International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 3866-3875

https://doi.org/10.70517/ijhsa463328

Research on multimodal data analysis and fault prediction of speech text in power industry based on LSTM network

Haitao Yu[1], Xuqiang Wang[2,*], Jian Zheng[2], Tianyi Liu[1] and Yongdi Bao[1]
[1] State Grid TJ Information & Telecommunication Co., Ltd., Tianjin, 300140, China
[2] State Grid Tianjin Electric Power Company, Tianjin, 300010, China
Corresponding authors: (e-mail: mypaper2021@126.com).

Abstract China is building a new type of power system mainly based on new energy. The large-scale access of new energy, the complex structure of AC-DC hybrid grid, and the application of power electronic devices make the grid faults show more complex modes. Fast and accurate diagnosis of grid faults is a necessary condition to guarantee the reliable operation of power grid. In this paper, based on the theory of equipment operation state diagnosis algorithm of multimodal data fusion analysis of power system, we add the bidirectional long and short-term memory network with attention mechanism, respectively, through Word2Vec and Fast Fourier Transform, extract text and audio data features of power industry, and fuse the extracted multimodal data features. The XGBoost decision tree algorithm is used to achieve the training objectives such as data prediction and pattern recognition. From the extracted time-domain plots, it can be seen that the amplitude of the windings fluctuates between -0.9~0.6m·s^{-2} before loosening, and the vibration signals after loosening are between -1~1 m·s^{-2}, with obvious amplitude variations when a fault occurs. The error of the prediction curve when a fault occurs suddenly becomes larger, and the error values of the two are 60.491 and 45.469, respectively, and the prediction model proposed in this paper has a high monitoring accuracy for normal operation of the power system.

Index Terms Multimodal data fusion, Long and short-term memory network, Word2Vec, Fast Fourier transform, XGBoost

I. Introduction

Driven by the wave of new generation artificial intelligence, the State Grid Corporation has put forward the development strategy of “three types and two grids”, and is committed to building a highly intelligent, automated and informationized smart grid [1], [2]. With the continuous development of the power system, the diagnosis of equipment operation state has become one of the key technologies to ensure the safe and stable operation of the power system [3]. The traditional diagnostic methods for power system equipment operating status mainly rely on single-modal data, which have problems such as limited diagnostic accuracy, lack of robustness and reliability, and insufficient environmental adaptability due to factors such as incomplete information, susceptibility to noise, and complex and variable working environment [4]-[6]. For this reason, a power system multimodal data fusion analysis method is investigated, aiming at realizing the accurate diagnosis of equipment operation status [7]. Power systems are characterized by extensive, diverse and rich data sources, data types and data structures, including unstructured data such as images, videos, audios and texts from various devices [8]-[10]. By fully exploring and utilizing the valuable features embedded in these multimodal data, the fault prediction capability and intelligence level of the power system can be significantly improved, providing real-time auxiliary decision support for the development of the power industry [11]-[13].

Collecting the rich multimodal data of the power system, maximizing the extraction of effective features of each modality while maintaining the diversity and integrity of the information, and minimizing the loss of information that may be generated in the process of processing, and effectively fusing them is one of the keys to playing the role of its core production elements. Alqudah, M. et al. constructed an automatic power system fault prediction method based on multimodal data analysis by combining the existing power measurement data and other related database data, which can realize the predisposition prediction of the occurrence of power system faults by jointly learning and representing the data from multiple sources [14]. Xing, Z. and He, Y. proposed that fast and effective analysis of time-series multimodal data is superior and efficient in real power transformer fault diagnosis by addressing the multimodal heterogeneity of power data and the missing sample proportion problem in the process of power transformer fault diagnosis [15]. Afrasiabi, S. et al. investigated a multimodal fault identification methodology for solar photovoltaic (PV) power generation systems by incorporating residual convolutional neural networks and

gated recurrent units so as to improve the robustness and accuracy of the diagnostic system and to effectively deal with a wide range of faults of the power system including grid connections [16]. Alsaif, K. M. et al. constructed a fault detection and diagnosis framework based on a multimodal large language model, whose dynamic, accurate and context-aware fault detection capabilities, as well as the diagnostic knowledge base synthesized through the model, effectively improve the accuracy of this diagnostic framework for fault detection and diagnosis in unbalanced scenarios [17]. Ke, L. et al. proposed a new multimodal attention fusion (MAF) model that significantly improves the comprehensive fault diagnosis performance of power systems by fusing different modal fault prediction results extracted from power system datasets using a joint fault feature extraction method [18].

