitations when considered from an IMS/VoLTE-oriented perspective. First, many works focus on generic VoIP environments rather than telecom-specific IMS/VoLTE signaling procedures. Second, several studies address isolated attack types, especially flooding attacks, while broader SIP anomalies such as abnormal dialog sequences, excessive error responses, and entity-level irregularities have received less integrated attention. Third, many approaches rely on single-level features, such as message rate, statistical thresholds, or dialog patterns, rather than combining multiple SIP abstraction levels.

Another important limitation is the lack of reproducible IMS/VoLTE-oriented dataset generation. Public SIP or VoIP datasets are useful, but they may not fully represent IMS/VoLTE procedures, telecom signaling entities, and controlled attack scenarios. For future experimental validation, controlled telecom emulation environments, such as Open5GS-based IMS/VoLTE testbeds, can provide a repeatable way to generate normal and anomalous SIP traces under known conditions. However, the existing literature does not sufficiently connect controlled IMS/VoLTE emulation, SIP procedure reconstruction, multi-level feature representation, and machine learning-based anomaly detection into a unified research framework.

This paper does not present a deployable SIP intrusion detection system, and it does not report validation on real operator traces. Its scope is narrower. It reviews how SIP anomaly detection has been studied so far, examines why several existing methods are not sufficient for IMS/VoLTE signaling scenarios, and proposes a multi-level representation of SIP behavior that can guide future validation work.

The review covers previous studies on SIP, VoIP, IMS, and VoLTE anomaly detection, including rule-based and statistical methods, supervised learning, deep learning, federated learning, and dataset-oriented research. The discussion is not limited to flooding attacks. In practice, some relevant anomalies are reflected in protocol state, dialog progression, response-code patterns, or entity behavior, rather than in traffic volume alone.

The paper also organizes SIP signaling anomalies into a taxonomy that considers both security behavior and SIP procedure logic. The taxonomy includes flooding and denial-of-service cases, registration abuse, malformed messages, abnormal transactions and dialogs, excessive error responses, and IMS/VoLTE-specific signaling issues. This classification helps show which anomaly types have been widely studied and which ones still need more systematic treatment.

Finally, the paper proposes a conceptual multi-level feature view for SIP anomaly detection. In this view, SIP behavior is separated into message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level features. This distinction is important because different anomalies become visible at different levels. For example, flooding behavior may be observed through message rates, while an incomplete dialog or a missing ACK requires reconstruction of the signaling procedure.

The paper closes by identifying the main steps needed for future validation, including controlled IMS/VoLTE testbed development, labeled SIP trace generation, transaction and dialog reconstruction, comparative model evaluation, and explainability analysis. These items are presented as future work, not as completed experimental results.

The remainder of this paper is organized as follows. Section 2 describes the review methodology, including the search strategy, inclusion and exclusion criteria, and classification dimensions. Section 3 presents SIP signaling and IMS/VoLTE security background. Section 4 introduces a taxonomy of SIP signaling anomalies. Section 5 reviews existing SIP anomaly detection approaches, including rule-based, statistical, machine learning, deep learning, federated, and dataset-oriented studies. Section 6 provides a comparative analysis and identifies research gaps. Section 7 presents the proposed conceptual multi-level machine learning framework. Section 8 discusses future research directions. Finally, Section 9 concludes the paper.

II. Review Methodology

This paper follows a structured narrative review methodology to analyze existing studies on SIP signaling anomaly detection, with particular attention to IMS/VoLTE environments. The purpose of the review is not only to summarize previous studies, but also to classify them according to target environment, anomaly type, detection approach, dataset usage, and level of SIP signaling awareness. This classification is then used to identify research gaps and motivate the proposed conceptual multi-level machine learning framework.

A. Search Strategy

Relevant studies were identified through major academic databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, MDPI, and Google Scholar. The search was conducted using combinations of keywords related to SIP security, SIP anomaly detection, IMS intrusion detection, VoLTE security, SIP flooding and DDoS attacks, SIP dialog abnormality, VoIP intrusion detection, machine learning-based telecom signaling security, network anomaly detection, and intrusion detection datasets.

The selected literature covers studies ranging from early SIP vulnerability analysis to studies on IMS/VoLTE security, SIP and VoIP anomaly detection, dialog-based signaling analysis, deep learning-based detection, adaptive learning methods, federated learning, and dataset-oriented intrusion detection research. The reviewed papers were grouped into five main categories: SIP/IMS/VoLTE vulnerability analysis, rule-based and statistical detection, machine learning-based detection, deep learning and sequence-based detection, and dataset/testbed-oriented studies.

B. Inclusion and Exclusion Criteria

The inclusion criteria were defined to select studies relevant to SIP signaling security, anomaly detection, telecom signaling protection, and learning-based security monitoring. A study was included if it satisfied at least one of the following conditions: it analyzed SIP, IMS, VoIP, or VoLTE security vulnerabilities; proposed or evaluated a SIP flooding, DoS, DDoS, or anomaly detection method; applied machine learning, deep learning, adaptive learning, or federated learning to security detection in SIP, VoIP, IMS, or network environments; studied abnormalities in SIP dialogs, transactions, or signaling patterns; introduced a dataset, testbed, or experimental environment for SIP/VoIP/IMS or intrusion detection evaluation; or discussed IMS/VoLTE-specific attacks, weaknesses, or operational security issues.

Studies were excluded if they focused only on general IP network intrusion detection without methodological relevance to anomaly detection, addressed unrelated mobile network layers without signaling relevance, or did not provide sufficient technical detail regarding the detection method, anomaly type, dataset, or evaluation environment. Studies from the broader fields of cybersecurity and network anomaly detection were considered only when they contributed methodologically to machine learning-based detection, dataset evaluation, adaptive learning, or anomaly analysis.

Based on these criteria, representative studies were selected from different research directions, including SIP vulnerability surveys [1], [6], IMS/SIP flooding detection methods [5], [8], [9], machine learning-based SIP/IMS detection [2]–[4], SIP dialog and abnormal signaling studies [18]–[20], IMS/VoLTE security analyses [11]–[16], [21], SIP/VoIP DDoS learning approaches [17], [24], dataset and intrusion detection evaluation studies [7], [22], [25], [26], general ML-based cyber threat detection studies [27], [28], network anomaly detection using deep learning [29], and federated learning-based SIP-DDoS detection [23].

C. Classification Dimensions

To support a structured comparison, the reviewed studies were classified according to six dimensions.

First, the **target environment** was considered. Studies were grouped according to whether they focused on generic SIP systems, VoIP networks, IMS networks, VoLTE environments, enterprise VoIP systems, operational mobile networks, or broader network-security environments. This distinction is important because IMS/VoLTE signaling includes telecom-specific procedures and entities that are not always represented in generic VoIP or general intrusion detection studies.

Second, the **anomaly or attack type** was identified. The reviewed works were classified according to whether they

addressed SIP flooding, REGISTER flooding, INVITE flooding, DDoS attacks, malformed messages, abnormal dialogs, amplification attacks, call denial-of-service, privacy leakage, general SIP signaling anomalies, or broader cyber and network anomaly detection scenarios.

Third, the **detection method** was analyzed. Existing approaches were grouped into rule-based methods, threshold-based methods, statistical techniques, stateful or procedure-aware models, traditional machine learning methods, deep learning methods, adaptive learning methods, and federated or collaborative detection approaches.

Fourth, the **feature representation level** was examined. Some studies rely mainly on traffic-level or message-rate indicators, while others consider SIP transaction states, dialog sequences, response-code behavior, temporal patterns, or general network-level features. This dimension is central to the present paper because it helps identify whether previous studies use single-level indicators or multi-level SIP signaling representations.

Fifth, the **dataset and evaluation environment** were considered. Studies were classified according to whether they used public datasets, generated datasets, simulation environments, controlled testbeds, enterprise VoIP traces, operational mobile network observations, or general intrusion detection benchmark datasets. This dimension is important because one of the key challenges in SIP anomaly detection is the limited availability of reproducible IMS/VoLTE-oriented datasets.

Finally, the reviewed works were assessed in terms of **procedure awareness**. A method was considered procedure-aware if it explicitly modeled SIP transactions, dialog states, message ordering, or signaling procedure validity. This criterion is particularly relevant for IMS/VoLTE networks, where anomalies may appear not only as abnormal traffic volume but also as invalid or incomplete signaling procedures.

These classification dimensions provide the basis for the comparative analysis presented in later sections and support the identification of the main research gap: the lack of a unified approach that combines controlled IMS/VoLTE-oriented emulation, SIP procedure reconstruction, multi-level feature representation, and machine learning-based anomaly detection.

III. SIP Signaling and IMS/VoLTE Security Background

A. SIP Role in IMS/VoLTE Networks

The Session Initiation Protocol is the main signaling protocol used in IMS and VoLTE networks for multimedia session control. In IMS-based mobile networks, SIP is responsible for procedures such as user registration, authentication, session initiation, session modification, call termination, and service control. Unlike ordinary packet traffic, SIP signaling follows predefined message exchanges between user equipment and IMS core entities. For this reason, the security and reliability of IMS/VoLTE services depend not only on network resource availability, but also on the correctness, timing, and ordering of SIP messages.

In a typical IMS/VoLTE architecture, SIP messages may be exchanged between user equipment and IMS entities such as the Proxy-Call Session Control Function (P-CSCF), Interrogating-CSCF (I-CSCF), and Serving-CSCF (S-CSCF). The P-CSCF is usually the first SIP contact point for user equipment, while the I-CSCF and S-CSCF support routing, registration, authentication coordination, and service control. Since these entities process large volumes of signaling traffic, abnormal SIP behavior can affect both individual subscribers and core network functions.

The importance of SIP security has been recognized in earlier studies, where SIP-based systems were shown to be vulnerable to impersonation, message manipulation, flooding, malformed messages, and denial-of-service attacks [1], [6]. In IMS/VoLTE environments, these issues become more critical because SIP signaling is directly connected to mobile voice service availability and call control. Recent studies on VoLTE and IMS have also shown that signaling weaknesses, implementation problems, and operational vulnerabilities may lead to service disruption, call denial-of-service, amplification attacks, or privacy leakage [11]–[16], [21].

B. Main IMS/VoLTE Signaling Procedures

IMS/VoLTE signaling behavior can be described through several main SIP procedures. The first is user registration. During registration, the user equipment sends SIP REGISTER messages to the IMS core, and authentication-related responses such as 401 Unauthorized and final success responses such as 200 OK may be exchanged. This procedure allows the IMS network to authenticate the user and bind the user identity to a reachable contact address. Because registration is fundamental to service access, attacks such as REGISTER flooding or registration abuse can create excessive signaling load and affect the ability of legitimate users to register.

