IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 February 2023, accepted 6 March 2023, date of publication 9 March 2023, date of current version 14 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3255101

RESEARCH ARTICLE

Cybersecurity Alert Prioritization in a Critical High Power Grid With Latent Spaces

JUAN RAMÓN FEIJOO-MARTÍNEZ[1], ALICIA GUERRERO-CURIESES[2], FRANCISCO GIMENO-BLANES[3,4], MARIO CASTRO-FERNÁNDEZ[1], AND JOSÉ LUIS ROJO-ÁLVAREZ[2,4], (Senior Member, IEEE)

[1]Red Eléctrica de España, Alcobendas, 28109 Madrid, Spain
[2]Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Fuenlabrada, 28943 Madrid, Spain
[3]D!lemmaLab Ldt Startup, Fuenlabrada, 28943 Madrid, Spain
[4]Departamento de Ingeniería de Comunicaciones, Universidad Miguel Hernández de Elche, 03202 Elche, Spain

Corresponding author: Juan Ramón Feijoo-Martínez (jrfeijoo@ree.es)

ABSTRACT High-Power electric grid networks require extreme security in their associated telecommunication network to ensure protection and control throughout power transmission. Accordingly, supervisory control and data acquisition systems form a vital part of any critical infrastructure, and the safety of the associated telecommunication network from intrusion is crucial. Whereas events related to operation and maintenance are often available and carefully documented, only some tools have been proposed to discriminate the information dealing with the heterogeneous data from intrusion detection systems and to support the network engineers. In this work, we present the use of deep learning techniques, such as Autoencoders or conventional Multiple Correspondence Analysis, to analyze and prune the events on power communication networks in terms of categorical data types often used in anomaly and intrusion detection (such as addresses or anomaly description). This analysis allows us to quantify and statistically describe high-severity events. Overall, portions of alerts around 5-10% have been prioritized in the analysis as first to handle by managers. Moreover, probability clouds of alerts have been shown to configure explicit manifolds in latent spaces. These results offer a homogeneous framework for implementing anomaly detection prioritization in power communication networks.

INDEX TERMS Telecommunication security, intrusion detection, deep learning, high power, power communication, latent variables, alert prioritization, alert manifolds.

I. INTRODUCTION

Red Eléctrica de España (REE, or Spanish Power Grid in English) is the transmission and management operator of the Spanish high-voltage power grid. One of the main responsibilities in this electric grid operation is ensuring the high reliability of the vast set of associated telecommunication assets, thanks to a highly reliable telecommunication network containing around 30 000 km of optical-fiber cables built throughout the country. This network covers all needs arising from the conveyed information continuously interchanged among relays and installations. This associated fiber network constitutes the backbone of the company telecommunication network.

REE holds a maintenance team whose mission is to ensure the continuity and quality of the telecommunication services. In addition, highly qualified personnel manage and supervise the REE network, acting as a single command for incidents, failures, redirection of traffic, security attacks, provision and service start-up, problem resolution, and work coordination. One of its primary functions is monitoring, evaluating, registering, and managing any alerts related to intrusion attempts and security breaches throughout the year. This information is obtained by implementing an intrusion detection system (IDS) that monitors the extensive network of thousands of computerized devices and IP-connected resources extending over hundreds of kilometers. A sheer number of events is

The associate editor coordinating the review of this manuscript and approving it for publication was Wencong Su.

generated throughout the day on one single network of this size. Identical circumstances can be found in other industries and networks, such as distribution companies, transmission companies, or large entities that manage public infrastructures. In the case of REE, the alerts are generated by the IDS system and then recorded for subsequent monitoring. Finally, a report is automatically generated for each one of the events. However, the extensive reports generated and the massive number of them do not facilitate an adequate interpretation or proper use to prioritize the necessary actions to be carried out. For these reasons, organizations with this type of reality should refrain from restricting themselves to producing summary reports based on existing systems and reports, rather, they require additional actions to be taken. Therefore, this is an actual and shared problem in all large network infrastructures. Proper management requires either large deployments of human resources or additional algorithmic developments, allowing more efficient alert handling and prioritization.

Over the last few years, organizations, companies, and institutions have created human teams devoted to detecting cybersecurity threats. These groups are commonly organized around divisions referred to as Security Operation Centers (SOC). A SOC is a department or unit focusing on security at a technical level. The goal of a SOC is to detect security incidents through network monitoring for abnormal behavior that may reveal that security has been compromised. SOC activities may include reverse engineering to study the incidents, as well as many other proposed tools for statistical analysis of the dynamic characteristics of the network. This reality and its challenges have been collected and published in some works [1], [2]. This literature tells us how the activities and initiatives taken by the work members of the SOC are exhausting and stressful. Moreover, given the enormous amount of information and the absence of adequate tools for its analysis and management, often they do not lead to being able to anticipate breakdowns caused by these eventualities. To the extent that this analysis is overloaded with clutter and tedious tasks, there is a need to create more sophisticated tools to scale detected events, as automated or semi-automated premature alarm detection becomes virtually impossible inside SOC. In particular, REE management states that finding several thousand waves of alerts in a single day is relatively standard. An identified bottleneck is the human cognitive ability to process information, and a pernicious side effect is the consequent burnout of staff working in these units. For these reasons, we analyzed here the events generated in the communications network that supports the Spanish power grid backbone, where practical and adequate management becomes essential to supply an essential service such as electricity countrywide.

Deep Learning techniques are considered valuable tools to deal with the inherent messiness of the heterogeneous data available in these and in other scenarios. Deep Learning is a set of structured algorithms that perform automatic learning to gain insight and knowledge. It stands out because it does not require programmed rules, as the algorithms can learn to perform a task through a training phase. It is also characterized by usually being composed of intertwined layers for information processing. It is mainly used for the automation of classification and regression problems learning from datasets available in the organizations [3], [4].

To carry out this work, we evaluated a significantly large dataset with actual records obtained from the REE system. The information in this dataset has been obtained from the IDS tool deployed per international and sectorial standards applicable to the discipline. The stored information consists of a set of records based on maintenance forms that reflect anomalies in the telecommunications network, where a set of attributes is associated with each record. These attributes include the time of occurrence, their related IP and MAC addresses, their geographical location, and the equipment involved. The systematic and advanced exploitation of this knowledge, despite the inherent complexities of the data typology, has been the main objective of this work. For this study, we applied simple Deep Learning techniques, namely several different autoencoders (AE) implementations, and compared their performance with a classical and extended analysis method, the Multiple Correspondence Analysis (MCA), for categorical variables. We determined that low-dimensional latent variable spaces can provide us with efficient representations of IDE-generated alerts, which allow us to prioritize those alerts and improve their management and reporting.

The main findings of this work are highlighted here below:

- It is possible to improve and prioritize the thousands of alerts of potential security and intrusion risks generated by the powerful management tools of large electricity infrastructures
- The analysis of the heterogeneous information, mostly categorical, based on anomaly detection strategies, has been of interest when: (i) on the one hand, detecting events of potential risk, (ii) and on the other to prioritize the thousands of events that occur daily and make their segmentation unfeasible by the human team responsible for network security
- The visualization of latent spaces in the intermediate stages has proven useful for the classification of varieties of alerts that will eventually lead to a more efficient analysis
- The identification of anomalies based on automatic encoders improved traditional multivariate methods (MCA), while beyond the detection of potential security risks, they have also expressed interest in detecting other types of network incidents of interest for infrastructure management.

This document is structured as follows. In Section II, we include a systematic review of the literature related to this work. In Section III, we present the framework and application techniques used in the proposed experiments. Section IV describes the dataset and collects the experiments

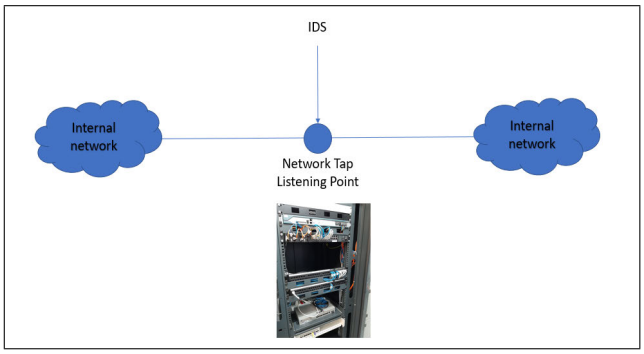


FIGURE 1. Network tap listening point.

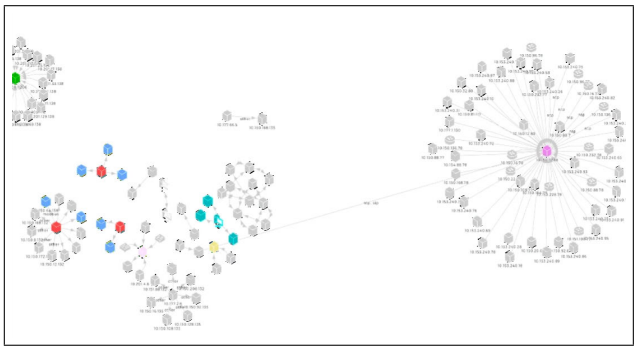


FIGURE 2. REE IDS graph.

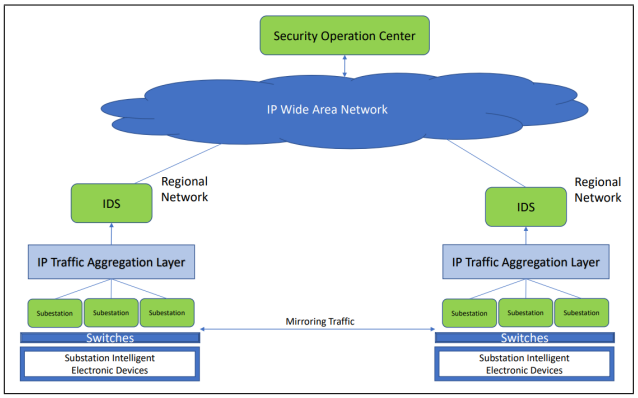


FIGURE 3. REE aggregation scheme.

and results, while Section VI compiles the main conclusions and discussion of the work.

