Abnormal Behavior Detection Based on Traffic Pattern Categorization in Mobile Cellular Networks

Jonathan M. DeAlmeida, *Student Member, IEEE*, Camila F. T. Pontes, Luiz A. DaSilva, *Fellow, IEEE*, Cristiano B. Both, Joao J. C. Gondim, Celia G. Ralha, Marcelo A. Marotta, *Member, IEEE*

***Abstract*—Abnormal behavior in mobile cellular networks can cause network faults and consequent cell outages, a major reason for operational cost increase and revenue loss for operators. Nonetheless, network faults and cell outages can be avoided by monitoring abnormal situations in the network and acting accordingly. Thus, anomaly detection is an important component of self-healing control and network management. Network operators may use the detected abnormal behavior to quantify numerically their intensity. The quantification of abnormal behavior assists the characterization of potential regions for infrastructure updates and to support the creation of public policies for local connectivity enhancements. We propose an unsupervised learning solution for anomaly detection in mobile networks using Call Detail Records (CDR) data. We evaluate our solution using a real CDR data set provided by an Italian operator and compare it against other state-of-the-art solutions, showing a performance improvement of around 35%. We also demonstrate the relevance of considering the distinct traffic patterns of diverging geographic areas for anomaly detection in mobile networks, an aspect often ignored in the literature.**

***Index Terms*—5G, next generation networks, anomaly detection, Call Detail Record (CDR), self-healing control, network analytics, traffic pattern analysis, Artificial Intelligence (AI).**

I. Introduction

CELLULAR networks increasingly incorporate support for Artificial Intelligence (AI)-based network management and control. It is possible, for example, to employ AI-enhanced applications to collect network operation data and perform self-healing control using the analytics function capabilities present in 5G and beyond (B5G) [1]. These network management tools seek to identify unexpected or abnormal behaviors, reflecting anomalies in the network operation [2], and to take appropriate action. Anomalies trigger the escalation of network faults and consequent cell outages, which is one of the major reasons for operational cost increase and revenue loss for operators [3]. Detecting and filtering anomalies can also enhance the performance of cellular traffic predictions [4]. In this context, anomaly detection has emerged as an important research direction for 5G/B5G [5][6].

Network faults and cell outages can be avoided by the timely detection of anomalies and responsive actions from the network (*e.g.,* unmanned aerial vehicles have been proposed for mobile service provisioning to deal with demand hotspots in urban environments [7]). For anomaly detection, AI-driven solutions using Machine Learning (ML) algorithms [8][9][10] and statistical models [11][12] based on entropy have been proposed. Most solutions are developed on inferred models to identify the occurence of anomalies based on training procedures and statistical analysis of simulated data or real data traces. Nonetheless, mobile networks are complex systems, making it challenging to devise effective predictive models due to several uncontrollable influencing factors, such as user mobility and behavior, and channel quality.

Recently, ML has been applied to detect abnormal behavior in mobile networks, relying on subscriber Call Detail Record (CDR) data recorded by the network operator for billing purposes. Several publications have been exploring the usage of ML in the context of anomaly detection and cellular traffic prediction [13][4][14][15]. However, the practical application of ML-based solutions in this space is challenging: most works in the literature assume that different geographic areas observe similar traffic patterns and that labeled data is always available to train the model. These assumptions do not hold in real network scenarios: traffic patterns from distinct regions vary significantly [10][16] and the data is usually not labeled [8]. Additionally, ML-based solutions in the literature classify anomalies in a binary manner, without considering that such anomalies vary in degree. This article advances the state-of-the-art by exploiting AI-driven statistical solutions to classify anomalies in mobile networks and assess their severity, without requiring pre-labeled data.

We propose an unsupervised statistical solution, called Energy-based Flow Classifier (EFC), which uses a classifier based on the inverse statistics of the Potts model (a generalization of the Ising model [17]) for anomaly detection in CDR data. The solution is evaluated using real traffic data from the city of Milan made available by Telecom Italia. Additionally, we compare the performance of our proposed solution against state-of-the-art solutions that use the K-means algorithm [18] for anomaly detection.

Our main contributions include:

- A flexible and unsupervised AI/ML-driven solution, using CDR data, based on the inverse statistics for anomaly detection in mobile networks.

Manuscript submitted on April 12, 2021. Major revisions submitted on September 5, 2021. Accepted on October 29, 2021.
This work was conducted with partial financial support from the National Council for Scientific and Technological Development (CNPq), Brazil, under grant numbers 311301/2018-5 and 130555/2019-3. It was also supported by the Commonwealth Cyber Initiative (CCI).
Jonathan M. DeAlmeida, Camila F. T. Pontes, João J. C. Gondim, Celia G. Ralha, and Marcelo A. Marotta are with the University of Brasilia, Brazil. (e-mail: jonathanalmd@gmail.com, cftpontes@gmail.com, gondim@unb.br, ghedini@unb.br, marcelo.marotta@unb.br)
Luiz A. DaSilva is with the Commonwealth Cyber Initiative, at Virginia Tech, United States of America. (e-mail: ldasilva@vt.edu)
Cristiano B. Both is with University of Vale do Rio dos Sinos, Brazil. (e-mail: cbboth@unisinos.br)

Table I
RELATED WORK COMPARISON. *GROUND TRUTH VALIDATION BASED ON SUPPOSITIONS

Reference	CDR data	Requires pre-labeling	Ground truth validation	Traffic pattern categorization	Traffic pattern evaluation	Evaluates degrees of anomaly
[3]	Real	✓	✗	✗	✗	✗
[8]	Real	✓	✗	✗	✗	✗
[10]	Real	✗	✓*	✓	✗	✗
[19]	Real	✓	✗	✗	✗	✗
[20]	Real	✓	✗	✗	✗	✗
[4]	Real	✗	✗	✗	✗	✗
[13]	Real	✗	✓*	✗	✗	✗
[21]	Simulated	✓	✗	N/A	N/A	✗
[22]	Real	✓	✗	✗	✗	✗
This work	**Real**	✗	✓	✓	✓	✓

- Ground truth validations relying on real traffic data to demonstrate our solution's performance and flexibility compared against the state-of-the-art solutions.
- An empirical evaluation of the impact of traffic pattern classification of distinct geographic regions on the performance of anomaly detection solutions.

This article is structured as follows. Section II summarizes the state-of-the-art solutions in anomaly detection based on CDR data. Section III introduces our AI-statistical anomaly detection solution for mobile networks. Section IV describes our research methodology and use case. Section V presents and discusses the experimental results. Section VI highlights conclusions and future work directions.

II. RELATED WORK

In this section, we discuss the anomaly detection state-of-the-art for mobile networks management systems, focusing on AI/ML-based solutions rather than heuristic implementations, *i.e.,* solutions based on predefined and specific rules created ad hoc by experts. The discussion here results from an extensive literature review from which we selected the most relevant works, summarized in Table I. Readers can refer to Ahmed *et al.* [23] for an exhaustive literature survey of network anomaly detection techniques.

We focus on work that applies AI in anomaly detection to manage mobile networks using CDR data. The vast majority of these studies use the Milano Grid CDR data set provided by Telecom Italia [24]. In particular, supervised/semi-supervised learning algorithms, such as deep neural networks and their variations, are predominant among those ML-based solutions. For example, Trinh *et al.* [20] present an approach using long short-term memory neural networks. Hussain *et al.* [22] [3] present a solution using feed-forward deep neural networks and later introduce a more scalable solution using deep convolutional neural networks [8]. Jaffry *et al.* [19] present a semi-supervised statistical-based solution using a Gaussian distribution.

One limitation of these approaches is that supervised learning solutions require labeled data, which usually is not available for real mobile network scenarios, as manual labeling is impractical for large areas [8][10]. This problem also occurs in semi-supervised learning solutions, since those approaches assume that the labels created do reflect the ground truth without further validation. In particular, they consider that anomalies detected based on the Euclidean distance are 100% true and use them for training the supervised model. In contrast to those approaches, our AI-driven statistical solution is capable of analyzing anomalies without requiring artificial data labeling.