In this paper, we construct a multimodal model of bi-directional long and short-term memory network with added attention mechanism, and adopt feature-level multimodal feature fusion method, which is sequentially realized by Word2Vec's text data feature extraction, and Fast Fourier Transform for faulty audio feature extraction. The model goes through forgetting, selective memory and output stages to complete the selective encoding of historical and current information. The XGBoost decision tree model is used, with the power industry fault data prediction and pattern recognition, etc. as the training objectives, and the final output is obtained through the combination of multiple decision trees, and the recurrent neural network of LSTM is applied to decode the encoded input information. The stochastic gradient descent algorithm is used for model training, and the model is applied to the actual fault prediction of power equipment.

II. Multimodal data fusion in power industry based on LSTM network

II. A. Algorithmic principles

In this section, the theory of equipment operation state diagnosis algorithm based on multimodal data fusion analysis of power system is given. Specifically, the bi-directional long short-term memory network (LSTM) multimodal model BC-AT-LSTM (where BC denotes bi-directional and AT denotes attention) with the addition of an attention mechanism is firstly introduced, as well as the feature-level multimodal feature fusion method employed [19]. A Word2Vec-based feature extraction method for text data and a spectrogram-based audio feature extraction method [20] were applied in turn.

II. A. 1) Multimodal data fusion based on BC-AT-LSTM modeling

Figure 1 shows the multimodal feature fusion, and the methods of multimodal feature fusion are divided into four types: feature-level fusion, decision-level fusion, hybrid-level fusion, and model-level fusion [21]. Among them, feature-level fusion and decision-level fusion are the two commonly used methods. Feature-level fusion belongs to tight coupling, which is a method of extracting features from different modalities and connecting them into a single high-dimensional feature vector, which can eliminate redundant information, but cannot model complex relationships. Decision-level fusion is loosely coupled, which integrates the predictions based on each modality by applying algebraic combination rules of multiple prediction class labels after obtaining the predictions based on each modality, and this method is unable to capture the interconnections between different modalities. Since this paper utilizes multimodal data fusion analysis techniques to achieve power equipment operation state diagnosis, and the relationships among modalities are not complex and interrelated, the feature-level fusion method is directly adopted to map the feature vectors of each modality into a common subspace, which is later fed into the multimodal model for training.

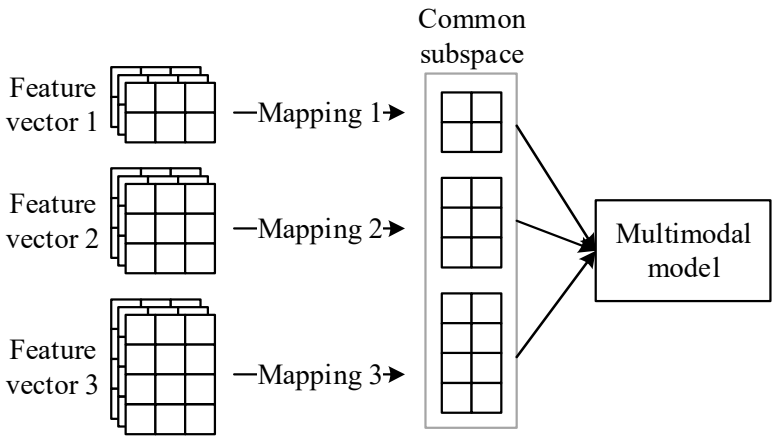


Figure 1: Multimodal feature fusion

II. A. 2) Word2Vec-based feature extraction for text data

Word2Vec is a model for generating word vectors, at the core of which is a shallow and two-layered neural network structure designed to be trained by reconstructing the text of linguistic words, mapping each word to a high-dimensional vector that expresses the relationship between words and in fact reflects the state of the hidden layer of the neural network.

The formula for its softmax layer is:

$$softmax(v_c^T, v_w) = \frac{\exp(v_c^T v_w)}{\sum_{i=1}^{V} \exp(v_c^T v_i)} \tag{1}$$

where: v_c denotes the vector of context words, v_w denotes the vector of target words, and V denotes the total number of words in the vocabulary. The meaning of equation (1) is to calculate the probability that the target word is a certain word given the context, and train the vector corresponding to each word by maximizing the sum of the conditional probabilities of all target words.