The second procedure is session establishment. In VoLTE call setup, SIP INVITE messages are used to initiate voice sessions, followed by provisional and final responses such as 100 Trying, 180 Ringing, 183 Session Progress, and 200 OK. After successful negotiation, ACK confirms session establishment. This procedure is highly sequence-dependent. Abnormal message ordering, incomplete dialogs, repeated INVITE messages, or excessive failed call attempts may indicate signaling abuse or attack behavior.

The third procedure is session termination and modification. SIP BYE messages are used to terminate sessions, while other SIP methods may be involved in session update or service-related signaling. Abnormal termination patterns, missing messages, repeated retransmissions, or unexpected responses can indicate signaling instability or malicious activity. Therefore, IMS/VoLTE anomaly detection should not only count SIP

messages but also examine whether transactions and dialogs follow expected procedure patterns.

From a detection perspective, SIP signaling can be analyzed at multiple abstraction levels. At the message level, individual SIP requests and responses are counted and classified. At the temporal level, message rates, inter-arrival times, and burst behavior are examined. At the transaction level, request-response relationships are reconstructed. At the dialog level, complete SIP call or registration sequences are analyzed. At the error level, abnormal response-code distributions are examined. At the entity level, the behavior of users, IP addresses, or IMS nodes is monitored. These levels provide the conceptual basis for the multi-level framework proposed later in this paper.

C. Security Challenges in SIP-Based IMS/VoLTE Signaling

SIP signaling in IMS/VoLTE networks introduces several security challenges. First, SIP is procedure-dependent, meaning that the validity of signaling behavior depends on the correct exchange of messages. An attacker may therefore create anomalies not only by increasing message volume, but also by disrupting transaction states, generating incomplete dialogs, or producing unexpected response patterns. Stateful approaches such as SIPAD highlight the importance of modeling SIP behavior beyond simple packet or message counting [10].

Second, IMS/VoLTE networks are sensitive to signaling overload. Flooding attacks, including REGISTER flooding and INVITE flooding, can consume processing resources at IMS entities and reduce service availability for legitimate users. Previous studies have investigated SIP and IMS flooding using rule-based, statistical, and anomaly detection techniques [5], [8], [9]. However, high-rate flooding is only one part of the problem. Lower-rate or procedure-aware anomalies may be harder to detect if the monitoring system relies only on thresholds or traffic volume.

Third, SIP signaling may expose telecom-specific vulnerabilities in VoLTE and IMS deployments. Studies on VoLTE systems have shown that implementation weaknesses, hidden channels, amplification behavior, and call denial-of-service scenarios can affect mobile voice services [11]–[14]. More recent research has also shown that voice service signaling in LTE and 5G environments can be associated with privacy and identity leakage risks [15], [16], [21]. These findings indicate that SIP anomaly detection for IMS/VoLTE should consider telecom-specific signaling behavior rather than relying only on generic VoIP assumptions.

Finally, dataset availability remains a major research challenge. Real IMS/VoLTE signaling traces are difficult to obtain because they may contain sensitive subscriber, operator, and service information. As a result, many studies rely on generic VoIP datasets, limited experimental traces, or specific attack scenarios. This motivates the need for controlled IMS/VoLTE-oriented emulation environments, such as Open5GS-based testbeds, where normal and anomalous SIP signaling behavior can be generated under controlled and repeatable conditions. This issue is discussed further in the comparative analysis and in the proposed conceptual framework.

IV. Taxonomy of SIP Signaling Anomalies

SIP signaling anomalies in IMS/VoLTE networks can appear at different levels of signaling behavior. Some anomalies are visible through abnormal traffic volume, while others are related to message syntax, transaction states, dialog progression, error responses, or entity behavior. Therefore, a taxonomy is required to organize SIP anomalies according to how they affect IMS/VoLTE procedures and how they can be observed by detection systems. In this paper, SIP signaling anomalies are grouped into six categories: flooding and denial-of-service attacks, registration abuse, malformed SIP messages, abnormal transaction and dialog sequences, excessive error-response behavior, and IMS/VoLTE-specific vulnerabilities.

A. Flooding and DoS Attacks

Flooding and denial-of-service attacks are among the most studied SIP signaling anomalies. In these attacks, an adversary generates a large number of SIP messages to overload SIP servers, IMS core components, or VoLTE service functions. REGISTER flooding and INVITE flooding are two common examples. REGISTER flooding targets the registration procedure and may overload IMS entities involved in authentication and user registration, while INVITE flooding targets session establishment and may affect call setup performance or service availability [5], [8], [9].

Flooding attacks are usually observable through message-level and temporal-level indicators, such as a sudden increase in SIP request rate, reduced inter-arrival time, excessive retransmissions, or abnormal request-response ratios. However, low-rate or distributed flooding may be harder to detect using simple thresholds. This limitation motivates the use of statistical, machine learning, and collaborative detection methods for SIP-based DoS and DDoS scenarios [17], [23].

B. Registration Abuse

Registration abuse refers to abnormal behavior related to SIP REGISTER procedures. In IMS/VoLTE networks, registration is required before users can access voice and multimedia services. An attacker may abuse this procedure by sending repeated REGISTER messages, attempting unauthorized registrations, manipulating identity-related fields, or generating excessive authentication challenges. Since registration involves interaction between user equipment and IMS core entities, abnormal registration behavior can increase signaling load and affect legitimate user access.

This anomaly category is closely related to REGISTER flooding, but it is broader. While REGISTER flooding mainly focuses on volume-based overload, registration

abuse may also involve invalid credentials, abnormal registration frequency, suspicious user identity behavior, or repeated failed authentication. Such behavior can be observed through transaction-level and error-level features, including REGISTER success ratio, authentication challenge ratio, failed registration ratio, and repeated 401 or 403 responses [8].

C. Malformed SIP Messages

Malformed SIP messages occur when SIP requests or responses contain invalid syntax, missing headers, abnormal header values, inconsistent fields, or unexpected message structures. These anomalies may be generated intentionally to test implementation weaknesses, bypass security controls, or cause abnormal behavior in SIP parsers and IMS components. Earlier SIP security studies identified malformed signaling and message manipulation as important categories of SIP-related threats [1], [6].

In IMS/VoLTE environments, malformed SIP messages are important because SIP messages are processed by several entities, including P-CSCF, I-CSCF, S-CSCF, Application Servers (AS), and border control functions. Malformed messages may not always produce high traffic volume, but they may trigger parsing errors, abnormal response codes, transaction failures, or unexpected dialog behavior. Therefore, detecting malformed SIP signaling requires more than traffic-rate analysis; it also requires message-level validation, header consistency checking, and response-code monitoring.

D. Abnormal Transaction and Dialog Sequences

SIP is a procedure-dependent protocol. A valid SIP session depends on the correct order of requests and responses. For example, a typical call setup may involve INVITE, provisional responses such as 100 Trying or 180 Ringing, a final 200 OK response, and ACK. Similarly, registration involves REGISTER messages, authentication-related responses, and final success or failure responses. An abnormal transaction or dialog sequence occurs when this expected signaling behavior is incomplete, invalid, repeated, or inconsistent.

Figure 1 compares normal SIP signaling with representative anomalies, including repeated registration, incomplete INVITE transactions, and a missing ACK.

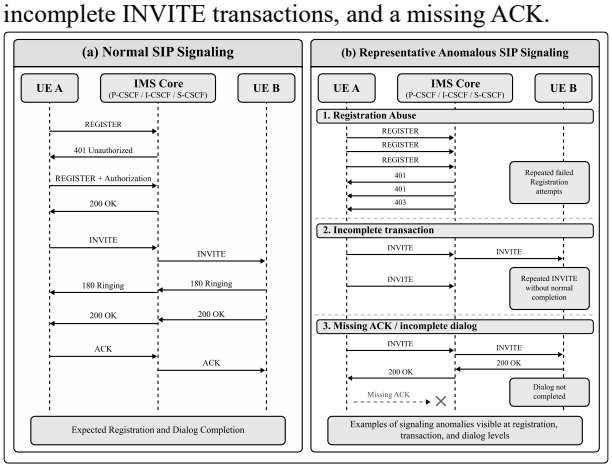


Fig. 1. Comparison of (a) normal IMS/VoLTE SIP signaling and (b) representative anomalous signaling behavior, including registration abuse, incomplete transactions, and missing ACK.

Abnormal transaction and dialog behavior may include missing ACK messages, incomplete INVITE dialogs, unexpected BYE messages, repeated retransmissions, invalid response ordering, or unusual call termination patterns. Stateful approaches such as SIPAD show the importance of modeling SIP behavior through procedure states rather than only counting messages [10]. More recent studies on SIP dialog prediction and abnormal signaling classification also highlight the relevance of dialog-level and sequence-based analysis [18]–[20].

This category is particularly important for IMS/VoLTE anomaly detection because many signaling attacks are not purely volumetric. Some anomalies may remain hidden if the detection system only monitors packet rate or message count. Procedure-aware monitoring can therefore provide a deeper understanding of whether SIP transactions and dialogs follow valid IMS/VoLTE signaling behavior.

E. Excessive Error Responses

Excessive error-response behavior refers to abnormal distributions of SIP response codes. In normal IMS/VoLTE signaling, some error responses may occur due to authentication challenges, unavailable users, rejected calls, or temporary network conditions. However, a sudden or persistent increase in specific response classes may indicate abnormal behavior. For example, repeated 401 Unauthorized, 403 Forbidden, 404 Not Found, 408 Request Timeout, 486 Busy Here, or 5xx server error responses may indicate registration abuse, scanning, failed call attempts, overload, or malformed signaling.

Error-level analysis is useful because some anomalies are reflected more clearly in response behavior than in request volume. For instance, an attacker may generate moderate-rate signaling traffic that does not exceed traffic thresholds but still causes abnormal failure ratios. Therefore, error-level features such as error-response rate, 4xx/5xx ratio, authentication failure ratio, and transaction failure ratio can support more reliable SIP anomaly detection.

F. IMS/VoLTE-Specific Vulnerabilities

Some SIP signaling anomalies are specific to IMS/VoLTE deployments and cannot be fully understood using generic VoIP assumptions. VoLTE systems include mobile-specific architecture, subscriber identities, SIM-based authentication, IMS core components, policy control, and interaction with LTE/5G access networks. As a result, vulnerabilities may arise from implementation weaknesses, misconfigurations, hidden channels, amplification behavior, call denial-of-service scenarios, or privacy leakage [11]–[16], [21].

