Hindawi
Security and Communication Networks
Volume 2022, Article ID 1182315, 9 pages
https://doi.org/10.1155/2022/1182315

WILEY | Hindawi

Research Article

Aggregated Traffic Anomaly Detection Using Time Series Forecasting on Call Detail Records

Arian Mokhtari,[1] **Niloofar Ghorbani**,[2] **and Behnam Bahrak**[1]

[1]*School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran*
[2]*Department of Mathematical Sciences, High Point University, High Point, NC, USA*

Correspondence should be addressed to Arian Mokhtari; a.mokhtari1992@alumni.ut.ac.ir

Received 31 October 2021; Revised 15 January 2022; Accepted 5 February 2022; Published 2 March 2022

Academic Editor: Shahram Babaie

Mobile network operators store an enormous amount of information like log files that describe various events and users' activities. Analysis of these logs might be used in many critical applications such as detecting cyber attacks, finding behavioral patterns of users, security incident response, and network forensics. In a cellular network, call detail records (CDRs) is one type of such logs containing metadata of calls and usually includes valuable information about contacts such as the phone numbers of originating and receiving subscribers, call duration, the area of activity, type of call (SMS or voice call), and a timestamp. With anomaly detection, it is possible to determine abnormal reduction or increment of network traffic in an area or for a particular person. This paper's primary goal is to study subscribers' behavior in a cellular network, mainly predicting the number of calls in a region and detecting anomalies in the network traffic. In this paper, a new hybrid method is proposed based on various anomaly detection methods such as GARCH, K-means, and neural network to determine the anomalous data. Moreover, we have discussed the possible causes of such anomalies.

1. Introduction

Today, a great deal of data is being produced by people and their interactions. In cellular networks, many continuously changing network parameters and measurements are obtained from subscribers. Mobile operators use these measurements and other information to improve the performance of their network. Call detail records (CDR) is one of these measurements that is widely employed to discover the behavioral patterns of subscribers in a network [1].

In the telecommunication network, the anomalies are those behaviors of the user in the network that are different or unusual from their usual or expected actions. Anomaly detection methods based on data mining techniques, such as statistical inference and machine learning, are extensively utilized in many industries and services such as financial systems, health insurance and healthcare, and cyber defense [1].

Anomaly detection has many applications in mobile networks, such as security incident detection, resource allocation, and load balancing [2]. Additionally, the anomaly detection of CDR data can play an essential role in improving municipal services, such as public transportation planning and traffic management. Many of the anomaly detection methods are based on forecasting techniques [3]. Forecasting problems are often classified into three categories: short term, medium term, and long term [3]. Short and medium-term forecasting problems are usually based on identification, modeling, and extrapolation of patterns found in previous data. Due to the lack of significant changes in these earlier data, statistical methods are useful for short-term and mid-term forecasting.

1.1. Contribution. In this paper, we utilized the CDR dataset from a real mobile cellular network, an example of short-time forecasting, which includes the prediction of future

events in short periods of time, such as days, weeks, and months. Time-space information in these CDR helps us analyze aggregated subscriber's behavior in a specific area on a particular date and time. Anomalies in the performance of a network can take place due to many reasons, such as sleeping cells, hardware failures, the surge in traffic, network attacks, and special occasions like national celebrations. In this paper, we propose a new method for anomaly detection in the time series of subscriber usage (measured by the number of calls) in a cellular network. Our approach is based on a combination of well-known methods, such as generalized autoregressive conditional heteroscedasticity (GARCH), K-means, and neural networks, and outperforms all of them. We call this model a hybrid model.

Our contributions towards anomaly detection in the telecommunication domain are as follows:

(i) We try to detect the unusual behavior of the users using a hybrid model that utilizes the benefits of three methods: GARCH, K-means, and neural networks

(ii) We use logistic regression for causality inference

(iii) We compare the results of the hybrid model with the previous works

1.2. Paper Organization. The remainder of the paper is organized as follows. Section 2 describes the related work. In Section 3, anomaly detection algorithms are discussed and the dataset is represented. In this section, various methods used for anomaly detection and the errors of each way are discussed and compared with the previous works. Finally, Section 4 concludes the paper.

2. Related Work

Anomaly detection methods based on machine learning and neural networks have been used in many research works [1–4]. Besides, methods based on statistical models such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), autoregressive conditional heteroscedasticity (ARCH), and GARCH models have been used as well [5, 6]. In reference [7], a framework for the large-scale classification of contact details is proposed in various networks.

Anomaly detection using CDR data has already been extensively studied in various investigations, including in reference [8], where anomaly detection was performed using fuzzy logic on the duration of the calls in the CDR dataset. In [9, 10], the K-means clustering method was used for CDR for purposes such as the identification of administrative areas, parks, and commercial areas. K-means clustering was also used in reference [11] to detect anomalies in the traffic data. The data included unlabeled records separated by the K-means algorithm into normal and abnormal traffic. In reference [2], K-means clustering and hierarchical clustering methods have been used to detect anomalies as well as neural network techniques for prediction. The paper [12] analyzes the main categories of abnormal diagnostic procedures, including classification, statistical methods, information theory, and clustering that were used for the network intrusion detection dataset. In references [13, 14], CDR-based anomaly detection using a rule-based technique and user-contact activity has been analyzed. In this article, the abnormal behavior of the user's activity in a cellular network was detected using some CDR attributes such as LAC ID, cell ID, call date, and call time. Also, in reference [15], anomaly detection on mobile networks was investigated using billing information. In reference [16], the time series anomaly detection methods have been studied based on statistical purposes, clustering, deviation, distances, and densities.

