Research on Deep Learning Model of Multimodal Heterogeneous Data Based on LSTM

Li Dongping, Yang Yingchun*, Shen Shikai, He Jun, Shen Haoru, Yue Qiang, Hong Sunyan and Deng Fei

Abstract—**To solve the problems of multimodal heterogeneous data fusion and feature learning, a pattern recognition method based on long short-term memory (LSTM) is proposed to improve the classification accuracy. By fusing the deep learning models corresponding to different data types, a shared pattern recognition model is generated based on association analysis. Firstly, classification models for different data type are trained. Long-term memory ability of LSTM is used for the time-dependent data characteristics. Then the fusion strategy is analyzed. And an adaptive determination method for weight fusion is proposed. Finally, the algorithm flow is given, including data preprocessing, model training and heterogeneous data fusing. The experimental results show that the classification accuracy of the proposed method is higher than the models using the single data type alone.**

Index Terms—**deep learning, time series, text classification, big data, pattern recognition**

I. Introduction

In recent years, IT and communication technologies have developed rapidly, such as internet of things, cloud computing, triple play, etc. The rapid data growth has become a severe challenge and valuable opportunity faced by many industries. The information society has entered the era of big data. The feature learning technology should be effectively used for big data. The hidden laws should be discovered in big data. The potential value of big data should be mined. Big data is used to predict the future development. As a typical technology of feature learning, deep learning uses supervised and unsupervised strategies to automatically learn the multi-layer representation of data. It has been successfully applied in image recognition [1-2], fault diagnosis [3-4], time series prediction [5-6] and other fields.

Although deep learning has made some progress in data feature learning, it still faces many scientific challenges in feature learning for big data. One of which is the feature learning of multimodal heterogeneous data. Big data has a wide range of sources and types. Meanwhile it contains a variety of heterogeneous data. However, the existing deep learning models can only learn the characteristics of one single data type. It is difficult to capture the complex associations between heterogeneous data. It is hard to learn the characteristics of multimodal big data. To learn the characteristics of multimodal data, researchers propose a multimodal learning method by combining the effective information from various modes. Literature [7] proposed two multimodal recognition methods based on deep learning. Various types of data are used. Satellite and street images are used to measure urban unbalanced areas. Literature [8] proposed a multimodal deep learning model to predict the interaction between lncRNAs and protein subtypes. Literature [9] established a multi-mode fusion network for comparing objects outside the vehicle. The features of gaze, head posture and finger pointing are used to accurately predict the reference objects in different car postures.

Although the current improved methods can learn the characteristics of multimodal data, these methods do not make full use of the data time-dependent characteristics. They can not solve the problems of data fusion and feature learning for time-dependent sequence. As a learning model of time series classification, long short-term memory (LSTM) can solve the problems of gradient disappearance, gradient explosion of cyclic neural network, and the lack of long-term memory ability. LSTM can effectively use the information of long-term time series data. LSTM has many successful applications for time series data, such as text classification [10-14], speech recognition [15], time series prediction [16], etc.

Aiming at the problem of feature fusion, a multi-modal data classification method is proposed based on LSTM. The experimental results show the superior performance of the proposed model in data fusion and classification.

Manuscript received February 16th, 2022; revised August 28th, 2022. This work was supported in part by the General Project of Universities in Yunnan 201801CF00022, named “Research on Feature Learning Methods Based on Deep Computing in Big Data Environment” and the National Natural Science Foundation of China 62066023, named “Research on Automatic Annotation Technology of Yi Speech Data Based on Poverty Alleviation Log”.

Li Dongping is an associate professor of Institute of Information Engineering, Kunming University, Kunming, China. (e-mail: ynkm1971@sina.com).

Yang Yingchun, the corresponding author, is a senior engineer of Yunnan Branch of China Telecom Co., Ltd., Kunming, China. (e-mail: gddx89@21cn.com).

Shen Shikai is a professor of Institute of Information Engineering, Kunming University, Kunming, China. (e-mail: liamlychee@163.com).

He Jun is a professor of Institute of Information Engineering, Kunming University, Kunming, China. (e-mail: 2021kmxy@sina.com).

Shen Haoru is a lecturer of Institute of Information Engineering, Kunming University, Kunming, China. (lu20190816@126.com).

Yue Qiang is an associate professor of Institute of Information Engineering, Kunming University, Kunming, China. (e-mail: yueqiang2021@126.com).

Hong Sunyan is a lecturer of Institute of Information Engineering, Kunming University, Kunming, China. (e-mail: kmuhongsunyan@126.com).

Deng Fei is a senior engineer of Institute of Information Engineering, Kunming University, Kunming, China. (e-mail: feigeoffice@163.com).