IMS/VoLTE-specific anomalies may involve abnormal interaction between user equipment and IMS entities, unexpected behavior in operational call systems, or privacy-related signaling leakage. These anomalies are important for

research because they show that SIP anomaly detection should not be limited to laboratory VoIP scenarios. Instead, detection models should consider telecom-specific procedures, signaling entities, and operational monitoring points such as P-CSCF-adjacent probes, Session Border Controller (SBC) monitoring layers, and Security Operations Center (SOC)/Network Operations Center (NOC) analytics platforms.

TABLE I Taxonomy of SIP signaling anomalies in IMS/VoLTE networks

Category	Example anomalies	Main observation level
Flooding/DoS	REGISTER flood, INVITE flood, SIP DDoS	Message, temporal
Registration abuse	repeated registration, failed authentication	Transaction, error, entity
Malformed messages	invalid headers, malformed syntax	Message, error
Abnormal transactions/dialogs	incomplete call setup, invalid sequence	Transaction, dialog
Excessive error responses	high 4xx/5xx rate, repeated 401/403	Error, transaction
IMS/VoLTE-specific issues	call DoS, amplification, privacy leakage	Procedure, entity, system

This taxonomy shows that SIP signaling anomalies are multi-dimensional. Some are best detected through traffic-volume indicators, while others require transaction reconstruction, dialog analysis, response-code monitoring, or entity-level behavioral analysis. This observation supports the need for a multi-level feature representation, which is discussed in the proposed framework section.

V. Existing SIP Anomaly Detection Approaches

Existing SIP anomaly detection approaches can be classified into rule-based and threshold-based methods, statistical detection methods, machine learning-based methods, deep learning and sequence-based methods, federated or collaborative detection methods, and dataset/testbed-oriented studies. Each category provides a different perspective on SIP signaling security. Rule-based and statistical methods are generally simple and interpretable, while machine learning and deep learning methods are more flexible for learning complex patterns. However, many existing approaches remain limited by generic VoIP assumptions, narrow attack coverage, or insufficient IMS/VoLTE-specific procedure awareness.

A. Rule-Based and Threshold-Based Methods

Rule-based and threshold-based approaches are among the earliest methods used for SIP anomaly detection. These methods usually define a set of predefined conditions, signatures, or thresholds to detect abnormal behavior. For example, a monitoring system may raise an alert when the number of SIP REGISTER or INVITE messages exceeds a predefined threshold within a specific time interval. Such methods are especially useful for detecting high-rate flooding attacks because these attacks often create sudden increases in SIP signaling volume.

Several studies have investigated SIP and IMS flooding attacks using rule-based or threshold-based detection principles. Rebahi et al. studied flooding attacks against IMS networks and considered how abnormal signaling load can affect IMS components [5]. Asgharian et al. also investigated denial-of-service message flooding attacks in SIP-based services [9]. These works show that volumetric SIP attacks can be detected by monitoring abnormal request rates, repeated messages, or excessive signaling attempts.

The main advantage of rule-based and threshold-based approaches is their simplicity. They are easy to implement, computationally lightweight, and understandable for network operators. In operational environments, interpretability is important because security analysts need to understand why an alert is generated. However, these methods have significant limitations. Static thresholds are difficult to tune because normal SIP traffic may vary depending on time, number of users, service conditions, and network events. A low threshold may generate many false positives, while a high threshold may miss low-rate or stealthy attacks. Moreover, rule-based approaches usually depend on known attack patterns and may not generalize well to new or evolving anomalies.

Another limitation is that threshold-based methods often focus on message volume rather than SIP procedure correctness. In IMS/VoLTE networks, some anomalies may not produce extreme traffic rates. Instead, they may appear as invalid message ordering, incomplete transactions, unusual response-code distributions, or abnormal behavior from specific user entities. Therefore, while rule-based and threshold-based methods are useful as baseline mechanisms, they are not sufficient for comprehensive SIP anomaly detection in IMS/VoLTE environments.

B. Statistical Detection Methods

Statistical detection methods improve on simple thresholds by modeling normal traffic behavior and identifying deviations from expected patterns. Instead of relying only on fixed limits, statistical methods may use moving averages, variance, cumulative sum techniques, probability distributions, or anomaly scores. These approaches are suitable for detecting changes in SIP traffic behavior, especially when attacks cause measurable deviations from normal signaling patterns.

Nosrati et al. proposed an adaptive z-score CUSUM algorithm for detecting REGISTER flooding attacks in IMS networks [8]. This type of approach is useful because it can detect changes in signaling behavior over time and adapt to variations in normal traffic. Statistical techniques are particularly relevant for SIP flooding detection because flooding often changes message arrival rates, inter-arrival times, and request distributions.

The strength of statistical methods is that they provide a more adaptive detection mechanism than fixed thresholds. They can model normal behavior and detect deviations even when absolute traffic values change. They are also relatively

lightweight compared with complex machine learning models, making them attractive for real-time monitoring environments.

However, statistical approaches also have limitations. First, they often assume that normal behavior can be represented by simple statistical patterns. In real IMS/VoLTE networks, SIP signaling behavior may be affected by many factors, including user mobility, call patterns, re-registration intervals, network congestion, and service configuration. Second, many statistical methods focus on a limited set of features, such as message count, message rate, or inter-arrival time. This makes them effective for volumetric anomalies but less effective for procedure-aware anomalies. Third, statistical methods usually require careful parameter tuning, and their performance may degrade when traffic patterns change or when attacks are low-rate and distributed.

Therefore, statistical methods are valuable for detecting traffic-level deviations, but they need to be complemented with transaction-level, dialog-level, error-level, and entity-level analysis to support a more complete IMS/VoLTE anomaly detection framework.

C. Stateful and Procedure-Aware Detection Methods

SIP is a stateful and procedure-dependent protocol. Therefore, some approaches attempt to detect anomalies by modeling SIP transactions, dialog states, and expected message sequences. These methods are more procedure-aware than simple traffic-volume techniques because they examine whether SIP messages follow valid signaling logic.

Seo et al. proposed SIPAD, a SIP/VoIP anomaly detection approach based on a stateful rule tree [10]. This type of method is important because many SIP anomalies can be detected only when the relationship between messages is considered. For example, an INVITE transaction should normally be followed by provisional and final responses, and a successful session setup should include an ACK. Missing, repeated, or unexpected messages may indicate abnormal behavior even if the overall message rate is not high.

Procedure-aware methods are especially relevant for IMS/VoLTE networks because telecom signaling depends heavily on correct procedure execution. Registration, authentication, call setup, call termination, and service-related signaling each follow expected sequences. Therefore, monitoring only the number of SIP messages may be insufficient. A detection system should also consider whether transactions are completed successfully, whether dialogs follow expected states, and whether response codes are consistent with normal behavior.

The main advantage of stateful methods is their ability to detect logical anomalies. They can identify incomplete dialogs, invalid message sequences, unexpected state transitions, and abnormal procedure behavior. However, these approaches also have limitations. They often require expert-defined rules or state models, which may be difficult to maintain as SIP services evolve. They may also be sensitive to legitimate variations in signaling behavior caused by network configuration, retransmissions, service features, or vendor-specific implementation differences. As a result, purely rule-based stateful models may lack flexibility when applied to diverse IMS/VoLTE environments.

D. Machine Learning-Based Methods

Machine learning-based methods detect SIP anomalies by learning patterns from data rather than relying only on manually defined rules or fixed thresholds. These approaches usually involve feature extraction from SIP messages, traffic flows, or signaling logs, followed by classification using algorithms such as k-nearest neighbors, support vector machines, decision trees, random forests, gradient boosting, or other supervised and unsupervised models.

Akbar et al. compared anomaly detection algorithms for detecting SIP flooding in IMS environments [2]. Farooqi et al. proposed an intrusion detection system for IMS using the k-nearest neighbor classifier [3]. Nassar et al. used support vector machines for monitoring SIP traffic [4]. These studies show that machine learning can be applied to SIP and IMS signaling data to detect abnormal behavior more flexibly than fixed rule-based methods.

Beyond SIP-specific studies, supervised machine learning has also been applied to cyber threat prediction, hacking-breach detection, and cybercrime-oriented threat detection [27], [28]. Although these studies are not specific to SIP signaling or IMS/VoLTE procedures, they show the broader role of learning-based models in security analytics and support a shift from static rule-based monitoring toward data-driven detection.

The main strength of machine learning methods is their ability to learn complex relationships between features. For example, an ML model can combine message rates, request-response ratios, error-code distributions, transaction failures, and entity behavior to classify normal and abnormal traffic. This makes ML suitable for anomalies that are not easily described by a single threshold or rule.

However, ML-based SIP anomaly detection also faces several challenges. First, model performance depends heavily on dataset quality and representativeness. If the dataset does not reflect realistic IMS/VoLTE procedures, the trained model may not generalize well to telecom environments. Second, many ML studies rely mainly on traffic-level or statistical features, while giving less attention to SIP transaction reconstruction, dialog progression, and procedure validity. Third, supervised ML requires labeled data, but labeled IMS/VoLTE attack traces are difficult to obtain from real networks because of privacy, operational, and security constraints. Fourth, some ML models may behave as black boxes, making it difficult for operators to understand why a specific SIP flow, transaction, or dialog is classified as anomalous.

Therefore, machine learning is promising for SIP anomaly detection, but it should be combined with careful feature

engineering, procedure-aware reconstruction, interpretable evaluation, and controlled dataset generation.

E. Deep Learning and Sequence-Based Methods

Deep learning and sequence-based methods have been introduced to capture complex temporal and sequential behavior in SIP signaling. Unlike traditional ML models, which often depend on manually engineered features, deep learning models can learn representations from ordered sequences or high-dimensional feature vectors. Recurrent neural networks and related architectures are particularly relevant for SIP analysis because SIP dialogs naturally consist of ordered message exchanges.

Nazih et al. applied recurrent neural networks to counter DDoS attacks in SIP-based VoIP networks [17]. Pereira et al. studied machine learning-based prediction of SIP signaling dialogs and later investigated deep learning-based classification of abnormal SIP dialogs [18], [19]. Pereira and Oliveira also examined the detection of abnormal SIP signaling patterns [20]. These studies are important because they shift the detection focus from simple traffic indicators toward dialog behavior and signaling sequence analysis.

Hybrid and adaptive learning models have also been explored in related network-security contexts. For example, CRNN-based architectures have been investigated for firewall anomaly detection [29], while transfer learning and boosted meta-learning have been applied to adaptive DDoS detection in software-defined SIP-VoIP environments [24]. The adaptive approach in [24] is useful because transfer learning can reduce dependence on a completely fixed training distribution when attack behavior changes. However, its evaluation focuses mainly on DDoS detection in a software-defined SIP-VoIP environment and does not address IMS/VoLTE registration, transaction, or dialog procedures. These studies indicate the growing role of deep, hybrid, and adaptive learning models in anomaly detection. However, their applicability to IMS/VoLTE signaling still depends on whether SIP transactions, dialog states, and telecom-specific procedures are properly represented in the input data.

