IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 October 2024, accepted 18 November 2024, date of publication 22 November 2024,
date of current version 9 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3504830

RESEARCH ARTICLE

Transform-Based Multiresolution Decomposition for Unsupervised Learning and Data Clustering of Cellular Network Behavior

JUAN CANTIZANI-ESTEPA[1], SERGIO FORTES[1], (Senior Member, IEEE), JAVIER VILLEGAS[1], JAVIER RASINES[2], RAÚL MARTÍN CUERDO[3], AND RAQUEL BARCO[1]

[1]Telecommunication Research Institute (TELMA), E.T.S. Ingeniería de Telecomunicación, University of Málaga, 29010 Málaga, Spain
[2]Ericsson-GAIA Sweden, 164 40 Kista, Sweden
[3]Ericsson-NDO SW Research and Development, 28045 Madrid, Spain

Corresponding author: Sergio Fortes (sfr@ic.uma.es)

This work was supported in part by the“Desarrollo de casos de uso para el diseño, optimización y dimensionado de redes móviles-Líneas B1 y D1,” as part of the Agencia de Innovación y Desarrollo de Andalucía (IDEA) incentives, under Contract 8.06/5.59.5705-3 IDEA; in part by the “Ministerio de Universidades” under Grant FPU21/04472; in part by the Spanish Government “Ministerio de Asuntos Económicos y Transformación Digital y la Unión Europea-NextGenerationEU,” through the Project MAORI inside the plan entitled “Plan de Recuperación, Transformación y Resiliencia y el Mecanismo de Recuperación y Resiliencia”; and in part by the University of Málaga by the “II Plan Propio de Investigación y Transferencia de la Universidad de Málaga.”

ABSTRACT The growing complexity of cellular networks makes it harder for network operators to control and manage the system. To ease the management and automatically detect network problems, unsupervised techniques have been put to use. This work proposes a novel method that combines Multi-Resolution Analysis (MRA) by wavelet transforms and unsupervised clustering with pre-initialized Gaussian Mixture Models (GMMs) for the totally unsupervised grouping of cellular network behaviors using different metrics. The application of multi-resolution decomposition, allows the much simpler clustering technique to take into account temporal information that would require of a much complex method otherwise, being useful for cluster analysis by experts as different duration issues are now segregated and automatically labeled. The generated labels are indicative of the intensity and duration of the anomalies, such labeling can be linguistic or visual, providing faster issue identification. The proposed approach has been tested with real network data, successfully separating different behaviors analyzed in the evaluation section of the manuscript.

INDEX TERMS Anomaly detection, cellular networks, clustering, multiresolution analysis (MRA), Gaussian mixture model (GMM).

I. INTRODUCTION

Cellular networks have constantly evolved in capabilities and complexity from the beginning of 2G to the current 5G. Due to this increasing complexity, it is necessary to study and develop methods that allow the rapid detection and clustering of anomalies in the network. A common approach to tackle this is to use Machine Learning (ML) or Artificial Intelligence (AI) techniques, these help simplify the issue detection and understanding. Some of the main challenges when using cellular network metrics are: the necessity to analyze the temporal correlation between multiple metrics to discover complex behaviors, data granularity, or the strong seasonality caused by the users. These issues make it difficult for clustering algorithms to detect and classify problems accurately.

However, other problems arise when trying to evaluate the results of such automatic methods. If the methods are supervised, a set of network experts need to analyze and label the data, whereas if the methods are unsupervised, the issues present in the data need to be well identified to generate an accurate evaluation of the method.

This poses the requirement of many hours of expert knowledge just to get to know your data and characterize the different issues that are contained within.

The associate editor coordinating the review of this manuscript and approving it for publication was Miguel López-Benítez.

For the previous reasons, this manuscript focuses on the issue of cellular network anomaly detection and classification in cellular network datasets where several network behaviors or anomalies are present.

As stated previously, in works like [1], where supervised methods are applied, the issue of first facing a new or unknown dataset requires great knowledge, experts need to analyze it in order to identify the multiple issues or behaviors contained in it. Supervised methods are not adequate for new datasets due to label requirement.

Due to this, unsupervised methods are a better option to work with new datasets, since they do not need the anomalies to be labeled prior to their execution. However, the best performing methods, like Variational AutoEncoder (VAE)-based methods, require a great effort to be put into hyperparameter tuning, generating big models that also take a lot of time and resources to tune and train.

Here, and to the best of our knowledge, this is the first work that applies the combination of wavelet decomposition plus simple clustering, not requiring expert knowledge nor hyperparameter tuning, to the issue of cellular networks anomaly detection and classification. Specifically, using the wavelet decomposition values and not just as a denoising agent.

Then, the key contribution of this work lies in going beyond the denoising capacities of wavelet decomposition [2] and its applications in anomaly management, which has been primarily detection-centric [3], thus contributing to the current literature related to network issue detection and classification. Besides, the present work proposes a new dual automatic labeling, both visual and linguistic, providing insight about the duration and intensity of the detected anomalies.

The proposed framework is completely automatic and has been tested over an unlabeled cellular dataset containing 16 different metrics, proving capable of identifying the most common behaviors found in the data while providing insight on the duration of such issues thanks to the wavelet decomposition. This framework is useful in the exploratory phase of a new dataset, providing experts with insight on the most common issues found in the network, accelerating analysis. This way, a network operator that may be trying to create an anomaly detection and classification tool, fine-tuned to new data from his network, can use this method to accelerate the understanding of experts of the issues found in such new dataset.

The structure of the paper is as follows: Section I presents a brief introduction to network anomaly detection and classification, as well as this manuscript contribution and structure. In Section II different works related to the current paper are presented and commented. Section III presents the current framework, in order to later describe the different steps in detail throughout the following four sections, Sections IV, V, VI and VII, where the Multi-Resolution Analysis (MRA), the automatic algorithm for the selection of the number of groups, the clustering, and the automatic dynamic method for the linguistic labeling of samples are described. Finally, Section VIII summarizes the results obtained closing the manuscript with Section IX with the achieved conclusions.

II. RELATED WORKS

The field of ML/AI anomaly detection in cellular networks has been extensively explored during the last years with a variety of methods. In [4] there is a wide bibliography on methods that are classified in groups based on the selected approach, clustering-based methods, statistical methods, classification-based methods or information theory based methods.