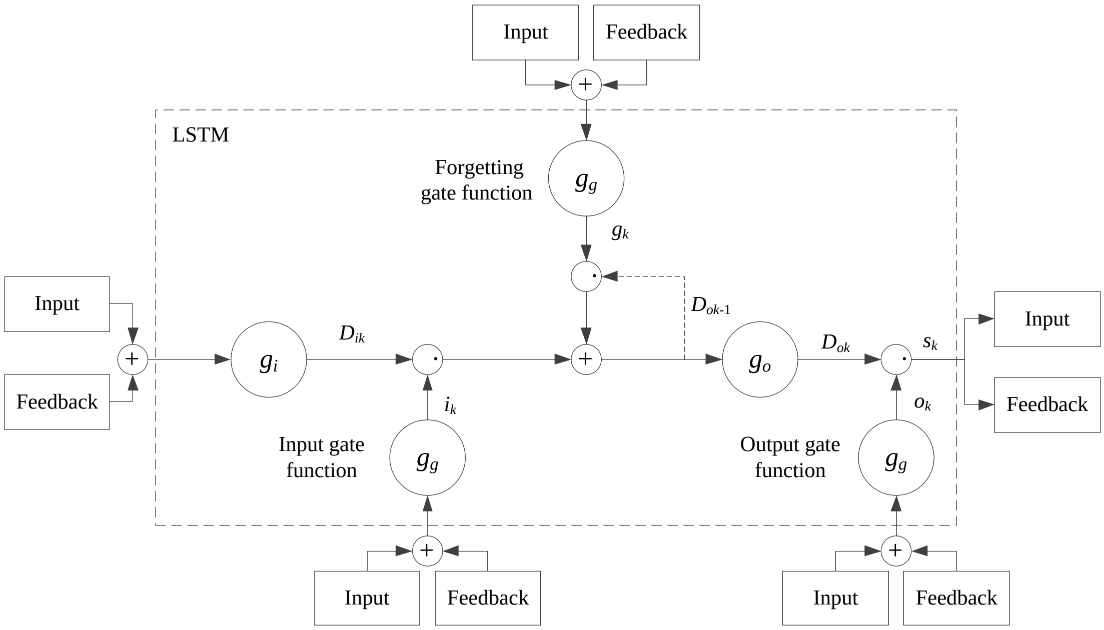


Fig. 1 Composition and structure of LSTM

II. LONG SHORT-TERM MEMORY (LSTM)

For the problem of time series related classification, LSTM memory unit can maintain and transmit the key information of time series data. LSTM has great advantages in time series classification [14].

The LSTM structure is shown in Fig. 1. LSTM includes three types of activation functions, namely $g_i()$、$g_g()$ and $g_o()$. $g_i()$ represents the input activation function. $g_g()$ represents the gate activation function. $g_o()$ represents the output activation function. Generally, sigmoid function is selected as the activation function for $g_i()$、$g_g()$ and $g_o()$.

The activation function output of $g_i()$、$g_g()$ and $g_o()$ can be calculated according to the following formula respectively.

$$i_k = g_g(\alpha_{x_i} x_k + \alpha_{s_i} s_{k-1} + \beta_i) \tag{1}$$

$$g_k = g_g(\alpha_{x_g} x_k + \alpha_{s_g} s_{k-1} + \beta_g) \tag{2}$$

$$o_k = g_g(\alpha_{x_o} x_k + \alpha_{s_o} s_{k-1} + \beta_o) \tag{3}$$

where x_k represents the model input at time k. s_k represents the hidden layer state at time k. i_k represents the output state of the input gate at time k. g_k represents the output state of the forgetting gate at time k. o_k represents the output state of the output gate at time k. α_{x_i} , α_{x_g} and α_{x_o} represent the input gate, forgetting gate and input weight of the output gate respectively. α_{s_i} , α_{s_g} and α_{s_o} represent the input gate, forgetting gate and feedback weight of the output gate respectively. β_i , β_g and β_o represent the offset of the input gate, forgetting gate and output gate respectively.

The output state D_{i_k} and D_{o_k} are corresponding to the input activation function $g_i()$ and the output activation function $g_o()$ at time k. D_{i_k} and D_{o_k} can be calculated according to the following formula.

$$D_{i_k} = g_i(\gamma_{x_i} x_k + \gamma_{s_i} s_{k-1} + \varphi_i) \tag{4}$$

$$D_{o_k} = g_k D_{o_{k-1}} + i_k D_{i_k} \tag{5}$$

where γ_{x_i} 、 γ_{s_i} and φ_i are the input weight, feedback weight and bias of the activation function respectively. Seen from Fig. 1, the output state is multiplied by the output state i_k at time k. Then the input gate is updated at time k. The update of LSTM unit is carried out with the kth time input and the k-1th feedback. The output state D_{o_k} is calculated according to equation (5). The update of the output state D_{o_k} of the hidden layer is calculated using equation (5).

$$s_k = o_k g(D_{o_k}) \tag{6}$$

LSTM uses the calculation process of its own input gate, forgetting gate and output gate function. Input activation function retains the key feature information in the model input. The output activation function transfers the feature information. It is especially suitable for data feature extraction and classification of time correlation. For model training with long-period sequence data, LSTM can extract the key feature information in the time series. And LSTM improves the model classification accuracy.

