Computer Communications 201 (2023) 116–122

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

A novel approach for flow analysis in software-based networks using L-moments theory

Jesús Galeano-Brajones [a,*], Mihaela I. Chidean [b], Francisco Luna [c], Javier Carmona-Murillo [a]

[a] Department of Computing and Telematics System Engineering, Universidad de Extremadura, Mérida, 06800, Extremadura, Spain
[b] Department of Signal Theory and Communications, Universidad Rey Juan Carlos, Fuenlabrada, 28942, Madrid, Spain
[c] School of Computer Science and Engineering, Universidad de Málaga, Málaga, 29071, Andalucía, Spain

ARTICLE INFO

Keywords:
Network traffic analysis
L-moments theory
Intelligent network management
6G
Machine Learning

ABSTRACT

The continuous increase in the number of devices connected to the Internet, together with the growth of applications and services, has made the tasks of network traffic analysis and classification essential in any environment. The deployment of 5G networks has prompted the research community to establish the pillars of Next-Generation Networks. These include intelligent systems, providing the network with intelligence in management and security tasks. In addition, these tasks require mechanisms capable of characterizing traffic in order to make network decisions. In this context, this paper proposes a novel methodology for processing network traffic using the L-moments theory and Machine Learning algorithms. This methodology is robust to outliers, requires few data to characterize the flows and subsequently fit the classification models. The results show that L-moments are particularly useful for processing network flows, and the classification algorithms obtain very high-quality results. Moreover, we show that the considered statistical tools also allow for a better understanding of the attack behaviour, leading the way to the improvement of the feature selection in similar problems.

1. Introduction

The major leap towards intelligent network management is thanks to 5G, mainly due to the introduction of software-defined, virtualization and slicing, among other techniques. These techniques allow services and applications to be virtualized on the network so that intelligent systems can be deployed as applications for both network management and security purposes. Moreover, the massive and continuous increase in network traffic makes the need to analyse and classify it even more essential. The sixth generation (6G) is not yet completely defined, but the research community agrees that these technologies will remain crucial. Furthermore, as intelligent systems evolve towards network self-management, Artificial Intelligence (AI) becomes much more important in Next-Generation Networks (NGNs) being the key characteristic of 6G autonomous networks [1].

Network and service management in 5G, Beyond 5G and especially 6G networks, including network security, are expected to be completely autonomous [2]. To achieve this, these networks will be driven on the Zero-touch network and Service Management (ZSM) concept defined by *European Telecommunications Standards Institute* [3]. This paradigm aims to integrate AI into the network as a key technology supported by software-defined and virtualization techniques. In this way, networks will be able to manage themselves by taking decisions without the need for human intervention [4], thereby optimizing capital expenditure and operating expenses [5]. For achieving this automation, network traffic analysis and classification techniques are crucial to provide networks with relevant information to guide them in taking accurate decisions.

In this scenario, network traffic analysis is a hot topic for the scientific community, specifically from the network security assessment point of view. In this area, different techniques have been employed for threat detection by analysing network traffic and flows: (i) port-based analysis is the simplest and no longer useful technique due to the large proliferation of new services and applications using non-IANA well-defined ports [6]; (ii) Deep Packet Inspection (DPI) emerges as an alternative, but its major limitations are that it is only applicable to non-encrypted packets and the problems regarding the user's privacy, leading the way for the proposal of Machine Learning (ML) or Deep Learning (DL) techniques to mitigate these drawbacks [7]; (iii) payload-based technique uses only the information contained in the application layer payload and is usually deployed together with DPI [6]; (iv) statistical-based approaches use payload-independent parameters (e.g., `flow duration`, `inter-arrival time`, `header length`, etc.), which can be used as input to different statistical, ML or DL

[*] Corresponding author.
E-mail addresses: jgaleanobra@unex.es (J. Galeano-Brajones), mihaela.chidean@urjc.es (M.I. Chidean), flv@lcc.uma.es (F. Luna), jcarmur@unex.es (J. Carmona-Murillo).

https://doi.org/10.1016/j.comcom.2023.01.022
Received 6 September 2022; Received in revised form 16 December 2022; Accepted 31 January 2023
Available online 2 February 2023

