International Journal of Research in Engineering, Science and Management
Volume-2, Issue-1, January-2019
www.ijresm.com | ISSN (Online): 2581-5792

Global Mobile Data Traffic Prediction and Anomaly Detection in Cellular Network

N. Suruthi[1], R. Kavitha[2]
[1]*M.E. Student, Dept. of Computer Science and Engg., Parisutham Inst. of Tech. and Science, Thanjavur, India*
[2]*Associate Professor, Dept. of Computer Science and Engg., Parisutham Inst. of Tech. and Sci., Thanjavur, India*

***Abstract*: Mobile networks possess information about the users as well as the network. Such information is useful for making the network end-to-end visible and intelligent. Big data analytics can efficiently analyse user and network information, unearth meaningful insights with the help of machine learning tools. Utilizing the big data analytics and machine learning, this work contributes in three ways. First, we utilize the call detail records (CDR) data to detect anomalies in the network. For authentication and verification of anomalies, we use k-means clustering, an unsupervised machine learning algorithm. Through effective detection of anomalies, we can proceed to suitable design for resource distribution as well as fault detection and avoidance. Secondly, we prepare anomaly-free data by removing anomalous activities and train a neural network model. By passing anomaly and anomaly-free data through this model, we observe the effect of anomalous activities of the model and also observe mean square error of anomaly and anomaly free data. We use an autoregressive integrated moving average (ARIMA) model to predict future traffic for a user and fit these models into the time series model to provide better understanding of the model. Through simple visualization, we show that anomaly free data generalizes the learning models and performs better on prediction task.**

***Keywords*: Anomaly, Call Data Records, k-means Clustering, Data Analytics.**

1. Introduction

Location information is an important feature in users' profiles in cellular mobile networks. Anomaly detection has always been the focus of researchers and especially, the developments of mobile devices raise new challenges of anomaly detection. For example, mobile devices can keep connection with Internet and they are rarely turned off even at night. This means mobile devices can attack nodes or be attacked at night without being perceived by users and they have different characteristics from Internet behaviors. Mobile technologies and cellular networks are getting smarter day by day, so mobile phone devices such as smartphones, tablets, wearable devices as well as mobile phone subscribers are increasing rapidly. According to report presented by Ericsson, the mobile devices have surpassed the world population [1]. Due to such a huge growth in mobile devices and mobile phone subscribers, the congestion of mobile network is not unusual. Hence the provision of best quality of services for such a huge number of mobile phone subscribers is challenging. With the massive growth of mobile devices and mobile phone subscribers, the data generated from these devices is increasing explosively. According to CISCO survey, the data increased 4000-fold during the last ten years [2]. From CISCO report, global mobile data traffic is generating 24 Exabyte (EB) data per month and this trend is continuously rising [2]. This huge data has following 4Vs characteristics making it different from the traditional data.

A. Volume

The data has very large volume; of order of Pico bytes. 12 Terabytes data is generated by Twitter every day [3]. On average, 1.2 Zeta bytes of data is being produced every year since 2012 and this value is continuously rising [4], [5].

B. Velocity

The flow rate of data at which the data goes in or out from mobile devices and mobile network is termed velocity of data. This determines the dynamic nature of the data and big data is highly dynamic.

C. Variety

This huge and dynamic data comes from various sources and occurs in different formats such as structured, unstructured and semi-structured.

D. Value

According to IDC report, a very famous group for big data research activities, big data technologies describes a new generation of technology, designed to extract value from a huge volume of a wide variety of data [6].

2. Literature survey

Naboulsi et al., [9] presented a framework, in which a large scale CDR dataset was sub categorized according to the history of activities. The framework reported in [9] determines the irregular and unexpected activities termed as anomalies. The authors in [10], [11] used k-means clustering techniques for determining regions of interest such as commercial areas, residential areas, office areas, and recreational areas etc. The authors in [12], [13] also used k-means clustering for anomaly detection purpose. The authors divided data into clusters of anomalous data and normal data. The authors in [14] analyzed