II. BACKGROUND AND PROBLEM SCOPE

A. CRITICAL-INFRASTRUCTURE PROTECTION

Essential infrastructures are currently facing new circumstances and risks that they must face from a global perspective. On the one hand, the fundamental advances and undoubted innovations offered by new technologies in all facets of business activity are no exception in the case of electrical infrastructures. Furthermore, neither are the intrinsic vulnerabilities that come with these new technologies [5]. As a consequence, cybersecurity has become a severe concern for electricity companies, given that these *a priori* non-essential new elements of the power grid, but undoubtedly now required for better management and efficient development, have revealed the opening of security gaps that hackers take advantage of to put entities, companies, and administrations in check. An example of this can be found in the events that recently took place in Ukraine [6]. It should be known that in the recent past, the planning and design of networks did not foresee the existence of multiple networks and related services. For this reason, the systems were developed under the perspective of a single isolated network. Therefore, the due precautions for potential concurrences were not established. The rapid technological evolution and the long useful life of these devices tell us that these initial premises are no longer valid nowadays. Current networks incorporate additional protection layers on top of the existing infrastructure to cope with this reality. However, they do not always manage to eliminate vulnerabilities, and eventual risks of cyberattacks [7].

New and more powerful tools are being developed in different industries to respond to this reality. In the case of the power industry, these IDS tools combine software and hardware to identify potential anomalies or incidents that may or may not be related to fraudulent or pernicious activities. Figure 1 presents a diagram of the IDS deployment implemented in REE. IDS provides comprehensive inspection of packets and information units, thus protecting them against attacks. Additionally, IDS allows the detection of network elements, denial of service attacks, and access attacks. IDS tools also detect possible anomalies based on traffic analysis. For this purpose, the system studies the existing pattern of the traffic, considered normal and commonly recognized as the baseline. It identifies anomalies or potential attacks based on dissimilarities from the baseline. This network monitoring and analysis modality is considered a non-intrusive system.

B. DETECTION ARCHITECTURE

Different frameworks and regulations are focused on Cybersecurity implementations for critical infrastructures. A particular practical framework is shown in the North America Electric Reliability Corporation (NERC) regulations. The NERC is responsible for the secure and reliable functioning of the electric grid of North America and is the regulatory body for all users, owners, and operators of the Bulk Energy Supply (BES) system. The standards defined by this organization and their cybersecurity applications are well considered worldwide as a guide for the electric industry and for almost any mature organization seeking a secure digital ecosystem [8].

In response to a good number of the recommendations presented by the aforementioned organization, REE defined a security architecture that establishes an electronic security perimeter for each critical set of IEDs. In the same way, IDS was deployed to the network, making a set of alerts and incident information available to the team of professionals,

as shown in Fig. 2. Particular interest has been placed on the trunk nodes of the IP network to ensure a complete network analysis. Additionally, the topology established at the hierarchical level offers the IDS system visibility over each significant element in the REE network. Our architecture scheme is depicted in Fig. 3.

The IDS whole system architecture is passive with listening span ports. These span ports have no logical address associated with being undetected by any potential intruder and are used solely for monitoring, never for management or transmission. Hence the system is immune to injection attacks or tampering and reduces the attack surface. Moreover, this ensured a fail open condition, leaving the corporate communications unaffected in case any device failed. The IP traffic is collected in the power substations using switches with mirroring ports enabled. Afterward this traffic is aggregated in a middle layer through TAPs (Traffic Aggregation Points), devices that assemble the whole IP traffic without the filtering that a conventional switch would do, such as broadcasting messages. As the number of installations is several hundred, with this intermediate aggregation layer, the targeted traffic is conveyed to a regional IDS. Subsequently, final aggregated traffic is carried upwards from the regional IDS into the SOC to be consolidated and redundancies eliminated.

C. SECURITY ALERT PROCESSING

The sensitivity and specificity of security alert systems are crucial and a primarily debated issue. The rationale behind it is twofold. On the one hand, more significant sensitivity levels can imply low predicting capabilities of false negatives while ensuring effective use of the always limited resources in SOC. On the other hand, attempting to prioritize specificity could limit the number of false negatives and dramatically increase the number of false positives, thus limiting the effective segmentation and management of the alerts inside the SOC. The need to guarantee network security encourages those managers responsible for these departments to approach this second problem statement and, consequently, to a massive volume of alerts to which the technicians must respond. And the ever-increasing demands in terms of security and the emergence of new techniques for breaching almost any system are forcing SOC managers to take said second approach regardless of the limitations of human resources available for this matter and triggering a real challenge.

This work aims to respond to this reality by using novel and powerful machine learning tools to prioritize the efforts of security service operators to manage appropriately the vast number of generated alarms.

D. RELATED WORKS

Event analysis based on packet scrutinization has been recognized as a helpful tool for managing security and constitutes an effective instrument for anomaly detection in telecommunication networks. An example of this has been published in [14], where the authors present a model for intrusion detection. Similarly, authors in [15] describe how deep learning can help to identify main events related to cyberattacks. Autoencoders have also been applied to predict successfully several kinds of harmful intrusions [4], [16]. An overview of intrusion detection can be seen in [17].

As seen from the growing number of publications, deep learning offers a promising paradigm for modeling network intrusion [9], [18]. This fact manifests itself in all domains and is no less accurate in network security, where its impact is just beginning. A classical attack modeled with deep learning is the Mirai malware [3], [19]. Several other works illustrate this new reality and the potential that these techniques offer in telecommunications, where these methodologies are extended beyond classical prediction to specific security applications [20], [21].

From the more specific point of view of alert classification noteworthy works has been made. In [22], a whole set of methods quantifying IDS accuracy have been analyzed. Remarkably the ratio employed has been derived from false and true positives and negatives. Another approach in grouping alerts from an IDS is elaborated in [12] with the results of alert clustering aiming to discover attack scenarios. In [23] are presented various deployments related to using machine learning classifiers and prioritization in IDS.

On Table 1, several recent and representative works related to the present one are compiled and summarized [9], [10], [11], [12], [13]. It can be noted that, to our knowledge, no previous work has been driven with the orientation and scope of our proposal, as far as it is a new risk management discipline in data networks associated with critical infrastructures. We have not found precedents that either analyzed information extracted from tools from an IDS cluster as depicted them or dealt with instance volumes of about half a million from the current activity of a critical power system operator, whose registers can be far from synthetic ones. The limitations of applicability of the past works for this present application could be related to the inherent multidimensional complexity of the supervision data and the difficulty to extract patterns in this peculiar case. This effect is to be partially eluded by a tailored methodology conforming an operational technology framework that strengthens the results.

III. ALGORITHMIC FRAMEWORK AND METHODS

Anomaly detection is a machine learning process that identifies different events in any system that diverge from the expected normal behavior. This well-known set of techniques has been exposed frequently today due to the intense development of digitization and the systematic registration of information. Anomaly detection algorithms entail alerts, including the additional information that gives rise to the anomaly. The volume and characteristics of the anomalies or alarms may drive the need for secondary efforts to verify the risk derived from the detected anomalies and the actions to be taken, where said actions may differ depending on how critical the incidents are. Priority will be different if we run a critical

TABLE 1. Related works during the last years.

Ref.	Year	Method	Objectives	Database	Motivation
[9]	2022	Support Vector Machines (SVM) Decision Trees	Reduce false positives in zero-day attacks	IDS 2018 Intrusion CSVs (1 048 575 entries)	Classification
[10]	2021	Naïve Bayes (NB) Support Vector Machines (SVM) K- Nearest Neighbor (KNN)	Feature selection	NSL-KDD Dataset (1 074 992 entries)	Prioritization
[11]	2020	Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) Latent Semantic Analysis (LSA)	Correlate and filter alerts without feature extraction	Malware Capture Facility Project (5 548 539 entries) + CIC-IDS2017 (431 860 entries)	Prioritization
[12]	2021	Stream Clustering	Classify Networ IDS alerts	Academic institution IDS output (22 327 086 entries)	Prioritization
[13]	2020	Kademlia-based Distributed Hash Table	Prioritize alerts in IoT	DARPA 1999 dataset + DShield (600 millions entries)	Prioritization

service fault or some anomalous behavior of a key element. Still, its detection does not limit the functionality or whether it is a malicious intrusion with unforeseeable consequences.

This previous essential classification, together with the increasing number of alarm records to avoid false negatives, makes it necessary to develop new automatic or semi-automatic tools that carry out this earlier segmentation. This section describes the techniques evaluated here to achieve this goal, namely, the conventional MCA and deep learning AE with different configurations. Their mathematical foundations are summarized next.

A. MCA FOR CATEGORICAL FEATURE DESCRIPTION

MCA is a data analysis technique for categorical data used to detect and represent underlying structures in a low-dimensional Euclidean space. MCA is a method in multivariate statistical analysis that can be seen as the equivalent of Principal Component Analysis (PCA) in metric features since both are based on matrix eigendecompositions of the input data. The main difference between them is that the former represents a variance orthonormalization, whereas the latter represents a probability-space orthonormalization. MCA is a multivariate statistical analysis technique for categorical data that detects hidden relationships in a dataset. MCA searches to provide an alternative description of the discrete observation matrix on a space of lower dimensionality. This new representation would retain as much information as possible about the original data matrix, having a lower number of dimensions.

MCA transforms the categorical input variables to construct a binary matrix $\mathbf{X} \in \mathbb{R}^{I\times J}$, being I and J the total number of samples and categories in the dataset, respectively. J is the sum of all the categories related to the categorical features. Then, MCA decomposes this $\mathbf{X}$ matrix into a pair of sets of projection matrices (also known as factor scores), one for the dimension of the samples (rows) and another for the dimension of the categories (columns), to project the samples and features to points in a low-dimensional space, allowing to establish intuitive relationships among them through distances in this projection space. In addition, let

$$\mathbf{B} = \frac{1}{M}\mathbf{X}^{\top}\mathbf{X} \tag{1}$$

denote the corresponding relative frequency matrix. B is known as the Burt matrix, being $M = I \times J$ the number of elements of the $\mathbf{X}$ matrix.

