2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

A hybrid ARIMA–ANN model for Internet Food safety risks

Wei Wang
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail:321810255@qq.com

Jun Sun
School Of Artificial Intelligence and Computer Science,Jiangnan University
WuXi ,China
E-mail:junsun@jiangnan.edu.cn

***Abstract*—As Internet technology becomes more and more mature, Internet food sales have become more and more problematic. problems from raw materials to food processing to food distribution are not easily regulated because of online sales. Firstly , crawl the take-out sales data of cities across the country, based on the data the time series prediction model is established to realize the prediction of business and regional risk values, From this obtains the merchant early warning as well as reminds the consumer's function. According to the data analysis, the ARIMA+ANN model is used for short-term prediction. Then the data analysis results of different regions are compared, and the data analysis of both sides of time and space is conducted, The value at risk ranking allows consumers to make a clear comparison of effects, In this way, we can realize the real and effective supervision of store owners from the economic level, realize the real and effective supervision of Internet food sales, and ensure food safety.**

***Keywords-component;* Internet Food safety ; Data acquisition and cleaning; spatiotemporal series; Arima; ANN**

I. INTRODUCTION

With Food safety incidents have occurred frequently In recent years, causing all walks of life to pay attention to food safety issues. With the gradual penetration of Internet thinking into traditional industries, the Internet will play an increasing role in the field of food safety [1], 2019, The number of Internet food delivery users has reached 421 million in china, The increase of 15.16 million compared to 2018, accounting for 49.3% of the total number of Internet users. The number of takeaway users has reached 417 million, Compared with the increase of 20.37 million in 2018, accounting for 49.3% of mobile Internet users, various drug takeout incidents emerge in endlessly, so Internet food safety is an urgent problem to be solved. Through the collection, screening and analysis of consumers' evaluation of various businesses, this project can have a specific and digital understanding of the food safety and health problems of businesses in this period, and then make a prediction of food safety of businesses through these data, By publicizing the results to the public[2], the public can have an intuitive understanding and comparison of food after obtaining information through various media.

From a theoretical point of view: according to the project planning, first of all, according to the "The PRC Food Safety Law" and "Food Safety Supervision And Management Measures of Online Catering Service", the indicators of food safety violations are determined [3], and the data of Internet food are classified reasonably, so as to have a more accurate cognition of each store.

In fact, the results of the experiment have guiding significance for both businesses and consumers. Businesses can grasp the safety status of their own stores according to the relevant scores and the recent scoring trend, and timely pay attention to adjust the food safety problems in the stores [4]. For consumers, the level of food safety score can be used as an important indicator of choosing consumption which is very important for food safety The prediction of food safety can also give consumers a relatively clear understanding of the recent food safety trend of a store [5], and the regional analysis also has a certain reference value for government supervision.

Big data technology has been applied to various fields of food safety in China, and scholars have been committed to establishing a huge food safety database and a visual model of food safety analysis [6].

Combined with computer technology, there are many applications in the direction of food safety, such as detection technology in the field of food safety, including infrared spectrum, hyperspectral image, computer vision, artificial olfaction, biochip and other detection technologies [7]. In the aspect of food safety information traceability, there is also a food information traceability system based on radio frequency identification (RFID), which can send and receive chip information, so as to obtain relevant food information and establish a traceability system [8], so as to detect the purpose of food safety on the other hand

There are some loopholes in China's current food safety legal regulation in the legal system, law enforcement and other aspects. This experimental model can play an early warning role to a certain extent [9].

The rest of this paper is arranged as follows: the first chapter is the data source and pretreatment, the second chapter is the construction of Internet food risk prediction model, the third chapter is the experimental results and analysis, the fourth chapter is the comparison of arima-ann, ARIMA, LSTM model effect, and the last chapter is the conclusion and reference literature.

II. DATA SOURCE AND PRETREATMENT

This chapter mainly introduces the data source and data preprocessing of this article. First, the analytic hierarchy process is used to obtain the weights, and then the Bert model analysis is used to obtain the food safety risk value, and then the data preprocessing includes time series stationarity detection and unit root detection.

A. Date Source

Use the Scrapy framework to obtain takeaway information in 46 major cities across the country, totaling approximately 230,000 takeaway stores and 38 million

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00048

review information, and filter out stores with too small data. The final data used in the experiment is about 200,000 stores and 3,600. Million comments.