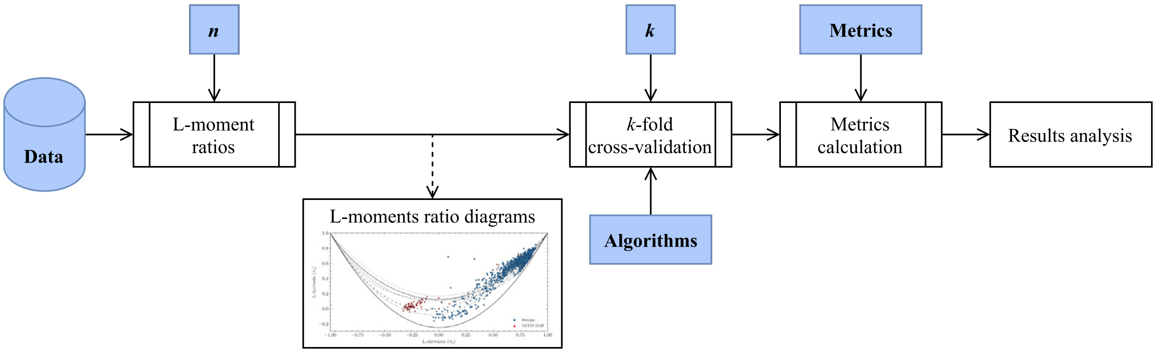

Fig. 1. Complete framework. The stages of the proposed methodology are represented with rectangles connected with arrows. Blue elements represent the modifiable parameters in this framework.

models. Finally, in recent years, there has been a growth in DL models applied to the network traffic classification [8].

L-moments have been widely used in different research fields since their proposal in 1990 [9], and network security and management has not been one of the most significant ones. The field with the most applications is climate analysis, especially regional frequency analysis [10]. Some specific examples include modelling probability distributions of wind and precipitation [11]. However, they have also been used in other fields like bioengineering for target classification in radar applications [12] and in the context of complex network theory [13]. Some additional examples include financial data and stock analysis [14,15], reliability disciplines [16], mathematical modelling of mechanical processes [17] or medical data [18]. Finally, as far as authors know, L-moments have only been used in two works related to network traffic analysis: (i) in [19] L-moments are used to fit the generalized Pareto distribution to network traffic data, especially to a heavy-tailed data sample; (ii) in [20] L-moments are used to characterize network flows. The latter is one of the first approaches of the authors to this methodology and a preliminary work of the present article.

This article proposes a novel methodology to classify network traffic data using L-moments and ML algorithms. L-moments allow the use of higher-order statistical moments avoiding the restrictions regarding the required amount of data for the estimation procedure. This advantage, together with the requirement of low computational resources, allows real-time data processing. Being this the first formal proposal of this methodology, the ML algorithms considered are k-Nearest Neighbours (kNN) and Support Vector Machines (SVM). In order to show the applicability of the proposed methodology, the experimentation has been performed with the CIC-DDoS2019 dataset [21]. This dataset contains scenarios with different up-to-date realistic DDoS and DrDoS attacks.

There are significant differences between [20] and the present work: (i) in this work, traffic data are analysed in a realistic way, i.e. data flows are not previously divided into benign/attack flows; (ii) in this work, we actually analyse and classify the traffic data using different state-of-the-art algorithms; (iii) in this work we consider a more realistic and state-of-the-art database, focusing on a specific attack. In short, [20] is just an exploratory work where the authors shown that network traffic data could be analysed with the L-moment statistical theory, while this is a complete analysis in a realistic scenario.

The rest of the document is organized as follows. Section 2 focuses on the theoretical background and technological basis of the proposed methodology; Section 3 details the set-up and the experimental evaluation conducted to validate our proposal; Section 4 shows and discusses some results obtained after the application of the proposed methodology, and also provides future directions for research. Finally, Section 5 concludes this article.

2. Framework and methods

This section describes the complete framework as well as each stage of the proposed methodology. Fig. 1 shows these stages, indicating in blue the inputs that can be modified. These inputs are described as follows:

- **Data** — input dataset. A cybersecurity-related dataset in this work, however, this methodology can be applied in other fields.
- ***n*** — amount of samples used to estimate each L-moment ratio, i.e., each point of the L-moment ratio diagram (LmomRD).
- ***k*** — number of folds used in cross-validation.
- **Algorithms** — network traffic classification algorithm. This methodology allows the usage of multiple algorithms for the classification task.
- **Metrics** — evaluation metrics used for results analysis.

The following subsections provide a more in-depth description of each of the stages of this methodology. First, L-moments and LmomRD are briefly described. Then, the two ML algorithms used in this article are defined, although any type of classification or clustering algorithms can be used in this methodology.

2.1. L-moments

In data analysis, statistical moments are used to characterize the geometry of distributions and summarize samples. Standard statistical practise is based on "classical" or "conventional" moments, also known in the literature as product moments. However, product moments are just one of the available moment definitions, being the L-moments theory [9] the selected framework for this work.

In short, the L-moments are calculated by means of a linear combination of the expected values of order statistics. L-moments are suitable for data with large skew, large or long tails, or outliers [10, 22], characteristics that several variables obtained from network flow data fulfil [20]. Furthermore, L-moment estimators are unbiased, robust to outliers and with low sampling variability [9,10], leading to more accurate and precise estimations than product moments. Also, the sample size required to accurately estimate L-moments is significantly lower than for the product moments [9]. Further details regarding L-moments, such as their formal definition as well as their basic properties and estimators, can be found in [9].