Solutions based on unsupervised learning that do not require labeled data have also been proposed in recent works. Parwez, Rawat, and Garuba [13] present a solution using K-means and hierarchical clustering. Similarly, Sultan, Ali, and Zhang [4] apply K-means clustering for anomaly detection in different CDR data sets, focusing on eliminating anomalies from the data to improve traffic predictions without presenting any ground truth validation, *i.e.,* their focus is on removing outliers to improve the traffic prediction. This same approach could be used in the context of cellular traffic prediction presented by Zhang *et al.* [15], which mentions that the abnormal traffic volume is hard to predict, which negatively impacts the traffic predictor's performance. Zhu and Sun [10] also explore traffic pattern classification in different regions of Milan using K-means clustering.

Although the authors highlight the relevance of considering each region's general traffic behavior before detecting anomalies, they do not evaluate the effects of observing the different traffic patterns present in the city.

Conversely, based on ground truth validations, we empirically illustrate the importance of considering different traffic patterns when detecting mobile traffic network anomalies. Moreover, all the proposed solutions give a hard-decision result indicating whether there is an anomaly or not. In contrast, our solution combines AI and statistical analysis to both detect anomalies and assess their severity.

Papadopoulos, Drosou, and Tzovaras [21] present a statistical solution based on entropy for anomaly detection in mobile networks. The authors evaluate their solution as a strict binary classification of anomalies using labeled data. In this context, we employ a statistics-based solution with enhanced processing capabilities adapted from quantum mechanics. Our solution detects levels of anomalies presented across traffic samples collected at different times and in different locations.

Like most of the recent works mentioned in this section, we have also used the CDR data set from Telecom Italia. Although

from 2013, this data set is still one of the most popular open data sets available from a real telecommunication operator. We designed an unsupervised anomaly detection algorithm that trains a model to detect abnormal cell activity using traffic pattern categorization. Table I summarizes the main characteristics of each aforementioned work and positions our solution in the current literature. Unlike the previous works, our AI-statistical model can detect different levels of anomalies considering the characterized behavior of each region and type of day.

Lastly, it is worth mentioning the importance of autoencoders in the context of anomaly detection in big data applications. This technique is well suited for applications that deal with a huge amount of data, reducing the dimensionality of data and detecting outliers (or anomalies). Considering that the use case presented in Section IV is not characterized as a big data solution, we focus on comparing our proposed methods against a directly related state-of-the-art solution. The reader can refer to [25] for more details on applications and limitations of autoencoders. We describe our solution in the following section.

III. Energy-based Flow Classifier for Anomaly Detection in Mobile Networks

A management system can gather information from the mobile network substrate, measuring traffic in different regions along the time. In our work, this measured traffic can be characterized and modeled using intuitions from quantum mechanics, the Potts model [26] (Figure 1). The reader can refer to the work of Pontes *et al.* [27] to apply the same model in the context of anomaly detection in network intrusion detection systems. Traffic data is represented by a specific graph configuration $G_k(\eta,\varepsilon)$. Instead of particle spins, to produce a statistical model for the detection of cellular traffic anomalies, we take into consideration the amount of traffic measured in each time T, in all regions of interest. Let $(T_1,...,T_N)$ be an N-tuple of time slots during a day, and let $s_{d,sq} = (t_1,...,t_N)$ be a sample of discretized traffic volumes collected on date d, in region sq, where v_i is the discretized amount of traffic during a time slot T_i, $i = 1,...,N$. The continuous values of traffic volume present in the data set are discretized in Q quanta, coded in the alphabet $\Omega = \{1,...,Q\}$. Meanwhile, $\varepsilon = \{(i,j)|i,j \in \eta; i \neq j\}$ is the set of edges determined by all possible pairs of timeslots, creating a fully meshed graph that can represent different traffic samples through their common features. Each edge has an associated coupling value determined by the function $e_{ij}(v_i,v_j)$. Depending on how many quanta are considered for quantization, the loss of resolution may impact results. The number of quanta needed depends on the application.

Before calculating the couplings, single and joint empirical frequencies $f_i(v_i)$ and $f_{ij}(v_i,v_j)$ are obtained from the set of all training samples $s_{d,sq} \in S_{training}$ by counting occurrences of a given traffic volume v_i or traffic volume pair (v_i,v_j), respectively, and dividing by the total number of samples in $S_{training}$. Since the set $S_{training}$ is finite and small compared to the universe of all possible traffic volume distributions during one day, inferences based on $S_{training}$ are subject to undersampling effects. Following the theoretical framework proposed by Morcos *et al.* [28], we add pseudocounts to empirical frequencies to limit undersampling effects by performing the following operations:

$$f_i(v_i) \leftarrow (1-\alpha) f_i(v_i) + \frac{\alpha}{Q}, \tag{1}$$

$$f_{ij}(v_i,v_j) \leftarrow (1-\alpha) f_{ij}(v_i,v_j) + \frac{\alpha}{Q^2}, \tag{2}$$

where $(v_i,v_j) \in \Omega^2$, $0 \leq \alpha \leq 1$ is a parameter defining the weight of the pseudocounts, and Q stands for the cardinality of Ω. The introduction of pseudocounts is equivalent to assuming that $S_{training}$ is extended with a fraction of traffic with uniformly sampled features.

The direct couplings between each traffic volume at each time period are calculated as follows:

$$e_{ij}(v_i,v_j) = -(C^{-1}_{i,j})(v_i,v_j), \tag{3}$$
$$\forall (i,j) \in \{1,...N\}^2, \forall (v_i,v_j) \in \Omega^2, v_i,v_j \neq Q,$$

where

$$C_{ij}(v_i,v_j) = f_{ij}(v_i,v_j) - f_i(v_i) f_j(v_j) \tag{4}$$

is the covariance matrix obtained from single and joint empirical frequencies. Taking the inverse of the covariance matrix is a well known procedure in statistics to remove the effect of indirect correlation in data [29]. It is important to clarify that the number of independent constraints considered for the inference of this statistical model is $\frac{N(N-1)}{2}(Q-1)^2 + N(Q-1)$, even though the model has $\frac{N(N-1)}{2}Q^2 + NQ$ parameters. Without loss of generality, we set:

$$e_{ij}(v_i,Q) = e_{ij}(Q,v_j) = 0. \tag{5}$$

This corresponds to a situation where all couplings are measured for state Q. The number of remaining parameters now matches the number of constraints, and the solution of the maximum-entropy model is unique [28]. Hence, it is not necessary to calculate $e_{ij}(v_i,v_j)$ when v_i or v_j is equal to Q.

Lastly, we can now use these direct couplings to calculate the Hamiltonian "energy" of a given sample $s_{d,sq} = (t_1,...,t_N) \in S_{testing}$ in the testing set at a given timeslot T_i:

$$\mathcal{H}_i(s_{d,sq}) = -\sum_{j|i\neq j} e_{ij}(v_i,v_j), i = 1,...,N. \tag{6}$$

This "energy" will be inversely proportional to the probability of sample $s_{d,sq}$ presenting an anomaly at timeslot T_i. We consider that a given test sample abnormal at timeslot T_i if the energy is above the threshold:

$$\tau_1 = \mathcal{H}_i(s_{d,sq}) > avg(\mathcal{H}_i(s_{d,sq})) + 4std(\mathcal{H}_i(s_{d,sq})), \tag{7}$$

of all $s_{d,sq} \in S_{training}$ [19].

