International Journal of Machine Learning and Cybernetics (2022) 13:3159–3172
https://doi.org/10.1007/s13042-022-01586-8

ORIGINAL ARTICLE

Detection and quantification of anomalies in communication networks based on LSTM-ARIMA combined model

Sheng Xue[1] · **Hualiang Chen**[1] · **Xiaoliang Zheng**[2]

Received: 6 August 2021 / Accepted: 15 May 2022 / Published online: 17 June 2022

Abstract
The anomaly detection for communication networks is significant for improve the quality of communication services and network reliability. However, traditional communication monitoring methods lack proactive monitoring and real-time alerts and the prediction effect of a single machine learning model on communication data containing multiple features is not ideal. To solve the problem, A prediction-then-detection anomaly detection method was proposed, and quantitative assessment of network anomalies was developed. Specifically, anomaly-free data was obtained by eliminating outliers, and the long short-term memory (LSTM) and autoregressive integral moving average (ARIMA) were combined via residual weighting to predict the future state of the key performance indicators (KPI) without outliers. Anomalies were identified using the error comparison between the prediction and actual values, and the network condition was quantified using the scoring method. It is observed that the proposed LSTM-ARIMA hybrid model has better prediction effect, which can well represent the performance of KPIs of the future state, and the prediction-then-detection anomaly detection method has excellent performance on both precision and recall.

Keywords Anomaly detection · Prediction · Hybrid model · Quantification

1 Introduction

Communication network has long become an important way of information exchange. With the rapid development of intelligence and information technology, the security requirements of communication network are getting more and more important. Network management is an important technology for maintaining the security of communication networks. Meanwhile, network monitoring is the most fundamental part of network management, aiming to improve communication service quality, resource utilization, and network reliability. Traditional communication network monitoring is achieved through reporting alarms, but with the complication and diversification of the network, only reporting alarms cannot satisfy the requirements of proactive monitoring and real-time alarms.

Analyzing the anomalies of network performance index data allows proactive monitoring for abnormal or defective conditions in the communication network. The threshold method is the most commonly used method in network anomaly detection, Ref. [1] applied statistical methods to set thresholds and successfully detected anomalies in the traffic flow. Ref. [2] utilized ARMA (Autoregressive moving average model) to predict confidence intervals to set thresholds for network outlier detection and achieved better results. Ref. [3] adopted firewalls, intrusion detection systems and intrusion prevention systems to calculate baseline and mean and standard deviation to define thresholds. The threshold method is essentially a method based on statistical hypothesis testing, which is too static and cannot reflect the dynamic changing characteristics of network performance indicators. The general threshold value is determined based on the experience of the operation and maintenance personnel, however, a narrow threshold range will lead to false alarms and a wide will lead to missed alarms. The setting of the threshold value is the key to the threshold method, which directly affects the accuracy of anomaly detection.

✉ Hualiang Chen
hualiang_chen520@163.com

[1] School of Safety Science and Engineering, Anhui University of Science and Technology, Tianjiaan District, Huainan 232001, Anhui, China

[2] School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001, China

Intrusion detection technology is an effective way to deal with network threats, Anomaly-based network intrusion detection is an important research and development direction of intrusion detection [4], and machine learning techniques are becoming the primary tools for network intrusion detection with the popularity of machine learning.Machine learning-based anomaly detection methods for communication networks can be divided into three types: supervised learning, unsupervised learning and semi-supervised learning. According to the supervised approach, labeled data are exploited to single out a feature subset to classify network anomalies [5]. As in Ref. [6], support vector machines (SVM) was used to distinguish disk operating system (DOS) attack patterns and detect anomalous network conditions. Ref. [7] applied particle swarm algorithm and gravitational search algorithm to optimize random forest for intrusion detection. Furthermore, many solutions have been investigated by supervised learning [8–12]. Unsupervised learning uses unlabeled samples to train the model and is a better solution in the absence of sufficient prior knowledge and labels [13, 14]. Ref. [15] adopted clustering methods for anomaly detection in networks and proposed the data-driven distance metric to deal with clustered network anomalies. Ref. [16] utilized a sample of log files of the network for unsupervised learning to monitor the network in real time. In Ref. [17], an unsupervised deep learning framework was constructed to monitor and visualize network anomalous discrete sequences such as payload and SYSCALL traces. Semi-supervised approach is based on a mixed strategy, striving to enrich an unlabeled set with some labeled data, so as to improve the feature selection phase [5]. For example, Ref [18] proposes the SemiADC model for semi-supervised anomaly detection in dynamic communication networks, and experimental evaluation on real-world datasets demonstrates the effectiveness of SemiADC. Unsupervised and the semi-supervised approaches exhibit the drawback of neglecting potential correlations among features, resulting in the loss of crucial (as well as deterministic) piece of information. On the other hand, a supervised approach can offer optimal results, provided that the data are correctly labeled. This case typically occurs in a controlled network environment, where, with the help of network analyzers, it is possible to automatically label the type of passing data traffic [5]. Therefore, this paper uses supervised learning for anomaly detection.

In recent years, work on network traffic (a kind of KPI in communication network) prediction has attracted the attention of researchers. Hanyu Yang et al. [19] utilized the (simulated annealing) SA-optimized (autoregressive integrated moving average model) ARIMA-BPNN (back propagation neural network) model to predict network traffic with some improvement in accuracy over the traditional network traffic prediction model. Guo et al. [20] designed a dynamic modification neural network model to select different neural networks and predict network traffic, which decreased the prediction value error and time skew of the inflection points. ARIMA model is a classical forecasting model, and many investigations have indicated that it has good performance in stationary sequence forecasting. For instance, Kim et al. [21] applied an ARIMA model for dynamic bandwidth provisioning in prior prediction of traffic. Salman [22] adopted ARIMA to weather forecasting tasks with excellent results. Extensive investigations have confirmed the excellent performance of LSTM models for nonlinear and seasonal data prediction, especially for time series prediction. Luo et al. [23] long short-term memory (LSTM) combined with XGboost to predict the number of COVID-19 infections in the United States. Song et al. [24] utilized (convolutional neural networks) CNN-LSTM to predict the heating load of smart district heating system, the CNN was used to extract the spatial characteristics of the heating load and the LSTM was used to extract the temporal characteristics of the heating load, which has evident accuracy advantage. In addition, LSTM has been applied to the prediction of power generation [25], remaining engine life [26], and coal mine safety [27] with excellent results. The above literature review demonstrates that both ARIMA and LSTM have unique advantages in time series processing. References [28–30] proved that hybrid methods could effectively improve forecasting accuracy and complement each other and disadvantages of previously proposed models, enabling the actual data to be captured and the forecasting accuracy to be improved [31]. In this paper, the advantages of ARIMA and LSTM were combined to predict the network KPIs.

The rest of this paper is organized as follows: The second part introduces the algorithmic models used in this paper, including new communication network anomaly detection methods, LSTM, ARIMA and combined models. In the third part, the hybrid model and the proposed anomaly detection method are experimented and analyzed. Our paper closes with summary of the researches methods in the article.

2 Model description

2.1 LSTM

Recurrent neural networks (RNN) are able to connect previous information to the current task and have a unique advantage in dealing with time series. However, when the time interval becomes large, RNN will have problems such as gradient disappearance and explosion. Long and short-term memory network (LSTM) solves such problems by using long and short-term memory units, and the principle of LSTM is described below.

Figure 1 shows the structure diagram of LSTM. The LSTM is consisted of a cell state and three gates, where the forgetting gate, shown in Fig. 1a, is responsible for discarding some information from the cell state C_{t-1}, it will read the information from h_{t-1} and x_t, output a number from 0–1 to the cell state C_{t-1}, and decide how much information to keep. As in Eq. (1):

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

where, f_t is the output of the forgetting gate, W_f is the weight matrix of the forgetting gate, b_f is the bias, $\sigma(x)$ is the activation function, which is generally a sigmod function, h_{t-1} is the input of the cell at the previous moment, and x_t is the input of the cell at the current moment.

