Journal of Modern Crop Science (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jmcs.2022.010105
ISSN 2616-2075 Vol. 1 Num. 1

Study on Prediction Method of Grain Pollutants Based on LSTM and Stochastic Forest

Li Wang[1,*], Lang Zheng[1], Xuebo Jin[1], Xiaoyi Wang[1,2], Jiabin Yu[1], Yuting Bai[1]

[1]*School of Artificial Intelligence, Beijing Technology and Business University, Beijing, China*
[2]*Beijing Institute of Fashion Technology, Beijing, China*

Keywords: Random forest, long short-term memory, fungal contamination, foodstuff

Abstract: Crops such as wheat and peanuts are the main food crops in the world. Due to the complex process of pollutant changes and discrete data in the process of food supply chain processing, accurate prediction is very important for food quality. Most of the existing methods are applicable to continuous systems, and the prediction accuracy of discrete systems such as grain pollutants is not high. To solve this problem, this paper proposes a modeling method based on Long Short-Term Memory (LSTM) network and stochastic forest algorithm. The research contents of higher precision prediction for the discrete system of grain pollutants mainly include: 1) using Random Forest algorithm to predict the pollutants in the grain supply chain; 2) The LSTM Network algorithm is used for the same prediction, and the prediction results of the two methods are compared. Taking the peanut oil supply chain data as an example, the results show that the Random Forest algorithm is better than the LSTM Network in comprehensive prediction, and the prediction accuracy of the test set reaches 99.7%. It can realize the accurate prediction of the pollutants in the grain supply chain.

1. Introduction

Rice, millet, peanuts and other agricultural products are important food crops in various countries in the world. For example, peanuts are one of the most important food crops in Asian countries and regions. According to relevant survey data [1], China's annual output of peanuts has reached 20 million tons, playing an important role in global staple food consumption. Food pollution is an important issue of general concern. The pollution of grain mainly includes heavy metal pollution, radioactive pollution and microbial pollution, of which the microbial pollution accounts for at least 40% [2], the main microorganisms that affect the change of grain quality in the process of grain processing and production are fungi. In the supply chain of grain production, processing, transportation, storage, marketing and other links, a large number of mycotoxins are easily produced, which seriously threaten the quality and safety of grain. Therefore, looking for appropriate prediction methods, discovering the growth trend of fungal pollutants in advance, and realizing timely and efficient prediction of fungal pollutants content will help prevent and control the harm of mycotoxin pollution in grain. However, there are still many deficiencies in China's accumulated experience and methods in the prediction of food pollutants, accelerating the research and development of mycotoxin detection and prediction technologies, Therefore, it is of great

significance to strengthen the research on mycotoxin prediction in the food supply chain.

Because of the great significance of forecasting grain fungal pollution, a large number of scholars have studied the relevant forecasting methods. The research of forecasting methods has gone through the development process from the initial microbial growth dynamics model, to the mathematical statistical model, and now to the in-depth learning algorithm to achieve effective prediction of grain pollutants. At present, there are two main prediction methods for grain pollutants, namely, mechanism driven prediction method and data driven prediction method. The physical concept of mechanism driven method modeling is clear and accurate, which can reflect the relationship between input and output parameters of the system. It is called "white box model". For example, Wang Hongxing [3] discovered the growth law of rice mold and established a dynamic growth model. Lan Xueping, Chen Jinying [4] and others tested the factors that affect Vomitoxin (DON), and obtained the DON risk early warning model. Rossi [5] and others studied the factors such as temperature and humidity, and found that these factors had a certain impact on DON. Data driven model refers to learning the model by using data after getting a set of data, obtaining system features from the data, and describing the change trend of the system. For example, Jia Xiaoyong et al. [6] used the least square method to obtain the predicted concentration of grain pollutants by solving the equation set. Deng Yurui [7] and others used the traditional Bayesian model to predict the mildew, established a simple conventional Bayesian algorithm model, and the accuracy of the prediction of rice mildew probability reached 94.8%. Support vector machine [8], regression model [9], etc. also belong to traditional machine learning. The deep learning and fitting ability is strong, and it can fit the relationship between deep data, but its calculation amount is large, and the model has poor interpretability, such as GRU [10], recurrent neural network [11], etc.

Most of the existing grain pollutant prediction modeling methods are only applicable to continuous systems, while the grain supply chain is composed of a variety of different technological links, so it is a batch process and belongs to a discrete system, so the method applicable to discrete systems should be adopted for analysis. Through many theoretical and empirical analyses, it is shown that for discrete systems, the method of stochastic forest [12] has high prediction accuracy and good tolerance, and will not involve many complex problems like nonlinear models. A large number of decision trees contained in a Random Forest are the key to solving discrete system modeling. At the beginning of feature selection, the classification can be made in advance to judge whether it can continue to be used. The number of decision trees makes the algorithm run more efficiently. In addition, the method of LSTM network is also an effective method to predict food pollutants. Although this method is mainly applicable to continuous system modeling, it can solve the problem of modeling long time series and avoid the gradient explosion problem of traditional neural network algorithm, so it can be compared with the Random Forest method.

