IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE POWER & ENERGY SOCIETY SECTION

Received 24 June 2022, accepted 9 July 2022, date of publication 12 July 2022, date of current version 19 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3190398

RESEARCH ARTICLE

Data-Driven Anomaly Detection and Event Log Profiling of SCADA Alarms

J. R. ANDRADE[1], C. ROCHA[1], R. SILVA[1], J. P. VIANA[1], RICARDO J. BESSA[1], (Senior Member, IEEE), C. GOUVEIA[1], B. ALMEIDA[2], R. J. SANTOS[2], M. LOURO[2], P. M. SANTOS[2], AND A. F. RIBEIRO[2]

[1]INESC Technology and Science (INESC TEC), 4200-465 Porto, Portugal
[2]E-REDES, 1050-044 Lisbon, Portugal

Corresponding author: Ricardo J. Bessa (ricardo.j.bessa@inesctec.pt)

This work was supported by the National Funds National Funds through the Portuguese Funding Agency, FCT-Fundação para a Ciência e a Tecnologia, under Project LA/P/0063/2020.

ABSTRACT Network human operators' decision-making during grid outages requires significant attention and the ability to perceive real-time feedback from multiple information sources to minimize the number of control actions required to restore service, while maintaining the system and people safety. Data-driven event and alarm management have the potential to reduce human operator cognitive burden. However, the high complexity of events, the data semantics, and the large variety of equipment and technologies are key barriers for the application of Artificial Intelligence (AI) to raw Supervisory Control and Data Acquisition (SCADA) data. In this context, this paper proposes a methodology to convert a large volume of alarm events into data mining terminology, creating the conditions for the application of modern AI techniques to alarm data. Moreover, this work also proposes two novel data-driven applications based on SCADA data: (i) identification of anomalous behaviors regarding the performance of the protection relays of primary substations, during circuit breaker tripping alarms in High Voltage (HV) and Medium Voltage (MV) lines; (ii) unsupervised learning to cluster similar events in HV line panels, classify new event logs based on the obtained clusters and membership grade with a control parameter that helps to identify rare events. Important aspects associated with data handling and pre-processing are also covered. The results for real data from a Distribution System Operator (DSO) showed: (i) that the proposed method can detect unexpected relay pickup events, e.g., one substation with nearly 41% of the circuit breaker alarms had an 'atypical' event in their context (revealed an overlooked problem on the electrification of a protection relay); (ii) capability to automatically detect and group issues into specific clusters, e.g., SF6 low-pressure alarms and blocks with abnormal profiles caused by event time-delay problems.

INDEX TERMS SCADA, power system protection, data-driven, digital substation, alarm message, contextual knowledge.

ACRONYMS

AI	Artificial Intelligence.
CCU	Central Control Unit.
DSO	Distribution System Operator.
HV	High Voltage.
IDF	Inverse Document Frequency.
IED	Intelligent Electronic Device.
MV	Medium Voltage.
NLP	Natural Language Processing.
NTP	Network Time Protocol.
PCA	Principal Component Analysis.
RTU	Remote Terminal Unit.
SAS	Substation Automation Systems.
SCADA	Supervisory Control and Data Acquisition.
SS	Substations.
TF-IDF	Term Frequency-Inverse Document Frequency.

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

I. INTRODUCTION

The evolution of the energy utility digital ecosystem led to the generation of large volumes of data that must be analyzed

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

to extract actionable insights. Artificial Intelligence (AI) is quickly redefining how utilities manage their infrastructures and is being applied to different use cases with positive effects [1].

In the control center, human operators depend on the alarm events generated in Substations (SS) and network equipment for grid supervision, outage detection and diagnosis. However, the evolution of protection relays, the quick adoption and growing monitoring capacity of internet-of-things assets and new devices (e.g., Phasor Measurement Unit) has exponentially increased the volume of data that human operators need to analyze in short periods of time [2], [3]. As an example, in E-REDES, the Portuguese Distribution System Operator (DSO), historical data from 2020 shows that a daily average of 295,000 events were registered on the Supervisory Control and Data Acquisition (SCADA) system, considering only assets in the High Voltage (HV) and Medium Voltage (MV) grid. HOPS, the Transmission System Operator (TSO) of Croatia, reported, for one month, multiple cases with more than 100 alarms in 10-min periods [4]. This makes real-time analysis of network state very complex and time-consuming. For instance, Feng *et al.* proposed an alarm optimization strategy combining time-delay mechanism, decision trees and a hash algorithm to reduce the number of invalid alarms due to voltage/current very short-term fluctuations around upper/lower limits, auto-enclosure and repeated alarms due to permanent equipment malfunction [5].

Intelligent alarm processing in power systems is not new; the first works were mainly rule-based expert systems that filtered and prioritized alarms to provide information to human operators [6]. According to Sun *et al.*, current alarm-classification methods are mainly based on rules and, thus, the authors propose a new classification method based on information theory (using four definitions for information entropy) and on an analytic hierarchy process to classify alarm messages (using alarm message entropy as indices) [7]. Moreover, as indicated in [8], the majority of said early works were essentially fault location systems. In order to address this issue, expert systems have been widely applied at the research and industry levels [9], but more recent works focus on artificial neural networks [10], rough set theory [11], Petri Nets [12] and Bayesian networks [13], among others. Nevertheless, alarm data can be exploited for other use cases. Miao *et al.* described a logic-based methodology to identify malfunctioning relays and breakers [14], where the logic expressions are not learned from data (i.e., built with domain knowledge). Hor and Crossley proposed an unsupervised rough classification technique to select the reference Intelligent Electronic Devices for each fault type and reduce the volume of information displayed to the operator [15]. This method also showed the ability to assess the operating performance of the protection system. However, the extracted rules only covered a limited number of simulated fault scenarios. Wang *et al.* combined spiking neural P systems with rough set theory for fault equipment identification of SS, including uncertainties regarding the status information of protective devices [16]. An alternative approach, named analytic model-based methods, was proposed in [17] to measure the mismatch between the expected and actual alarm event, employing time constraint networks to capture the temporal logic among event occurrences; it was used to cluster alarms into related groups, identify abnormal or missing alarms and relate causes with alarm events. A similar principle was applied in the mixed integer linear programming (MILP) model formulated in [18], focusing on detecting malfunctioning circuit breakers or relays. Wang *et al.* formulated a MILP model for identification of false alarms sent by remote bi-directional or unidirectional fault indicators via linearized mathematical expressions [19]. Xu *et al.* proposed a supervised data-driven method that uses dynamic time warping distance to measure similarity between observed and hypothetical alarm sequence and detect faulty protection devices [20].

The aforementioned approaches presume that alarm data is available in a structured format and is fully readable by data-driven methods. However, despite the adoption of the IEC 61850 standard, a major challenge identified in [21] is to turn semi-structured/unstructured alarm information into algorithm-readable semantic, without neglecting its hierarchical and topological structure of data. In this context, recent advances in Natural Language Processing (NLP), such as Word2vec combined with Convolutional Long Short-Term Memory Networks [22], are being applied to pre-process and extract knowledge from unstructured data (e.g., fault classification). Zhang *et al.* mapped fault alarms to fault diagnosis models by using an analytical method (based on the protection configuration and protective relays setting principles) and semantic alarm data analysis with preinstalled semantic templates [23]. Another work is the data processing method from [24], based on the semantic framework of alarm information and that uses the horspool algorithm for string matching. Moreover, knowledge graphs are a promising technique to combine multi-source heterogeneous information such as alarm and metering data, social networks, operational rules, etc. [25].

The work in the present paper focuses on exploring historical SCADA data of outage events and the corresponding operators' actions, proposing two data-driven functions capable of supporting the human operator decision-making. The first function, called *Alarm2Insights*, uses data mining techniques to create a simplified topological map and identify the anomalous operation of directional and non-directional HV and MV line protections close to real-time. The second function, called *EventProfiler*, uses unsupervised learning to identify and group similar historical events (i.e., with similar messages) from control panels of HV lines (protection, switching, abnormal measurements, etc.), A description is then assigned to each group by the dispatch center operator and a classification process, based on the clusters obtained from historical data and a membership grade using the Mahalanobis-Wasserstein distance, is applied

to classify new event logs. This second tool aims to support the *post-mortem* analysis of events required to determine a probable cause to a given outage.

Compared to the state-of-the-art, the main innovative contributions from this work are:

- To the best knowledge of the authors, this is one of the first works to use raw operational data from a DSO (in contrast to [14]–[16]) and propose techniques capable of converting a large volume of alarm events into data mining terminology (e.g., vector representation – *embedding*) and a compact dataset with useful (or 'smart') data. The most similar work is [24] that proposed an alarm data processing technique to work with natural language data like in [22] and specifically designed for one use case (i.e., fault diagnosis and trace). In contrast to [23], the present work does not require the definition of semantic templates or analytical rules to extract knowledge from alarm data.
- Unsupervised data-driven methodology for protection system malfunction assessment. The developed method follows the same principle of analytic methods like [17], [18], but it is fully data-driven and does not require solving an optimization problem or construction of mathematical expressions that mimic protection system expected operation like in [18], [19]. Moreover, it does not require a hypothetical time series of alarms, like in the supervised learning approach from [20].
- Data-driven methodology (not based in optimization problems like in [17]) to segment and categorize historical event log data, helping human operators to identify different groups of occurrences in HV lines automatically, and exclusively based on their event profiles (i.e., without using a pre-existing class in a supervised learning fashion, e.g., [5]). The present work can also leverage from the information-based approach proposed in [7] to evaluate the information value of alarm messages and alarm pre-processing.

The remainder of this paper is organized as follows: Section II describes the data analytics framework, including the main aspects and challenges of the raw dataset and an overview of the data pre-processing pipeline. Section III describes the algorithm for protection functions normality modeling and section IV includes the alarm segmentation approach and detection of rare events. Section V presents results for real data from a DSO. The conclusions and topics for future works are discussed in section VI.