The input gate, shown in Fig. 1b, is responsible for allowing a certain amount of information from the current moment to be added to the cell state. It determines the updated information i_t through the function σ, and forms the alternative update message $\tilde{C}_t$ via tanh. As in Eqs. (2) and (3):

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{3}$$

where, W_i and W_C are the input gate weight and the weight matrix of the tanh activation function, and b_i and b_C are the bias vectors.

The updated cell state C_t is obtained from Eqs. (2) and (3) as:

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{4}$$

The output gate determines what information will be output, and its equation is shown in Eqs. (5) and (6):

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{6}$$

where, W_o is the output gate weight matrix and b_o is the bias vector.

2.2 ARIMA

ARIMA is a typical time series model consisting of three components, AR (auto-regression), MA (moving average), and differencing degree d [19, 32]. It is suitable for dealing with stable time series problems and can be expressed as:

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} \delta_j \varepsilon_{t-j} \tag{7}$$

where, p represents the auto-regression parameters, q represents the moving average parameters, y_t represents the model parameters to be estimated, a_i is the autocorrelation

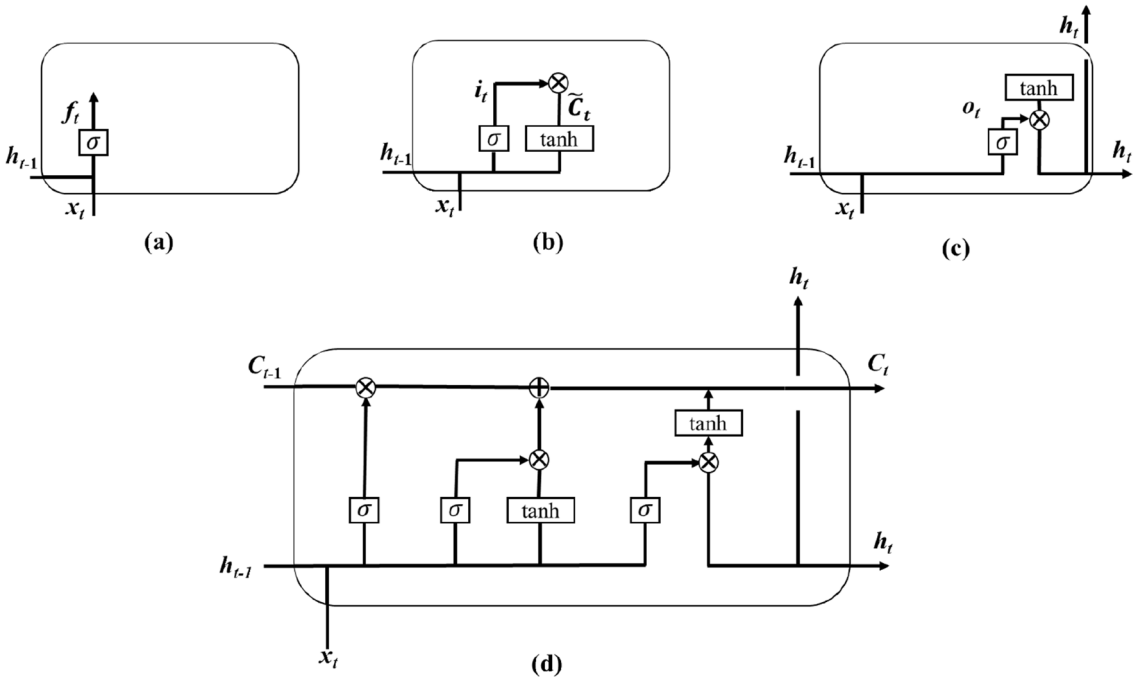

Fig. 1 Structure diagram of LSTM

coefficient and ε_{t-j} is the white noise. When the sequence is unstable, the sequence should be stabilized by data differencing, d is the number of times the sequence becomes stable by differencing.

Seasonal ARIMA (SARIMA) [33] adds seasonal components (P, D, Q) to ARIMA, and this paper uses the SARIMA model for rolling forecasts [34, 35].

2.3 Model of LSTM-SARIMA combination based on residual assignment

For time series forecasting of multiple indicators, a single model often does not predict multiple indicators well, so we apply a hybrid model to solve this problem. In this paper, we applied a hybrid model based on residual weighting to predict network performance metrics. The residual weighting gives the corresponding weight to the model at that moment based on the magnitude of the predicted residuals of a single model at the previous moment, the calculation formulation is:

$$\begin{cases} f(x_t) = \frac{1}{n}\sum_{i=1}^{n} \omega_i(t-1) f_i(x_t) \\ \omega_i(t-1) = \frac{\frac{1}{\bar{\varepsilon}_i(t-1)}}{\sum_{i=1}^{n}\frac{1}{\bar{\varepsilon}_i(t-1)}} \\ \text{s.t.} \sum_{i=1}^{n} \omega_i(t-1) = 1, \omega_i(t-1) \geq 0. \end{cases} \tag{8}$$

where, $\omega_i(t-1)$ is the weight of the i-th model at moment t-1, $\bar{\varepsilon}_i(t-1)$ the squares of the prediction errors of the i-th model at moment t-1, and n is the number of models.

2.4 Anomaly detection method

In the field of communication networks, the fixed-threshold method is often used for anomaly detection of network KPIs. Although the process is simple, the fixed-threshold method often leads to unsuitable thresholds for detection of time-series data with obvious periodicity in communication networks, such as traffic and uplink PRB utilization. As shown in Fig. 2: threshold 1 selects the maximum value of the workday's KPI, which causes most of the data from the week days and festivals to be mistakenly assumed to be abnormal. Threshold 2 selects the maximum value for festivals, which also leads to detection of no abnormalities and is also unrealistic.

In this paper, we adopt the anomaly detection approach of prediction-then-detection. As shown in Fig. 3, historical data is used to predict KPIs for future period, and the predicted value is compared with the actual value, when the status is far apart from the predicted status at one timestamp, it is considered an anomaly. Process chart of anomaly detection is shown in Fig. 4. This method solves the shortcoming that

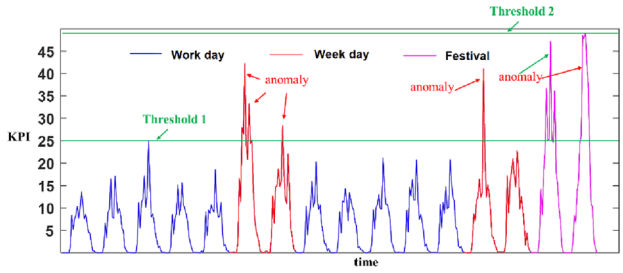

Fig. 2 Fixed-threshold method

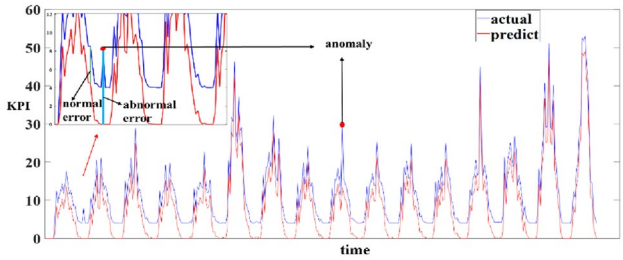

Fig. 3 Prediction-then-detection method

the fixed-threshold method cannot handle periodic problems, while the accuracy of anomaly detection depends on the performance of the prediction model. Specifically:

1. Data pre-processing, including outlier processing and normalization, is used to predict KPI data in the absence of anomalies.
2. The LSTM and SARIMA models were used to predict the pre-processed data, respectively.
3. The hybrid model was obtained using the residual assignment method. Use the hybrid Model to predict the pre-processed data.
4. Calculate the error and anomaly level of each point compared with the real data.
5. The anomaly classification model is constructed using random forest to obtain the weights of each indicator.
6. The weights are multiplied with the anomaly level of each indicator to get the anomaly level of the communication network.

