Short-Term Relay Quality Prediction Algorithm Based on Long and Short-Term Memory

XUE Wendong*, Chai Yuan, LI Qigan, HONG Yongqiang, ZHENG Gaofeng

Department of Instrumental and Electrical Engineering, Xiamen University, Xiamen 361102

Abstract: The fraction defective of semi-finished products is predicted to optimize the process of relay production lines, by which production quality and productivity are increased, and the costs are decreased. The process parameters of relay production lines are studied based on the long-and-short-term memory network. Then, the Keras deep learning framework is utilized to build up a short-term relay quality prediction algorithm for the semi-finished product. A simulation model is used to study prediction algorithm. The simulation results show that the average prediction absolute error of the fraction is less than 5%. This work displays great application potential in the relay production lines.

Key words: Relay Production Line; Long and Short-Term Memory Network; Keras Deep Learning Framework; Quality Prediction

1 Introduction

The development of quality prediction methods has gone through the process from statistical process control to statistical/machine learning, and now the deep learning is the novel development trend[20,2,13]. The method of statistical process control, which used controlling diagram based on sampling inspection to realize the quality control of products, is used wildly in the manufacturing enterprises[18,16]. But statistical process control is a hysteresis method that can't be used in the real-time controlling of product quality. The online quality prediction algorithm has become the key to the product quality controlling[12, 6].

The methods of statistical or machine learning depend on the statistical data characteristics to realize the online quality prediction. But the multi-dimensional data characteristics are difficult to obtain from the close-coupled statistical data. Thanks to the development of calculate capability and optimization algorithm, the deep learning has been applied in various industrial fields[19, 11]. With the ability of automatic rule setting and data characteristics choosing, deep learning technology has displayed great advantages over traditional machine learning. The operation mode of deep learning is more similar to thinking behaviors of human brain than conventional methods, and has been applied in the fields of online quality and failure prediction[21, 14]. Chen[3] et al proposed a Long and Short-Term Memory (LSTM) neural network architecture to realize the short-term electrical load prediction. Ge[5] et al utilized deep confident neutral that comprised of single or multi-sensors to achieve feature detection and state recognition, which has been used to realize the bearing fault diagnosis. Bruin[1] et al proposed a long and short-term memory based on recurrent neural networks (LSTM-RNN) to investigate fault diagnosis of train railway.

In this paper, a short-termquality prediction algorithm based on LSTM is proposed for the relay production lines. This algorithm extracts multi-parameter characteristics from the production data as the input vector, by which fraction defective of semi-finished products is predicted. The prediction results can be used to foresee the productor quality and increase the productivity. The modelling and computation process of this algorithm has been discussed.

2 Long-Term Memory Network

2.1 Recurrent Neural Networks

Compared with the traditional forward neural network, Recurrent Neural Networks (RNN) adopted a circular information flow that stored information data of the previous time step in the hidden layer. The current output of the network is defined by the previous state, which is similar to the working mode of memory in the human brain[15], as shown in Fig. 1.

The RNN network displays a chain structure after it is expanded, and there is also a corresponding relationship between the RNN network and time series. Then, C_t, the state of nerve cell A at the time of t, can be given in equation (1):

$$C_t = f(W C_{t-1} + U x_t) \tag{1}$$

Where, f is the nonlinear activation function, W and U are weight factors for the previous state of nerve cell C_{t-1} and input x_t *respectively*. Some common activation functions include functions of *Tanh*, *Sigmoid*, *Softmax*, and *ReLU* et al[8]. With the partial derivative of error was being calculated according to the chain rule and backpropagation algorithm, the error gradient could be gained by multiplying continuously a series of differentials that are smaller than 1. With the increase of process time t, the error gradient tends to zero quickly, and the weight of network cannot be trained and renewed, the network stops learning and the error gradient disappears[10, 4].