In reference [17], first, a graphic is provided for displaying a voice call. Then, using the cipher query language, CDR data are imported to the Neo4j graph database to understand subscriber behavior and abnormal behaviors.

Lower accuracy and high false positive rates (FPRs) allude to the loss of rare resources, which eventually results in increased operational expenditure (OPEX) while interrupting the network's quality of service (QoS) and user's quality of experience (QoE). High FPR implies that false alarms may squander a substantial amount of OPEX and network resources. In the following, we want to highlight the efforts made to improve accuracy and FPR. Parwez et al. [2] proposed K-means and hierarchical clustering algorithms to indicate rising traffic (that may lead to congestion) in a cell by analyzing past one-week data. They obtained 90% accuracy. Imran et al. achieved 94% accuracy for the detection of sleeping cells [18]. Hussain et al. [19] applied a semisupervised machine learning algorithm to discover the anomalies in one-hour data using the CDR dataset that had information about the past several weeks' user interactions. Their proposed method can achieve an accuracy of about 92.79%; however, they also obtained 14.13% FPR.

The study proposed by Hussain et al. [20] is the first study that applies deep learning for the detection of anomalies. The authors utilized a comprehensive investigation of the L-layer deep feedforward neural network fueled by a real CDR dataset. They achieved 94.6% accuracy with a 1.7% FPR, which are remarkable improvements, and overcome the limitations of the previous studies. Hussain et al. and Sui et al. [21, 22] proposed a framework that utilizes a feedforward deep neural network to detect anomalies in a single cell of a cellular network. It preprocesses real CDR to extract a 5-feature vector corresponding to user activities of a cell, that it accepts as an input. The output is a binary number indicating zero as usual and one as an anomaly. Their framework achieved 98.8% accuracy with 0.44% FPR. These results for accuracy and FPR are summarized in Table 1.

Anomaly detection for large-scale cellular networks can be used by network operators to optimize network performance and enhance mobile user experience. Some research studies aim at detecting user anomalies from spatiotemporal cell phone activity data. Actually, they design an approach combining time series analysis and machine learning to extract the traffic patterns of areal units [23, 24]. In references [25, 26], a spatiotemporal convolutional

Table 1: Summarized results for accuracy and FPR.

Literature	Accuracy (%)	FPR
Parwez et al. [2]	90	—
Imran et al. [18]	94	—
Hussain et al. [19]	92.79	14.13%
Hussain et al. [20]	94.4	1.7%
Hussain et al. [21]	98.8	0.44%

network is presented that uses an attention mechanism to solve spatiotemporal modeling and predict wireless network traffic.

Our work introduces a new method for anomaly detection based on various methods of data forecasting. GARCH, neural network, K-means, and logistic regression techniques are used on mobile network data. This type of information is well studied in the literature in terms of anomaly detection. The novelty of this paper is in using the prediction algorithm in a hybridized way. Data are predicted using GARCH and neural network techniques and evaluated in the hybrid model. This model is examined from two perspectives. In the first mode, each record will be identified as an anomaly if at least one of the methods detected it as an anomaly. In the second mode, a record must be recognized as an anomaly in all ways in order to be considered as an anomaly. By applying the proposed methods, proper solutions can be reached for minimizing the FPR and maximizing accuracy. Our approach delivered an FPR of 0.01% for the first mode and 0.012% for the second mode, which is significantly lower than the reported rates. Also, we achieve an accuracy of 99.72% for the first mode and 99.68% for the second mode. Both methods have a significant improvement as compared with the reported results in Table 1. Furthermore, we use logistic regression for causality inference.

In the following, we provide the technical background on different anomaly detection algorithms required to understand the rest of this paper.

2.1. Statistical-Based Anomaly Detection. In this section, statistical methods such as ARIMA and GARCH are explained.

2.1.1. ARIMA Model. ARIMA is a generalization of the ARMA model. ARIMA models are used because they can reduce a nonstationary series to a stationary series utilizing a sequence of differencing steps. ARIMA models are applied in some cases where the data show evidence of nonstationarity. It is common to use ANOVA when the mean is stationary. The ANOVA is the generalized model of the t test and is an adequate method for the comparison of mean in the time series. We can utilize the Leven test or Bartlett test stationary of variance. The nonstationary data can be converted to stable data by the several uses of the differentiation technique, so it is possible to assess an ARMA model for the transformation data. The ARMA (p,q) model for the transformation data is the same as the ARIMA (p,d,q) model for the primary data with parameters p, d, and q where p is the repetition number of utilizing the technique of differentiation, d is the degree of autoregressive, and q is the moving average. It can be used in other transformation techniques such as Box-Cox when the data remain nonstationary after several uses of differentiation [6].