II. DATA ANALYTICS FRAMEWORK

A. DATA DESCRIPTION

Figure 1 presents an overview of E-REDES standard Substation Automation System (SAS) architecture composed of four different layers, namely: remote management, local management, bay level and process level. This system provides protection, control, automation, monitoring and communication functionalities that are critical for a successful

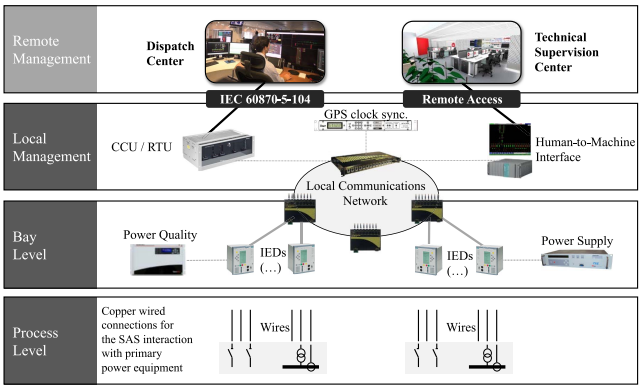


FIGURE 1. Standard substation automation system architecture of E-REDES (adapted from [26]).

supervision and operation of the distribution grid. A detailed description of the SAS architecture can be found in [26].

The dataset used in this work is composed of 8,631,091 historical events, retrieved from the DSO's dispatch center central SCADA database (see Figure 1 'Remote Management' group). The data spans from January 1st, 2014 to June 30th, 2020, containing events from 22 primary SS geographically spread across mainland Portugal. Two types of events can be found in an outage event log, namely:

- *Remote Terminal Unit (RTU) event log*: Equipment state changes, measurements or alarm signals occurring at any given physical equipment of a SS (e.g., power transformers, capacitor banks, feeders and busbars); some examples of alarm signals, on which we focused in particular, include the opening or closure of a circuit breaker, the triggering of a maximum current or voltage protection relay or a drop in the pressure of an SF6 gas-insulated transformer. However, most of the historical events are merely informative, such as the switching of the illumination at the SS or a measurement of the current intensity at a specific phase of a transformer.
- *Command event log*: Control actions that dispatch center operators may take during an event. For each action, a pair of events is always generated, one indicating the beginning of the action and another with the status of its conclusion, naturally with some time delay between them. Some examples include an attempt on manually closing a circuit breaker or (de)activating a specific panel's protection.

The following structure (illustrated in Table 1) is available for each event:

- *Timestamp (RTU)*: Timestamp of events from Intelligent Electronic Device (IED) and other SS equipment at the SS local management layer, namely RTU or Central Control Unit (CCU), depending on the SAS architecture;
- *Timestamp (SCADA)*: Timestamp of SS events and operator commands' events in the SCADA database;

TABLE 1. Simplified example of an event log from the SCADA database.

Timestamp (RTU)	Timestamp (SCADA)	Event description	Event tag	Operator ID
2021-01-01 00:01:16.430	2021-01-01 00:01.16	Substation XX - Panel YY - Alarm 1 Description - State	XX-YY-EV1	-
2021-01-01 00:01:16.521	2021-01-01 00:01.16	Substation XX - Panel YY - Alarm 2 Description - State	XX-YY-EV2	-
-	2021-01-01 00:01.18	Substation XX - Panel YY - Operator Command - Initiated	XX-YY-CMD1	E******
-	2021-01-01 00:01.18	Substation XX - Panel YY - Operator Command - Concluded	XX-YY-CMD1	E******

- *Event description*: Summary of the signal (alarm or command) that triggered the event creation. Identifies the SS, equipment panel, event's description and respective equipment state;
- *Event tag*: Structured key representation of the event's description;
- *Operator ID*: This field indicates an operator ID for events originated at SCADA level, since these are triggered by an operator manual command; it is used to distinguish between events generated at SS level and at dispatch center SCADA level.

Note that the *Event description* and respective *Event tag* are normalized between similar substations and panels, which means that they are not restricted to a single event. For example, the *Event tag* indicating the status of a circuit breaker (open or closed) is always like "XX-YY-DJEST", where "XX" indicates the Substation and "YY" the panel. The total number of unique *Event description* found in the historical raw dataset, considering them to be SS and panel agnostic, is 937. The normalized event descriptions and states are available in the functional specifications of E-REDES' protection system [27].

B. DATA PRE-PROCESSING

The historical raw dataset is composed of many heterogeneous records that, despite sharing the basic structure of Table 1, presented a series of challenges to their correct interpretation and use:

- Time delays due to synchronization problems between SS equipment RTU and SCADA system prevent the clear establishment of a sequential data reordering process. Different reasons lead to this problem. The considered SS are at different stages of modernization, which does not guarantee the synchronization of all events, which is the case in modern Distribution Automation Systems (e.g., IEC 61850 based) where a master clock is used to synchronize all SS devices. Time delays greater than 2 seconds were detected for nearly 10% of the total number of records, spanning from a few seconds to years. Differences of 1-2 seconds and smaller are considered normal and inherent to communication times, but larger differences arise exclusively at RTU/CCU level, signaling the malfunctioning of either the internal GPS clock of the SS or the communication hardware (i.e., failures in Network Time Protocol (NTP) syncing or delays between the SS and the dispatch center time references). For instance, some events were registered in dispatch center SCADA database days after they happened; in other cases, the SS internal clock was stuck at the Unix epoch time, i.e., 1970-01-01 00:00:00, or other events were placed in 2034. This is critical since it affects the establishment of a sequential data reordering between events from SS equipment and dispatch center operators' remote commands. As will be explained below, this problem was circumvented through the establishment of a pre-processing sorting logic that made these events usable.
- The event description and respective tag's nomenclature for equivalent records (i.e., the same event occurring at the same panel of the equivalent SS) change across different equipment vendors and age. This has a direct impact when performing an intra-substation analysis since recent event messages might have small differences compared to older ones.
- Establishing an unequivocal connection between the SCADA/RTU records and complementary information in metadata files was not always possible given the lack of an update to the latter, which would reflect the different equipment's technology and newly installed assets.

Considering the aforementioned challenges, a data processing pipeline was developed (see Fig. 2). First, the data was sorted to represent, as accurately as possible, the real temporal sequence of events generated during an outage. This was accomplished through a combination of RTU and SCADA timestamps with an incremental integer field representing each event entry order in the SCADA system database.

Then, a rule-based algorithm was applied to a) identify changes in event descriptions and tags (i.e., due to technology vendor updates) and b) create a standard nomenclature for event descriptions and tags with similar meaning. These rules were defined based on a sensitivity analysis performed together with E-REDES and were only applied to the event log from HV lines' panels. Keep in mind that these procedures do not remove major differences among technologies that could encrypt certain behavior of SS equipment, since it only normalizes the nomenclature of equivalent event messages. Finally, the equipment's state for each description is normalized, indicating its normal or abnormal state (e.g., for a circuit breaker that is normally closed, a registry

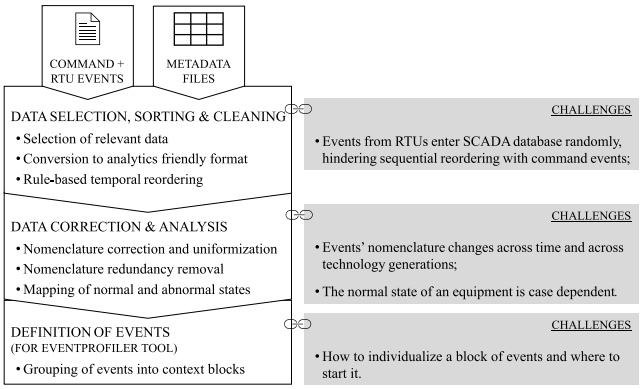


FIGURE 2. Data pre-processing pipeline.

indicating an 'open' state would be classified as abnormal). The state considered as normal was determined for each SS equipment, by assessing the most frequent state in all historical events related to that specific equipment. This approach was validated by the DSO, particularly for circuit breakers, since network reconfigurations took place at MV level, where the normal state of some of the equipment is switched.

An extra step is also performed for the *EventProfiler* function, which is explained in section IV.

C. MODEL CHAIN

Fig. 3 depicts a simplified diagram of the main steps from data-preparation to functions' execution. The functions described in sections III and IV aim to extract meaningful insights from the pre-processed data (section II), through a combination of data mining and AI techniques, thus helping dispatch center operators to rapidly identify:

- Anomalous behaviors regarding the performance of protection relays functions associated with HV and MV lines (i.e., abnormal or missing protection functions' pickups at any given bay).
- Similar events (i.e., with similar log messages) in large historical datasets to classify new outages' events into previously defined clusters and detect unique or rare events.

III. ALARM2INSIGHTS: DETECTION OF ABNORMAL BEHAVIORS IN SS PROTECTION RELAYS FUNCTIONS

The data-driven approach is illustrated in Fig. 4 and divided in three phases, described below.

A. PHASE 1: INPUT CREATION

An input episode is created every time a circuit breaker opening alarm is detected in either an HV or MV line panel. Each episode is composed of the circuit breaker alarm event description and the sequence of events created in the time (minutes) preceding that alarm (see Fig. 4 'Input Episode' group).

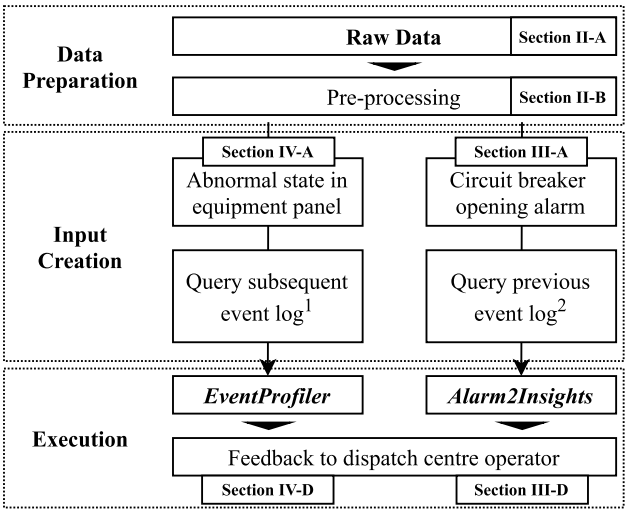


FIGURE 3. Model chain for the data-driven analysis of SCADA alarms.