B. Data Preprocessing

First, the time series stationarity test is performed to test the stationarity of the sequence, and the difference is when the sequence is not stable. Since the data is stable mainly by subjective experience, it is necessary to perform unit root test to confirm whether the data is stable after the difference, and there is an objective judgment.

1)Using the analytic hierarchy process [10-13] to obtain the weights of 16 categories, the steps are as follows: invite a number of people familiar with the situation as scorers, ask them to fill in the designed questionnaire, then the judgment matrix table is obtained, and the characteristic vector of the judgment matrix is calculated . The steps are: Calculate the product of each row of the judgment matrix:

$$M_i = \prod_{j=1}^{n} b_{ij} \quad (2)$$

Calculate Mi=n times square root Vi:

$$V_i = \sqrt[n]{M_i} \quad (3)$$

Normalize Wi: $W_i = V_i \sum V_i$

$$W_i = v_i \sum v_i \quad (4)$$

It is obtained that Wi=(W1, W2, W3,..., Wn) is the eigenvector of the judgment matrix, that is, the corresponding weights are shown in Table 1.

Table 1 Indicator weight map

index	***ranking***
C1Use perverted raw materials	0.309975
C4Merchants operate in an unsothy environment	0.1382
C3There is a mix of foreign objects in the food	0.109617
C6Provide unqualified cutlery	0.084708
C5Use nonconforming packaging materials	0.061817
C2Food processing procedures are improper	0.052617
C9The sanitary condition of the distribution container	0.051267
C11Sell out-of-date food	0.046342
C8Special food quality	0.035192
C7The packaging is not standardized	0.032475
C10The product label is not qualified	0.016517
C15Merchants provide false information	0.015858
C16Merchants swipe orders and fake comments	0.01265
C13Merchants do not process or delay handling complaints to report incidents	0.012458
C12Inso- and offline inse and out of line	0.01025
C14The merchant does not provide proof of sale	0.010067

The data is labeled and substituted into the Bert[14-16] training model. The Bert model is an NLP model developed by the Google AI team in 2018. The main process is as follows: each word (token) in the comment is sent to the embedding layer, plus Self-attention mechanism, and then training in the network, and finally get the probability of 16 classifications (this experiment uses the Bert model as part of the data preprocessing, so it is briefly described), and then multiply by the index weight obtained by the analytic hierarchy process, then A score can be obtained for each review, all review information is summarized, and the weighted average is finally used to obtain the risk value of the merchant in each time period through the model.

2) Time Series Stationarity Detection

First judge the stability of the data, and judge that the group of data has no obvious upward and downward trend, and there is a continuing trend, that is, whether there is a finite fluctuation around a certain value. If the data is not stable, the data needs to be differentiated. Until the data is stable

3) Unit Root Detection

Then objectively use the unit root test to test whether the time series is stable after the score, and whether there is a unit root in the risk value to test whether the data is stable. If it does not exist, it is stable. Otherwise, it is not stable. The formula is as follows:

$$\Delta yt = (\rho - 1) yt - 1 + ut$$
$$= \delta yt - 1 + ut \quad (5)$$

The whole formula is a regression model, Δ is the first difference, t is the time, ρ is the coefficient, and ut is the error term. Testing whether there is a unit root is equivalent to testing whether = 0

4) After the processed data is stable, the data needs to be tested for white noise. If the data is white noise, it is a set of purely random numbers. The white noise monitoring in this experiment is divided into two parts: raw data white noise monitoring , Residual data white noise monitoring, using the Ljung-Box method to monitor white noise, the specific process is described in the experiment.

III. Internet Food Risk Prediction Model Construction

The risk prediction model is ARIMA+ANN[17-20] combined prediction. The main process is ARIMA[21-24]. First make predictions, and then put the data residual values into the ANN model. The two model results are combined to obtain the final prediction value.

A. ARIMA Model

The ARIMA model is also known as the differential integrated moving average regression model. First of all, the premise of this model is that the data must be a stationary sequence and have a continuation trend. First, the sequence to be processed is subjected to d-order difference until the data is stable. The sequence model is ARIMA(p,d,q), p is the order of the autoregressive model, and q is the moving average order. P is solved by plotting the autocorrelation coefficient ACF and partial autocorrelation coefficient PACF of the sequence, which are specifically determined by analyzing the graph.The structure of the ARIMA time series model is shown in Figure 1:

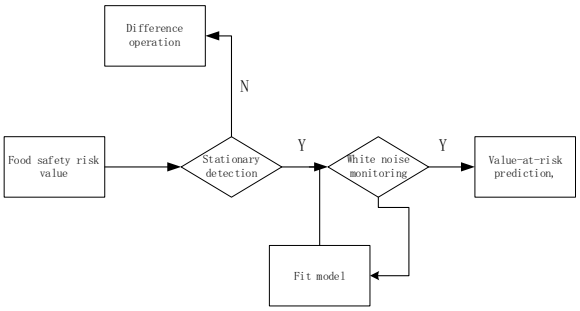

Figure 1 ARIMA time series model structure

B. *ARIMA-ANN model*

Biological neural networks are connected by a large number of biological neurons. Similarly, ANN is also composed of multiple neuron models connected according to certain rules.

Use Keras to build a neural network, mainly including input layer, hidden layer, and output layer. Through the training of sample data, the network weights and thresholds are constantly modified to make the error function fall along the negative gradient direction, and the hidden layer can be one or more layers.

In this article, the ANN neural network is used to predict the residual data generated by the ARIMA model. Then the residual prediction results and the ARIMA prediction results are fused, the hidden layer uses the tanh transfer function, and the output layer uses a linear function.

The residual prediction steps of the ANN network in this article are as follows:

(1) Determine the number of nodes according to the output results of the ARIMA model, and determine the input data and output data

(2) Determination of the number of hidden layer nodes: In fact, enough hidden layer nodes can fit any non-linear function, but the corresponding calculation amount will increase correspondingly, and the phenomenon of overfitting will affect the model Effect, on the other hand, too small number of nodes will also lead to poor model performance. In fact, the number of hidden nodes is related to the complexity of the actual problem, the number of inputs and outputs, and the expected error. The calculation formula for the selected hidden layer used in this article is:

$$L=\sqrt{\mathrm{m+n}}+a \qquad (5)$$

Among them, n is the number of nodes in the input layer, m is the number of nodes in the output layer, and a is a constant between [1,10] to determine the range of the number of nodes in the hidden layer.

The basic structure of the ARIMA-ANN fusion model is shown in Figure 2:

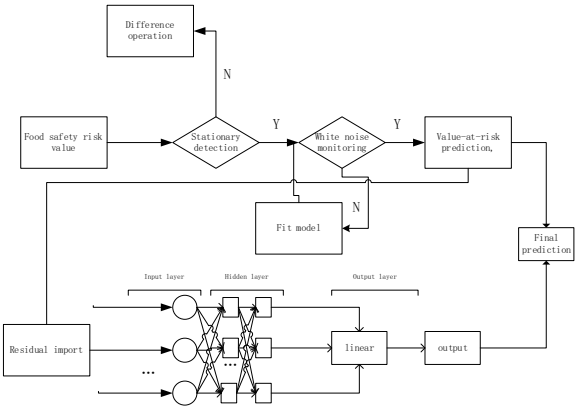

Figure 2 Mixed model diagram

The main process is as follows:

(1) After obtaining the merchant's risk value, the sequence is brought into the ARIMA model. First, the data stationarity is detected. If it is not stable, the difference is performed until it is stable, and then it is placed in the trained ARIMA model to predict the preliminary result y1.

(2) Make the difference between the ARIMA model prediction result and the true value to obtain the residual value, and import the residual value into the ANN,

Get the residual prediction value y2.

(3) Combine the ARIMA model result y1 with the residual prediction result y2 obtained by ANN to obtain y:

$$Y_{\mathrm{f}}=Y_1+Y_2 \qquad (6)$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

First, take out a store for model testing. Take the time period from July 2016 to July 2019 as an example, the interval period is one week, and there are a total of 145 data. As shown in Figure 3

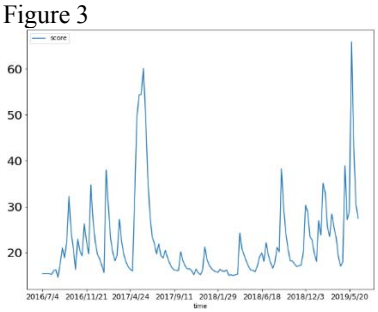

Figure 3 Data timing

The first 66% of the data is used as the training set, and the remaining 34% is used as the test set, that is, the first 95 pieces of data are used as the training set, and the last 50 pieces of data are used as the test set. Carry on the stationarity detection of the data, the difference detection effect is shown as in Fig. 4.