Furthermore, we define $\mathbf{r}$ and $\mathbf{c}$ as the vectors of the horizontal and vertical sum of $\mathbf{B}$, respectively, i.e., $\mathbf{r} = \mathbf{1}^{\top}\mathbf{B}$ and $\mathbf{c} = \mathbf{B}\cdot\mathbf{1}$, where $\mathbf{1}$ is a column vector of ones. Then, let $\mathbf{D_c}$ and $\mathbf{D_r}$ represent diagonal matrices whose diagonal entries are the elements of $\mathbf{c}$ and $\mathbf{r}$, respectively. The subsequent decomposition of the normalized Burt matrix achieves the projection matrices:

$$\mathbf{D_r}^{-\frac{1}{2}}(\mathbf{B} - \mathbf{r}\mathbf{c}^{T})\mathbf{D_c}^{-\frac{1}{2}} = \mathbf{P}\Delta\mathbf{Q}^{T}, \tag{2}$$

being $\mathbf{P}$ and $\mathbf{Q}$ the left and right singular vector matrices, respectively, and Δ a diagonal matrix with the singular values organized in decreasing order. Then $\Lambda = \Delta^2$ is the eigenvalues matrix. Since $\mathbf{B}$ is a symmetric matrix, $\mathbf{P}$ and $\mathbf{Q}$ are identical matrices, being $\mathbf{V} = \mathbf{P} = \mathbf{Q}$ the eigenvector matrix, and thus $\mathbf{F}$ is also equal to $\mathbf{G}$.

Once eigenvectors and eigenvalues are obtained, MCA results can be studied. According to each category, normalized eigenvalues could be used to display the dispersion surrounding the gravity center. The eigenvectors represent the different projecting directions, and their coefficients indicate the relative relevance of each category for each factor.

To increase the interpretability of the MCA, Bootstrap resampling techniques are used to obtain a point cloud that implicitly represents the empirical distribution of these projections in the projected space. Visualizing overlapping confidence regions allows for interpreting the statistical correlation relationship between the analyzed categories. In contrast, separate and distant regions indicate independent categories.

Bootstrap resampling [24], [25] is a technique that allows the generation of a new list of statistical measurements from a sampling with replacement of a dataset. If we denote the operator that obtains the MCA eigenvector matrix by $\Theta(\cdot)$, the computations described in Eqs. (1)–(2), as follows,

$$\mathbf{V} = \Theta(\mathbf{X}). \tag{3}$$

Applying Bootstrap resampling, where the asterisk * indicates the resampled statistical element, we can define the resampled data matrix as $\mathbf{X}^* \in \mathbb{R}^{I\times J}$ after sampling with

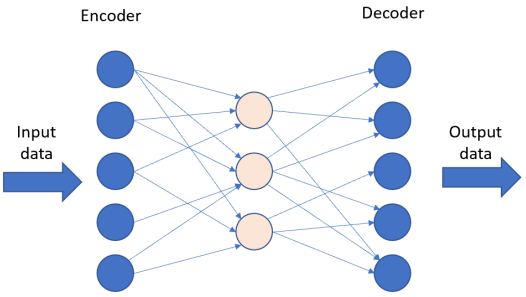


FIGURE 4. General architecture for an AE with a 3-unit intermediate layer.

replacement the $\mathbf{X}$ rows I times. This way, the $\mathbf{X}^*$ matrix is composed of rows of $\mathbf{X}$, whose rows may appear once, several, or zero times. Resampling B times, denoting $\mathbf{X}^*(b)$ as the resampled matrix in the b-th iteration bootstrap, where $b = 1, \ldots, B$, we obtain a bootstrapped representation of the projection matrix over the resampled population, $\mathbf{X}^*(b)$, as follows

$$\mathbf{V}^*(b) = \Theta(\mathbf{X}^*(b)). \tag{4}$$

Hence, using these B repetitions, we can estimate the empirical distribution function of the statistical element [24], in this case, of the projection vectors.

B. AUTOENCODERS

As mentioned before, Deep Learning technologies have undergone strong development in recent years, and today they constitute a promising battery of methods for extracting relevant information from large data volumes. Said technologies include a particular set of techniques called the AE [26], which are neural networks designed to learn complex and intrinsic relationships in the data. Typically, an AE consists of multiple neural network layers trained to reconstruct the input at the output. In contrast, different strategies are proposed in the intermediate layers to approach a certain objective [27]. Specifically, the AE can be contractive or expansive depending on the number of dimensions of their middle layers related to the input. One of the main advantages of using AE is their expressive capacity to transform complex high-dimensional input data into lower-dimensional structures in the intermediate layers. This reality, together with the expressive effect of the successive layers that allow the initial dimensional space to be generated at the output, offers a succinct and compressed representation model of the reality of the input in a lower-dimensional space. These spaces are known in the literature as latent spaces, bottlenecks, embeddings, or feature spaces. An example of a simple contractive AE can be seen in Fig. 4, where the intermediate layer works like a bottleneck to force some compression of the input data and to avoid the input values being memorized through the network. This AE has two components: the encoder (mapping the input data to the intermediate layer) and the decoder (mapping the intermediate layer to output data. Notice that within this approach, and since the AE has to reconstruct the input using a reduced number of nodes, it will try to retrieve only the most essential aspects of the input, that is, and ignore meager variations such as noise [27]. Although AE are trained to attempt to copy its input to its output, they are designed to achieve a copy that is approximate but not perfect. Using an intermediate layer of lower dimension than the input data (which is known as undercomplete AEs [28]), the model can learn useful properties of the data because it is forced to capture the most salient features of the training data, that is, those correlated with patterns of normal behavior which tend to form groups in a lower dimension space. Thus, new events which show statistical deviation from these normal patterns will be identified as anomalies.

AE architectures can be used for supervised and unsupervised learning. Unsupervised learning is a self-learning set of techniques trying to discover the intrinsic features of the input samples on its own, and no initial set of known categories is used for this purpose. This kind of approach has interest for its application in specific fields such as cybersecurity, where the wide variety of possible attacks makes their reality unknown concerning the characterization of regular network traffic. In this work, we propose establishing a framework based on intrinsic statistical descriptors of the available IDS data. However, the heterogeneity of the existing variables and their prevalent categorical nature makes it necessary to adequately manage and transform them before their analysis [26]. Also, for this particular application, it was deemed appropriate to use unsupervised learning as the problem of selecting the most critical alerts is new without proper knowledge, as the IDS implementation is brand new. Labels, which indicate explicitly whether an alert can be crucial, are not available. Therefore, we aimed to establish a similar framework for all of them while making the results and interpretability readily available to network administrators. Recall that the application field of the proposed machine learning solution is a telecommunication network embedded in a high-power grid and used to convey critical services such as electrical protection and supervision, whose security against intrusions needs to be extremely high. The availability of an essential collection of events from IDS forms has made possible its systematic analysis here.

The equation describing the first half of the AE, known as the encoder, can be expressed as follows,

$$\mathbf{h}_i = f(\mathbf{x}) = \phi(\mathbf{W}_e\mathbf{x}_i + b_e), \tag{5}$$

where $\mathbf{h}_i \in \mathbb{R}^d$ is the resultant vector in the feature space that maps the input $\mathbf{x}_i \in \mathbb{R}^d$, $f(\mathbf{x})$ is the transformation from the input to the new space, $\phi(\cdot)$ is the activation function, $\mathbf{W}_e$ is the weight matrix, and b_e is the bias. Seamlessly, the second side of the AE, called the decoder, can be expressed mathematically using the following expression,

$$\mathbf{z}_i = g(\mathbf{h}_i) = \varphi(\mathbf{W}'\mathbf{h}_i + b'), \tag{6}$$

where $g(\cdot)$ is the nonlinear transformation from the present space to the original space $\varphi(\cdot)$ denotes the nonlinear

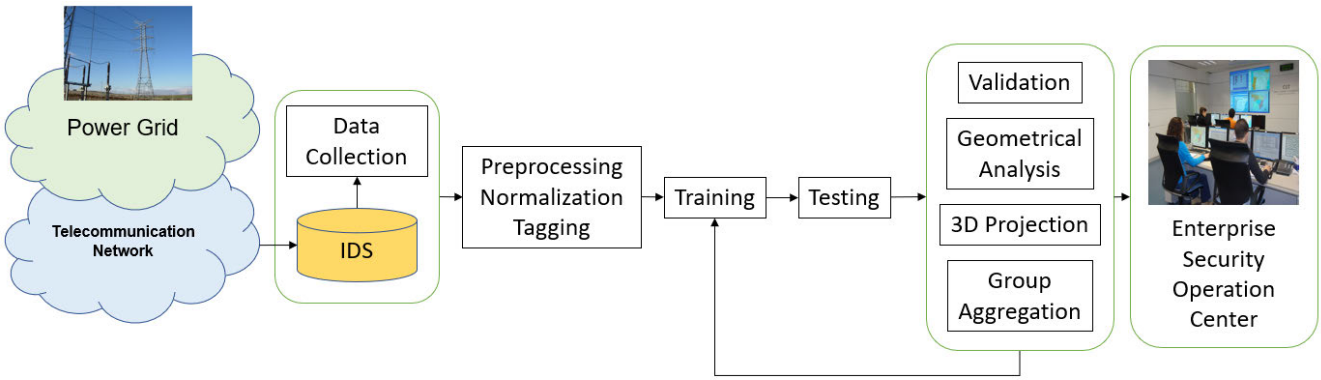


FIGURE 5. Overall scheme of the processing system.

activation, $\mathbf{W}'$ is the weight vector, b' is the bias vector, and $\mathbf{z}_i$ is the estimated output. The objective is to estimate $f(\cdot)$ and $g(\cdot)$ from a set of samples, therefore, the estimated weights and biases $\mathbf{W}$, $\mathbf{b}$ so that $\mathbf{z}_i = \hat{\mathbf{x}}_i$.

AE have been used for many different machine-learning tasks, and in this paper, we are interested in using autoencoders for alert prioritization. The main objective has been to train an autoencoder with an inner layer of three dimensions so that an intuitive representation of data was also possible in addition to the descriptive capability. As mentioned, with the appropriate constraints, AE can learn nonlinear and non-Euclidean projections of the data that are more interesting than other fundamental techniques. Our implemented AE network consists of an encoder and a decoder, where the encoder maps the input to a hidden representation that implies dimensional reduction. The decoder endeavors to map this representation back to the original space. Once the system is adequately trained using customary standard data, special consideration should be raised if a relevant difference between the coded and decoded data is observed. In other words, the data with a relevant reconstruction error become suitable candidates for further analysis as they do not match the standard patterns.