Sequence-based approaches are suitable for detecting anomalies such as missing ACK messages, incomplete INVITE dialogs, abnormal response ordering, repeated transactions, or invalid procedure paths. In IMS/VoLTE networks, such anomalies are particularly relevant because the validity of a call or registration procedure depends on message order, transaction completion, and dialog consistency.

Despite their advantages, deep learning methods have several limitations. They often require larger datasets than traditional ML models, and their performance depends strongly on the diversity and representativeness of training data. They may also require more computational resources, which can be a concern for real-time telecom monitoring. Another issue is explainability. While deep learning models may achieve high detection performance, it can be difficult to interpret their decisions in a way that is useful for SOC/NOC analysts. In addition, if the dataset is generated from generic VoIP environments, the learned patterns may not fully represent IMS/VoLTE-specific signaling behavior.

For these reasons, deep learning and sequence-based methods should be viewed as promising but not sufficient by themselves. They need to be supported by well-defined datasets, procedure-aware labeling, SIP transaction and dialog reconstruction, and interpretable evaluation methods.

F. Federated and Collaborative Detection Methods

Federated and collaborative detection methods have recently gained attention because they allow multiple entities to participate in model training or anomaly detection without directly sharing raw traffic data. This is particularly relevant for telecom security because real SIP and IMS/VoLTE signaling traces may contain sensitive subscriber and operator information.

Sbai et al. proposed a SIP-DDoS framework based on federated learning for collaborative anomaly detection [23]. The framework allows multiple SIP domains to contribute to a shared detection model without exchanging raw traffic data, thereby improving privacy and supporting collaborative detection across distributed environments. The study also showed that knowledge sharing between participating nodes can improve the detection of INVITE flooding attacks. However, the evaluation focuses mainly on SIP-DDoS and INVITE flooding scenarios, and does not examine broader IMS/VoLTE procedures such as registration behavior, transaction completion, or dialog-state consistency.

Related work has also considered the security of the federated learning process itself. Jahani et al. proposed a Byzantine fault-tolerant federated learning framework in which PCA-based anomaly detection is used to identify abnormal local model updates [33]. The method reduces the impact of malicious participants without requiring centralized raw data. However, it was evaluated in a predictive maintenance setting rather than a SIP or IMS environment. Therefore, its use in telecom signaling would require SIP-specific features and validation on IMS/VoLTE traffic.

The main advantage of federated detection is that it can address the limited availability of centralized SIP datasets. In principle, different operators, testbeds, or monitoring domains could contribute to a shared detection model without exposing their raw SIP traces. This is especially useful for DDoS and distributed anomaly detection scenarios, where attack behavior may vary across locations.

However, federated learning also introduces challenges. It requires coordination between participants, secure aggregation mechanisms, and careful handling of heterogeneous data distributions. In telecom environments, different networks may have different IMS configurations, SIP procedures, vendor implementations, traffic volumes, and security policies. These differences may affect model

convergence and generalization. Moreover, federated learning does not remove the need for careful feature engineering. If each participant extracts weak or inconsistent features, the collaborative model may still perform poorly.

Therefore, federated and collaborative detection should be considered a future direction that can benefit from the same multi-level SIP feature representation proposed in this paper.

G. Dataset and Testbed-Based Studies

Dataset availability is one of the most important challenges in SIP anomaly detection research. Nassar et al. introduced a labeled VoIP dataset for intrusion detection evaluation [7], while Alvares et al. provided a dataset of attacks on a live enterprise VoIP network for ML-based intrusion detection and prevention systems [22]. These datasets are valuable because they support reproducible evaluation and comparison of detection methods.

Dataset quality is also important for anomaly detection research. Dataset-oriented studies emphasize the importance of label integrity, scenario realism, traffic diversity, practical relevance, and clear separation between normal and attack scenarios [25], [26]. These issues are directly relevant to SIP anomaly detection because a model trained on poorly labeled or weakly representative data may achieve good experimental results but fail to generalize to realistic telecom signaling environments.

However, generic VoIP, enterprise VoIP, or general intrusion detection datasets may not fully represent IMS/VoLTE signaling behavior. IMS/VoLTE networks include telecom-specific components, subscriber identities, SIM-based authentication, registration procedures, policy interactions, and operator-controlled signaling paths. Therefore, a model trained only on generic VoIP or general network traces may not capture the full behavior of IMS/VoLTE signaling anomalies.

Operational IMS/VoLTE studies have shown that real mobile networks may contain vulnerabilities related to call denial-of-service, amplification behavior, implementation weaknesses, and privacy leakage [11]–[16], [21]. The study in [21] is particularly valuable because it examines IMS security issues in real 4G/5G mobile environments and provides practical evidence of vulnerabilities that may not appear in laboratory VoIP settings. However, the work mainly focuses on identifying and demonstrating security weaknesses rather than developing a general-purpose SIP anomaly detection model. These works are important because they demonstrate real-world relevance. However, operational traces are difficult to reuse for ML research because they may contain sensitive subscriber, operator, and service information and are not usually publicly available.

For this reason, controlled telecom emulation environments are important for future research. Testbeds based on platforms such as Open5GS and IMS-related components such as Kamailio or OpenSIPS can be used to generate normal and anomalous SIP traces under repeatable conditions. Such environments can support controlled execution of registration, authentication, call setup, call termination, flooding attacks, malformed messages, abnormal dialogs, and error-response scenarios. They also allow researchers to define labels, attack intervals, traffic intensities, scenario boundaries, and observation points more precisely than in uncontrolled real networks.

Nevertheless, emulated datasets also have limitations. They may not fully capture the diversity of operational mobile networks, user behavior, vendor-specific implementations, and large-scale traffic dynamics. Therefore, controlled emulation should be considered a necessary foundation for reproducible research, but future work should also validate models against more diverse and realistic traces when possible.

H. Summary of Existing Approaches

The reviewed approaches show that SIP anomaly detection has evolved from rule-based and statistical detection toward machine learning, deep learning, adaptive learning, and collaborative detection methods. However, no single approach fully addresses the requirements of IMS/VoLTE anomaly detection. Rule-based and threshold-based methods are interpretable but rigid. Statistical methods are adaptive but often limited to traffic-level deviations. Stateful methods are procedure-aware but depend heavily on predefined rules. Machine learning methods are flexible but require representative labeled datasets. Deep learning and sequence-based methods can model signaling sequences but need large datasets and may lack explainability. Adaptive and transfer learning approaches can improve flexibility in DDoS-oriented scenarios, but they still require careful dataset design and domain-specific validation [24]. Federated methods support collaborative learning but introduce deployment, coordination, and data heterogeneity challenges [23].

Broader cybersecurity and network anomaly detection studies also confirm the increasing use of supervised, deep, and hybrid learning models for security analytics [27]–[29]. However, these approaches are not sufficient by themselves for SIP anomaly detection in IMS/VoLTE networks because they do not directly model telecom-specific procedures such as registration, authentication, SIP transaction completion, dialog progression, error-response behavior, and IMS signaling entities.

Overall, the literature suggests that SIP anomaly detection in IMS/VoLTE networks requires an integrated perspective. A suitable framework should combine controlled dataset generation, dataset quality control, SIP transaction and dialog reconstruction, multi-level feature engineering, interpretable model evaluation, and machine learning-based detection. This observation motivates the comparative gap analysis in the next section and supports the conceptual framework proposed later in this paper.

VI. Comparative Analysis and Research Gaps

The previous sections introduced the main SIP signaling anomaly categories and reviewed representative detection approaches. This section synthesizes those findings by comparing existing studies across key dimensions and identifying the remaining gaps that motivate the proposed conceptual framework.

A. Comparative Summary of Existing Studies

Table II compares representative studies according to their coverage of SIP/VoIP/IMS signaling, IMS/VoLTE specificity, learning-based detection, procedure awareness, dataset or testbed support, multi-level feature representation, and unified framework design.

TABLE II Comparative summary of existing SIP/IMS/VoLTE anomaly detection studies

Ref.	SIP/VoIP/IMS	IMS/VoLTE	ML/DL	Procedure-aware	Dataset/Testbed	Multi-level	Unified framework
[1], [6]	✓	✗	✗	✗	✗	✗	✗
[5], [8], [9]	✓	✓	✗	✗	✗	✗	✗
[10]	✓	✗	✗	✓	✗	✗	✗
[2]–[4]	✓	✓	✓	✗	✗	✗	✗
[7], [22], [25], [26]	✓	✗	✓	✗	✓	✗	✗
[17], [24]	✓	✗	✓	✗	✗	✗	✗
[18]–[20]	✓	✗	✓	✓	✗	✗	✗
[27], [28]	✗	✗	✓	✗	✗	✗	✗
[29]	✗	✗	✓	✗	✗	✗	✗
[11]–[16], [21]	✓	✓	✗	✓	✗	✗	✗
[23]	✓	✗	✓	✗	✗	✗	✗
Proposed	**✓**	**✓**	**✓**	**✓**	**✓**	**✓**	**✓**

As shown in Table II, the existing methods address different parts of the SIP anomaly detection problem rather than the complete signaling process. Rule-based and statistical approaches are mainly suitable for anomalies that produce observable changes in message rate or traffic distribution, while stateful and sequence-based methods are better suited to abnormalities in transaction completion and dialog progression. Machine learning can combine several indicators, but its performance depends on representative labeled data and appropriate feature design. Deep and federated approaches provide additional flexibility for sequence modeling and distributed learning, although they introduce higher data, computational, and coordination requirements. The main gap is therefore not the absence of detection algorithms, but the lack of an IMS/VoLTE-oriented framework in which these methods can be evaluated using the same procedure-aware and multi-level representation.

B. Key Limitations of Existing Research

The first limitation is the dependence on generic SIP, VoIP, or broader network-security environments. Several datasets and detection methods are useful for anomaly detection research, but they do not fully represent IMS/VoLTE-specific entities, registration procedures, authentication behavior, subscriber-related signaling, or mobile voice service paths. This limits the direct applicability of models trained or evaluated outside telecom-oriented signaling environments.

The second limitation is the concentration on narrow anomaly classes, particularly flooding and DDoS scenarios. Although flooding attacks are important, IMS/VoLTE signaling anomalies may also involve abnormal dialogs, malformed messages, excessive error responses, registration abuse, incomplete transactions, and entity-level irregularities. Detection models that focus mainly on volumetric behavior may therefore miss anomalies that appear through procedure violations or abnormal signaling sequences.