2. Prediction Method Based on Random Forest and Long Short-Term Memory Network

The grain supply chain is composed of many different technological links, such as processing and production processes. It is a batch process. Its working steps are usually carried out at different times at the same location. Its operating state is unstable and its parameters are discontinuous. Therefore, it belongs to a discrete system. It should be analyzed using methods suitable for discrete systems. Random forest have the advantages of simplicity and high prediction accuracy, strong adaptability to data sets, low requirements for data standardization, and strong anti-interference capability of models. The accuracy of models is often higher than that of single decision trees. Even discrete or large data systems can make accurate predictions, so they are suitable for modeling and prediction of discrete systems. LSTM was used for comparison.

2.1. Theory of Random Forest

Random Forest belongs to one of machine learning algorithms [13]. Compared with traditional machine learning algorithms, Random Forest runs faster, has higher accuracy, and has strong learning ability. It can be used to make high-dimensional data feature selection. Random Forest have been widely used in medical insurance [14], marketing [15] simulation modeling, communication, biology, management, economics and other aspects.

In the Random Forest algorithm, it is necessary to input different samples in each tree. Only a small number of excellent decision trees can make good predictions. By evaluating the classification results of multiple weak classifiers, a powerful classifier is formed. This is the idea of Bootstrap aggregation, Bagging[16]. Stochastic forest can be used to solve both classification and regression problems. It can effectively handle missing values and thousands of input data, with good dimension reduction effect. The schematic diagram of Random Forest is shown in Figure 1.

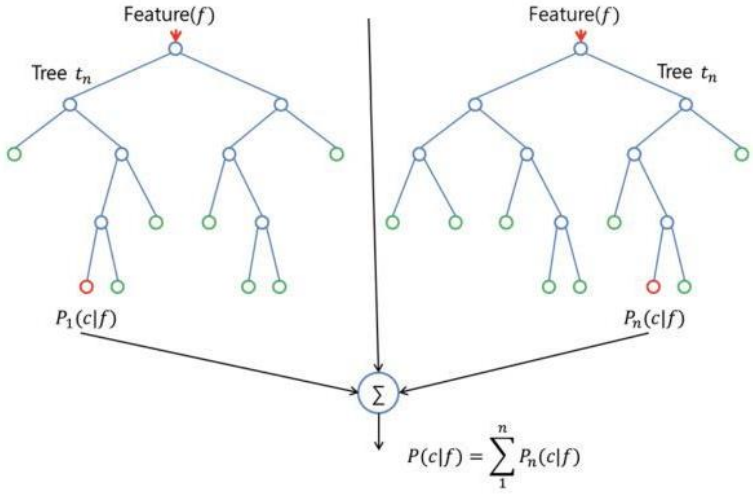


Figure 1: Schematic Diagram of Random Forest

In machine learning, a single weak classifier has poor adaptability and is easily affected by data sets, resulting in over fitting. Therefore, the Random Forest algorithm combines these weak learners with strategies to form a strong learner, thus improving the accuracy of model prediction [17]. The machine learning performance is expressed by k, and the formula is as follows:

$$k = log_2 N \text{ or } k = log_2 N + 1 \quad (1)$$

The training method of Random Forest model is sampling with return, predicting the average value of all sample training sets, and obtaining the final voting result through voting mechanism [18]. The random sampling of Random Forest is divided into data random selection and feature random selection. First, the sampling method with placement is adopted from the original data set. Secondly, the generated sub data set is used to construct a sub decision tree [19], and these data are placed in each sub decision tree to obtain an output value. When constructing a decision tree, we must select all features according to a specific order, train and predict multiple decision trees, and then convert them into classifiers. There are a lot of decision trees in the Random Forest, and each decision tree is different, which enriches the diversity of the system and improves the classification efficiency.

2.2. Theory of Long Short-Term Memory

It is difficult for traditional statistical prediction methods to use mathematical expressions to describe the relationship between the generation of food pollutants and the factors affecting food

pollutants. Existing machine learning and deep learning research only focus on their own characteristics and algorithms for optimization and calculation, such as feedforward neural network and recurrent neural network models. However, for long sequence problems, gradient explosion and gradient disappearance will occur, at the same time, some important data will be omitted. Long Short-Term Memory (LSTM) [20] is mainly used to solve the problem of gradient disappearance and gradient explosion in long sequences during training [21]. The chain structure composed of forgetting gates, input gates, output gates, memory units and activation functions, compared with the Recurrent Neural Network (RNN) neural network, which contains a cell processor, uses gated loop units to process information, allowing information to pass selectively, and discarding information is forgotten through the forgetting gates, avoiding the problem of large prediction errors of long sequence information, It is the main reason that the LSTM network can process the long time series [22] information, so the prediction accuracy is high. LSTM is an improved time cycle artificial neural network, which is gradually used to build a prediction model [23]. LSTM schematic diagram is shown in Figure 2:

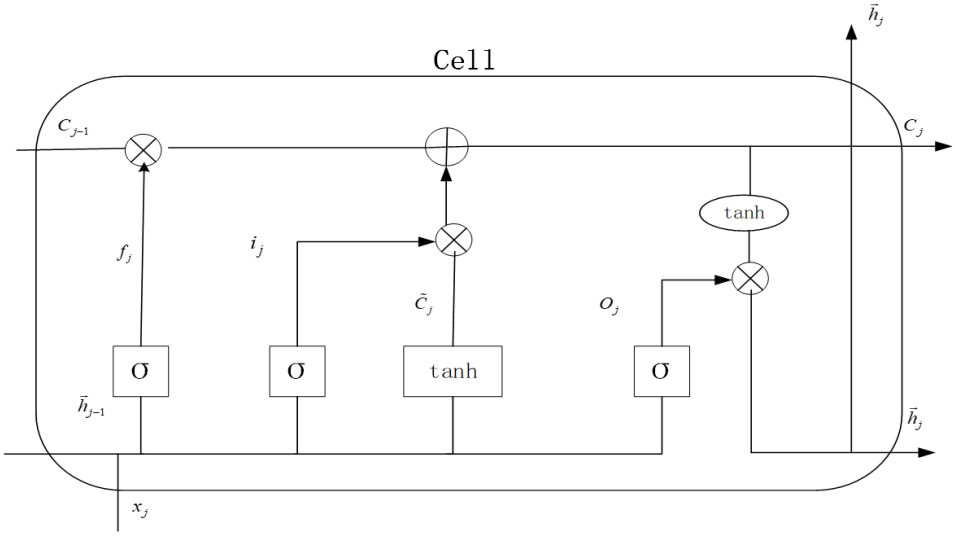


Figure 2: Schematic Diagram of LSTM

Wherein, C_{t-1} represents the input of memory unit at time $t-1$,ht-1 represents the hidden state at time $t-1$,X_t represents the input vector at time t, i_j represents the input gate, ft represents the forgetting gate, and O_t represents the output gate, σ and tanh denote sigmoid activation function and hyperbolic tangent activation function.

The three gates that make up the LSTM structure, namely, the forgetting gate, the input gate and the output gate, are very similar in structure, and can be obtained by processing the parameter matrix, sigmoid function and offset quantity according to the state of the last time and the data of the current time, as follows:

Parameter matrix:

$$I_t = \sigma(X_t W_{xi} + H_t W_{hi} + b_i) \tag{2}$$

Sigmoid function:

$$F_t = \sigma(X_t W_{xf} + H_t W_{hf} + b_f) \tag{3}$$

Offset:

$$O_t = \sigma(X_t W_{xo} + H_t W_{ho} + b_o) \tag{4}$$

Wherein, X_t is the input vector, the input gate isI_t, the forgetting gate is F_t, O_t is the output gate, σ is the sigmoid activation function, H_t is hidden, W is the weight vector, and b is the offset term.

The model parameters of LSTM include time step and loss function. The step size is the size of the time span. Common loss functions include mean square error loss function, cross entropy loss function and log likelihood loss function. This paper uses the mean square error loss function to calculate the average value after the square difference between the predicted value and the true value. LSTM should first preprocess the data of the training set, then create the LSTM model, set the loss function and optimize the algorithm, use the data of the training set to train the model, then use the trained model to predict the data of the test set, and then use the real value and the predicted value data as the mean square error and the mean absolute error.

3. Prediction of Food Pollutants Based on Random Forest and LSTM and Its Application

3.1. Data Sources

The experimental data comes from the aflatoxin monitoring data of the peanut oil supply chain of COFCO Feixian Zhongzhi Oil Co., Ltd. There are 12 links in the peanut oil supply chain, namely, acquisition, storage, screening, warehousing, crushing, embryo rolling, steaming and frying, pressing, fine filtering, packaging, warehousing and sales. In the experiment, 1000 groups of data were randomly selected, each group of 12 data corresponds to 12 links, and the actual amount of data used was 12000. The supply chain of peanut oil is shown in Figure 3.

