IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.xxxx/ACCESS.20xx.DOI

Correlation-Based Abnormal SIP Dialog Identification: A Performance Comparison with Bayesian and Deep Learning Approaches

CLARISSE FEIO[1,2], DIOGO PEREIRA[1,2], RODOLFO OLIVEIRA[1,2], (Senior Member, IEEE), PEDRO AMARAL[1,2],

[1]Departamento de Engenharia Electrotécnica e de Computadores, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[2]Instituto de Telecomunicações, 1049-001 Lisbon, Portugal

Corresponding author: Clarisse Feio (c.feio@campus.fct.unl.pt).

This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by project reference UIDB/50008/2020, and DOI identifier https://doi.org/10.54499/UIDB/50008/2020, and by project reference 2022.08786.PTDC and DOI identifier https://doi.org/10.54499/2022.08786.PTDC.

ABSTRACT The Session Initiation Protocol (SIP) is crucial in establishing, maintaining, and terminating multimedia sessions. It is particularly vital for the operation of 4G/5G networks, where the network's low latency and high reliability enable advanced services such as real-time video streaming and Internet of Things applications. The widespread use of SIP in various network generations emphasizes the need for robust security mechanisms to protect against potential vulnerabilities. SIP is susceptible to various attacks, including registration hijacking, call tampering, and denial of service. A specific threat arises from exploiting unknown or abnormal SIP dialogs to uncover weaknesses in the different SIP implementations running on servers. In this paper, we propose an innovative methodology for anomalous SIP dialog detection based on prior knowledge of observed correct and anomalous SIP dialogs. The proposed approach leverages cross-correlation techniques to score the similarity of the SIP dialogs and the use of statistical metrics to classify the anomalous ones. Our method achieves an accuracy of approximately 98.91%. We compare its performance with the optimal Bayesian solution, a deep learning-based approach, and a hybrid method using both deep-learning and statistical methods. While our solution is close to the optimal accuracy, it does not achieve the lowest false alarm rate. However, it offers a significant advantage in computational efficiency, being over 1000 times faster than both the optimal Bayesian and deep learning methods. These findings underscore the potential of the proposed technique for real-time detection of abnormal SIP dialogs in high-performance network environments. Cross-correlation was also employed to predict the SIP ID of ongoing SIP dialogs before their full arrival. Although this method was faster than the other studied methods, its predictive performance was suboptimal, achieving high accuracy only when over 90% of the data was available. Based on these findings, we conclude that the proposed method has high performance in classification tasks with faster computational times than alternative methods, while it is less effective for prediction tasks were other methods achieve higher performance.

INDEX TERMS Session Initiation Protocol, Security, Anomalous SIP Dialogs, Performance Analysis.

I. INTRODUCTION

The Session Initiation Protocol (SIP) [1] is a fundamental protocol used to initiate, maintain, and terminate real-time communication sessions in multimedia services, such as voice, video, and messaging. SIP's versatility has made it a key component of modern communication networks, particularly in 4G, where it enables a range of applications, including Voice over IP (VoIP) [2], video conferencing, and push-to-talk services. As networks evolve to 5G, SIP remains critical [3], facilitating the ultra-reliable low-latency commu-

nications (URLLC) and massive machine-type communications (mMTC) necessary for advanced services such as real-time video streaming, Internet of Things (IoT) integration, and augmented reality. Looking ahead, SIP is expected to play an even more prominent role in future generation networks [4], supporting novel applications such as immersive communications, tactile internet, and seamless integration of human-centric services.

The security of the SIP protocol is crucial because SIP serves as the backbone of many real-time communication services, including voice calls, video conferencing, and instant messaging, which are essential for both personal and business use [5]. The protocol's widespread adoption in critical infrastructures, such as telecommunication networks and emergency services, makes it a prime target for malicious activities. Any compromise in SIP security can lead to severe consequences, such as service disruption, unauthorized access to sensitive communications, financial losses due to toll fraud, and breaches of privacy. Moreover, the growing use of SIP in modern networks, including 4G and 5G, expands its attack surface, making it a critical component to secure in increasingly complex and interconnected environments [6]. Ensuring robust SIP security protects the integrity, confidentiality, and availability of communication services, thereby maintaining user trust and safeguarding the underlying network infrastructure from potential threats.

Despite its widespread adoption, SIP is susceptible to various security threats. The openness and extensibility of SIP make it vulnerable to several types of attacks. Attacks encompass a variety of strategies that aim to disrupt communication services or exploit system vulnerabilities [7]. Common types of SIP attacks include registration hijacking, where attackers manipulate registration requests to redirect calls to unauthorized devices; call tampering, involving the interception or modification of SIP messages to alter session parameters; message spoofing, in which attackers forge SIP messages to impersonate legitimate users, and denial of service (DoS) attacks, which overwhelm the server with excessive SIP requests to degrade or disable services.

SIP dialog attacks pose a significant threat due to the protocol's complex and stateful nature, which can be exploited to perform sophisticated attacks that evade traditional security mechanisms. A particularly challenging issue is the identification of abnormal or unexpected SIP dialogs, which may exploit vulnerabilities in different SIP implementations running on servers to cause disruptions or compromise the system's integrity [8]. Attackers often exploit SIP dialog attacks because they target the sequence of messages exchanged during a session, allowing them to uncover potential implementation flaws in the server's handling of SIP dialogs. By injecting, modifying, or reordering SIP messages, attackers can exploit those weaknesses to bypass security controls, manipulate call routing, or disrupt the establishment and maintenance of communication sessions. Abnormal dialogs can be used to discover and exploit weaknesses, highlighting the need for robust mechanisms to detect potentially harmful SIP activity in real time.

In this paper, we propose an innovative approach for detecting abnormal SIP dialogs based on prior knowledge of past observed SIP messages in the server. The technique leverages cross-correlation between SIP dialog patterns and its statistics to classify ongoing SIP communications, allowing for the identification of dialogs that deviate from expected behavior and may pose a potential threat. The proposed methodology achieves an accuracy of 98.91% with a false alarm rate of approximately 8.6% on a realistic benchmark dataset, demonstrating its capability to accurately distinguish between normal and abnormal SIP dialogs. Although prior works have already focused on detecting abnormal SIP dialogs, existing approaches have primarily relied on optimal Bayesian techniques [9] or deep learning methodologies [10]–[12] to identify anomalies in SIP communications. These methods, while effective, often involve substantial computational complexity or require extensive training datasets to achieve high detection accuracy. In contrast, this paper introduces a novel method based on the cross-correlation of SIP dialogs and the statistics derived from these correlations. By analyzing the relationships and patterns in the sequences of SIP messages, our approach aims to detect deviations from a prior knowledge base of SIP dialogs, with a high level of accuracy and computational efficiency. This technique differs from traditional methods by offering a simpler yet effective alternative for real-time detection, leveraging the statistical properties of SIP dialogs to identify potential security threats without needing complex modeling or intensive computation.

The main contributions of this paper are as follows:

- We introduce a new approach to classify SIP dialogs as abnormal, representing cases where the dialog may be indicative of a potential security threat.
- We conduct a performance evaluation of the proposed classification scheme using a realistic SIP dataset, demonstrating their effectiveness in practical scenarios.
- We benchmark the proposed methods against the optimal Bayesian solution in [9], the deep learning-based approach in [11] and the hybrid approach which combines deep learning with statistical methods in [12], identifying the advantages and limitations of the proposed methodology, particularly in terms of computational efficiency and detection accuracy.
- The proposed approach achieves a similar classification performance as state-of-the-art methods [9], [11] while introducing a major improvement: a 1000-fold reduction in computation time. This significant efficiency gain makes our method highly scalable and practical for real-time applications.

The remainder of this paper is organized as follows: Section II provides an overview of related work in the field of SIP security and anomaly detection. Section III details the proposed methodology for predicting and classifying SIP dialogs. Section IV presents the experimental dataset and the

performance achieved by the proposed methodology. Finally, Section V concludes the paper and outlines future work directions.

Notation: The notation adopted in this paper is as follows.

Vectors are represented in lowercase, upright boldface type. A vector of k elements is represented by $\boldsymbol{v} = \{v_1, v_2, \ldots, v_k\}$. An ordered sequence is represented by the vector of k elements represented by $\boldsymbol{v} = \langle v^1, v^2, \ldots, v^k \rangle$. Matrices are represented in upper case bold-face type $\boldsymbol{M}$, and the element in the i row and j column is represented by m_{ij}. Finally, sets are represented in calligraphic font, e.g., $\mathcal{S}$.