Another great benefit of using L-moments is that this theory is parallel to the product moment theory also in terms of interpretation. That is, the first L-moment (λ_1) is defined as *L-location* and equals the mean of the distribution or average value of the dataset; this is the only case where the values are the same for both statistical theories. The second L-moment (λ_2) is known as *L-scale* and gives insight into the scale of dispersion, the third one (λ_3) describes the asymmetry, the

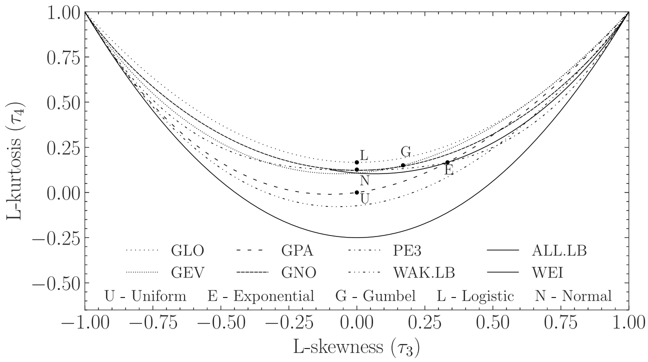

Fig. 2. LmomRD of some common distributions (GLO: Generalized Logistic; GEV: Generalized Extreme Value; GPA: Generalized Pareto; GNO: Generalized Normal; PE3: Pearson Type 3 or Gamma; WEI: Weibull; WAK.LB: Lower bound of the Wakeby distribution; ALL.LB: Lower threshold for any distribution) [9].

fourth L-moment (λ_4) is related to the tails of a given distribution, and so on.

The standardized versions of λ_3 and λ_4 are named as *L-skewness* (τ_3) and *L-kurtosis* (τ_4), respectively. They also have the same interpretation as the skewness and kurtosis in classical statistics, e.g., $\tau_3 > 0$ (< 0) indicates positive (negative) symmetry and $\tau_4 > 0$ (< 0) indicate positive (negative) kurtosis. L-skewness and L-kurtosis are both lower and upper-bounded by definition for all distributions, a very interesting property that allows, for example, the direct comparison between distributions with significantly different locations and scales.

In this work, τ_3 and τ_4 will be estimated for the selected network traffic parameters and will be the input to the classification algorithms.

2.2. LmomRD

The L-moment theory provides also an extremely useful graphical tool: the L-moments ratio diagram, previously defined as the acronym LmomRD. This tool is mainly used for exploratory analysis as well as for distribution selection tasks, however in this work enables a visual result presentation, interpretation and comparison.

The LmomRD plots tuples (usually pairs) of L-moment ratios, each element in one axis. The most common pair of L-moment ratios to be related using this diagram is the $\{\tau_3, \tau_4\}$ one. It is also common to include in the LmomRD the theoretical L-moment ratios for some common distributions (see Fig. 2). In order to facilitate the interpretation and result comparison, all figures presented in this work will also include these theoretical lines, following the same legend.

2.3. Algorithms

As previously mentioned, the proposed methodology can include any clustering algorithm, classification technique, and even more complex ML or DL models. Basically, this framework can include any type of algorithm capable of classifying the points of the LmomRD. Given all available algorithms that fulfil the previous requirement, in this work two different yet state of the art representative algorithms are considered. In the following, these are briefly described; please refer to the original publications for further details.

The first considered algorithm is kNN, the non-parametric classification method proposed in 1951 [23]. kNN is a method used for both regression and classification since in both cases the algorithm takes as input the k samples closest to the dataset. If used for regression, the output of the algorithm is the average of the values of the k nearest neighbours. If used for classification, the output is a property class of the object based on its k neighbours.

The second considered algorithm is SVM. These are a set of supervised learning algorithms that analyse data in order to perform classification or regression tasks and outliers detection. Proposed in 1992 [24], it has become one of the most robust prediction methods available currently. An SVM is a model that represents the data samples in space, separating the classes by a hyperplane or set of hyperplanes. Each hyperplane is defined as the vector between the points of the two nearest classes, which is called the support vector.

2.4. Evaluation metrics

In this work, results are quantitatively analysed using the *balanced accuracy* metric, which is defined as follows:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{1}$$

where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives and FP is the number of false positives. The balanced accuracy is a widely accepted metric in the scientific literature and it is suitable for unbalanced datasets, like the one considered in this work (see Section 3.1 for details regarding the dataset). Recall that the proposed framework allows any evaluation metric.