Finally, the complexity terms are summarized in Table II. The classifier training complexity is

$$O(K'[N + M^3Q^3 + NM^2Q^2]),$$

where N is the number of instances (samples of the training set), K' is the number of employed regions, M is the number of features and Q is the size of the alphabet. Meanwhile, the complexity for the classification phase is quadratic on the number of used features M and linear for the test data set size

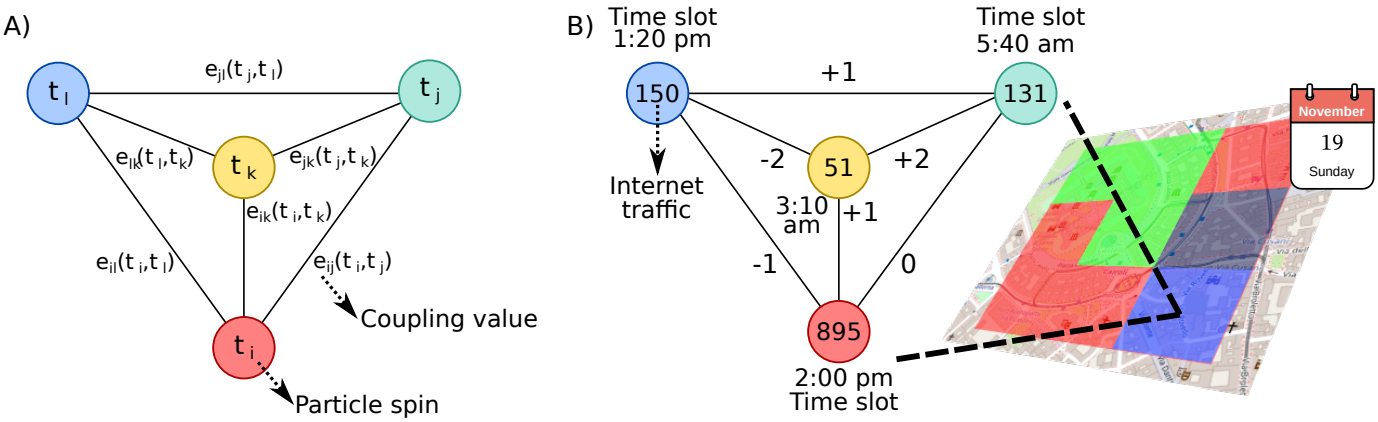


Figure 1. EFC intuition - A) Interacting spins on a crystalline lattice. B) Network traffic mapped into a graph structure.

N and for the number of employed regions K':

$$O(KNM^2).$$

In our work, features stand for traffic volume and timestamp. Therefore, the number of features, as well as the classification time, can be kept small. K-means, the most representative algorithm in the literature used for anomaly detection in mobile networks, presents a linear complexity in the number of data objects:

$$O(KN),$$

where K is the number of clusters and N is the number of data objects.

Table II
COMPLEXITY TERMS DESCRIPTION

Term	Complexity
EFC Training phase	$O(K'[N+M^3Q^3+NM^2Q^2])$
EFC Classification phase	$O(NM^2)$
K-means	$O(KN)$

When the city is divided into different regions, each one containing clusters of squares with similar behavior, one statistical model is inferred for each context, *e.g.,* cluster 1 during weekdays, cluster 2 during weekends, etc. To infer each one of these different models, only samples belonging to that context are included in the training set. We employ k-fold cross-validation for each region (cluster), where k equals 10. The full dataset is partitioned into ten different disjoint training sets to avoid overfitting. Each set contains squares of one of the five regions (clusters) either on weekdays or weekends.

Observing the energy level resulting from the model, it is possible to differentiate the type of anomaly occurring, from a severe cell outage to under-usage. Our solution presents quadratic computational complexity and a high degree of flexibility in considering different types of anomalies. It also provides flexibility in terms of cutoff definition, since the anomalies can be ranked and the top-k (in both ends of the spectrum, *i.e.,* increased traffic volume, or decreased traffic volume) can be selected by the network manager to focus on. Though ML-based and statistical solutions can potentially detect anomalies in the entire region of interest, the inferred model can be further improved by training that considers regions with similar behavior.

As characterized by De Almeida *et al.* [16] and Zhu and Sun [10], the traffic activity varies considerably across urban, suburban, and rural areas. Therefore, we combine our solution with an AI clustering algorithm to detect degrees of anomalies considering temporal and spatial traffic patterns. Spectral, mean shift, agglomerative hierarchical and K-means clustering are examples of algorithms that can be used for characterizing traffic patterns. In summary, our AI-statistical solution has an unsupervised ML component to characterize and consider the region and time of day before detecting anomalies. We provide more details regarding this approach and a use case in the next section.

IV. METHODOLOGY AND USE CASE

In this section, we present our methodology and the use case employed to evaluate our solution. First, we provide information regarding the data set employed in our experiments, the metrics to evaluate our results, and the data we used to establish the ground truth. Next, we outline the unsupervised learning approach used for traffic pattern classification for region clustering. Lastly, we describe how our solution integrates the network's abnormal behavior detection and the traffic pattern classification components.

A. Scenario and Data Set Description

The data set consists of CDRs gathered from the management system of the Telecom Italia network in Milan and currently available in the Harvard Dataverse [24].

Telecom Italia has provided a *geojson* file that specifies each region covered by the network along with the CDR data [30]. In this case, Milan is represented as a 100×100 square grid comprised of squares with a side of 0.235 Km and an area of 0.055 Km^2, as depicted in Figure **??**. According to their coordinates, we have mapped each cell's spatial locations using the *geojson* file for this data set. It is worth mentioning that the spatio-temporal CDR data does not specify the traffic activity units, to remove business-sensitive information. Nonetheless,

the activities present in the CDR data are proportional to the amount of real traffic. The exact location and mobile phone number are also omitted to preserve the users' privacy. Readers can refer to Barlacchi *et al.* [31] for more details on how the CDR data is obtained.

This data set contains activity from Telecom Italia's subscribers in Milan, generated and collected for 62 days, from November 1, 2013, to January 1, 2014. The data set is temporally split into 10-minute slots (Table III describes each feature available in the data set). We extract the following features from the CDRs: ID (an integer that identifies each cell), timestamp (raw time stamp recorded in milliseconds), and Internet activity by subscribers. The time granularity is relevant when detecting abnormal behavior in networks since it will make the algorithm's outcome more precise and, consequently, render the actions to deal with cell outages more effective. Using such a low time granularity increases the execution time and processing power requirements, emphasizing the need for a low complexity solution.

B. Data Pre-processing

Since the data set provided is available in a raw format, we must first pre-process the data before applying our solution. The data is stored in separate files, for a total of 62 files, one for each day. We read the files for each day and convert the timestamps units from milliseconds to minutes, aggregating the CDR records, resulting in each cell's total activity during a time slot of 10 minutes. Moreover, we classify days into weekends or weekdays and create an additional column for this feature in the data set. We identify clusters for this spatially distributed data using a clustering algorithm, considering each cell's position and aggregated traffic.

Different from previous works, we do not use the clustering algorithm to detect anomalies. Instead, we employ K-means clustering to characterize and evaluate each region's traffic pattern. To implement the clustering algorithm, we use the kmeans++ initialization strategy to choose initial points that are as distant from each other as possible. To select the optimal number of clusters (the value of K), we use the Elbow test [32]. This test is based on the within-cluster sum of squares (WSS), which indicates the cohesion within each group by evaluating the algorithm with different K values. According to the Elbow method, K should be selected considering that adding a cluster does not significantly decrease the average WSS. Finally, the result of this approach is a geographical map divided into clusters, classified as urban/suburban from Milan city center and its surrounding sub rural/rural areas.