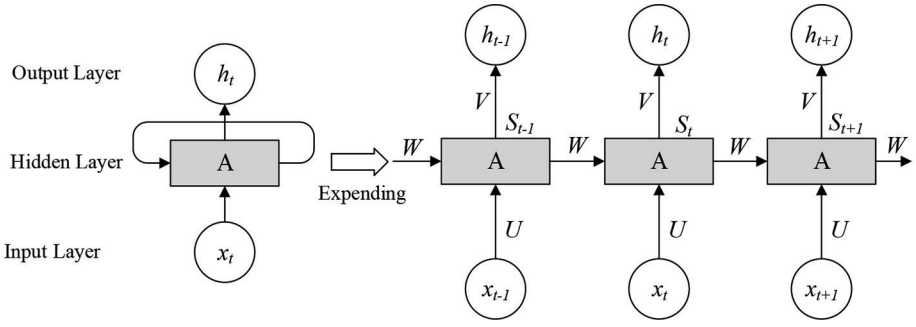


Fig. 1 The Expanding Structure of RNN Network

2.2 Long-Term Memory Network

The LSTM is one special RNN structure that was invented byHochreiter and Schmidhuber in 1977[9]. In recent years, LSTM has been used in the fields of voice recognition[7], handwriting recognition, text generation[17] and disease prediction. A typical LSTM expanding structure is shown in Fig. 2.

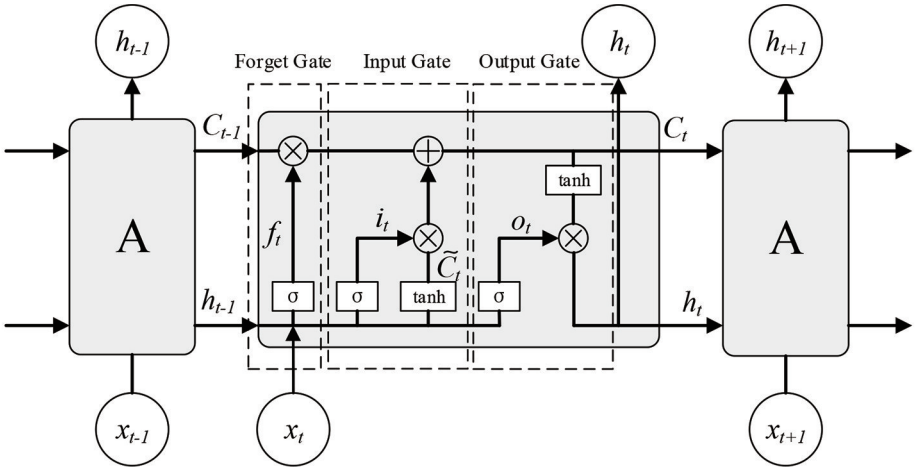


Fig. 2 The Typical LSTM Expanding Structure

Compared with RNN network, three special gate structures-forget gate, input gate, and output gate are introduced into the LSTM.

(1) Forget Gate

The forget gate is used to gain the previous state information ofthe nerve cell, and determine the information to be abandoned. It reads the output of previous state h_{t-1} and the input of current state x_t. The outputs of *Sigmoid* layer in forget gate, which distribute in the range of 0 to 1, are used to adjust the influence coefficient of previous state on the current state. If the output is 0, the information of previous state will be abandoned completely, and the previous state has no effect on the current state. If the output is 1, the information of previous state will be fully retained to describe the effect of previous state.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

(2) Input Gate

After the process ofthe forget gate, the effect of state information is determined by the nerve cell's own demand. Whether the previous state information needs to be abandoned or updated is determined by the state time series $i_t$gained from the Sigmoid layer. Then, a new state information variable $\widetilde{C}_t$ is created by the Tanh Layer based on the output of previous state h_{t-1} and the input of current state x_t. The time series i_t is the factor that controls the current state information updating of nerve cell.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

The state information updating which is determined by the output of forget gate f_t and the time series of input gate i_t can be described as:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{5}$$

(3) Output Gate

The output gate controls the network output, which works as the same way as the forget gate and the input gate. Firstly, a coefficient o_t in the range of 0 and 1 is generated by the Sigmoid layer. Then, the final output h_t of network can be gained by multiplying coefficient o_t and the state information that transferred throughout the Tanh layer. The final output h_t of network is written as:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * tanh\ (C_t) \tag{7}$$

By comparing equations (1) and (5), it is obvious that the state information of nerve cell in RNN network is renewed by overwriting. The overwriting method is depicted as:

$$C_t = f(C_{t-1}, x_t) \tag{8}$$

where the gradient is described in the way multiplication to gain the derivation.