3. Experimental setting

In order to validate the usefulness of the proposed framework, we evaluate it using a state-of-the-art cybersecurity-related dataset. This section describes the experimental setting and the considered dataset.

3.1. CIC-DDoS2019 dataset

There are multiple network traffic datasets available for the research community, each considering specific scenarios and applications, and even with a variety of DDoS attacks. As the attacks are continuously evolving and presenting new challenges, new datasets are created that contain the latest information about the attacks.

In this work, we use the CIC-DDoS2019 dataset [21], as nowadays can be considered as the state-of-the-art dataset for any work that analyses network threats, specifically DDoS threats. This dataset has been generated by the Canadian Institute for Cybersecurity (CIC) with the aim of remedying all current deficiencies related to DDoS attacks. The dataset contains traffic flows belonging to different types of DDoS attacks that resemble actual real-world data.

In addition to the captured traffic, the authors of the dataset provide labelled CSV files generated by the CICFlowMeter-V3 tool. CICFlowMeter-V3 is a tool designed by CIC to perform analysis of the captured flows. The flow features obtained by this tool are based on the time stamp, source and destination IPs and ports, protocols, packets, inter-arrival time between packets, etc.

The CIC-DDoS2019 dataset contains an abstract behaviour of 25 users using the HTTP, HTTPS, FTP, SSH and email protocols. It includes network flows and CSV files for 10 DrDoS and 12 DDoS attacks captured in two days.

3.2. Experimental application

The experimental application implements all the stages included in the considered framework, starting from the initial parameters configuration and ending with the result representation. It automates the process of analysing the network flows of the input dataset. Besides calculating the L-moments and L-moment ratios to train classification algorithms, the application can perform an automatic analysis to establish which features of the dataset are the most promising to obtain the best classifications. This approach is very useful before deploying the trained models in the intelligent network since these models will be trained with the most promising features and will be as efficient as possible.

Table 1
Balance accuracy scores obtained for all the considered scenarios. Columns indicate the scenario and rows indicate the classification algorithm. The last row includes accuracy results from [26] for comparison purposes.

	(a)	(b)	(c)	(d)	(e)	(f)
kNN-unif	.9994	.8708	**.9991**	**.9989**	.6660	.9370
kNN-dist	**.9995**	.9666	**.9991**	.9800	.7549	.9545
SVM-lin	**.9995**	.9791	**.9991**	.8584	.8438	.6125
SVM_RBF	.9994	.9791	**.9991**	.9795	**.9556**	**.9995**
SVM-poly	.9924	**.9916**	.9978	.9784	.6660	.9820
DIDDOS [26]	.9952	.9997	.9997	.9987	.9996*	.9998

First, the CIC-DDoS2019 dataset is loaded and properly organized using the Python available variable representation. Following, the L-moment ratios are calculated using the $n = 200$ value, meaning that for each point of the LmomRD a total of 200 data packets are used. The data packets are analysed by means of a non-overlapping sliding window. The $n = 200$ value was empirically determined during the initial tests and it is a trade-off between moment estimation accuracy and delay. Using lower n values lead to less accurate moment estimation, while using larger n values imply larger delays in the analysis. Recall that the considered L-moments are third and fourth-order statistical moments, therefore using such a low amount of data packets to properly estimate them is one of the main reasons the L-moment theory was included in this methodology.

At this point, the LmomRDs are plotted. This step is used mainly as an auxiliary step to visually observe the input to the classification algorithms increasing the user-friendliness of the application. Afterwards, the classification task is performed with either kNN and SVM algorithm, together with the cross-validation technique [25]. In this work, 5-fold cross-validation is performed (input parameter $k = 5$) and the folds are made by preserving the percentage of samples for each class. Finally, the balanced accuracy metric of the classifications for the trained models is computed with the test subset.

Regarding the algorithm-related parameters, both kNN and SVM are configured attending their particular features. For kNN, we consider $\mathrm{k} = \sqrt{N}/2$, being N is the size of the training set. The choice of an adjustable value for k is a consequence of the dataset characteristics, where each scenario has a different amount of data. With this k the kNN algorithm is able to properly adapt to each scenario, obtaining better accuracy and avoiding both under and over-fitting. We also consider the following weights functions: uniform and distance. The first one considers the same distance between neighbours, and the second one takes into account the distance between points in the classification space. These two cases will be labelled as "kNN-unif" and "kNN-dist", respectively, in the rest of this document.

On the other hand, for SVM we consider the following three kernels: linear, polynomial and Radial Basis Function (RBF). The first one creates a linear hyperplane; the second one uses a polynomial (degree 3) function; the last one uses $\gamma = 1/(n\_features \cdot \sigma^2)$, where γ is a scalar that defines how much influence a single training example has, $n\_features$ is the number of features and, σ is the variance. These three cases will be labelled as "SVM-lin", "SVM-poly" and "SVM-RBF", respectively, in the rest of this document.