From this body of knowledge, we will focus on those works related to network anomaly detection and classification based on the usage of cellular metrics and Key-Performance Indicators (KPIs). Inside this set of works, there are mainly two investigation lines based on the ML/AI classifier nature, mainly, supervised classification or unsupervised classification is used.

Among the most recent supervised network anomaly classification works we find [1], where five different supervised methods are applied on a dataset with the purpose of identifying the best performing one. The importance of exploratory data analysis and prior expert labeling is clearly stated in the article, since the efficiency of the supervised algorithms is strongly dependent on such information. One issue of this work is that it focuses on a single metric, the Paging Success Rate, greatly simplifying the classification problem with a binary classification (anomalous or regular behavior).

Other works like [5], based on a Support Vector Machine (SVM), or approaches that use different types of Neural Networks (NNs) like [6] or [7] achieve good results, but being supervised techniques they still require a huge amount of labeled data for their training, as well as proper hyperparameter tuning for the NN-based methods. Some other supervised techniques, try context information, like location in [8] or social events in [9], to improve results and facilitate failure management and prediction.

In [10] an unsupervised method based on transfer learning and VAEs is used, techniques based on this approach prove to have very good results, but the amount of hyperparameters to define are huge compared to the approach presented here, since the possible implementations of a VAE are immense. In [11] and [12] a combination of much simpler techniques is used. In [11], feature extraction and density-based clustering are combined to determine whether a sample presents normal or anomalous behavior (though it does not deep into the classification of such behavior). In [12] a rather similar scheme is applied, but instead of grouping metric samples directly it applies Density Based Spatial Clustering of Applications with Noise (DBSCAN) to the regular daily shapes of the dataset to differentiate regular daily behavior from anomalous ones, once again not describing the different anomalies found.

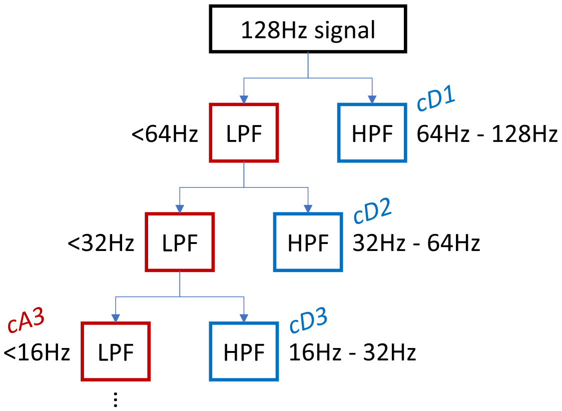


FIGURE 1. Diagram of iterated 2-band filter banks representing the results of multiresolution decomposition by wavelets.

About the usage of MRA, it has been greatly explored in multiple fields like energy grid failure analysis [13], imaging techniques [14], image analysis [15]…But to the knowledge of the authors, the only two related works that apply this decomposition for cellular network anomalies use it as a way of denoising the metrics [2] or uniquely as a detector of anomalies [3], without going deeper into the classification of the multiple anomalies or behaviors present in the data.

Other more recent works, but in this case focused on the Internet of Things (IoT) scenario, are [16] and [17]. In [16] the Haar wavelet transform is used to do feature extraction prior to the application of an AutoEncoder (AE), suffering from the same issues as [10]. In [17] an even more complex network is presented, achieving very good results on anomaly detection, but without any classification of the anomalies or issue labeling as proposed here.

This manuscript introduces an innovative methodology based on MRA and wavelet decomposition for unsupervised classification of anomalies based on the wavelet decomposition of cellular network metrics. MRA decomposes a metric into a multitude of sub-metrics, each of which encapsulates information spanning different temporal and frequency spectrums, like shown in [3]. The methodology is characterized by its ability to identify a variety of patterns, taking into account their temporal aspects and the changes of metric values over different time frames.

III. GENERAL FRAMEWORK

The algorithm presented in this work, delineated in Fig. 2 and Algorithm 1, introduces an innovative approach to the detection and clustering of cellular metrics. It integrates several computational methods into a single, coherent procedure designed to improve the first approach to unlabeled data by experts, providing groups representing different network issues.

In the first place, a series of steps are applied to ensure the accuracy of the data by identifying and removing measurement errors (data imputation via interpolation of lacking or invalid samples). This steps are followed by a min-max normalization procedure to standardize the range of values in the dataset. Then, MRA is used for decomposition, to further refine the data. After the application of the MRA, the Automatic Cluster Number Selection by Finding Density Peaks (ACNSFDP) algorithm is used to initialize the grouping technique and establish the number of clusters. The next stage involves the application of a Gaussian Mixture Model (GMM) to enable effective clustering of the data. The final stage of the algorithm is the analysis of the results along with textual label generation, where insights are extracted and interpreted.

Algorithm 1 Process Steps

```
Input: Metrics data
Output: Clusterized metrics (with labels)
  1) Filter measuring errors
  2) Apply Min-Max normalization
  3) MRA decomposition of metrics in levels
  4) Execution of the ACNSFDP algorithm
  5) GMM clustering
  6) Result analysis
  7) Label generation (optional)
```

IV. WAVELET DECOMPOSITION FOR MULTIRESOLUTION ANALYSIS (MRA)

MRA, specifically using wavelet transforms, provides a powerful method for analyzing a wide range of cellular behaviors considering their temporality [3]. This approach enables the decomposition of time series at multiple levels, capturing effects of different duration at each level. This differentiates between transient issues (of 1-2 hours) from more prolonged ones (8-16 hours). Deeper decomposition levels correspond to longer event duration.