The third limitation is the lack of an integrated feature representation. Many studies focus on one dominant perspective, such as traffic rate, statistical deviation, network-level indicators, or dialog sequence. Broader cybersecurity and network anomaly detection studies demonstrate the usefulness of supervised, deep, and hybrid learning models [27]–[29], but they do not directly model SIP-specific procedures. Few approaches combine message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level indicators within a unified IMS/VoLTE-oriented framework.

The fourth limitation is the shortage of reproducible IMS/VoLTE-oriented datasets. Real operator traces are difficult to access because of privacy, confidentiality, and operational sensitivity, while public VoIP or intrusion detection datasets may not sufficiently reflect telecom-specific signaling procedures. Dataset-oriented studies also show that label quality, scenario realism, traffic diversity, and practical relevance are critical for reliable anomaly detection evaluation [25], [26]. This creates a need for controlled emulation-based dataset generation.

The fifth limitation concerns operational evaluation. Existing studies often emphasize detection accuracy, while telecom monitoring also requires attention to false positive rate, detection latency, interpretability, deployment location, computational cost, and integration with SOC/NOC workflows. Without these criteria, a detection model may appear effective in experimental evaluation but remain difficult to deploy in practical IMS/VoLTE monitoring environments.

C. Research Gap

Based on the comparative analysis, the main research gap can be stated as follows:

Existing SIP anomaly detection studies have addressed SIP vulnerabilities, flooding attacks, statistical detection, machine learning classification, dialog modeling, IMS/VoLTE security weaknesses, dataset development, adaptive learning, and federated detection. However, there is still a lack of a unified framework that combines IMS/VoLTE-oriented controlled emulation, SIP transaction and dialog reconstruction, multi-level feature engineering,

learning-based anomaly detection, dataset quality control, and operational evaluation criteria.

This gap is important because broader cybersecurity and network anomaly detection studies show the value of supervised, deep, hybrid, and adaptive learning models [24], [27]–[29], while dataset-oriented studies emphasize the need for realistic, well-labeled, and representative evaluation data [25], [26]. However, these contributions do not fully address the procedure-dependent nature of SIP signaling in IMS/VoLTE networks.

A controlled IMS/VoLTE emulation environment, such as an Open5GS-based setup, can support the repeatable generation of normal and anomalous SIP traces. These traces can then be analyzed through message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level features, enabling future evaluation of machine learning models under reproducible, telecom-oriented conditions.

D. Implications for the Proposed Framework

The identified gaps suggest that a suitable SIP anomaly detection framework for IMS/VoLTE networks should satisfy five requirements.

First, it should be **procedure-aware**, meaning that it should reconstruct SIP transactions and dialogs rather than relying only on packet or message counts.

Second, it should be **multi-level**, combining message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level features.

Third, it should be **testbed-oriented**, allowing normal and anomalous SIP traces to be generated under controlled and repeatable conditions.

Fourth, it should consider **dataset quality**, including scenario realism, label consistency, traffic diversity, and separation between normal and anomalous signaling scenarios.

Fifth, it should support operational evaluation based not only on accuracy but also on false positive rate, detection latency, interpretability, computational cost, and suitability for SOC/NOC monitoring workflows.

These requirements form the basis of the conceptual multi-level machine learning framework presented in the next section.

VII. Proposed Conceptual Multi-Level Framework

Based on the research gaps identified in the previous section, this paper proposes a conceptual multi-level framework for SIP signaling anomaly detection in IMS/VoLTE networks. The framework is not presented as an implemented detection system; rather, it defines a realistic and testable path for later empirical evaluation in a controlled telecom emulation environment. This direction is consistent with the need for reproducible SIP/IMS signaling generation, controlled labeling, and AI/ML-based evaluation under isolated laboratory conditions.

The proposed framework connects six main activities: controlled dataset generation, SIP trace collection, SIP parsing, transaction and dialog reconstruction, multi-level feature extraction, and machine learning-based anomaly detection. For empirical evaluation, a controlled test environment based on Open5GS and IMS components such as Kamailio IMS can be used. Controlled user equipment or SIP traffic generators may produce SIP signaling, while packet captures and SIP logs can provide the data required for feature extraction and model evaluation. Normal and anomalous signaling scenarios should be executed only in an isolated laboratory environment and should not be connected to an operational network.

A. Framework Overview

The proposed framework is organized as a modular pipeline for SIP signaling anomaly detection in IMS/VoLTE networks. In a future validation setup, normal and anomalous SIP signaling scenarios can be generated in a controlled testbed. SIP traces can then be collected from suitable monitoring points, for example near the P-CSCF, at an SBC monitoring interface, or through an equivalent probe in the test environment. The collected SIP messages can be parsed into structured records, and related transactions and dialogs can be reconstructed using information such as Call-ID, CSeq, SIP method, response code, and message direction. After this step, features can be extracted at message, temporal, transaction, dialog, error, and entity levels. These feature vectors can then be used as input to learning-based models to distinguish normal and anomalous signaling behavior.

The conceptual workflow is illustrated in Fig. 2.

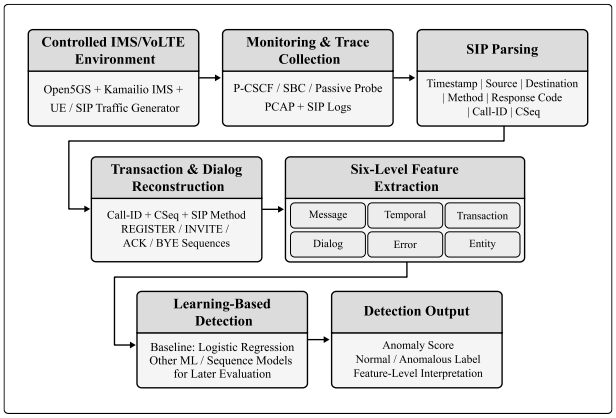


Fig. 2. Conceptual architecture of the proposed multi-level SIP signaling anomaly detection framework for IMS/VoLTE networks.

One advantage of this modular structure is that the different stages can be evaluated separately. For example, SIP parsing can be checked against captured traces, while transaction and dialog reconstruction can be compared with known REGISTER and INVITE procedures. Feature extraction can also be examined before the data are used for model training. In the first stage of future evaluation,

relatively simple and interpretable machine learning models can be considered. More complex sequence-based or deep learning models may be examined later if enough labeled data are available.

The same structure can also help with dataset quality control. When the signaling scenarios are produced under controlled conditions, the scenario labels, time intervals, traffic intensity, and observation points can be recorded more clearly than in operational traces. This can make later experiments more reproducible and keep the analysis focused on IMS/VoLTE-specific signaling behavior.

B. Controlled Telecom Emulation-Based Dataset Generation

A central requirement of the framework is reproducible dataset generation. Since real IMS/VoLTE traces are difficult to obtain because of privacy, confidentiality, and operational restrictions, a controlled telecom emulation testbed provides a practical alternative. Such a testbed allows normal procedures and anomaly scenarios to be defined in advance, scenario intervals to be recorded, and labels to be assigned more accurately than in uncontrolled operational traffic. This is important because dataset quality, label consistency, traffic diversity, and scenario realism strongly affect the reliability of anomaly detection evaluation [25], [26].

The dataset generation process should include both normal and anomalous SIP signaling scenarios. Normal scenarios may include IMS registration, authentication challenge and response, successful registration, VoLTE call establishment, ringing, session acceptance, call termination, re-registration, and failed but legitimate signaling attempts. Anomalous scenarios may include REGISTER flooding, INVITE flooding, malformed SIP messages, abnormal transaction behavior, incomplete dialogs, excessive error responses, and repeated failed registration or call attempts.

Each scenario can be repeated under controlled conditions. Relevant parameters may include the number of users, scenario duration, signaling rate, anomaly start time, anomaly duration, anomaly intensity, SIP method distribution, and observation point. Recording these parameters can improve the reproducibility of the dataset generation process and support later evaluation with supervised, semi-supervised, or anomaly score-based methods.

TABLE III Scenario groups for future dataset generation

Scenario group	Example procedures	Expected label
Normal signaling	Registration, call setup, call release, re-registration	Normal
Flooding anomaly	Repeated REGISTER or INVITE messages	Anomalous
Message anomaly	Malformed or incomplete SIP messages	Anomalous
Procedure anomaly	Incomplete transaction or abnormal dialog order	Anomalous
Error anomaly	Excessive 4xx or 5xx responses	Anomalous
Entity anomaly	Abnormal behavior from one user, IP address, or SIP identity	Anomalous

For binary classification, each observation window can be labeled as:

$$y_k = \begin{cases} 0, & normal \\ 1, & anomalous \end{cases} \quad (1)$$

For multiclass classification, the label may represent the specific scenario type, such as normal signaling, REGISTER flooding, INVITE flooding, malformed messages, abnormal dialogs, excessive error behavior, or entity-level anomalies. Labels can be assigned using scenario execution logs, predefined anomaly intervals, and known traffic-generation parameters.

C. SIP Message Parsing and Reconstruction

After trace collection, each SIP message should be converted into a structured event record. The i-th SIP event can be represented as:

$$e_i = (t_i, src_i, dst_i, method_i, code_i, callid_i, cseq_i, size_i, dir_i) \quad (2)$$

where t_i is the timestamp, src_i and dst_i are the source and destination entities, $method_i$ is the SIP method, $code_i$ is the SIP response code where applicable, $callid_i$ is the Call-ID value, $cseq_i$ is the CSeq value, $size_i$ is the message size, and dir_i is the signaling direction.

The collected SIP trace dataset can then be defined as:

$$D = \{e_1, e_2, \dots, e_n\} \quad (3)$$

where D denotes the complete set of collected SIP events and n is the total number of events.

Analyzing isolated SIP messages is not sufficient for procedure-aware anomaly detection. Therefore, SIP transactions and dialogs should be reconstructed. A SIP transaction can be identified using Call-ID, CSeq, and SIP method:

$$T_j = (callid_j, cseq_j, method_j) \quad (4)$$

where T_j represents the j-th SIP transaction.

A SIP dialog can be represented as an ordered sequence of related SIP methods and response codes:

$$G_j = \langle m_1, m_2, \dots, m_l \rangle \quad (5)$$

where G_j denotes the j-th dialog and $m_1, m_2, \dots, m_l$ are the ordered SIP messages or response codes in that dialog. A simplified successful call setup dialog can be represented as:

$$G_j = \langle INVITE, 100, 180, 200, ACK \rangle \quad (6)$$

This reconstruction step supports the detection of incomplete call setup, missing ACK messages, repeated INVITE attempts, unexpected BYE messages, failed transactions, and abnormal message ordering.