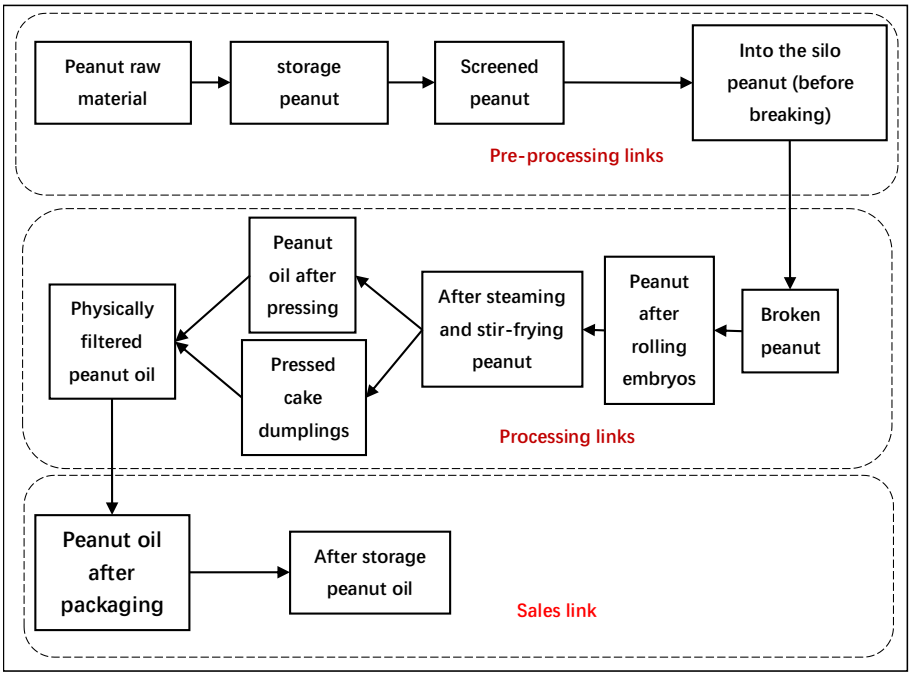


Figure 3: Peanut oil supply chain

3.2. Experimental Process

In this paper, Random Forest and LSTM algorithms are applied to the mycotoxin prediction of each link of peanut oil supply chain, which can efficiently predict the aflatoxin content of peanut oil products in the whole supply chain, thus ensuring the health and safety of peanut oil products. The implementation flow chart of Random Forest and LSTM algorithm is shown in Figure 4.

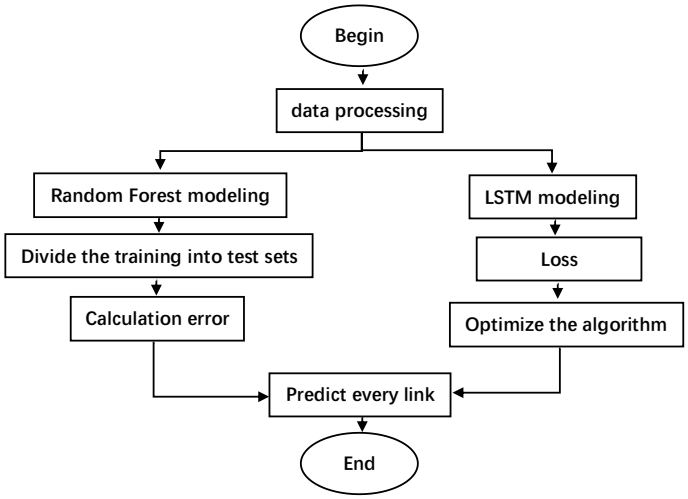


Figure 4: Flow Chart of Random Forest and LSTM Algorithm Implementation

There are 12 links in the supply chain. Each link has a grain pollutant value, namely aflatoxin. First, according to these data, Random Forest and LSTM models are established to predict. For example, if the pollutants in the remaining links 3 to 12 are predicted according to the pollutant data in links 1 to 2, a model needs to be built with the values in links 1 to 2 as input and links 3 to 12 as output. After training, the prediction model, learn the relationship between the links. When the new data of links 1 to 2 is input again, the predicted values of links 3 to 12 will be output. The prediction model can predict the data of the first two phases and the data of the last 10 phases. Input phases 1-2 and output phases 3-12 are recorded as 2to10 phases. It is also possible to predict the data of the remaining 9 links based on the data of links 1 to 3, and so on.

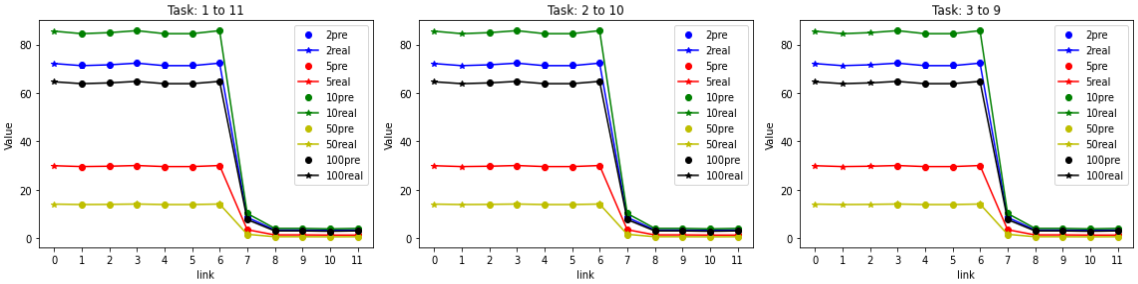
3.3. Experimental Methods

The known data is stored in the file. First, assign the read data. The previous 11 links are used as input, and the 12th link is used as output. Assign the data in the first 11 columns to X, that is, as input, and the value in the 12th column is used as output Y.