In the LSTM system, however, the renew state information can be calculated as:

$$C_t = \sum_{\tau=1}^{t} \triangle C_\tau \tag{9}$$

where the accumulative method is used to replace the multiplication process, by which the disappearance of the gradient can be avoided.

3 Model Building

3.1 Data Preprocess

The LSTM network is utilized to analyze the process data of relay production. The parameters of semi-finished product are acquired from the actual relay production line. The primary data is shown in Table 1.

As depicted in Table 1, the dimension ranges of different parametersvaries, which should be nondimensionalized and adjusted into the same dimension range. In order to increase the solving speed of the optimal solution in the method of gradient descent, the parameter data is uniformized firstly. As the gate activation function of LSTM network is *tanh*, the parameter data range is adjusted into the range$^{[-1, 1]}$. The uniformization results of parameter data are shown in Fig 3. The number of samples used in the work is about 13,000. Here, Fig. 3 (a) ~ (f) refers to the parameters of Dynamical Pressure, Interval, Super-path, Coil Resistance, Pick-up Voltage, Release time in Table 1.

Table 1 The Parameter Data of Relay Semi-Finished Product (Selected Data)

No.	Time	Dynamical Pressure (N)	Interval (mm)	Super-path (mm)	Coil Resistance (Ω)	Pick-up Voltage (V)	Release time (V)	Qualified
1	20170408-0:22:9	0.373	0.384	0.220	79.900	5.880	2.080	OK
2	20170408-0:22:14	0.380	0.388	0.213	79.900	5.950	2.340	OK
3	20170408-0:22:19	0.378	0.364	0.230	79.700	6.070	2.380	OK
4	20170408-0:22:25	0.373	0.384	0.225	78.900	5.730	2.040	OK
5	20170408-0:22:30	0.412	0.381	0.216	79.000	5.990	2.150	OK
…	…	…	…	…	…	…	…	…
100	20170408-0:31:17	0.401	0.381	0.231	79.900	6.590	2.380	OK
101	20170408-0:31:22	0.391	0.388	0.227	80.000	6.360	2.190	OK
…	…	…	…	…	…	…	…	…