III. MULTIMODAL HETEROGENEOUS DATA FUSION

A. *Fusion strategy*

There are usually two ways for multimodal heterogeneous data fusion. One is the data feature fusion method. Heterogeneous data features are standardized and unified into a common form. Then the data features are fused according to the weight. And a classification model is used to classify the fused data features. The second is the classification result fusion method. Various types of data are used to independently train the classification model. The classification results are obtained. Then the classification results of different data models are aggregated to obtain the results. For the data feature fusion method, it is difficult to find the common feature forms of different data types. Taking heterogeneous data composed of text data and time series data for example, it is difficult to find the data features with common characteristics. For the classification result fusion method, the corresponding features can be extracted according to their respective data types. Such method is not limited by the requirements of common data features. It can effectively use the existing classification models for different data types. Therefore, the classification fusion method is used to learn the features of multimodal data. Considering the advantages for time-related data, LSTM is used as the feature learning model. LSTM extracts the feature vectors or original data of different data types respectively. Different data types are sent to the corresponding LSTM for model learning. The specific calculation process is as following.

$$y_i = f_i(x_i, \theta_i),\ i = 1, 2, \cdots, N \tag{7}$$

$$y = \sum_{i=1}^{N} w_i y_i \tag{8}$$

$$\tilde{y} = f_{\text{judge}}(y) \tag{9}$$

where N represents the number of data types. x_i represents the original data or extracted feature vector of the ith data type. f_i () represents the feature classification model of the ith data type. f_i () can be different deep learning model. The LSTM model is adopted for the feature learning of time-related data. θ_i is the parameter of f_i () model. w_i is the fusion weight coefficient of the ith data type. $f_{\text{judge}}()$ is the classification decision function. $\tilde{y}$ is the classification result of multimodal heterogeneous data.

B. Adaptive determination of fusion weight

For the fusion weight coefficient w_i in equation (8), the average assignment method can be adopted, i.e., $w_i = 1/N$. However, such method does not consider the training accuracy of different models. It can not reflect the learning results of different models. It affects the classification accuracy after fusion. Therefore, the training accuracy of different models is introduced into the w_i calculation process. The w_i calculation is carried out according to the following formula.

$$w_i = \frac{accuracy_i}{\sum_{i=1}^{N} accuracy_i} \tag{10}$$

where $accuracy_i$ is the training accuracy of model f_i ().

IV. Multimodal Heterogeneous Data Deep Learning Algorithm Flow

The algorithm flow of multimodal heterogeneous data deep learning model is proposed, which mainly includes three stages, namely data preprocessing, model training and heterogeneous data fusion. The flow chart is shown in Fig. 2.

A. Data preprocessing

For N different types of input data set $\{x_1, x_2, \ldots, x_N\}$, data preprocessing is used for data standardization. For example, text data should be converted into a numerical vector. The common word vector processing model is word2vec. Image should be converted into two-dimensional or three-dimensional numerical information. Normalization processing is also required to avoid large differences for data value. Normalization processing may affect the model convergence speed and classification accuracy.

B. Model training

$x_i = \{r_i^j, t_i^j\}$ is the preprocessed data set x_i of the ith type, $j = 1, 2, \cdots, M$. r_i^j represents the model input vector of LSTM$_i$. t_i^j represents the corresponding classification label of r_i^j. M represents the sample number. The training objective function can be expressed as following.

$$p(t_i \mid r_i) = \arg\max_{\theta_i} f(t_i \mid r_i; \theta_i),\ i = 1, 2, \cdots, N \tag{11}$$

where θ_i represents the model parameter of LSTM$_i$. LSTM$_i$ is trained by the ith data type set. Using N different types of data, N LSTM$_i$ models are trained respectively. The training accuracy acc$_i$ of N LSTM$_i$ models are recorded.