To elucidate this concept, a comparison can be drawn with the Short-Time Fourier Transform (STFT). The continuous form of the STFT equation is given as:

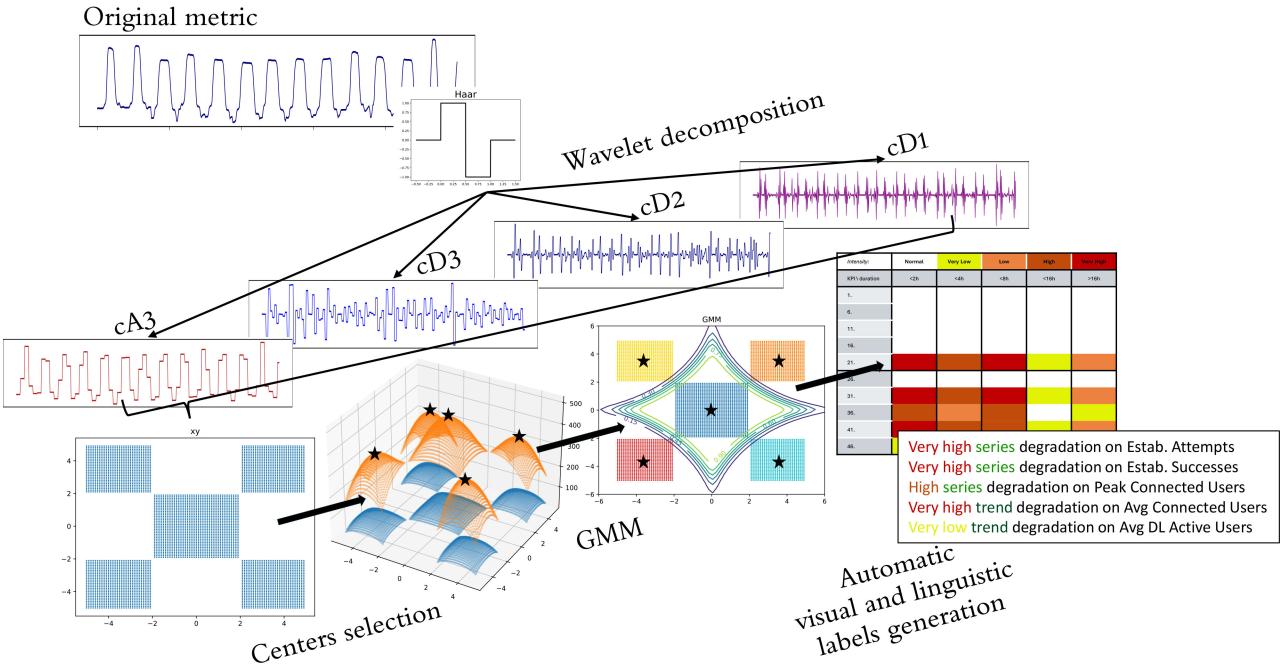
$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{i\omega t}dt, \tag{1}$$

where $x(t)$ represents the signal to be transformed, $w(t-\tau)$ is the selected window function shifted in time and $e^{i\omega t}$ is the complex exponential term, which represents oscillations at a particular frequency ω. Although STFT is widely used for analyzing a signal's frequency behavior, its time and frequency resolutions are determined by the width of the analysis window ($w(t-\tau)$). Enhancing the resolution in time invariably leads to a decrease in frequency resolution.

Otherwise, MRA involves transforming the signal using wavelets [18]. The continuous wavelet transform is given by Eq. (2) as:

$$X(\Psi) = \int_{-\infty}^{\infty} x(t)\Psi^{*}_{a,b}(t)dt, \tag{2}$$

where $x(t)$ represents the signal to be transformed and $\Psi^{*}_{a,b}(t)$ is the chosen wavelet function with scale a and translation b.The latter determines where the wavelet is applied along



Very high series degradation on Estab. Attempts
Very high series degradation on Estab. Successes
High series degradation on Peak Connected Users
Very high trend degradation on Avg Connected Users
Very low trend degradation on Avg DL Active Users

FIGURE 2. Scheme of the implemented algorithm depicting the different steps: wavelet decomposition, center selection, clustering and automatic visual and linguistic labelling. The original metric is decomposed using a wavelet (in this case the Haar wavelet) into multiple subcomponents, these subcomponents are then fed to the clustering steps (centers selection and GMM) and then automatically labeled, visually (heatmap style table) and linguistically.

the time axis, whereas the scale sets the wavelet frequency, a larger a corresponds to a broader (low-frequency) wavelet, and a smaller a corresponds to a narrower (high-frequency) wavelet.

Contrary to the STFT, the wavelet transform utilizes wavelets, enabling an analysis where the frequency and time resolutions can vary. For signals of lower frequency, the time resolution decreases whereas frequency resolution improves. In the case of higher-frequency signals, the time resolution increases whereas frequency resolution decreases.

This approach is ideally suited for analyzing metrics exhibiting persistent daily and weekly patterns along with rapid, non-periodic changes (anomalies). While a variety of wavelets can be chosen for the transform, the D2 or Haar wavelet is preferred for this application as it allows for more effective detection of rapid transitions in the signal. After the MRA, the absolute maximum value for each component is selected to reduce the dimensions of the features.

Consequently, each transform component encapsulates the maximum values with different duration and frequency windows. If a 64-sample signal undergoes five levels of decomposition, events of the shortest duration and highest frequency (1-2 samples) are analyzed at level 1, slightly longer events (2-4 samples) at level 2, and so on, increasing consecutively in powers of 2 (4-8, 8-16, 16-32 samples and >32 samples). To illustrate this idea, a comparison can be done with a bank filter as depicted in Fig. 1, where each step of the system segregates the signal in two, one containing the high frequency elements, and the other containing the lower frequency elements. As indicated, the subsequent higher frequency portions will correspond with the discrete components, whereas the low frequency remainder is left as the approximation component.

V. AUTOMATIC SELECTION OF THE NUMBER OF GROUPS

As indicated in the algorithm steps, once the metric signals have been decomposed in their different levels, the next step is to automatically group the samples based on the MRA results. One of the issues of unsupervised classification when the number of existing classes is not known a priori, is the selection of the number of groups to be generated. In order to make the generated algorithm as automatic as possible, the ACNSFDP [19] algorithm, based on [20], has been implemented to solve this issue.

This algorithm selects the optimal number of clusters by analyzing the local density (ρ) of each sample and their distances (δ). It is based on the supposition that cluster centers are highly dense points that are relatively separated from others. With these two parameters, a Cluster Selectivity score (CS) is obtained for each sample. Using this score, a graph is generated where the samples are sorted according to the score obtained.

To compute the mentioned parameters the distance between samples must be calculated first. In this case the euclidean distance is being used. After this step, the local density (ρ) of each sample is calculated by Eq. (3) as:

$$\rho_i = \sum_{j=1}^{n-1} e^{\frac{-d_{ij}}{d_c}}, \tag{3}$$

using different distances and a gaussian kernel.