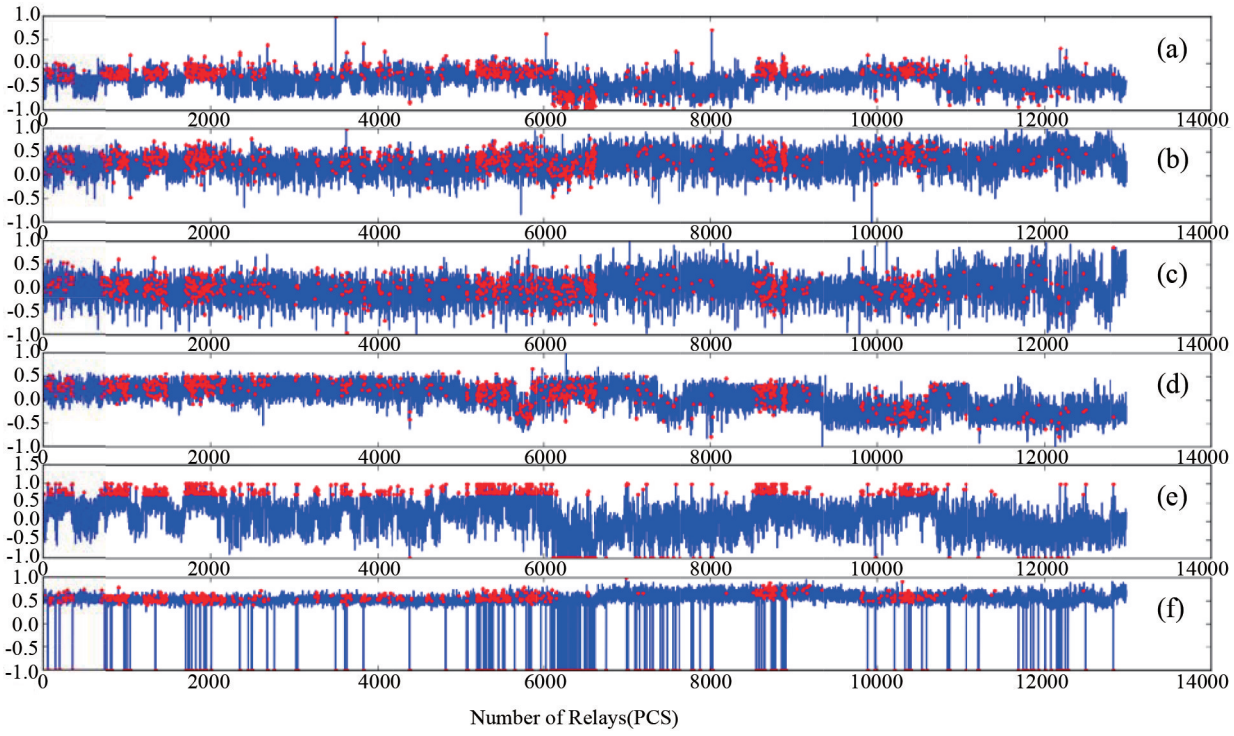


Fig. 3 The Uniformization Results of Relay Parameters

The red points in the Fig.3 represent the quantity of defective products during the production time. The red points distribute intensively in some time periods, which are the unstable periods of relay producing line. In this work, a proportion coefficient P_{ng} is introduced to describe the unstable state of production line in the presented data windows. The proportion coefficient P_{ng} is written as:

$$P_{ng} = \frac{N_{ng}}{N_{window}} \tag{10}$$

The fraction defective in the slip time window can be calculatedusing equation (10), where N_{ng} is the number of defective products and N_{window} is the number of all the products. Then the fraction defec-

tive curve of relay can be gained by adjusting the time windows, the curve of fraction defective vs time is depicted in Fig. 4. The fraction defective is normalized into the range$^{[0, 1]}$, the sample number in each time window is 50.

3.2 Model Selection

There are 4 common structures in LSTM network, as shown in Fig. 4, and 3 typical models of input and output- one-to-many, many-to-one and many-to-many. Figure 4(a) shows the input-output model of one-to-many which is usually used to realize the character recognition. Figure 4(b) shows the one-to-many model that can be used to gain sentiment analysis, there are multiple outputs that can be used to describe the diverse results at the same time. Figure 4(c) and (d) show the input-output many-to-many model that is appropriate for the operation process of serialized parameters. Compared with the structure in Fig. 4(d), the structure of Fig. 4(c) has more time lag. And, structure in Fig. 4(c) is suitable for the machine translation; while the structure in Fig. 4 (d) is more ideal for the video frame-by-frame tagging or voice recognition. In this work, the production information of relay is a typical kind of time series parameter that varies with the production time. There are 6 detecting parameters to predict the quality of relay that demand multiple input network, and the quality of relay is the single output for this controlling network. So, the structures in Fig. 4 (b) of many-to-one is used in our work to build up prediction model.