To train the model and obtain word vectors using Word2Vec, a large corpus is required. The specific training steps are as follows: first, the vocabulary is encoded as word vectors in one-hot form and input into a one-layer neural network. The number of neuron nodes in the input layer should match the dimension of the one-hot word vector. Then, the association probability of the target word with other words is calculated by the mapping layer and its activation function in the neural network. In this process, negative sampling technique is used to improve the training speed and correctness. The loss is then calculated using Stochastic Gradient Descent (SGD) optimization algorithm to evaluate the accuracy of model prediction. Finally, the weights and biases of the neurons are updated by a back-propagation algorithm to optimize the performance of the model. This process enables Word2Vec to learn and generate word vectors that express relationships between words based on a large corpus.

II. A. 3) Spectrogram-based audio feature extraction

Audio features can be categorized in several dimensions, including the source of the features and the time range. From the source of features, time-domain features (e.g., duration, mean, variance, energy, etc.) and frequency-domain features (e.g., spectrograms, power spectral density maps, spectral center of mass, etc.) can be computed directly from the audio signal. The Fast Fourier Transform (FFT) converts the time domain signal to frequency domain signal and is calculated as [22]:

$$X_k = \sum_{n=0}^{N-1} (x_n \cdot e^{-j \cdot 2\pi \cdot k \cdot n/N}) \tag{2}$$

where: X_k denotes the result of the discrete Fourier transform, which represents the magnitude of the k th frequency component of the signal in the frequency domain, and it is a complex number containing magnitude and phase information. x_n denotes the input signal, a sequence of complex numbers of length N, which represents the magnitude and phase of the signal at different points in time n in the time domain, and $e^{-j \cdot 2\pi \cdot k \cdot n/N}$ denotes the complex rotating factor, where j is the imaginary unit, and the key role of the rotating factor is to be responsible for converting the each sample point in the time domain signal x_n to a different frequency component in the frequency domain.

For acoustic spectral features, such as Meier spectral coefficients (MFCC), power spectral density (PSD), etc., the audio signal needs to be processed by short-time Fourier transform (STFT), etc. before extraction. In terms of time scale, features can represent transient or global information of the audio signal. Transient features are usually measured in frames and can capture local changes in the audio signal, while global features cover a longer time dimension and are used to reflect the overall characteristics of the audio signal.

When extracting audio features, it is necessary to choose the appropriate feature extraction method according to the specific application scenario. For example, in the audio event detection task, it may be necessary to focus on the local changes of the audio signal and the occurrence of specific events, so transient features and underlying features may be more applicable. Audio event detection is an important branch of audio surveillance systems that aims to identify the occurrence of specific events from audio signals. Due to the complexity and diversity of audio signals, there are some technical difficulties in audio event detection, among which, how to effectively extract audio features to distinguish different events is one of the key issues.

II. B. LSTM-based multimodal data extraction and fusion

Long Short-Term Memory Networks (LSTM) are good for extracting semantic dependencies in long range.Structure of LSTM recurrent module.

The LSTM recurrent module conveys information about the neural unit state C_t with the hidden layer information h_t and realizes the selective encoding of historical and current information through the following three stages.

II. B. 1) Oblivion stage

The forgetting stage determines how the input neural unit state information C_{t-1} from the previous moment is retained, and is calculated as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3}$$

where W_f, U_f and b_f are the weights and bias of the linear relationship in the forgetting stage, f_t is the output of the forgetting stage at moment t, x_t is the input information at moment t, and h_{t-1} is the information of the hidden layer at moment $(t-1)$. It contains the history information from the initial moment to the $(t-1)$ moment, and $\sigma(\cdot)$ is the Sigmoid activation function, which can map the input value to the interval range of [0, 1]. When mapping to 0, it means discarding all historical information, and when mapping to 1, it means retaining all historical information.

II. B. 2) Selective Memory Phase

This stage determines the retained portion of the input information x_t at moment t, as expressed in equation (4):

$$\begin{cases} i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ C_t^{'} = \tanh(W_c x_t + U_c h_{t-1} + b_c) \end{cases} \tag{4}$$

where W_i, W_c, U_i, U_c, b_i and b_c are the weights and their bias of the linear relationship in the selection memory stage, i_t is the retention ratio, $C_t^{'}$ is the input x_t is the retained information content, and $\tanh(x)$ is the nonlinear activation function with the following expression:

$$\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{5}$$

Then the state information C_t of the neural unit at the moment of t is:

$$C_t = f_t C_{t-1} + i_t C_t^{'} \tag{6}$$

II. B. 3) Output phase

The task of the output stage is to select some or all of the information as the output information of the hidden layer with the following expression:

$$\begin{cases} o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ h_t = o_t \tanh(C_t) \end{cases} \tag{7}$$

where W_o, U_o and b_o are the weights and biases of the linear relationship in the forgetting phase, respectively.