2.1.2. GARCH Model. When the ARMA model is used for error variance, it will be the GARCH model that conditional difference at any moment depends on data and conditional variances of previous moments. In GRACH (p, q) model, parameter q is the number of delays of error, and parameter p is the number of delayed series. The variance is defined as follows [6]:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \tag{1}$$

where p is the order of the GARCH terms σ^2 and q is the order of GARCH terms ε^2. α_i and β_j are the coefficients for the GARCH model. It can be proven that the stochastic process based on the GARCH model is broad sense stationary when the following equation is established:

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1. \tag{2}$$

2.2. Machine Learning-Based Anomaly Detection. In this section, different methods of machine learning, such as K-means, clustering, and neural network are introduced, which are used for anomaly prediction and detection.

2.2.1. K-Means Clustering. K-means clustering is one of the most straightforward unsupervised clustering techniques used to solve clustering problems, especially when there are lots of data. The purpose of using the K-means clustering method is splitting n observations into K clusters where every observation belongs to the cluster with the closest mean. It is supposed that the parameter K is deterministic. Various methods, such as the elbow method, can be used for calculating parameter K [2].

2.2.2. Neural Network-Based Anomaly Detection. Artificial neural networks are predictive methods functioning based on modest mathematical models of the brain. Neural networks can be considered as a network of neurons that consists of several layers. The predictor consists of the lower layers (inputs) and predictions (outputs) of the upper layers. Also, the middle layers include hidden neurons. The simplest networks, which are linear regression, are without hidden layers. With time series data, delayed time series can be employed as inputs for a neural network. Given that the delayed values are used in the linear autoregressive model, they are called neural network autoregressive (NNAR). The NNAR (p, k) represents the latency of p input and the k nodes in the hidden layer [27–29].

2.3. Logistic Regression. Logistic regression is a causality inference method for categorical variables and is one type of the generalized linear model (GLM). Here, GLM can be fitted by choosing the features as the explanatory variables and the anomaly as the categorical response variable. Each GLM has the following characteristics:

(i) probability distribution describing the outcome variable

(ii) A linear model

$$\eta = \beta_0 + \sum_{i=1}^{n} \beta_i X_i. \tag{3}$$

(iii) A link function that relates the linear model to the parameter of the outcome distribution:

$$\begin{aligned} g(p) &= \eta, \\ p &= g^{-1}(\eta). \end{aligned} \tag{4}$$

Because the response variable is binomial distribution, the common link function that connects η to p is the following logit function:

$$\eta = \text{logit}(p) = \log\left(\frac{p}{1-p}\right) 0 \le p \le 1. \tag{5}$$

Based on equation (5), the odd ratio of success to failure will be Euler's number to the power of coefficients of the fitted model [30–33].

3. Call Detail Record Analysis

The data are divided into two sets: training data and test data, in which 48% of data are training data, and the rest 42% are test data. All simulations of this paper are performed with *R* and MINITAB software. Then, a suitable statistical model is chosen for the time series. In the next step, the predicted data and the detected anomaly can be acquired using this statistical model and techniques of K-means clustering and neural network. In most anomaly detection methods, the forecasted values are compared with the test data, and the difference between these two series is calculated as an outlier score. Finally, anomalies are detected based on these outlier scores. We consider the anomaly detection for two modes. First, in a less cautious manner, where the anomaly detection is being conducted less guardedly, each record that is identified as an anomaly by at least one of the methods would be considered as an anomaly. In the second mode, which detects the anomalies more accurately, a record is considered anomaly only if it is identified as an anomaly by all the detection methods.

3.1. Dataset. In this paper, to recognize the anomaly behavior of users, we study the CDR dataset from a particular mobile phone operator over a period of 3 months. The data used in this paper are the anonymized CDR from one of the largest mobile phone operators in Iran. These records are gathered from 21 December, 2016, to 20 March, 2017, in a commercial area of a large city. CDR data are utilized for understanding the activity pattern of the user and identifying the abnormal behavior. The dataset had the activity logs for every five minute interval separately for call in and call out. We summed up the activities to calculate the log details for one-hour time interval.

3.2. Model Selection. First, we represented data as a time series (see Figure 1). It seems that the mean and the variance are not constant over time, so the Leven test and ANOVA are used for investigating the stationary of these moments. Figure 2 illustrates that the variance is not constant because all lines do not overlap with each other, and also the number of time series data was 2160 points. To use the Leven test, we divided this number of data into 54 groups of 40. SSS is the number of groups. It also can be seen that the p value is equal to zero, so the null hypothesis (equality of variance) is rejected. In Figure 3, it is clear that the mean of the time series is increased over time, so we conclude that the mean is not stationary. Due to this instability, data transformation is needed. The data are not still fixed after several uses of the differentiation technique, so Box-Cox transformation is applied. It is seen that the data remain unstable when the Leven test is carried out, so AR, MA, ARMA, and ARIMA models are not suitable for this data. In this situation, more advanced methods, such as the GARCH model, should be used. This method only stabilize the mean but also because of its structure that automatically makes the variance stationary.

3.3. GARCH Model. The GARCH model is utilized for the training data. In this situation, predicted time series and test time series are compared with each other, and their difference is considered as an anomaly point. Then, the threshold level is defined. We chose a threshold based on minimum error. Drawing an error plot in the threshold, we saw a linear decrease in error by decreasing the threshold until we reached a point where the reduction in threshold led to an increase in error. We stopped at this point and considered it as a threshold. We compared the difference between the predicted time series and test time series with this threshold; if this difference is more than the threshold level, it will be an anomaly. In Figure 4, the black line is the threshold level, red points are differences between predicted and test time series, and blue triangular points above the threshold level line are the anomaly.