To separate the different regions based on their Internet traffic behavior, we employ the following feature vectors from 2,500 squares (separated in time slots of 10 minutes) in the clustering process: Internet activity, activity time, and position (x-axis and y-axis). We apply the clustering algorithm with the Euclidean distance to categorize each regions' traffic patterns. As an output of the clustering algorithm, Figure 2 depicts different clustered regions, in which the number of clusters ($K = 5$) is based on the Elbow test. Each square is colored according to its cluster. Moreover, the traffic intensity and variation are higher for clusters 1 and 2 and lower for clusters 3 to 5 (representing the suburban/rural area around Milan). Clusters 1 to 5 comprise 132, 364, 638, 905, and 1287 squares, respectively. The data traffic heatmap that served as input to the clustering algorithm is presented in the blue image in the background. Considering these clusters, we train our solution and assess its performance based on the ground truth data.

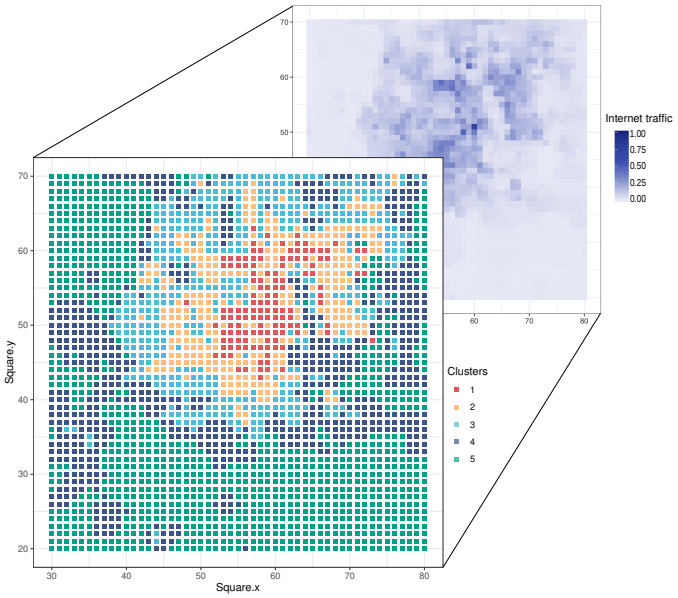


Figure 2. Internet activity from Milan is categorized into five different groups of traffic patterns. The heatmap for the normalized Internet activity around Milan is shown in the background with shades of blue.

C. Ground Truth

Before assessing our solution's performance, we must first determine how to recognize a correct detection of abnormal behavior (true positive) in the network environment. To do so, we analyzed information from a particular area of Milan to obtain ground truth regarding periods when large abnormal behaviors in network traffic are believed to have occurred. The area selected houses Italy's largest stadium, the Stadio Giuseppe Meazza, commonly known as San Siro Stadium. Its capacity is 80,018, and it is used by two of the biggest soccer teams from Italy (AC Milan and FC Internazionale Milano), as well as by the Italian national team. Major events in the stadium between November 1 and December 23, 2013, are presented in Table IV. Almost all events are large soccer matches, with an estimated attendance varying between 12,000 and nearly 80,000 (close to the stadium's capacity). Besides soccer matches, one small event occurred on November 24, a protest by the Milan "Ultras", the largest group of hardcore fans among the AC Milan supporters [33]. The protest occurred due to a draw between AC Milan and a small team on the day before, on November 23.

It is worth mentioning that the occurrence of soccer matches and other events (*e.g.,* protests and concerts) in the stadium are not fixed, *i.e.,* it will not necessarily occur every week on the same day or periods. Moreover, although the soccer matches are human comprehensible and predictable by checking calendars and other sources of information, the same cannot be

Table III
DATA SET: FEATURES AND DESCRIPTION

Feature	Description
ID	Integer to identify each cell
Timestamp	Raw timestamp recorded in milliseconds within an interval of 10 minutes
Country code	Integer to identify the user's country code
Internet Activity	Indicates the Internet usage by users in a grid within a time slot of 10 minutes
Inbound SMS Activity	Indicates the incoming SMS activity from users in a grid within a time slot of 10 minutes
Outbound SMS Activity	Indicates the outgoing SMS activity from users in a grid within a time slot of 10 minutes
Inbound Call Activity	Indicates the incoming Call activity from users in a grid within a time slot of 10 minutes
Outbound Call Activity	Indicates the outgoing Call activity from users in a grid within a time slot of 10 minutes

stated for an AI/Statistical algorithm solely considering CDR data without any other source of input. Therefore, using the stadium region for validation purposes is adequate and broadly employed in recent literature [10] [13].

The city of Milan was divided into five areas by applying the aforementioned clustering algorithm with a traffic pattern classification approach. For each grouped area, we use our EFC solution to detect anomalies in the CDR data. After detection, we analyze the anomalies that occurred in the San Siro area, including the metro station close to the stadium. In this case, anomalies detected on the days specified in Table IV are considered true positives in our ground truth validation. Conversely, detected anomalies on a day without any event are considered false positives. False negatives and true negatives occur if there are no detected anomalies on an event day and a day without events, respectively.

Table IV
GROUND TRUTH INFORMATION FROM SAN SIRO STADIUM [34]

Date	Event	Part of the day	Attendance
Nov. 2	National soccer match	Evening	44,261
Nov. 9	National soccer match	Evening	38,933
Nov. 15	International soccer match	Evening	>40,000
Nov. 23	National soccer match	Evening	34,848
Nov. 24	Protest	Evening	>400
Dec. 1	National soccer match	Afternoon	43,706
Dec. 4	National soccer match	Evening	12,714
Dec. 8	National soccer match	Evening	33,732
Dec. 11	International soccer match	Evening	61,744
Dec. 16	National soccer match	Evening	37,987
Dec. 22	National soccer match	Evening	79,311

To provide additional results using a different region, we evaluated our solution in the Mediolanum Forum, an indoor sports arena that is located in Assago (near Milan, Italy) [35]. With a seating capacity of 12,700, the arena is the home of the Italian League professional basketball team Pallacanestro Olimpia Milano. The Mediolanum Forum is used for different activities, such as ice hockey, basketball, tennis, and live concerts. Additionally, there is a metro station right next to the arena. Therefore, it is a great region to evaluate potential abnormal behaviors in the network traffic. It is important to mention that the arena is used in a broad range of activities. Thus, we were not able to precisely map every single event that occurred there between November and December 2013 as we were able to do with the San Siro stadium. However, to provide more pieces of evidence of our solution's coverage, we will present our results considering some of the confirmed events that occurred in the Mediolanum Forum, without comparing performance metrics.

D. Experiment Design and Performance Metrics

The total number of time slots in which abnormal traffic activity was detected varied from zero to almost 5,000. Figure 3 presents a heatmap for the detected anomalies in each region of the city during the 52 days. The validation presented in the next section is based on the regions from squares highlighted in green inside the region outlined by the red dashed lines in Figure 3 (San Siro Stadium area in squares 5737 and 5638, and the metro station located between squares 5639 and 5640).

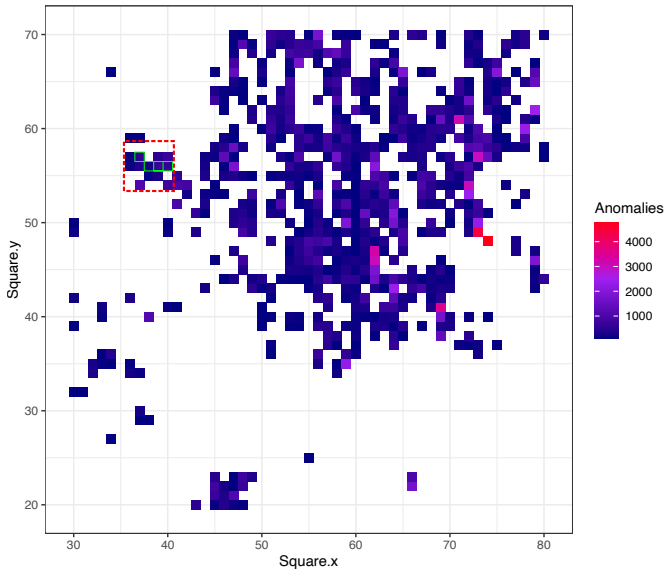


Figure 3. Heatmap for the number of detected anomalies between November 1 and December 22, 2013, in Milan filtering outliers (squares with only one abnormal 10-minute slot in the entire period).