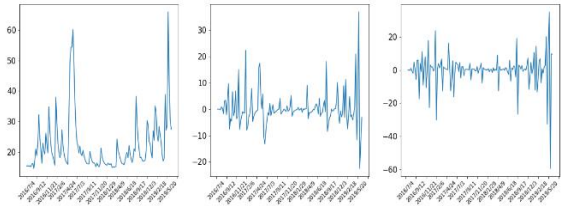

Figure 4 Differential effect

A. *Stationarity detection of data series*

Figure 4 shows the original data, the first-order difference, and the second-order difference. The time series

data is basically stable after the first-order difference. The ADF (unit root) detection of the first-order difference results is shown in Table 2.

The result shows that the data fluctuates relatively smoothly after the difference. The unit root test value adf is less than one percent, five percent, and 13 percent. The P-Value value is 4.6341*10-5, which is close to 0. , Data stationary series no longer need difference operation.

Table 2 ADF detection

adf	cValue			p 值
	1%	5%	10%	
-4.836	-3.477	-2.882	-2.577	0.00004

B. *White noise detection of original data*

The Ljung-Box method is used to test, and the Ljung-Box test is the LB test, which is a method to test the autocorrelation of the series in the time series analysis. The Q statistic of the LB test is:, where T is the sample size, m is an artificially selected number, and ρi is the autocorrelation coefficient of the i-order lag. The experimental results are shown in Table 3. The p value obtained from the original data is much less than 0.05, so the data is a stationary non-white noise sequence.

Table 3 Raw data white noise monitoring

Stat	P
80.45290731	2.97716222*10-19

C. *ARIMA Model*

First determine the three parameters p, d, and q required by the ARIMA model. The first difference has been determined above, so the value of d is 1.

Plot ACF (autocorrelation coefficient graph) and PACF (partial autocorrelation coefficient).

The autocorrelation coefficient compares an ordered sequence of random variables with itself. It reflects the correlation between the same sequence in different time series. In fact, the autocorrelation coefficient is not only the relationship between two elements, but also its influencing factors. There are all the elements between these two elements, and the partial autocorrelation coefficient removes the intermediate influencing factors, which is a strict correlation between the two elements. As shown in Figure 5

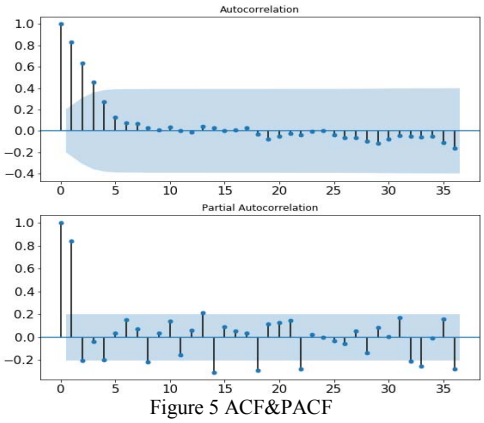

Figure 5 ACF&PACF

From the ACF and PACF diagrams, we can get the ACF diagram is censored, and the PACF diagram is trailing, so p and q are (1, 0), that is, we can try to fit the data sequence with the ARIMA (1, 1, 0) model.

However, considering the large number of stores and human identification will be accompanied by great subjectivity and historical experience, the optimal model identification method is used to calculate AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and select the smallest AIC, BIC values, and finally the ARIMA (1, 1, 0) model is used to fit the data sequence, and the summary information of the final model is shown in Table 4.

Table 4 ARMA Model Results

Dep. Variable:	y	No. Observations:	294
Model:	ARMA(1, 0)	Log Likelihood:	-988.256
Method:	css-mle	S.D. of innovations:	6.970
AIC:	1982.513	BIC:	1993.564
Sample:	0	HQIC:	1986.938

The model prediction results are shown in Figure 6:

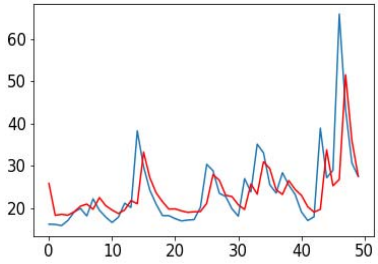

Figure 6 ARIMA_PREDICT

D. *ARIMA-ANN model prediction*

The ARIMA (1,0,1) model is used above to model the takeaway risk value sequence of a store in Wuxi from July 2016 to July 2019, fitting the linear information of the time series, the ARIMA prediction result is y1, and the ANN model prediction The result is y2, the final predicted result. The effect diagram of model prediction results is shown in Figure 7.