II. RELATED WORK

The security aspects of the SIP protocol [1] have been widely studied across various research works [13]–[15] to avoid misuse of SIP resources allocated for legitimate purposes, service disruption, and multimedia resource exhaustion. SIP attacks can be divided into different types, including but not limited to flooding, malformed SIP messages, authentication breaches, spoofing, and SIP signaling [5]. Authentication is being addressed in several works by proposing more robust authentication schemes, often multi-factor authentication methods aimed at improving security [16]. Another prevalent attack vector involves flooding attacks, which can lead to service interruptions due to denial of service by overwhelming the SIP infrastructure. Several mitigation strategies, such as threshold-based mechanisms [17], have been proposed to detect these attacks by analyzing deviations in SIP traffic patterns compared to normal operational baselines, which can be described as a single dimension [18], or in multiple dimensions [19]; or by stopping the transmission of SIP messages over the SIP path [20]. Malformed SIP messages can also interrupt SIP servers and originate DoS attacks, which can be avoided by a myriad of different methods, such as, firewalls enforcing specific rules [21]–[23], specific machine learning [24] and deep learning[25] methods, or statistical approaches that compare message patterns to detect malformation [26], [27]. Other solutions also adopt the classification of the incoming SIP messages before they are processed by the server to filter out potentially harmful content [28], [29]. The ability to spoof SIP's caller-ID identification can be easily performed due to the numerous SIP servers over the internet and the lack of regulation in specific regions [30]. The detection of this type of attack is a challenging task due to the complexity of telephony networks, and a possible solution relies on the verification of the caller-ID at different stages of the call using, for example, blockchain technology [31].

In addition to these attack types, vulnerabilities can also be exploited in the signaling processes, particularly through faulty protocol implementations. SIP signaling attacks manipulate the protocol's logic to bypass security measures, allowing improper authentication or call termination. Examples include BYE-attacks, CANCEL-attacks, or REFER-attacks, as outlined in [13]. The detection of these attacks often relies on identifying unique patterns within SIP traffic that correlate with specific attack behaviors. Techniques such as rule-based monitoring, as presented in [32], use event graphs constructed from protocol activities to detect known vulnerabilities. However, these signature-based solutions have limitations when encountering novel attack patterns, necessitating more adaptable methods. Research on SIP signaling vulnerabilities, such as in [33], has introduced debugging tools that analyze incoming SIP message flows to categorize them into compliant and non-compliant dialogs. In [32], a mitigation approach was proposed based on the contextual characteristics of SIP traffic, similar to the technique in [34], which compares peer interactions and timing to historical data to identify significant deviations.

Contrary to the approaches presented in [33], [34], the works in [9]–[12] develop an automated detection and prediction system capable of identifying both known and novel SIP dialogs. In these works, the detection of known dialogs benefits from prior labeling and assessment of its vulnerability. In [9], the authors propose a sequential analysis of the SIP dialog, and a solution for its detection and prediction is proposed through the adoption of an improved implementation of the optimal Viterbi algorithm over a Bayesian network. Although the proposed solution achieves the optimal detection and prediction rates, its computational complexity is a non-linear function of the number of different SIP dialogs, thus not scalable for a high number of dialogs. In an attempt to decrease the computational complexity of the methodology in [9], the deep learning-based solutions in [10]–[12] achieved improved computation performance although not achieving the optimal detection rates. In this paper, we propose a cross-correlation-based methodology for SIP dialogs' prediction and detection, which increases the detection rates when compared to the schemes in [10]–[12] and decreases the computation time in a magnitude of 3 times.

III. SIP DIALOGS' PREDICTION AND CLASSIFICATION

This section presents a model for SIP dialog inference based on a cross-correlation statistical analysis applied to a dataset of SIP dialogs. We then present an algorithm that performs SIP dialog classification in the presence of the full dialog and dialog prediction for partial observable dialogs.

A. SYSTEM MODEL

Table 1 provides a summary of the symbols adopted throughout this paper, along with their corresponding descriptions, and serves as a reference for the used notation.

The SIP protocol signals multimedia sessions between multiple participants through the exchange of SIP messages.

Definition 1. *A SIP message, denoted by m_t, with $t \in \mathcal{M}$, refers to a particular type of SIP request or response. $\mathcal{M}$ is the set of all possible types of SIP requests and responses. The cardinality $|\mathcal{M}|$ of the set $\mathcal{M}$ is the number of different methods and responses that exist.*

An interaction begins with a peer sending a SIP request message that indicates its type via the SIP method field in the message header. The receiving peer then responds with a SIP response message, which includes a reply code. A SIP request and its associated response form a SIP transaction. A series of such transactions constitutes a SIP dialog that is uniquely identified by the Call-ID field in the SIP message header.

Definition 2. *A sequence of consecutive SIP messages forms a SIP dialog denoted by* $\boldsymbol{d}_k = \langle m_{t_1}^1, \ldots, m_{t_h}^h, \ldots, m_{t_k}^{L_d^k} \rangle$, *where* m_t^h *represents the* h*-th message of the SIP dialog with a given type* $t_h \in \mathcal{M}$. L_d^k *is used to represent the length of SIP dialog* k. *All messages belonging to the same dialog share the same SIP Call ID denoted by* $\gamma_{\boldsymbol{d}_k}$, *as well as identical sender and receiver addresses*

In this work, we assume that the SIP peers and intermediary servers along the communication path can access the exchanged SIP messages and read their headers to identify the Call-ID and associate the requests and responses that belong to a given dialog. This can be used to obtain an observation of SIP traffic representing a specific sequence from the beginning of an SIP dialog.

TABLE 1. Table of Symbols

Symbols	Definitions
$\mathcal{D}$	Dataset (collection of padded SIP dialogs).
$\lvert\mathcal{D}\rvert$	Dataset size.
$\mathcal{D}_A$	$\mathcal{D}$ subset containing anomalous dialogs.
$\mathcal{D}_N$	$\mathcal{D}$ subset containing normal dialogs.
$\boldsymbol{d}_k$	SIP dialog k.
$\gamma_{\boldsymbol{d}_k}$	SIP Call ID of a SIP dialog $\boldsymbol{d}_k$.
Γ	Decision threshold for classification.
$\mathcal{H}_x$	Hypothesis x.
L_d^k	Length of a SIP dialog $\boldsymbol{d}_k$.
L_o^k	Length of an observation $\boldsymbol{o}_k$.
L_S	Length of padded sequences $\boldsymbol{s}_k$ and padded dialogs $\boldsymbol{p}_k$.
l	Cross-correlation lag.
m_t	SIP message with type t.
$m_{t_h}^h$	h-th message of a SIP dialog $\boldsymbol{d}_k$ with type t_h.
$\mathcal{M}$	Set with all the possible SIP messages.
$\lvert\mathcal{M}\rvert$	Number of possible SIP methods and responses.
μ_n	Central statistical moment of order n.
$\boldsymbol{C}$	Matrix of cross-correlation values.
n^k	Number of zeros added for padding purposes to the dialog $\boldsymbol{d}_k$ or observation $\boldsymbol{o}_k$.
$\boldsymbol{o}_k$	Observation k (sequence of SIP messages observed in a dialog).
$\boldsymbol{p}_k$	Padded SIP dialog $\boldsymbol{d}_k$.
$\boldsymbol{r}$	Vector used to store cross-correlation values.
$R_{\boldsymbol{s}_k\boldsymbol{p}_k}$	Cross-correlation between a padded sequence $\boldsymbol{s}_k$ and a padded SIP dialog $\boldsymbol{p}_k$.
$\boldsymbol{s}_k$	Padded sequence obtained from adding zeroes to an observation $\boldsymbol{o}_k$.
$y_{\boldsymbol{p}_k}$	Label of an observation $\boldsymbol{o}_k$, or padded SIP dialog $\boldsymbol{p}_k$.

Definition 3. *An observation* k *can be captured by a SIP user agent or server and is composed of a series of SIP messages represented as* $\boldsymbol{o}_k = \langle m_{t_1}^1, \ldots, m_{t_h}^h, \ldots, m_{t_k}^{L_o^k} \rangle$, *with* $t \in \mathcal{M}$ *and* $h \in \{1, 2, \ldots, L_o^k\}$. *The symbol* L_o^k *denotes the length of the observation. Within the same observation, all the SIP messages have the same SIP Call ID denoted by* $\gamma_{\boldsymbol{o}_k}$.

The reason we refer to it as an "observation" is that it allows us to gather critical data from the SIP communication in a particular server context and over time, as the dialog evolves. We use these observations to analyze and predict the type of SIP dialogs that are taking place. To allow the comparison of observations with variable lengths, we introduce a padding method to convert the observations into sequences of equal length, which we call padded sequences.

Definition 4. *An observation* $\boldsymbol{o}_k$ *is transformed into a padded sequence* $\boldsymbol{s}_k$ *by adding zeros at the end of the observation. The total length of the padded sequence is represented by a constant* L_S, *with* $L_S = L_o^k + n^k$, *being* n^k *the number of zeros added, resulting in the padded sequence* $\boldsymbol{s}_k = \langle \boldsymbol{o}_k, \underbrace{0, 0, \ldots, 0}_{(n^k)} \rangle$.

The proposed algorithm compares the padded observation of a SIP dialog $\boldsymbol{s}_k$ with a labeled knowledge base of complete dialogs that forms a dataset that contains known examples of SIP dialog patterns along with their corresponding classifications (i.e., normal or malicious).

Definition 5. *A dataset* $\mathcal{D}$ *is a collection of labeled SIP dialogs* $\boldsymbol{d}_k$ *that were padded with zeroes, where the total length of every padded dialog is* $L_S = L_d^k + n^k$. *Therefore, a dataset* $\mathcal{D}$ *with* $\lvert\mathcal{D}\rvert$ *elements, contains the padded SIP dialogs represented by* $\boldsymbol{p}_k = \langle \boldsymbol{d}_k, \underbrace{0, 0, \ldots, 0}_{(n^k)} \rangle$ *and their respective class labels* $y_{\boldsymbol{p}_k}$ *that mark it as normal* $y_{\boldsymbol{p}_k} = N$ *or anomalous* $y_{\boldsymbol{p}_k} = A$.