B. PHASE 2: KNOWLEDGE DISCOVERY

In this phase, two statistically-based insights are extracted from the historical input episodes: 1) Simplified topological map for SS equipment; 2) Identification of typical protection relays functions behavior during circuit breaker opening alarms in MV line panels.

1) SIMPLIFIED TOPOLOGICAL MAP

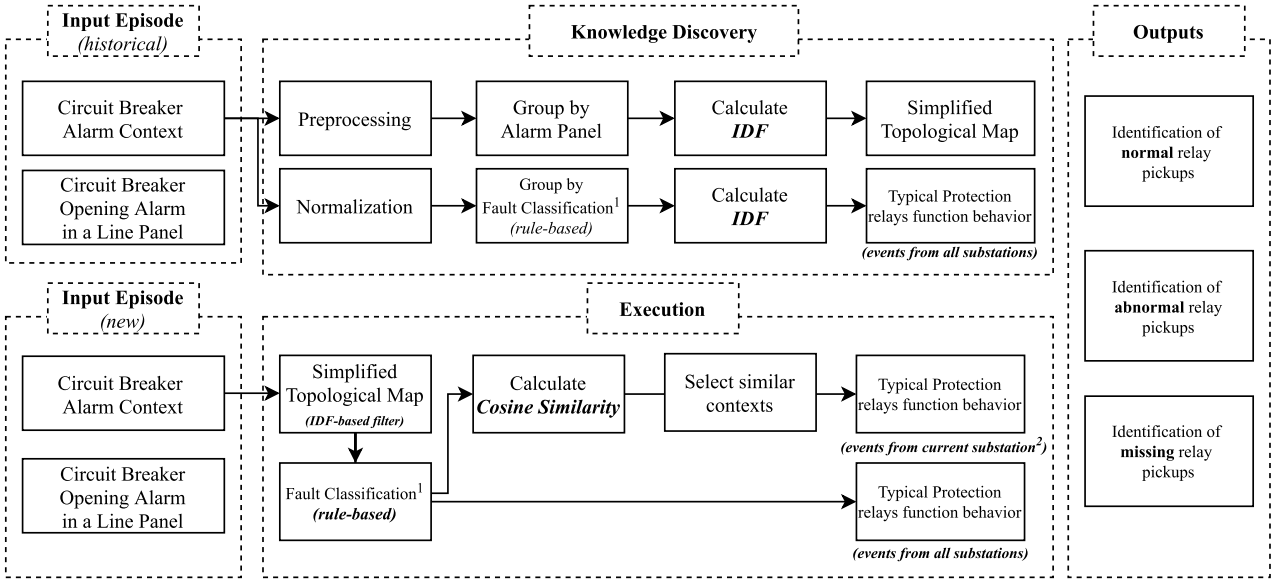
First, each input episode is represented by the set of SS panels referenced at least once in the original context. This initial pre-processing step removes information regarding the event description in each panel (see Table 1), but retains information about which panels logged in the alarm context.

Then, the new sequences of events are grouped according to the circuit breaker alarm panel (see 'Group by Alarm Panel' reference in Fig. 4), and the Inverse Document Frequency (Inverse Document Frequency (IDF)) [28], used in NLP, is calculated for each group:

$$\text{IDF}(tag)_j = \log \frac{N_j}{n_{tag,j}} \tag{1}$$

For each circuit breaker opening alarm j, N_j represents the total number of historical input episodes for that alarm and $n_{tag,j}$ the number of episodes containing a specific *tag* (i.e., panel identifier). For each panel identifier, an IDF reference value is calculated.

This IDF-based approach shows which SS panels frequently (i.e., with lower IDF values) log in the context of each circuit breaker alarm. This statistical information comprehends topological relationships between SS equipment panels during alarm events; thus, in this case, it is considered as a simplified topological map representation. Based on the IDF values, we then create an operational filter for contexts from past and new alarm episodes, which keeps the event log for



1 - Process only applied for circuit breaker opening alarms in MV lines panels
2 - Considered context data from the current substation and other relevant substations (as defined in the simplified topological map)

FIGURE 4. Processes flow for the *Alarm2Insights* function.

SS panels that frequently log in the historical contexts of such circuit breaker alarm. Note that information-theoretic metrics to represent joint informational content, such as mutual information [29], can also be used to construct/reconstruct electrical network topology from SCADA measurements.

2) TYPICAL PROTECTION RELAYS FUNCTIONS BEHAVIOR

First, the historical data is normalized by replacing the SS panel identifier with a general equipment type identifier (e.g., 'substation A panel B - event description 1' is replaced by 'MV line panel - event description 1') and only events regarding protection relays functions are kept. This normalization step is applied to the historical input episodes from all SS.

Then, the IDF is calculated over the normalized sequences to display the typical behavior of protection relay functions' pickups. A 'common' classification is assigned to frequent protection functions' pickup events (i.e., with IDF values lower than 0.5), while 'uncommon' or 'rare' classifications are assigned to protection functions' pickup events with higher IDF values (i.e., between 0.5 and 1.2, and higher than 1.2 respectively). The IDF thresholds were defined based on a sensitivity analysis, using the methodology presented in Fig. 5 and can be quickly adjusted by the operators to increase or decrease the sensitivity of the tool to rare events. This functionality is especially relevant as some installation or panel may show specific behaviors that are frequent (and apparently common) in that specific installation (e.g., caused by an old and persistent problem), but rare when compared with the predominant (and normal) behavior of the remaining distribution grid SS.

C. PHASE 3: EXECUTION

Each new input episode is initially pre-processed using the simplified topological map IDF-based filter created in the 'Knowledge Discovery' phase, and the events for panels with higher IDF values are discarded. Then, a Term Frequency-Inverse Document Frequency (TF-IDF) approach is used to obtain a vector representation of both historical and new event sequences for the current circuit breaker alarm. Subsequently, the cosine similarity metric is used to calculate the similarity between the current context vector representation and the remaining historical contexts. The subset of most similar contexts (i.e., cosine similarity superior to 0.7) is used as reference to profile the typical relay pickups behavior - achieved by calculating the IDF value for each event tag in the context.

D. OUTPUTS

For each circuit breaker alarm, the following feedback is provided to the control center operator:

1) Abnormal relay pickup events detected by comparing current events with the typical behavior of historical contexts for that fault-type. This output is only available for MV lines.
2) Typical relay pickups events (i.e., on a global and SS level) that do not appear in the context of HV and MV lines' alarms.

It is important to underline that, for the first output, two types of typical relay protection function behaviors are considered, retrieved from historical episodes of:

- Circuit breaker opening alarms registered in all the SS.

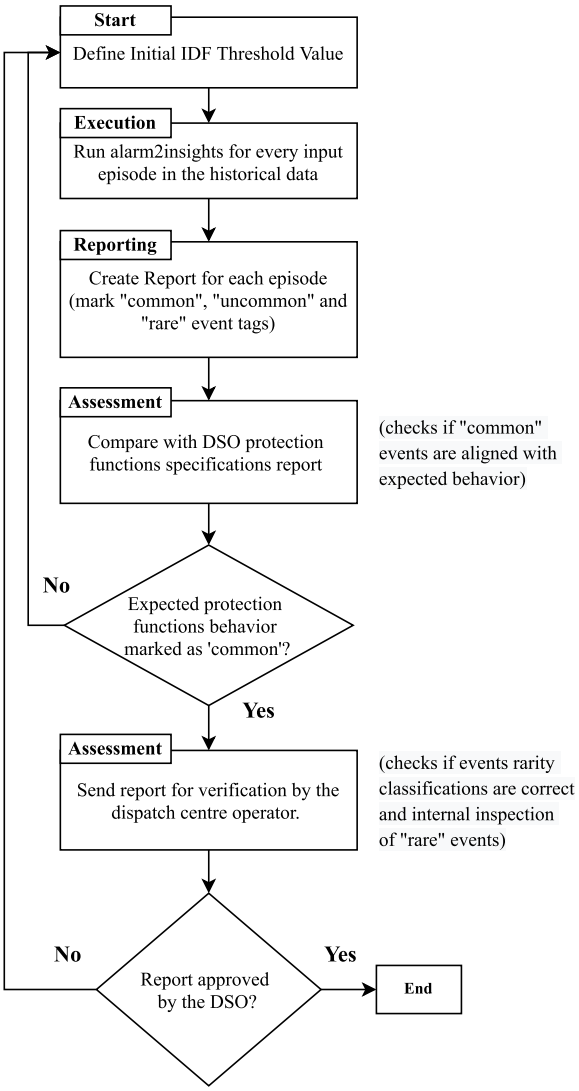


FIGURE 5. Methodology used for the definition of IDF thresholds for alarm2insights. Note: E-REDES protection functions functional specifications are available in [27].

- The current circuit breaker opening alarm (i.e., in the same SS).

The combination of these two types of behaviors enables the detection of overlooked problems in specific SS (i.e., by comparing with the global typical behaviors) or even the identification of SS with abnormal dynamics in protection pickup events, as presented in Section V. For MV line panels, a rule-based classification was used to aggregate historical data per fault-type, which contributed greatly to improve the definition of typical relay pickups behavior during each fault.

IV. EVENTPROFILER: EVENT LOG PROFILING AND CLASSIFICATION

This data-driven function explores historical events which occurred in HV lines and is composed of four separate phases, described below.

A. PHASE 1: DEFINITION OF BLOCKS

As mentioned in section II-B, this function requires an additional data pre-processing step to identify context blocks (or simply blocks). A block is a set of events that represent one occurrence or 'problem' in the SS. Hence, each block begins once an abnormal state is detected on the event log (e.g., relay pickup event) and ends once every equipment returns to its normal state, or if there is an action by a dispatch center operator. There are two types of blocks: 1) blocks without any operator intervention (e.g., faults solved through automatic re-closing or other grid automatic system) and 2) blocks where an operator intervention is required. In both cases, the number of events per block range between a minimum of 3 and a maximum close to 7000. In this phase, all information regarding the SS and the panel where the event occurred is removed from the event tags, due to its irrelevance for both processes - clustering and classification.