- d_{ij} is the distance between point i and point j and $i \neq j$.

- d_c is the cut-off distance.
- n is the total of samples.

This equation requires a cutoff distance (d_c). Generally, (d_c) can be automatically set as percentile five of all d_{ij} values. As stated in [19], the method results are not really sensitive to this value setting.

After computing ρ, δ follows by applying Eq. (4) as:

$$\delta_i = \begin{cases} max_(d_{ij}), & \text{if } \rho_i = max(\rho) \\ min_{(j:\rho_j>\rho_i)}(d_{ij}), & \text{otherwise}, \end{cases} \tag{4}$$

where, for each sample, the formula of δ changes depending on whether the sample has the maximum local density (ρ) or not. On one hand, if the local density (ρ) of a point is maximum, its δ value equals the maximum distance found to any other sample. On the other hand, if local density (ρ) is not maximum, δ is equal to the minimum distance to a sample with greater local density (ρ).

As the final parameter computation step of the ACNSFDP algorithm, the CS is computed via Eq. (5) as:

$$CS_i = \rho_i \cdot \delta_i, \tag{5}$$

The generated score is sorted from highest to lowest.

To select the optimal number of groups the elbow of the score must be located. This is achieved by fitting two least squares lines to the data and sequentially modifying the cutoff point. The pair of lines that minimize the fitting error are selected, using the highest scoring samples as centers.

Fig. 3 shows an example where the optimal number of groups is found after the sequential adjustment at k=5. The point where the elbow is located is marked with a vertical red line. On the left are the five samples with the highest scores, which are the center candidates when applying GMM or any other clustering method.

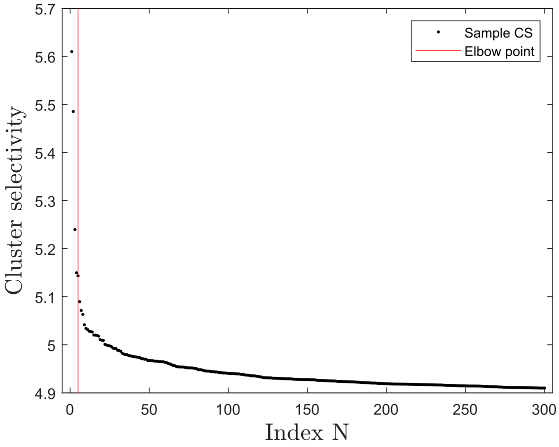


FIGURE 3. **Cluster selectivity score of the best scoring 300 samples with the elbow marked in red.**

VI. CLUSTERING VIA GAUSSIAN MIXTURE MODEL (GMM)

GMM has been chosen as the clustering algorithm for this task due to its superior performance compared to k-means, despite a higher computational overhead [21]. GMM, paired with its adjustment stage, the Expectation-Maximization (EM) algorithm, facilitates the automatic selection of the number of clusters. If an excessively high number of clusters is selected, some clusters will simply remain vacant, thereby avoiding the generation of extremely small or singular clusters.

The principle of GMM revolves around fitting a set of gaussian distributions to the dataset. Consequently, each sample is attributed to the gaussian distribution where it has the maximum likelihood of membership. When dealing with multidimensional data, the gaussian distributions are expressed as:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \tag{6}$$

which models the probability of x occurring, given μ and Σ, where x is a D-dimensional vector representing a point in the space, μ represents an D-dimensional vector encompassing the mean values of the gaussian (center of the distribution) and Σ denotes the $D \times D$-dimensional covariance matrix. The value $|\Sigma|$ is the determinant of Σ and D is the number of dimensions $x \in \mathbb{R}^D$. When there is a modification of μ, the gaussians are shifted on the D-dimensional plane, whereas changing the covariance matrix Σ changes the shape of the gaussian.

The EM algorithm is an iterative method that determines the mean and covariance matrix for each gaussian distribution. The initialization of these values is non-trivial and significantly impacts the results [22]. In our implementation, the means are set to the parameter values of the samples selected by ACNSFDP as cluster centers. The covariance matrix is set diagonal with random initial values to ease convergence.

The fitting procedure consists of two iterative steps:

- Expectation, where the probability that each gaussian assigns to each sample is calculated, allowing the estimation of each sample's group membership based on similarity.
- And maximization, where the coefficients for each gaussian are recalculated using the newly assigned group memberships.

After several iterations of these two steps, a point of minimum deviation is reached where the coefficients from the maximization step no longer differ from their previous values. The results are then obtained and subject to further analysis.

VII. AUTOMATIC VISUAL AND LINGUISTIC LABEL GENERATION

As a last step, to facilitate the understanding of the behavior or issues found in the different samples, a labeling system is applied based on the analysis of the different decomposition level values.

These labels are a composition of 3 elements, an intensity label, a time label and the name of the metric.

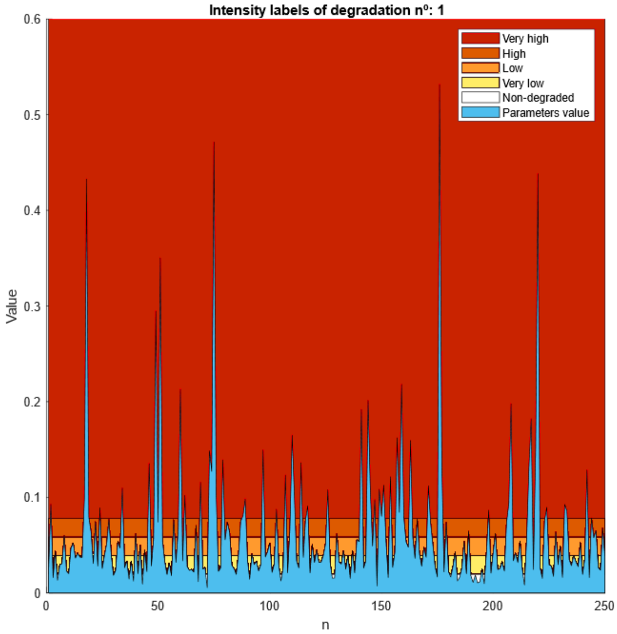


FIGURE 4. Intensity level of the label based on the number of deviations from the mean for the fastest level of decomposition of a measured metric.