II. B. 4) XGBoost-based multimodal fusion algorithm

XGBoost is an integrated algorithm based on decision tree with the advantages of good fitting and fast training [23]. The decision tree algorithm continuously divides the data set by data attribute metrics to generate a classification tree. The algorithm decides the optimal classification attributes by maximizing the information entropy gain. The information entropy is a measure of the randomness of a variable, and the greater the information entropy, the greater the randomness of the variable. The information entropy for a random variable S is calculated as follows:

$$\begin{aligned} Ent(S) &= [-p(s_1)\log_2(p(s_1))] + \cdots \\ &+ [-p(s_n)\log_2(p(s_n))] \\ &= -\sum_{i=1}^{n} [p(s_i)\log_2(p(s_i))] \end{aligned} \tag{8}$$

where $s_1, s_2, \cdots, s_n$ are the n possible values of the random variable S, and $p(s_1), p(s_2), \cdots, p(s_n)$ are the probabilities of $s_1, s_2, \cdots, s_n$ probabilities.

When the random variable S is affected by some influence factor H, then the conditional information entropy of the random variable S under the influence of H is:

$$Ent(S \mid H) = \sum_{j=1}^{m} p(h_j) Ent(S \mid h_j) = -\sum_{j=1}^{m} \sum_{i=1}^{n} p(s_i, h_j) \log_2 p(s_i, h_j) \tag{9}$$

where $p(h_j)$ is the probability that h_j occurs, and $p(s_i, h_j)$ is the probability that s_i and h_j occur simultaneously.

When the dataset X is partitioned according to the attribute A, which results in a possible partition into the class Z_i, the conditional entropy of the dataset is:

$$Ent(X \mid A) = -\sum_{j=1}^{s} p(a_j) \sum_{i=1}^{h} p(Z_i \mid a_j) \log_2 p(Z_i \mid a_j) \tag{10}$$

where a_j is the possible values of attribute A.

Then the information entropy gain $G(X \mid A)$ after node branching according to attribute A is:

$$G(X \mid A) = -Ent(X) - Ent(X \mid A) \tag{11}$$

The structure of the XGBoost algorithm is shown in Figure 2. It consists of multiple independent decision trees, each of which classifies features and fits residuals to a randomly selected subset of the original training set, thus realizing the training objectives of data prediction and pattern recognition. Different decision trees focus on different features of the original data, and multiple decision trees can be computed concurrently, so they have obvious advantages in fitting effect and computational speed. Finally, the final output is obtained by combining multiple decision trees.