A relevant architecture of AE is the so-called Variational autoencoder (VAE), which can learn a continuous and statistically-principled latent variable model from its input data ([29]). For doing so, instead of letting the neural network learn an arbitrary function, the parameters of a probability distribution modeling the data are learned in the latent space. If points from this distribution are sampled, it is possible to generate new input data samples. A VAE is a generative model, meaning that VAE can encode inputs as distributions instead of points, and the obtained latent space structure is regularised to be parameterized by standard Gaussian distributions [30].

The following steps were implemented in the present application to benchmark the VAE capabilities. First, an encoder network turns the input samples x into two parameters in a latent space, which we will note z_{mean} and z_{log_sigma}. Then, we randomly sample similar points from the latent normal distribution that is assumed to generate the data via $z = z_{mean} + \varepsilon \cdot e^{(z_{log\_sigma})}$, where ε is a random Gaussian tensor. Finally, a decoder network maps these latent space points back to the original input data. The VAE model parameters can be trained via two loss functions: a reconstruction loss forcing the decoded samples to match the initial inputs (like in other AE); and the Kullback-Leibler (KL) divergence between the learned latent distribution and the prior distribution, which is acting as a regularization term. Being $\hat{y}$ and y two different probability distributions, their KL divergence [31] is defined as follows on their observations,

$$KL(\hat{y}||y) = \sum_{c=1}^{M} \hat{y}_c \log \frac{\hat{y}_c}{y_c}, \tag{7}$$

Note that a KL divergence close to 0 indicates that the two distributions in question have strongly similar information.

C. SYSTEM MODELING

The following phases were followed for system modeling purposes: data collection, data preprocessing, training and testing, algorithm application, and exploitation. Figure 5 shows the architecture of the running processing system. It consists of several subprocesses: the interface with the IDS for data collection, data preprocessing, data analysis, training, the algorithmic fabric, and the eventual exploitation in the company context. Some crucial issues are spotting the graphic results and adjusting the model to ensure practical results.

The core of the system is the machine learning algorithmia used to create the latent spaces with the IDS-generated alerts and to determine the prioritization in terms of the latent coordinates. Note that the experiments were performed to identify the best algorithmic implementation supporting the prioritization system among all the benchmarked and scrutinized algorithms. A final validation by REE supervisory personnel is needed. Tactical analyses by cybersecurity teams are made, conclusions are drawn and substantially false positives are elucidated.

TABLE 2. Data setup and description of the main fields.

Fields	Description
Alert id	Numerical used as primary key.
Type	Categorical that points to the type of alert.
Associated risk	Numerical used as an indicator of risk.
Description	String describing the alert.
Initial Time	Date/time value that marks the alert beginning.
IP direction	String with IP and mac direction.
Mac direction	String with IP and mac direction.
Protocol	String with the protocol used.
Installation	Location where the anomaly happens.
Equipment	Device or asset affected.

IV. DATA DESCRIPTION

Our dataset consisted of nearly half a million security alerts detected by the IDS. Information on these episodes was collected in the IP telecommunication network in REE. This sample included present active alerts at a specific time. Specifically, data selected for this work consisted of registers with 53 fields each that are currently stored in a *csv* file with the possibility to export to standard spreadsheets. Its structure is one-dimensional in a single table with no relation to other repositories. Table 2 shows a high-level description of some fields associated with the most relevant variables.

Most of the variables recorded in our database are inherently categorical or integer-valued numerical treated as categorical. In both cases, quantitative association measures are of interest. In the feature processing phase, text values were encoded by creating a mapping dictionary that mapped categories to a token. Then categorical variables (except risk) were replaced with their tokens, and data were converted to list format to match the network structure. A re-scaling was also made to adjust the data range. Redundant fields were filtered.

Data were split into train, validation, and test sets. The train set allows us to adjust the parameters of our model directly. The validation set is the one that allows us to adjust the hyperparameters of every model and to compare a large number of different models to select the best of them. The test set allows us to evaluate the final model performance.

V. EXPERIMENTAL RESULTS

In this section, we present the experimental results driving the design of a system for prioritizing alerts. Special attention is paid to the intrinsic structure of the alert vectors through several methods for generating latent spaces. Given that related features are intrinsically categorical, we start by scrutinizing the scope and limitations of MCA applied to this field, which represents a reference of the performance and information that can be retrieved from our problem using classical multivariate methods. Subsequently, we present the results with several AE structures, which can simultaneously highlight the intrinsic structure of the vector data on low-dimensional latent spaces, as well as provide information about some of the alerts as being different from the most usual states for these embedded spaces, thus representing usual and atypical alerts. As different AE architectures provide us with different system performances, the results with conventional AE, based on error reconstruction of the input vectors, are first presented. Then the advantages provided by using VAE are scrutinized and summarized in a set of representative experiments.

A. MCA AND DETECTION PERFORMANCE

As described before, MCA builds an eigenvector-based decomposition of the observed vectors. Given that MCA can be sensitive to loosely populated categories in the variables, we first revised the input variables of the complete dataset by category and grouped for each of them all the loosely populated categories into a single category labeled as *others*, which allowed the reduction in categories for variables *Type id* (starting from 18 initial categories and retrieving finally 5 categories), *IP source direction* (from 114 to 40), *IP destination direction* (from 189 to 10), *MAC source direction* (from 139 to 40), *MAC destination direction* (from 146 to 15), *Protocol* (from 36 to 6), and *Name* (from 18 to 5). This means that our data matrix finally included 649 categories from 7 variables.

As far as a large number of observations were available (499 013 alert vectors), building a single Burt matrix and inverting it was not viable. Instead, we can take advantage of the large numbers available and generate statistical instances of the Burt matrix and scrutinize the variability of the eigenvectors and the projections. We present the results of building a Burt matrix with 5000 randomly sampled alerts, obtaining the statistical fluctuations of its eigendecomposition and its corresponding elements (eigenvectors, projected features, and projected vector alerts), and repeating this process 300 times.

A slowly decreasing spectrum of eigenvalues was obtained (not depicted), showing strong correlations among categories. Fig.6 shows the first six eigenvectors, restricting the representation to a subset of categories to visualize the statistical properties. The red vertical lines denote the significant categories for each eigenvector. We used up to three eigenvectors to subsequently project the categories and the alert vectors, noting that they had 89, 63, and 0 significant categories each. Interestingly, the eigenvector number three of the representation and the fourth number often show symmetrical confidence intervals across the horizontal axis. This effect could be attributed to the rotation of the eigendecomposition axes through different random samples. It can also be observed that the confidence intervals in the two first components are narrow, thanks to the high number of input vectors used to build the Burj matrices. This fact also indicates that these first eigendirections are robust concerning rotations.

For each realization, the projection of each category was obtained on the three-dimensional latent space, as usual in MCA studies. Fig. 7 depicts each projected category on a different color, showing the structure of the alert vectors in terms of the mutual statistical information of each category.

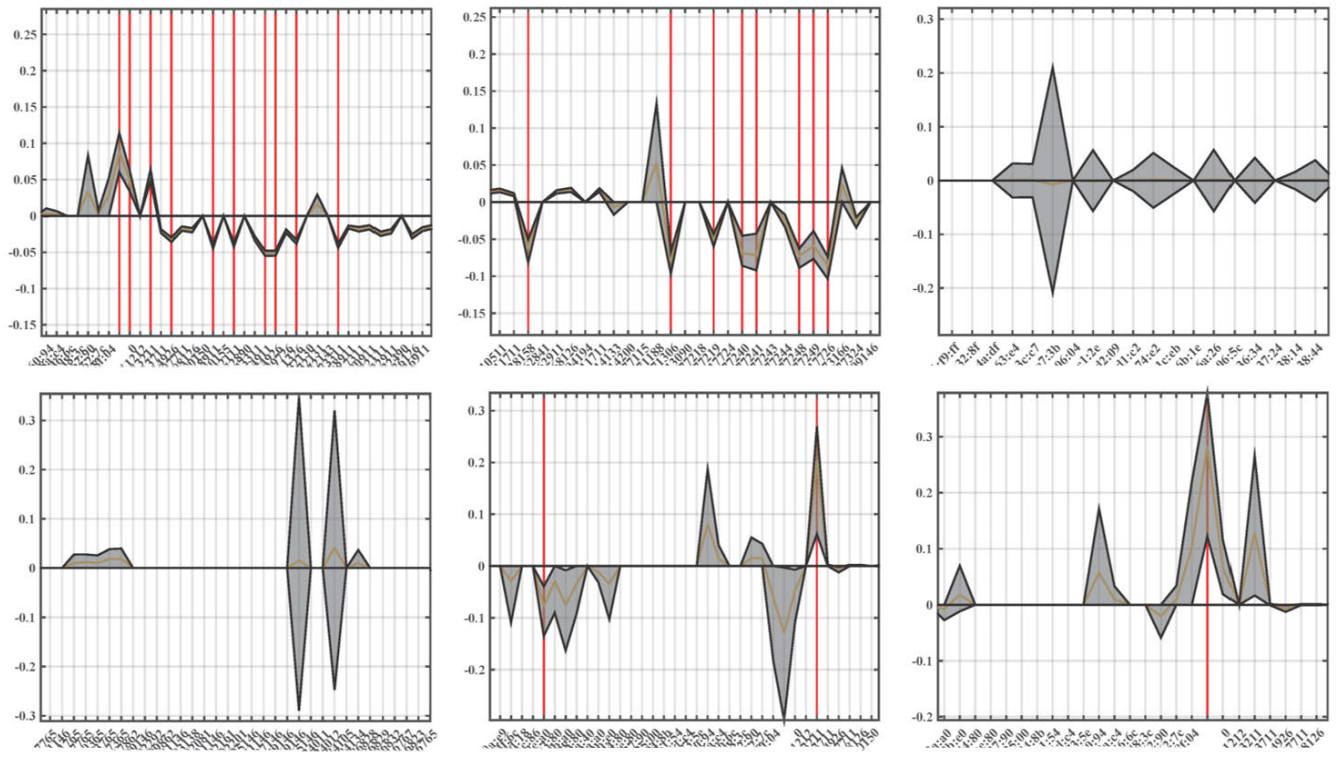


FIGURE 6. Eigenvectors and confidence intervals (grey bands) in MCA. Some categories are represented for each of the first 6 eigenvectors. Vertical red lines denote the statistically significant categories on each eigenvector, this is, those whose confidence interval does not overlap zero.