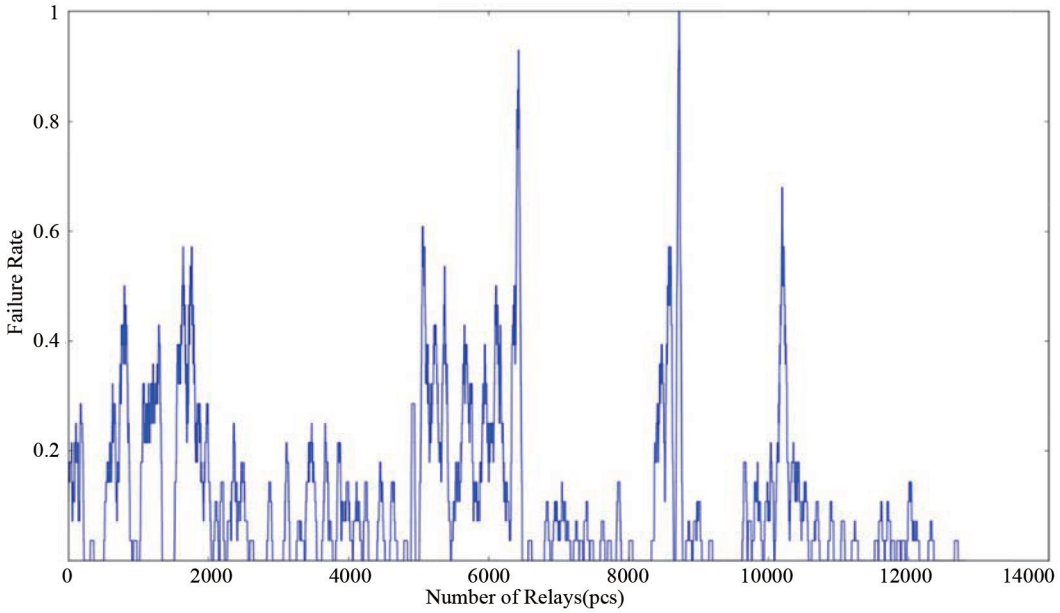


Fig. 4 The Curve of Fraction Defective vs Time

2.3 Model Building

In this work, *Keras* frame based on Python language is utilized to build up the prediction model. The *Keras* frame, which is a high-level programming interface for neural network, can be used as black box model to help user to focus on the quick experiment of the algorithm.

There are two parts in the controlling network structure, as shown in Fig. 6. The left part in Fig. 6describes the working process of characteristics learning from the relay parameter tensor in the time window. The tensor is composed of sample number (Field name: *n_samples*), time step (Field name: *time_steps*), parameter number (Field name: *dim_input*). The dimension of relay parameter tensor is (1024, 100, 6). The input of control network is a