D. Multi-Level Feature Representation

The proposed framework represents SIP signaling behavior at six complementary levels: message, temporal, transaction, dialog, error, and entity levels. The purpose of this representation is to avoid relying on a single traffic indicator. Some anomalies mainly change the number or distribution of SIP messages, while others become visible only after transactions or dialogs are reconstructed.

A time-based observation window is defined as:

$$W_k = [t_k, t_k + \Delta] \tag{7}$$

where W_k denotes the k-th observation window, t_k is its starting time, and Δ is the window duration. Let n_k denote the total number of SIP messages observed in W_k.

For each observation window, the final feature vector can be expressed as:

$$\mathbf{X}_k = [\mathbf{X}_k^{msg}, \mathbf{X}_k^{temp}, \mathbf{X}_k^{tx}, \mathbf{X}_k^{dlg}, \mathbf{X}_k^{err}, \mathbf{X}_k^{ent}] \tag{8}$$

where the six components represent message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level features, respectively. The following definitions provide representative examples of how features from these levels can be calculated. They are not intended to define an exhaustive feature set.

TABLE IV Proposed multi-level SIP feature representation

Feature level	Example indicators	Main purpose
Message-level	Method counts, method proportions $q_{k,m}$, request/response ratio	Detect changes in SIP message composition
Temporal-level	Message rate λ_k, inter-arrival time, arrival entropy H_{arr}, burstiness	Detect timing changes and flooding behavior
Transaction-level	Completion ratio R_{tx}, failure ratio, retransmission ratio	Detect incomplete or failed transactions
Dialog-level	Transition matrix $\mathbf{P}_k$, dialog completion, invalid transitions, missing ACK ratio	Detect abnormal SIP procedure flow
Error-level	Response ratio $R_{k,c}$, 4xx/5xx ratios, timeout ratio	Detect abnormal response behavior
Entity-level	Source concentration C_{src}, unique sources, per-user activity	Detect abnormal source or identity behavior

1. Message-Level Features

Message-level features describe the composition of SIP requests and responses within an observation window. Let $n_{k,m}$ denote the number of messages belonging to SIP method m, such as REGISTER, INVITE, ACK, BYE, OPTIONS, or MESSAGE. The relative frequency of method m can be defined as:

$$q_{k,m} = \frac{n_{k,m}}{n_k} \tag{9}$$

where $q_{k,m}$ represents the proportion of method m in W_k. Changes in these proportions can indicate abnormal signaling behavior even when the total message rate is not unusually high.

Other message-level indicators may include the total number of requests and responses, request-to-response ratio, retransmitted message count, and the relative frequency of selected SIP methods. These features are particularly useful for identifying flooding, scanning-like behavior, or sudden changes in SIP message composition.

2. Temporal-Level Features

Temporal-level features describe how SIP messages are distributed over time. A basic feature is the signaling rate

$$\lambda_k = \frac{n_k}{\Delta} \tag{10}$$

where λ_k represents the average number of SIP messages per unit of time in W_k.

To capture the distribution of message arrivals within the window, W_k can be divided into B equal sub-intervals. Let $n_{k,b}$ denote the number of SIP messages observed in sub-interval b. The corresponding arrival probability is:

$$p_{k,b} = \frac{n_{k,b}}{n_k} \tag{11}$$

and the message-arrival entropy can be defined as:

$$H_{arr}(W_k) = -\sum_{b=1}^{B} p_{k,b} \log_2 p_{k,b} \tag{12}$$

where terms with $p_{k,b} = 0$ are treated as zero. This measure reflects how uniformly signaling messages are distributed within the observation window. A concentrated burst of messages and a more evenly distributed signaling pattern may therefore produce different entropy values even when their total message counts are similar.

Other temporal features may include mean inter-arrival time, variance of inter-arrival time, burst duration, and short-term signaling intensity.

3. Transaction-Level Features

Transaction-level features are obtained after reconstructing SIP request-response relationships. Let T_k denote the set of transactions observed in W_k, and let T_k^c denote the subset of transactions that reach an expected completion state. The transaction completion ratio can be defined as:

$$R_{tx}(W_k) = \frac{|T_k^c|}{|T_k|} \tag{13}$$

where $|T_k|$ is the total number of reconstructed transactions in the window.

A low completion ratio may indicate repeated failures, incomplete exchanges, abnormal retransmissions, or procedure disruption. Additional transaction-level features may include transaction failure ratio, retransmission ratio, REGISTER success ratio, INVITE failure ratio, and authentication challenge ratio.

These features provide information that cannot be obtained from individual SIP messages alone because they consider the relationship between requests and their corresponding responses.

4. Dialog-Level Features

Dialog-level features describe the logical progression of SIP procedures. A dialog can be represented as a sequence of signaling states or ordered message events. Let s_i and s_j

denote two possible dialog states, and let $N_k(i,j)$ represent the number of observed transitions from s_i to s_j in W_k.

The transition probability from state s_i to state s_j can be defined as:

$$P_k(i,j) = \frac{N_k(i,j)}{\sum_{l=1}^{M} N_k(i,l)} \quad (14)$$

where M is the number of considered dialog states. The transition probability matrix is then written as:

$$\mathbf{P}_k = [P_k(i,j)]_{M\times M} \quad (15)$$

This representation provides a compact description of SIP procedure progression inside an observation window. For example, normal call establishment is expected to follow known transitions involving INVITE, provisional responses, 200 OK, and ACK. Missing ACK messages, unexpected BYE messages, repeated states, or invalid response ordering may change the observed transition structure.

The transition matrix can be used directly as part of the feature representation or summarized through features such as dialog completion ratio, invalid transition count, missing ACK ratio, and incomplete INVITE dialog ratio.

5. Error-Level Features

Error-level features describe the distribution of SIP response codes. Let $N_{k,c}$ denote the number of responses belonging to a selected response code or response class c, and let N_k^{resp} denote the total number of SIP responses in W_k. The response ratio for c can be defined as:

$$R_{k,c} = \frac{N_{k,c}}{N_k^{resp}} \quad (16)$$

The value of c may represent a complete response class, such as 4xx or 5xx, or an individual SIP response such as 403, 408, or 503.

Error-level features may therefore include the 4xx response ratio, 5xx response ratio, timeout ratio, authentication failure ratio, and frequency of selected error responses. Individual SIP codes should be interpreted in the context of the SIP procedure because some responses, such as authentication challenges, may also occur during legitimate IMS procedures.

A persistent change in these ratios may indicate registration problems, malformed signaling, overload, abnormal call attempts, or other forms of signaling misuse.

6. Entity-Level Features

Entity-level features describe the behavior of individual sources or SIP identities. Let U_k denote the set of active source entities in W_k, and let $n_{k,u}$ denote the number of SIP messages generated by entity u. A source concentration index can be defined as:

$$C_{src}(W_k) = \sum_{u\in U_k} \left(\frac{n_{k,u}}{n_k}\right)^2 \quad (17)$$

This index increases when a larger proportion of SIP signaling is generated by a small number of sources. It can therefore help distinguish distributed traffic from behavior dominated by one or a few entities.

Depending on the monitoring point, the entity identifier may correspond to a source IP address, SIP URI, IMS identity, user equipment, or another observable signaling entity. Additional entity-level features may include the number of unique sources, per-user registration rate, per-entity failure ratio, and the number of transactions initiated by each source.

Combined Feature Representation

The six feature groups provide different views of the same SIP signaling activity. Message and temporal features are useful for identifying changes in traffic volume and timing, while transaction and dialog features provide information about procedure completion and signaling order. Error-level features describe abnormal response behavior, and entity-level features indicate whether suspicious activity is concentrated around specific sources or identities.

The resulting feature vector $\mathbf{X}_k$ can include both directly measured indicators and derived features such as $H_{arr}(W_k)$, $R_{tx}(W_k)$, $\mathbf{P}_k$, $R_{k,c}$, and $C_{src}(W_k)$. These definitions provide a formal basis for the multi-level representation while allowing the exact feature set to be adjusted during later empirical evaluation.

E. Machine Learning-Based Detection Pipeline

The proposed framework is not tied to a single learning model. Different models may be suitable depending on the availability of labeled data, class imbalance, feature representation, computational requirements, and the need for explainable decisions. Table V summarizes several candidate models for future evaluation. Logistic regression is retained as a baseline because it is simple, interpretable, and directly applicable to the structured multi-level feature vector.

TABLE V Candidate Models for SIP Anomaly Detection

Model	Suitable when	Main limitation
Logistic Regression	Labeled structured features; interpretable baseline	Limited nonlinear modeling
Random Forest / Gradient Boosting	Labeled data with nonlinear feature relations	Less interpretable than LR
One-Class SVM	Mainly normal data; few labeled anomalies	Parameter-sensitive; limited scalability
Isolation Forest	Rare anomalies; limited labels	Does not model SIP sequence order
Autoencoder	Large mainly normal dataset	Low explainability; threshold selection
RNN / LSTM / GRU	Ordered SIP transaction or dialog sequences	High data and computational requirements

Model selection depends on both data availability and operational requirements. Logistic regression and tree-based models are suitable when labeled structured features are available, while One-Class SVM and Isolation Forest are useful when anomalous samples are limited or highly

imbalanced. Autoencoders can be considered when a larger amount of normal traffic is available without reliable attack labels. Sequence models are more appropriate when the order of SIP messages and dialog progression are important. In operational IMS/VoLTE monitoring, model selection should also consider explainability, computational cost, scalability, and false-positive behavior.

For the k-th observation window W_k, the extracted multi-level feature vector is denoted by X_k. The anomaly score s_k is computed using a logistic regression function:

$$s_k = P(y_k = 1|X_k) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^T X_k)}} \tag{18}$$

where $y_k = 1$ denotes an anomalous window, X_k is the feature vector extracted from W_k, β_0 is the intercept, and β is the learned coefficient vector. This formulation follows the standard logistic regression model for probabilistic binary classification [30]. The value of s_k is between 0 and 1 and represents the estimated probability that the SIP signaling window is anomalous. In anomaly detection, such a score-based representation is commonly used to support decision making before applying a final classification threshold [31].

The predicted class is obtained by applying a decision threshold τ to the anomaly score:

$$\hat{y}_k(\tau) = \begin{cases} 1, & s_k \geq \tau \\ 0, & s_k < \tau \end{cases} \tag{19}$$

where $\hat{y}_k = 1$ indicates an anomalous SIP signaling window and $\hat{y}_k = 0$ indicates normal signaling behavior. The threshold τ controls the sensitivity of the detection system. Score-producing classifiers can be converted into binary classifiers by applying different thresholds, which also produces different ROC operating points [32].