3 Experiments and analysis

3.1 Data processing

The experimental data in this paper are obtained from Huainan City China Mobile. It is cell-level network history

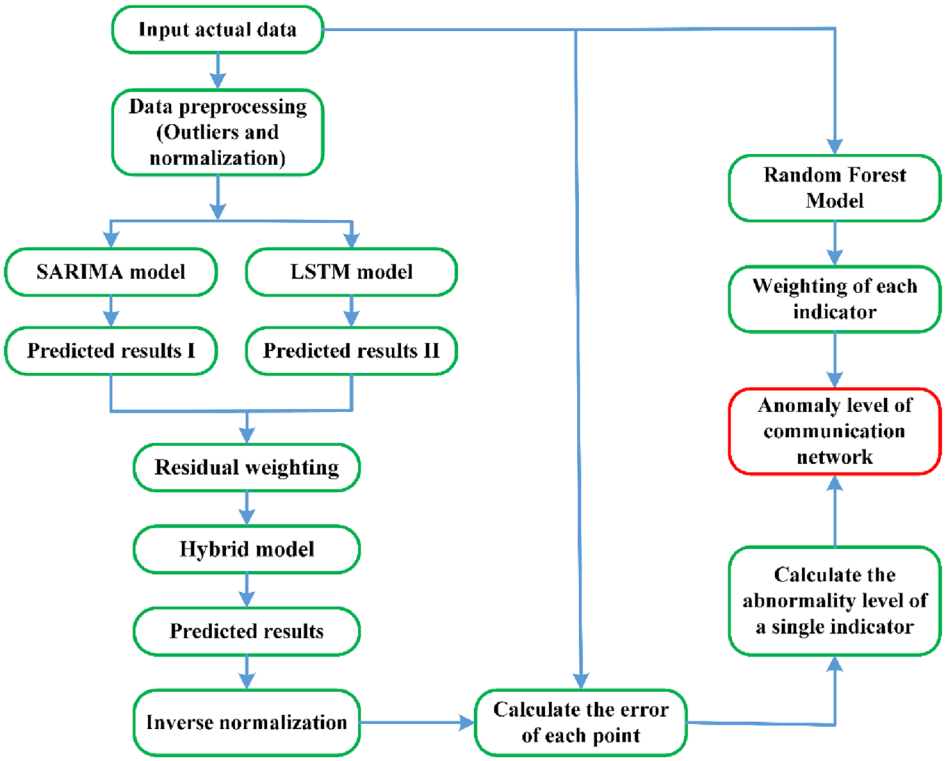

Fig. 4 Process chart of anomaly detection

performance for the time period of March 10, 2020–June 11, 2020, with data collected at a granularity of 1 h, i.e., 24 times per day, and the moment of collection is every hour on the whole hour. The main KPIs are RRC connection success rate, handover success rate, traffic, wireless connection rate, RRC reconstruction rate, uplink PRN utilization, etc., as shown in Fig. 5.

From Fig. 5, we observed that the traffic has the same trend as the uplink PRB utilization, and the RRC connection success rate has the same trend as the wireless connection rate, showing that we only need to select four of these KPIs for anomaly quantitative evaluation. In this paper, RRC connection success rate, handover success rate, RRC reconstruction success rate, and uplink PRB utilization were selected as inputs for anomaly evaluation.

Before forecasting, we need to process the outliers in the original data to better predict the KPIs in the normal state. Given the periodicity and tendency of the KPI time series, outliers were handled in the following manner:

$$\begin{cases} y(t) = \frac{y(t-1)+y(t+1)}{2} y(t) \text{ ii outlier}, y(t-1) \text{ and } y(t+1) \text{ is normal} \\ y(t) = y(t-24) y(t) \text{ is outlier}, \ y(t-1) \text{ or } y(t+1) \text{ is also outlier} \end{cases} \tag{9}$$

After outlier processing, four KPIs were obtained as illustrated in Fig. 6.

In order to eliminate the increase in model training time caused by anomalous sample points and to make the model learn the laws of time series better, the normalization method was used as follows:

$$x^1 = \frac{x - min(x)}{max(x) - min(x)} \tag{10}$$

where, x^1 is the normalized data, x is the sample data, and $max(x)$ and $min(x)$ are the maximum and minimum values in the sample data. The normalized data is divided into training set and test set according to the ratio of 7:3.

3.2 Predcition

For the four KPIs mentioned in Sect. 3.1, a single model does not predict each KPI well. Take the example of PRB utilization and wireless success rate. Experiments founded that LSTM is suitable for the prediction of uplink PRB utilization, On the contrary, the SARIMA rolling prediction model has better performance in handover success rate. Rationale is that SARIMA rolling regression is suitable for data with excellent stationarity, and LSTM is suitable for nonlinear data with prominent temporal characteristics. The stationarity of the handover success rate is better than that of the uplink PRB utilization, while the time periodicity of the uplink PRB utilization is more obvious. Figure 7 illustrates the prediction effect of the two prediction models on the uplink PRB utilization and handover success rate.