For performance evaluation, it is important to mention that we employ cross-validation (10-fold) to prevent model overfitting, which occurs when a function strongly fits the training process's data. To compare our approach, EFC, with a state-of-the-art solution, we also perform experiments with solutions proposed by Zhu and Sun [10] and Parwez, Rawat, and Garuba [13], which are based on K-means clustering used

as an anomaly detection algorithm. To avoid biasing and to make the results fair for both EFC and the solutions from [10] and [13], the algorithms are trained and tested with clustered regions output from the clustering algorithm and also without considering them.

For each solution, we compare the accuracy, precision, recall, and F1 score. These metrics are widely used in the literature to evaluate machine learning classifiers and require ground truth validation to be computed. Precision and recall are defined by the following quotients:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, \tag{8}$$

where TP represents true positives, FP false positives, and FN false negatives. The F1 score analyzes the trade-off between precision and recall with the harmonic mean of them:

$$F1 = \frac{2}{Precision^{-1} + Recall^{-1}}. \tag{9}$$

A CDR dataset in granularity of minutes or hours is required to properly apply our proposed method for anomaly detection on Internet, SMS, and call activity from mobile-cellular networks. The fundamental features are summarized in Table III. If the data is geographically distributed, each cell's traffic must be characterized in the data set, *i.e.,* a feature determining the cell identification for each CDR is mandatory (in our case, it is the feature ID). This identification feature is used to characterize geographic areas or individual base stations. A clustering step is required to obtain better results considering several distinct cells. We provide an extensive discussion in Section V on the relevance of characterizing traffic patterns among different cells. Suppose there are a small number of distinct cells. In that case, it is not required to partition them into groups, *i.e.,* applying a clustering technique will not impact the anomaly detection performance. Using an enriched data (clustered), we apply our method for each cluster and return the abnormal activity, informing the energy for each. This entire process is depicted in Figure 4.

Lastly, to compare the execution time of our solution with the state-of-the-art, all of our experiments were executed using the same computer, with a 1.6 GHz Intel Core i5-8210Y processor and 16 GB 2133 MHz LPDDR3. Our source code used in the experiments reported in this paper is available on GitHub[1]. We discuss the obtained results in the following section.

V. Experimental Results: Analysis and Discussion

In this section, we first introduce an analysis of the EFC threshold hyper-parameter to detect anomalies of different severity. Then, the anomaly detection performance of EFC is compared against the performance of solutions from the literature.

A. EFC Threshold Analysis

Considering the data from the region for which we have ground truth information about anomalies, an energy density plot, such as displayed in Figure 5, can be created by applying EFC. As described in Section III, EFC uses a threshold to differentiate normal from abnormal traffic (an anomaly), which is calculated in Equation 7 and presented in the x-axis as $\tau_1 = -27.57$. In this case, any sample $s_{d,sq}$ with energy above this threshold is considered an anomaly.

To enable EFC to detect anomalies of different severity levels, we can filter further by readjusting the threshold with the energy values of the detected anomalies that resulted from the first analysis. In this case, to filter out anomalies with lower energy to focus only on the ones with high energy, we can employ the following threshold:

$$\tau_2 = E(\mathcal{H}_i), \tag{10}$$

where $E()$ stands for the expected value of the energies $\mathcal{H}_i$ of the detected anomalies resulted from the first analysis. With this new threshold $\tau_2 = -17.64$, only anomalies with a high degree of energy are detected, such as depicted in Figure 5. We present our results considering the first threshold as "EFC Low degree Filter" (EFC-LF) and the second one as "EFC High degree Filter" (EFC-HF).

B. Evaluation

A mobile operator may be interested in tracking places where the infrastructure must be rethought to support the presence of flash events. In this case, for illustrative purposes, the foreground of Figure 6 shows the detected anomalies in the region employed in our experiments (depicted in the background) for one specific day, December 22, highlighting the possibility of exploiting our solution for management systems. These anomalies are depicted in the bar plots presented in the foreground. The y-axis represents the frequency of abnormal 10-minute time slots detected using three different approaches: the light gray bars represent detected anomalies using EFC-LF, dark gray bars represent the detected anomalies using EFC-HF, and the black bar represents the detected anomalies using the solution from Zhu and Sun [10] (K-means used as an anomaly detection algorithm). The same evaluation for the whole region depicted in the foreground of Figure 6 was reproduced for each day from November 1 to December 22, resulting in the detected anomalies presented in Figure 7.

In Figure 7, the days colored in green represent the true-positive anomalies (dates from Table IV). The day colored in red represents the false-positive anomalies (*i.e.,* they are abnormal activities not confirmed by available reported events to be used as a ground-truth) from our solution variations (Clustering + EFC-LF and Clustering + EFC-HF). By analyzing the bar plots, it is possible to verify that our solution was able to detect all events in Table IV. As expected, the number of abnormal time slots is always higher for the day when the event occurred. On average, 47% of the detected anomalies on an event day are concentrated in the period of the event.

Both EFC solution variations (EFC-LF and EFC-HF) can detect "less intense" anomalies during the event day, including before and after the scheduled time for the event. For instance, considering that the soccer match on December 1 started at 3 PM, our solution was able to detect abnormal traffic

[1]https://github.com/jonathanalmd/anomaly-detection-in-mobile-networks Last accessed on 2021-09-03.

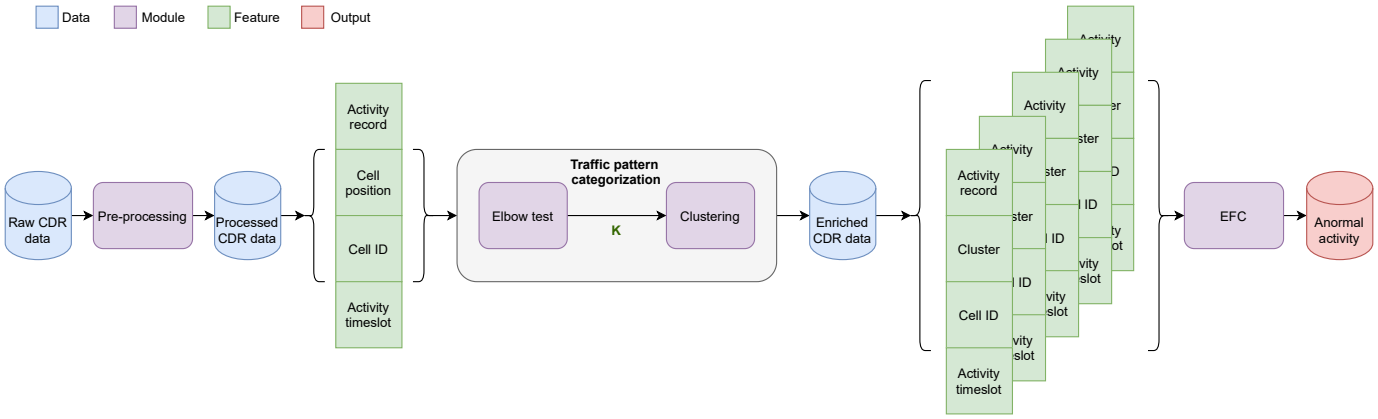


Figure 4. EFC flow chart.