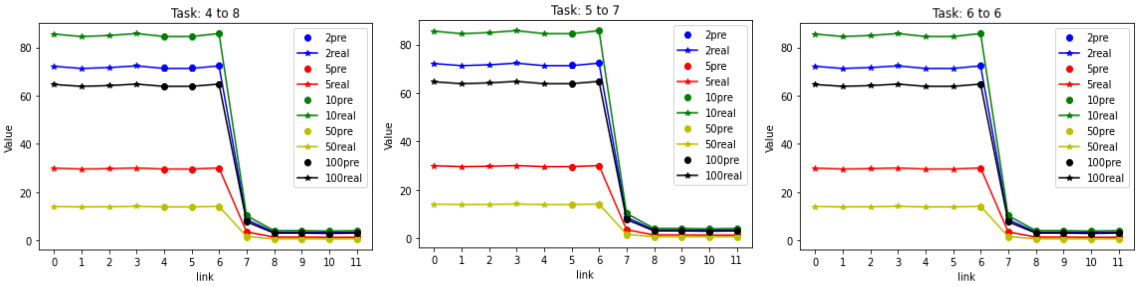
First of all, based on the mycotoxin prediction of Random Forest method, 1000 groups of data are randomly sampled, and the data are divided into two groups, training set and test set. This paper takes 30% of the data as the test set, and assigns the divided data to the input of the training set, the input of the test set, the output of the training set and the output of the test set, and it is a random process to build the model, generate the data set, and split the data set, The random number seed is taken to avoid the contingency of the experiment, so that the data generated each time is not exactly the same. Therefore, the random number seed is taken as 5, and then the Random Forest regressor is called. The output of the training set obtained is compared with the real output. The predicted value and the real value are compared to calculate the mean square error to obtain the prediction accuracy, and the experimental results are shown on the graph.

Secondly, for mycotoxin prediction based on LSTM method, first create LSTM model, set loss function and optimization algorithm, use gated loop unit modeling and use two-layer GRU network to build. Use LSTM to train the data, predict the classification results based on the characteristics of the training set, and then call the output of the training set obtained by your own model to compare with the real output. Calculate the accuracy. The characteristics of the test set also need to predict the classification results, and then call the output of the training set obtained by your own model to compare with the real output. Compare the predicted value with the real value to calculate the mean square error, and get the prediction accuracy, the experimental results are shown on the graph.

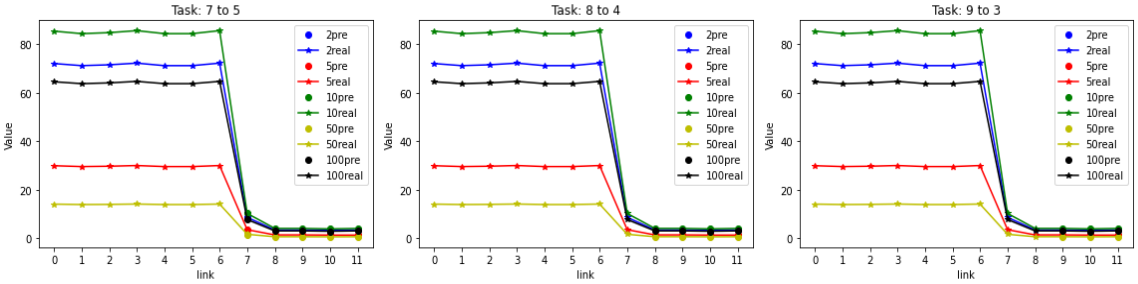
3.4. Experimental Result



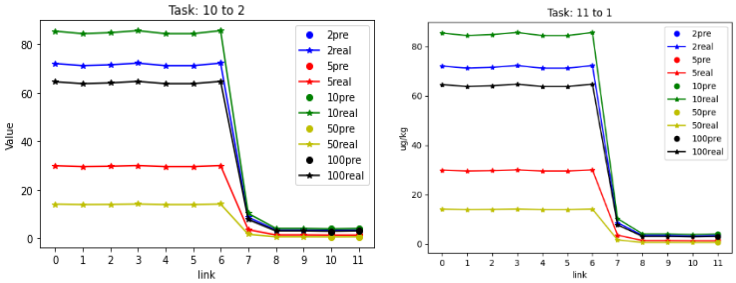
a. input 1 predict 2-12 b. input 1-2 predict 3-12 c. input 1-3 predict 4-12



d. input 1-4 predict 5-12 e. input 1-5 predict 6-12 f. input 1-6 predict 7-12



g. input 1-7 predict 8-12 h. input 1-8 predict 9-12 i. input 1-9 predict 10-12



j. input 1-10 predict 11-12 k. input 1-11 predict 12

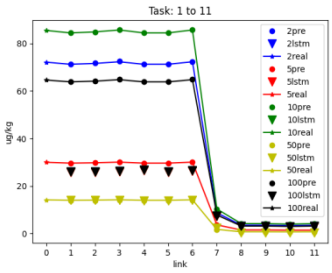
Figure 5: Aflatoxin in each link predicted based on Random Forest

For the prediction of each link, it is necessary to change the initial assignment of X and Y, select 100 groups of data, and the initial size of input data is 100 × 11, the output is 100 × 1. If you want to change to the first 10 links and forecast the next 2 links, change the size to 100 × 10, the output is 100 × 2. Get the results, as shown in Figure 5. The 11 result graphs are: the values of the first 11 links are known, and the values of the 12th link are predicted; Given the values of the first 10 links, predict the values of the 11th and 12th links; and so on. In order to make the comparison between the predicted value and the real value more obvious, multiple groups of data results are extracted from the same graph to facilitate the analysis of the prediction effect.