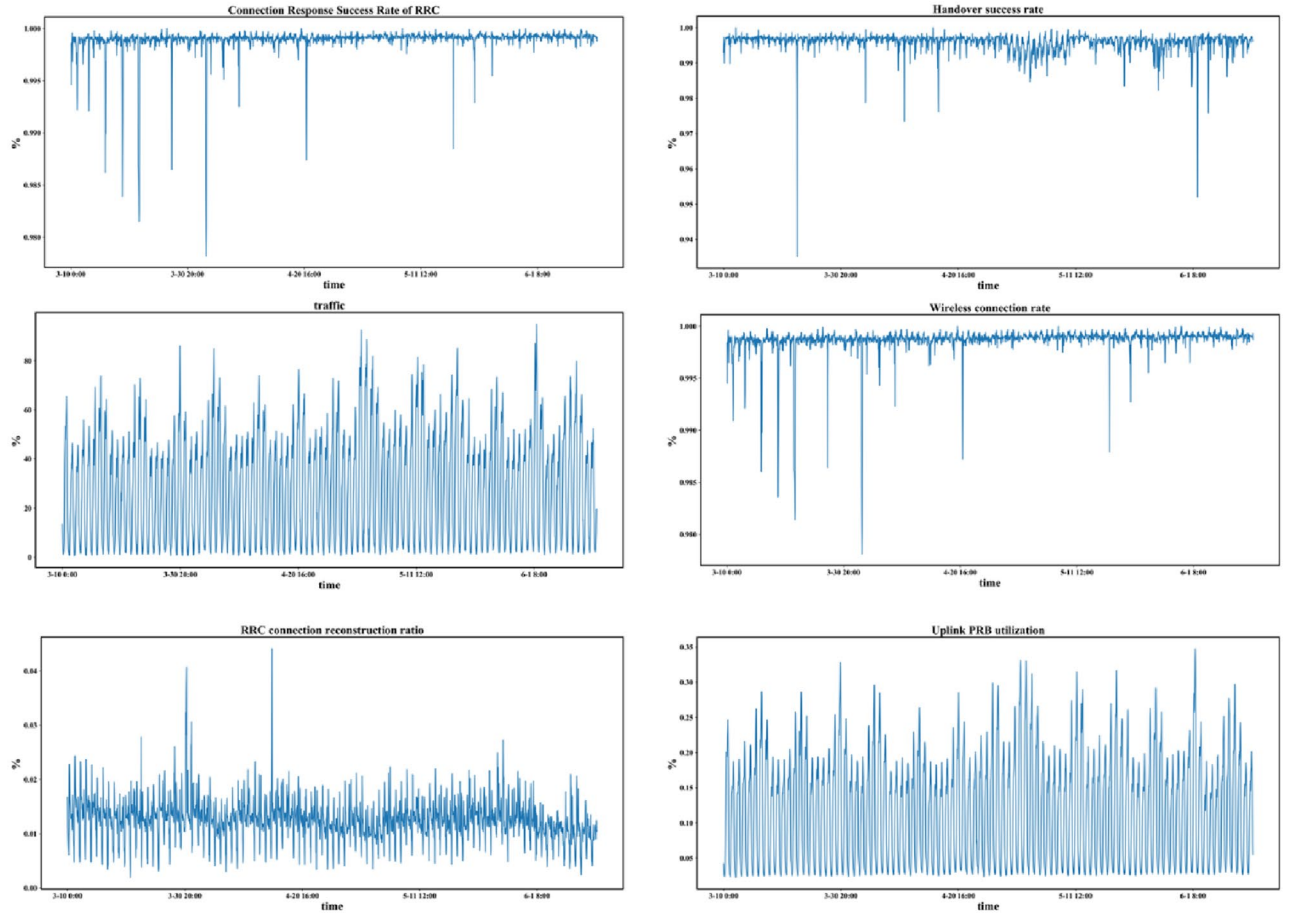

Fig. 5 KPI data

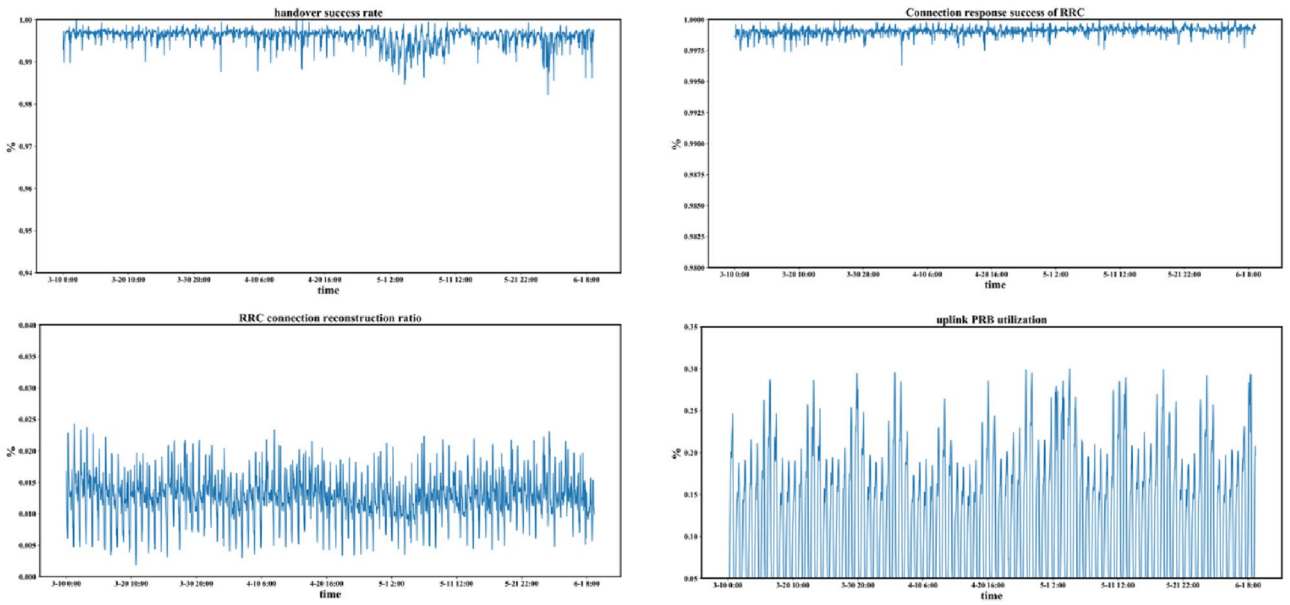

Fig. 6 KPIs after outlier processing

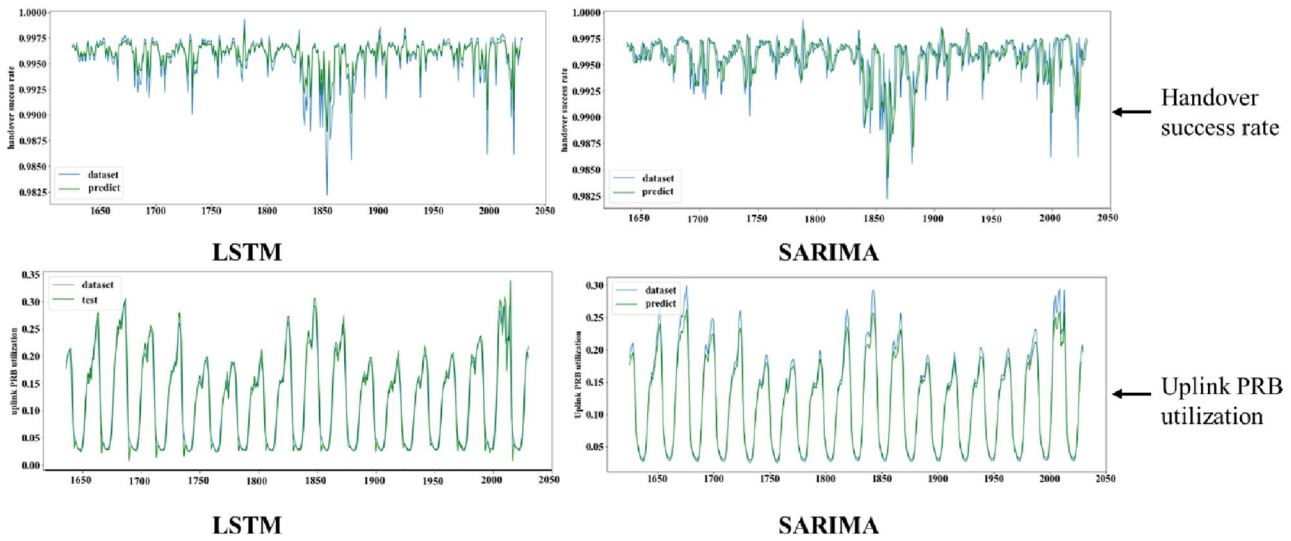

Fig. 7 Prediction effect of LSTM and SARIMA for two KPIs

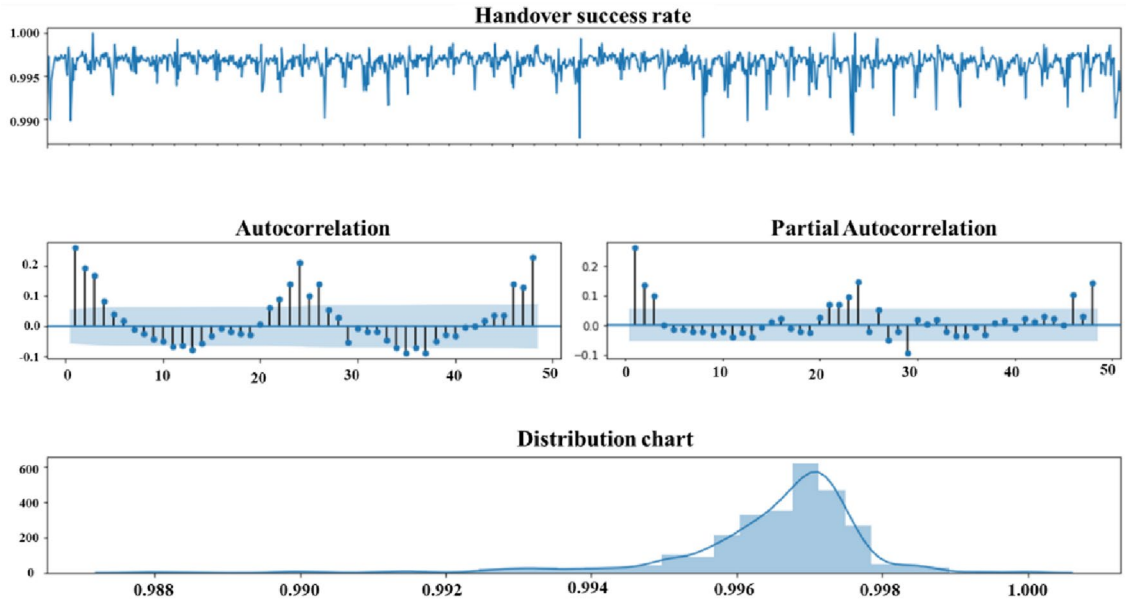

Fig. 8 Smoothness test chart of handover success rate

Aiming at the problem that the single model cannot be applied to the prediction of four types of KPIs. In this paper, a hybrid model based on residual weighting LSTM-rolling SARIMA is proposed. Various parameters of both models need to be determined before the hybrid model predictions to be made.