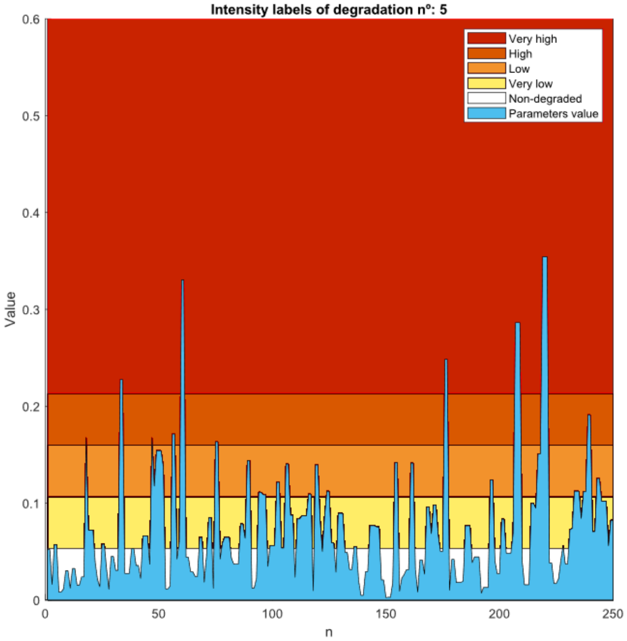


FIGURE 5. Intensity level of the label based on the number of deviations from the mean for the slowest level of decomposition of a measured metric.

The intensity label is created dynamically by the analysis of the deviation of the maximum value, separately in each level of degradation, for each 24 hours of measures. Taking the mean plus the standard deviation of a certain metric throughout the dataset as baseline, four intensity labels are possible based on the strength of the deviation. This four deviation strengths are labeled as follows:

$$\text{Strength label} = \begin{cases} \text{``Very Low''}, \text{if} \in [\mu + \sigma, \mu + 2\sigma) \\ \text{``Low''}, \text{if} \in [\mu + 2\sigma, \mu + 3\sigma) \\ \text{``High''}, \text{if} \in [\mu + 3\sigma, \mu + 4\sigma) \\ \text{``Very High''}, \text{if} \geq \mu + 4\sigma. \end{cases} \tag{7}$$

In Fig. 4, the four different ranges are displayed along with the value of the metric for multiple consecutive 24h samples in the case of the fastest decomposition level, whereas in Fig. 5, the same is represented for the slowest decomposition level. It is clearly seen how the dynamic thresholds are changed for each decomposition level while being the same metric. Also, the duration of the changes can be noticed between the decomposition levels, notice the flat tips in Fig. 5 in contrast to the fast level.

The time label is really simple to assign thanks to the decomposition of the MRA, since each level is related to a certain duration, the label is directly selected based on the level of the detected degradation of the previous intensity label. In this case, three duration labels have been set: peaks, for degradation found in the fastest levels comprising one to four samples in duration; series, for degradation found in medium levels, comprising four to 16 samples in duration; or tendency, for those anomalies found in the slowest level, with effects between 16 and 32 samples.

TABLE 1. Example of linguistic labels generated for a random sample describing the intensity, duration period and associated metric of the degradation.

Intensity	Duration		Metrics
Very high High Low	peak peak series	degradation on degradation on degradation on	Drops Connected users Uplink users

Colors are also assigned to each label to facilitate a faster understanding of the labels, as an example, a label assigned to a random portion with drops caused by an excess in user connections results in Table 1.

Such as the sample labels, generic labels could be generated for each cluster taking the most repeated labels in its portions, serving as a textual representation of the cluster behavior.

To keep a reduced number of labels, high intensity changes are prioritized over lower intensity ones, having chosen in this case to prioritize peaks over long duration issues, although this priority could be easily modified. This way, a single label per metric is the maximum number of labels per sample or group.

VIII. EVALUATION

To test the proposed method, a dataset encompassing 16 unique network metrics collected from 1000 cells over a 30-day span was utilized. The list of metrics is presented in Table 2. Annoyingly, the Signal to Interference and Noise Ratio (SINR) of the Downlink (DL) was not available in the dataset which would have been useful to analyze compared to the packet loss in the DL.

The setup used for the implementation and testing of the system was a computer with an AMD Ryzen 5 5600X CPU with 12 threads, a NVIDIA RTX 3070 and 32GB of RAM. The system has been tested in python 3 making use of multiple libraries the most relevant ones being pywavelets, sklearn, scipy, numba, numpy and pandas.

TABLE 2. Set of measured metrics for the evaluation of the system.

Index	Metric
1.	Downlink Packet Loss Rate
6.	Uplink Packet Loss Rate
11.	Drops
16.	Spectral Efficiency
21.	Estab Attempts
26.	Estab Failures
31.	Estab Successes
36.	Peak Connected Users Per Cell
41.	Avg Connected Users
46.	Avg DL Active Users
51.	Avg UL Active Users
56.	Uplink Control SINR
61.	Uplink SINR
66.	Latency
71.	Downlink Cell Throughput
76.	Uplink Cell Throughput

To such metrics, normalization and wavelet decomposition were applied. Undertaken at five levels, it generates 5 parameters per metric, creating a total of 80 classification parameters differentiated by the value of the decomposed metric throughout a certain time period. As an example, the first metric would produce the following 5 parameters presented in Table 3, hence the numeration of the metrics on the previous table, Table 2.

TABLE 3. MRA 5 level decomposition of metric n°1 (Downlink packet loss rate).

Index	Metric
1.	Downlink Packet Loss Rate 1-2h
2.	Downlink Packet Loss Rate 2-4h
3.	Downlink Packet Loss Rate 4-8h
4.	Downlink Packet Loss Rate 8-16h
5.	Downlink Packet Loss Rate 16-32h

After these steps, the 80 parameters were processed by the ACNSFDP algorithm, looking for the number of clusters to generate. After computation, eight key samples were considered as initial group centers and proposed to the GMM clustering algorithm.

To facilitate a three-dimensional visualization of the results, Principal Component Analysis (PCA) [23] was employed for data dimensionality reduction. This technique reduces the number of variables whilst preserving the maximal amount of information. New variables, termed Principal Components (PCs), are constructed as linear combinations of the initial variables. These PCs account for the maximum variance in orthogonal directions. By selecting a subset of the most significant PCs —three in this instance— information loss in data representation is minimized.