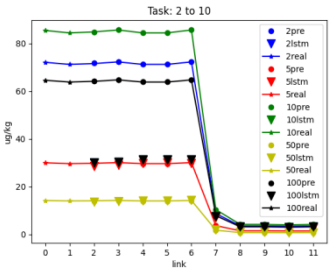
Among them, 2Pre is the second group of prediction data, and 2real is the second group of real

data. By analogy, the second, fifth, tenth, fiftieth, and hundredth groups of data are selected to form a comparison. Through comparison, it can be seen that the difference between the predicted value and the real value is very small, approximately coincident. Because the data is discrete, the difference between the aflatoxin concentration and content of each group of data is large.

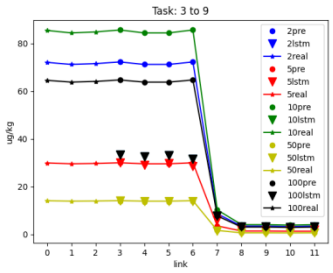
In order to compare with the Random Forest prediction method, the comparison chart of the prediction results using Random Forest and LSTM method is shown in Figure 6.



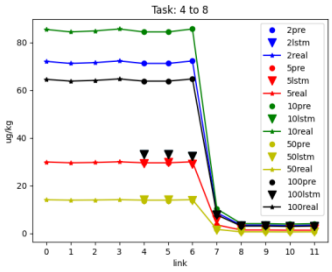
a. input 1 predict 2-12



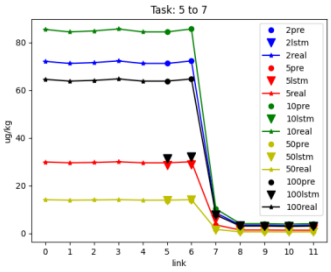
b. input 1-2 predict 3-12



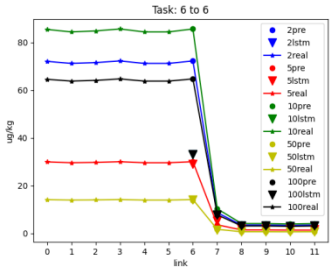
c. input 1-3 predict 4-12



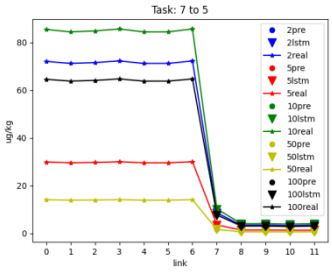
d. input 1-4 predict 5-12



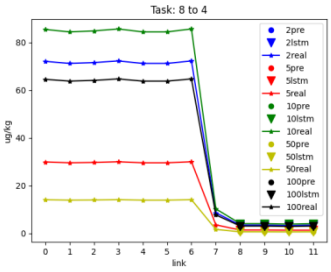
e. input 1-5 predict 6-12



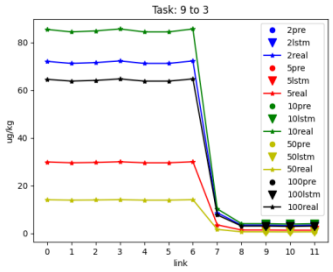
f. input 1-6 predict 7-12



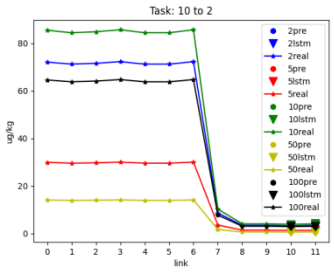
g. input 1-7 predict 8-12



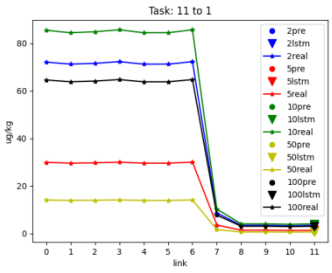
h. input 1-8 predict 9-12



i. input 1-9 predict 10-12



j. input 1-10 predict 11-12



k. input 1-11 predict 12

Figure 6: Comparison Effect between Random Forest and LSTM

The data is discrete, and the data volume is small for LSTM, so LSTM cannot learn enough rules. As a result, only 11 to 1, 10 to 2, 9 to 3, 8 to 4, and 7 to 5 groups of data are successfully fitted, because there is enough historical data in the early stage, but as the data volume provided in the later stages decreases, and more prediction data are available, LSTM can obviously see that there is a large difference between the selected data of the 100th group and the aflatoxin concentration

predicted by the Random Forest. It can be seen that for the discrete system

data in this paper, the LSTM prediction effect suitable for continuous system time series analysis is not ideal, while the Random Forest itself is an integrated algorithm with strong adaptability to data sets, and is good at processing missing data and discrete data. The data standardization requirements are relatively low, and it is not easy to fall into over fitting. Therefore, the Random Forest method is more ideal to achieve the data prediction of the discrete system in this paper, The prediction effect of grain pollutants is better than that of LSTM.