A stationarity test is required before using SARIMA, after which the values of p, d and q are determined based on ACF and PACF plots. The following is a stationarity analysis using the handover success rate as an example. As indicated in Fig. 8, the stationarity of the handover success rate data is not satisfactory. After performing the first-order difference for the handover success rate to have a better prediction by SARIMA, the results are shown in Fig. 9. The ADF test and white noise test (Ljung-Box test) were performed after the first-order difference of the data, the results are shown in Table 1. As can be seen from Table 1, T value are less than 1% value, 5% value and 10% value, P value < 0.05, so there is no unit root and it is a stationary data. All LB_P value < 0.05 means that the data is not white noise data.

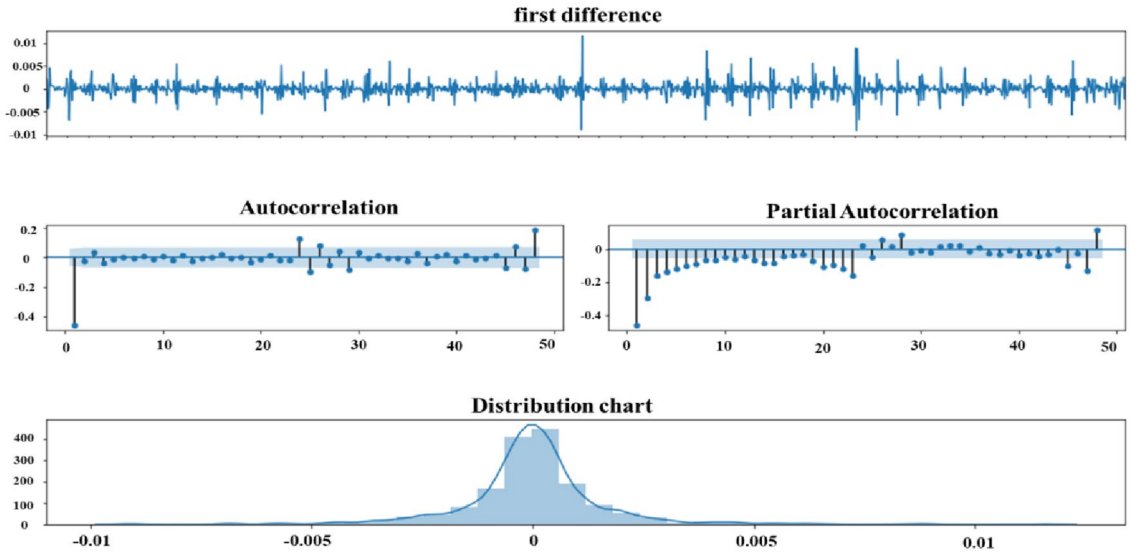

Fig. 9 Smoothing test chart after first-order differencing

Table 1 The results of ADF test and white noise test

ADF test		White noise test	
T value	− 3.962659	LB_P value1	1.73806130e−170
P value	0.001619	LB_P value2	6.52940306e−175
1% value	− 3.340311	LB_P value3	1.70124017e−177
5% value	− 2.863166	…	…
10% value	− 2.567635	LB_P value	1.35309266e−199

As indicated in Fig. 9, the data after first-order differencing is stable with $d=1$. Then the trailing and ending of the ACF and PACF graphs show that $p=0$ and $q=1$. Both ACF and PACF have anomalous changes at lag = 24, and both are positive, then it is given that $P=1$, $Q=0$, and the period is 24, and generally D is chosen to be 1. The model can be derived as SARIMA $(0,1,1)\ (1,1,0)_{24}$. The following parametric significance test residual white noise test were performed on the model and the results are shown in Tables 2 and 3. From the P>|z| column in Table 2, the P value of each variable is less than 0.01, so the plus hypothesis is rejected at the significance level of 0.01, and the coefficient of each variable in the model passes the significance test, and it can be considered reasonable to include these variables in the fitted model. From the test results in Table 3, it can be seen that the p-values of the Q-statistics are greater than 0.05 at the delayed order 1–12 of the residual series, and the original hypothesis is not rejected at the significance level of 0.05, which means that the residuals are white noise series, indicating that the

Table 2 Parameter significance test

	Coef	Std error	z	P>|z|	[0.025	0.975]
ma. L1	− 0.7367	0.027	− 26.997	0.000	− 0.790	− 0.683
ar.S.L24	− 0.6461	0.014	− 44.937	0.000	− 0.674	− 0.618
sigma2	4.073e−06	1.41e−07	28.907	0.000	3.8e−06	4.35e−06

Table 3 Residual white noise tests

Lag	AC	Q	Prob(<Q)	Lag	AC	Q	Prob(<Q)
1	− 0.061219	2.193142	0.151131	7	− 0.002974	5.93314	0.648524
2	0.016337	2.423013	0.282142	8	− 0.030623	6.52314	0.70324
3	− 0.036942	3.532142	0.302156	9	0.007131	7.31565	0.73654
4	0.023654	3.992541	0.412301	10	0.029412	8.74152	0.63214
5	0.013516	4.135683	0.536113	11	0.039412	9.10335	0.56414
6	0.007351	5.394511	0.61564	12	− 0.016032	9.98354	0.53101

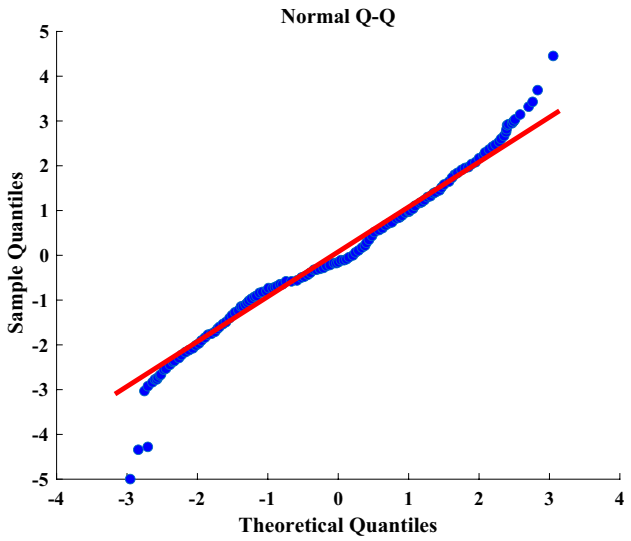

Fig. 10 Q–Q plot of handover success rate

fluctuations of the residual series no longer have any statistical pattern. Therefore, it can be considered that the fitted model has sufficiently extracted the information in the time series. In addition, it is also illustrated from the Q-Q plot in Fig. 10 that the residuals obey a normal distribution.

From the afore mentioned method, the parameters of SARIMA for the four KPIs are obtained in the Table 4.

The LSTM model parameters are as follows (Table 5):

The experiments of the hybrid model were conducted with the uplink PRB utilization rate as an example, and a comparison of the predicted results is shown in Fig. 11.

It is illustrated in Fig. 11 that the combined model combines the advantages of both LSTM and rolling SARIMA. It is able to predict the uplink PRB utilization smoothly as LSTM model, and it can also predict the spikes well which cannot be predicted by LSTM. Table 3 presents the error comparison of the combined model with other models.