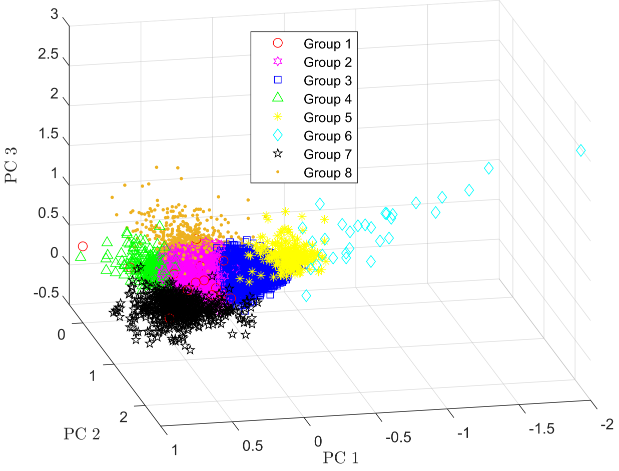


FIGURE 6. Principal Component Analysis of the resulting clustering.

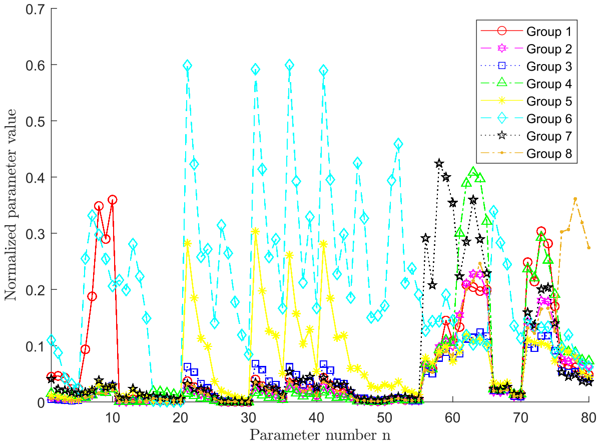


FIGURE 7. Mean of the parameter values of each group generated in the clustering process. There are eight different groups represented in different colors and 80 different parameters (n), result of the 5 level decomposition of the 16 available metrics.

In Fig. 6, the eight resulting clusters are shown, providing a visual representation of the groups' PCA. To understand the clustering algorithm's separation, the parameter values were analyzed by the mean of each cluster values, as depicted in Fig. 7. Upon examining the various groups, the following characteristics were noted thanks to the MRA five level decomposition:

1) Group 1 (in red), is composed by samples where the Uplink (UL) and DL packet loss rate are high, and, in the case of the UL packet loss, maintained during whole days (series and trend degradation). This could be explained by the higher traffic in the DL.
2) Group 2 (in magenta), is similar to the next group, but has slightly less users, with more traffic, and higher SINR in the UL. It does not have packet loss.
3) Group 3 (in blue), comprises samples with a medium number of connected users but with low traffic, there are no degradations except for a lower SINR in the UL.

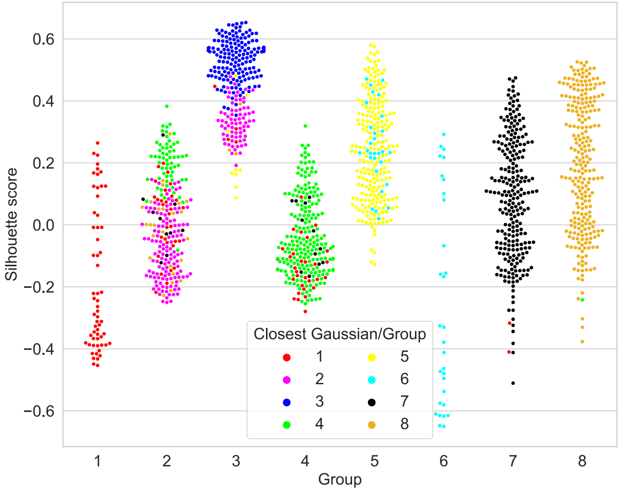


FIGURE 8. Results of the silhouette score of the generated groups and closest group indicated via a color map. (The represented points per cluster is limited to a maximum of 250 samples per cluster).

4) Group 4 (in green), represents samples with the biggest SINR values in the UL, with a duration of multiple hours and with high traffic, specially in the DL (similar to that of the first group). Spectral efficiency of this samples is the highest of all groups.
5) Group 5 (in yellow), comprises a group characteristic of this dataset with huge peaks of users of great intensity and short duration, in this case no issue is caused by these users since activity is kept low. A summary of the labels can be seen in Table 4.
6) Group 6 (in cyan) is the other user peak related group, but in this case the peak of users is maximum, with great activity (active users), causing clear degradation in the network, incremented packet loss rate in both UL and DL, drops, connection failures, huge increment in latency, etc... This group labels are also summed up in Table 4.
7) Group 7 (in black), contains samples characterized with big fluctuations in SINR of both channels of the UL (control and shared).
8) Group 8, (in orange) is composed by samples with intense use of the cell UL, with medium SINR fluctuations and typical values in the rest of metrics.

Note that in Table 4 only groups 5 and 6 are graphically represented, the underlying motivation of this representation is to remark the simplification of analysis achieved thanks to the resulting labels between an anomalous case (group 5), and a similar anomalous case with degradation (group 6), either graphically, as in this table, or linguistically.

In order to analyze the quality of the generated groups, Fig 8 is included. This figure shows the silhouette score (based on the squared euclidean distance) for the different clusters. A color mapping is used for the indication of the closest gaussian to each one of the samples based on the Mahalanobis distance as a second quality measurement.

In this quality analysis it can be appreciated that there are multiple low-scoring samples, this is probably due to the gradualness of the samples. For example, group six could be contemplated as an extreme case of group five due to the extreme number of connections evolving into drops and connectivity degradation. As such, many of the samples from group six score low, since they may be really close to group five. A similar thing happens with group one in relation to groups two and four, to whom it shares a traffic increase compared to, for example, group three (in Fig. 6 group one is actually behind these two groups).