C. Heterogeneous data fusion

(1) Classification probability calculation of N LSTM$_i$ model

The samples $\{\bar{r}_i, \bar{t}_i\}$ of the ith type are substituted into the trained LSTM$_i$ model. $\{\bar{r}_i, \bar{t}_i\}$ are used to calculate the prediction probability p_i of different classifications. The calculation process can be expressed as

$$\bar{p}_i = LSTM_i(\bar{r}, \theta_i) \tag{12}$$

where $\bar{p}_i = [p_{\bar{t}_i=1}, p_{\bar{t}_i=2}, \cdots, p_{\bar{t}_i=q}]$, $i = 1, 2, \cdots, N$. q is the number of category labels.

(2) Fusion weight calculation

Replace the training accuracy acc$_i$ of N LSTM$_i$ models into equation (10). Calculate the fusion weight w_i, $i = 1, 2, \cdots, N$.

(3) Classification probability fusion

Aggregate the classification prediction probability of N LSTM$_i$ models. The calculation formula is as follows.

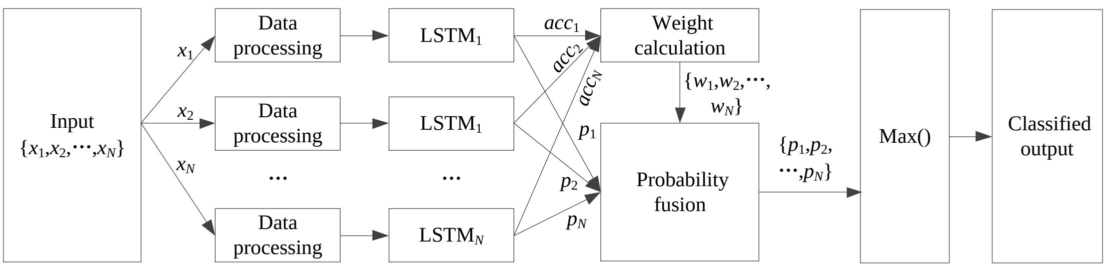


Fig. 2 Algorithm flow chart

$$P=\sum_{i=1}^{N} w_i p_i=\sum_{i=1}^{N}\left(\frac{acc_i}{\sum_{i=1}^{N} acc_i}\overline{p}_i\right) \tag{13}$$

where $P=[p_1, p_2, \cdots, p_q]$. p_j represents the probability that the classification result belongs to the jth label. Substitute P into the max () function to find the label corresponding to the maximum probability. The output is the fused classification result. The calculation formula can be expressed as following.

$$output=\max(p_1, p_2, \cdots, p_q) \tag{14}$$

V. EXPERIMENTAL ANALYSIS

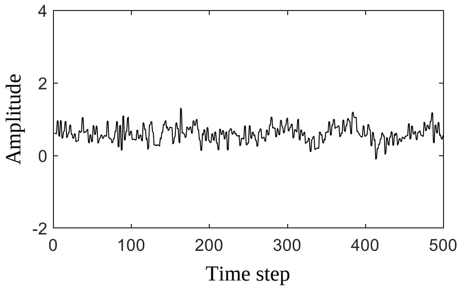
Combined with an industrial case in the field of fault diagnosis, the proposed method is performed for experimental verification.

A. Experimental data

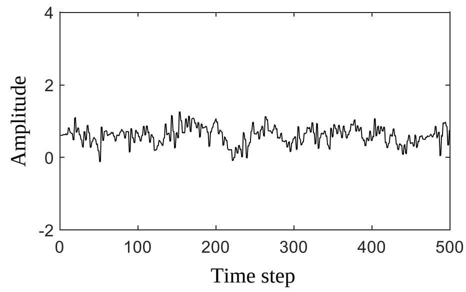
The experimental data set comes from the fault records of a factory. The data set includes four fault types. Each fault data includes two types of data, namely text sequence with the description of factory events and time series data. Each text sequence describes different faults in the form of text. Table I shows some examples of text sequence sample. Each time series data contains 52 variables, which are used to characterize different fault types. 52 variables are collected every 3 minutes during 25 hours. And each variable consists of 500 points with the same time interval. Fig.3 and Fig.4 show the time series examples corresponding to variable 1 and variable 2 under different fault types.

TABLE I
EXAMPLE OF TEXT SEQUENCE IN FAULT SAMPLE

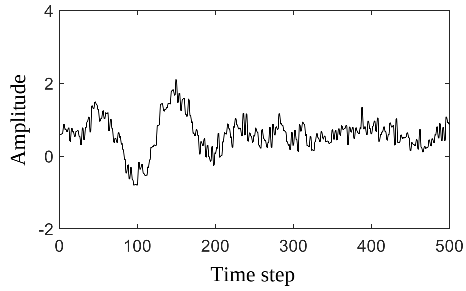
Num	Fault type	Text descriptions
1	Type 1	High pitch coil whine heard in the mixer.
2	Type 1	Fuse to the mixer assembly tripped.
3	Type 2	A few dribbles of fluid are showing up underneath the constructing agent.
4	Type 2	Some of the output accumulates liquid in the bottom of the assembler.
5	Type 3	The products occasionally leave the scanner cracked.
6	Type 3	Breaks appear in the scanner roller.
7	Type 4	Mixing software keeps freezing.
8	Type 4	Scanner program is frozen.