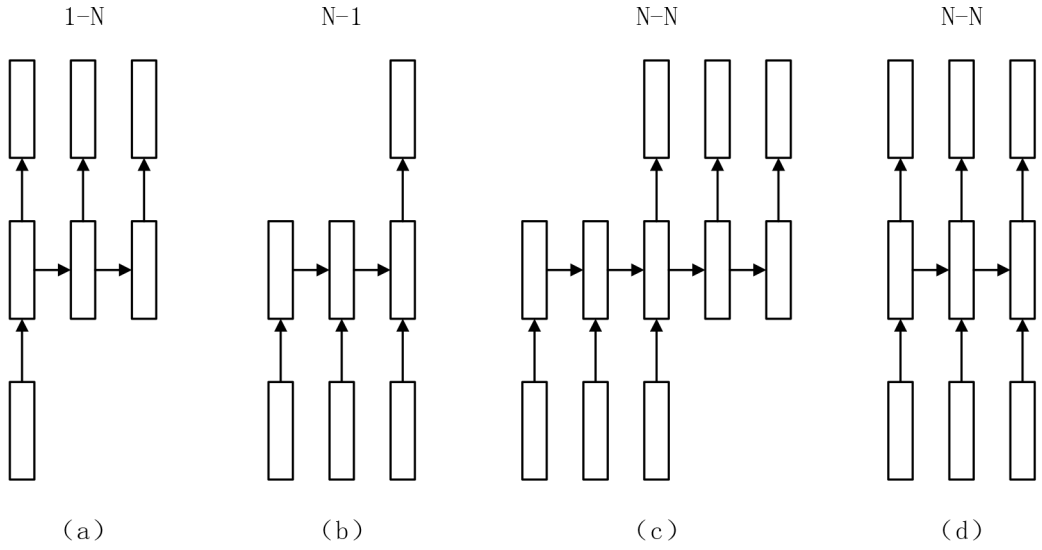


Fig. 5 The Common Structures Used in LSTM Network

100×6 two-dimensional vector, there are 100 samples data in each time window. In order to improve the learning effect of control network, the control network comprises two LSTM layer of LSTM1 and LSTM2. The right part shows the characteristic learning process from the tensor of fraction defective, of which the dimension is (1024, 100, 1). There is a single LSTM3 layer in the right part of this network structure for the tensor of fraction defective. The Dropout technology is utilized in the hidden layer of LSTM3 network to overcome the disadvantage of overfitting. The activation function *Hard _ sigmoid* is used as recurrent activation in LSTM1~LSTM3, by which the learning process can be speeded up and the gate coefficient can be generated; the activation function *Tanh* is used for activation, and works as the state information and final input of nerve cell. The left and right parts of network are connected to the connection layers of *Dense*1 and *Dense*2 respectively, which were activated by the *ReLU* function to avoid gradient diffusion and gradient explosion of the network. *Dense*1 and *Dense*2 and are connected by the merged layer *Merge*. The results of both parts are linearly added through the merged layer *Merge* and pass to the connection layer Dense3 to generate the final output. In this model, optimizer *Adam* is used to realize the network training.

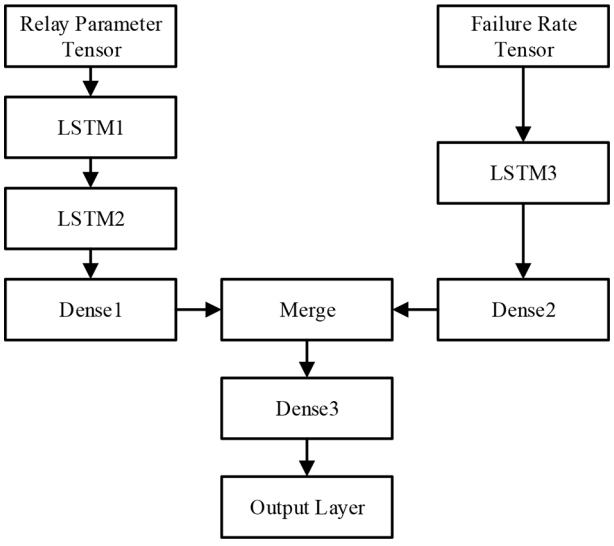


Fig. 6 Short-Term Prediction Model of Relay Quality

4 Results and Discussion

The experimental platform of LSTM is Window 10 operation system, i7 4700MQ (2.4GHz) CPU, GT755M GPU, 8GB System Memory, 2GB Video Memory. The production data is divided into two parts of testing set and validation set. The percents of testing set and validation set are 80% and 20%, respectively. The data is loaded to the system for batch training. The training process is shown in Fig. 7. Here, *iterator* is the iteration times in one data training process. It is defined by the size of data batch

batchSize, and is 1024 in this work. And, *epoch* is the rounds of the data training processes, which describes the total training times required to complete the data input into the model for all of production samples, in this work *epoch*=20. Mean square error function *mse* is adopted as the loss function *loss* during model training. The sequence of training set should be randomly shuffled and rearranged to start a new training process after each *epoch* iteration round. The curve of loss function is shown in Fig. 8, and the calculated predicted fraction defective is shown in Fig. 9.