3.4. K-Means Clustering. Parameter K is defined equal to 2 because there are both sets of normal and anomaly. Figure 5 shows the number of calls versus the time that anomalies are shown with blue color, which is acquired by the K-means method; likewise, red color data are normal.

3.5. Neural Network Autoregressive. Like the previous section, the data are divided into two parts: the training and test data. First, a neural network model is fitted to the training data. The fitted model is NNAR (29,15) which has fifteen

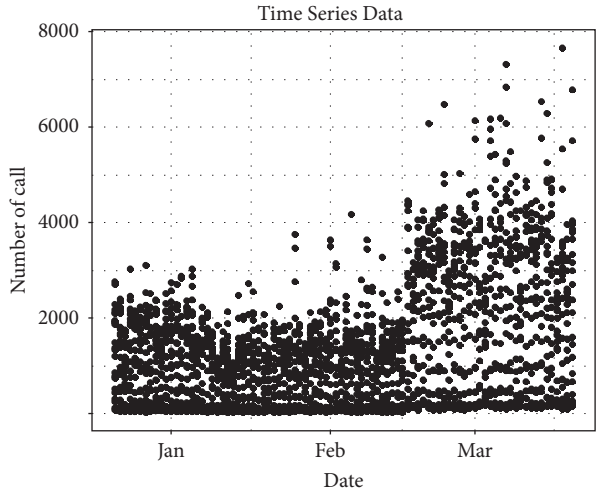

Figure 1: Time series data.

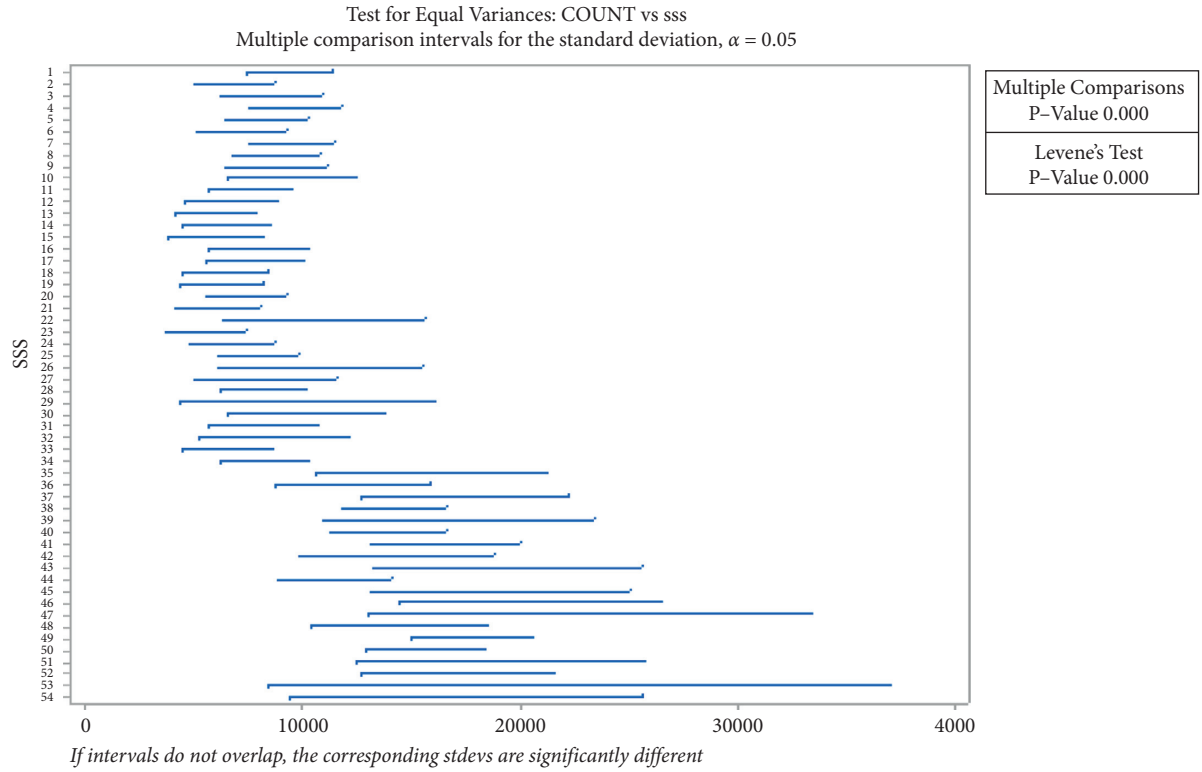

Figure 2: Leven test for evaluating the stationary of variance.

neurons in the hidden layer, and 29 last observations $(x_{t-1}, \ldots, x_{t-29})$ are used as internal data. In the next step, the neural network model uses training data to predict. Then, the predicted data are compared to the test data, and their differences are considered the anomaly. According to the previous section, the first threshold level is defined, and all points above the threshold level are the anomaly, as shown in Figure 6.

3.6. Hybrid Model. The hybrid model uses three methods: GARCH, K-means, and neural network. This method can detect anomalies in two different ways. Firstly, the detection of abnormality is done cautiously, and each record, which is recognized as an anomaly by at least one method, is considered an anomaly. Still, in the second type, a record can be an anomaly if all three methods detect it as an anomaly.