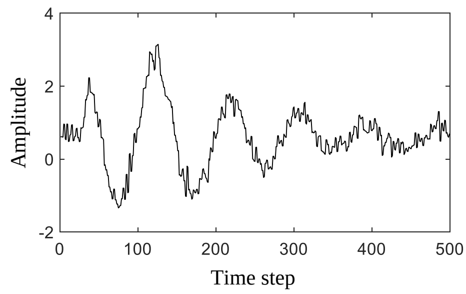
(a) Fault type 1



(b) Fault type 2

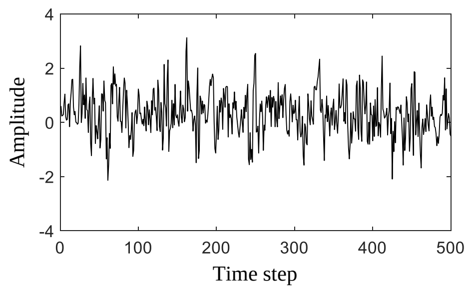


(c) Fault type 3

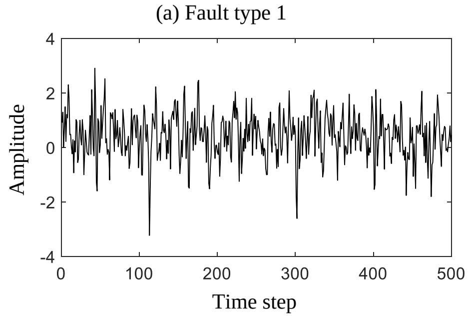


(d) Fault type 4

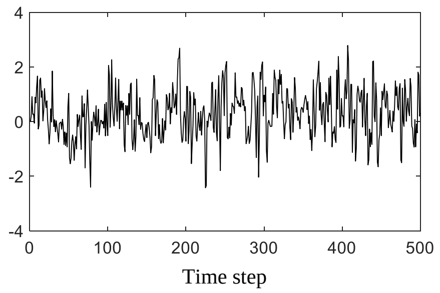
Fig. 3 Example of variable 1 in time series samples



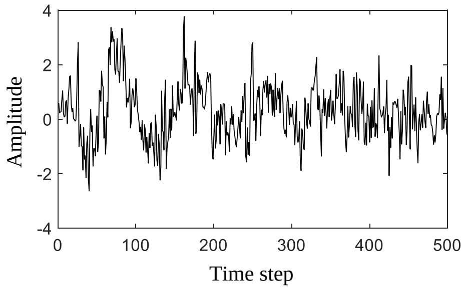
(a) Fault type 1



(b) Fault type 2



(c) Fault type 3



(d) Fault type 4

Fig.4 Example of variable 2 in time series samples

The fault data contains 2000 samples in total. Each fault type corresponds to 500 samples. 80% of the fault samples are used for model training. The remaining 20% are used to test model performance. The number of training and testing samples is shown in Table II.

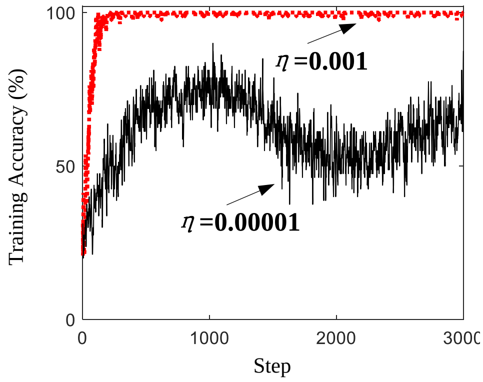
TABLE II
NUMBER OF TRAINING AND TESTING SAMPLES

Sample type	Fault type 1	Fault type 2	Fault type 3	Fault type 4	Total
Training sample	400	400	400	400	1600
Testing sample	100	100	100	100	400

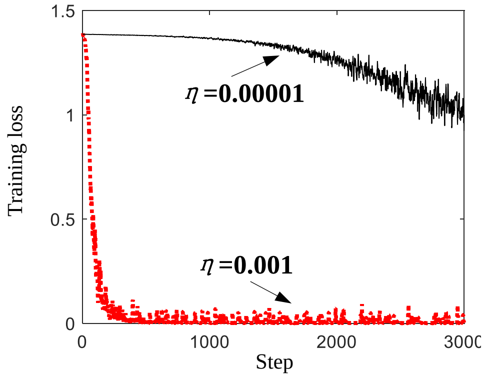
B. *Model training*

For the standardized text sequence and time series data, LSTM classification models are established respectively. The text classification model structure includes one input layer, one word embedding layer, one LSTM hidden layer (composed of 80 hidden units), four full connection layers, one softmax layer and one classified output layer. Word embedding layer converts text characters into numerical indexes. The structure of time series classification model includes one input layer, three LSTM hidden layers (composed of 52, 40, 25 hidden units respectively), four full connection layers, one softmax layer and one classification output layer.