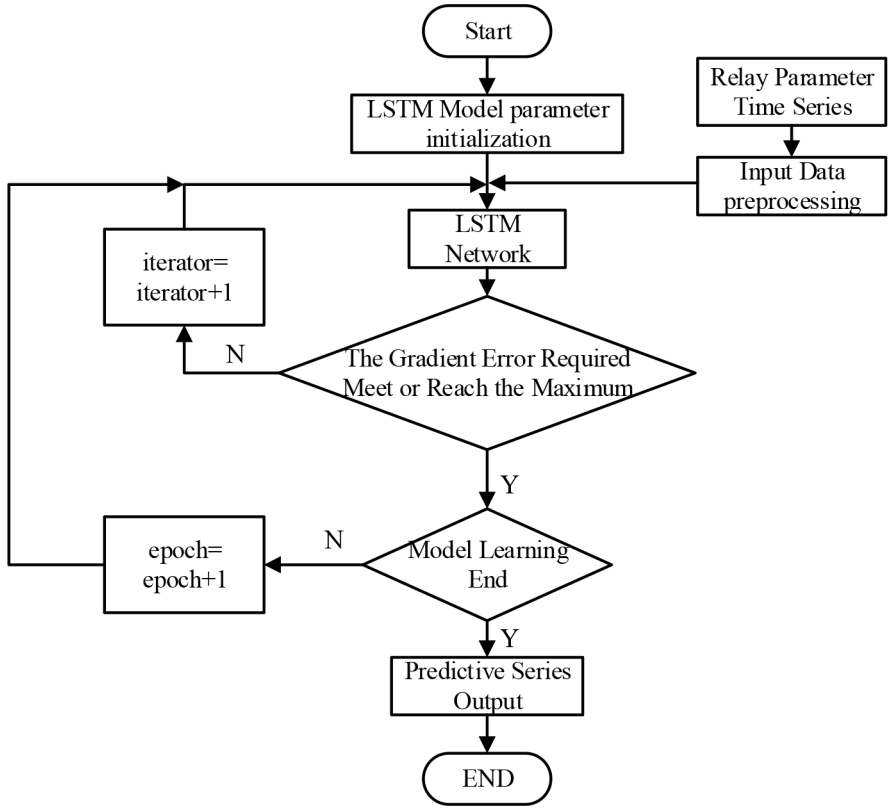


Fig. 7 Flow Chart of LSTM Algorithm Training Process

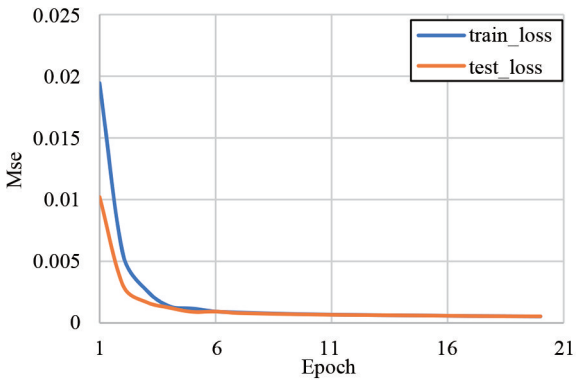


Fig. 8 Loss Function Curve of Model Training

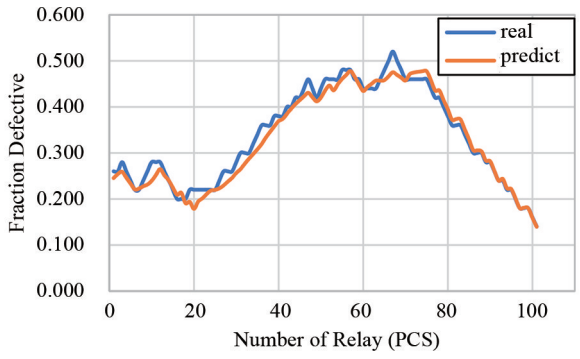


Fig. 9 Calculated Results of Predicted Fraction Defective

The loss curves of mean square error ofthe training set and verification set are shown in Fig. 8. Then, the model will step into convergence stability after 6 rounds of *epoch* iteration. The real fraction defective and predicted fraction defective are shown in Fig 9, which depicts that the LSTM model net-

work can predict fraction defective precisely. The model performance evaluation of Root-Mean-Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are presented in Table 2. The results show that the prediction network can gain precise results, and the MAPE is less than 5% that is suitable for the industrial applications of quality prediction.

Table 2 Model Performance Evaluation

Performance	Results
RMSE	0.017491
MAE	0.013889
MAPE	4.1876%

5 Conclusion

The deep learning algorithm based on LSTM is introduced to build up relay quality prediction model. In this work, *Keras* framework is utilized to build up short-term quality predictive model based on LSTM theory. The actual production data has been used as training and verification data for this predictive algorithm. Through the analyses of production data, a quality predictive model is built up. The simulation results show that the MAPE of this model is less than 5% , which is suitable for the industrial application for quality prediction. The novel short-term quality prediction algorithm has displayed great application potential in the industrial production line, which can provide technical support to quality control staff in the enterprise.

ACKNOWLEDGMENT

The thesis was funded by Fujian Science and Technology Key Project (No. 2016H6022, 2018J01099,2017H0037).

References

[1] BRUIN, T. D. (2016). Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks, *IEEE Transactions on Neural Networks & Learning Systems*, 28(3), pp. 523 - 533.