For each threshold value, the true positive rate (TPR) is calculated as:

$$TPR(\tau) = \frac{TP(\tau)}{P} \tag{20}$$

where TP(τ) is the number of anomalous windows correctly detected at threshold τ, and P is the total number of actual anomalous windows.

Similarly, the false positive rate (FPR) is calculated as:

$$FPR(\tau) = \frac{FP(\tau)}{N} \tag{21}$$

where FP(τ) is the number of normal windows incorrectly classified as anomalous at threshold τ, and N is the total number of actual normal windows. The use of TPR and FPR follows standard ROC-based classifier evaluation [32].

Therefore, each threshold produces one ROC operating point:

$$ROC(\tau) = \big(FPR(\tau), TPR(\tau)\big) \tag{22}$$

This formulation also allows the decision threshold to be adjusted according to operational requirements. After multi-level SIP features are extracted from each observation window W_k, the logistic regression model can assign an anomaly probability s_k. This probability can then be compared with a selected threshold τ to classify each window as normal or anomalous. Changing τ provides a trade-off between sensitivity to suspicious signaling behavior and the number of false alarms. The resulting decisions can be assessed using $TPR(\tau)$, $FPR(\tau)$, and ROC analysis to identify suitable operating points for IMS/VoLTE monitoring.

F. Recommended Evaluation Criteria

Although this paper does not present experimental results, the proposed framework defines evaluation criteria for future experimental implementation. Standard classification metrics such as accuracy, precision, recall, and F1-score should be used to evaluate detection performance. However, telecom monitoring also requires operational metrics such as false positive rate, detection latency, and feature importance.

False positive rate is important because excessive false alarms reduce the usefulness of a detection system in SOC/NOC environments. Detection latency is important because signaling attacks may affect service availability within a short period. Feature importance analysis is also necessary because it helps determine whether the model relies mainly on message volume, timing behavior, transaction failures, dialog abnormalities, error responses, or entity-level activity.

A future experimental study should compare single-level feature groups against the combined multi-level feature vector. This comparison will directly test whether the proposed multi-level representation provides additional value compared with message-level, temporal-level, transaction-level, dialog-level, error-level, or entity-level features alone. The future evaluation should also consider dataset balance, scenario-based splitting, and testing across different anomaly intensities to reduce overfitting and improve reproducibility.

VIII. Future Research Directions

Building on the taxonomy and conceptual framework presented in this study, future research should move toward controlled implementation and empirical validation. In particular, further work should focus on emulation-based IMS/VoLTE testbed development, reproducible dataset generation and labeling, multi-level feature evaluation, real-time monitoring, and collaborative SIP anomaly detection.

A. Open5GS-Based IMS/VoLTE Testbed Implementation

A primary direction for future work is the construction of a controlled IMS/VoLTE-oriented testbed. An Open5GS-based mobile core can provide the foundation for emulating LTE/EPC network behavior, while IMS-related components such as Kamailio IMS can be integrated to support SIP-based registration and VoLTE signaling procedures. This environment would allow normal and anomalous signaling

scenarios to be executed under controlled and repeatable conditions.

The testbed should include defined subscriber profiles, controlled user equipment or SIP user agents, IMS signaling components, packet capture points, and logging mechanisms. Where available, COTS user equipment and dedicated test SIM configurations can improve the realism of signaling traces. However, even when physical devices are not available, controlled SIP user agents and emulated traffic sources can still support early-stage validation of registration, call setup, call release, and abnormal signaling scenarios.

The testbed must remain isolated from operational networks. This is essential to ensure safe experimentation, ethical data generation, and reproducible validation of SIP anomaly detection methods.

B. Controlled Dataset Generation and Labeling

A second research direction is the development of a reproducible SIP signaling dataset generated from controlled IMS/VoLTE scenarios. The dataset should include normal procedures such as registration, authentication, successful call establishment, call release, re-registration, and legitimate failed attempts. It should also include selected anomaly scenarios, such as REGISTER flooding, INVITE flooding, malformed SIP messages, incomplete dialogs, excessive error responses, and abnormal entity behavior.

Dataset generation should be scenario-based. For each scenario, important parameters should be recorded, including scenario type, start time, end time, number of users, signaling rate, anomaly duration, and anomaly intensity. These records can then support accurate labeling of SIP traces.

The labeling process should support both binary and multiclass classification. Binary labels distinguish normal and anomalous traffic, while multiclass labels identify specific anomaly categories such as REGISTER flooding, INVITE flooding, malformed signaling, abnormal dialog behavior, or error-response anomalies. To reduce data leakage, future studies should split datasets at the scenario-run level rather than only at the individual-window level.

C. Feature Engineering and Model Evaluation

Future empirical work should evaluate the proposed multi-level feature representation. Message-level, temporal-level, transaction-level, dialog-level, error-level, and entity-level features should be extracted from the collected SIP traces and evaluated separately and jointly.

A key experiment should compare single-level feature groups with the combined multi-level feature vector. This comparison would show whether combining several SIP behavior levels provides measurable improvement compared with using only message counts, timing indicators, transaction statistics, or dialog patterns.

Initial model evaluation can use classical machine learning algorithms such as decision trees, random forests, support vector machines, k-nearest neighbors, and gradient boosting. These models are suitable for structured tabular features and provide a practical starting point. Adaptive learning, transfer learning, and hybrid deep learning models may also be examined, particularly for DDoS-oriented or sequence-based SIP anomaly detection [24], [29]. However, such models should be evaluated against interpretable classical methods to determine whether their added complexity provides practical benefit.

Evaluation should include accuracy, precision, recall, F1-score, false positive rate, detection latency, confusion matrix analysis, and feature importance. Feature importance analysis is especially useful because it can indicate which SIP behavior levels contribute most to anomaly detection.

D. Real-Time Monitoring and Operational Deployment

Another direction is the adaptation of the proposed framework for real-time or near-real-time SIP monitoring. In practical IMS/VoLTE environments, an anomaly detection system should not only perform well in offline evaluation but should also support timely alert generation and operational interpretation.

Potential monitoring locations include P-CSCF-adjacent probes, SBC monitoring layers, passive packet capture points, and SOC/NOC analytics platforms. A real-time implementation would require efficient SIP parsing, sliding-window feature extraction, low-latency classification, and alert prioritization.

Operational deployment also requires careful control of false positives. Excessive alarms can reduce analyst trust and increase workload. Therefore, future work should evaluate different window sizes, alert thresholds, and anomaly scoring strategies. Alerts should also be interpretable, indicating whether the detected anomaly is mainly related to traffic volume, transaction failure, abnormal dialog behavior, excessive errors, or entity-level irregularity.

E. Federated and Collaborative SIP Anomaly Detection

Federated and collaborative learning represent another promising direction for SIP anomaly detection. IMS/VoLTE signaling traces may contain sensitive subscriber, operator, and service information, which makes raw data sharing difficult. Federated learning can help address this limitation by allowing distributed training without transferring raw SIP traces between domains.

Future studies can examine whether multiple testbeds, operators, or monitoring domains can train local anomaly detection models and share only model parameters or updates. This direction is especially relevant for distributed SIP-DDoS detection and collaborative telecom security monitoring.

However, federated SIP anomaly detection introduces several challenges. Different environments may use different IMS configurations, SIP implementations, traffic patterns, user behavior, and security policies. These differences can create heterogeneous datasets and affect model

generalization. Therefore, collaborative approaches should be supported by standardized feature extraction, consistent labeling rules, privacy-preserving aggregation, and evaluation across diverse signaling scenarios.

These research directions can be followed as a gradual validation path. The first stage is the implementation of a controlled IMS/VoLTE testbed and the generation of reproducible SIP traces. The second stage is the evaluation of the proposed feature groups and the comparison of detection models using the same labeled scenarios. The third stage should examine class imbalance, false positives, detection latency, explainability, and generalization across different signaling conditions. After these steps, the framework can be extended toward real-time monitoring and collaborative or federated detection across multiple environments. This progression allows each part of the conceptual framework to be evaluated before moving toward more complex deployment scenarios.

IX. Conclusion

SIP signaling plays a central role in IMS/VoLTE networks by supporting registration, authentication, session establishment, call control, and service continuity. At the same time, its procedure-dependent behavior makes it vulnerable to different forms of signaling anomalies, including flooding attacks, registration abuse, malformed messages, abnormal transaction and dialog sequences, abnormal error responses, and entity-level irregularities. These anomalies can affect service availability, signaling stability, and user experience in mobile operator environments.

This paper reviewed existing studies on SIP, VoIP, IMS, and VoLTE security and classified them according to their target environment, anomaly type, detection approach, dataset or testbed usage, procedure awareness, and feature representation. The review showed that previous studies have made important contributions to SIP vulnerability analysis, flooding detection, statistical monitoring, machine learning-based detection, deep learning-based dialog analysis, and IMS/VoLTE vulnerability assessment. However, the analysis also indicated that many existing approaches focus on generic VoIP environments, isolated attack types, or limited feature levels.

To address these limitations, this paper proposed a conceptual multi-level machine learning framework for SIP signaling anomaly detection in IMS/VoLTE networks. The framework emphasizes controlled telecom emulation, SIP message parsing, transaction and dialog reconstruction, and feature extraction across six complementary levels: message, temporal, transaction, dialog, error, and entity levels. This structure provides a procedure-aware foundation for future anomaly detection research and supports more realistic evaluation than approaches that rely only on traffic volume or single-level indicators.

Future work should focus on implementing the proposed framework in a controlled IMS/VoLTE-oriented testbed, such as an Open5GS-based environment integrated with IMS-related components. Further research should also focus on developing reproducible SIP signaling datasets, evaluating single-level and multi-level feature representations, comparing machine learning models, and assessing operational criteria such as false positive rate, detection latency, feature importance, and deployment feasibility. These directions can support the development of more practical, reproducible, and telecom-oriented SIP anomaly detection solutions for IMS/VoLTE security monitoring.

REFERENCES

[1] D. Geneiatakis, T. Dagiuklas, G. Kambourakis, C. Lambrinoudakis, S. Gritzalis, S. Ehlert, and D. Sisalem, “Survey of security vulnerabilities in Session Initiation Protocol,” IEEE Communications Surveys & Tutorials, vol. 8, no. 3, pp. 68–81, 2006, doi: 10.1109/COMST.2006.253270.