Under different learning rates（η=0.001、η=0.00001), Fig.5 shows the training accuracy and loss of text classification model. When η=0.001, the training accuracy is large, while the training loss is small. The training accuracy is close to 100%, while the training loss is about 0.01. When η=0.00001, the training accuracy is about 80%, while the training loss is about 1.



(a) Training accuracy



(b) Training loss

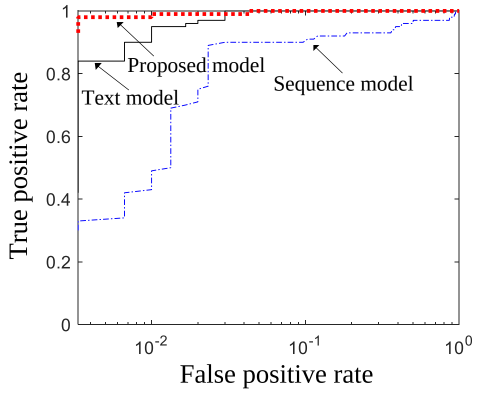
Fig. 5 Training accuracy and loss curve under different learning rates

C. *Comparison of experimental results*

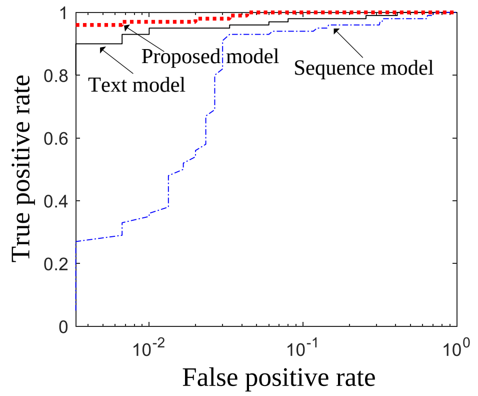
To verify the advantages of the proposed method, the text model and sequence model are used as the comparison methods. The text model is established only by text, while the sequence model is established only by time series. The two methods also use LSTM as the classification model.

Firstly, the receiver operating characteristic (ROC) curves of the three methods are compared. The area under curve (AUC) is defined as the area surrounded by the coordinate axis under the ROC curve. And the area value is not greater than 1. The AUC is equivalent to the probability of randomly sorting positive samples before negative samples. If the area of ROC is close to 1, the superior performance of the classifier is obtained.

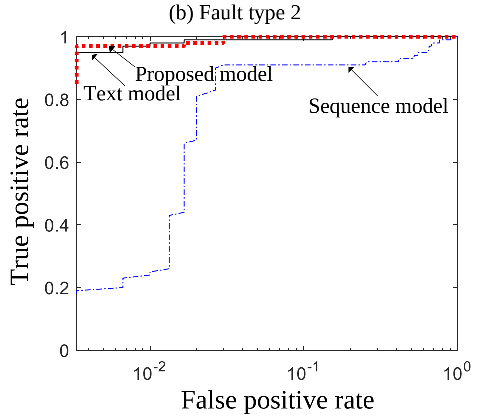
The ROC curves of the three methods are shown in Fig.6. The corresponding AUC values are shown in Table III. Seen from Fig.6 and Table III, the AUC difference of the proposed method is small under different fault categories. Seen from Table III, the AUC of the proposed method has reached more than 0.998. The AUC of the text model is slightly small, which reaches more than 0.99. The AUC of the sequence model is the smallest. The lowest value is only 0.9339 for fault type 3. It shows the superior classification performance of the proposed method.