B. PHASE 2: WORD EMBEDDING MODEL

At this point, each block is a set of event tags, like text, that we need to convert into a numeric representation to apply a clustering method. Since the goal is not only to preserve the tags' information but also their relative position in the blocks, this phase resorts to a word embedding model to represent the tags. In this context, similarly to NLP tasks, each block is seen as a text and each tag as a word. This means that our dataset is a collection of texts and all the historical records are used to fit a word embedding model. In this work, we use the Word2vec algorithm, proposed by [30] with the continuous Skip-Gram model, in order to obtain the word embedding model. This model is then used to represent each record into a new dense representation with predefined and fixed dimension, in which similar records are characterized by similar encoding values.

C. PHASE 3: CLUSTERING

At this stage, the main goal is to identify similar groups of context blocks. One ought to select a clustering method, or methods, and a distance to be used as similarity measure.

1) SIMILARITY MEASURE

Each block is represented by a numeric matrix with a fixed number of correlated attributes (i.e., defined by the word embedding model) but a variable number of lines (equal to the number of events in the block). This means that, for each block, we have several values for each one of the correlated variables. Therefore, we use the Mahalanobis-Wasserstein distance [31] since it is suitable to work with correlated variables and is based on the variables' distributions.

Considering the unidimensional case, we have the Wasserstein distance between two random variables f and g, with F and G distribution functions. This metric is considered an extension of the Euclidean distance between quantile functions and is defined in [31] as:

$$d_W(F, G) := \left(\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt \right)^{1/2} \quad (2)$$

where F^{-1} and G^{-1} are the quantile functions of the two distributions. In an analogous way, the Mahalanobis-Wasserstein distance is a generalization of the Mahalanobis distance for p correlated variables and is defined as:

$$d^2_{MW}(\mathbf{F_i}, \mathbf{F_{i'}}) = \sum_{h=1}^{p}\sum_{k=1}^{p}\int_0^1 s^{-1}_{hk}\left(F^{-1}_{ih} - F^{-1}_{i'h}\right)\left(F^{-1}_{ik} - F^{-1}_{i'k}\right)dt, \quad (3)$$

where F^{-1}_{ih} is the quantile function of the distribution F for variable h and of the i^{th} block; and $\left[s^{-1}_{hk}\right]_{p\times p}$ is the inverse of the codeviance matrix where each element is the codeviance of a dataset described by the two distribution variables (h and k). This matrix, the codeviance of a dataset, $CODEV_F$, described by two distribution variables is given by:

$$CODEV_F(X_j, X_{j'}) = \sum_{i=1}^{n}[\alpha_i - \beta_i - \gamma_i] + n\delta + \sum_{i=1}^{n}\mu_{ij}\mu_{ij'} - n\mu_j\mu_{j'} \quad (4)$$

with

- $\alpha_i = \rho_{QQ}(F^{-1}_{ij}, F^{-1}_{ij'})\sigma_{ij}\sigma_{ij'}$;
- $\beta_i = \rho_{QQ}(F^{-1}_{ij'}, \bar{F}^{-1}_{j})\sigma_{j}\sigma_{ij'}$;
- μ_{ij} is the first moment of F_{ij} (the same for $\mu_{ij'}$);
- $\rho_{QQ}(F^{-1}_{ij}, F^{-1}_{ij'})$ is the QQ correlation between the j^{th} distribution and the $j^{'th}$ distribution of the i^{th} individual;
- $\gamma_i = \rho_{QQ}(F^{-1}_{ij}, \bar{F}^{-1}_{j'})\sigma_{ij}\sigma_{j'}$;
- $\rho_{QQ}(F^{-1}_{ij'}, \bar{F}^{-1}_{j})$ is the QQ correlation between the barycenter distribution of the j^{th} variable and the $j^{'th}$ distribution of the i^{th} individual;
- $\delta = \rho_{QQ}(\bar{F}^{-1}_{j}, \bar{F}^{-1}_{j'})\sigma_{j}\sigma_{j'}$;
- $\rho_{QQ}(F^{-1}_{ij}, \bar{F}^{-1}_{j'})$ is the QQ correlation between the barycenter distribution of the $j^{'th}$ variable and the j^{th} distribution of the i^{th} individual;
- σ_{ij} is the standard deviation of F_{ij} (the same for $\sigma_{ij'}$);
- σ_j is the standard deviation of $\bar{F}_j$ (the same for $\sigma_{j'}$); and
- $\rho_{QQ}(\bar{F}^{-1}_{j}, \bar{F}^{-1}_{j'})$ is the QQ correlation between the barycenter j^{th} distribution and the barycenter $j^{'th}$ distribution.

More information about this distance can be found in [31].

The Mahalanobis-Wasserstein distance was initially proposed to measure distances between histograms. However, as mentioned in [32], besides the fact that the Wasserstein distance can be computed to compare any two empirical distributions (discrete version), f and g, with n quantiles, and is defined in [33] as:

$$d_M(f, g) = \left(\frac{1}{n}\sum_{i=1}^{n}\left|F^{-1}_{(i)} - G^{-1}_{(i)}\right|^2\right)^{1/2}, \quad (5)$$

the Mahalanobis-Wasserstein distance can also be easily adapted to multivariate empirical distributions, which is the case of the data in this work.

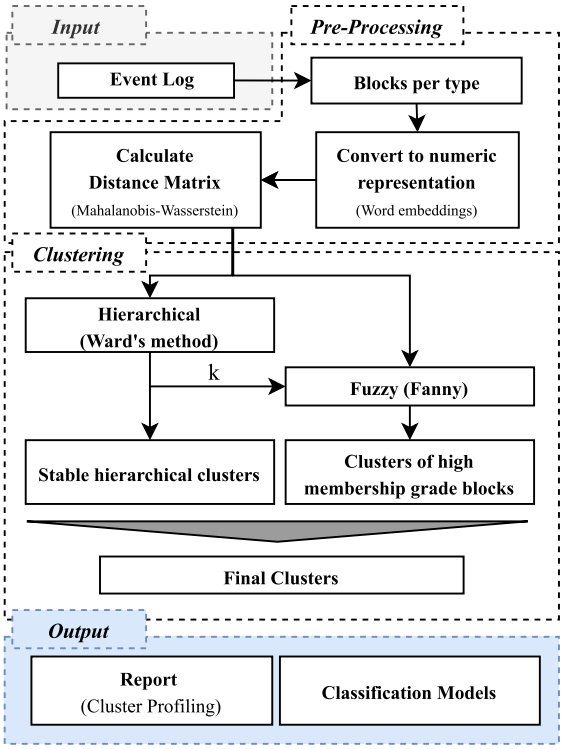


FIGURE 6. Process diagram for EventProfiler clustering phase.

2) CLUSTERING METHODS

Knowing that the events that characterize a specific problem may or may not be unique to said problem raises questions about the method used in the clustering process. This means that, if on the one hand, we can have some characteristics that overlap between groups, which require soft clustering methods, on the other hand, we can have some distinct characteristics for other groups, which require hard clustering methods. Based on this, we decided to combine the two approaches in the clustering process, as illustrated in Fig. 6, and two clustering algorithms were used: a) hard clustering type, i.e., hierarchical clustering with Ward's agglomerative method [34]; b) soft clustering type, i.e., fuzzy clustering fanny algorithm [35].

To assess the stability of the results from the hierarchical clustering, we applied the bootstrap hierarchical clustering, using a different number of clusters and bootstrapping 1000 samples for each run. This method is carried out first, and only the clusters with a high Jaccard similarity (greater than 0.95 for blocks of type 1 (without operator action); 0.83 for type 2 (with operator intervention) are considered for inclusion in the final clusters. Since we are interested in highly stable clusters and only a Jaccard similarity value smaller or equal to 0.5 is referred in the literature as an indication of a "dissolved cluster" [36], these two values were empirically determined, based on three random samples of size 100 extracted from the set of blocks described in Section IV-A, and with the support of a domain expert (human operator). Initially we compared the clusters obtained for the

three samples using as threshold the values of 0.8, 0.85 and 0.95. The value that led, in the three samples, to the highest number of clusters with a clear profile, and the lowest number of clusters with an ambiguous profile (from the domain expert's point of view) was then used as center around which we added and subtracted multiples of 0.01 to find the most suitable threshold for each case. In the end, we obtained the threshold of 0.95 for blocks of type 1 and 0.83 for blocks of type 2. The resulting number of clusters plus one is used as the k parameter's value (number of clusters) in the fanny algorithm (see Fig. 6).

The results from both algorithms are combined in a way that only the most stable clusters are preserved. More specifically, sets of context blocks that are grouped at least 95% (or 83%, for type 2 blocks) of the times in the hierarchical clustering or have a membership grade over at least 95% (75% for type 2) in the Fuzzy clustering. These blocks are the core of each cluster, and the clusters reflect the typical patterns in the data.

The main output of this phase is a report with 1) the number of blocks that form the cluster and 2) the IDF for each unique event that appears in the cluster.