[2] M. A. Akbar, Z. Tariq, and M. Farooq, “A comparative study of anomaly detection algorithms for detection of SIP flooding in IMS,” in Proc. 2nd International Conference on Internet Multimedia Services Architecture and Applications (IMSAA), Bangalore, India, 2008, pp. 1–6, doi: 10.1109/IMSAA.2008.4753934.

[3] A. H. Farooqi and A. Munir, “Intrusion detection system for IP Multimedia Subsystem using K-nearest neighbor classifier,” in Proc. 2008 IEEE International Multitopic Conference (INMIC), 2008, pp. 423–428, doi: 10.1109/INMIC.2008.4777775.

[4] M. Nassar, R. State, and O. Festor, “Monitoring SIP traffic using support vector machines,” in Recent Advances in Intrusion Detection, Lecture Notes in Computer Science, vol. 5230, Berlin, Germany: Springer, 2008, pp. 311–330, doi: 10.1007/978-3-540-87403-4_17.

[5] Y. Rebahi, M. S. Ramzan, and T. Magedanz, “Detecting flooding attacks against IP Multimedia Subsystem (IMS) networks,” in Proc. 2008 IEEE/ACS International Conference on Computer Systems and Applications (AICCSA), 2008, pp. 128–135, doi: 10.1109/AICCSA.2008.4493627.

[6] S. Ehlert, D. Geneiatakis, and T. Magedanz, “Survey of network security systems to counter SIP-based denial-of-service attacks,” Computers & Security, vol. 29, no. 2, pp. 225–243, Mar. 2010, doi: 10.1016/j.cose.2009.09.004.

[7] M. Nassar, R. State, and O. Festor, “Labeled VoIP data-set for intrusion detection evaluation,” in Networked Services and Applications—Engineering, Control and Management, Lecture Notes in Computer Science, vol. 6164, Berlin, Germany: Springer, 2010, pp. 97–106, doi: 10.1007/978-3-642-13971-0_10.

[8] E. Nosrati, A. S. Kashi, Y. Darabian, and S. N. H. Tonekaboni, “Register flooding attacks detection in IP Multimedia Subsystems by using adaptive z-score CUSUM algorithm,” in Proc. 2011 International Conference on Information Technology and Multimedia (ICIMU), 2011, doi: 10.1109/ICIMU.2011.6122765.

[9] Z. Asgharian, H. Asgharian, A. Akbari, and B. Raahemi, “Detecting denial of service message flooding attacks in SIP based services,” AUT Journal of Electrical Engineering, vol. 44, no. 1, pp. 75–85, Spring 2012, doi: 10.22060/eej.2012.44.

[10] D. Seo, H. Lee, and E. Nuwere, “SIPAD: SIP-VoIP anomaly detection using a stateful rule tree,” Computer Communications,

vol. 36, no. 5, pp. 562–574, Mar. 2013, doi: 10.1016/j.comcom.2012.12.004.
[11] H. Kim, D. Kim, M. Kwon, H. Han, Y. Jang, D. Han, T. Kim, and Y. Kim, "Breaking and fixing VoLTE: Exploiting hidden data channels and mis-implementations," in Proc. 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), 2015, pp. 328–339, doi: 10.1145/2810103.2813718.
[12] E. Ko, S. Park, S. Kim, K. Son, and H. Kim, "SIP amplification attack analysis and detection in VoLTE service network," in Proc. International Conference on Information Networking (ICOIN), 2016, pp. 334–336, doi: 10.1109/ICOIN.2016.7427126.
[13] S. Zhang, L. Zhou, M. Wu, Z. Tang, N. Ruan, and H. Zhu, "Automatic detection of SIP-aware attacks on VoLTE device," in Proc. IEEE 84th Vehicular Technology Conference (VTC-Fall), 2016, pp. 1–5, doi: 10.1109/VTCFall.2016.7880916.
[14] Y.-H. Lu, C.-Y. Li, Y.-Y. Li, S. H.-Y. Hsiao, T. Xie, G.-H. Tu, and W.-X. Chen, "Ghost calls from operational 4G call systems: IMS vulnerability, call DoS attack, and countermeasure," in Proc. 26th Annual International Conference on Mobile Computing and Networking (MobiCom), 2020, pp. 96–109, doi: 10.1145/3372224.3380885.
[15] Z. Cheng, M. Ordean, F. D. Garcia, B. Cui, and D. Rys, "Watching your call: Breaking VoLTE privacy in LTE/5G networks," Proceedings on Privacy Enhancing Technologies, vol. 2023, no. 2, pp. 282–297, 2023, doi: 10.56553/popets-2023-0053.
[16] J. Shi, T. Xie, G.-H. Tu, C. Peng, C.-Y. Li, A. Hou, S. Wang, Y. Hu, X. Lei, M.-Y. Chen, L. Xiao, and X. Liu, "When good turns evil: Encrypted 5G/4G voice calls can leak your identities," in Proc. 2023 IEEE Conference on Communications and Network Security (CNS), 2023, pp. 1–9, doi: 10.1109/CNS59707.2023.10288900.
[17] W. Nazih, Y. Hifny, W. S. Elkilani, H. Dhahri, and T. Abdelkader, "Countering DDoS attacks in SIP based VoIP networks using recurrent neural networks," Sensors, vol. 20, no. 20, p. 5875, Oct. 2020, doi: 10.3390/s20205875.
[18] D. Pereira, R. Oliveira, and H. S. Kim, "A machine learning approach for prediction of signaling SIP dialogs," IEEE Access, vol. 9, pp. 44094–44106, Mar. 2021, doi: 10.1109/ACCESS.2021.3065660.
[19] D. Pereira, R. Oliveira, and H. S. Kim, "Classification of abnormal signaling SIP dialogs through deep learning," IEEE Access, vol. 9, pp. 165557–165567, Dec. 2021, doi: 10.1109/ACCESS.2021.3135195.
[20] D. Pereira and R. Oliveira, "Detection of abnormal SIP signaling patterns: A deep learning comparison," Computers, vol. 11, no. 2, p. 27, Feb. 2022, doi: 10.3390/computers11020027.
[21] J. Shi, S. Wang, M.-Y. Chen, G.-H. Tu, T. Xie, M.-H. Chen, Y. Hu, C.-Y. Li, and C. Peng, "IMS is not that secure on your 5G/4G phones," in Proc. 30th Annual International Conference on Mobile Computing and Networking (MobiCom), 2024, pp. 513–527, doi: 10.1145/3636534.3649377.
[22] C. Alvares, D. Dinesh, S. Alvi, T. Gautam, M. Hasib, and A. Raza, "Dataset of attacks on a live enterprise VoIP network for machine learning based intrusion detection and prevention systems," Computer Networks, vol. 197, p. 108283, Oct. 2021, doi: 10.1016/j.comnet.2021.108283.
[23] O. Sbai, B. Allaert, P. Sondi, and A. Meddahi, "SIP-DDoS framework based on federated learning for collaborative anomaly detection," Cybersecurity, vol. 9, no. 1, p. 109, Jan. 2026, doi: 10.1186/s42400-025-00527-6.
[24] R. E. Yoro, M. D. Okpor, M. I. Akazue, E. A. Okpako, A. O. Eboka, P. O. Ejeh, A. A. Ojugo, C. C. Odiakaose, A. P. Binitie, and R. E. Ako, "Adaptive DDoS detection mode in software-defined SIP-VoIP using transfer learning with boosted meta-learner," PLOS One, vol. 20, no. 6, p. e0326571, Jun. 2025, doi: 10.1371/journal.pone.0326571.
[25] A. Villafranca, I. Tasic, and M.-D. Cano, "TRUSTLab dataset: A real-world CICFlowMeter dataset for IoT/edge intrusion detection," Frontiers in Computer Science, vol. 8, p. 1803271, May 2026, doi: 10.3389/fcomp.2026.1803271.
[26] O. Apejoye, N. Ajienka, J. He, and X. Ma, "Critical review of network intrusion detection benchmark datasets for practical IoT security," Computer Networks and Communications, vol. 3, no. 2, pp. 182–208, Dec. 2025, doi: 10.37256/cnc.3220257228.
[27] S. Kamble, S. Patil, M. Alone, et al., "Cyber hacking breaches prediction and detection using machine learning," in Proc. 2025 12th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET-SIP), 2025, doi: 10.1109/ICETETSIP64213.2025.11156203.
[28] M. V. Doraiswamy and P. K. Adakane, "AI powered threat detection in cybercrime using supervised ML models," in Proc. 2025 12th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET-SIP), 2025, doi: 10.1109/ICETETSIP64213.2025.11156267.
[29] A. S. S. Mukhtar and R. N. Jugele, "Optimizing network security: A CRNN architecture for firewall anomaly detection," in Proc. 2025 12th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET-SIP), 2025, doi: 10.1109/ICETETSIP64213.2025.11156857.
[30] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.
[31] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, Art. no. 15, pp. 1–58, Jul. 2009, doi: 10.1145/1541880.1541882.
[32] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[33] K. Jahani, B. Moshiri, and B. H. Khalaj, "Secure PDM: A novel Byzantine Fault Tolerant federated learning framework using a robust PCA-based anomaly detection approach," International Journal of Industrial Electronics, Control and Optimization (IECO), vol. 8, no. 4, 2025, doi: 10.22111/ieco.2025.51226.1668.

Appendix A



Ali Kuhestani (Member, IEEE) received the Ph.D. degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 2017. He is currently an Associate Professor with the Department of Communications and Electronics, Faculty of Electrical and Computer Engineering, Qom University of Technology, Qom, Iran. His research interests include physical-layer security of wireless communications, the Internet of Things, millimeter-wave communications, massive MIMO systems, and space-time coding.



Mojtaba Karimi received the B.Sc. degree in Software Engineering from Shahid Shamsipour Technical and Vocational College (STVC), Tehran, Iran. He is currently working toward the M.Sc. degree in Communications Systems at Qom University of Technology (QUT), Qom, Iran. He is currently a Telecommunication Network Security Researcher, with a particular focus on the security analysis and assessment of mobile communication networks. His research experience includes security studies of 2G, 3G, 4G/LTE, IMS, and VoLTE networks, with emphasis on signaling security and vulnerability assessment. His research interests include mobile network security, IMS/VoLTE security, SIP signaling analysis, 5G security, and the application of AI/ML to anomaly detection, traffic analysis, and intrusion detection in mobile communication networks.



Mohammadreza Keshavarzi received the Ph.D. degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 2009. Since 2012, he has been with the Iran Telecommunication Research Center (ITRC), Tehran, Iran, where he is currently an Assistant Professor. His research interests include physical-layer security of wireless communications, the Internet of Things, and MIMO systems.