[2] CHEN, X. (2019). A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning, *Medical Physics*, 46(1), pp. 56-64.

[3] CHENG, Y. (2017). *Short-Term Electricity Demand Forecasting Based on Artificial Neural Network.* Zhejiang University.

[4] CHUNG, J. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, *Eprint Arxiv*, pp.

[5] GE, Q. (2016). *Research on data-drive fault diagnosis method based on deep belief network.* Harbin Institute of Technology.

[6] GHOSH, P. (2017), Published. Learning human motion models for long-term predictions. 3D Vision (3DV), 2017 International Conference on. IEEE, 458-466.

[7] GRAVES, A. (2014), Published. Hybrid speech recognition with Deep Bidirectional LSTM. Automatic Speech Recognition and Understanding. 273-278.

[8] GUO, J. (2010). Real-Time Short-Term Traffic Speed Level Forecasting and Uncertainty Quantification Using Layered Kalman Filters, *Transportation Research Record Journal of the Transportation Research Board*, 2175(2175), pp. 28-37.

[9] HOCHREITER, S. (1997). Long Short-Term Memory, *Neural Computation*, 9(8), pp. 1735-1780.

[10] JOZEFOWICZ, R. (2015), Published. An empirical exploration of recurrent network architectures. International Conference on International Conference on Machine Learning. 2342-2350.

[11] LECUN, Y. (2015). Deep learning, *Nature*, 521 (7553), pp. 436 - 444.

[12] LIU, Y. (2018). Ensemble deep kernel learning with application to quality prediction in industrial polymerization processes, *Chemometrics and Intelligent Laboratory Systems*, 174(3), pp. 15-21.

[13] MEYER, A. (2018). Machine learning for real-time prediction of complications in critical care: a retrospective study, *Lancet Respiratory Medicine*, 6(12), pp. 905-914.

[14] NIAN, F. (2018). Viewpoints about the prognostic and health management, *Chinese Journal of Scientific Instrument*, 39(8), pp. 1-14.

[15] PALM, R. B. (2012). *Prediction as a Candidate for Learning Deep Hierarchical Models of Data.* Techni-

cal University of Denmark.

[16] SILVA, A. F. (2019). In-Depth Evaluation of Data Collected During a Continuous Pharmaceutical Manufacturing Process: A Multivariate Statistical Process Monitoring Approach, *Journal of pharmaceutical sciences*, 108(1), pp. 439-450.

[17] SUNDERMEYER, M. (2012), Published. LSTM Neural Networks for Language Modeling. Interspeech. 601-608.

[18] TôRRES, A. R. (2018). Multivariate statistical process control in annual pharmaceutical product review, *Journal of Process Control*, 69(9), pp. 97-102.

[19] YANG, L. (2018). Deep learning based weld and flange identification in pipeline, *Chinese Journal of Scientific Instrument* 39(2), pp. 193-202.

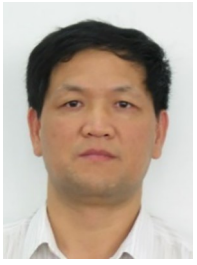
[20] YAO, J. (2019). Deep Learning From Noisy Image Labels With Quality Embedding, *IEEE Transactions on Image Processing*, 28(4), pp. 1909-1922.

[21] YU, H. (2017). Survey of compressed sensing technology for signal and data of power system, *Chinese Journal of Scientific Instrument*, 38(8), pp. 1943.

Author Biographies

XUE Wendong received his master´s degree in mechanical and electronic engineering and doctor´s degree in measuring and testing technologies from Xiamen University, China, in 2009 and 2012 respectively. He is now working as an Associate Professor at the Department of Mechanical and Electrical Engineering of Xiamen University. His research interests include Intelligent instrument and power electronic technology.

E-mail: xwd@xmu.edu.cn



HONG Yongqiang received his master´s degree in mechanical and electronic engineering from Jilin University of Technology. He is now working as a professor at the Department of Mechanical and Electrical Engineering of Xiamen University. His Research interests include intelligent control and industrial automation.

E-mail: hongyq@xmu.edu.cn