The length of L_o^k in Definition 4 and L_d^k in Definition 5 depend on the SIP dialog k, while the length of the padded sequences in the dataset $\mathcal{D}$ is always L_S for all padded dialogs.

The knowledge dataset $\mathcal{D}$ can be partitioned into two subsets. One that contains all normal padded SIP dialogs.

Definition 6. $\mathcal{D}_N$ *is the subset of* $\mathcal{D}$ *that contains the padded dialogs with the normal label* $\mathcal{D}_N = \boldsymbol{p}_k \in \mathcal{D} \mid y_{\boldsymbol{p}_k} = N$.

Another is the set of anomalous padded dialogs.

Definition 7. $\mathcal{D}_A$ *is the subset of* $\mathcal{D}$ *that contains the padded dialogs with the anomalous label* $\mathcal{D}_A = \boldsymbol{p}_k \in \mathcal{D} \mid y_{\boldsymbol{p}_k} = A$.

The knowledge datasets allow us to analyze an observation $\boldsymbol{o}_k$ in order to classify it as anomalous or normal. The classification is based on the cross-correlation similarity metric.

Classification of anomalous dialogs occurs when the observation obtained by a server corresponds to the full SIP dialog (i.e., o_k contains all the SIP messages of a Call ID $\gamma_{\boldsymbol{o}_k}$). Prediction, on the other hand, tests the similarity of the current observation $\boldsymbol{o}_k$ with the knowledge base datasets every time a new SIP message $m_{t_h}^h$ arrives until the observation corresponds to the full dialog $\boldsymbol{d}_k$ for a given call ID $\gamma_{\boldsymbol{o}_k}$.

B. CLASSIFICATION OF SIP DIALOGS

The first goal is to distinguish anomalous dialogs from normal dialogs.

To detect anomalous dialogs, we need to understand the similarities (or dissimilarities) between the padded sequence $\boldsymbol{s}_k$ of an observation $\boldsymbol{o}_k$ and the padded dialogs $\boldsymbol{p}_k$ in the database. Cross-correlation, denoted by $R_{\boldsymbol{s}_k\boldsymbol{p}_k}$, is calculated as

$$R_{\boldsymbol{s}_k\boldsymbol{p}_k}[l] = \sum_h \boldsymbol{s}_k[m_{t_h}^h]\boldsymbol{p}_k[m_{t_{h-l}}^{h-l}], \tag{1}$$

where l represents the lag, $\boldsymbol{s}_k$ represents a padded sequence k, $\boldsymbol{p}_k$ represents a padded SIP dialog k from a dataset $\mathcal{D}$ and m_t^h is the h-th message. Eq. (1) computes the correlation between $\boldsymbol{s}_k$ and $\boldsymbol{p}_k$ by summing the products of all their h messages. The correlation values utilized in this work are calculated with zero time lag ($l = 0$) unless otherwise stated since we are focused on the relationship between $\boldsymbol{s}_k$ and $\boldsymbol{p}_k$ at the same message h.

Two different approaches were used to analyze the correlation values. In the first approach, the correlation values were used without any modifications, i.e., $R_{\boldsymbol{s}_k\boldsymbol{p}_k}[0]$ values are adopted. In the second approach, the correlation values were normalized using L2-normalization. Full sequence L2-normalization is computed as follows

$$||R_{\boldsymbol{s}_k\boldsymbol{p}_k}[0]||_2 = \frac{R_{\boldsymbol{s}_k\boldsymbol{p}_k}[0]|}{\sqrt{\sum_l |R_{\boldsymbol{s}_k\boldsymbol{p}_k}[l]|^2}}, \tag{2}$$

where $R_{\boldsymbol{s}_k\boldsymbol{p}_k}[0]$ is the value to be normalized (i.e., the cross-correlation value for null lag) and l denotes all lags between the SIP messages of $\boldsymbol{s}_k$ and $\boldsymbol{p}_k$.

Considering the scenario where one SIP dialog is fully received by a server observing the dialog, the observation $\boldsymbol{o}_k$ that contains that dialog is padded with zeroes, following Definition 4, leading to the padded sequence $\boldsymbol{s_k}$. The cross-correlation between $\boldsymbol{s_k}$ and each padded dialog $\boldsymbol{p}_k$ of the knowledge base $\mathcal{D}$ is then computed according to (1) for all $\boldsymbol{p}_k$ padded dialogs in $\mathcal{D}$. The values are stored in a vector $\boldsymbol{r}$ with a total of $|\mathcal{D}|$ entries, represented by

$$\boldsymbol{r} = \{R_{\boldsymbol{s}_k\boldsymbol{p}_1}[0], R_{\boldsymbol{s}_k\boldsymbol{p}_2}[0], \ldots, R_{\boldsymbol{s}_k\boldsymbol{p}_{|\mathcal{D}|}}[0]\},$$

or

$$\boldsymbol{r} = \{||R_{\boldsymbol{s}_k\boldsymbol{p}_1}[0]||_2, ||R_{\boldsymbol{s}_k\boldsymbol{p}_2}[0]||_2, \ldots, ||R_{\boldsymbol{s}_k\boldsymbol{p}_{|\mathcal{D}|}}[0]||_2\},$$

depending on whether non-normalized or normalized correlation is assumed, respectively. For the sake of clarity, we name the classification as non-normalized or normalized if the non-normalized or normalized correlation is assumed, respectively.

After computing the vector $\boldsymbol{r}$, the central moments, μ_n, are calculated as follows

$$\mu_n = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(r_i - \mu(\boldsymbol{r}))^n, \tag{3}$$

where r_i is i-th cross-correlation value $R_{\boldsymbol{s}_k\boldsymbol{p}_i}[l]$, stored in vector $\boldsymbol{r}$, n is the order of the moment, and $\mu(\boldsymbol{r})$ is the mean of the vector $\boldsymbol{r}$ (i.e., the mean of all cross-correlation values).

We evaluated various statistical moments, specifically, mean, variance, skewness, kurtosis, and higher-order moments. Among these statistical moments, the 4$^{\text{th}}$-order central moment, kurtosis, provided the best results in distinguishing between normal and anomalous dialogs due to its properties. Kurtosis allows the study of the "tailedness" of a distribution [35], which makes it particularly effective in identifying outliers in data. Kurtosis has been proven to exhibit a high sensitivity to outliers. It has been used in several other domains, including forensic analysis [36], geophysics [37], and meteorology [38]. These multiple applications further support the relevance of kurtosis as a discriminative measure, supporting its adoption in the context of dialogue classification.

The decision for classifying an observation $\boldsymbol{o}_k$, padded in the sequence $\boldsymbol{s}_k$, based on the knowledge base $\mathcal{D}$, involves testing the following hypotheses

$$\mathcal{H}_0 : M^4 \leq \Gamma, \tag{4}$$
$$\mathcal{H}_1 : M^4 > \Gamma, \tag{5}$$

where Γ represents the decision threshold used to determine the similarity or fit between the observation $\boldsymbol{o}_k$ and the labels in the knowledge base $\mathcal{D}$. The classifier assigns to the observation $\boldsymbol{o}_k$ the label $y_{\boldsymbol{p_k}} = N$ or $y_{\boldsymbol{p_k}} = A$. Considering $\mathcal{D} = \mathcal{D}_N$, the label's assignment is based on the following decision rule:

1) If $\mathcal{H}_0$ is true, $\boldsymbol{o}_k$ is labeled as $y_{\boldsymbol{p_k}} = N$.
2) In contrast, if $\mathcal{H}_1$ holds, $\boldsymbol{o}_k$ is labeled as $y_{\boldsymbol{p_k}} = A$.

We highlight that the threshold Γ is defined by characterizing the kurtosis of the knowledge base under consideration, i.e., $\mathcal{D} = \mathcal{D}_N$, and being greater than the highest kurtosis value achieved from all dialogs in $\mathcal{D}_N$.

C. PREDICTION OF SIP DIALOGS

For predicting SIP dialogs the goal is to use the observation $\boldsymbol{o}_k$ that is collected over time in a SIP server, in order to predict the complete SIP dialog $\boldsymbol{d}_k$ corresponding to the partial dialog of SIP messages in $\boldsymbol{o}_k$ already received by the server.