4. Results and discussion

This section includes the presentation and discussion of the results, as well as a brief analysis of the main benefits and drawbacks of the proposed methodology.

The experimentation has been conducted with the complete CIC-DDoS2019 dataset. In this work, we show the results for a total of six different scenarios, in order to show the potential of this methodology. These scenarios differentiate one from another in terms of the considered attack (either DDoS or DrDoS), the flow feature and/or different traffic capture of the dataset, i.e., different capture days. The following list details the characteristics of each scenario:

(a) DrDos attack using a Network Time Protocol (NTP) vulnerability to amplify UDP traffic to the victim and benign traffic; *packet length mean* feature; attack captured on the first day of the dataset.
(b) DrDoS attack amplified by the Trivial File Transfer Protocol (TFTP) and benign traffic; *destination port* feature; attack captured on the first day of the dataset.
(c) Scenario with the same characteristics as scenario (b) except for the feature; in this case, the feature is *maximum packet length*.
(d) DrDoS attack amplified by Portmap and benign traffic; *packet length mean* feature; attack captured on the second day of the dataset.
(e) DrDoS, amplified by NetBIOS and by LDAP, and benign traffic; *packet length mean* feature; attack captured on the second day of the dataset. This is a scenario where two different attacks are considered and multi-class classification is applied.
(f) DDoS attack with TCP SYN flood, where the attackers initiate massive TCP connections to the victim without terminating the connection consuming the victim's resources hindering the ability to not respond to legitimate traffic; *minimum forwarding inter-arrival time* feature; attack captured on the second day of the dataset.

The obtained results are represented in Fig. 3, where each inset represents an LmomRD (τ_3 vs. τ_4), calculated using n data packets, identifying benign flows and attacks with different colours and markers. In order to better understand the classification results, each point is labelled as benign (attack) when the majority of the n data packets used for each L-moment calculation are labelled as benign (attack). These points are the input to each considered classification algorithm, and the obtained balanced accuracy scores for all cases are showed in Table 1. In both figures, each scenario is identified by the label used in the previous list and, in the following, we discuss the results for each scenario.

4.1. LmomRD

In most cases, benign and attack markers are blended in the LmomRD. In particular, these are the points where the proportion of n benign and attack data flows used to estimate each L-moment are similar. This mix will be the source of errors for the classification algorithms, an expected situation in these kinds of problems.

Starting with the considered features, let us start with scenarios (a), (d) and (e) where the same feature (*packet length mean*) is used to analyse different attacks. This feature selection is not casual and helps with the method validation. We can observe that, as expected, clusters corresponding to benign traffic, although with a small number of points, are concentrated around similar values of τ_3 and τ_4 across the three insets.

Regarding the cluster localization and shape in general, they entirely depend on the traffic type (attack or benign) and the selected feature. In general, benign traffic tends to have positive L-skewness, indicating that the data distribution follows a probability distribution where most of the data are concentrated in the lower range. Also, benign traffic tends to have positive L-kurtosis, indicating that distribution tails are heavier than for a Normal distribution, therefore outliers are more likely.

Attack behaviour in terms of the LmomRD shows two quite some different situations: low and high cluster dispersion. On one hand, scenarios (c), (d) and DrDoS-NetBIOS from (e) reveal a significantly high range for L-skewness. This fact indicates that the values of the measured feature do not necessarily concentrate around a "gravity point" and can also indicate changes in the data statistics over time. The dispersed cluster behaviour can be explained by the way DrDoS works and its impact on the considered features, as through the network travel both short-length request packets as long-length response packets. The