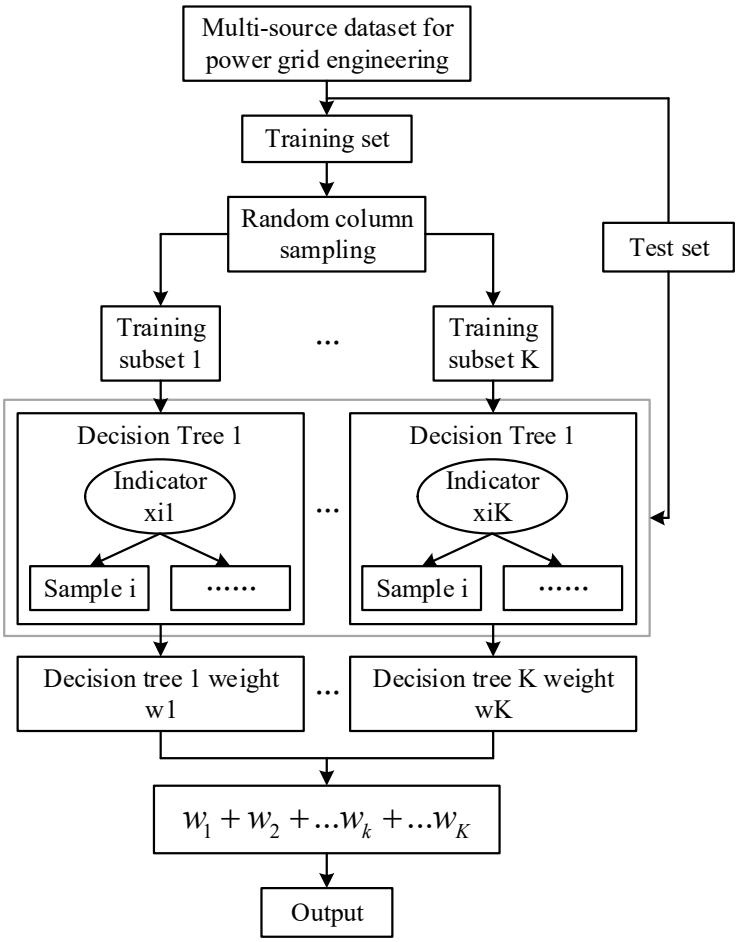


Figure 2: Multimodal fusion algorithm based on XGBoost

II. C. Power system fault prediction

In order to predict power system faults, LSTM-based recurrent neural networks are used to decode the encoded input information. However, as the length of the input sequence increases, the performance of the encoder and decoder degrades. To solve this problem, a temporal attention mechanism is used in the decoder to adaptively select the encoder hidden state associated with the current decoding time step. Specifically, based on the previous decoder hidden states and LSTM units, the attention weights for the i th time step are computed, and then these weights are applied to the encoder hidden states, and finally this is used to compute the weights of each encoder hidden state at the i th moment:

$$l_j^t = v_d^T \tanh(W_d[d_{i-1}; s_{i-1}^{'}] + U_d h_j), (1 \le j \le T) \tag{12}$$

$$\beta_i^j = \frac{\exp(l_i^j)}{\sum_{k=1}^{T} \exp(l_i^k)} \tag{13}$$

where $[d_{i-1}; \tilde{y}_{i-1}] \in R^{p+1}$ represents the neuronal connectivity, which is required to adaptively select the relevant encoder hidden states using a temporal attention mechanism during decoding to cope with long sequences of inputs. The attention weight β_i^j represents the importance of the j th hidden state for prediction, and since each encoder hidden state h_j is mapped to the temporal component of the input, in the computation of the context vectors c_i, all the encoder hidden states $\{h_1, h_2, ..., h_l\}$ are taken into account:

$$c_i = \sum_{j=1}^{l} \beta_i^j h_j \tag{14}$$

The context vector c_i will be updated at each time step. The c_i will be combined with the target sequence $(y_1, y_2, ..., y_{l-1})$, and this combining operation can be done with a simple feed-forward neural network that receives as input the context vector and the previously generated target sequence and outputs the predicted value of the next target sequence y_i:

$$\tilde{y}_{i-1} = \tilde{w}^T [y_{i-1}, c_{i-1}] + \tilde{b} \tag{15}$$

where $[y_i - 1, c_{i-1}] \in R^{m+1}$ is the splice of the decoder input $\tilde{y}_{i-1}$ and the vector c_{i-1}. The parameters $\tilde{w} \in R^m + 1$ and $\tilde{b} \in R$ map the splice to the size of the decoder input. The newly computed $\tilde{y}_{i-1}$ can be used to update the hidden state of the decoder at the i th moment. The nonlinear function f_2 is chosen as the unit. This function is widely used for long-term dependencies. d_i can be updated as:

$$d_i = f_2(d_{i-1}, \tilde{y}_{i-1}) \tag{16}$$

$$f_t^{'} = \sigma(W_{f'}[d_{i-1}; \tilde{y}_{i-1}] + b_f^{'}) \tag{17}$$

$$K_i^{'} = \sigma(W_j^{'}[d_{i-1}; \tilde{y}_{i-1}] + b_j^{'}) \tag{18}$$

$$T_i^{'} = \sigma(W_o^{'}[d_{i-1}; \tilde{y}_{i-1}] + b_o^{'}) \tag{19}$$

$$s_i^{'} = f_i^{'} \square\ s_{i-1}^{'} + K_i^{'} \square\ \tanh(W_s^{'}[d_{i-1}; \tilde{y}_{i-1}] + b_s^{'}) \tag{20}$$