D. PHASE 4: CLASSIFICATION

The classification process is based on the clusters obtained in the previous process (classification models). First, the block is converted into a numeric matrix (using the same word embedding model) and then, based on the Mahalanobis-Wasserstein distance in Eq. (3), the membership grade in Eq. (7), u_j, is computed for each cluster j through

$$tu_j = \frac{1}{\sum_{c=1}^{k} \left(\frac{dist(j)}{dist(c)}\right)^{\left(\frac{2}{r-1}\right)}} \tag{6}$$

with a normalization step

$$u_j = \frac{tu_j}{\sum_{c=1}^{k} tu_c}, \tag{7}$$

where $dist(c)$ is the minimum distance to cluster c, k is the number of clusters, and r is the membership exponent ($r = 1.2$). It is worth mentioning that the membership grades do not allow to identify blocks that are outliers, i.e., blocks so different that they end up far from all the clusters. Therefore, in addition to the membership grades, we also computed a normalized control parameter, *control*, that quantifies the relative overall distance to all the clusters:

$$control = \sum_{c=1}^{k} \frac{dist(c)}{10d(1,2)}. \tag{8}$$

Due to the presence of data codeviance in the calculation of the distance between two blocks, Eq. (3), it is necessary to standardize the distances, in order to compare them with a threshold value. In fact, when a block is added to the set of clusters' blocks, the distances between these blocks are affected, but the relationship between them remains the same. Therefore, one can use the distance between two clusters' fixed blocks as a reference and obtain normalized distances, i.e., comparable with the threshold. Hence, in Eq. (8), $d(1, 2)$ is the distance between the first pair of blocks belonging to the clusters and the scalar 10 is used to reduce the magnitude and give the control parameter an interpretation similar to an average value.

Considering that the control parameter enables distinguishing new blocks with a profile similar to the ones represented in the clusters from those who are very different, we may interpret the set of membership grades as the fraction of similarity between the new block and the clusters, i.e., the set of membership grades of a new block represents the distribution of similarities to all clusters. For this reason, when one of the elements of the membership grades vector is at least 0.5, it means that the new block has at least 50% of its similarity associated with that cluster and the remaining similarity is distributed by the other clusters. So, when that happens, we may assume that the block has the same profile of that cluster.

The output of this process is consolidated in a report which provides the following information:

- If the control parameter is higher than the clusters' blocks' mean value plus three standard deviations, a warning is issued in the report.
- If one of the membership grades is higher than 0.5, then the block is classified as belonging to the corresponding cluster.
- If the membership grades are all less than 0.5, then the report presents the two clusters with highest membership grade, but does not classify the block.
- If the block lacks records that are common to all blocks of the cluster it has been assigned to, an alert message is sent to the operator.
- If any record in the block is new to the word embedding model, a warning is issued in the report.

V. RESULTS AND DISCUSSION

This section presents the numerical results obtained with the *Alarm2Insights* and *EventProfiler* functions for real data from E-REDES (described in section II).

A. ALARM2INSIGHTS RESULTS

The results of this function are obtained with a leave-one-out cross-validation approach over the entire historical data (i.e., all the substations event log data). Each execution (see Section III-C) is applied over an input episode extracted from the pool of historical episodes being the 'Knowledge Discovery' phase (see Section III-B) ran over the pool of historical episodes minus the episode being analyzed.

1) IDENTIFICATION OF RARE RELAY PICKUP EVENTS

An initial inspection of the historical data revealed a small subset of SS with distinct event log patterns during faults in MV lines. This occurs due to specific event tag terminologies

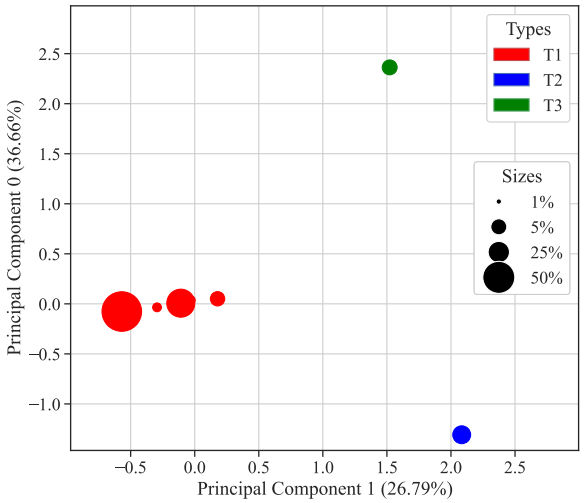


FIGURE 7. PCA-based visualization for phase-to-phase outage events in MV lines.

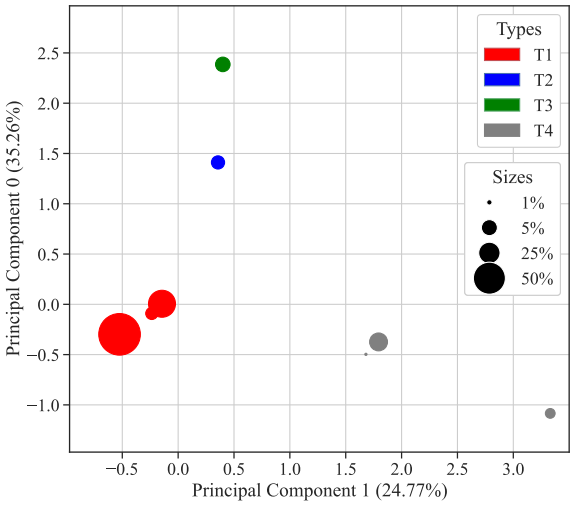


FIGURE 8. PCA-based visualization for phase-to-ground outage events in MV lines.

originated from differences in protective relays equipment vendors and technology. Since this data-driven function relies on historical data from all SS, it is important to analyze these differences first, and assess their potential impact on the final output. Fig. 7 and Fig. 8 capture and demonstrate the terminology differences that exist in circuit breaker alarm contexts for phase-to-phase and phase-to-ground faults in MV lines respectively, as each fault type also has specific protective relay pickup events (originating different sequences of events). This analysis was performed by applying the following steps over historical event log data from all SS:

1) Select contexts for circuit breaker alarms in MV lines;
2) Calculate the IDF over the selected contexts in order to obtain a rarity value for each event;
3) Apply k-means clustering over the IDF matrix to create event terminology clusters. Each cluster represents a different terminology. The number of clusters was empirically defined according to the number of terminologies identified in the historical data;
4) Use Principal Component Analysis (Principal Component Analysis (PCA)) for dimensionality reduction and plot the first 2 components (i.e., with highest variance). Each point has a color attribute, representing a terminology type, and a size attribute, representing a percentage of the total number of outage events where that terminology is used.

As depicted by the figures, there are up to four terminology types, namely T1, T2, T3 and T4. The predominant type (i.e., T1) is used in 13 out of 22 SS, containing more than 75% of the total number of occurrences in the historical data for these two fault types. This was an important finding due to the data-driven nature of the *Alarm2Insights* function. Events from the minor terminology types (i.e., T2, T3 and T4) deviate from the typical system behavior (mainly driven by terminology T1) and might be misclassified as

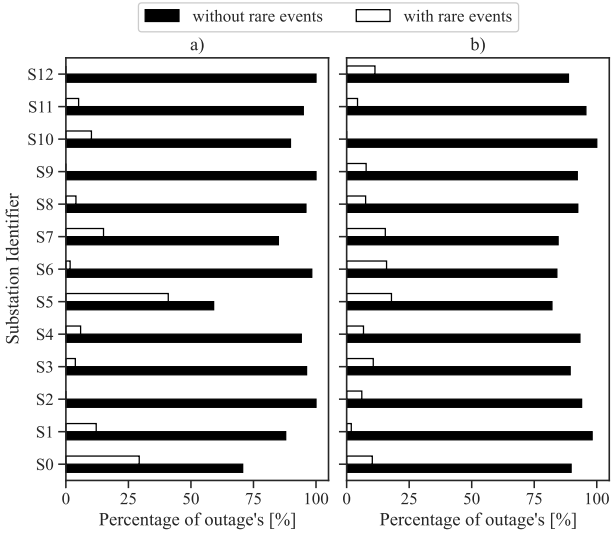


FIGURE 9. Comparison between percentage of outages with and without 'rare' events in the event log of a) phase-to-ground faults and b) phase-to-phase faults.

'rare' by the function. In this analysis, the focus is on SS that use terminology T1, providing specific examples of real anomalies, as well as potential false alarms due to terminology changes.

Fig. 9 depicts the results obtained by applying the identification of rare relay pickup events over the historical data from the SS with 'T1' terminology. Both figures compare the percentage of circuit breaker alarm contexts per SS with: a) at least one event classified as 'rare' (see 'with rare events' label in the figures); b) without any event classified as 'rare' (see 'without rare events' label), representing the expected normal behavior.

An analysis of Fig. 9 reveals that, for phase-to-ground faults (see left plot), three SS (i.e., S0, S5, S7) have more than 15% alarm occurrences with at least one rare event in its context. The following conclusions were drawn:

Description	Tag / State
MV LINE VERMOIL - MAX I>> INST PICKUP	3301-PRI2I pickup
MV LINE VERMOIL - MAX I>> TEMP TRIP	3301-PRI2T trip
MV LINE VERMOIL - RECLOSURE CYCLE ACTIVE	3301-RLCIC active
MV POWER TRANSFORMER 1 - MIN U< INST PICKUP	3TP1-PRUNH pickup
HV POWER TRANSFORMER 1 - PROT DIF TP INST PICKUP	5TP1-PRCA- pickup
HV POWER TRANSFORMER 1 - MAX I> INST UP2 DIF PICKUP	5TP1-PRI1I pickup
MV POWER TRANSFORMER 1 - MAX I> INST UP1 PICKUP	3TP1-PRI1I pickup
HV POWER TRANSFORMER 1 - MAX I> INST UP1 PICKUP	5TP1-PRI1I pickup
MV LINE VERMOIL - MAX I> INST PICKUP	3301-PRI1I pickup
MV LINE VERMOIL - MAX I>>> TEMP TRIP	3301-PRI3T trip
MV LINE VERMOIL - MAX I>>> INST PICKUP	3301-PRI3I pickup
MV POWER TRANSFORMER 1 - MIN U<< INST PICKUP	3TP1-PRUN0 pickup
MV LINE LEIRIA - MAX I> INST PICKUP	3311-PRI1I pickup
MV LINE VERMOIL - CIRCUIT BREAKER OPEN	**3301-DJEST open**

FIGURE 10. ***Alarm2Insights*** **input episode example for a circuit-breaker opening alarm in MV line VERMOIL (bay identifier 3301) from a SS in Ranha, Portugal.**

- For S0, only 29% of occurrences contained anomalies. A close inspection showed that this SS included MV lines for two grid voltage levels, 15kV and 30kV, with different terminologies per level. In fact, events from the 15kV line panels use terminology 'T3'. These 'rare' events were therefore considered as misclassifications.
- In S5, nearly 41% of the circuit breaker alarms had a 'rare' event in their context. An in-depth analysis revealed an overlooked problem on the electrification of a protection relay that was originating the pickup of an unwanted backup protection function. These 'rare' events were considered as correctly classified.
- S7 included most of occurrences without any 'rare' event. However, analogous to the findings in S5, a similar electrification problem was also detected, but during a shorter time span, therefore affecting a smaller number of occurrences (15%). These 'rare' events were also considered as correctly classified.