CDR information of wireless network and detected anomalies by rule based approach. Authors in [15], [16] presented the significance of CDR data analysis in case of natural disasters. The CDRs being generated daily are huge in number. The CDR data contains valuable insights that can be used for the benefit of the network operators as well as subscribers. A milestone for such a huge CDRs data analysis is presented in [17].The authors in [17] have presented a stream processing model which is able to analyses 6 billion CDRs generated per day. The model has the ability to support higher throughput, lower latency and fault tolerance. Parallel processing, deduplication and easy to use platform for network operators are the main aspects of the model. The authors in [18] presented a big data analytics based model for optimizing 5G networks and showed that the network will be faster and proactive with the aid of big data analytics. The authors in [19] showed that bandwidth can be efficiently distributed with the aid of big data analytics. The authors in [20] presented that self-organizing network (SON) which will be used for enabling 5G, can be efficiently implemented with big data analytics. Such a framework based on SON was named a BSON [20]. The authors in [21] showed that big data analytics will be helpful for proactive caching which is very important for empowering 5G. Motivated from the literature, we use k-means clustering algorithm and detect the anomalous behavior of the users. Our work is different from the previously reported work which was limited to anomaly detection only. We perform verification of the anomaly detection through comparison with ground truth data. After successful anomaly detection and verification, we prepare anomaly-free data. We also train a neural network model for observing mean square error of anomaly and anomaly-free data. Finally, we train ARIMA prediction model for predicting users' future activities. Furthermore, we discuss that such type of insights are also helpful towards 5G networks' requirements such as proactive caching, maximum throughput and close-to-zero latency.

3. Proposed system

The CDR data of a cellular network can be used for analyzing user or network behavior. With the aid of CDR data analysis, one may extract information and detect unusual events of critical significance e.g., a terrorism activity, earthquakes, floods, Christmas Eve, soccer world cup, black Friday etc. If there is flood, earthquake or any terrorism activity in a particular region, the CDR activities will be increased in that region as mobile phone subscribers will give calls, send SMS to other people to inform about such activity in that region and will ask for rescue and help. Because of such type of abnormal behaviour or anomalous activities, the performance of network will be down and mobile phones subscribers will have poor QoS. However, after detecting and predicting such type of anomalous activities for the next time frame, the network operator can provide some extra resources for specific time frame on a particular area.

The hypothetical representation in Fig. 1 shows that there are three types of areas, low activity area, average activity area and high activity area. This work utilizes the unsupervised k means technique for classifying the Call Details Records to find the exact call anomalies. Here we use ARIMA model to detect the anomalies in the network. ARIMA time series forecasting model is used to predict the user traffic. It is the class of model that captures a suite of different standard temporal structures in the time series data. Auto Regressive model that uses the different relationship between an observation and some number of lagged observations. Integrated model used to provide differences of raw observations in order to make the time-series stationary. Moving Average model uses the dependency between observation and a residual error from a moving average model applied to lagged observation.

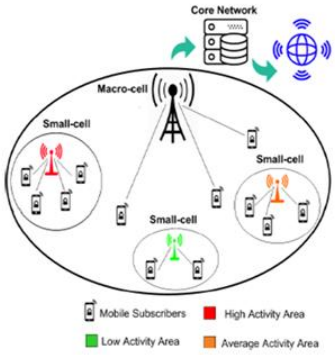

Fig. 1. System model

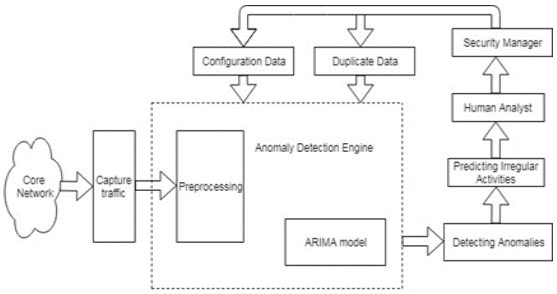

Fig. 2. System architecture

4. System module description

A module description provides detailed information about the module and its supported component, which is accessible in this system. Here we used four types of system modules.

A. Data acquisition & pre processing

Data Acquisition is the process of data from variety of data sources. Here we have collected the data from three different datasets of cellular networks. The first CDR Dataset is obtained from CRAWDAD community. CRAWDAD is a wireless network data resource for the research community [22]. This dataset contains the mobile phone records of 142 days from September 2010 to February 2011.Users' activity is in the form of incoming call, outgoing call, incoming SMS and outgoing SMS. The second CDR Dataset is obtained from Nodobo, which is a suite of software developed at the University of

Strathclyde, and allows precise capture and replay of smartphone user interactions sessions [23], [25]. This dataset was collected during a study of mobile phone usage of 27 high school students. The dataset contains the mobile records for a period of six months from September 2010 to February 2011.This dataset includes 13035 voice call records, 83542 messages records and other related data. User field contains the ID of the subscriber or caller. Other field contains ID of the receiver. The third dataset is obtained from open big data database of Dandelion API web forum. The dataset which we use from this open big data database is for Telecom Italia. This CDR available from the Telecom Italia is for the city of Milano. The dataset includes CDRs for a period of two months from November 2013 to January 2014. Table III represents sample of this dataset. This field contains the identification number of the grid. The data from these sources contain missing entries or noise causing misleading pattern. Data pre-processing step removes such type of irregularities. Through pre-processing, we obtain data and ready for further analysis. Data cleaning and Data splitting can be done in data pre-processing. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and lacking in certain behaviors or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. Data pre-processing is used database-driven applications such as customer relationship management and rule-based applications.