$$d_i = T_i^{'} \square\ \tanh(s_i^{'}) \tag{21}$$

where $[d_{i-1}; \tilde{y}_{i-1}] \in R^{p+1}$ is a splice of the previously hidden state d_{i+1}, decoder $\tilde{y}_i$, $W_f^{'}, W_k^{'}, W_T^{'}, W_s^{'} \in R^{p \times (p+1)}$, $b_f^{'}, b_K^{'}, b_T^{'}, b_s^{'} \in R^p$ are the parameters to be learned the parameters. The σ is a logistic s-type function and elementwise multiplication. For the prediction of industrial rotating equipment failures, the model is utilized to approximate the function F to obtain an estimate of the current output $\tilde{y}_i$ with the combination of inputs and previous outputs. Specifically, it can be obtained via Eq. (12):

$$\tilde{y}_l = F(y_1, ..., y_{l-1}, x_1, x_2, ..., x_l) = v_y^T (W_y[d_l; c_l] + b_w) + b_v \tag{22}$$

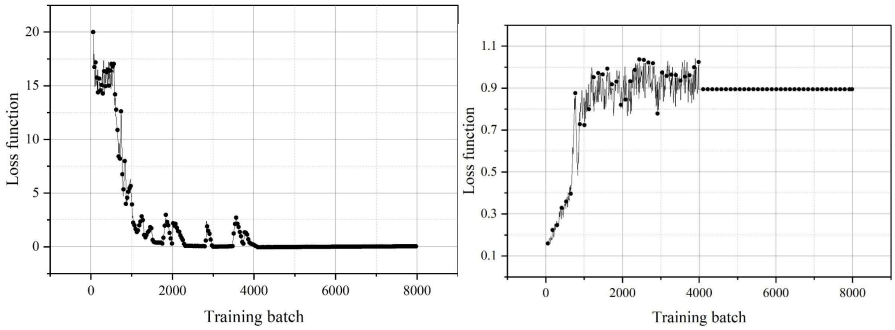
where $[d_l; c_l] \in R^{p+m}$ is the connection vector of the decoder's hidden state and context vectors. The parameters $W_y \in R^{p \times (p+m)}$ and $b_w \in R^{p+m}$ map the cascade to the size of the decoder's hidden state. A linear function of weights $v_y \in R^{p+m}$ and deviations $b_v \in R$ is used to predict the final result. In the encoder, ProbSparse is used instead of attention. The standard self-attention is replaced with ProbSparse. Attention reduces the extracted features as the number of network layers increases to avoid an explosion in the network size. The decoder receives a long string of inputs, computes the attention of the weighted features and outputs a prediction.

III. Research on multimodal data analysis and fault prediction of power system

III. A. Algorithm Performance Analysis

III. A. 1) Algorithm convergence

In this paper, the stochastic gradient descent (SGD) algorithm is used for model training, the learning rate is 0.025, and the convergence process is shown in Fig. 3, where Fig. (a) shows the loss function and Fig. (b) shows the training accuracy. Where Figure 3 (a) is the loss function for each training batch, the fluctuation phenomenon of the loss function is more obvious at the beginning of training, on the one hand, because of the high feature dimension of the input data, on the other hand, because of the close spatial-temporal coupling of the input data, which leads to the possibility that the model may fall into the local optimal point at the beginning of training, which can be avoided by setting a smaller learning rate. After 4000 batches of training, the loss function is basically stabilized at about 0, at which time the model is considered to converge. Figure 3(b) shows the change of training accuracy with training batches, which shows that when the model converges, the training accuracy is more than 80%, and basically stabilized at about 90%~100%.



(a)Loss function (b)Training accuracy rate

Figure 3: Convergence of BC-AT-LSTM neural network

III. A. 2) Forecast accuracy

The prediction accuracy of the proposed method is validated in the test set using cross-checking and some of the results are shown in Fig. 4. The results show that the proposed method has a high prediction accuracy, and the average of the prediction accuracy for all grids is 82.064% when the prediction is performed on 500 grids. Comparing the predicted and labeled values of 128 of the grids, the mean value of the difference between the predicted and labeled values is 0.00346, which shows that the model predicts better for the faults in the power system.