3.5. Result Analysis

There are two kinds of calculation formulas commonly used to evaluate the prediction error, namely, root mean square error and average absolute error. The root mean square error (or standard error) is the square root of the ratio of the square of the deviation between the observed value and the true value and the number of observations n. In actual measurement, we can only observe n in a very small range, and only replace it with the most consistent one. Here is the root mean square error formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{5}$$

The absolute value of each single observation value and calculated operative mean value. Compared with the average error, the average absolute error is converted into absolute value and cannot be eliminated by positive and negative phases. Therefore, the average absolute error can reflect the reality of the prediction error with the highest efficiency. The average absolute error formula is as follows:

$$MAE = \frac{\sum_{i=1}^{N}|(y_i - x_i)|}{N} \tag{6}$$

From the results, it can be concluded that the accuracy of the results of the Random Forest model test set and training set is high, and the mean square error is small, which proves that the implementation is very successful. From the results of the code, it can be seen that the error between the predicted value and the real value is small, and the average absolute error result is more accurate, as shown in Table 1.

Table 1: Precision and mean square error results of Random Forest

Model	RMSE (Random Forest)	RMSE (LSTM)	MAE (Random Forest)	MAE (LSTM)
1to11	0.0733	20.3887	0.0460	14.9502
2to10	0.0667	17.0384	0.0425	12.0458
3to9	0.0607	14.6300	0.0383	10.2142
4to8	0.0569	12.3375	0.0343	8.5860
5to7	0.0490	9.7335	0.0284	6.8116
6to6	0.0393	6.3201	0.0208	4.5067
7to5	0.0214	0.0681	0.0082	0.0325
8to4	0.0190	0.0306	0.0061	0.0155
9to3	0.0083	0.0246	0.0046	0.0167
10to2	0.0089	0.0239	0.0047	0.0159
11to1	0.0126	0.0197	0.0060	0.0119

It can be seen from Table 1 that according to the obtained mean square error and mean absolute error data, the Random Forest algorithm performs well in predicting aflatoxin, and the results are

relatively ideal. For the mean square error and mean absolute error of 11 to 1, 10 to 2, 9 to 3, 8 to 4, and 7 to 5, it is found that the predicted value and the true value are highly consistent, but the other groups of data have large errors. When LSTM algorithm is used to predict aflatoxin with a large number of inputs, that is, 11 to 1, 10 to 2, 9 to 3, 8 to 4, and 7to5, the results of these models are similar to those of Random Forest. However, in the 1to11, 2to10, 3To9, 4To8, 5to7, and 6to6 models, we can see that the mean square error and mean absolute error have significantly increased, and the maximum is 20.3887. It can be seen that the effect of Random Forest algorithm is more ideal.

In the aflatoxin prediction based on Random Forest, aflatoxin in 12 links of peanut oil supply chain is discrete system data. A large number of decision trees contained in Random Forest can better improve the ability to describe discrete systems and have higher prediction accuracy for discrete systems. At the same time, because the Random Forest uses the integrated algorithm, its accuracy is higher than that of the single algorithm and traditional algorithm which are only applicable to continuous process prediction. In addition, the training of Random Forest model is fast, easy to form parallel operations, and can predict the interaction between different data.

In the prediction of aflatoxin based on LSTM, because there are four full connection layers in each LSTM cell, and the network of the algorithm itself is very deep, LSTM takes a lot of time to run the program. If LSTM needs to predict data with a large time span, it will take a lot of computing time.

It can be seen that Random Forest has more advantages than LSTM in the prediction of pollutants in the food supply chain. Because a large amount of data needs to be used in the neural network in order to truly achieve the effectiveness of data processing, LSTM neural network will greatly reduce the prediction accuracy in the case of less data in the input layer. When using a Random Forest to face this situation, multiple decision trees can be combined into a model to achieve better prediction results.

4. Conclusion

Aiming at the problems related to the variety and complexity of food pollutants in the research of food pollutant prediction, the traditional prediction methods for food pollutants consider few factors and cannot accurately predict the type and content of pollutants, according to the integrated learning theory applicable to discrete system modeling and the neural network theory applicable to time-space series prediction, the prediction methods for food pollutants based on Random Forest and LSTM are proposed respectively.

The main research work and innovation points of this paper include the following:1) Based on the theory of integrated learning, a grain pollutant prediction model based on stochastic forest algorithm is proposed to predict the content of grain pollutants and calculate the prediction accuracy.2) Based on the theory of neural network, a grain pollutant prediction model based on LSTM algorithm is proposed to realize the prediction of grain pollutant content, calculate the prediction accuracy, and compare the method with the Random Forest algorithm.

Through the application of aflatoxin data in peanut oil supply chain in this paper, the results show that stochastic forest algorithm has greater advantages in predicting food pollutants, and is superior to LSTM algorithm in terms of prediction accuracy and operation efficiency. Of course, Random Forest also have a lot of room for improvement. The Random Forest is composed of multiple decision trees, and because each decision tree needs an independent operation space, which just affects the efficiency of calculation. Assuming that the number of decision trees is appropriately changed, the number of decision trees and the operation time can reach an optimal combination, which can avoid wasting too much decision tree resources.

It is also known from the application results of examples that, for the case data in this paper, the prediction effect of LSTM is generally worse than that of Random Forest, and the effect of LSTM algorithm in predicting some links is relatively good, which is similar to that of Random Forest, indicating that LSTM has certain advantages in predicting time series, and can solve the problems of gradient disappearance and gradient explosion of traditional neural networks. The algorithm itself has strong self-learning ability, but for the model with small input data, the effect is not very good, and the mean square error and the average absolute error are both large.