3.6.1. First Mode. In this method, a record is anomalous if at least one of the three methods identified it as an anomaly. After detection and verification of anomalies, we can also determine the date and time where such abnormalities occur. For example, in Figure 7, anomalies at 17 o'clock on 2

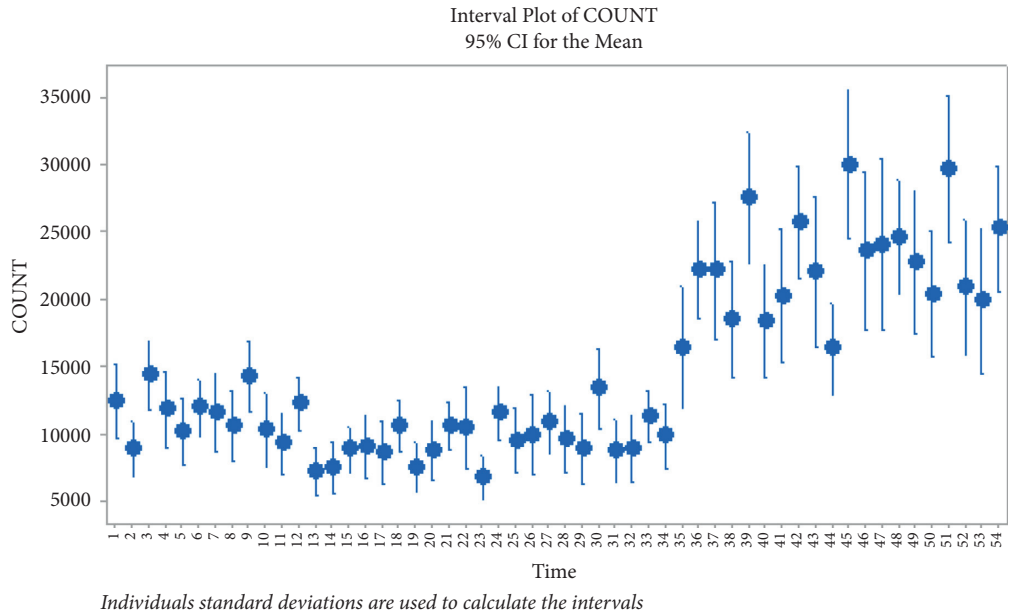

Figure 3: ANOVA test for evaluating the stationary of mean.

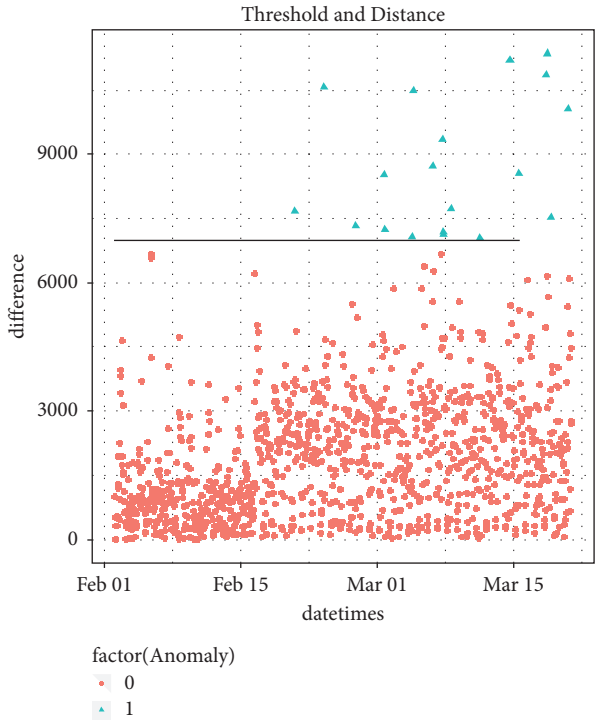

Figure 4: Anomaly detection for GARCH model; anomaly (triangle).

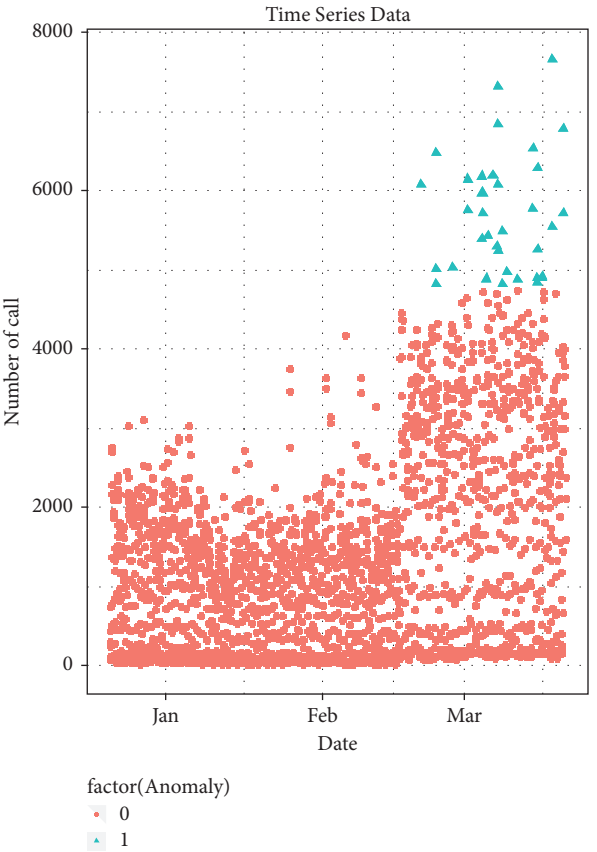

Figure 5: Anomaly detection by the use of K-means; anomaly (triangle) and normal data (circle).