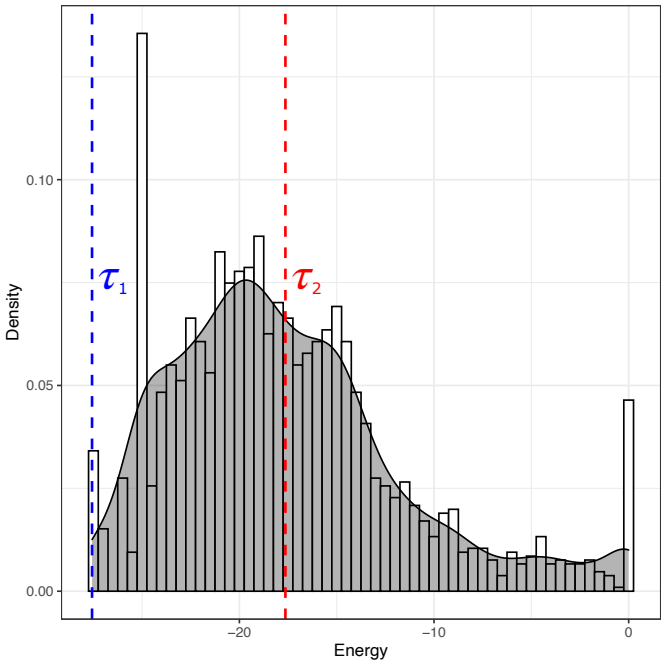


Figure 5. Density plot for the energy of each detected anomaly using EFC.

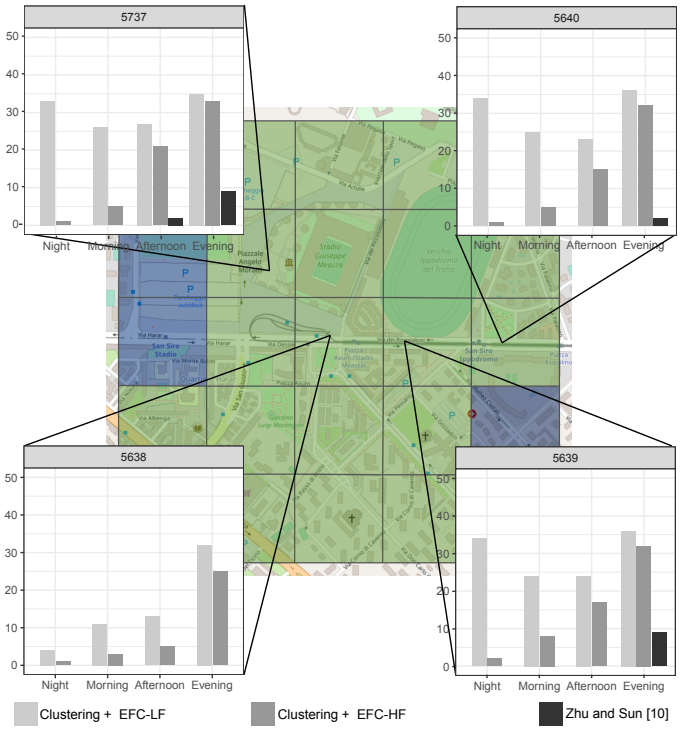


Figure 6. Anomalies from December 22 in the San Siro Stadium/Metro station area (squares 5737, 5638, 5639, and 5640).

behavior from the beginning of the day, achieved a peak in the afternoon, and then decreased in the evening without extending to the next day. This behavior is usually more evident when the estimated attendance at the event is larger, such as the soccer match between the two largest teams from Milan (AC Milan and Internazionale) on December 22, with an estimated attendance of almost 80,000, or the December 11 Champions League match between AC Milan and Ajax (the largest team from the Netherlands), with an attendance of 60,000.

The gradual increase/decrease in the number of detected anomalies can also be observed in some events with a smaller audience, such as the soccer match on December 4 between Internazionale and Trapani Calcio, a small team from the third division of the Italian soccer league. Conversely, this behavior is not evident for other smaller events, such as the protest by hardcore fans among the AC Milan supporters on November 24, during which very few time slots were detected as anomalies. It is expected to observe more abnormal time slots for larger events and fewer for smaller events. Nevertheless, our solution was able to detect anomalies even during those events with lower attendance, while the anomaly detection solution in [10] resulted in several false negatives.

It is evident that, by selecting only the highest energy anomalies (applying the threshold from Equation 10), they are more focused on the exact period of the event. For instance, this is observed in the detected anomalies on December 22, the largest event in our ground truth data. Overall, 73% of the detected anomalies, on average, are focused on the exact period of the event. Selecting the threshold from Equation 10 allows us to tune the precision of EFC to remove less severe anomalies.

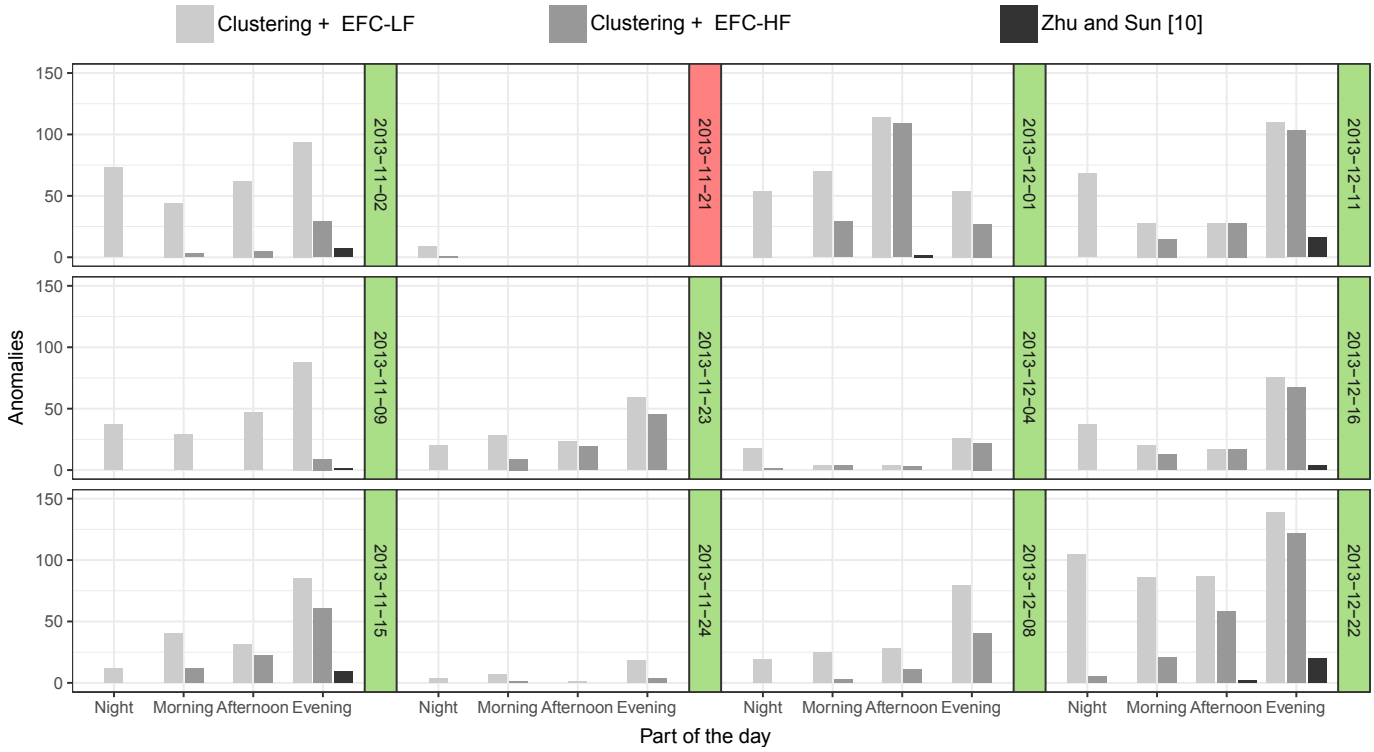


Figure 7. The number of detected anomalies in the San Siro Stadium area for each approach being compared per period of the day from November 1 to December 22 considering the traffic pattern categorization.