Acknowledgement

Thanks to the National Key Research and Development Program of China No. 2020YFC1606801-03.

References

[1] Purushtham S, Shetty H. Storage fungal invasion and deterioration of nutritional quality of rice. Journal of Mycol Plant Pathol, 2010, 40 (4): 581-585.
[2] Piotr M, Yann L. Statistical Machine Learning and Dissolved Gas Analysis: A Review. IEEE Transactions on Image Processing, 2017, 28: 2-18.
[3] Wang Hongxing, Tao Zhengguo. The harm of mycotoxins in feed and its prevention measures. Veterinary Medicine and Feed Additives, 2018, 5 (3): 19-20.
[4] Lan Xueping, Chen Jinying, Jiang Youjun, et al. Research of Building Grain Storage Quality Prediction Model Based on BP Neural Network Algorithm. Chinese Journal of Cereals and Oils, 2020, 35 (11): 147-151.
[5] Manstretta V, Rossi V. Comparison of three modelling approaches for predicting deoxynivalenol contamination in winter wheat. Toxins (Basel), 2018, 10 (267): 1-15.
[6] Jia Xiaoyong, Xu Chuansheng, Bai Xin. The Establishment of Least Square Method and Its Thinking Method. Journal of Northwest University (Natural Science Edition). 2006, 2006 (03): 507-511.
[7] Deng Yurui, Zhou Yong, Cong Wei, et al. Research on prediction model of grain mildew probability based on naive Bayesian algorithm. Chinese Journal of Cereals and Oils, 2019, 34 (S2): 35-38.
[8] Zheng Moli, Zhao Yanke, Yan Min, et al RDPSO-SVM based intelligent evaluation method for loss in grain post production storage. Computer and modernization, 2020, 2020 (3): 72-76.
[9] Wang Chunhui, Zhou Shenglu, Wu Shaohua, et al. Prediction of grain output in Jiangsu Province based on multiple linear regression model and grey correlation analysis. Urban geology, 2014, 2014 (04): 35-53.
[10] Peromingo B, Rodriguez A, Bernaldez V, et al. Effect of temperature and water activity on growth and aflatoxin production by Aspergillus flavus and Aspergillus parasiticus on cured meat model systems. Meat Science, 2016, 2016 (122): 76-83.
[11] Vanderfels-Klerx H, Olesen J, Madsen M, et al. Climate change increases deoxynivalenol contamination of wheat in north-western Europe. Food Additives and Contaminants Part A -Chemistry Analysis Control Exposure & Risk Assessment, 2012, 29 (10): 1593-1604.
[12] Tian K, Huang Z, Wang X, et al. Research progress on in vitro models for evaluating drug-induced neurotoxicity. Drug Evaluation Research, 2020, 43 (7): 1433-1443.
[13] Guo Yafei, Fan Chao, Yan Hongtao. Grain yield prediction based on principal component analysis and particle swarm optimization neural network jiangsu agricultural sciences, 2019, 47 (19): 241-245.
[14] Chen Dingyu, Wan Jian, Cheng Hanfeng. Prediction of China's grain output based on ARIMA model. Marketing, 2019, 2019 (13): 95-96.
[15] Ortega J, Ferr ś J, Berbel O, et al. Environmental neurotoxins (IV). Tobacco, alcohol, solvents, fluoride, food additives: adverse effects on the fetal and postnatal nervous system. Preventive measures. Acta Pediatrica Espanola, 2006, 64 (10): 493-502.
[16] Tian K, Huang Z, Wang X, et al. Research progress on in vitro models for evaluating drug-induced neurotoxicity. Drug Evaluation Research, 2020, 43 (7): 1433-1438.
[17] Pan Yan. Research on Application of Decision Tree Algorithm in Curriculum Association Analysis of Higher Vocational Colleges. Modern Information Technology, 2019, 3 (2): 159-161.
[18] Yan Zhengxu, Qin Chao, Song Gang Stock price prediction based on Pearson's feature selection of stochastic forest model. Computer engineering and application, 2021, 57 (15): 286-296.
[19] Liu Min, Lang Rongling, Cao Yongbin. Number of trees in random forests. Computer Engineering and Application, 2015, 51 (5): 126-131.

[20] Cheng Min, Zhang Yaowen, Jiang Jiyi, et al. Rainfall prediction analysis based on time series model. Water Science and Engineering Technology, 2019, 2019 (1): 1-5.
[21] Dobbins W. BOD and oxygen relationship in streams. Journal of the Sanitary Engineering Division, 1964, 90 (3): 53-78.
[22] Chen Y, Rangarajan G, Feng J, et al. Analyzing multiple nonlinear time series with extended Granger causality. Physics Letters A, 2004, 324 (1): 26-35.
[23] Devi A, Maragatham G, Boopathi K, et al. Hourly day-ahead wind power forecasting with the EEMD-CSO-LSTM-EFG deep learning technique. Soft Computing, 2020, 24 (16): 12391-12411.