Regarding phase-to-phase faults (see Fig. 9, right plot) most of the events classified as 'rare' appear due to simultaneous protection relay pickup triggers in other MV lines, mainly due to generation feed-in and line induction phenomena.

As previously stated, there are still some false positives in the outputs of this function, mainly due to event terminology type changes. A solution for future work is to consider creating models per SS terminology, i.e., use the k-means clustering method mentioned above.

2) DETECTION OF POTENTIAL MISSING PROTECTIVE RELAYS TRIGGERS

Fig. 10 depicts a real example of this function input episode, and Fig. 11 the respective output for a circuit-breaker alarm event log during a phase-to-phase fault, with the expected protection relay pickups.

An analysis of Fig. 11 shows that, for this specific circuit breaker opening alarm (i.e., in SS MV line panel '3301', VERMOIL), most of the events have IDF value lower than 0.5, meaning that all refer to 'common' pickup events of protection relays.

This example also shows this function's behavior when uncommon or rare events are detected in the alarm event log. A closer look shows that there is a level three overcurrent relay pickup event, classified as uncommon, which is expected since such high pickup currents are not as frequent as the lower level pickup currents. Also, there is a simultaneous level one overcurrent relay pickup event in another MV line (i.e., in SS MV line panel '3311', LEIRIA), which is one of the main 'rare' events detected in phase-to-phase faults.

Tag / State	IDF	Rarity Classification
3301-PRI2I__pickup	0.232	Common
3TP1-PRUNH__pickup	0.208	Common
5TP1-PRCA-__pickup	0.434	Common
5TP1-PRI1I__pickup	0.163	Common
3TP1-PRI1I__pickup	0.210	Common
5TP1-PRII1__pickup	0.163	Common
3301-PRI1I__pickup	0.247	Common
3301-PRI3I__pickup	0.615	Uncommon
3TP1-PRUN0__pickup	0.226	Common
3311-PRI1I__pickup	1.600	Rare

FIGURE 11. ***Alarm2Insights*** **output for input episode illustrated in Fig. 10.**

Description	Tag/State	idf
DISTANCE PROTECTION *PICKUP*	prza-__pickup	0,00
AUTO-RECLOSING CYCLE *ACTIVE*	rlcic__active	0,00
LINE VOLTAGE *Very Low*	0tu--__v. low	0,00
AUTO-RECLOSING CYCLE *END*	rlcic__end	0,00
LINE DISCONNECTOR PERMISSION TO OPEN *NORMAL*	l0abq__normal	0,00
LINE ISOLATOR PERMISSION TO OPEN *NORMAL*	sjabq__normal	0,00
DISTANCE PROTECTION ZONE 1 *NORMAL*	przd1__normal	0,00
DISTANCE PROTECTION *NORMAL*	prza-__normal	0,00
CIRCUIT BREAKER *OFF*	djest__off	0,00
MAX I> INST UP1 *PICKUP*	priig__pickup	0,00
DISTANCE PROTECTION ZONE 1 *TRIP*	przd1__trip	0,00
MAX I> INST UP1 *NORMAL*	priig__normal	0,00
MAX Io> INST UP1 *PICKUP*	prh1i__pickup	0,29
MAX Io> INST UP1 *NORMAL*	prh1i__normal	0,29
BUSBAR SECTIONALIZER PERMISSION TO OPEN *NORMAL*	s2abq__normal	0,29
MAX Io>D INST UP1 *NORMAL*	prdi1__normal	0,69
MAX Io>D INST UP1 *PICKUP*	prdi1__pickup	0,69
CURRENT *Very High*	0ii--__v. high	0,98
CURRENT *Normal*	0ii--__normal	0,98
BUSBAR1 SECTIONALIZER PERMISSION TO OPEN *NORMAL*	s1abq__normal	1,39

FIGURE 12. **IDF-based profile of an output cluster, for type 2 blocks. Events sorted by IDF values for reporting purposes.**

The missing tags detection mechanism is straightforward as it searches for events that i) have low IDF values (i.e., lower than 0.5), and ii) do not appear in the current alarm context. Therefore, any of the common relay pickup events (see Fig. 11 events with 'common' rarity classification) for this fault type, and that are missing in the alarm context, will be clearly pinpointed to the human operator.

B. EVENTPROFILER RESULTS

The first processing phase (see Section IV-A) led to 838 type 1 blocks and 420 type 2 blocks. In this analysis, we have considered 75% of the blocks to train the clustering methods and the remaining 25% to evaluate the classification method.

For each cluster a summary report was issued, listing the tags present in the blocks, their internal representation and also their IDF, indicating how prevalent they are in the blocks forming the cluster. An example can be seen in Fig. 12, for a type 2 cluster. The cluster designation was

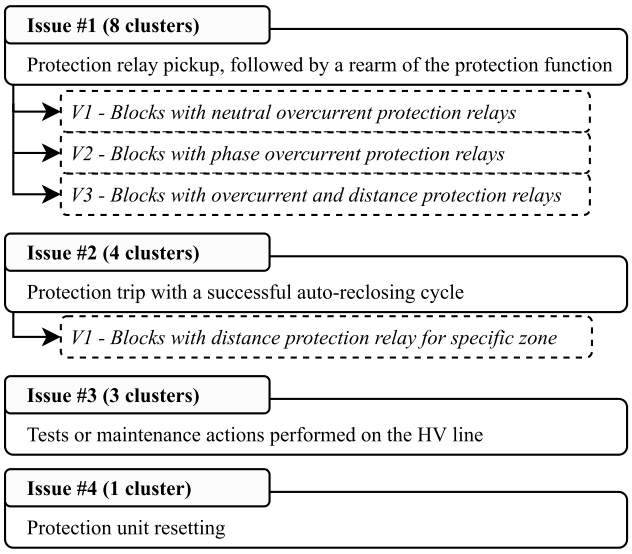


FIGURE 13. Issue descriptions provided by human operators for type 1 clusters.

assigned manually, upon inspection of the contents of the blocks by a domain expert. This particular cluster, manually designated by the domain expert as *protection trip with failed auto-reclosing*, includes outages events where a circuit breaker opens due to a distance protection trip, followed by an unsuccessful automatic reclosing cycle and a subsequent action by the human operator. It is important to emphasize that human domain knowledge remains fundamental to extract interpretable information from the clustering results, and it will not be fully replaced by autonomous AI functions (see [37] for an interesting discussion about human-AI interaction in power systems).

Fig. 13 shows the number of type 1 clusters per issue description (i.e., defined by the domain expert). The sixteen clusters were categorized according to four major categories, with two (e.g., issue 1 and 2) representing actual outages' events in HV lines. Besides separating real occurrences from test (or maintenance) actions, minor variants were also identified by different clusters within each issue (e.g., circuit breaker openings due to tripping of specific groups of protection functions).

Fig. 14 provides a similar summary for type 2 clusters. These were categorized into seven major issues and, similarly to the previous scenario, some included multiple clusters representing specific variants (e.g., different protection relay trips and outages with or without auto-reclosing cycles) while others were composed exclusively of one cluster. Other issues were automatically detected and grouped into specific clusters, such as SF6 low-pressure alarms and blocks with abnormal profiles caused by event time-delay problems (see Section II-B), which may help to discover such anomalies.

The classification of the remaining 25% of blocks not used in the clustering process supported the analysis of the types of situations in each SS. The comparison is feasible since these

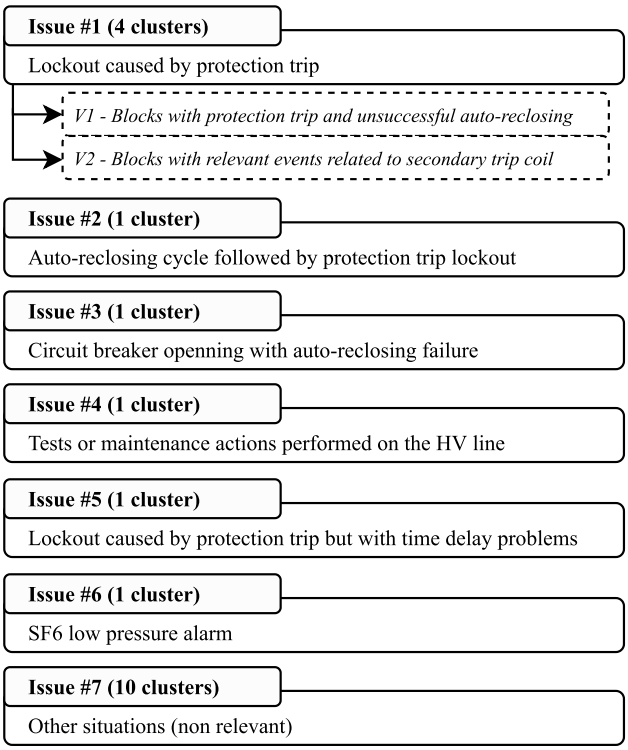


FIGURE 14. Issue descriptions provided by human operators for type 2 clusters.

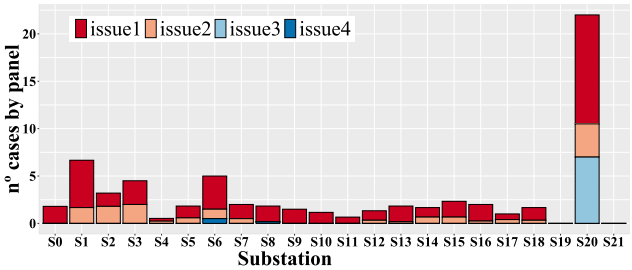


FIGURE 15. Distribution of the four type 1 issues, considering blocks created between January 2019 and June 2020.

blocks correspond to events that occurred in the same period (January 2019 to June 2020).