B. Anomaly detection and verification

In this method we detect the anomalies by visualizing ground truth data and comparing the anomalies through K-means clustering. In ground truth data we consider the activity of user over one week. The activity level is very high in some user. This higher levels are showing an unexpected and irregular behavior of user. So we consider these activity as anomalies. These are helpful to find the Region of Internet such as location, hospital, and natural disasters. We then verify anomalies through visualization of the ground truth data and comparison with the anomalies detected through clustering. Clustering is a process of partitioning a group of data into small number of clusters and sub-groups. After k-means clustering, we obtain number of clusters with different data points. Thus, the data is divided into different clusters. As one can anticipate, cluster of fewer objects or data points is the cluster of anomalous activities. As anomalous activities are unexpected and irregular making them unique and fewer in number than normal activities. Through k-means clustering algorithms, normal activities are grouped into the same clusters different than the one in which abnormal activities are placed. Objects or users which lie in the range of activity level 1-150 are grouped in cluster 1, users which lie in the range of activity level 150-400 are grouped in cluster 2 and remaining users with much higher activity level are grouped in cluster 3. Anomaly detection using ground truth data is useful to find anomalous activities which are also helpful to identify region of interest e.g., locations with high density of users such as shopping malls, hospitals, or stadium; or identify events of interest e.g., natural disaster, fatal road accidents, or terror activity. Anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

C. Anomaly filtering

After detection and verification of anomalies by ground truth data and machine learning algorithm, we clean the data from such anomalous and abnormal activities making data anomaly free. For preparation of anomaly-free data, we replace the anomalous activities of the users by average activities of all the users. We train neural network model with anomaly as well as anomaly free data and observe mean square errors. For observing error difference in anomaly and anomaly-free data, we train a neural network model. We pass anomalous and anomaly-free data through this model and observe mean square error. It is observed that the mean square error of test, train and validation data is high when anomalous data is passed through the model. On the other hand, when the model is trained with anomaly free data, the overall mean square error is decreased. In order to highlight the significance of the pre-processing step for data pre-preparation, it is important to use a numeric metric. We calculate the mean square error for the model training with both anomalous and anomaly-free data. Thus, the mean square error serves as a numeric parameter to ascertain the impact of outliers in the cellular data. The mean square error calculation for both the datasets shows that anomaly free data help us develop a better model. The mean squared error of an estimator measures the average of the squares of the errors that is the average of squared differences between the estimated values and what is estimated. Here it is used to find the differences between the values using both anomaly and anomaly free data. The effect of the mean square error can be severe, depending upon the target of the model. For example in the case of sleeping cell detection, if anomalous data is used in the model, then the model would not be able to detect sleeping cells correctly. Under worst scenarios, this may lead to network outage as a consequence of denial of service for newer devices.

D. Traffic prediction

ARIMA time series forecasting model is used to predict the future traffic of mobile networks. The CDR Datasets which are used in this work are time series data. CDR time series data can be used for predicting and detecting future anomalous behavior of the network and Subscribers. Among the time series forecasting models, ARIMA is a popular and widely used time-series forecasting model. ARIMA stands for Autoregressive Integrated Moving Average. It is generalized auto-regressive

model and adds the notion of integration. ARIMA model is fitted with the time series model to provide either the better understanding of data or to predict future points in series. Autoregressive feature uses the dependency relationship between current observations and a specified number of previous observations. Integrated feature is used for making a series stationary if it is non-stationary, done by subtracting raw observations. Moving Average feature uses dependency between an observation and a residual error from a moving average model applied to lagged observations.