In this paper, RMSE and MAPE are used as the evaluation indexes of the prediction model. The loss function of the model is MSE. The formulas are as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{11}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{12}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{13}$$

where, $\hat{y}_i$ is the predicted value, y_i is the true value, and n is the number of samples.

Table 4 Parameters table of SARIMA for KPIs

KPIs	参数
Handover success rate	SARIMA (0,1,1) $(1,1,0)_{24}$
Uplink PRB utilization	SARIMA (1,1,0) $(0,1,1)_{24}$
RRC connection success rate	SARIMA (1,0,0) $(0,1,1)_{24}$
RRC reconstruction success rate	SARIMA (1,1,0) $(0,1,1)_{24}$

Table 5 Parameters of LSTM

Layers	Learning rate	Epochs	Batch
5	0.001	50	64

It is illustrated in Table 6 that the residual weighting hybrid model has the best performance in both RMSE and MAPE, providing a reliable basis for anomaly detection.

3.3 Anomaly detection

Following is the example of handover success rate for anomaly detection, the predicted value of the hybrid model is compared with the true value. From our empirical studies of our data center environment, we found that when the single point average error is higher than 2 times MAPE, we can predict all the outliers, and we set the following scoring criteria.

$$\text{level} = \begin{cases} 0 & \frac{E_i}{MAPE} < 2 \\ 1\ 2 & \leq \frac{E_i}{MAPE} < 3 \\ 2\ 3 & \leq \frac{E_i}{MAPE} < 4 \\ 3 & \frac{E_i}{MAPE} \geq 4 \end{cases} \tag{14}$$

where, E_i is the absolute value of the error at point i and Level is the anomaly level.

The final anomaly of the handover success rate is obtained as shown in Fig. 12.

The data in Fig. 12 is the handover success rate from May 23 to June 9, 2020, in which there are 9 known anomalies, which are points 1, 3, 4, 5, 7, 8, 9, 10, 11. The algorithm in this paper detects 11 anomalies, and considers the common points 2 and 6 as anomalies, which shows that precision is $9/(9+2)=81.82\%$, Recall is $9/(9+0)=100\%$.

We compare the method of this paper with the fixed threshold method and the unsupervised learning method, and the results are shown in Table 7. It can be observed from Table 7 that although the other three have higher precision, Recall is too much lower than the method proposed in this paper. In this paper, the proposed method detects all the anomalies by reducing a small number of Precision. In

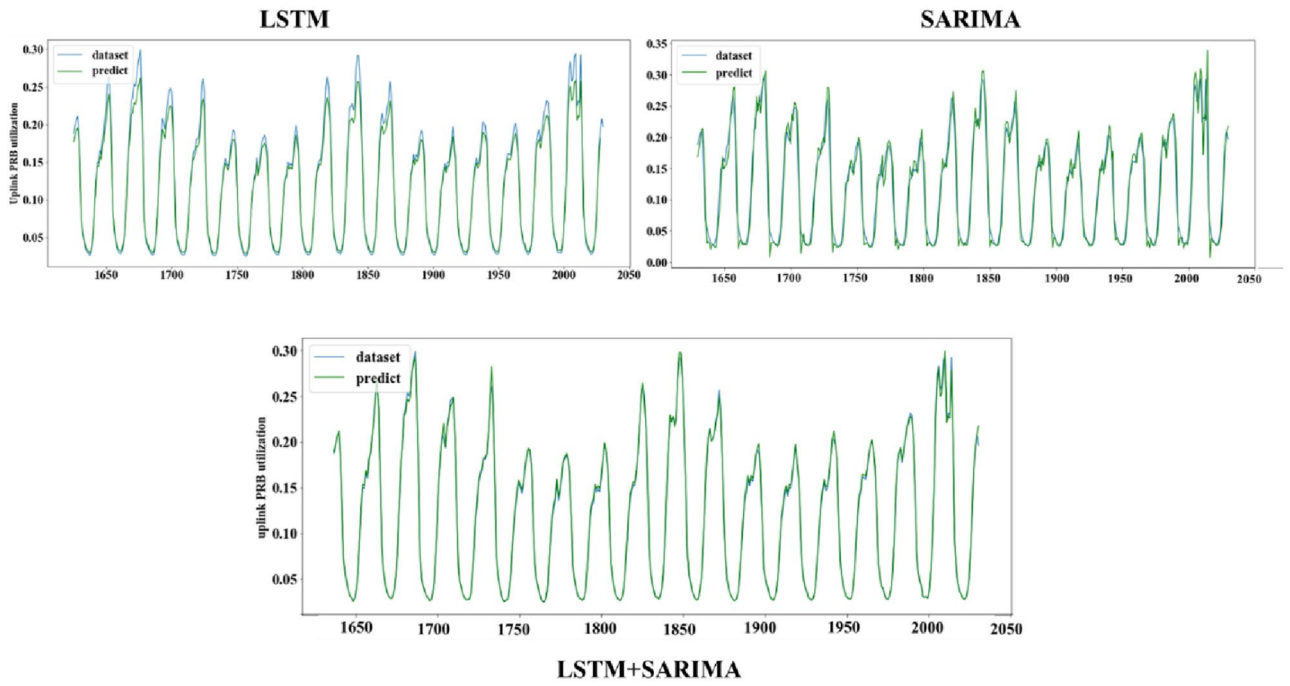

Fig. 11 Comparison of model prediction results

Table 6 Error comparison table of prediction models

	LSTM		SARIMA		Equivalent weighting		Residual weighting	
	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)
Uplink PRB utilization	0.002098	5.458	0.002680	7.181	0.001982	5.093	0.001869	4.787
Handover success rate	0.002345	6.761	0.001992	5.743	0.001004	4.089	0.000824	2.371
RRC connection success rate	0.002893	5.063	0.002848	4.689	0.002633	4.269	0.002035	3.496
RRC reconstruction success rate	0.002698	5.100	0.003057	5.935	0.002356	4.796	0.002035	4.069

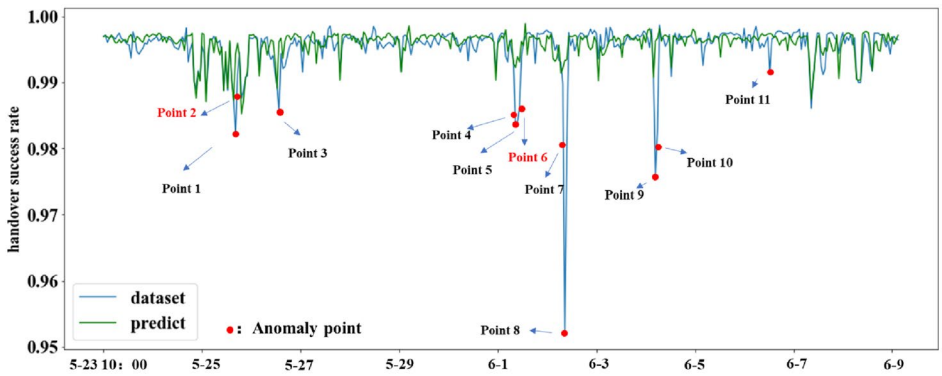

Fig. 12 Anomaly point chart of handover success rate

addition, with the increasing of data, the gap between the Precision of this paper's method and the other three methods will become less.

The abnormality level of each anomaly can be obtained from Eq. (14) as shown in Table 8.

The anomaly detection of the other three KPIs by the above method gives the following Tables 9, 10 and 11.

Following we need to consider the anomalies of each KPIs together to determine the anomaly status of the network. The weights of each KPIs need to be available first.