Fig. 15 and Fig. 16 illustrate the distribution of each issue description by SS and block type, respectively. It is important to mention that the number of HV lines may differ between SS, which would bias the results. Therefore, each plot bar refers to the total number of occurrences in the SS, divided by the respective number of HV lines connected to that SS.

In the case of type 1 blocks (see Fig. 15), most correspond to issue 1 (see Fig. 13), which is normal, as there are multiple variants belonging to this type of issue, covering most of the outages' events. This summary also shows that S20 has a large share of blocks corresponding to tests or maintenance actions in the HV lines, in contrast to the remaining SS.

TABLE 2. Designation assigned to each block that occurred at both ends of a line.

Result for block panel 1		Result for block panel 2	
Op?	Classification Result	Op?	Classification Result
Yes	Lockout caused by protection trip	No	Protection trip with a successful auto-reclosing cycle
No	Protection trip with a successful auto-reclosing cycle (Cluster with specific distance protection relay)	Yes	Lockout caused by protection trip
Yes	Lockout caused by protection trip, but with time delay problems	Yes/Yes	Lockout caused by protection trip / Lockout caused by protection trip
Yes	Auto-reclosing followed by protection trip lockout	Yes	Auto-reclosing followed by protection trip lockout
Yes	Lockout caused by protection trip, but with time delay problems	No/Yes	Protection trip with a successful auto-reclosing cycle / Auto-reclosing followed by protection trip lockout

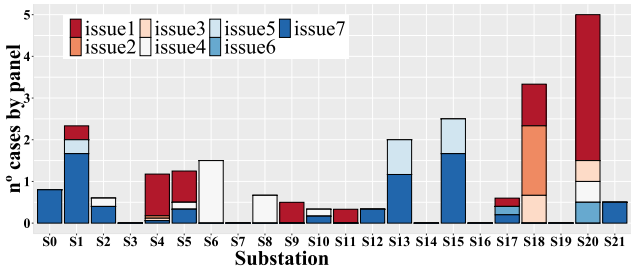


FIGURE 16. Distribution of the seven type 2 issues, considering blocks created between January 2019 and June 2020.

TABLE 3. Alarm2Insights computational processing times.

Phase	8 substations (3823 faults)	22 substations (7202 faults)
Input creation	959.22s	3145.89s
Knowledge Discovery	48.49s	96.82s
Execution	1s to 5s	

TABLE 4. EventProfiler computational processing times.

Phase	8 substations (30975 blocks)	22 substations (52419 blocks)
Definition of Blocks	600s	1500s
Train Embeddings	9900s	15300s
Clustering	1155s	4601s
Classification	5s to 9s	

S19 and S21 did not register any context block for the period under analysis.

Regarding type 2 blocks (see Fig. 16), most are classified into clusters representing less relevant situations for dispatch center operators (issues 4 and 7). Interestingly, most of the blocks corresponding to real outages events belong to outages 'with lockouts caused by protection trips' (i.e., issues 1 and 5), apart from S18, in which a large share of outage events also include auto-reclosing cycles prior to the protection trip lockout event. Also, S18 and S20 have a minor share of occurrences where the circuit breaker opened following an auto-reclosing failure event (different from issue 2, where there was still a successful auto-reclosing before the protection trip lockout). S3, S7, S14, S16 and S19 did not register any context block during the evaluation period.

In this work, we focused on providing insights for each HV line panel. However, some of the outages might affect both ends of HV lines (i.e., in different SS), even leading to simultaneous circuit breaker openings. This classification step is useful for dispatch center operators to quickly profile the system response behavior, paving the way for more complex analysis - combining, for instance, blocks created in both extremes of one HV line. Table 2 lists some of said combinations in the historical dataset. Each line represents an occurrence affecting extremes of an HV line. For a particular panel, if more than one block is included, the classification results for each will appear separated by a slash. They are an example of situations where the same classification in one panel is associated with different classifications on the opposite panel, showing a way to provide insights on the history of occurrences for a line or combination of panels, but also helping to reduce the cognitive load for an operator, avoiding the need to inspect a sequence of event logs and summarizing an occurrence through a set of cluster descriptions.

C. COMPUTATIONAL TIMES

Tables 3 and 4 present the computational times of *Alarm2Insights* and *EventProfiler* functions considering 8 and 22 SS. *Alarm2Insights* function tests were performed in a Intel(R) Core(TM) i7-6700 CPU, with 12 GB RAM. *EventProfiler* function tests were performed in a Intel(R) Core(TM) i7-2600 CPU, with 16 GB RAM. Both machines run on a 64-bit Windows 10 operating system.

In both functions there is an almost linear increase of the computational times with episode creation, knowledge discovery and clustering phases. Besides, in order to integrate new historical data, these phases only need to be performed on a monthly basis. For operational purposes, the most

relevant phases are software execution and classification, where the change in computational times was marginal.

VI. CONCLUSION

The results obtained in raw SCADA data from a DSO highlight the potential of applying data-driven functions (based on AI and machine learning methods), supporting the operators' decision-making process by categorizing the high volume of event log data (*EventProfiler* function), while also helping to uncover potential problems in SS protection relays (*Alarm2Insights* function).

Information presented in a much clearer, insightful and error-proof manner eventually leads to better informed decisions that empower and enhance the role of dispatch center operators. The detection of trends and anomalies in large event log datasets, otherwise overlooked, already proved beneficial to increase human operator situational awareness, both close to real-time and *post-mortem*. The results in section V-A show that the DSO could detect protection relays malfunction earlier, and quickly take the necessary actions to solve it, and the outputs from the *EventProfiler* function, as showed in section V-B, can help dispatch center operators to identify normal occurrences (e.g., fault with successful auto-reclosing cycle) and anomalies on the SCADA alarm data without having to manually analyze hundreds of individual events. Despite the different operational challenges identified and tackled by each function, the combination of both approaches provides different and complementary layers of information in both close to real-time and *post-mortem* event analysis.

The integration of both functions with the Distribution Management System requires further work (e.g., standardization of event tags, minimizing the impact of time delays between RTU and SCADA events, parallel processing), in order to improve the real-time reliability and favor its daily use by dispatch operators. Topics for future work include: a) integrate network topology information; b) create separate models according to the SS equipment generation or vendors (departing from the PCA and clustering process discussed in section V); c) generate synthetic data for SS or specific lines with few historical events. This work represents a first step towards AI-assisted operation of electrical grids [37].

REFERENCES

[1] M. Kezunovic, P. Pinson, Z. Obradovic, S. Grijalva, T. Hong, and R. J. Bessa, "Big data analytics for future electricity grids," *Electr. Power Syst. Res.*, vol. 189, p. 106788, Dec. 2020.

[2] L. Wei, W. Guo, F. Wen, G. Ledwich, Z. Liao, and J. Xin, "An online intelligent alarm-processing system for digital substations," *IEEE Trans. Power Del.*, vol. 26, no. 3, pp. 1615–1624, Jul. 2011.

[3] S. Pandey, A. K. Srivastava, and B. G. Amidan, "A real time event detection, classification and localization using synchrophasor data," *IEEE Trans. Power Syst.*, vol. 35, no. 6, pp. 4421–4431, Nov. 2020.

[4] N. Baranovic, P. Andersson, I. Ivankovic, K. Zubrinic-Kostovic, D. Peharda, and J. E. Larsson, "Experiences from intelligent alarm processing and decision support tools in smart grid transmission control centers," in *Proc. CIGRE Session*, Paris, France, Aug. 2016, pp. 1–10.

[5] C. Feng, L. Wang, R. Ye, J. Gu, L. Xie, Y. Wang, Q. Feng, and C. Cui, "Research and application of alarm optimization mechanism in power grid operation monitoring," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Chongqing, China, Mar. 2021, pp. 2352–2356.

[6] D. B. Tesch, D. C. Yu, L.-M. Fu, and K. Vairavan, "A knowledge-based alarm processor for an energy management system," *IEEE Trans. Power Syst.*, vol. 5, no. 1, pp. 268–275, Feb. 1990.

[7] G. Sun, X. Ding, Z. Wei, P. Shen, Y. Zhao, Q. Huang, L. Zhang, and H. Zang, "Intelligent classification method for grid-monitoring alarm messages based on information theory," *Energies*, vol. 12, no. 14, p. 2814, Jul. 2019.

[8] D. S. Kirschen and B. F. Wollenberg, "Intelligent alarm processing in power systems," *Proc. IEEE*, vol. 80, no. 5, pp. 663–672, May 1992.

[9] Z. A. Vale, A. Machado, M. F. Fernandes, and C. Ramos, "Sparse: An intelligent alarm processor and operator assistant," *IEEE Expert*, vol. 12, no. 3, pp. 86–93, May 1997.

[10] G. Cardoso, J. G. Rolim, and H. H. Zurn, "Application of neural-network modules to electric power system fault section estimation," *IEEE Trans. Power Del.*, vol. 19, no. 3, pp. 1034–1041, Jul. 2004.

[11] C. L. Hor, P. A. Crossley, and S. J. Watson, "Building knowledge for substation-based decision support using rough sets," *IEEE Trans. Power Del.*, vol. 22, no. 3, pp. 1372–1379, Jul. 2007.

[12] J. Sun, S.-Y. Qin, and Y.-H. Song, "Fault diagnosis of electric power systems based on fuzzy Petri nets," *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 2053–2059, Nov. 2004.

[13] Y. Zhu, H. Limin, and J. Lu, "Bayesian networks-based approach for power systems fault diagnosis," *IEEE Trans. Power Del.*, vol. 21, no. 2, pp. 634–639, Apr. 2006.