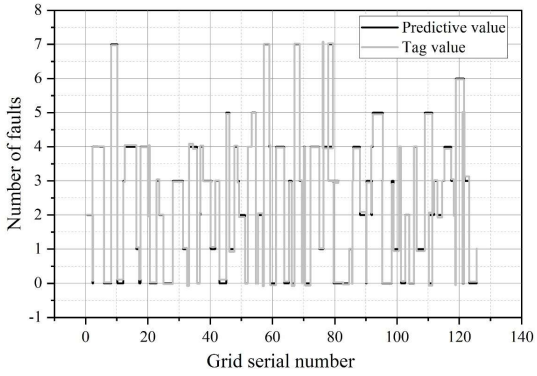


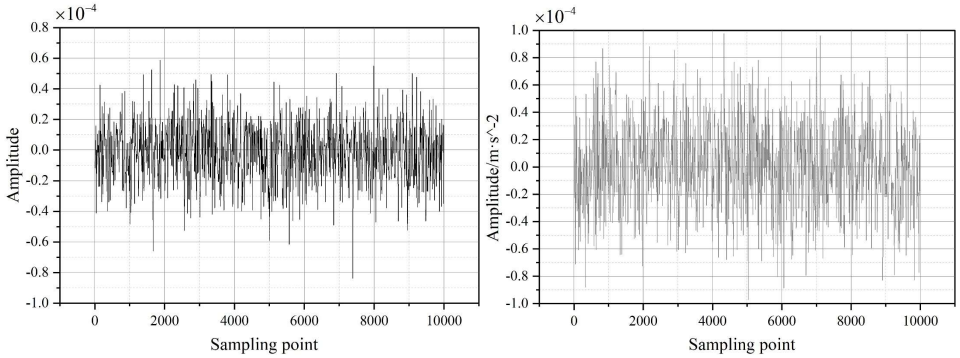
Figure 4: Prediction results of BC-AT-LSTM neural network

III. B. Multimodal data signal analysis

Due to the location of the power system measurement points and the differences in structure, the amplitude of the vibration signals of each channel in the experiment is different. In this paper, the data of the 3rd channel is selected, which is the closest to the winding fault set point, and the amplitude of vibration of the signal is the largest. The signal characteristics of different working conditions are all taken from the same current measurement. Fig. 5 shows the vibration signals before and during the loosening of the windings, Fig. (a) shows before loosening and

Fig. (b) shows during loosening. The vibration signals of the windings in the normal and faulty states at a load current of 110 A are depicted, respectively.

The fast Fourier transform is applied to extract the vibration time-domain signals at the time of system failure, and it can be seen from the time-domain diagrams that there is a significant change in the amplitude of the windings before loosening and under normal conditions, the amplitude of the windings before loosening fluctuates between -0.9~0.6m·s-2, and the vibration signals after loosening fluctuates between -1~1 m·s-2.



(a)Before loosening (b)When loose

Figure 5: The vibration signal of the winding before and when loose

III. C. Example equipment prediction results

III. C. 1) Detection result curve

Figure 6 shows the actual gearbox oil temperature sequence and the predicted temperature sequence of the BC-AT-LSTM model for Unit 26. The actual case of Unit 26 of a wind farm is collected to verify the monitoring performance of BC-AT-LSTM model on the gearbox of a normally operating wind turbine.The gearbox oil temperature overruns on May 15 and June 28 were caused by the failure of the gearbox temperature control valve and radiator, which were lubrication and cooling faults of the gearbox, respectively. The normal operation data from February to March before the failure were selected to train BC-AT-LSTM, CEEMDAN-AE, XGBoost, CNN-LSTM, and LSGAN, respectively, and the actual operation data from March 6 to July 1 were selected to test the model. Among them, BC-AT-LSTM and CEEMDAN-AE are multi-part warning models, and the rest are single-part warning models. In order to ensure the relative fairness of the comparison, a sliding window preprocessing operation is used to control the approximate model training time, hyperparameters, etc. Before May 1, the actual value of the gearbox oil temperature in the part of the gearbox that operates normally has a small deviation from the predicted value of the proposed model in general, and the error values of the two are 60.491 and 45.469 respectively when the failure occurs[24]-[26] .