5. Preparation of anomaly free data using ground truth data

Ground truth refers to information collected on location. Ground truth allows image data to be related to real features and materials on the ground. The collection of ground-truth data enables calibration of remote-sensing data, and aids in the interpretation and analysis of what is being sensed. The collection of ground-truth data enables calibration of remote-sensing data, and aids in the interpretation and analysis of what is being sensed. Examples include cartography, meteorology, analysis of aerial photographs, satellite imagery and other techniques in which data are gathered at a distance. More specifically, ground truth may refer to a process in which a pixel on a satellite image is compared to what is there in reality in order to verify the contents of the "pixel" on the image. In the case of a classified image, it allows supervised classification to help determine the accuracy of the classification performed by the remote sensing software and therefore minimize errors in the classification such as errors of commission and errors of omission.

A. K- Means clustering

It is the popular clustering algorithm in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter's of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done.

6. ARIMA time series forecasting model

A popular and widely used statistical method for time series forecasting is the ARIMA model. ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data. ARIMA model have four steps to detect anomalies and provide anomaly free data. In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an ARIMA model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series. ARIMA models are applied in some cases where data show evidence of non-stationary, where an initial differencing step can be applied one or more times to eliminate the non-stationery. The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible. Non-seasonal ARIMA models are generally denoted ARIMA (p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. Seasonal ARIMA models are usually denoted ARIMA(p,d,q)(P,D,Q)m, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model. Differencing in statistics is a transformation applied to time-series data in order to make it stationary. A stationary time series' properties do not depend on the time at which the series is observed. ARIMA model contains four steps to predict future traffic. It will fitted with time series model to provide better understanding of the data.

- Data Stationary
- Optimal Parameters Selection
- Build the ARIMA model
- Make Prediction

A. Data stationary

We visualize the data and observe increasing or decreasing trends of the data. We take necessary action to make the data stationary as required for time series forecasting model. It is typically assumed for time series forecasting models that the input data is stationary. If time series data is not stationary, it

should be made stationary before training a time series forecasting model1. In summary statistics, it is observed that for stationary time series data, mean and variance should be constant for observations. For stationary confirmation of the data, The Augmented Dicky-Fuller (ADF) statistical test is used. ADF test is also called unit root test. The ADF test has two hypothesis; Null hypothesis and Alternate hypothesis. Null hypothesis suggests that time series data has a unit root, so data is non-stationary. Alternate hypothesis suggests that time series data does not have any unit root implying that data is stationary. These hypothesis are interpreted by p value of the ADF test. If p value is greater than 0.05, null hypotheses is accepted and data is non-stationary. Similarly, if p-value is less than or equal to 0.05, null hypothesis is rejected and data is stationary. We have applied ADF test on one day's CDR activities of the users for confirmation of data stationarization. ADF statistics of available dataset shows that p-value is 0.6177, implying that the data is non-stationary. Hence after confirmation of non-stationary behavior of data, we make the data stationary by differencing series and the lower part represents the stationary time series. A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary through the use of mathematical transformations. A stationaries series is relatively easy to predict: you simply predict that its statistical properties will be the same in the future as they have been in the past. The predictions for the stationaries series can then be "untransformed," by reversing whatever mathematical transformations were previously used, to obtain predictions for the original series. Thus, finding the sequence of transformations needed to stationaries a time series often provides important clues in the search for an appropriate forecasting model. Stationarizing a time series through differencing (where needed) is an important part of the process of fitting an ARIMA model.

B. Optimal parameters selection

After determining and confirming the stationery of the time series, the next step is to determine the optimal value of model's parameters. The optimal values of p and q is determined with the help of two plots. Those are,

- Autocorrelation Function
- Partial Autocorrelation Function

C. Build the ARIMA model

After defining the optimal values and the order of model's parameters (p; d; q), we build ARIMA model and fit data in the model. The dataset is divided into train and test sequences. The model is trained on 70 percent data and tested on remaining 30 percent data. The parameters of the ARIMA model are defined as follows:

- p: The number of lag observations included in the model, also called the lag order.
- d: The number of times that the raw observations are differenced, also called the degree of differencing.
- q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove trend and seasonal structures that negatively affect the regression model. A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model. Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious, but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