[14] H. Miao, M. Sforna, and C.-C. Liu, "A new logic-based alarm analyzer for on-line operational environment," *IEEE Trans. Power Syst.*, vol. 11, no. 3, pp. 1600–1606, Aug. 1996.

[15] C.-L. Hor and P. A. Crossley, "Unsupervised event extraction within substations using rough classification," *IEEE Trans. Power Del.*, vol. 21, no. 4, pp. 1809–1816, Oct. 2006.

[16] T. Wang, W. Liu, J. Zhao, X. Guo, and V. Terzija, "A rough set-based bio-inspired fault diagnosis method for electrical substations," *Int. J. Electr. Power Energy Syst.*, vol. 119, p. 105961, Jul. 2020.

[17] W. Guo, F. Wen, Z. Liao, L. Wei, and J. Xin, "An analytic model-based approach for power system alarm processing employing temporal constraint network," *IEEE Trans. Power Del.*, vol. 25, no. 4, pp. 2435–2447, Oct. 2010.

[18] Y. Jiang and A. K. Srivastava, "Data-driven event diagnosis in transmission systems with incomplete and conflicting alarms given sensor malfunctions," *IEEE Trans. Power Del.*, vol. 35, no. 1, pp. 214–225, Feb. 2020.

[19] C. Wang, K. Pang, M. Shahidehpour, and F. Wen, "MILP-based fault diagnosis model in active power distribution networks," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 3847–3857, Sep. 2021.

[20] B. Xu, C. Wang, F. Wen, I. Palu, and K. Pang, "Fault diagnosis and identification of malfunctioning protection devices in a power system via time series similarity matching," *Energy Convers. Econ.*, vol. 1, no. 2, pp. 81–92, Nov. 2020.

[21] Z. Liao and S. Liu, "Substation alarm information processing based on ontology theory," in *Proc. 5th Int. Conf. Electr. Utility Deregulation Restructuring Power Technol. (DRPT)*, Changsha, China, Nov. 2015, pp. 2377–2382.

[22] Z. Bai, G. Sun, H. Zang, M. Zhang, P. Shen, Y. Liu, and Z. Wei, "Identification technology of grid monitoring alarm event based on natural language processing and deep learning in China," *Energies*, vol. 12, no. 17, p. 3258, Aug. 2019.

[23] X. Zhang, J. Wei, S. Yue, and X. Zha, "An analytical method for mapping alarm information to model of power grid fault diagnosis," *IEEJ Trans. Elect. Electron. Eng.*, vol. 13, no. 6, pp. 823–830, 2018.

[24] Z. Wang, Q. Chen, L. Wang, L. Lin, G. Li, and L. Niu, "Processing and application of substation alarm information based on semantic framework," in *Proc. IEEE/IAS Ind. Commercial Power Syst. Asia (I&CPS Asia)*, Weihai, China, Jul. 2020, pp. 1023–1028.

[25] J. Wang, X. Wang, C. Ma, and L. Kou, "A survey on the development status and application prospects of knowledge graph in smart grids," *IET Gener., Transmiss. Distrib.*, vol. 15, no. 3, pp. 383–407, Feb. 2021.

[26] P. Santos, H. Barbosa, A. Blanquet, B. Fischer, and B. Espírito Santo, "EDGE digital substation–A disruptive automation field project," in *Proc. 25th Int. Conf. Elect. Dist. (CIRED)*, Madrid, Spain, Jun. 2019, pp. 1–4.
[27] *Instalações AT e MT. Sistemas de Proteção, Comando e Controlo Numérico (SPCC). Funções de Proteção—Especificação Funcional, EDP Distribuição*, Standard DEF-C13-570/N, Ed. 4, Jul. 2020. [Online]. Available: https://www.e-redes.pt/sites/eredes/files/2020-09/DEF-C13-570.pdf
[28] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 1–37, Jun. 2008.
[29] J. Krstulovic and V. Miranda, "Selection of measurements in topology estimation with mutual information," in *Proc. IEEE Int. Energy Conf. (ENERGYCON)*, Cavtat, Croatia, May 2014, pp. 589–596.
[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 3111–3119.
[31] R. Verde and A. Irpino, "Comparing histogram data using a Mahalanobis–Wasserstein distance," in *Proc. COMPSTAT: Comput. Statist.*, 2008, pp. 77–89.
[32] C. Rocha and P. Q. Brito, "Profiles identification on hierarchical tree structure data sets," *J. Appl. Statist.*, vol. 45, no. 15, pp. 2848–2863, Nov. 2018.
[33] E. Levina and P. Bickel, "The Earth Mover's distance is the Mallows distance: Some insights from statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 251–256.
[34] C. Hennig, "Cluster-wise assessment of cluster stability," *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 258–271, Sep. 2007.
[35] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 2009.
[36] C. Hennig, "Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods," *J. Multivariate Anal.*, vol. 99, no. 6, pp. 1154–1176, Jul. 2008.
[37] A. Marot, A. Kelly, M. Naglic, V. Barbesant, J. Cremer, A. Stefanov, and J. Viebahn, "Perspectives on future power system control centers for energy transition," *J. Mod. Power Syst. Clean Energy*, vol. 10, no. 2, pp. 328–344, 2022.

J. R. ANDRADE received the M.Sc. degree in electrical and computer engineering from the Faculty of Engineering of the University of Porto (FEUP), Portugal, in 2016. Currently, he is a Researcher with the Center for Power and Energy Systems, INESC TEC. His research interests include energy analytics, renewable energy and electricity prices forecasting, and the potential applications of natural language processing to the energy sector.

C. ROCHA received the degree in chemistry engineering from the Faculty of Engineering of the University of Porto (FEUP), Porto, Portugal, in 1995, the degree in physics from the University of Science (FCUP), Porto, in 2008, the M.Sc. degree in quantitative methods in economics and management from the Faculty of Economics (FEP), Porto, in 2011, and the Ph.D. degree in applied mathematics from the University of Porto, Portugal, in 2014. Currently, she is a Researcher with the Center for Power and Energy Systems, INESC TEC. Her research interests include data validation, data collecting techniques, modelling, statistics, predictive modelling, and text mining/data mining.

R. SILVA received the M.Sc. degree in electrical and computer engineering from the Faculty of Engineering of the University of Porto (FEUP), Porto, Portugal, in 2018. Currently, he is a Researcher with the Center for Power and Energy Systems, INESC TEC. His research interests include optimal management of micro-grids and energy communities, data-driven modeling, and the design of local electricity markets.

J. P. VIANA received the degree in mathematics from the University of Minho, in 2008, and the M.Sc. degree in computer science from the Faculty of Sciences of University of Porto, in 2019. Currently, he is a Researcher with the Center for Power and Energy Systems, INESC TEC. His research interests include data engineering and modeling and load forecasting—in low voltage.

RICARDO J. BESSA (Senior Member, IEEE) was born in Viseu, Portugal, in 1983. He received the Licenciado degree in electrical and computer engineering from the Faculty of Engineering of the University of Porto (FEUP), Porto, Portugal, in 2006, the M.Sc. degree in data analysis and decision support systems from the Faculty of Economics of the University of Porto (FEP), Porto, in 2008, and the Ph.D. degree in sustainable energy systems (MIT Portugal) from FEUP, in 2013. Currently, he is Co-Ordinator with the Center for Power and Energy Systems, INESC TEC. He worked at several international projects, such as the European Projects FP6 ANEMOS.plus, FP7 SuSTAINABLE, FP7 evolvDSO, Horizon 2020 UPGRID, InteGrid, InterConnect and Smart4RES, and an international collaboration with the Argonne National Laboratory for the U.S. Department of Energy. His research interests include renewable energy, energy analytics, smart power systems, and electricity markets. He is the Editor of the IEEE TRANSACTIONS ON SUSTAINABLE ENERGY.

C. GOUVEIA received the M.Sc. and Ph.D. degrees in electrical engineering from the Faculty of Engineering, University of Porto (FEUP), in 2008 and 2015, respectively. She has been a member with the Centre for Power and Energy systems, INESC TEC, since 2011, where she is currently a Senior Researcher. Her research interests include operation of distribution networks within smart grid context, considering the large-scale integration of distributed energy resources, and microgrid concepts. She has been involved in several national and European projects, such as MERGE, SENSIBLE, and UPGRID project, namely in the development and demonstration activities in INESC TEC Smart Grids and the Electric Vehicles Laboratory of control and management strategies to enable the safe integration of distributed energy resources in distribution networks, particularly when operating islanded from the main grid.

B. ALMEIDA received the M.Sc. degree in electrical and computer engineering from the Instituto Superior Tecnico (IST), University of Lisbon, Portugal, in 2013. He is currently a Project Manager with the Innovation and Development Department, E-REDES, with previous experience in the development of power protection coordination studies and power system incidents analysis and prevention.

R. J. SANTOS received the M.Sc. degree in electrical engineering from the Instituto Superior Técnico (IST), University of Lisbon, Portugal, in 2007. He is currently an Associate Director with the Innovation Department, E-REDES, where he is managing innovation projects for electrical power grids, mostly focused in smart-grids and micro-grids, energy storage, and demand response. He also manages a wide variety of other innovation projects with scopes ranging from augmented reality to data analytics and AI, to the IoT sensorization for asset management.

M. LOURO received the bachelor's and M.Sc. degrees in electrical engineering from the Instituto Superior Técnico (IST), University of Lisbon, Portugal, in 2008. He is currently the Deputy Director of distribution system optimization with E-REDES, and supervising large-scale incident analysis, distribution power system analysis and protection, load and generation forecast, and distribution power network optimization.

P. M. SANTOS received the M.Sc. degree in electrical and computer engineering from the Instituto Superior Técnico (IST), University of Lisbon, in 2014. He is currently a Dispatcher with EDP Produção, with previous experience at the E-REDES High-Voltage Grid Operation Center.

A. F. RIBEIRO received the Ph.D. degree in applied mathematics from the Faculty of Sciences of Porto University, in 2014. She is currently an Area Product Owner with the Digital Boost Department, E-REDES, with previous experience with Big Data and BI-SCADA platform which supports decision-making and operational management of the distribution networks.