As explained elsewhere, confidence volumes were obtained using a domain description technique. This representation allows us to identify strongly related categories, as they are represented as overlapped confidence volumes. Also, those categories that are farther from the origin are less frequent. Whereas this information is sometimes evident and natural from an alert point of view, sometimes this represents relevant information for the manager. In general, we consider that this view represents a valuable tool for managers. For instance, multimodalities can be observed according to unconnected confidence volumes in the same color.

We also obtained the projection of each input alert vector onto the latent space build using the three first eigenvectors. Fig. 7 also shows the cloud point obtained in this way. We can appreciate that alerts on this projected space keep some strong spatial similarity properties and tend to group into non-Gaussian and manifold-like clouds. Note also that due to the strong effect of discretization, many of the single points in the figure correspond to many alert vectors that are initially equal or very similar in most of the features. We represented in the same figure the severity for each alert as provided by the IDS in terms of a color code, which can represent a qualitative proxy to detect those alerts as candidates to be prioritized. It should be mentioned at this point that the finding of the existence of morphologies and especially the concentration of risks in certain regions must be assessed for further interpretation.

From a result perspective based on graphical representation, we can appreciate that the study synthesized here allows us to estimate the method used and qualify its applicability. First, the method stability can be appreciated by the solidity, over the hundreds of realizations, of the values and vectors of the MCA transformation. This effect can be easily observed in Fig. 6, where the reduced confidence intervals of the coefficients (identified with red lines), especially in the first eigenvectors, indicate so. Secondly, the projection over the three first dimensions (first 3 eigenvectors) allow us to scrutinize the groupings of instances, as shown in Fig. 7. This representation allowed us to visualize clusters related to the different categories although overlapping (upper part of Fig. 7). This representation also allowed us to observe how the low-risk index obtained from IDS is concentrated in a region of the space (see the lower part of Fig. 7). It is important to notice that the categorical nature of the source guides the fact that the seemingly unique points may correspond to hundreds of them. Although these experiments have allowed us to conclude the existence of latent relationships between the categories, the difficulty of jointly processing many samples and the overlap of the identified groups limits the applicability.

B. SIMPLE AND MULTILAYER AE PERFORMANCE

The experiment data set was structured to scrutinize the scope of latent variable extraction from AE architectures. The

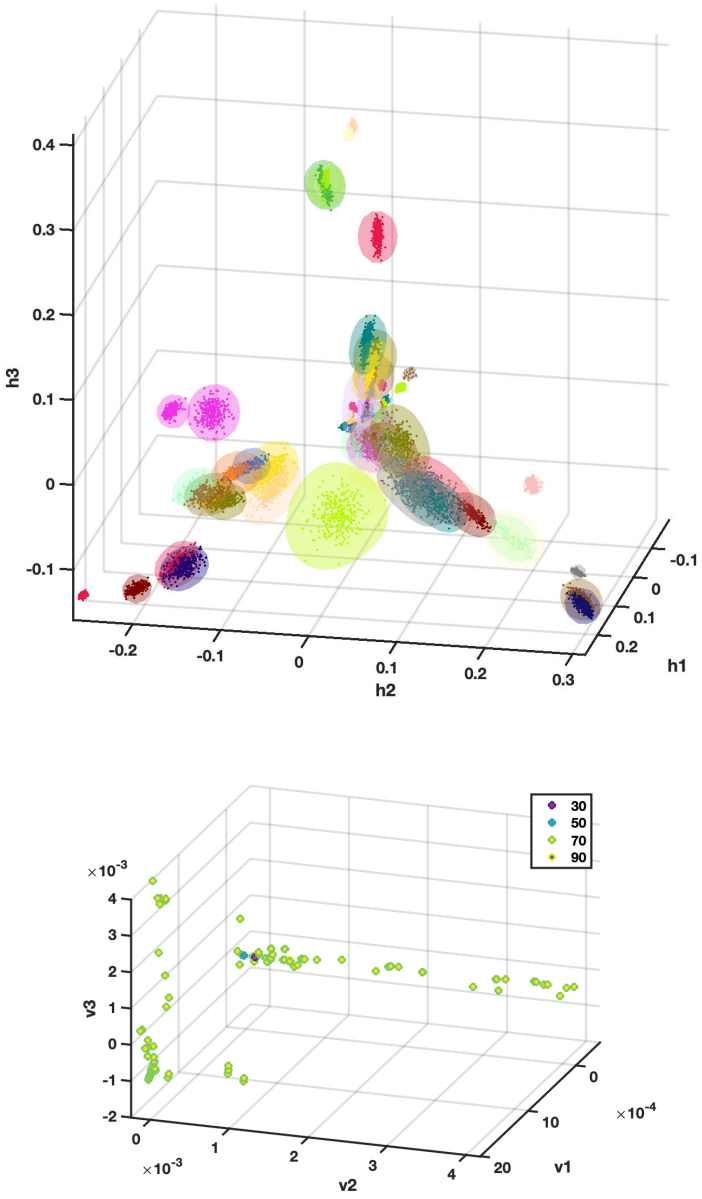


FIGURE 7. MCA Results. Up: Projected categories using MCA. The cloud points represent the aggregated projections through the 300 repetitions, and the colored volumes represent the confidence volumes for each category. Down: Projected alert vectors in MCA using the first three components as latent variables.

available alerts of the first dataset (up to 92508) were split into training (80%) and validation (20%). A second dataset, including over 499000 alerts, was used for testing. The following features were used: *MAC source direction*, *MAC destination direction*, *IP source direction*, *IP destination direction*, *Protocol*, and *Type*. This last one was considered optional, given that labeling the IDS alerts could affect the latent spaces and alter their structure. All of them were categorical features, so one hot encoding was applied. In this case, no filtering of the low-populated categories was performed, as AE architectures have been shown to exhibit robustness concerning this, in contrast with MCA implementations.

Initially, a single fully-connected neural layer was used as an encoder and a decoder, called *simple AE* in the following. The final model maps an input vector to the reduced latent variables and from them to its reconstruction. A plot of the encoded data obtained in the inner autoencoder layer (again using a 3-dimensional latent space) is shown in Fig. 8(a) when the *Type* feature that identifies the kind of alert is or is not considered during the training phase. The different colors represent the values of the data set *IP source direction*. Using *Type*, featured groups transformed the input data into different manifold structures. Therefore, the following paper results are obtained with and without the *Type*

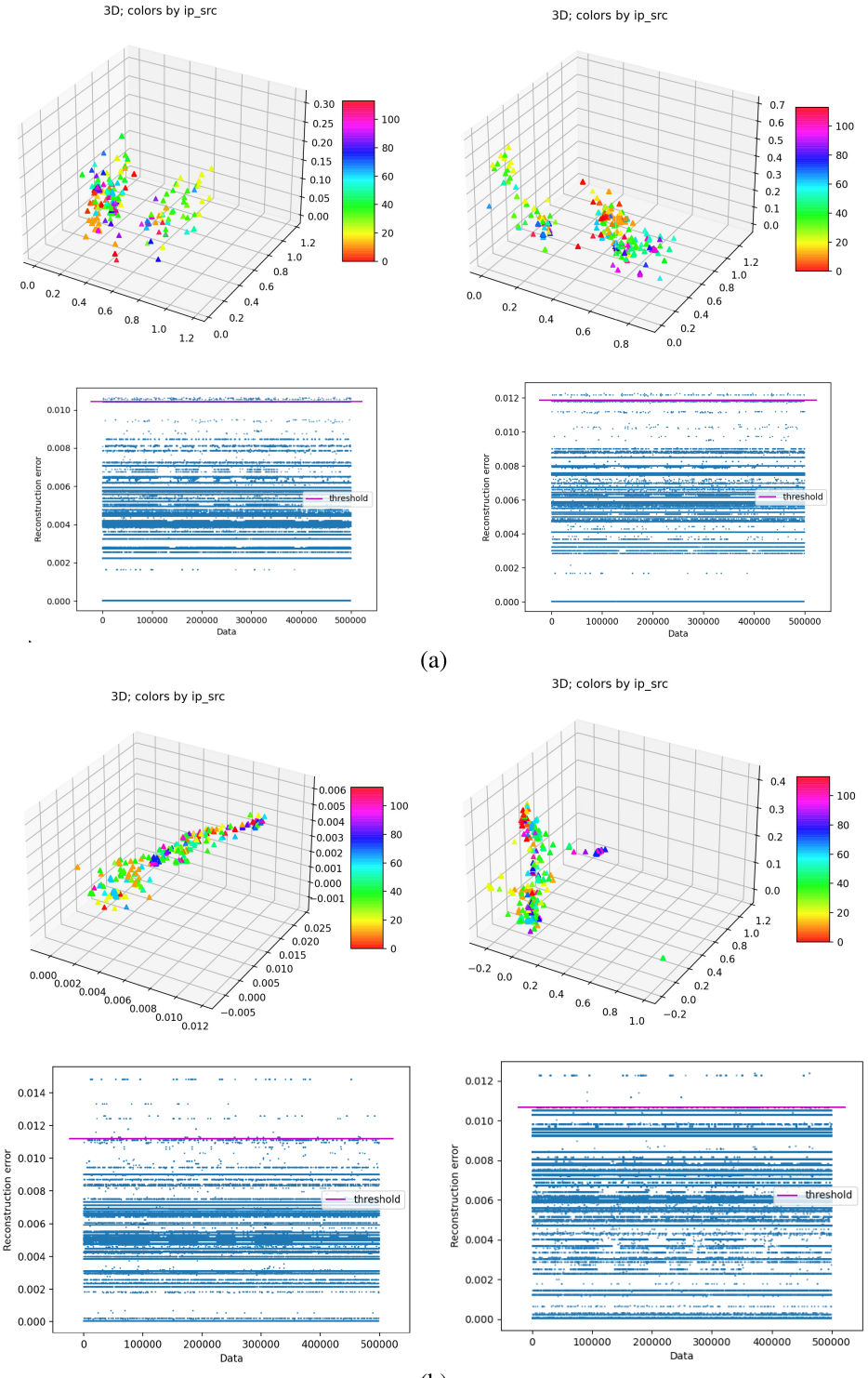


FIGURE 8. 3D-encoded data (up) and reconstruction errors (down), without (left) and with (right) the *Type* feature, in a single layer AE (a) and in the multilayer AE (b).

feature to compare and show the importance of this feature in the anomaly detection problem addressed here. Figure 8(a) also shows the reconstruction error obtained by this simple AE in both conditions. An operative threshold of 0.1% was established to discriminate the urgent alerts. According to the result summary in Table 3(a), the inclusion of *Type* during the training phase achieved two positive points: (1) the system detected as anomalies those alerts of type

TABLE 3. Type and number of alerts detected by simple AE (a) and by multilayer AE (b).