February to 20 March are shown. This figure demonstrates that at 17 o'clock, three anomaly points are known. These anomalies occurred on March 1, 6, and 7. This is because March is the last month in Iran's yearly calendar. Afterward, the New Year is celebrated, which might be a reason for encountering such anomalies in the number of calls in a commercial area where people go for shopping. All the predicted values are higher than the real values, indicating that the reason for these anomalies was not the failure of the telecommunication systems, but the more significant number of people who attend the area, the possible reason for which was mentioned above.

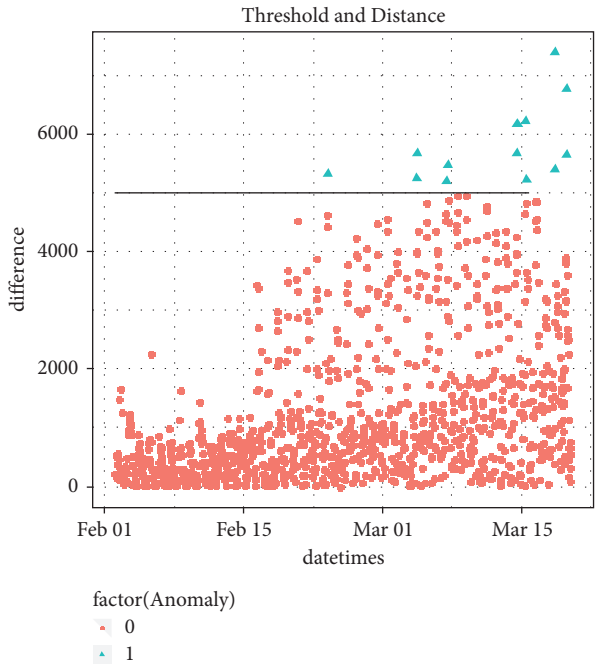

Figure 6: Anomaly points (triangle) and normal points (circle) in the neural network model.

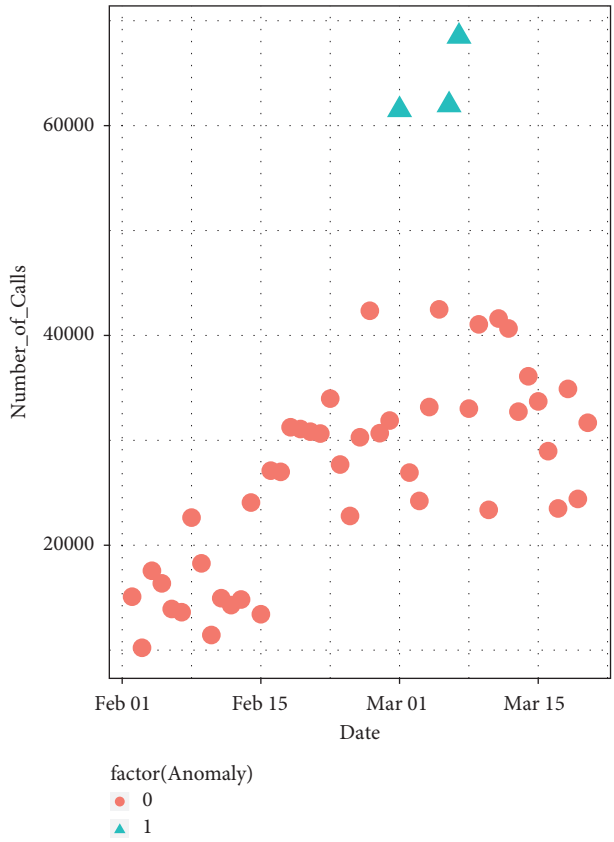

Figure 7: Anomaly detection at 17 o'clock for the first mode.