The color also helps understand how the clustering is working. We can clearly notice the gradualness from groups three to two and two to four. Overall, the different groups characterize different situations but the boundaries of many of the groups are not clear due to the mentioned gradualness of metrics.

TABLE 4. Heatmap table of visual labels of metrics mean values in group 5 and 6. Metric and duration of the anomalies are presented with the color indicating anomaly intensity.

Color Code	Normal	🟨 Very Low	🟧 Low	🟫 High	🟥 Very high

Metric	Group 5				
	<2h	<4h	<8h	<16h	>16h
1.					
6.					
11.					
16.					
21.	🟥	🟫	🟥	🟨	🟧
26.					
31.	🟥	🟫	🟥	🟨	🟧
36.	🟫	🟧	🟫		🟨
41.	🟥	🟫	🟫	🟨	🟧
46.	🟨				
51.					
56.					
61.					
66.					
71.					
76.					

Metric	Group 6				
	<2h	<4h	<8h	<16h	>16h
1.	🟨	🟧			
6.	🟥	🟥	🟥	🟥	🟥
11.	🟥	🟥	🟧	🟥	🟥
16.					
21.	🟥	🟥	🟥	🟥	🟧
26.					
31.	🟥	🟥	🟥	🟥	🟧
36.	🟥	🟥	🟥	🟥	🟨
41.	🟥	🟥	🟥	🟥	🟫
46.	🟥	🟥	🟥	🟧	🟥
51.	🟥	🟥	🟥	🟥	🟥
56.	🟧	🟧	🟨		
61.					
66.	🟥	🟥	🟥	🟥	🟥
71.	🟨				
76.					

Information can be extracted just by analyzing the mean values of the metrics of the different groups. But a more complete and faster interpretation can be achieved by looking at the tables presented here, which are the direct results of the labeling process explained in Section VII.

This label can either be visually represented, like in Table 4, or based on text. As an example, in this case, the labels of groups 5 and 6 would result in Table 5 and Table 6.

TABLE 5. **Resulting linguistic labels containing the intensity, duration period and corresponding metrics of the detected degradations of group 5.**

Intensity	Duration		Metric
Very high	series	degradation on	Estab Attempts
Very high	series	degradation on	Estab Successes
High	series	degradation on	Peak Connected Users
Very high	trend	degradation on	Avg Connected Users
Very low	trend	degradation on	Avg DL Active Users

TABLE 6. **Resulting linguistic labels containing the intensity, duration period and corresponding metrics of the detected degradations of group 6.**

Intensity	Duration		Metric
Low	series	degradation on	DL Packet Loss Rate
Very high	peak	degradation on	UL Packet Loss Rate
Very high	peak	degradation on	Drops
Very high	peak	degradation on	Estab Attempts
Very high	peak	degradation on	Estab Successes
Very high	peak	degradation on	Peak Connected Users
Very high	peak	degradation on	Avg Connected Users
Very high	peak	degradation on	Avg DL Active Users
Very high	peak	degradation on	Avg UL Active Users
Low	series	degradation on	Uplink Control SINR
Very high	peak	degradation on	Latency
Very low	trend	degradation on	DL Cell Throughput

On the one hand, regarding the presented labels of group 5, it is easily observed that there is a really high number of user connections per cell (‘‘Estab attemps’’, ‘‘Estab successes’’ and ‘‘Peak of users’’) in samples contained in this group, and this vast amount of users is maintained over periods of more than 4h. Besides, the average number of connected users is really high during even longer periods of more than 16h, and the DL activity is slightly higher than usual for a similar time. The interesting thing of this group is that no issues are found due to this user anomaly, cells keep handling the large number of users without drops or connectivity interruptions.

On the other hand, we find group 6, whose labels clearly indicate the issues caused by a similar, yet greater, user anomaly. An immense peak of users of short duration (less than 2h) that causes packet loss in both UL and DL, drops and latency, all of this due to the increase in users and their activity in both UL and DL. The previous group also presented an increase in user amount, but not in user activity, which occurred over a longer period of time, avoiding service interruption or degradation.

IX. CONCLUSION

A system capable of identifying and classifying different daily network behaviors autonomously has been introduced in this work. The evaluation was carried out on real network metrics, leading to the classification of cell data into eight groups based on the behavioral patterns manifested in their metric values.

The algorithm that has been implemented affords a preliminary filtering of the samples, enabling a focused examination of groups demonstrating network degradation by experts. Moreover, just analyzing the mean values of each group of the different components generated, the expert can obtain information on the duration of the degradation based on the degraded level of the metric in question.

The system also allows the generation of linguistic labeling, thanks to the decomposition levels, that shows the duration and intensity of the degradation or changes in the metrics.

As future work, the implemented algorithm could be joined or applied with more complex detection and classification algorithms, for example those that work with labeled data (VAEs, NNs…), to see if the application of the decomposition and the generated labels facilitates the task of classifying different issues.

REFERENCES

[1] M. R. Ahasan, M. S. Haque, and M. G. R. Alam, ‘‘Supervised learning based mobile network anomaly detection from key performance indicator (KPI) data,’’ in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jul. 2022, pp. 1–6.

[2] S. Wang, M. Lu, S. Kong, J. Ai, J. Wang, and W. E. Wong, ‘‘Anomaly detection via kpis for software performance failures,’’ SSRN, 2022. Accessed: Mar. 11, 2023, doi: 10.2139/ssrn.4054805.

[3] S. Fortes, P. Muñoz, I. Serrano, and R. Barco, ‘‘Transform-based multiresolution decomposition for degradation detection in cellular networks,’’ *Sensors*, vol. 20, no. 19, p. 5645, Oct. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/19/5645

[4] M. Ahmed, A. Naser Mahmood, and J. Hu, ‘‘A survey of network anomaly detection techniques,’’ *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804515002891

[5] Y. Lu, J. Wang, M. Liu, K. Zhang, G. Gui, T. Ohtsuki, and F. Adachi, ‘‘Semi-supervised machine learning aided anomaly detection method in cellular networks,’’ *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8459–8467, Aug. 2020.

[6] B. Hussain, Q. Du, and P. Ren, ‘‘Deep learning-based big data-assisted anomaly detection in cellular networks,’’ in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6, doi: 10.1109/GLOCOM.2018.8647366.