Anomalies	without *Type*	with *Type*
WRONG-TIME	344	344
CLEARTEXT-PASSWORD	2	
CONFIGURATION-CHANGE	13	13
INVALID-IP	55	44
MULTIPLE-ACCESS-DENIED	12	12
MULTIPLE-UNSUCCESSFUL-LOGINS	125	125
NETWORK-MALFORMED	4	4
SCADA-MALFORMED	137	
TCP-SYN-FLOOD	1	1

(a)

Anomalies	without *Type*	with *Type*
CLEARTEXT-PASSWORD	2	
NETWORK-SCAN		2
PASSWORD:WEAK		8
NETWORK-MALFORMED	4	
SCADA-MALFORMED	697	1176
TCP-SYN-FLOOD	62	4
WEAK-ENCRYPTION		1

(b)

MULTIPLE-UNSUCCESSFUL-LOGINS, which belong to the ones with higher potential real attacks according to experience; and (2) the system also reduced the total amount of not potentially-dangerous alerts that are detected compared with training the simple AE without using *Type*.

Another set of experiments was made with a fully-connected neural stack up to 5 layers (called from now on the *multilayer AE*). A plot of the encoded data obtained in the 3d latent space is shown in Fig. 8(b). Again, the encoded data showed different manifold structures when including the *Type* feature. Figure 8(b) also shows the reconstruction errors obtained under the same conditions as the ones used with the simple AE. After setting a heuristical threshold, the multilayer AE singularizes 765 anomalies without using *Type* and 1191 when using *Type*. Although the number of events considered as anomalies using *Type* is larger than without, we checked that none of those alerts would be considered possible attacks by the responsible human operator. More specifically, the simple AE identified about a 0.11% variables as potentially different on the test data set versus 0.24% provided by the multilayer AE. However, the sets of alerts related to this difference were negligible warnings that were not severe actually and had been filtered. We checked that there was a noticeable prevalence of high-severity events in data with high reconstruction errors. Still, there were also some low-severity rare events corresponding to unusual IP source address with high reconstruction errors, and finally, some high-severity events were not taken into account. Occasionally, alerts with high risk tagged by the IDS as anomalous packets were also discriminated by the method. Similar results were obtained using another multilayer AE with more than 5 layers.

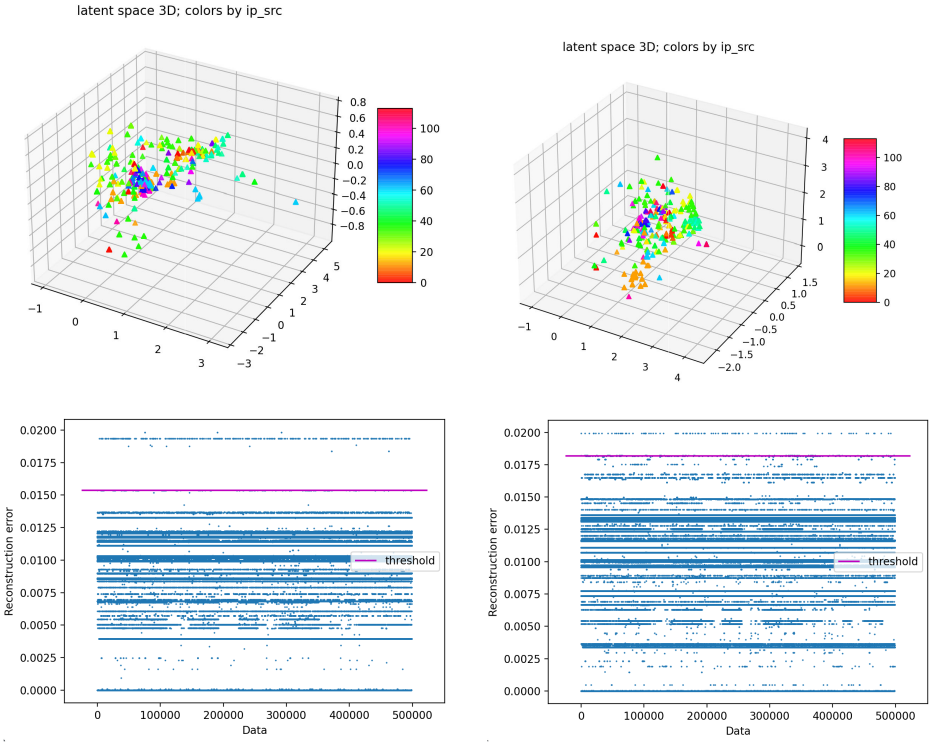
In summary, in an attempt to consolidate the results achieved and represented graphically in this work, we can say that the analysis carried out using simple and multilayer AE has allowed visualizing the latent space and the reconstruction error to study the potential incidence under a risk perspective. The analysis was performed by incorporating and isolating various combinations of variables to validate the most effective model. The best and most differential results were obtained with all variables and incorporating (right side of Fig. 8) or not (left side of Fig. 8) the event type variable. The different varieties, although again overlapping, both for simple (upper part of Fig. 8) and multilayer (lower part of Fig. 8) allow us to observe some consistency with the categorization of the IDS, still not being conclusive. The use of reconstruction error (represented below each 3D latent space portrait) as a risk indicator confirmed a much larger detection capability by simple AE versus multilayer AE (upper part of the Fig. 8).

C. VAE PERFORMANCE

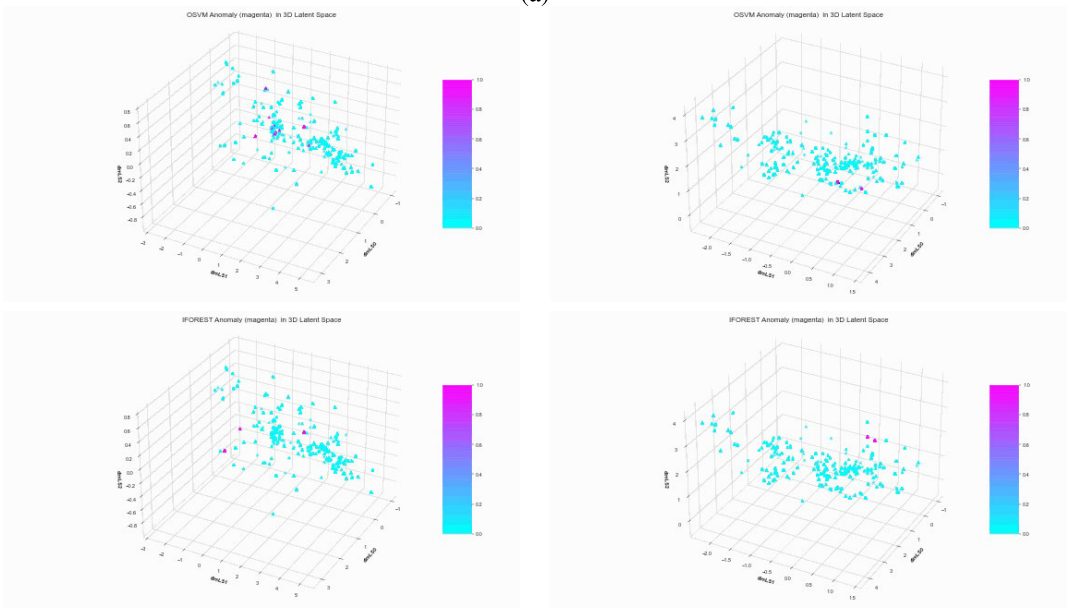
In the preceding experiments, we were able to check that low-dimensional latent spaces represented with advantage the heterogeneous set of input vectors representing alerts provided by the IDS. The embedded point clouds provided both by the simple AE and by the multilayer AE show that non-Gaussian shapes emerged in these spaces, which is an interesting-to-exploit characteristic. In addition, the classically used criterion to identify the atypical alerts using the reconstruction error exhibited noticeable dependence on the choice of the heuristic threshold. Therefore, the preceding results indicate that anomaly detection should be better addressed in the latent space rather than the reconstruction error. The use of some principled approach for identifying those data points far from the usual-traffic points appears as a necessity to be fulfilled. We decided to use VAE architectures to cope with all these requirements.

For the VAE training, the settings were set at a batch size of 256, intermediate dimension of 16 units, latent-space dimension of 3 units, and 150 epochs. The reconstruction errors obtained for both scenarios (with and without using the IDS-provided *Type* feature) are shown in Fig. 9(a). The VAE could detect 765 anomalies without *Type* and 1, 191 with it. On a dataset of 499, 013 events, the VAE detected 27, 485 events as outliers without *Type*, whereas the events considered anomalies were reduced to 1, 076 with *Type*.

Note that the approach of identifying the anomalies using the error of the reconstructed output given by an AE can have several limitations. One of them is that high reconstruction error could be due to a non-observed case, but it also could be due to noise. Another is that the threshold set on each case cannot be easily adjusted with a criterion equally useful across different AE. We set here an operative threshold allowing the network analyst to scrutinize an operative number of cases being possible candidates for anomalies. Hence



(a)



(b)

FIGURE 9. VAE experiments. (a) 3D latent space (up) and reconstruction error (down) in the VAE. (b) OSVM (up) and IFOREST (down) on the VAE-generated 3D latent space, without (left) and with (right) the *Type* feature.

we decided to do it by inspecting the residual plot in each scheme. Another criterion, such as choosing a small and fixed percentage of cases, or a fixed number of cases, would have been possible to set the thresholds, though it also would have been heuristic choices.

Again to analyze the relevance of this reduction, Table 4(a) shows the type of alert associated with each case. Without using *Type*, there is a clear-cut group of alerts of type *NETWORK-MALFORMED*, which will require further analysis in the discussion section as it might be related to reported critical alerts.

As one of the main objectives of this work is to reduce to the minimum the false alerts suggested by the IDS, the application of well-known anomaly detection methods can

TABLE 4. Results in VAE latent spaces. Types and amounts of alerts detected by the VAE (a), by OSVM (b), and by IFOREST (c). (d) Degree of alerts considered relevant by VAE and OSVM or IFOREST.