Solutions based on the K-means algorithm used to detect anomalies, such as the solutions from Zhu and Sun [10] and Parwez, Rawat, and Garuba [13], are not able to identify the beginning nor the end of an event since these periods' Internet traffic slowly transitions from low to high across different areas and at different times, resulting in less intense anomalies. In these solutions, all detected anomalies are already part of the event with massive traffic demand, and results fall exclusively in the period of the ongoing event. In contrast to the binary classification obtained with those approaches, our solution is more flexible in evaluating the degree of each anomaly and adjusting the energy threshold according to different objectives.

Table V presents the accuracy, precision, recall, and F1 score of four solutions: our proposed Clustering + EFC-LF solution; our EFC-LF solution without considering clustering according to different traffic patterns; the Zhu and Sun solution [10]; and Parwez, Rawat, and Garuba solution [13].

Table V
EFC AND K-MEANS PERFORMANCE RESULTS

Solution	Accuracy	Precision	Recall	F1 Score
Clustering + EFC-LF	0.98	0.92	1.00	0.96
EFC-LF	0.85	1.00	0.33	0.50
Zhu and Sun [10]	0.88	1.00	0.55	0.71
Parwez, Rawat, and Garuba [13]	0.83	1.00	0.27	0.43

The precision is the only metric for which the solutions previously proposed in the literature outperformed our solution, by a small difference of 0.08. Upon close inspection of the data, we determined that this was due to false-positives in four time slots between 0 AM and 1 AM on November 21 in the metro station close to the stadium, represented as the red-labeled day in Figure 7. Nonetheless, it is considered preferable to have some false-positive alarms than missing an anomaly (false-negatives). From the perspective of a manager tracking cyberattacks, detecting isolated anomalies is a most relevant feature, as infrequent anomalies are harder to detect [36][37]. Considering the ground truth from Table IV, our solution obtained better results in general. In particular, the experiments show that our solution is superior in accuracy, recall, and F1 score. For our use case, our solution's (Clustering + EFC-LF) F1 score is more than 35% higher than the F1 score for the solution from Zhu and Sun [10].

It is worthy mentioning, that except from the events displayed in Table IV and since there is no evidence from the literature to map the anomalies on the entire data set, we do not have the ground truth for other timestamps and geographical positions. Therefore, we cannot be sure that detected anomalies from other timestamps and places are false positives or not. In turn, as a step forward in this subject we created an anomaly mask of geographical points that can be incorporate in the data set for further studies with EFC pushing new investigations in this direction. The anomaly mask is available on Github

From Figure 8 it is possible to notice that both variations of our solution were able to detect anomalies in all confirmed events and other potential events at the Mediolanum Forum (square id 2145 of the data set). In particular, for the basketball

[1]https://github.com/jonathanalmd/anomaly-detection-in-mobile-networks/tree/main/anomalies_mask_geojson Last accessed on 2021-09-03.

Figure 8. The number of detected anomalies in the Mediolanum Forum (square 2145) for each approach being compared per period of the day from November 1 to December 22, considering the traffic pattern categorization. Entries colored in green represent days with confirmed events [35], and entries colored in gray represent days with potential events.

match (Olimpia Milano against Regiana) that occurred on November 24, the EFC variations and the literature solution were able to detect abnormal behaviors in the network. Nonetheless, only EFC was capable to detect the events from November 2, 3, 12, 15, 23 and December 3, 16, 18, 19, 21, 22 [35]. For instance, on November 23 there was a live performance of Avenged Sevenfold (a famous American heavy metal band) and on December 12 a large musical event (X-Factor Italia Finale), including a live performance of One Direction (a famous British pop band). Given space limitations, we will not exploit the anomalies from the other days in depth. Nevertheless, we encourage the reader to take a look at [35], where it is possible to track all major events that happened in Mediolanun during the period of interest.

Further, we have strong pieces of evidence that some events occurred on November 9, 10, and 30 and also on December 1 and 6. For the days that our solution has detected few anomalies during the entire day (from a maximum of 144 possible anomalies per day), the majority of them are detected at the end of the work day, in the late afternoon and early evening. We conjecture that these few isolated anomalies detected by our solution may have been caused by agglomerations in the metro station (since most of them are detected between 5 PM and 7 PM). In contrast, the state-of-the-art solution was able to detect anomalies in only six out of the 52 days that were evaluated in the Mediolanum Forum area, in which the major fraction of the detected anomalies were from days with larger events.

Since EFC can have its model trained in advance, it can easily be adapted to perform anomaly detection for certain regions in real-time. In this case, management systems capable of employing dynamic adjustment through self-healing capabilities can leverage from our detection to readjust resources available in a region to provide better communication conditions during a flash crowd event or under the occurrence of network node failures and resource scarcity. Conversely, for under usage of resources due to low traffic in certain regions, these resources can be redistributed or subleased by operators, enhancing their use and profit from leasing, such as described in the work of de Almeida *et al.* [16].

Our results present evidence of the relevance of considering the different traffic patterns from a region when detecting anomalies for the management of mobile networks. As discussed above, this aspect hugely impacts most of the solutions' performance metrics, in particular the F1 score. Hence, to achieve accurate results, the traffic pattern must be considered in anomaly detection solutions, which requires a traffic pattern classification before detecting abnormal behaviors in the network traffic.

VI. Concluding Remarks and Future Work

In this work, we proposed a solution based on unsupervised learning for anomaly detection with traffic pattern classification. We evaluated our proposed solution using a real CDR data set provided by an operator in Italy and compared it against other state-of-the-art anomaly detection solutions. Our solution presented better performance and flexibility, with an F1 score of 0.96, which is more than 35% better than the leading state-of-the-art solution, which has a similar computational complexity. Furthermore, we also demonstrated that considering the distinct traffic patterns of different geographic regions is a key factor for the accuracy of anomaly detection in mobile networks. This aspect is usually ignored in the literature.

Considering the capabilities of the analytics function present in 5G/B5G networks, managers and network operators can employ our solution in real-time to improve their understanding of the user demand and enhance their network performance, for example reconfiguring their network to avoid resource outage or tracking potential cyberattacks (according to the detected anomalies). As future work, we aim to perform ground truth validation in other regions and provide a labeled CDR data set, starting from regions where we detected more anomalies, which are potential regions of interest. These validations will allow the adoption and evaluation of supervised learning algorithms for anomaly detection in mobile-cellular networks.

References

[1] 3rd Generation Partnership Project (3GPP), "3GPP TS 29.520 version 16.4.0 Release 16 - 5g; 5g system; network data analytics services; stage 3", Tech. Rep. 4, August 2020.

[2] D. Turner *et al.*, "California Fault Lines: Understanding the Causes and Impact of Network Failures", in *Proceedings of the ACM SIGCOMM Conference*, p. 315–326, Association for Computing Machinery, 2010.

[3] B. Hussain *et al.*, "Mobile edge computing-based data-driven deep learning framework for anomaly detection", *IEEE Access*, vol. 7, pp. 137656–137667, 2019.

[4] K. Sultan, H. Ali, and Z. Zhang, "Call Detail Records Driven Anomaly Detection and Traffic Prediction in Mobile Cellular Networks", *IEEE Access*, vol. 6, pp. 41728–41737, 2018.

[5] B. Ma, W. Guo, and J. Zhang, "A Survey of Online Data-Driven Proactive 5G Network Optimisation Using Machine Learning", *IEEE Access*, vol. 8, pp. 35606–35637, 2020.

[6] D. Mulvey *et al.*, "Cell fault management using machine learning techniques", *IEEE Access*, vol. 7, pp. 124514–124539, 2019.

[7] B. Galkin, J. Kibiłda, and L. A. DaSilva, "A Stochastic Model for UAV Networks Positioned Above Demand Hotspots in Urban Environments", *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6985–6996, 2019.

[8] B. Hussain *et al.*, "Artificial Intelligence-powered Mobile Edge Computing-based Anomaly Detection in Cellular Networks", *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 1–1, 2020.