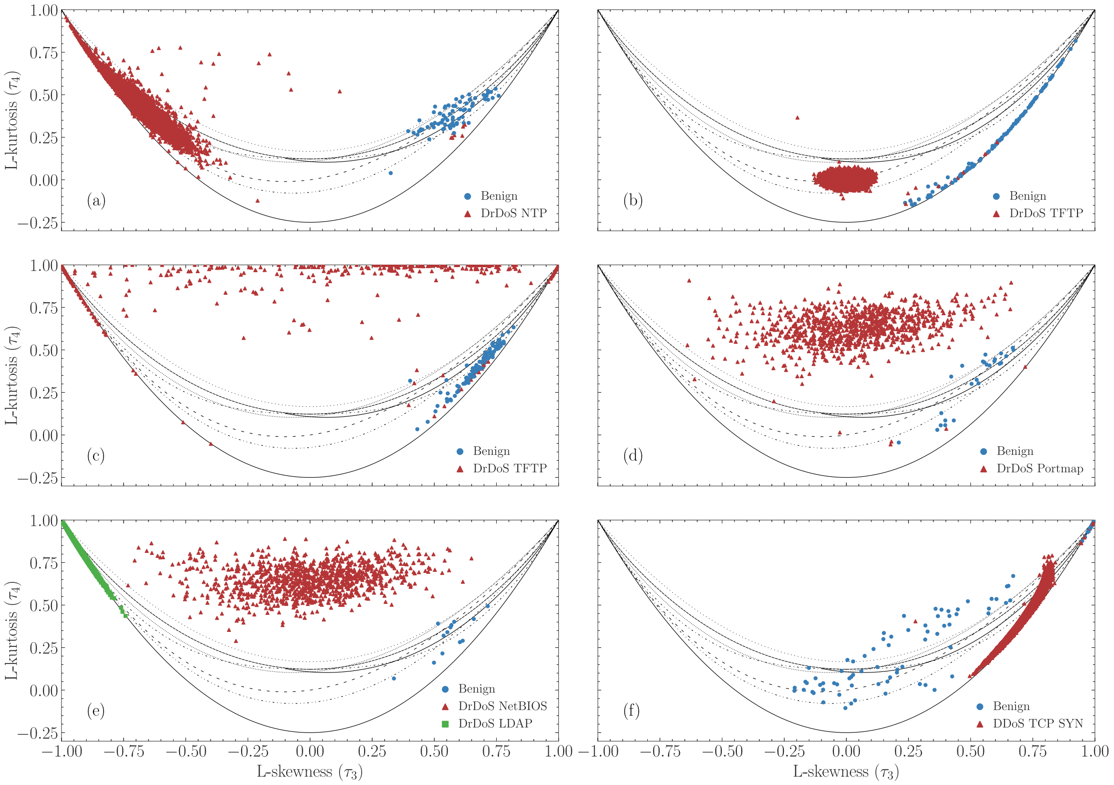

Fig. 3. LmomRD for the six considered scenarios. Each inset identifies the corresponding scenario in the left-lower corner. Attack and benign data are identified with different markers and colours, see legend for each scenario in the right-lower corner.

temporal behaviour would be very interesting to be analysed with more detail in future work, searching for example if there is some kind of relation between L-skewness and the duration of the attack and/or the temporal organization of the attack. In these three cases, the L-kurtosis is positive in all cases with rather high values, revealing heavy tails and therefore high outlier probability.

On the other hand, scenarios (a), (b), DrDoS-LDAP from (e) and (f) show less dispersion in the obtained clusters for attack traffic, but also with different behaviour. For example, DrDoS-NTP from (a) and DrDoS-LDAP from (e) show high negative L-skewness, while DDoS-TCP-SYN from (f) reveal high positive L-skewness, while the tree cases show slightly high positive L-kurtosis. These facts reveal that these features for these specific attacks concentrate around a “gravity point” (either at the lower or the higher side of the range) and have quite some heavier tails than the Normal distribution, i.e., high probability for outliers. The extreme values obtained for the L-skewness in the three cases are also due to the attack/feature combination: (i) attacks from scenarios (a) and DrDoS-LDAP from (e) do not get any response from the victim and only long-length packets generated by the amplification mechanism travel through the network; (ii) attack from scenario (f) establishes a high amount of connections with the victim that lead to a significant increase in the packet transmission rate and this fact is reflected in the considered feature. Finally, DrDoS-TFTP from (b) has both low L-skewness and low L-kurtosis, meaning that these features could be easily adjusted to a Uniform distribution. Again, this statistical behaviour is the expected one for this feature, as the DrDoS attack tries to collapse a device by flooding a specific port. The variation around the $(0, 0)$ point is due to the mix of benign and attack in the n packets used to estimate each L-moment.

4.2. Balanced accuracy

Once understood the scenario behaviours, the classification results are now analysed in terms of the balanced accuracy metric. These results are shown in Table 1 for all the considered algorithms (with their respective settings) and scenarios. Precision, recall, and F_1-score values have also been obtained to validate the results of the accuracy and can be found in the supplementary material. Best scores are marked in bold font, however, it can be observed that the balanced accuracy is quite high for the majority of the cases. From all considered scenarios, the SVM-RBF algorithm is the one that obtains a better-balanced accuracy score, even for the (e) scenario where other algorithms perform rather poor.

Table 1 also shows the results obtained for the accuracy score in a different work from the literature that analyses the same CIC-DDoS2019 dataset, but with a Gated Recurrent Unit (GRU), a type of Recurrent Neural Network (RNN) [26]. In order to properly compare these scores, recall that in [26] the authors balance the dataset by means of Synthetic Minority Oversampling Technique (SMOTE) and compute the accuracy, while in this work we analyse the original dataset and compute the balanced accuracy. Also note that the * in scenario (e) of Table 1 is due to the fact that DIDDOS classifies the two attacks involved individually, while our proposal performs a multi-class classification considering both attacks at the same time.