Anomalies	without *Type*	with *Type*
INVALID-IP	15	
CLEARTEXT-PASSWORD		6
NETWORK-MALFORMED	27433	1
NETWORK-SCAN:	9	1042
SCADA-MALFORMED	28	22
TCP-SYN-FLOOD		5

(a)

Anomalies	without *Type*	with *Type*
WRONG-TIME:	344	
CLEARTEXT-PASSWORD	9	
CONFIGURATION-CHANGE:	5	
MULTIPLE-UNSUCCESSFUL-LOGINS:	125	
NETWORK-SCAN:	8	1002
SCADA-MALFORMED	10	

(b)

Anomalies	without *Type*	with *Type*
WRONG-TIME:	344	
INVALID-IP:	161	879

(c)

Method	Prioritization
Only VAE without *Type*	5.5%.
Only VAE with *Type*	0.2%.
OSVM on VAE latent space with *Type*	0.2%.
IFOREST on VAE latent space with *Type*	0.18%.
OSVM on VAE latent space without *Type*	0.1%.
IFOREST on VAE latent space without *Type*	0.1%.

(d)

be a solution. Thus, we applied two well-known algorithms of anomaly detection, namely, the One-class Support Vector Machines (OSVM) [32] and Isolation Forest (IFOREST) [33], on the VAE-generated 3d latent space, instead of using the reconstruction error as a criterion.

On the one hand, Figure 9(b) shows the VAE-encoded test data, with the considered anomalies in magenta by the OSVM algorithm. The OSVM-detected anomalies correspond to the types of alerts registered by Table 4(b). Without using *Type*, we checked that intuitively, a combination of the alerts *NETWORK-SCAN*, *MULTIPLE-UNSUCCESSFUL-LOGINS,* and *CONFIGURATION-CHANGE*. This singular fact is consistent with security issues reported in the literature, as covered in the discussion section.

On the other hand, Figure 9(b) also shows the VAE-encoded test data considered anomalies (in magenta) by the IFOREST algorithm. The IFOREST anomalies corresponded to the types of alerts registered by Table 4(c). The *WRONG-TIME* anomaly means that the timestamp specified in the alert is not the same as the network time.

In general terms, we can see that the alert profiles exhibit differences that can be scrutinized with the proposed method. The results on the specific analyzed dataset show that average alerts are often grouped in probability clouds mutually adjacent in the probability space. Alternatively, it can be said that alerts with low occurrence probability are often grouped in the same class, far away from the other classes, thus pointing at their possible severity importance. It has also been observed that distant and isolated classes contain about a small percentage of alerts, distinguishing an objective group where to start the intrusion analysis. Table 4(d) shows a high-level description of comparative results from the different methods. The use of flexible anomaly detection algorithms in latent spaces generated by input spaces without using the *Type* feature tends to better prioritize the number of alerts and the content in the possible alert as evaluated by expert human managers.

As in the previous experiments and in an attempt to consolidate the results achieved and represented graphically in this work, we can say, by way of synthesis, that the exploration of possible risk analysis and its classification leveraged on the graphic representations of the results of the VAE (see the first row of Fig. 9(a)), allow us to observe how the representation of the latent space that was relevant in the case of AE, it is even more so in the case of VAE, with the groupings being observed more clearly. Again, the characterization through the reconstruction error (lower row of Fig. 9(a)) was identified as of interest for the categorization of the different types of events regardless of the variable type. Even further, the double analysis through domain descriptors such as OSVM or IFOREST represented in Fig. 9(b), allowed to reveal anomalies bundled to events identified as of great interest for the management.

VI. DISCUSSION AND CONCLUSION

Entering now into the conclusions, the focus of the present work has been to provide the network supervisor with an alert assortment through a mixed approach, including many methods. In particular, the alert aggregation has been considered the key indicator for prioritization, and its significance is provided by a three-dimensional representation built from different approaches. A particular aspect of this work is the availability of large amounts of data, forcing an adaptation of the proposed methods.

From a Machine Learning point of view, our work provides evidence that alerts generated by an IDS can be processed and subsequently projected to low-dimensional spaces of latent variables, thus generating manifolds (geometrical structures). On the one hand, the existence of said manifolds paves the way for using Machine Learning for other supervised and unsupervised data-based models, for instance, classification and regression, in cybersecurity problems based on traffic or alert data. On the other hand, we implemented an anomaly detection system based on machine learning, which allowed

us to prioritize the IDS-generated alerts. Whereas the use of conventional MCA for categorical variables did not provide a clear and operative tool, relevant findings that require additional discussion were observed. As they relate to the MCA method, we can say that as the alerts associated by the IDS with more considerable severity are condensed in a spatial and specific region of latent space but intertwined with the general structure, it is not isolatable using spatial segregation strategies. This fact and the difficulty of simultaneously working with large volumes of alerts discourage its application in this discipline. However, the confirmation of the presence of spatial structure in the MCA latent space indicates that other multiple learning methods could better support the alert prioritization and therefore opens the way to other methods that share this type of morphological analysis.

Much better results were obtained using AE structures. In this regard, we can point out that the results of the simple AE modeling provided reasonable results when it comes to prioritizing the agent's effort in their search for possible security risks. On the other hand, the multi-layer AE mainly warned about apparently unimportant but repetitive alerts, which could be essential to detect some attacks using tunnels with standard protocols such as ICMP or DNS. This multilayer model could associate the probability of misconfiguration by attaching apparent DoD alerts such as TCP SYN FLOODING with typical IP addresses. Other than that, few more significant trends were observed, such as several clusters of low severity. Still, they seemed to correspond more to isolated cases of random significance than to the presence of some structural patterns. In a comparative analysis of the autoencoders, the results obtained did not prove to be better for the case of the multilayer AE compared to the simple one. For this reason, and considering the increase in computational complexity of multilayer, the use of simple versus multilayer is suggested, the former being capable of a good detection of a high number of potential attacks with lower computational resources.

Incrementally, the results from VAE architecture offered the best results. It should be mentioned here that regarding the VAE technique, and specifically in the case where the Type variable was not included, the combination of alerts found is compatible with a privilege escalation attack pattern [34], which should be taken seriously. In this kind of event, an attacker could scan the network and find a default or inappropriate application or password, thus achieving privilege escalation to generally inaccessible assets. On the other hand, when using Type, we can see that a very simple aggregation is detected, which could point to a malicious scan. Network scanning involves detecting active hosts on a network and assigning them to their IP addresses, and it is a prior step before launching an attack on a system. In this very same analysis, the NETWORK-MALFORMED alert was found significant. According to the literature, this alert is related to detecting malformed packets that violate a protocol check during the packet inspection phase. A malformed packet sequence could indicate that some hidden process is in progress [35]. If it was the case, the communication might be established between both ends and encoded in a way that is not recognized. That should be considered attentively as a hint of a possible attack. However, with *Type*, there is a prevalence of *NETWORK-SCAN* category detected, making it possible to be a case of enumeration and reconnaissance, which could be used to gather and covertly discover as much information as possible about a target system [36]. It should be noted that reconnaissance is essential in achieving a breach in an information system.

In summary, the use of AE structures gave better results. Also, simple AE outperformed multilayer AE, and VAE surpassed both in providing latent manifolds useful for prioritization. The use of domain description techniques in the latent space yielded better results in this task than the standard reconstruction error, and more, the second one required an empirical threshold to be tuned heuristically, whereas the first one can use systematic and natural-to-tune thresholds. Other manifold learning algorithms could be used for alert prioritization problems in the future, as this is an evolving field [37], [38].

The proposed analysis is to be used by the supervisory team in alerts management if values of the significant features are reached so that the severity estimated can be contrasted with the incorporated by the IDS and relevant situations can be detected. However, it is possible that this approach still generates false positives, so specifications should be worked together with end users. The traditional experience-based judgment has a growing need to be updated. This way, efforts would be well spent in refining the data. The designed algorithms were observed to fit adequately with the intrinsic messiness present in the high-level data an IDS provides. The proposed method can be used to interpret a wide variety of information, including some points which are not directly related to attacks. These results show the analysis power of the methods when driven by large amounts of data and enhance the use of multivariate techniques [39].

The aggregation of alerts is a crucial aspect of attack prevention [40], [41], [42]. One singular incident, such as a gratuitous ARP or a single scan, could hint that an attack is in progress if it belongs to a sequence. Another possibility is tunneling detection. Attackers may tunnel network communications to and from a victim system within a separate protocol to avoid detection and enable access to otherwise unreachable systems. For example, attackers may perform SSH, HTTP, or even DNS tunneling, forwarding arbitrary data over an encrypted tunnel. Separately, this will be considered an innocuous action, but the aggregation is a severe indicator to be considered.

The objective of this work is not to show that the VAEs outperform other AEs or other machine learning (ML) methods. Our objective is to show the advantages of models based on AEs when applied as attack diagnostic aid methods for human network operators. AEs and VAEs can be an excellent

solution to reduce the substantial false positives given by the intrusion detection systems (IDS) used to monitor networks. Among the different ML models, we have chosen those based on AEs because, according to the literature, they have shown good enough results for anomaly detection applications such as the one studied here. The proposed scheme is a useful statistical tool that helps security operators improve decision-making by providing relevant information to the security process. Likewise, this work opens the door to formalizing statistical learning analysis methods based on low-dimensional latent spaces to boost the treatment of cybersecurity events.

Regarding the possible next steps, the findings regarding the ability to classify anomalies and the clustering effect of different varieties or classes in the latent space encourage us to consider that it is possible to further improve the results by using classification tools, either with clustering or with exploratory refinement techniques. Alert detection improvement may help to a more accurate revealing of potential risks, while on the other hand, the study of the characterization of all other identified categories may eventually allow us to better analyze and classify the remaining assemblies for the appropriate management.

REFERENCES

[1] M. Vielberth, F. Bohm, I. Fichtinger, and G. Pernul, "Security operations center: A systematic study and open challenges," *IEEE Access*, vol. 8, pp. 227756–227779, 2020.

[2] F. David Janos and N. Huu Phuoc Dai, "Security concerns towards security operations centers," in *Proc. IEEE 12th Int. Symp. Appl. Comput. Intell. Informat. (SACI)*, May 2018, pp. 000273–000278.