Every message m_t^h that arrives is added to the observation $\boldsymbol{o}_k$ that is then padded to the size L_S of the dialogs in the database $\mathcal{D}$. This produces a new padded sequence $\boldsymbol{s_k}^h$, for each message h received. The result is a set of padded sequences with the size of the number of messages in the

Algorithm 1 Classification Algorithm

Input: s_k, p_k
Output: y_{p_k}

▷ Calculate cross-correlation:
for each padded dialog p_k, $k = 1, 2, \ldots, |\mathcal{D}|$ **do**
  $R_{s_k p_k}[l]$=calc_cross_correlation(s_k, p_k)
  if normalization==True **then**
    $R_{s_k p_k}[0]$=normalize_cross_corr($R_{s_k p_k}[l]$)
  end if
  r=r.append($R_{s_k p_k}[0]$)
end for
▷ Calculate Moment:
μ_n=calc_moment(r,n)
▷ Decision Making:
if $M^4 \leq \Gamma$ **then**
  $y_{p_k} = N$ ▷ Normal dialog
else if $M^4 > \Gamma$ **then**
  $y_{p_k} = A$ ▷ Anomalous dialog
end if

dialog, L_d^k, one for each new message. Each padded sequence is written as

$$s_k{}^h = \langle o_k, m_t^h, \ldots, 0 \rangle, h \in \{1, 2, \ldots, L_d^k\},$$

where L_d^k is the length of the dialog k being received, and the number of zeroes is in the interval $[L_S - 1, L_S - L_d^k]$.

Each of the $s_k{}^h$ sequences is correlated using (1) with all padded SIP Dialog p_k in $\mathcal{D}$, to achieve $R_{s_k p_k}[0]$, resulting in a matrix C of $R_{s_k p_k}[0]$ values with dimension $|\mathcal{D}| \times L_S$. Each element $c_{k,h}$ of C contains the cross-correlation value between the $k - th$ p_k padded dialog in $\mathcal{D}$ and the partial padded sequence $s_k{}^h$ that corresponds to an observation k up to message h.

A column h of the matrix C corresponds to all cross-correlation values between the partial sequence $s_k{}^h$ and every other padded SIP Dialog p_k in $\mathcal{D}$. The prediction procedure is computed when a new message h is received at the SIP server, and the outcome of the prediction is the SIP dialog $d_k, k \in \{1, \ldots, |\mathcal{D}|\}$, with k given by

$$\arg\max_k c_{k,h}. \tag{6}$$

Regarding the validation of the prediction, whenever the predicted dialog for $s_k{}^h$ matches the correct one it is counted as a success, otherwise, it is deemed a prediction failure.

IV. PERFORMANCE EVALUATION

In this section, we present the results obtained from both classification and prediction of SIP dialogs. These results are analyzed and discussed to provide insights into the performance of the developed classifiers in comparison to other methods. Performance evaluation is conducted using a variety of metrics, complemented by graphical representations to facilitate visual interpretation.

Table 2 summarizes the symbols employed in this section, along with their respective definitions.

Algorithm 2 Prediction Algorithm

Input: o_k, p_k
Output: $pred_h$

▷ Calculate cross-correlation:
for each message m_l^h in o_k, $h = 1, 2, \ldots, L_d^k$ **do**
  $s_k{}^h$=add_message(m_l^h)
  for each padded dialog p_k, $k = 1, 2, \ldots, |\mathcal{D}|$ **do**
    $R_{s_k{}^h p_k}[l]$=calc_cross_correlation($s_k{}^h$, p_k)
    if normalization==True **then**
      $R_{s_k{}^h p_k}[0]$=normalize_cross_corr($R_{s_k{}^h p_k}[l]$)
    end if
    $\mathcal{C}[h, k]$=($R_{s_k{}^h p_k}[0]$)
  end for
  ▷ Calculate column h argmax
  max_id=get_id($\arg\max_k c_{k,h}$)
  if max_id == get_id($s_k{}^h$) **then** ▷ Check IDs
    $pred_h$ = Success
  else
    $pred_h$ = Failure
  end if
end for

TABLE 2. Table of Symbols

Symbols	Definitions
C_{NA}	Matrix of correlation values between normal and anomalous dialogs.
C_{NN}	Matrix of correlation values between normal and normal dialogs.
$\mathcal{D}_A^u$	$\mathcal{D}$ subset containing unique anomalous dialogs.
$\mathcal{D}_N^u$	$\mathcal{D}$ subset containing unique normal dialogs.
tn	True negative - An anomalous dialog that was correctly classified as anomalous.
tp	True positive - A normal dialog that was correctly classified as normal.
fn	False negative - A normal dialog that was incorrectly classified as anomalous.
fp	False positive - An anomalous dialog that was incorrectly classified as normal.

A. METRICS

To evaluate our model's performance and compare it with other classifiers, we selected a total of 10 evaluation metrics. In this paper, we define normal dialogs as the "positive" class and anomalous dialogs as the "negative" class.

True Negative (tn) refers to the rate of anomalous dialogs that the classifier correctly identified as anomalous. This metric is also referred to as the correct rejection rate.

True Positive (tp) is the rate of normal dialogs that the classifier was able to correctly identify as normal. Other names include hit, and correct detection rate.

False Positive (fp) also referred to as incorrect detection, is the rate of anomalous dialogs that were incorrectly classified as normal dialogs.

False Negative (fn) represents the rate of normal dialogs

that were incorrectly labeled as anomalous dialogs by the classifier. Other names include miss or underestimation.

Specificity is also called true negative rate, and selectivity. This metric provides insight about how well the classifier can identify true negatives (anomalous dialogs correctly identified). In other words, it reflects the model's capacity to avoid false positives (anomalous dialogs incorrectly classified as normal). Specificity is mathematically represented by

$$\text{Specificity} = \frac{tn}{tn + fp}. \tag{7}$$

Sensitivity is similar to Specificity but focuses on the classifier's ability to correctly identify normal dialogs (true positives), i.e., how well it avoids false negatives (normal dialogs incorrectly classified as anomalous). It is also referenced by true positive rate and recall, and is calculated as follows

$$\text{Sensitivity} = \frac{tp}{tp + fn}. \tag{8}$$

Precision measures the accuracy of positive predictions (normal dialogs correctly identified), i.e., how much we can trust the classifier when a dialog is labeled as normal. It is also known as positive predictive value, and is calculated as

$$\text{Precision} = \frac{tp}{tp + fp}. \tag{9}$$

Accuracy provides an overall assessment of the classifier's performance, as it incorporates all four elements of the confusion matrix. Accuracy measures the proportion of correct predictions (both true positives and true negatives). It is calculated as follows

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}. \tag{10}$$

F1-Score is a commonly used metric that combines precision and recall into a single value. F1-Score is the harmonic mean between these two metrics, representing them symmetrically. The F1-Score is particularly useful in situations where the dataset is imbalanced, as accuracy alone may provide skewed results. It is mathematically represented by

$$\text{F1-Score} = \frac{2tp}{2tp + fp + fn} \tag{11}$$

Mathews Correlation coefficient (MCC), also known as phi coefficient, is another important metric, especially when dealing with imbalanced datasets. One common critique of the F1-Score is that it does not account for true negatives (anomalous dialogs correctly identified), which may lead to biased results in certain contexts. In contrast, MCC considers all four components of the confusion matrix and is generally regarded as a more reliable measure than F1-Score. MCC is calculated as

$$\text{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp) \times (tp + fn) \times (tn + fp) \times (tn + fn)}}. \tag{12}$$

B. DATA HANDLING

The performance evaluation is based on the SIP dataset denoted by $\mathcal{D}$ made available by Nassar et al. [39], which employs two open-source SIP servers: Asterisk [40] and Opensips [41]. The dataset includes details such as the URI of sender and receiver user agents, SIP messages exchanged, packet counts, timestamps for each SIP dialog, and the final session state (successful, rejected, or canceled). Each SIP dialog is identified using the URI user agent, SIP messages, and timestamps, with SIP message types encoded as integer values.

Two different datasets are taken from $\mathcal{D}$, $\mathcal{D}_N$ containing 14801 normal SIP dialogs of various lengths, with a total of 1043 unique dialogs; and $\mathcal{D}_A$ containing 8141 anomalous dialogs, from which 152 are unique.

Each dialog can have a maximum of 56 SIP messages, although not all dialogs contain as many messages. Dialogs whose length is shorter than 56 messages were padded with zeros as described in Definition 5 to ease data manipulation and analysis processes.

To implement the clarification methodology, each dataset was first striped of any repeated dialogs. This is done by eliminating dialogs with equal sequences of SIP messages. The datasets of unique dialogs are denoted by $\mathcal{D}_N^u$ and $\mathcal{D}_A^u$.

Normalized and non-normalized values of the cross-correlation are then calculated with these new datasets $\mathcal{D}_N^u$ and $\mathcal{D}_A^u$. Each entry of $\mathcal{D}_N^u$ is considered as a padded sequence ($\boldsymbol{s}_k$) and is correlated with each entry of $\mathcal{D}_A^u$, which is considered as a padded SIP dialog $\boldsymbol{p}_k$ in (1). The correlation values computed with (1) are stored in a matrix $\boldsymbol{C}_{NA}$. Since there is a total of 1043 unique normal dialogs and 152 unique anomalous dialogs, $\boldsymbol{C}_{NA}$ has the dimension 1043×152. This process is then repeated but using $\mathcal{D}_N^u$ only, correlating each of its entries with the remaining ones, i.e., each SIP dialog is considered as a padded sequence ($\boldsymbol{s}_k$), and is correlated with the other SIP dialogs of $\mathcal{D}_N^u$ that are considered as padded SIP dialogs $\boldsymbol{p}_k$ in (1). After computing (1) for all SIP dialogs of $\mathcal{D}_N^u$, we obtain the correlation matrix $\boldsymbol{C}_{NN}$ of size 1043×1043. In what follows, we refer to the non-normalized classifier or normalized classifier, when the correlation values in $\boldsymbol{C}_{NN}$ and $\boldsymbol{C}_{NA}$ are not normalized, or normalized using (2), respectively.