[7] B. Hussain, Q. Du, A. Imran, and M. A. Imran, ‘‘Artificial intelligence-powered mobile edge computing-based anomaly detection in cellular networks,’’ *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 4986–4996, Aug. 2020.

[8] S. Fortes, C. Baena, J. Villegas, E. Baena, M. Z. Asghar, and R. Barco, ‘‘Location-awareness for failure management in cellular networks: An integrated approach,’’ *Sensors*, vol. 21, no. 4, p. 1501, Feb. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1501

[9] J. Villegas, E. Baena, S. Fortes, and R. Barco, ‘‘Social-aware forecasting for cellular networks metrics,’’ *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1931–1934, Jun. 2021.

[10] S. Zhang, Z. Zhong, D. Li, Q. Fan, Y. Sun, M. Zhu, Y. Zhang, D. Pei, J. Sun, Y. Liu, H. Yang, and Y. Zou, ‘‘Efficient KPI anomaly detection through transfer learning for large-scale web services,’’ *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2440–2455, Aug. 2022.

[11] G. Yu, Z. Cai, S. Wang, H. Chen, F. Liu, and A. Liu, ‘‘Unsupervised online anomaly detection with parameter adaptation for KPI abrupt changes,’’ *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 3, pp. 1294–1308, Sep. 2020.

[12] N. Zhao, J. Zhu, Y. Wang, M. Ma, W. Zhang, D. Liu, M. Zhang, and D. Pei, ‘‘Automatic and generic periodicity adaptation for KPI anomaly detection,’’ *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 3, pp. 1170–1183, Sep. 2019.

[13] K. Buthelezi, M. Kabeya, and M. Leoaneka, ‘‘A review of fault location algorithms utilising travelling wave, wavelet transform and multi-resolution analysis techniques,’’ in *Proc. 30th Southern Afr. Univ. Power Eng. Conf. (SAUPEC)*, Jan. 2022, pp. 1–6.

[14] A. M. Molaei, S. Hu, V. Fusco, and O. Yurduseven, "A multi-resolution analysis-based approach to accelerate data acquisition for near-field MIMO millimeter-wave imaging," in *Proc. SPIE*, Jun. 2022, pp. 90–101.

[15] A. Gudigar, U. Raghavendra, T. R. San, E. J. Ciaccio, and U. R. Acharya, "Application of multiresolution analysis for automated detection of brain abnormality using MR images: A comparative study," *Future Gener. Comput. Syst.*, vol. 90, pp. 359–367, Jan. 2019.

[16] X. Xie, X. Li, L. Xu, W. Ning, and Y. Huang, "HaarAE: An unsupervised anomaly detection model for IoT devices based on Haar wavelet transform," *Int. J. Speech Technol.*, vol. 53, no. 15, pp. 18125–18137, Aug. 2023, doi: 10.1007/S10489-023-04449-Z.

[17] S. Xie, L. Li, and Y. Zhu, "Anomaly detection for multivariate time series in IoT using discrete wavelet decomposition and dual graph attention networks," *Comput. Secur.*, vol. 146, Nov. 2024, Art. no. 104075.

[18] D. B. Percival and A. T. Walden., *Wavelet Methods for Time Series Analysis* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge, U.K.: Cambridge Univ. Press, 2000.

[19] J. Wang, Y. Zhang, and X. Lan, "Automatic cluster number selection by finding density peaks," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. 2016, pp. 13–18.

[20] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1242072

[21] E. Patel and D. S. Kushwaha, "Clustering cloud workloads: K-means vs Gaussian mixture model," in *Proc. 3rd Int. Conf. Comput. Netw. Commun. (CoCoNet'19)*, 2020, pp. 158–167. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920309820

[22] T. Su and J. G. Dy, "In search of deterministic methods for initializing K-means and Gaussian mixture clustering," *Intell. Data Anal.*, vol. 11, no. 4, pp. 319–338, Jul. 2007.

[23] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Principal component analysis," *Nature Methods*, vol. 14, pp. 641–642, Jul. 2017.

JUAN CANTIZANI-ESTEPA was born in Lucena, Andalucía, Córdoba, Spain, in November 1997. He received the B.Sc. degree in electronics, robotics and mechatronics engineering and the M.Sc. degree in mechatronics engineering from the University of Málaga (UMA), in 2019 and 2020, respectively, where he is currently pursuing the Ph.D. degree in communications engineering. His main research interests include mobile communication networks and machine learning techniques as well as the IoT networks.

SERGIO FORTES (Senior Member, IEEE) received the M.Sc. and Ph.D. degree in telecommunication engineering from the University of Málaga. He began his career being part of main European space agencies (DLR, CNES, and ESA) and Avanti Communications plc, where he participated in various research and consultant activities on broadband and aeronautical satellite communications. In 2012, he joined the University of Málaga, where his research is focused on self-organizing networks for cellular communications, the IoT, and aerospace applications. He is currently an Associate Professor with the University of Málaga.

JAVIER VILLEGAS received the degree in telecommunications systems engineering and the M.Sc. degree in telecommunication engineering and in telematic engineering from the University of Málaga, Spain, where he is currently pursuing the Ph.D. degree with the Department of Communications Engineering. He is a Research Assistant with the Department of Communications Engineering, University of Málaga.

JAVIER RASINES received the bachelor's degree in electrical and computer engineering from the Universidad Politécnica de Madrid and the M.Sc. degree in embedded systems, robotics, control theory, AI, and software development from the Kungliga Tekniska Högskolan (KTH). He began his career as an Embedded System Engineer with CSIC. He joined Ericsson, in 2018, with a focus on the development of AI systems. He is currently a Data Scientist with Ericsson with more than five years of experience.

RAÚL MARTÍN CUERDO began his career as an Access Network Engineer. In 2008, he joined Ericsson with a focus on RAN design and optimization. He is currently a ML and AI Product Development Leader with Ericsson with more than 20 years of experience in the sector. He holds the title of a Telecommunications Engineer with the Universidad Politécnica de Madrid.

RAQUEL BARCO is currently a Full Professor of telecommunication engineering with the University of Málaga. Before joining the university, she was with Telefonica and the European Space Agency (ESA). As a Researcher, she is specialized in mobile communication networks and smart-cities, having led projects funded by several million euros, published more than 100 papers in high impact journals and conferences, authored five patents, and received several research awards.