[3] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," 2018, *arXiv:1802.09089*.

[4] S. J. Lee, P. D. Yoo, A. T. Asyhari, Y. Jhi, L. Chermak, C. Y. Yeun, and K. Taha, "IMPACT: Impersonation attack detection via edge computing using deep autoencoder and feature abstraction," *IEEE Access*, vol. 8, pp. 65520–65529, 2020.

[5] T. Nguyen, S. Wang, M. Alhazmi, M. Nazemi, A. Estebsari, and P. Dehghanian, "Electric power grid resilience to cyber adversaries: State of the art," *IEEE Access*, vol. 8, pp. 87592–87608, 2020.

[6] A. Bindra, "Securing the power grid: Protecting smart grids and connected power systems from cyberattacks," *IEEE Power Electron. Mag.*, vol. 4, no. 3, pp. 20–27, Sep. 2017.

[7] J. Yen Siu and S. Kumar Panda, "A review of cyber-physical security in the generation system of the grid," in *Proc. 46th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2020, pp. 1520–1525.

[8] North American Electric Reliability Corporation, *Critical Infrastructure Protection (NERC-CIP) Standard*, Standard CIP-002 Through CIP-014, 2022. [Online]. Available: https://www.nerc.com/pa/Stand/Pages/ReliabilityStandards.aspx

[9] P. Pitre, A. Gandhi, V. Konde, R. Adhao, and V. Pachghare, "An intrusion detection system for zero-day attacks to reduce false positive rates," in *Proc. Int. Conf. Advancement Technol. (ICONAT)*, Jan. 2022, pp. 1–6.

[10] N. R. Sai, B. S. Chandana, S. P. Praveen, S. S. Kumar, and M. J. Kumar, "Improving performance of IDS by using feature selection with IG-R," in *Proc. 5th Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud) (I-SMAC)*, Nov. 2021, pp. 1–8.

[11] E. Kidmose, M. Stevanovic, S. Brandbyge, and J. M. Pedersen, "Feature-less discovery of correlated and false intrusion alerts," *IEEE Access*, vol. 8, pp. 108748–108765, 2020.

[12] R. Vaarandi, "A stream clustering algorithm for classifying network IDS alerts," in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, Jul. 2021, pp. 14–19.

[13] M. Nasir, K. Muhammad, P. Bellavista, M. Y. Lee, and M. Sajjad, "Prioritization and alert fusion in distributed IoT sensors using kademlia based distributed hash tables," *IEEE Access*, vol. 8, pp. 175194–175204, 2020.

[14] S. Ponmaniraj, R. Rashmi, and M. V. Anand, "IDS based network security architecture with TCP/IP parameters using machine learning," in *Proc. Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Sep. 2018, pp. 111–114.

[15] W. Choukri, H. Lamaazi, and N. Benamar, "RPL rank attack detection using deep learning," in *Proc. Int. Conf. Innov. Intell. Inform., Comput. Technol. (3ICT)*, Dec. 2020, pp. 1–6.

[16] K. Faber, L. Faber, and B. Sniezynski, "Autoencoder-based IDS for cloud and mobile devices," in *Proc. IEEE/ACM 21st Int. Symp. Cluster, Cloud Internet Comput. (CCGrid)*, May 2021, pp. 728–736.

[17] H. N. Bhor and M. Kalla, "An intrusion detection in Internet of Things: A systematic study," in *Proc. Int. Conf. Smart Electron. Commun. (ICOSEC)*, Sep. 2020, pp. 939–944.

[18] M. Yousefi-Azar, V. Varadharajan, L. Hamey, and U. Tupalula, "Autoencoder-based feature learning for cyber security applications," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3854–3861.

[19] C. McDermott, F. Majdani, and A. Petrovski, "BotNet detection in the Internet of Things using deep learning approaches," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[20] S. Shende and S. Thorat, "A review on deep learning method for intrusion detection in network security," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Mar. 2020, pp. 173–177.

[21] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for IoT security based on learning techniques," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2671–2701, 3rd Quart., 2019.

[22] U. Dixit, S. Bhatia, and P. Bhatia, "Utilizing ML and DL algorithms for alert classification in intrusion detection and prevention systems: A detailed review," in *Proc. 2nd Int. Conf. Advance Comput. Innov. Technol. Eng. (ICACITE)*, Apr. 2022, pp. 1199–1205.

[23] U. S. Musa, M. Chhabra, A. Ali, and M. Kaur, "Intrusion detection system using machine learning techniques: A review," in *Proc. Int. Conf. Smart Electron. Commun. (ICOSEC)*, Sep. 2020, pp. 149–155.

[24] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.

[25] F. Alonso-Atienza, J. L. Rojo-Álvarez, A. Rosado-Muñoz, J. J. Vinagre, A. García-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1956–1967, 2012.

[26] C. Soguero-Ruiz, K. Hindberg, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, K. M. Augestad, and R. Jenssen, "Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 5, pp. 1404–1415, Sep. 2016.

[27] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera, "A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines," *Inf. Fusion*, vol. 44, pp. 78–96, Nov. 2018.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[29] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[30] Q. Xu, Z. Wu, Y. Yang, and L. Zhang, "The difference learning of hidden layer between autoencoder and variational autoencoder," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 4801–4804.

[31] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2013.

[32] B. Schölkopf, J. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[33] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[34] D. M. Espinosa, D. C. Vidal, and C. B. Huidobro, "Methodological proposal for privilege escalation in windows systems," in *Telematics and Computing*, M. F. Mata-Rivera and R. Zagal-Flores, Eds. Cham, Switzerland: Springer, 2021, pp. 138–150.

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

[36] M. I. Al-Saleh, Z. A. Al-Sharif, and L. Alawneh, "Network reconnaissance investigation: A memory forensics approach," in *Proc. 10th Int. Conf. Inf. Commun. Syst. (ICICS)*, Jun. 2019, pp. 36–40.
[37] E. M. Chakir, M. Moughit, and Y. Idrissi Khamlichi, "An efficient method for evaluating alerts of intrusion detection systems," in *Proc. Int. Conf. Wireless Technol., Embedded Intell. Syst. (WITS)*, Apr. 2017, pp. 1–6.
[38] K. Alsubhi, E. Al-Shaer, and R. Boutaba, "Alert prioritization in intrusion detection systems," in *Proc. IEEE Netw. Oper. Manage. Symp. (NOMS)*, Apr. 2008, pp. 33–40.
[39] J. Camacho, P. Garcia-Teodoro, and G. Macia-Fernandez, "Traffic monitoring and diagnosis with multivariate statistical network monitoring: A case study," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2017, pp. 241–246.
[40] M. Dusi, M. Crotti, F. Gringoli, and L. Salgarelli, "Detection of encrypted tunnels across network boundaries," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 1738–1744.
[41] I. Homoliak, D. Ovsonka, K. Koranda, and P. Hanacek, "Characteristics of buffer overflow attacks tunneled in HTTP traffic," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2014, pp. 1–6.
[42] N. Ishikura, D. Kondo, V. Vassiliades, I. Iordanov, and H. Tode, "DNS tunneling detection by cache-property-aware features," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1203–1217, Jun. 2021.

JUAN RAMÓN FEIJOO-MARTÍNEZ received the degree in telecommunication engineering from the Universidade de Vigo, Spain, in 1991, and the Ph.D. degree from the Universidad Carlos III de Madrid, Madrid, Spain, in 2007. He was with France Telecom in telecommunication transmission network planning. Since 2003, he has been with the Telecommunication Division, Red Eléctrica de España, mainly in network design and maintenance. He has also been an Adjunct Professor with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid. He is currently with Universidad Rey Juan Carlos, Spain. He has participated in several research and development projects involving the reliability and resiliency of large critical infrastructures. His main research interests include network reliability improvement, estimation, and prediction methods based on statistical learning theory, support vector machines, Bayesian methods, cybersecurity, and big data analytics.

ALICIA GUERRERO-CURIESES received the degree in telecommunication engineering from the Universidad de Valladolid, Spain, in 1998, and the Ph.D. degree from the Universidad Carlos III de Madrid, Madrid, Spain, in 2003. She is currently an Associate Professor with the Department of Signal Theory and Communications, Telematics, and Computing, Universidad Rey Juan Carlos, Madrid. Her main research interests include statistical learning theory and pattern recognition and their applications to communications, image processing, bioengineering, and remote sensing.

FRANCISCO GIMENO-BLANES received the Diploma degree in advanced studies in taxes and business administration and the Ph.D. degree in communications technologies. He is currently a Professor with Miguel Hernández University. He is also a telecommunications engineer. With over 25 years of professional experience, and almost half of it in the industry, both in Spain and Abroad, he occupied a number of management positions, namely strategy planning director with a major cable operator, the head of the strategy planning with Telefonica DataCorp, the chairman's officer with Telefonica, Corporate Development, Fuertes Group, the deputy vice-chancellor, and an assistant director with the School of Engineering. On the academic side and with over 15 years of experience as a teacher and a researcher, his scientific production includes 50 contributions (papers/conferences), 33 research/technology transfer projects, two patents, one Know-How contract, and two university startups.

MARIO CASTRO-FERNÁNDEZ received the degree in telecommunication engineering from the Universidad de Vigo, Spain, in 1996. He specialized in infrastructure maintenance management and has led multidisciplinary teams distributed at a national level, ensuring the correct management of execution, budget control, planning of large projects, meeting deadlines, and estimating costs. He is currently in charge of the Telecommunication Division, Red Eléctrica de España, and conducts investments in transmission, switching, optical fiber, and IP networks. He has led his division into different network deployment types, focusing on innovation and technological development. His special work lines have been targeted for telecommunication cybersecurity improvement associated with critical services in the Spanish high-voltage electrical grid.

JOSÉ LUIS ROJO-ÁLVAREZ (Senior Member, IEEE) received the B.Sc. degree in telecommunication engineering from the University of Vigo, in 1996, and the Ph.D. degree in telecommunication engineering from the University Politécnica de Madrid, in 2000. He is currently a Professor with the Department of Signal Theory and Communications, University Rey Carlos, Spain. He has coauthored more than 150 international papers and contributed to more than 180 conference proceedings. His research interests include statistical learning methods for signal and image processing, arrhythmia mechanisms, and robust signal processing methods for cardiac repolarization.