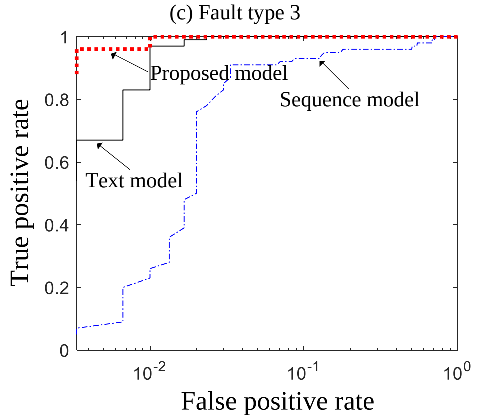
(a) Fault type 1



(b) Fault type 2



(c) Fault type 3



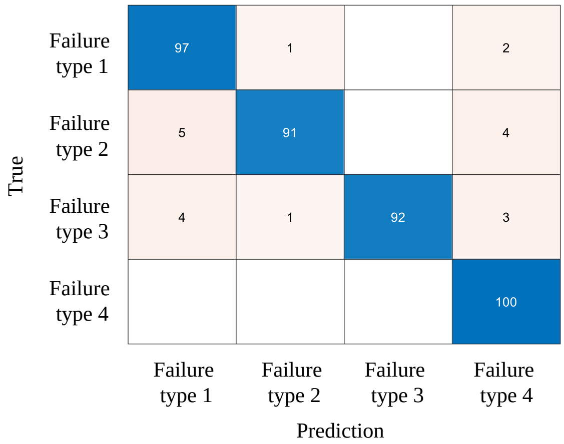
(d) Fault type 4

Fig.6 ROC curves of different methods

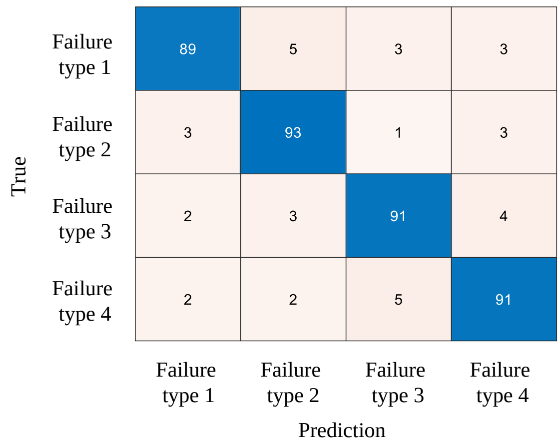
TABLE III
AUC OF DIFFERENT METHODS

	Sequence model	Text model	Proposed model
AUC of fault type 1	0.9432	0.9964	0.9993
AUC of fault type 2	0.9626	0.9909	0.9990
AUC of fault type 3	0.9339	0.9980	0.9988
AUC of fault type 4	0.9550	0.9965	0.9993

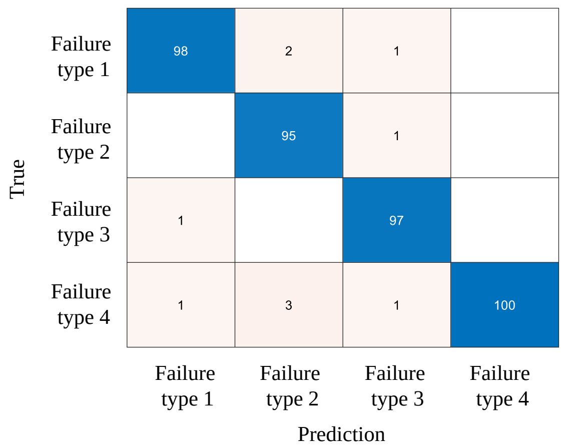
The classification confusion matrix of the three methods is shown in Fig.7. For the sequence model, the recognition accuracy of fault type 2 and fault type 3 is low. For the text model, the recognition accuracy of the four fault types is low. Because the proposed method integrates the fault information of time series data and text data, the classification accuracy of the four fault types has been improved. The experimental result proves the advantages of the proposed method in the classification of heterogeneous data. The classification accuracy of the three methods is shown in Table IV. The classification accuracy of the proposed method is the highest. The accuracy of the text model is the lowest.



(a) Sequence model



s
(b) Text model



(c) Proposed model

Fig. 7 Classification confusion matrix of different methods

TABLE IV
CLASSIFICATION ACCURACY OF DIFFERERNT METHODS

	Sequence model	Text model	Proposed model
Accuracy	95.00%	91.00%	97.50%

VI. Conclusion

A multimodal heterogeneous data classification method based on LSTM is proposed in this paper. The proposed method can make full use of the characteristic information of different types of data. The information of different data types is effectively fused. Then the classification accuracy is improved. The proposed classification method for multimodal heterogeneous data is verified by experiments. The proposed method expands the application scope of deep learning technology. It verifies the effectiveness of text data and temporal data fusion. In the next step, the key features should be extracted from image, voice and other data types.

References

[1] Toldinas J, Venkauskas A, and Damaeviius R, "A Novel Approach for Network Intrusion Detection Using Multistage Deep Learning Image Recognition, " *Electronics*, vol. 10, no. 15, pp1854-1855, 2021.