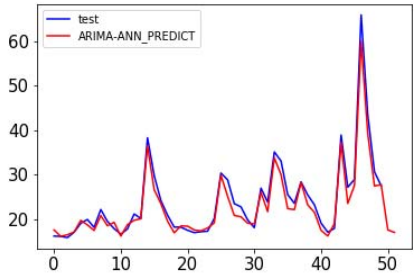

Figure 7 ARIMA-ANN model renderings

E. *Comparison of the effects of ARIMA-ANN, ARIMA, and LSTM models*

LSTM is a special form of RNN, mainly to avoid the problem of long-term dependence. It is solved by introducing a number of gates. The LSTM architecture has an input layer, a hidden layer and an output layer. The hidden layer is composed of memory cells. Each cell has three gates, namely: input gate, forget gate and output gate. Each gate can access the current input and previous output [25-27].

The comparison of the effects of ARIMA-ANN, ARIMA, and LSTM models is shown in Figure 8. In the experimental results shown in the figure, the blue line is the actual experimental value, the red is the predicted value of the ARIMA-ANN model, the green is the predicted value of the ARIMA model, and the yellow is the LSTM The prediction value of the model can be concluded that the

ARIMA model can predict the trend of food safety risks, but the prediction error is large. The prediction effect of the LSTM model is similar to the result of the ARIMA model, and there is also a prediction lag effect. In fact, time series analysis is being performed. It's normal to have a lag at the time. Generally, it is caused by insufficient information. In this article, it may be because the risk value time series are more complicated, and there are more data in the nonlinear time series. After the ARIMA-ANN model further processes the residuals generated by the ARIMA model, the effect achieved Better, the delay phenomenon is better solved in this data sequence, and the prediction error value is also small.

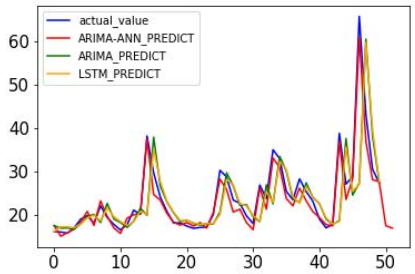

Figure 8 Comparison of prediction models

V. CONCLUSION

This paper uses the ARIMA+ANN model to realize the food safety risk value modeling, which greatly improves the lag of the ARIMA model and the LSTM model in this experiment. Through model testing, the ARIMA+ANN model has good predictive capabilities. The model can make short-term predictions of the merchant's risk value. It has certain guiding significance for government supervision, store self-recognition, and customer consumption. Spatio-temporal analysis The result forecast is also conducive to deepening the government's awareness of local food safety.

VI. REFERENCES

[1] Xu Wei,Zhang Zhipeng,Wang Hongxun,Yi Yang,Zhang Yanpeng. Optimization of monitoring network system for Eco safety on Internet of Things platform and environmental food supply chain[J]. Computer Communications,2020,151(C).

[2] Meng Chun,Sun Lin,Guo Xiaoni,Wu Miao,Wang Yuqi,Yang Lingping,Peng Bin. Development and Validation of a Questionnaire on Consumer Psychological Capital in Food Safety Social Co-governance [J]. Frontiers in Psychology,2021.

[3] Hernández San Juan Isabel. The Blockchain Technology and the Regulation of Traceability: The Digitization of Food Quality and Safety[J]. European Food and Feed Law Review,2020,15(6).

[4] Tsai Hsinyeh,Lee Yu Ping,Ruangkanjanases Athapol. Understanding the Effects of Antecedents on Continuance Intention to Gather Food Safety Information on Websites[J]. Frontiers in Psychology,2020.

[5] Francesco Meneguzzo,Federica Zabini. Agri-food and Forestry Sectors for Sustainable Development[M]:2021-02-26.

[6] Kharroubi Samer,Nasser Nivin A,ElHarakeh Marwa Diab,Sulaiman Abdallah Alhaj,Kassem Issmat I. First Nation-Wide Analysis of Food Safety and Acceptability Data in Lebanon.[J]. Foods (Basel, Switzerland),2020,9(11).

[7] Mojun Zou. Analysis of Food Safety Inspection Technology and Method[J]. International Journal of Education and Economics,2020,3(4).

[8] Drangert Jan-Olof. Correction to: Urban water and food security in this century and beyond: Resource-smart cities and residents[J]. Ambio,2021,50(3).