C. ABNORMAL SIP DIALOGS DETECTION

Using the non-normalized version of the matrices $\boldsymbol{C}_{NA}$ and $\boldsymbol{C}_{NN}$, the moments of order 2 (M^2, variance) and order 4 (M^4, kurtosis) are calculated and plotted in Figure 1.

The classifier, implemented by the conditions in (4) and (5), was parametrized with Γ slightly higher than the maximum kurtosis value obtaining for the normal SIP dialogs, i.e., the maximum kurtosis value computed for $\boldsymbol{C}_{NN}$, which in this case was parameterized to $\Gamma = 1 \times 10^{12}$.

As observed in the results in Figure 1, the condition in (4) allows the detection of all normal SIP dialogs, which we denote as a true positive (tp) rate of 100% and false negative (fn) rate of 0%. Regarding the anomalous SIP dialogs clas-

sified with the condition (5), the proposed methodology is capable of detecting 139 of the 152 abnormal SIP dialogs, missing the detection of 13, which we denote as a true negative (tn) rate of 91.45% and a false positive (fp) rate of 8.55%. These results are summarized in the form of a confusion matrix in Figure 2.

Although the number of false positives is relatively low, in a real-world application, even a small number of misclassifications could lead to significant issues. Some service degradation is to be expected as the false positive rate increases, since anomalous dialogs will erroneously be classified as normal, and therefore, avoid immediate detection. This delay in detection means that appropriate measures will not be taken immediately, increasing the likelihood of successful attacks. Additionally, since automatic measures will not be triggered, there will also be an increase in administrative overhead since it will require manual intervention and potentially additional audits and security scans to assess the impact of these undetected anomalies. Finally, as false positives accumulate, user trust may be eroded, especially in sensitive sectors such as healthcare and finance, where users will begin to notice service degradation and recognize that attacks are being overlooked.

Figure 3 presents the results obtained with the normalized classifier. The normalization of correlation values at null lag ($l = 0$) does not significantly improve classification performance, as the kurtosis and variance values computed for the normalized correlation in $\boldsymbol{C}_{NA}$ overlap with those in $\boldsymbol{C}_{NN}$.

To enhance performance, instead of setting the threshold Γ based on the maximum kurtosis value from $\mathcal{D}_N$ (as described in Subsection IV-C), we parameterized Γ using the minimum kurtosis value from $\boldsymbol{C}_{NA}$. This approach successfully detects all anomalous SIP dialogs. However, it also misclassifies approximately 46.4% of normal SIP dialogs as abnormal dialogs, highlighting the limited effectiveness of the normalized classifier. The confusion matrix of the normalized classifier can be visualized in Figure 4 and compared to the classifier without normalization. As described, and evident by the confusion matrices, L2 normalization presents a drawback in dialog classification. While the first classifier exhibits a false negative rate of 0%, the L2 normalization classifier, though presenting a false positive rate of 0% shows a significantly higher false negative rate of 53.60%.

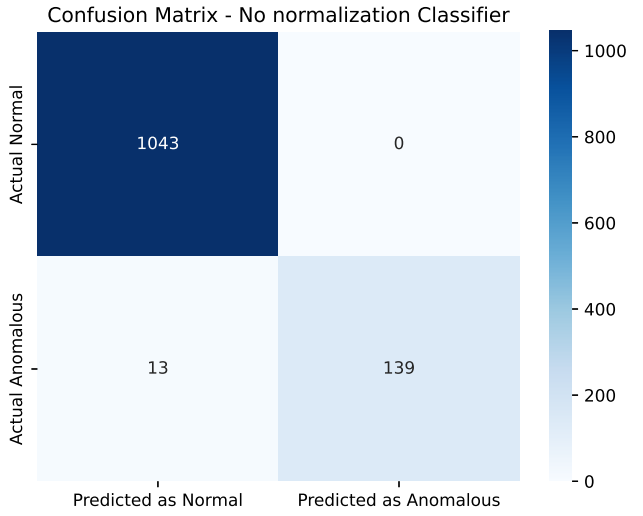


FIGURE 2. No normalization Classifier Confusion Matrix.

A high false negative rate can also significantly degrade system performance, given that normal dialogs will be incorrectly classified as anomalous, triggering unnecessary security measures. This will create major system delays, causing user frustration and diminishing trust in the service, as users perceive it to be slow and unreliable. Additionally, the high rate of false alarms also increases administrative overhead, since additional manual interventions will be needed to truly assess whether a flagged dialog is truly anomalous or simply a false positive.

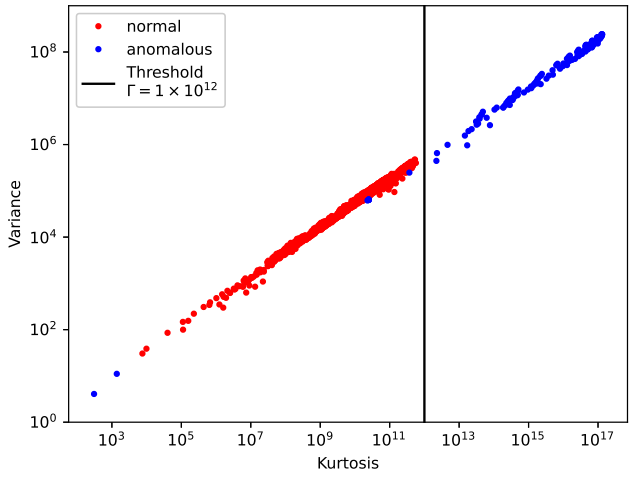


FIGURE 1. Non-normalized classification adopting the threshold $\Gamma = 1 \times 10^{12}$.

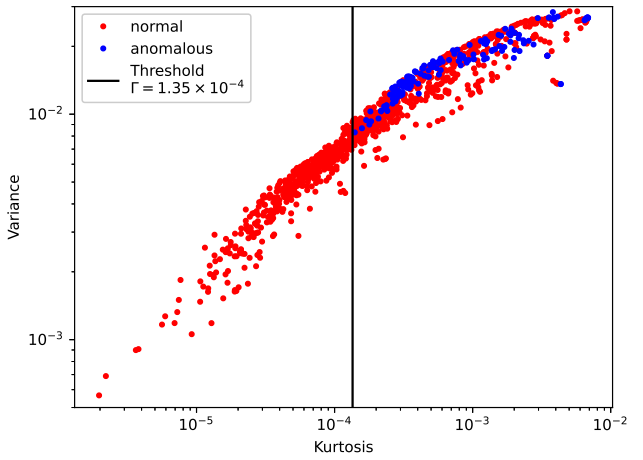


FIGURE 3. Normalized classification adopting the threshold $\Gamma = 1.35 \times 10^{-4}$.

TABLE 3. Performance metrics comparison of multiple SIP dialog classifiers.

Model	HMM [9]	LSTM RNN model 1 [11]	LSTM RNN model 2 [11]	CNN model [12]	Cross-correlation (no normalization)	Cross-correlation (L2 normalization)
True negative (tn)	1.000	0.9430	0.9451	0.9583	0.9145	1.000
True positive (tp)	1.000	0.9365	0.9189	0.7912	1.000	0.4640
False positive (fp)	0.000	0.0635	0.0811	0.0417	0.0855	0.000
False negative (fn)	0.000	0.0570	0.0549	0.2088	0.000	0.5360
Specificity	1.000	0.9430	0.9451	-	0.9145	1.000
Sensitivity	1.000	0.9365	0.9189	-	1.000	0.4640
Precision	1.000	0.9718	0.9723	0.9981	0.9877	1.000
Accuracy	1.000	0.9386	0.9274	0.7969	0.9891	0.5322
F1-Score	1.000	0.9538	0.9449	0.8827	0.9938	0.6339
MCC	-	-	-	-	0.9503	0.3149

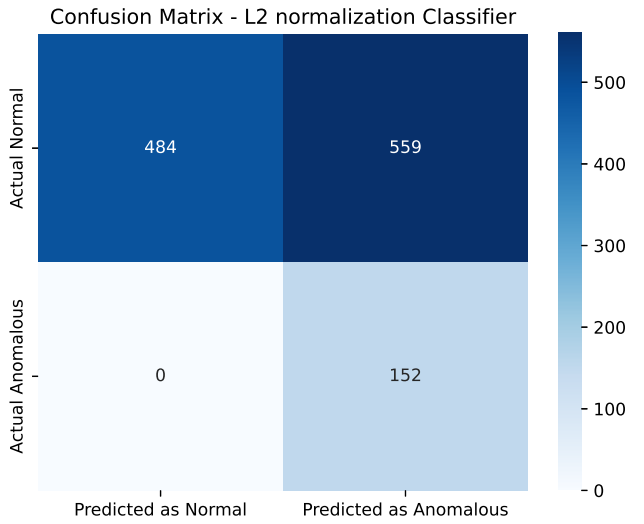


FIGURE 4. L2 normalization Classifier Confusion Matrix.