Table 7 Comparison of anomaly detection methods

	Fixed-threshold method		Unsupervised learning [15] (%)	Proposed method (%)
	Threshold1 (work day) (%)	Threshold2 (week day) (%)		
Precision	83.33	80.00	85.71%	81.82
Recall	55.55	44.44	66.67	100

Table 8 Handover success rate anomaly level table

Anomaly points	Anomaly level	Anomaly points	Anomaly level
Point 1 (5-26 5:00)	3	Point 7 (6-2 4:00)	3
Point 2 (5-26 6:00)	1	Point 8 (6-2 5:00)	3
Point 3 (5-27 3:00)	3	Point 9 (6-4 2:00)	3
Point 4 (6-1 3:00)	3	Point 10 (6-4 3:00)	3
Point 5 (6-1 4:00)	3	Point 11 (6-6 12:00)	2
Point 6 (6-1 5:00)	3		

Table 9 Uplink PRB utilization anomaly level table

Anomaly points	Anomaly level	Anomaly points	Anomaly level
5-25 12:00	3	6-8 11:00	3
5-26 6:00	2	6-8 12:00	3
5-27 13:00	2	6-8 13:00	3
6-1 11:00	1	6-8 14:00	3
6-1 12:00	1		
6-3 13:00	1		

Table 10 RRC connection success rate anomaly level table

Anomaly points	Anomaly level
5-30 12:00	2
6-3 13:00	1
6-8 11:00	1
6-8 12:00	1

We noticed that not only it is necessary to build an accurate model in machine learning, but the interpretability of the model is equally important. For example, determining the importance of input features to the model can help understand the logic of the model and select features more rationally, which is the same as the sense of indicator weights. In the random forest, the output of feature importance achieves this capability.

Table 11 RRC reconstruction success rate anomaly level table

Anomaly points	Anomaly level
5-24 12:00	2
5-24 15:00	2
6-8 11:00	1
6-8 14:00	1
6-9 1:00	1

Random forest is an algorithm that integrates multiple decision trees through the idea of ensemble learning. The principle is that a certain number of samples are randomly selected (Bootstrap) on the basis of the original data to form a sub-sample, and the base classifier (decision tree) is constructed respectively, the results of each base classifier are assembled by voting or finding the mean value. After constructing the random forest, variable importance measurement (VIM) can be applied to each feature by permutation. Traditional VIM methods are divided into two main categories, Gini impurity-based and permutation-based methods [36]. In this paper, the Gini impurity method is used to score the importance of the characteristic variables with the following formula.

$$GIn = \sum_{k=1}^{K} \sum_{k' \neq k} PnkPnk' = 1 - \sum_{k=1}^{K} P^2 nk \tag{15}$$

where GIn denotes the reduction in feature impurity for random forest nodes, K is the number of nodes, and Pnk is the proportion of features.

Construct data using four metrics (handover success rate, uplink PRB utilization, RRC connection success rate, and RRC reconstruction success rate) as input and anomaly status as output (where the abnormal status is labeled as 1 and the normal status label is 0). Construction of a classification model for communication network anomalies by random forest. The model parameters of the random forest are $n\_estimators = 45$, $max\_depths = 8$, $max\_features = 3$.

We obtain a random forest classification accuracy of 97.89%, and keep in mind that the more accurate the model is, the more reasonable the feature importance is. Then the importance of each KPI for the classification is obtained from the Feature_importance parameter as the weight of the KPI. The results are: $w_1 = 0.5151$, $w_2 = 0.1792$, $w_3 = 0.1347$, $w_4 = 0.1710$.w_1、w_2、w_3, and w_4 are the weights of handover success rate, uplink PRB utilization, RRC connection success rate, and RRC reconstruction success rate, respectively. After that, the anomaly score of the network is calculated according to the Eq. (16).

Table 12 Network anomaly conditions

Anomaly points	Anomaly score	Anomaly points	Anomaly score	Anomaly points	Anomaly score
5-24 12:00	0.342	6-1 5:00	1.5453	6-6 12:00	1.0302
5-24 15:00	0.342	6-1 11:00	0.1792	6-8 11:00	0.8433
5-26 5:00	1.5453	6-1 12:00	0.1792	6-8 12:00	0.6723
5-26 6:00	0.8735	6-2 4:00	1.5453	6-8 13:00	0.5376
5-27 3:00	1.5453	6-2 5:00	1.5453	6-8 14:00	0.7086
5-27 13:00	0.3584	6-3 13:00	0.3139	6-9 1:00	0.1347
6-1 3:00	1.5453	6-4 2:00	1.5453		
6-1 4:00	1.5453	6-4 3:00	1.5453		

$$score = \sum_{i=1}^{m} w_i \cdot level_i \tag{16}$$

where, *score* is the anomaly score of a single point, w_i is the weight of the KPI, $level_i$ is the anomaly level of each KPI, and m is the number of KPIs.

The network anomalies from 10:00 on May 23, 2020 to 5:00 on June 9, 2020 were obtained by combining the four indicators as shown in Tables 12 and 13.

The obtained network anomaly score quantifies the network anomalies, which can help operators take different measures according to different situations, rationalize the use of network resources and do a proper resource scheduling.

4 Conclusion and discussion

Real-time anomaly detection for communication networks is conducive to improving communication service quality, resource utilization, and network reliability. However, the traditional communication network reported alerts cannot meet the function of active detection and real-time alarm. The paper applied a prediction-then-detection approach to anomaly detection in communication networks. The accuracy of prediction directly affects the effectiveness of subsequent anomaly detection, so the paper combines the LSTM and SARIMA models with better prediction on the KPI's data by residual weighting. The experimental results illustrate that the hybrid model performs optimally on all four KPIs, providing a reliable basis for subsequent anomaly detection. Next, we use the error between predicted data and real-time data to detect anomalies, and quantify and classify the anomaly status into four levels. Finally, we use random forest to calculate the weights of KPIs to evaluate the whole network performance and give specific treatment measures for each anomaly state, which provides a new idea for communication network anomaly detection and quantification.

In this paper, we try to implement anomaly detection using the method of prediction before detection. Combining LSTM and SARIMA models to detect multiple communication indicators is essentially considered as a supervised learning. Unsupervised learning is also a way of anomaly detection [37], in the absence of sufficient prior knowledge and presence of labeling problems, unsupervised learning is a better solution. In addition, semi-supervised learning and reinforcement learning have been gradually applied to anomaly detection tasks in recent years. Constructing multiple use-case anomaly detectors requires especially analyzing the characteristic of the full data or making strong assumptions, e.g., the training data is anomaly-free which is yet unrealistic in most real datasets [38]. Reinforcement learning follows the incremental self-learning process that the agent autonomously learns a generic framework from interactions with the environment without any assumption and constraint [39]. Hence, it provides a novel way of solving the anomaly detection problem.

Table 13 Anomaly handling measures

Anomaly score	Safety measures and representative colors
0~1	Some risk, but no action required
1~2	Higher risk, need to alert, take some expansion measures
2~3	High risk, need to check network to avoid congestion

Acknowledgements This work was supported by the Key Program of National Natural Science Foundation of China (51934007) and Shandong Key S&T Innovation Engineering Project (2019JZZY020504).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared toinfluence the work reported in this paper.