D. Make rediction

After building the model, we use it for forecasting. Firstly we analyze the model with the available data, in this model precision is observed. We use the user activity data of one week to make prediction with ARIMA model applied on data for one day. We also apply the model on user's activity of a week. With such time series forecasting for CDR data, the trends in mobile network as well as mobile's subscribers are predicted well before time. This help in interpretation, and thus smart management of cellular networks in terms of spectrum management, fault detection/avoidance and provision of just in-time services. These applications have great potential use in next generation networks. Anomaly detection is an effective means of identifying unusual or unexpected events and measurements within a web application environment. As the term "unexpected" can also be read as "statistically improbable," it should be clear why anomaly detection depends heavily on deep knowledge of a system's baseline performance and behavior for its insights and load forecasts. This is why Dynatrace monitors entire technology stacks end-to-end within web-scale environments. Dynatrace monitors the baseline performance and behavior of applications, services, infrastructure components, and more. Dynatrace captures metrics related to availability, error rates, response times, service load, user traffic, and resource dependencies across millions of entities. Because there are differing assumptions involved in evaluating load anomalies than there are in evaluating performance anomalies, Dynatrace relies on a wide spectrum of measures and methodologies to identify anomalous events that affect customer experience and therefore require your attention. While multidimensional base lining is used to automatically detect anomalies in the response times and error rates of applications and services (response times should never rise to critical levels, even during high-load situations), a prediction-based methodology approach is used to detect

abnormalities in application traffic and service load. This is because traffic and load are entirely dependent on daily, seasonal, and business-cycle related patterns that are driven by an application's business model, related marketing efforts, and sociological factors. Examples of such cycles include weekends/workweeks, workday/evening hours, and holiday-driven customer activity. Black Friday is a great example of an extraordinary seasonal event that occurs on an annual cycle.

7. Conclusion

In this project, we have analyzed CDR data from mobile network. For CDR data analysis, we have used k-means clustering technique. The users' activities which exhibit unusual behavior are termed as anomalies. We verified anomalies by plotting ground truth data and analysis plot for k-means clustering algorithm. The goal of a detection system is to select time periods, where there is unusual behavior in the examined network element. These flagged periods can later be classified into different anomaly types either by human experts or by utilizing machine learning algorithms, extracting additional information about the network, and how these anomalies are formed. After detection and verification of anomalies, we can also identify the region where such anomalies occur. This helps in identification of the region of interest or event of interest. After identification of such regions or events, proper action such as resource distribution, sending drone small cells can be taken in advance and on time. Hence because of such actions, the user's requirements will be fulfilled and will have the best QoS as well as the avoidance of network congestion. After finding the region of interest we used ARIMA model to predict the future traffic. Further we can use regression method to classify anomalies based on their relationship. We can use Levenberg-Marquardt algorithm to train Neural Network to improve performance. We can use Python language to reduce elapsed time of the project.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1065-1082, May 2014.

[2] E. Bas¸tu˘g, M. Bennis, E. Zeydan, M. A. Kader, I. A. Karatepe, and M. Debbah, "Big data meets telcos: A proactive caching perspective," Journal of Communications and Networks, vol. 17, no. 6, pp. 549–557, January 2015.

[3] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti," PLoS Med, vol. 8, no. 8, pp. 1-10, June 2011.

[4] Bo Sun, KuiWu, Yang Xiao, and Victor C. M. Leung, "Enhancing Security Using Mobility Based Anomaly Detection in Cellular Mobile Networks," IEEE transactions on vehicular technology, vol. 55, no. 4, pp. 1385-1395, July 2017.

[5] Chunyong Yin, Sun Zhang, and Knwang-jun Kim, "Mobile Anomaly Detection Based on Improvement Self Organizing Maps" Research Article, volume 13, no. 5, pp.1-9, July 2017.

[6] B. Fan, S. Leng, and K. Yang, "A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks," IEEE Network, vol. 30, no. 1, pp. 6–10, July 2016.

[7] A. Forestiero, "Self-organizing anomaly detection in data streams," Information Sciences, vol. 373, no. 4, pp. 321–336, March 2016.

[8] J. Gantz and D. Reinsel, "Extracting value from chaos," IDC iview, vol. 11, no. 11, pp. 1–12, June 2011.

[9] Gerhard Munz, Sa Li, Georg Carle, "Traffic Anomaly Detection Using K-Means Clustering," journal of networks, vol.8, no.8, pp.2914-2925, August 2017.

[10] P. W. Gething and A. J. Tatem, "Can mobile phone data improve emergency response to natural disasters?," PLoS Med, vol. 8, no. 8, pp.10-85, August 2011.

[11] Ilyas Alper Karatepe, Engin Zeydan, "Anomaly Detection In Cellular Network Data Using Big Data Analytics," IEEE,Vol. 9, no. 9, pp.34-40, June 2017.

[12] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," IEEE Network, vol. 28, no. 6, pp. 27–33, February 2014.

[13] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," 20th European Wireless Conference; Proceedings of. VDE, vol. 30, no. 1, pp. 1–5, September 2014.