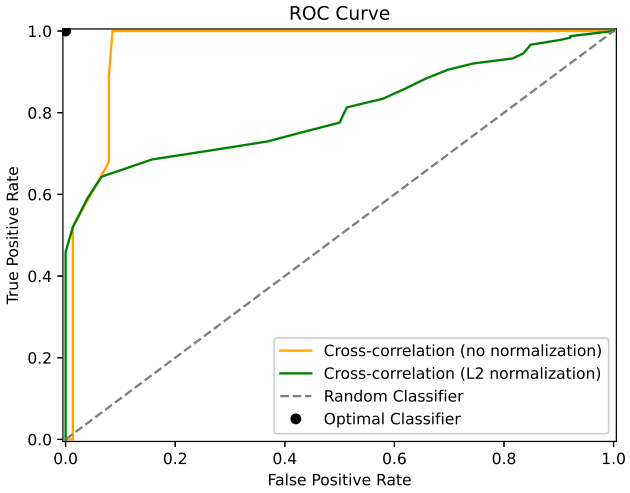


FIGURE 5. ROC Curve of both classifiers

We can further compare the performance of these two classifiers through the analysis of their Receiver operating characteristic (ROC) curves, as presented in Figure 5. In this graph, it once again becomes evident that the classifier with no normalization presents better results, since it is closer to the perfect classifier. In contrast, the ROC curve for the L2 normalized classifier is closer to that of a random classifier, indicating a less favorable performance.

Table 3 presents a comparison of various performance metrics for the normalized and non-normalized classifiers, alongside the classifiers proposed in [9], [11], and [12].

The HMM model from [9] achieves optimal classification performance but is impractical due to its high computational cost and lack of scalability as the number of dialogs in the knowledge base increases. Models 1 and 2 from [11], while not reaching optimal classification (F1-scores of 0.9538 and 0.9449, respectively), offer a significant computational advantage over the optimal model in [9]. The CNN model [12] combines deep learning with statistical methods, and although its performance is inferior to the aforementioned methods, it offers a significant reduction in computational time, even lower than LSTM models presented in [11].

Comparing the F1-score of the proposed non-normalized cross-correlation model, we find that it closely approaches the performance of the HMM model, slightly surpasses Models 1 and 2 presented in [9], and offers a clear advantage over the CNN model. These results show that the non-normalized classifier proposed in this paper achieves the closest performance to the optimal solution among the evaluated approaches.

Table 3 includes the Matthews Correlation Coefficient (MCC) for both classifiers to account for dataset imbalance. Specifically, there are 1,043 unique normal dialogs but only 152 unique anomalous dialogs. This imbalance can distort certain metrics, such as the F1-score, leading to overly optimistic or pessimistic results. Since MCC provides a more balanced assessment of performance in imbalanced datasets, it is a more reliable metric in this scenario. MCC shows the advantage of adopting the non-normalized version of the proposed classifier when compared to the normalized one.

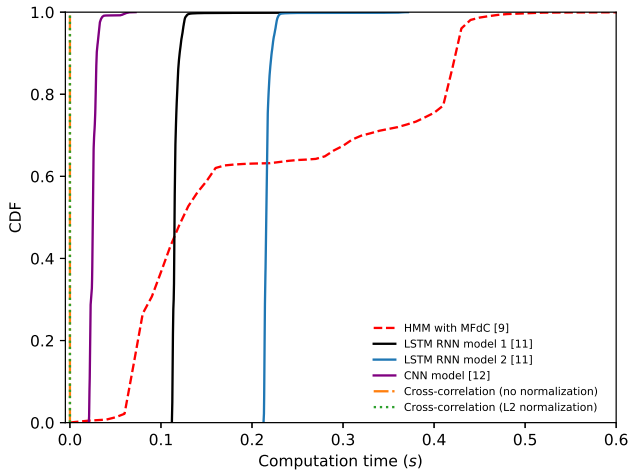
Figure 6 shows the cumulative distribution function

(CDF) of the computation times for the detection classifiers compared in Table 3. The results show that the cross-correlation classifiers achieve a significantly lower computation time. To better visualize, the graph adopting a logarithmic scale Figure 6 (b) reveals that the non-normalized cross-correlation classifier classifies an SIP dialog in approximately 1.87×10^{-5} s, and the normalized one in 2.21×10^{-5} s. In contrast, the HMM-based optimal solution presented in [9] exhibits a prohibitively high computation time, making it unsuitable for real-time applications and non-scalable as the number of dialogs increases. The LSTM-based models significantly reduce the computation time compared to the HMM model (by at least a factor of 3), with LSTM Model 1 performing slightly better than LSTM Model 2 due to its lower complexity. Additionally, the CNN model further reduces computational time, although it remains significantly higher than both the non-normalized and normalized cross-correlation classifiers.

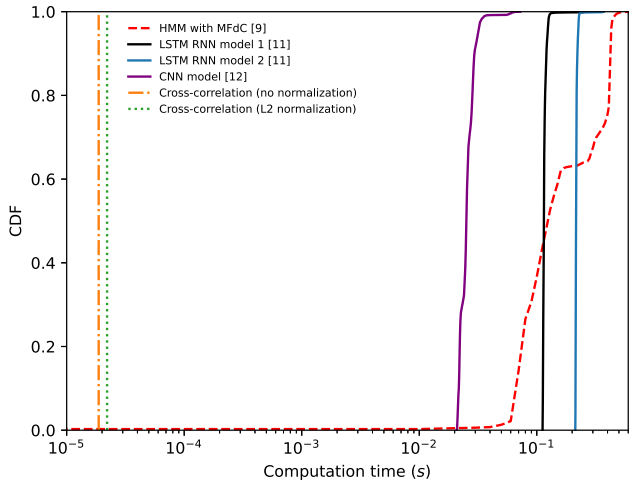
In Figure 6, the results plotted in logarithmic scale highlight the efficiency of the proposed approach: the non-normalized classifier achieves the lowest computation time, while the normalized version incurs a slight overhead due to the normalization step. Nonetheless, both versions of the proposed correlation-based classifier exhibit a significant improvement compared to the best-performing models in the literature (CNN model [12]), which takes approximately 0.02 s to classify each SIP dialog.

To facilitate the visualization of trade-offs between the performance of each classifier and their computation times, Figure 7 is presented next. The x-axis illustrates the average computation time of each model, measured in seconds. For the y-axis Accuracy was the selected metric, as it is available across all the models examined. Additionally, unlike the F1-score, accuracy incorporates all elements of the confusion matrix, thereby reducing susceptibility to biases that may arise from imbalanced datasets, as encountered in this study. As previously mentioned, the HMM model exhibits perfect accuracy (1.0), at the expense of one of the longest computation times (around 0.2 s). In contrast, the cross-correlation classifier (without normalization) achieves the second highest accuracy (98.91%), closely following the HMM model, while also demonstrating the fastest computation of approximately 1.87×10^{-5} s. Conversely, the cross-correlation model with L2-normalization, while highly efficient in terms of computation time, yields the lowest accuracy among the models explored, rendering it unsuitable for accurate classification of normal and anomalous SIP dialogs.

Besides computation time, algorithmic complexity is another factor that can be used to compare these multiple methods. However, this comparison is challenging for the models under consideration. The HMM model, for example, exhibits non-linear complexity that depends on the number of different dialogs [9], which makes it difficult to scale as the number of dialogs increases. Both the LSTM RNN models and the CNN model rely on machine learning techniques, where a significant overhead is needed to train the model,



(a)



(b)

FIGURE 6. CDFs of the classification computation times: (a) presents the results in a linear scale, and (b) illustrates the same results on a logarithmic scale.

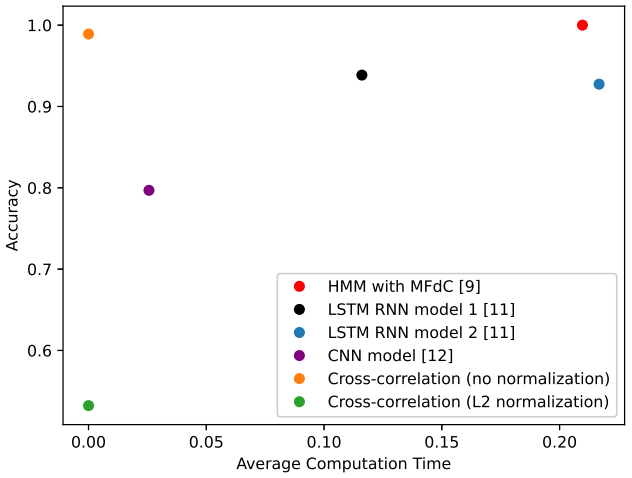


FIGURE 7. Trade-off between Average Computation Time and classification Accuracy.

and their complexities are also non-linear and troublesome to accurately calculate. In contrast, our model features a lower complexity of $\mathcal{O}(n)$ for SIP dialog classification, and $\mathcal{O}(n^2)$ for SIP dialog prediction, making it fast and easily scalable.