References

1. Soule A, Salamatian K, Taft N (2005) Combining filtering and statistical methods for anomaly detection. IMC
2. Yanhua Y, Meina S, Wenting Z, Junde S (2011) A dynamic compuation approach to determining the threshold in network anomaly detection. J Beijing Univ Posts Telecommun 34(2):45–49
3. Eswaran S, Honnavalli P, Honnavalli P (2021) A threshold-based, real-time analysis in early detection of endpoint anomalies using SIEM expertise. Netw Secur 4:7–16. https://doi.org/10.1016/S1353-4858(21)00039-8
4. Zhen Y, Liu XD, Li T, Wu D, Wang JJ, Zhao YW, Han H (2022) A systematic literature review of methods and datasets for anomaly-based network intrusion detection. Comput Secur 102675:0167–4048. https://doi.org/10.1016/j.cose.2022.102675
5. Mauro MD, Galatro G, Fortino G, Liotta A (2021) Supervised feature selection techniques in network intrusion detection: a critical review. Eng Appl Artif Intell 101:104216. https://doi.org/10.1016/j.engappai.2021.104216
6. Mukkamala S, Sung AH (2003) Detecting denial of service attacks using support vector machines. FUZZ-IEEE.02
7. Boahen EK, Elvire B, Wang C (2021) Network anomaly detection in a controlled environment based on an enhanced PSOGSARFC. Comput Secur 104(4):102225. https://doi.org/10.1016/j.cose.2021.102225
8. Zhou Y, Mazzuchi TA, Sarkani S (2020) M-AdaBoost-A based ensemble system for network intrusion detection. Expert Syst Appl 162(6):113864. https://doi.org/10.1016/j.eswa.2020.113864
9. Potluri S, Henry NF, Diedrich C (2017) Evaluation of hybrid deep learning techniques for ensuring security in networked control systems. In: 22nd IEEE conference on ETFA
10. Weijie H, Xue J (2019) Detecting anomalous traffic in the controlled network based on cross entropy and support vector machine. IET Inf Secur 13(2):109–116. https://doi.org/10.1049/iet-ifs.2018.5186
11. Ranshous S, Shen S, Koutra D, Harenberg S, Faloutsos C, Samatova NF (2015) Anomaly detection in dynamic networks: a survey. Comput Stat 7(3):223–247. https://doi.org/10.1002/wics.1347
12. Chalapathy R, Chawla S (2019) Deep learning for anomaly detection: a survey. arXiv:1901.03407v1
13. Wang F, Zhu L, Li J, Haibao C, Huaxiang Z (2021) Unsupervised soft-label feature selection. Knowl Based Syst 219(2):106847. https://doi.org/10.1016/j.knosys.2021.106847
14. Hui X, Jiaxing W, Hao L, Dengqing O, Jie S (2021) Unsupervised meta-learning for few-shot learning. Pattern Recognit 116(6):107951. https://doi.org/10.1016/j.patcog.2021.107951
15. Aliakbarisani R, Ghasemi A, Wu SF (2019) A data-driven metric learning-based scheme for unsupervised network anomaly detection. Comput Electr Eng 73(2019):71–83. https://doi.org/10.1016/j.compeleceng.2018.11.003
16. Zenfack V, Donghyun K, Daehee S, Ahyoung L (2021) An unsupervised anomaly detection framework for detecting anomalies in real time through network system's log files analysis. High Confid Comput 1(2):100030. https://doi.org/10.1016/j.hcc.2021.100030
17. Qin ZQ, Ma XK, Wang YJ (2020) ADSAD: An unsupervised attention-based discrete sequence anomaly detection framework for network security analysis. Comput Secur 99:102070. https://doi.org/10.1016/j.cose.2020.102070
18. Meng X, Wang S, Liang Z, Yao D, Zhou JH, Zhang YJ (2021) Semi-supervised anomaly detection in dynamic communication networks. Inf Sci 571:527–542
19. Yang H, Xutao L, Wenhao Q, Yuhan Z, Wei Z, Chang T (2021) A network traffic forecasting method based on SA optimized ARIMA-BP neural network. Comput Netw 193(3):108102. https://doi.org/10.1016/j.comnet.2021.108102
20. Guo D, Xingwen X, Lin Z, Yong Z (2021) Dynamic modification neural network model for short-term traffic prediction. Procedia Comput Sci 187(6):134–139. https://doi.org/10.1016/j.procs.2021.04.043
21. Hyun Woo W, Jun Hui L, Yong Hoon C, Young-UK C, Hyunkjoon L (2011) Dynamic bandwidth provisioning using ARIMA based traffic forecasting for mobile WiMAX. Comput Commun 34(1):99–106. https://doi.org/10.1016/j.comcom.2010.08.008
22. Salman AG, Kanigoro B (2021) Visibility forecasting using autoregressive integrated moving average (ARIMA) models. Procedia Comput Sci 179(2021):252–259. https://doi.org/10.1016/j.procs.2021.01.004
23. Junling L, Zhongliang Z, Yao F, Feng R (2021) Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results Phys 3:104462. https://doi.org/10.1016/j.rinp.2021.104462
24. Jiancai S, Liyi Z, Guixiang X, Yunpeng M, Shan G, Qingling J (2021) Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model. Energy Build 243(3):110998. https://doi.org/10.1016/j.enbuild.2021.110998
25. Agga A, Abbou A, Labbadi M, Yassine H (2021) Short-term self consumption PV plant power production forecasts based on hybrid CNN-LSTM. ConvLSTM Models Renew Energy 177:101–112. https://doi.org/10.1016/j.renene.2021.05.095
26. Junqiang L, Fan L, Chunlu P, Dongbin H, Hongfu Z (2021) Prediction of remaining useful life of multi-stage aero-engine based on clustering and LSTM fusion. Reliab Eng Syst Saf 214:107807. https://doi.org/10.1016/j.ress.2021.107807
27. Prasanjit D, Chaulya SK, Sanjay K (2021) Hybrid CNN-LSTM and IoT-based coal mine hazards monitoring and prediction system. Process Saf Environ Prot 152:249–263. https://doi.org/10.1016/j.psep.2021.06.005
28. Fan GF, Peng LL, Hong WC, Sun F (2016) Electric load forecasting by the SVR model with differential empirical mode decomposition and auto regression. Neurocomputing 173:958–970. https://doi.org/10.1016/j.neucom.2015.08.051
29. Feng YT, Zhang PX, Yang M, Li Q, Zhang AA (2019) Short term load forecasting of offshore oil field microgrids based on DA-SVM. Energy Procedia 158:2448–2455. https://doi.org/10.1016/j.egypro.2019.01.318
30. He FF, Zhou JZ, Feng ZK, Liu GB, Yang YQ (2019) A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with Bayesian optimization algorithm. Appl Energy 237:106–116. https://doi.org/10.1016/j.apenergy.2019.01.055
31. Fan GF, Yu M, Dong SQ, Yeh YH, Hong WC (2021) Forecasting short-term electricity load using hybrid support vector regression with grey catastrophe and random forest modeling. Util Policy 73:101294
32. Qiuying Y, Jie W, Hongli M, Xihao W (2020) Research on COVID-19 based on ARIMA model$^{\Delta}$-Taking Hubei, China as

an example to see the epidemic in Italy. J Infect Public Health 13(10):1415–1418. https://doi.org/10.1016/j.jiph.2020.06.019
33. Farsi M, Hosahalli D, Manjunatha BR (2020) Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data. Alex Eng J 60(1):1299–1316. https://doi.org/10.1016/j.aej.2020.10.052
34. Saiqun L, Qiyan Z, Guangsen G, Dewen S (2020) A combined method for short-term traffic flow prediction based on recurrent neural network. Alex Eng J 60(1):87–94. https://doi.org/10.1016/j.aej.2020.06.008
35. Chen-jui L, Jeng-Jong L, Chiao-Wun J, Ming-Chang T (2020) A rolling forecast approach for next six-hour air quality index track. Ecol Inform 60:101153. https://doi.org/10.1016/j.ecoinf.2020.101153
36. Nicodemus KK, Malley J, Strobl C, Ziegler A (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinform 11(1):110–122. https://doi.org/10.1186/1471-2105-11-110
37. Subutai A, Alexander, Scott P, Zuha A (2017) Unsupervised real-time anomaly detection for streaming data. Neurocomput 262:134-147. https://doi.org/10.1016/j.neucom.2017.04.070
38. Jiang P, Liu Z (2019) Variable weights combined model based on multi-objective optimization for short-term wind speed forecasting. Appl Soft Comput 82:105587
39. Mengran Yu, Shiliang Sun (2020) Policy-based reinforcement learning for time series anomaly detection. Eng Appl Artif Intel 95:103919