[9] B. Li *et al.*, "Anomaly detection for cellular networks using big data analytics", *IET Communications*, vol. 13, no. 20, pp. 3351–3359, 2019.

[10] Q. Zhu and L. Sun, "Big Data Driven Anomaly Detection for Cellular Networks", *IEEE Access*, vol. 8, pp. 31398–31408, 2020.

[11] S. K. Singh, R. Bose, and A. Joshi, "Entropy-based electricity theft detection in AMI network", *IET Cyber-Physical Systems: Theory & Applications*, vol. 3, no. 2, pp. 99–105, 2018.

[12] A. Jindal *et al.*, "Tackling Energy Theft in Smart Grids through Data-driven Analysis", *International Conference on Computing, Networking and Communications (ICNC)*, pp. 410–414, 2020.

[13] M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network", *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 2058–2065, 2017.

[14] Q. Zeng, Q. Sun, G. Chen, H. Duan, C. Li, and G. Song, "Traffic prediction of wireless cellular networks based on deep transfer learning and cross-domain data", *IEEE Access*, vol. 8, pp. 172387–172397, 2020.

[15] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data", *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389–1401, 2019.

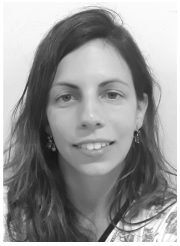
[16] J. M. De Almeida *et al.*, "Optimal Allocation of vBBUs Considering Distance Between MDC and RRH in F-RANs", *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020.

[17] F. Y. Wu, "The Potts model", *Rev. Mod. Phys.*, vol. 54, pp. 235–268, Jan 1982.

[18] J. Macqueen, "Some methods for classification and analysis of multivariate observations", in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.
[19] S. Jaffry, S. T. Shah, and S. F. Hasan, "Data-driven Semi-supervised Anomaly Detection using Real-World Call Data Record", in *IEEE Wireless Communications Networks Conference*, pp. 3–8, 2020.
[20] H. D. Trinh *et al.*, "Detecting Mobile Traffic Anomalies through Physical Control Channel Fingerprinting: A Deep Semi-Supervised Approach", *IEEE Access*, vol. 7, pp. 152187–152201, 2019.
[21] S. Papadopoulos, A. Drosou, and D. Tzovaras, "A Novel Graph-Based Descriptor for the Detection of Billing-Related Anomalies in Cellular Mobile Networks", *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2655–2668, 2016.
[22] B. Hussain, Q. Du, and P. Ren, "Deep Learning-Based Big Data-Assisted Anomaly Detection in Cellular Networks", *Proceedings of IEEE Global Communications Conference*, pp. 1–6, 2018.
[23] M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques", *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.
[24] T. Italia, "Harvard Dataverse - Telecommunications - SMS, Call, Internet - MI", 2015.
[25] I. Zafar, G. Tzanidou, R. Burton, N. Patel, and L. Araujo, *Hands-On Convolutional Neural Networks with TensorFlow: Solve Computer Vision Problems with Modeling in TensorFlow and Python*. Packt Publishing, 2018.
[26] F. Y. Wu, "The Potts model", *Reviews of Modern Physics*, vol. 54, pp. 235–268, jan 1982.
[27] C. F. T. Pontes, M. M. C. de Souza, J. J. C. Gondim, M. Bishop, and M. A. Marotta, "A new method for flow-based network intrusion detection using the inverse potts model", *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1125–1136, 2021.
[28] F. Morcos *et al.*, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. E1293–E1301, dec 2011.
[29] B. Giraud, J. M. Heumann, and A. S. Lapedes, "Superadditive correlation", *Physical Review E*, vol. 59, no. 5, p. 4983, 1999.
[30] T. Italia, "Harvard Dataverse - Milano Grid", 2015.
[31] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino", *Scientific Data*, vol. 2, p. 150055, Oct 2015.
[32] R. L. Thorndike, "Who belongs in the family", *Psychometrika*, pp. 267–276, 1953.
[33] Justin Davis, "Milan leave stadium after players meet angry 'ultras'". https://www.smh.com.au/sport/soccer/serie-a-angry-milan-ultras-lay-siege-to-san-siro-20131124-2y3l0.html, Last accessed on 2021-04-11.
[34] World Football, "Worldwide, AC Milan, and Internazionale Milano - Fixtures & Results from Season 2013/2014". https://www.worldfootball.net, Last accessed on 2021-04-11.
[35] Setlist.fm, "The setlist wiki - Find setlists for your favorite artists". https://www.setlist.fm/search?country=it&query=milan+italy&venue=33d7cccd&year=2013, Last accessed on 2021-04-11.
[36] M. Ahmed and A. S. S. M. Barkat Ullah, "Infrequent pattern mining in smart healthcare environment using data summarization", *The Journal of Supercomputing*, vol. 74, pp. 5041–5059, Oct 2018.
[37] M. Ahmed, "Infrequent pattern identification in SCADA systems using unsupervised learning", *Security Solutions and Applied Cryptography in Smart Grid Communications*, pp. 215–225, 2017.

BIOGRAPHIES

Jonathan M. DeAlmeida (jonathan@aluno.unb.br) holds a M.Sc. and a B.Sc. in Computer Science from the University of Brasilia, Brazil. Currently, he is a Ph.D. student at the same university and he is a Data Scientist at Pier. His research involves Radio Access Networks, Next Generation Networks, Cognitive Radio Networks, Optimization,Artificial Intelligence, and Machine Learning.



Camila F. T. Pontes (cftpontes@gmail.com) is a student at University of Brasilia (UnB), Brasilia, DF, Brazil. She received her M.Sc. degree in Molecular Biology in 2016 from UnB and is currently an undergrad student at the Department of Computer Science (CIC/UnB). Her research interests are Computational and Theoretical Biology, and Network Security.

Luiz A. DaSilva [F] (ldasilva@vt.edu) is the Executive Director of the Commonwealth Cyber Initiative and the Bradley Professor of Cybersecurity at Virginia Tech. He was previously at Trinity College Dublin, where he was the director of CONNECT, the Science Foundation Ireland Research Centre for Future Networks. He is an IEEE Fellow, and an IEEE Communications Society Distinguished Lecturer.

Cristiano B. Both (cbboth@unisinos.br) is an associate professor of the Applied Computing Graduate Program at the University of Vale do Rio dos Sinos (UNISINOS), Brazil. He has coordinated research projects funded by H2020 EU-Brazil, CNPq, FAPERGS, and RNP. His research focuses on wireless networks, next generation networks, softwarization and virtualization technologies for telecommunication network, and SDN-like solutions for the Internet of Things. He is participating in several Technical Programme and Organizing Committees for major international conferences.

João C. Gondim (gondim@unb.br) was awarded an M.Sc. in Computing Science at Imperial College, University of London, in 1987 and a Ph.D. in Electrical Engineering at UnB (University of Brasilia, 2017). He is an adjunct professor at Department of Computing Science (CIC) at UnB where he is a tenured member of faculty. His research interests are network, information and cyber security.

Celia G. Ralha (ghedini@unb.br) holds a Ph.D. from Leeds University, England, and a M.Sc. in Electronic and Computer Engineering from the Aeronautics Institute of Technology, Brazil. She is an associate professor at the Informatics Graduate Program, University of Brasilia, Brazil, and a senior member of the Brazilian Computer Society. She receives a research grant from the Brazilian National Council for Scientific and Technological Development (CNPq). Her current research interests include intelligent systems, multi-agent systems & planning, agent-based simulation, and cognitive networks.

Marcelo A. Marotta (marcelo.marotta@unb.br) is an adjunct professor of the Department of Computer Science of the University of Brasilia, Brazil. He holds a Ph.D. in Computer Sciences from the Institute of Informatics of the Federal University of Rio Grande do Sul, Brazil. His research involves Cloud Radio Access Networks, Wireless Networks, Next Generation Networks, Internet of Things, Software Defined Radio, and Cognitive Radio Networks.