D. SIP DIALOGS PREDICTION

In this subsection, we evaluate the SIP dialog prediction rate based on the amount of available information for predicting each dialog. Specifically, we use the ratio between the observed input sequence length and the total SIP dialog length, denoted as the ratio L_o^k/L_d^k. The prediction of the SIP dialogs is performed for the normal dialogs in the dataset $\mathcal{D}_N^u$. The prediction algorithm is applied progressively: starting with a single observation (the first SIP message in the dialog), then two observations (the first two messages), and continuing this process until the full SIP dialog (L_d^k messages) is considered.

Figure 8 presents the prediction results of the proposed algorithm, obtained through the computation of (6) considering that the matrix $\boldsymbol{C}$ is filled with non-normalized correlation values and normalized ones. The results also include the performance of the SIP prediction dialogs proposed in [9], [11], and [12]. The x-axis represents the ratio of observed input sequence length to total dialog length, expressed as a percentage.

The results in Figure 8 show that the proposed prediction method achieves a lower performance when compared to the ones proposed in [9], [11], and [12]. The LSTM-based predictor in [11] achieved a high prediction probability even with minimal available information, gradually improving as more data became accessible. The HMM-based prediction in [9] requires more input information than LSTM-based to reach similar accuracy. The CNN model performs worse than both HMM and LSTM model since it requires an even higher amount of information to correctly predict SIP dialogs. However, the CNN model still outperforms the predictor based on cross-correlation which demonstrated poor predictive performance as it only achieves a high probability of correctly predicting SIP dialogs when more than 90% of information (L_o^k/L_d^k) is available.

While the correlation-based classifiers show both improved classification accuracy and significantly reduced computation time, the results in Figure 8 show that this technique is not well-suited for SIP dialog prediction tasks. We have intentionally included this comparison to emphasize the poor predictive performance of the proposed prediction method, which stands in stark contrast to its strong performance in classification. This distinction highlights the limitations of the correlation-based approach when applied beyond classification, reinforcing the need for alternative techniques such as the ones proposed in [9] and [11] for SIP dialog prediction.

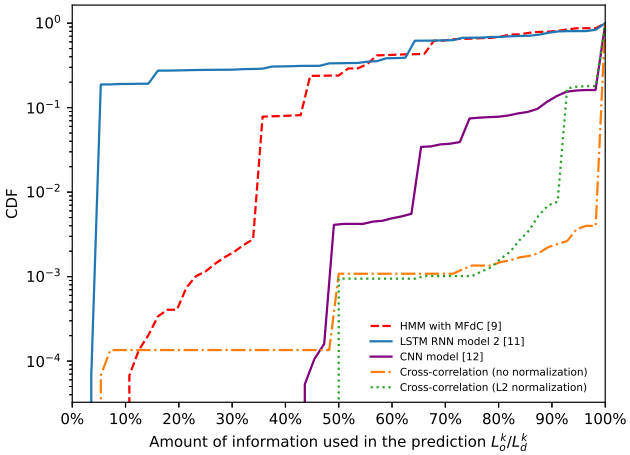


FIGURE 8. SIP dialogs prediction probability over the amount of available information, i.e., in terms of the number of SIP messages that integrate the SIP dialog.

V. CONCLUSIONS

In this paper, we presented a novel approach for detecting abnormal SIP dialogs, which leverages the cross-correlation of SIP dialog patterns and associated statistical metrics to classify ongoing communications. Our method demonstrated a 100% true positive rate and a false positive rate of approximately 8.6%, leading to an overall accuracy of 98.91% and showcasing its robustness in accurately distinguishing between normal and abnormal SIP dialogs. These results suggest that our approach is highly effective in identifying potential threats in SIP communications.

Unlike prior methods, which often rely on computationally intensive Bayesian or deep learning models, the proposed methodology offers a simpler yet efficient alternative. By focusing on the statistical properties and patterns within SIP dialogs, the proposed classification based on correlation eliminates the need for complex models or large training datasets. This provides a significant advantage in terms of computational efficiency, with our technique being over 1000 times faster than existing Bayesian and deep learning-based methods proposed in [9], [11], and [12].

In comparison to the optimal Bayesian and deep learning-based approaches, our method achieves a slightly improved classification performance but with the added benefit of lower computational cost, making it well-suited for real-time threat detection in high-performance network environments. Although the false positive rate achieved by our method is slightly higher than the optimal Bayesian method, the trade-off between the classification accuracy and the computational efficiency makes our approach a promising solution for practical deployment.

A. DISCUSSION AND FUTURE WORK

Our paper presents one step forward towards automatic detection of anomalous SIP dialogs by presenting a classifier that is both fast and accurate. The detection of anomalous SIP dialogs is of great importance since it can help prevent malicious activity and provide resistance to common attacks such as flood attacks and DoS. We demonstrate that the use of statistical methods is highly relevant in certain applications since it can achieve accuracy levels comparable to machine

learning methods, but at a fraction of the computation time and computational resources. Although our work is focused on SIP, future work could include the application of similar statistical methods to other protocols.

When applying the developed classifier in real-world settings, several challenges arise and must be considered, namely latency and overhead, memory and storage requirements, and handling of corrupted or incomplete messages. The integration with existing SIP systems must be simple and seamless without causing major latency and overhead. Since our classifier is based on statistical analysis, it has significantly lower computational requirements (memory and storage) and is much faster than other classifiers based on machine learning. This makes it easier to integrate with existing SIP systems even when resources are scarce, and since it is also faster than other methods, it causes very low overhead.

Once integrated in a SIP server, our classifier can quickly classify dialogs as normal or anomalous and immediately take appropriate action (e.g., terminating a session once an anomalous SIP dialog is detected) to prevent any damage. The classifier could be used as a standalone anomaly detection system by identifying anomalous SIP dialogs and immediately taking action without human intervention. Alternatively, it could be used in conjunction with different classifiers so that they complement each other and lead to better attack prevention. As we observed, our classifier excels at the classification of full SIP dialogs but is lacking in the prediction of incomplete SIP dialogs, and thus could be used to complement machine learning models to enhance overall detection accuracy.

Regarding the handling of corrupted or incomplete SIP messages, our classifier can still perform its function, although some accuracy degradation is to be expected as mainly false positives, but also false negatives, may increase due to the limited information. Although presenting these limitations, our classifier remains a cost-effective and fast solution with high accuracy.

Future work may focus on further optimizing the method to reduce the false positive rate while maintaining its efficiency, as well as exploring its scalability in larger network infrastructures. Additionally, integrating other statistical and machine learning techniques could enhance the robustness of the detection system, enabling more nuanced anomaly detection in dynamic environments.

REFERENCES

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "Sip: Session initiation protocol," *Internet Requests for Comments*, RFC Editor, RFC 3261, June 2002.

[2] A. Uzelac and Y. Lee, "Voice over ip (voip) sip peering use cases," *Internet Requests for Comments*, RFC Editor, RFC 6405, November 2011.

[3] Y.-H. Lu, S. H.-Y. Hsiao, C.-Y. Li, Y.-C. Hsieh, P.-Y. Chou, Y.-Y. Li, T. Xie, and G.-H. Tu, "Insecurity of operational ims call systems: Vulnerabilities, attacks, and countermeasures," *IEEE/ACM Trans. Netw.*, vol. 31, no. 2, pp. 800–815, 2023.

[4] X. Huang, W. Qi, X. Xia, Y. Sun, Z. Sun, and M. Peng, "Iot ntn for voice services: Architectures, protocols, and challenges," *IEEE Netw.*, vol. 38, no. 4, pp. 40–47, 2024.

[5] D. Sisalem, J. Floroiu, J. Kuthan, U. Abend, and H. Schulzrinne, *SIP Security*. Wiley Telecom, 2009.

[6] A. D. Keromytis, "A comprehensive survey of voice over ip security research," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 514–537, 2012.

[7] J. Seedorf, "Security challenges for peer-to-peer sip," *IEEE Netw.*, vol. 20, no. 5, pp. 38–45, 2006.

[8] K. S. Trivedi, M. Grottke, and J. A. Lopez, "Rethinking software fault tolerance," *IEEE Trans. Rel.*, vol. 73, no. 1, pp. 67–72, 2024.

[9] D. Pereira, R. Oliveira, and H. S. Kim, "A machine learning approach for prediction of signaling sip dialogs," *IEEE Access*, vol. 9, pp. 44 094–44 106, 2021.

[10] ——, "Abnormal signaling sip dialogs detection based on deep learning," in *2021 IEEE 93rd Veh. Technol. Conf. (VTC2021-Spring)*, 2021, doi: 10.1109/VTC2021-Spring51267.2021.9448664.

[11] ——, "Classification of abnormal signaling sip dialogs through deep learning," *IEEE Access*, vol. 9, pp. 165 557–165 567, 2021.

[12] D. Pereira and R. Oliveira, "Detection of abnormal sip signaling patterns: A deep learning comparison," *Comput.*, vol. 11, no. 2, 2022, doi: 10.3390/computers11020027.

[13] D. Geneiatakis, T. Dagiuklas, G. Kambourakis, C. Lambrinoudakis, S. Gritzalis, K. S. Ehlert, and D. Sisalem, "Survey of security vulnerabilities in session initiation protocol," *IEEE Commun. Surveys Tuts.*, vol. 8, no. 3, pp. 68–81, 2006.