The approach followed [26] requires significantly higher computational resources for the RNN model compared with the methodology proposed in this work. However, it can be observed that their accuracy results differ quite little from the ones obtained in this work. Therefore, we can conclude that the methodology proposed in this work obtains results comparable in quality with more complex solutions published in the literature, with the clear benefit of requiring less computational resources for its implementation.

The previous results show a total of six scenarios, although the considered database includes other attacks and each flow has many other features. The presented results show some positive cases, where the combination of attack and feature leads to an adequate classification. However, there are also cases where the balanced accuracy is not high enough to properly separate attack from benign traffic. This situation is common in this type of problem with datasets where the amount of available features is high and in an actual implementation, an initial analysis and feature selection is unavoidable. Nevertheless, we consider that the present results show a sufficient variety of cases to properly validate the proposed methodology.

4.3. Drawbacks and benefits

Regarding the main pros and cons of the presented methodology, in the following, we summarize them and propose several future research lines.

On the drawback part, one of the most relevant ones is the requirement to perform feature selection in order to obtain high-quality results, and therefore high attack detection accuracy. However, this also occurs in most classification problems, and it can be resolved by either using information about the attack characteristics, performing exploratory analysis over the available database or even with automated procedures. Another drawback of this methodology is that categorical features would require preprocessing in order to define numerical values that would allow computing the corresponding L-moments. In any case, network traffic databases usually include many more numerical than categorical features and, depending on the scenario, this drawback could be ignored. Finally, this is a new method that we validated using a limited amount of scenarios and a specific database. For a full validation and therefore usefulness, this method should be also validated in an actual 5G scenario in a real-time operation, being this one of our main future research and work lines.

On the positive part, this methodology has lower computational complexity compared to other state-of-the-art procedures, as both the L-moment estimation and the considered classification algorithms (kNN and SVM) have low computational requirements [10,23,24]. It is worth to mention that the total computational complexity depends on the considered classification algorithm, however, we have shown that even simple algorithms like kNN lead to high-quality classifications. This methodology also allows a better understanding of the statistical behaviour of the data and even to study the temporal attack behaviour, thanks to the usage of the LmomRD. This information can be useful for the proposal of mitigation actions in an actual software-defined scenario. Another benefit is that this methodology can be easily adapted to include multi-feature analysis. This can be achieved by either including a multivariate classification algorithm or by introducing the L-comoments [10], e.g., L-correlation, in the framework, being this idea another of our main future research lines. The last pro that we would like to mention is the possibility to introduce in the framework higher-order L-moments. This additional characteristic is straightforward from a programming point of view, however, it would require some additional theoretical support for the result interpretation. When considering higher-order L-moments, also multidimensional classification will be enabled and LmomRDs with more than two dimensions could be considered.

5. Conclusions

In 5G networks, the increase in connected devices and traffic volume has highlighted the need to analyse network traffic and classify it for both intelligent management and security purposes. Therefore, and in order to contribute to the progress towards Zero-touch networks, this article proposes a novel methodology for analysing and classifying network traffic. This methodology is based on the use of the L-moment ratios, a tool that has proven to be very useful for this task and that, to the best of our knowledge, has not been previously explored for this application. In order to validate the methodology, experimentation has been performed with the most up-to-date realistic dataset. The results allow us to validate the methodology, showing comparative results with another current proposal in the literature, and to propose various lines of future research.

CRediT authorship contribution statement

Jesús Galeano-Brajones: Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Mihaela I. Chidean:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision. **Francisco Luna:** Conceptualization, Validation, Writing – review & editing. **Javier Carmona-Murillo:** Conceptualization, Validation, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset used is referenced in the document.

Acknowledgements

This research was funded in part by the European Union NextGenerationEU/PRTR, grant TED2021-131699B-I00 (AEI/FEDER,UE), by the Spanish Ministry of Science and Innovation, grant numbers PID2020-112545RB-C54 and PDC2022-133900-I00, by the Regional Government of Extremadura, Spain, grant IB18003, and by the Univ. Rey Juan Carlos Program for Research Promotion and Development (Ref. F920 and “AYUDA PUENTE 2022, URJC” Ref. F931).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.comcom.2023.01.022. It contains the tables with the results of the evaluation metrics of the classifications as a complement to Table 1.

References

[1] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G.K. Karagiannidis, P. Fan, 6G wireless networks: Vision, requirements, architecture, and key technologies, IEEE Veh. Technol. Mag. 14 (3) (2019) 28–41.