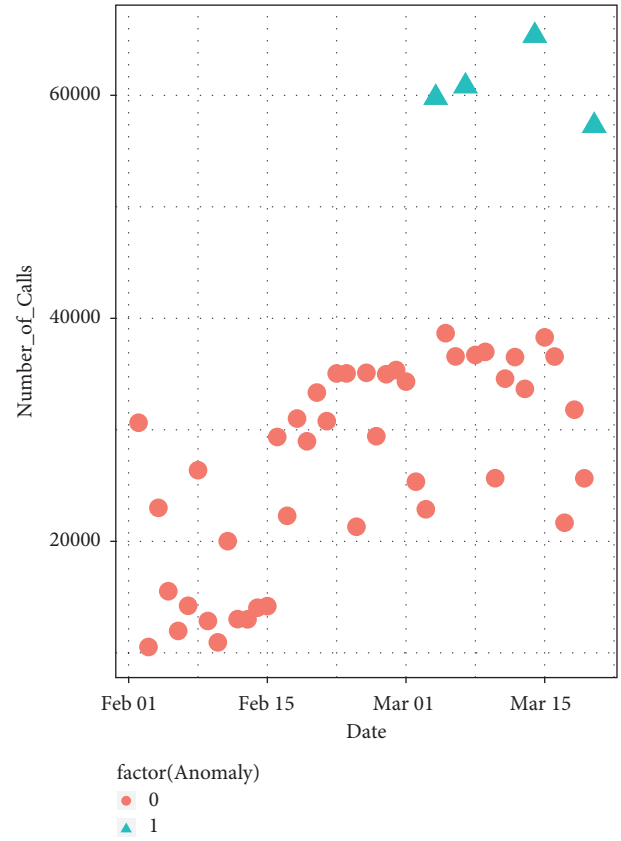

Figure 8: Anomaly detection at 15 o'clock for the second mode.

3.6.2. Second Mode. In this method, every record which is detected as an anomaly with all three methods (K-means, GARCH, and neural network) is considered an anomaly. After using this method, we can identify anomaly points and recognize the date and time in which these anomalies occur. For example, in Figure 8, anomalies at 15 o'clock are shown. The anomalies occurred at 15 o'clock on March 4, 7, 14, and 20. These anomalies are happened because of the nearness to the Iranian New Year.

3.7. Logistic Regression. Some features, such as days, nights, or day time, and the number of calls, are chosen for finding the causes of anomalies and what features are effective, so hypothesis testing is exploited. These features are selected based on domain expert knowledge and existing work on anomaly detection in telecommunication data usage. The null hypothesis is that the coefficient of each element is zero. Likewise, the alternative hypothesis is that the coefficient of every feature is not zero. The coefficients in which p values are very low can be effective in the response variable.

By applying logistic regression on the number of calls in every hour, we conclude that two features of Friday (weekend of Iranian people) and number of calls are effective

Table 2: Accuracy and FPR for hybrid model.

Hybrid model	Accuracy (%)	FPR (%)
Fist mode	99.72	0.01
Second mode	99.68	0.012

Table 3: Improvement of first mode in accuracy and FPR.

Literature	Accuracy (%)	FPR
Parwez et al. [2]	9.72	—
Imran et al. [18]	5.72	—
Hussain et al. [19]	6.93	14.12%
Hussain et al. [20]	5.32	1.69%
Hussain et al. [21]	0.92	0.43%

Table 4: Improvement of second mode in accuracy and FPR.

Literature	Accuracy (%)	FPR
Parwez et al. [2]	9.68	—
Imran et al. [18]	5.68	—
Hussain et al. [19]	6.89	14.118%
Hussain et al. [20]	5.28	1.688%
Hussain et al. [21]	0.88	0.428%

in anomalies. The effectiveness of the number of calls in anomaly is evident because the anomaly is defined based on this feature. On Friday, the coefficient was −2.397 that means the odd ratio on Friday to other days is equal to $e^{(-2.397)} = 0.091$, so most of the anomalies have happened on the days of the week except Friday.

3.8. Error. Lower accuracy and high FPR are two main limitations of the latest approaches for anomaly detection in cellular networks. By comparing acquired anomaly points with data labels, the accuracy and ratio of false positive are calculated. These results are shown in Table 2 for the first mode and the second mode. The preliminary results in Table 1 clarify the facility and superiority of our hybrid model for anomaly detection in terms of the first mode and the second mode. Tables 3 and 4 show the improvement in accuracy and FPR for the first mode and the second mode, respectively. These results are obtained due to comparing our hybrid model with the results in Table 1.

4. Conclusion

In this paper, we operated some CDR data (i.e., the hourly number of calls in the time series) to identify anomaly behavior patterns in subscribers' usage. Three methods (i.e., GARCH, K-means, and neural networks) have been adapted to suggest a prediction method. This type of information is well studied in the literature in terms of anomaly detection, and the innovation of this paper is in using the prediction algorithm in a combination of these three methods. The decision is made based on the conclusion of the three used predictors. Solely, the algorithms have been used as a voting classifier to make the final decision if there is an anomaly usage or not. We called the new method the hybrid model and investigated it in the first and second modes. We concluded that this method helps us to achieve high accuracy rates and low FPR. So, by the identification of unusual events, proper action such as resource distribution and sending small drone cells can be taken in advance and on time. Hence because of such actions, the users' requirements will be fulfilled and will have the best QoS, and network congestion will be avoided. Besides, by using logistic regression, we determined which features have a more significant role in the occurrence of anomalies in this type of data. The restrictions in conducting this study were the limited set of data. For future work, we can predict and detect anomalies with different methods such as bootstrapping, vector autoregressions, and complex seasonality.

Data Availability

The data used in this paper are the anonymized CDR from one of the largest mobile phone operators in Iran. So, data are not available due to commercial restrictions.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[2] 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.

[3] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, Wiley, Hoboken, NJ, USA, 2008.

[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] A. Yaacob, I. Tan, S. Chien, and H. Tan, "ARIMA based network anomaly detection," in *Proceedings of the Second International Conference on Communication Software and Network*, pp. 205–209, Bangalore, India, March 2010.

[6] T. Andrysiak, L. Saganowski, M. Maszewski, and A. Marchewka, "Detection of network attacks using hybrid ARIMA-GARCH model," in *Proceedings of the Twelfth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*, pp. 1–12, Brunow, Poland, July 2018.