[14] S. Ehlert, D. Geneiatakis, and T. Magedanz, "Survey of network security systems to counter sip-based denial-of-service attacks," *Comput. & Secur.*, vol. 29, no. 2, pp. 225 – 243, 2010, doi: 10.1016/j.cose.2009.09.004.

[15] D. Sisalem, J. Kuthan, and S. Ehlert, "Denial of service attacks targeting a sip voip infrastructure: attack scenarios and prevention mechanisms," *IEEE Netw.*, vol. 20, no. 5, pp. 26–31, 2006.

[16] S. H. Islam, P. Vijayakumar, M. Z. A. Bhuiyan, R. Amin, V. Rajeev M., and B. Balusamy, "A provably secure three-factor session initiation protocol for multimedia big data communications," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3408–3418, 2018.

[17] Murat Semerci and Ali Taylan Cemgil and Bülent Sankur, "An intelligent cyber security system against DDoS attacks in SIP networks," *Computer Networks*, vol. 136, pp. 137–154, 2018.

[18] I. M. Tas, B. G. Unsalver, and S. Baktir, "A novel sip based distributed reflection denial-of-service attack and an effective defense mechanism," *IEEE Access*, vol. 8, pp. 112 574–112 584, 2020.

[19] J. Tang and Y. Cheng and Y. Hao and W. Song, "SIP flooding attack detection with a multi-dimensional sketch design," *IEEE Trans. Depend. Sec. Comput.*, vol. 11, no. 6, pp. 582–595, 2014.

[20] A. Febro, H. Xiao, and J. Spring, "Telephony denial of service defense at data plane (tdosd@dp)," in *NOMS 2018 - 2018 IEEE/IFIP Netw. Oper. Manage. Symp.*, 2018, doi: 10.1109/NOMS.2018.8406281.

[21] Z. Tsiatsikas, G. Kambourakis, D. Geneiatakis, and H. Wang, "The devil is in the detail: Sdp-driven malformed message attacks and mitigation in sip ecosystems," *IEEE Access*, vol. 7, pp. 2401–2417, 2019.

[22] H. Li and H. Lin and H. Hou and X. Yang, "An efficient intrusion detection and prevention system against SIP malformed messages attacks," in *2010 Int. Conf. on Comput. Aspects of Social Networks*, 2010, pp. 69–73.

[23] Dongwon Seo and Heejo Lee and Ejovi Nuwere, "SIPAD: SIP–VoIP anomaly detection using a stateful rule tree," *Computer Communications*, vol. 36, no. 5, pp. 562–574, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366412004124

[24] Nassar, Mohamed and State, Radu and Festor, Olivier, "Monitoring SIP traffic using support vector machines," in *Recent Advances in Intrusion Detection*, Lippmann, Richard and Kirda, Engin and Trachtenberg, Ari, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 311–330.

[25] Nazih, Waleed and Hifny, Yasser and Elkilani, Wail S. and Dhahri, Habib and Abdelkader, Tamer, "Countering DDoS attacks in SIP based VoIP networks using recurrent neural networks," *Sensors*, vol. 20, no. 20, 2020.

[26] N. Hentehzadeh, A. Mehta, V. K. Gurbani, L. Gupta, T. K. Ho, and G. Wilathgamuwa, "Statistical analysis of self-similar session initiation protocol (sip) messages for anomaly detection," in *2011 4th IFIP Int. Conf. New Technol., Mobility and Secur.*, 2011, doi: 10.1109/NTMS.2011.5720662.

[27] R. Ferdous and R. L. Cigno and A. Zorat, "On the use of SVMs to detect anomalies in a stream of SIP messages," in *2012 11th Int. Conf. on Machine Learning and Applications*, vol. 1, 2012, pp. 592–597.

[28] S. Marchal, A. Mehta, V. K. Gurbani, R. State, T. Kam Ho, and F. Sancier-Barbosa, "Mitigating mimicry attacks against the session initiation protocol," *IEEE Trans. Netw. Service Manag.*, vol. 12, no. 3, pp. 467–482, 2015.

[29] Nguyen, Thuy T. T. and Armitage, Grenville and Branch, Philip and Zander, Sebastian, "Timely and continuous machine-learning-based classification for interactive IP traffic," *IEEE/ACM Trans. Netw.*, vol. 20, no. 6, pp. 1880–1894, 2012.
[30] A. Sheoran, S. Fahmy, C. Peng, and N. Modi, "Nascent: Tackling caller-id spoofing in 4g networks via efficient network-assisted validation," in *IEEE INFOCOM 2019 - IEEE Conf. Comput. Commun.*, 2019, pp. 676–684.
[31] I. M. Tas and S. Baktir, "Blockchain-based caller-id authentication (bbca): A novel solution to prevent spoofing attacks in voip/sip networks," *IEEE Access*, vol. 12, pp. 60 123–60 137, 2024.
[32] A. Lahmadi and O. Festor, "A framework for automated exploit prevention from known vulnerabilities in voice over ip services," *IEEE Trans. Netw. Service Manag.*, vol. 9, no. 2, pp. 114–127, 2012.
[33] D. Bao, D. L. Carni, L. De Vito, and L. Tomaciello, "Session initiation protocol automatic debugger," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 6, pp. 1869–1877, 2009.
[34] D. Golait and N. Hubballi, "Detecting anomalous behavior in voip systems: A discrete event system modeling," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 3, pp. 730–745, 2017.
[35] Peter H. Westfall, "Kurtosis as peakedness, 1905–2014. R.I.P." *The American Statistician*, vol. 68, no. 3, pp. 191–195, 2014, PMID: 25678714. [Online]. Available: https://doi.org/10.1080/00031305.2014.917055
[36] Pan, Xunyu and Zhang, Xing and Lyu, Siwei, "Exposing image splicing with inconsistent local noise variances," in *2012 IEEE Int. Conf. on Comput. Photography (ICCP)*, 2012, pp. 1–10.
[37] Liang, Zhiqiang and Wei, Jianming and Zhao, Junyu and Liu, Haitao and Li, Baoqing and Shen, Jie and Zheng, Chunlei, "The statistical meaning of kurtosis and its new application to identification of persons based on seismic signals," *Sensors*, vol. 8, no. 8, pp. 5106–5119, 2008. [Online]. Available: https://www.mdpi.com/1424-8220/8/8/5106
[38] Supraja, "Kurtosis in practice: Real-world applications and interpretations," 2024. [Online]. Available: https://www.analyticsinsight.net/tech-news/kurtosis-in-practice-real-world-applications-and-interpretations
[39] M. Nassar, O. Festor et al., "Labeled voip data-set for intrusion detection evaluation," in *Meeting of the Eur. Netw. Univ. Co. in Inf. Commun. Eng.* Springer, 2010, pp. 97–106.
[40] S. Technologies, "Asterisk project, asterisk.org," 2020. [Online]. Available: https://wiki.asterisk.org/wiki/display/AST/Home
[41] O. S. Foundation, "Opensips," 2020. [Online]. Available: https://opensips.org/

CLARISSE FEIO is currently enrolled in the Ph.D. program in Electrical and Computer Engineering at NOVA School of Science and Technology (FCT NOVA). She received B.Sc. and M.Sc. degrees in Electrical and Computer Engineering from FCT NOVA. In 2023 she obtained a second M.Sc. degree in Information Security and Cyberspace Law from IST ULisboa. Her main interests reside in telecommunications and cybersecurity.

DIOGO PEREIRA is currently enrolled in the Ph.D. program in Electrical and Computer Engineering at NOVA School of Science and Technology (FCT NOVA). He obtained his B.Sc. and M.Sc. degrees in Electrical and Computer Engineering from FCT NOVA. His research interests lie in the areas of stochastic processes applied to computer and data science, wireless mobile networks, and network modeling.

RODOLFO OLIVEIRA (S'04-M'10-SM'15) received the Licenciatura degree in electrical engineering from the Faculdade de Ciências e Tecnologia (FCT), Universidade Nova de Lisboa (UNL), Lisbon, Portugal, in 2000, the M.Sc. degree in electrical and computer engineering from the Instituto Superior Técnico, Technical University of Lisbon, in 2003, and the Ph.D. degree in electrical engineering from UNL, in 2009. From 2007 to 2008, he was a Visiting Researcher at the University of Thessaly. From 2011 to 2012 and in 2023, he was a Visiting Scholar at Carnegie Mellon University. Rodolfo Oliveira is currently with the Department of Electrical and Computer Engineering, at UNL. He is also affiliated as a Senior Researcher with the Instituto de Telecomunicações, where he researches Wireless Communications, Computer Networks, and Computer Science. He serves on the Editorial Board of Ad Hoc Networks (Elsevier), ITU Journal on Future and Evolving Technologies (ITU J-FET), IEEE Open Journal of the Communications Society, and IEEE Communications Letters.

PEDRO AMARAL Received the Phd. in Electric and Computer Engineering in 2013 and the M.Sc. in Computer Engineering in 2006 from Universidade Nova de Lisboa. He is an Assistant Professor at Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, a researcher at Instituto de Telecomunicações, Lisboa and a IEEE member. Current research interests include model free resource optimization algorithms in Networks, AI enhanced network managment and control and Network security.