[2] M. Bunyakitanon, X. Vasilakos, R. Nejabati, D. Simeonidou, End-to-end performance-based autonomous VNF placement with adopted reinforcement learning, IEEE Trans. Cognit. Commun. Netw. 6 (2) (2020) 534–547.

[3] ETSI, GSZS, Zero-touch network and Service Management (ZSM); Reference Architecture, Tech. Rep, 2019.

[4] C. Benzaid, T. Taleb, AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions, IEEE Netw. 34 (2) (2020) 186–194.

[5] M. Bagaa, T. Taleb, J.B. Bernabe, A. Skarmeta, Qos and resource-aware security orchestration and life cycle management, IEEE Trans. Mob. Comput. (2020).

[6] H.-K. Lim, J.-B. Kim, K. Kim, Y.-G. Hong, Y.-H. Han, Payload-based traffic classification using multi-layer LSTM in Software Defined Networks, Appl. Sci. 9 (12) (2019) 2550.

[7] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, J. Aguilar, Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey, IEEE Commun. Surv. Tutor. 21 (2) (2018) 1988–2014.

[8] S. Rezaei, X. Liu, Deep Learning for Encrypted Traffic Classification: An Overview, IEEE Commun. Mag. 57 (5) (2019) 76–81.

[9] J.R. Hosking, L-moments: Analysis and estimation of distributions using linear combinations of order statistics, J. R. Stat. Soc. Ser. B Stat. Methodol. 52 (1) (1990) 105–124.

[10] W.H. Asquith, Univariate Distributional Analysis with L-Moment Statistics using R (Ph.D. thesis), 2011.

[11] M. Fawad, T. Yan, L. Chen, K. Huang, V.P. Singh, Multiparameter probability distributions for at-site frequency analysis of annual maximum wind speed with L-moments for parameter estimation, Energy 181 (2019) 724–737.

[12] R. Ginoulhac, F. Barbaresco, J.-Y. Schneider, J.-M. Pannier, S. Savary, Target Classification Based On Kinematic Data From AIS/ADS-B, Using Statistical Features Extraction and Boosting, in: 2019 20th International Radar Symposium, IRS, IEEE, 2019, pp. 1–10.

[13] F. Mohd-Zaid, C.M. Schubert Kabban, R.F. Deckro, A test on the L-moments of the degree distribution of a Barabási–Albert network for detecting nodal and edge degradation, J. Complex Netw. 6 (1) (2018) 24–53.

[14] J.R.M. Hosking, L-Moments and their Applications in the Analysis of Financial Data, IBM Thomas J. Watson Research Division, 1999.

[15] E. Jurczenko, B. Maillet, P. Merlin, Efficient frontier for robust higher-order moment portfolio selection, 2008.

[16] N.U. Nair, B. Vineshkumar, L-moments of residual life, J. Statist. Plann. Inference 140 (9) (2010) 2618–2631.

[17] S. Cao, H. Lu, Y. Peng, F. Ren, A novel fourth-order L-moment reliability method for L-correlated variables, Appl. Math. Model. 95 (2021) 806–823.

[18] P. Royston, Which measures of skewness and kurtosis are best? Stat. Med. 11 (3) (1992) 333–343.

[19] J. Hosking, Some theory and practical uses of trimmed L-moments, J. Statist. Plann. Inference 137 (9) (2007) 3024–3039.

[20] M.I. Chidean, J. Carmona-Murillo, R.H. Jacobsen, Q. Zhang, Network Traffic Characterization Using L-moment Ratio Diagrams, in: 2019 Sixth International Conference on Internet of Things: Systems, Management and Security, IOTSMS, IEEE, 2019, pp. 555–560.

[21] I. Sharafaldin, A.H. Lashkari, S. Hakak, A.A. Ghorbani, Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy, in: 2019 International Carnahan Conference on Security Technology, ICCST, IEEE, 2019, pp. 1–8.

[22] R.M. Vogel, N.M. Fennessey, L moment diagrams should replace product moment diagrams, Water Resour. Res. 29 (6) (1993) 1745–1752.

[23] E. Fix, Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties, vol. 1, USAF school of Aviation Medicine, 1985.

[24] B.E. Boser, I.M. Guyon, V.N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992, pp. 144–152.

[25] F. Mosteller, J.W. Tukey, Data analysis, including statistics, in: Handbook of Social Psychology, vol. 2, 1968, pp. 80–203.

[26] S. ur Rehman, M. Khaliq, S.I. Imtiaz, A. Rasool, M. Shafiq, A.R. Javed, Z. Jalil, A.K. Bashir, DIDDOS: An approach for detection and identification of distributed denial of service (DDoS) cyberattacks using gated recurrent units (GRU), Future Gener. Comput. Syst. 118 (2021) 453–466.