[9] By Thomas W. Hertel,Uris L.C. Baldos,Keith O. Fuglie. Trade in Technology: A Potential Solution to the Food Security Challenges of the 21st Century[J]. European Economic Review,2020.

[10] Annas Firdaus,Ediana Dina,Kurniawan Asep,Wandira Raju,Zakir Supratman. Decision Support System in Detrmination of Project Tender Winner Using the Analytical Hierarchy Process (AHP) Method[J]. Journal of Physics: Conference Series,2021,1779(1).

[11] Qing Wang,Song Liu,Congcong Li,Chao Yu,Yanxi Liu. AHP Based Evaluation Model of Energy Efficiency for Energy Internet[J]. E3S Web of Conferences,2021,233.

[12] Pham Ngoc Thach,Do Anh Duc,Nguyen Quang Vinh,Ta Van Loi,Dao Thi Thanh Binh,Ha Dieu Linh,Hoang Xuan Truong. Research on Knowledge Management Models at Universities Using Fuzzy Analytic Hierarchy Process (FAHP)[J]. Sustainability,2021,13(2).

[13] Olabanji Olayinka Mohammed,Mpofu Khumbulani. Appraisal of conceptual designs: coalescing Fuzzy Analytic Hierarchy Process (F-AHP) and Fuzzy Grey Relational Analysis (F-GRA)[J]. Results in Engineering,2020(prepublish).

[14] Wang Junshu,Zhang Guoming,Wang Wei,Zhang Ka,Sheng Yehua. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT[J]. Journal of Cloud Computing,2021,10(1).

[15] [Engineering; Researchers at Chinese Academy of Sciences Target Engineering (Enhancing BERT Representation With Context-Aware Embedding for Aspect-Based Sentiment Analysis)[J]. Journal of Technology & Science,2020.

[16] Su Jing,Dai Qingyun,Guerin Frank,Zhou Mian. BERT-hLSTMs: BERT and hierarchical LSTMs for visual storytelling[J]. Computer Speech & Language,2021,67.

[17] Pavan Kumar Singh,Nitin Singh,Richa Negi. Wind Power Forecasting Using Hybrid ARIMA-ANN Technique[M].Springer Singapore:2019-03-31.

[18] Ümit Çavuş Büyükşahin,Şeyda Ertekin. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition[J]. Neurocomputing,2019,361.

[19] C Narendra Babu,Pallaviram Sure. Partitioning and interpolation based hybrid ARIMA–ANN model for time series forecasting[J]. Sādhanā,2016,41(7).

[20] Ümit Çavuş Büyükşahin,Şeyda Ertekin. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition[J]. Neurocomputing,2019,361.

[21]]Singh Ram Kumar,Rani Meenu,Bhagavathula Akshaya Srikanth,Sah Ranjit,Rodriguez-Morales Alfonso J,Kalita Himangshu,Nanda Chintan,Sharma Shashi,Sharma Yagya Datt,Rabaan Ali A,Rahmani Jamal,Kumar Pavan. Prediction of the COVID-19 Pandemic for the Top 15 Affected Countries: Advanced Autoregressive Integrated Moving Average (ARIMA) Model.[J].JMIR public health and surveillance,2020,6(2).

[22] .Periodic autoregressive models for time series with integrated seasonality[J]. Journal of Statistical Computation and Simulation,2021,91(4).

[23] Makala D,Li Z. Prediction of gold price with ARIMA and SVM[J]. Journal of Physics: Conference Series,2021,1767(1).

[24] COVID-19/SARS-CoV-2 News from Preprints; Brief Analysis of the ARIMA model on the COVID-19 in Italy (Published April 11, 2020)[J]. Medical Letter on the CDC & FDA,2020.

[25] Wei Li,Amin Kiaghadi,Clint Dawson High temporal resolution rainfall–runoff modeling[J],Neural Computing and Applications, 2020(prepublish), pp.1-18

[26] Mohamed Hosny,Minwei Zhu,Wenpeng Gao,Yili Fu. A novel deep LSTM network for artifacts detection in microelectrode recordings[J]. Biocybernetics and Biomedical Engineering,2020.

[27] Zhou Yutao,Wu Huayi,Cheng Hongquan,Qi Kunlun,Hu Kai,Kang Chaogui,Zheng Jie. Social graph convolutional LSTM for pedestrian trajectory prediction[J]. IET Intelligent Transport Systems,2021,15(3).