[2] Li S, Du Z, and Meng X, "Multi-stage Malaria Parasite Recognition by Deep Learning, " *GigaScience*, vol. 6, pp6-7, 2021.

[3] Zhang D, Li C, and Shahidehpour M, "A Bi-level Machine Learning Method for Fault Diagnosis of Oil-immersed Transformers with Feature Explain ability, " *International Journal of Electrical Power & Energy Systems*, vol. 134, pp107356, 2022.

[4] Jiang W, Wang C, and Zou J, "Application of Deep Learning in Fault Diagnosis of Rotating Machinery, " *Processes*, vol. 9, no. 6, pp919, 2021.

[5] Jie C A, Gqz B, and Wz A, "Wind Speed Forecasting Using Nonlinear-learning Ensemble of Deep Learning Time Series Prediction and Extremal Optimization – ScienceDirect," *Energy Conversion and Management*, vol. 165, pp681-695, 2018.

[6] Kumar R, Kumar P, and Kumar Y, "Analysis of Financial Time Series Forecasting using Deep Learning Model," *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2021.

[7] Suel E, Bhatt S, and Brauer M, "Multimodal Deep Learning from Satellite and Street-level Imagery for Measuring Income, Overcrowding, and Environmental Deprivation in Urban Areas," *Remote Sensing of Environment*, vol. 257, pp112339, 2021.

[8] Shaw D, Chen H, and Xie M, "DeepLPI: A Multimodal Deep Learning Method for Predicting the Interactions between LncRNAs and Protein Isoforms," *BMC Bioinformatics*, vol. 22, no. 1, 2021.

[9] Aftab A R, Michael V, and Rohrhirsch S, "Multimodal Fusion Using Deep Learning Applied to Driver's Referencing of Outside-Vehicle Objects, " *Proceedings of 2021 IEEE Intelligent Vehicles Symposium*, pp1-9, 2021.

[10] Zhou C, Sun C, and Liu Z, "A C-LSTM Neural Network for Text Classification, " *Computer Science*, vol. 1, no. 4, pp39-44, 2015.

[11] Zhang Y, "Research on Text Classification Method based on LSTM Neural Network Model," *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 2021.

[12] Yinyin Hu, Xiaoxia Zhang, Jiao Yang, and Shuai Fu, "A Hybrid Convolutional Neural Network Model based on Different Evolution for Medical Image Classification," *Engineering Letters,* vol. 30, no.1, pp168-177, 2022.

[13] Andi Nugroho, Harco Leslie Hendric Spits Warnars, Ford Lumban Gaol, and Tokuro Matsuo, "Trend of Stunting Weight for Infants and Toddlers using Decision Tree," *IAENG International Journal of Applied Mathematics*, vol. 52, no.1, pp144-148, 2022

[14] Prasad B R, "Classification of Analyzed Text in Speech Recognition Using RNN-LSTM in Comparison with Convolutional Neural Network to Improve Precision for Identification of Keywords," *Revista Gestão Inovação e Tecnologias*, vol. 11, no. 2, pp1097-1108, 2021.

[15] Karevan Z, and Suykens J, "Transductive LSTM for Time-series Prediction: An Application to Weather Forecasting," *Neural Networks*, vol. 125, pp1-9, 2020.

[16] Li Y J, Huang J J, and Wang H Y, "Study of Emotion Recognition based on Fusion Multi-modal Bio-signal with SAE and LSTM Recurrent Neural Network," *Journal on Communications*, vol. 38, no. 12, pp109-119, 2017. (in Chinese)

Li Dongping is an associate professor in the Institute of Information Engineering of Kunming University. Her major is computer science and technology engineering. Her research direction includes big data technology and data mining.

Yang Yingchun is a senior engineer in Yunnan Branch of China Telecom Co., Ltd. His major is computer science and technology. His interest is big data analysis and algorithm research.

Shen Shikai is a professor in the Institute of Information Engineering of Kunming University. His research direction includes mobile communication network performance analysis, sensor network node location, data fusion theory and technology research.

He Jun is a professor in the Institute of Information Engineering of Kunming University. His major is computer software engineering. His research direction includes machine learning, data analysis and software evolution.

Shen Haoru is a lecturer in the Institute of Information Engineering of Kunming University. His major is big data research of computer networks and education.

Yue Qiang is an associate professor in the Institute of Information Engineering of Kunming University. His major is software engineering. His research direction includes big data technology and data mining.

Hong Sunyan is a lecturer in the Institute of Information Engineering of Kunming University. Her research direction includes machine learning algorithm.

Deng Fei is a senior engineer in the Institute of Information Engineering of Kunming University. His research direction includes artificial neural network and data mining.