[7] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *Proceedings of the IEEE Conference on Computer Communications*, pp. 1806–1814, Toronto, Canada, May 2014.

[8] Nithi and L. Dey, "Anomaly detection from call data records," in *Proceedings of the International Conference on Pattern Recognition and Machine Intelligence*, pp. 237–242, New Delhi, India, December 2009.

[9] V. Soto and E. F. Martinez, "Automated land use identification using cell-phone records," in *Proceedings of the 3rd ACM International Workshop on MobiArch*, pp. 17–22, New York, NY, USA, June 2011.

[10] M. Amer, "Comparison of unsupervised anomaly detection techniques," B.Sc Thesis, Multimedia Analysis and Data Mining Competence Center German Research Center for Artificial Intelligence, Kassel, Germany, 2011.

[11] M. F. Lima, B. B. Zarpelao, L. H. Sampaio, J. J. PC. Rodregues, T. Abrao, and M. L. Proenca, "Anomaly detection using baseline and K-means clustering," in *Proceedings of the International Conference on Software Telecommunications and Computer Networks*, pp. 305–309, Kochi, India, September 2010.

[12] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2015.

[13] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," in *Proceedings of the 20th European Wireless Conference*, Maui, HI, USA, May 2014.

[14] D. Jiang, Y. Wang, Z. Lv, S. Qi, and S. Singh, "Big data analysis based network behavior insight of cellular networks for industry 4 applications," *IEEE Transactions on Industrial Informatics*, vol. 16, 2020.

[15] S. Papadopoulous, 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.

[16] H. S. Wu, "A survey of research on anomaly detection for time series," in *Proceedings of the International Computer Conference on Wavelet Active Media Technology and Information Processing*, Chengdu, China, December 2016.

[17] E. Geepalla, N. Abuhamoud, and A. Abouda, "Analysis of call detail records for understanding users behavior and anomaly detection using Neo4j," in *Proceedings of the 5th International Symposium on Data Mining Applications*, pp. 74–83, Riyadh, Saudi Arabia, March 2018.

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

[19] B. Hussain, Q. Du, and P. Ren, "Semi-supervised learning based big data-driven anomaly detection in mobile wireless networks," *China Communications*, vol. 15, no. 4, pp. 41–57, 2018.

[20] B. Hussain, Q. Du, and P. Ren, "Deep learning-based big data-assisted anomaly detection in cellular networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, December 2018.

[21] B. Hussain, Q. Du, S. Zhang, A. Imran, and M. A. Imran, "Mobile edge computing-based data-driven deep learning framework for anomaly detection," *IEEE Access*, vol. 7, pp. 137656–137667, 2019.

[22] T. Sui, X. Tao, S. Xia et al., "A real-time hidden anomaly detection of correlated data in wireless networks," *IEEE Access*, vol. 8, pp. 60990–60999, 2020.

[23] D. Cortés-Polo, L. I. J. Gil, J.-L. González-Sánchez, and J. Carmona-Murillo, "A quantitative and comparative evaluation of key points selection algorithms for mobile network data sets analysis," *IEEE Access*, vol. 9, pp. 92030–92042, 2021.

[24] Q. Zhu and L. Sun, "Big data driven anomaly detection for cellular networks," *IEEE Access*, vol. 8, pp. 31398–31408, 2020.

[25] S. Sudhakaran, A. Venkatagiri, P. A. Taukari, and A. Jeganathan, "Metropolitan cellular traffic prediction using deep learning techniques," in *Proceedings of the IEEE International Conference, Networks and Sattelite*, Batam, Indonesia, December 2020.

[26] M. Li, Y. Wang, Z. Wang, and H. Zhang, *A Deep Learning Method Based on an Attention Mechanism for Wireless Network Traffic*, Elsevier, Amsterdam, Netherlands, 2020.

[27] R. J. Hyndman and G. Athanasopoulos, *Forecasting Principle and Practice*, OTexts, Melbourne, Australia, 2018.

[28] D. M. Diez, C. D. Barr, and M. C. Rundel, *OpenIntro Statistics*, CreateSpace, Scotts Valley, CA, USA, 2015.

[29] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine learning for anomaly detection: a systematic review," *IEEE Access*, vol. 9, pp. 78658–78700, 2021.

[30] L. Li, S. Dai, Z. Cao, J. Hong, S. Jiang, and K. Yang, *Using Improved Gradient Boosted Decision Tree Algorithm Based on Kalman Filter (GBDT-KF) in Time Series Prediction*, Springer, New York NY, USA, 2020.

[31] G. V. Houngbonon, E. L. Quentrec, and S. Rubrichi, "Access to electricity and digital inclusion: evidence from mobile call detail records," *Humanities and Social Science Communication Journal*, vol. 8, 2021.

[32] G. Pestre, E. Letouze, and E. Zagheni, *The ABCD of Big Data: Assessing Biases in Call Detail Records for Development Estimates*, The World Bank Economic Review, 2020.

[33] G. Zhang, X. Rui, S. Poslad, X. Song, Y. Fan, and B. Wu, "A method for the estimation of finely-grained temporal spatial human population density distributions based on cell phone call detail records," *Remote Sensing*, vol. 12, no. 16, p. 2572, 2020.