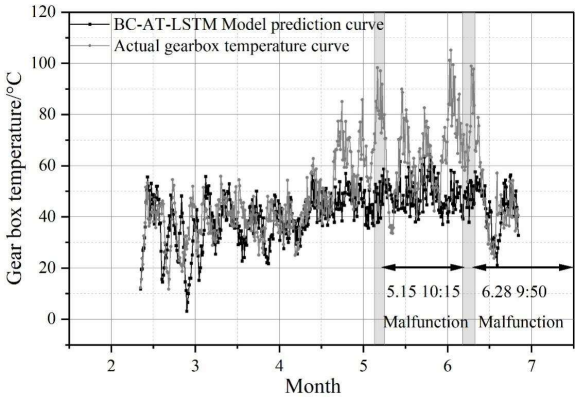


Figure 6: #26 Unit Gearbox Oil Temperature Monitoring Curve (February-July)

III. C. 2) Comparison of test results

Figure 7 shows the prediction curves of each comparison model from March 27 to 28.The mean values of the differences between BC-AT-LSTM, XGBoost, CEEMDAN-AE, CNN-LSTM and LSGAN and the true values are

1.142, 3.759, 2.589, 1.539 and 1.473, respectively, and the BC-AT-LSTM outperforms the other models, which verifies that the proposed BC-AT-LSTM prediction model has high monitoring accuracy for normal operation power system.

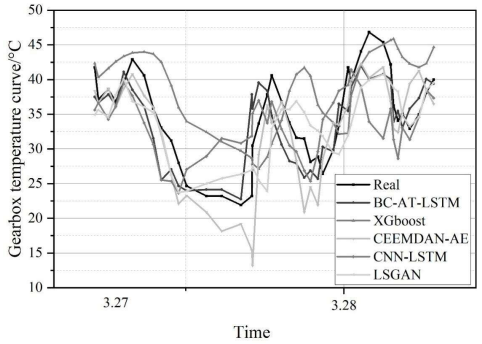


Figure 7: The monitoring curve of each comparison model

IV. Conclusions and outlook

IV. A. Conclusion

In this paper, Word2Vec technique and Fast Fourier Transform technique are applied to extract text data features and speech spectral features in the electric power industry, respectively. The extracted multimodal data are fused using BC-AT-LSTM model, and the stochastic gradient descent algorithm is used to put into model training. The loss function of the model constructed in this paper is basically stabilized around 0, and the training accuracy is more than 80%. Prediction of 500 grids, the average of the accuracy of all grids is 82.064%, the mean value of the difference between the predicted value and the labeled value of the two is 0.00346, and the model has a good prediction effect. The FFT technique is applied to extract the vibration time-domain signals when the system fails, and the vibration signals before and after the loosening of the windings are in the amplitude range of [-0.9,0.6]m·s-2, [-1,1]m·s-2, respectively, and there is a significant difference in the amplitude. Applying the model to the real world, the predicted dates of the faults are 5.15 and 6.28, and the error values of the two days are 60.491 and 45.469, respectively. Comparing the prediction curves of the different models with the model constructed in this paper, the mean values of the differences between the BC-AT-LSTM, XGBoost, CEEMDAN-AE, CNN-LSTM, and LSGAN and the true values are respectively 1.142, 3.759, 2.589, 1.539, 1.473, BC-AT-LSTM is better than the other models and has higher prediction accuracy.

IV. B. Outlook

The research in this paper realizes the construction and reasoning of the equipment system failure prediction model for the power industry, which can more accurately identify the potential failure factors and predict the possible failures of the equipment by constructing the failure prediction model of the power equipment, integrating the multi-source data such as historical failure data, equipment information, and the operating environment, and analyzing the correlation between the equipment and the influencing factors, so that it can provide a strong failure prediction of power equipment system by more accurately identifying the potential failure factors and predicting the possible failures of the equipment. It can provide strong data support and decision-making basis for power equipment system fault prediction. However, combining the research results accomplished in this paper, as well as some problems found in the research process and the summarization of references at home and abroad, we make the following outlook on the research development:

(1) When creating the power equipment fault prediction model, future research can consider adopting the methods of far-supervised learning, active learning, semi-supervised learning and domain expert participation to carry out the knowledge extraction research in order to improve the efficiency and quality of the construction of the knowledge graph of power equipment faults. How to efficiently construct and enrich the knowledge graph of power equipment is a key issue to improve the accuracy and efficiency of fault prediction.

(2) The operating conditions of electric power equipment are affected by a variety of factors, including the equipment's own characteristics, environmental conditions, operating loads, and so on. Therefore, how to effectively integrate multi-source information from voice, text and other sources in the power industry is the key to improve the fault prediction capability. In this paper, only 2 entity relationship types are summarized and designed, and more entity relationship types can be considered subsequently for multi-source data fusion and analysis, which can provide more comprehensive and in-depth data support for the fault prediction model and improve the accuracy and reliability of the prediction.

(3) Explore more intelligent and accurate fault prediction methods. The equipment fault prediction method proposed in this paper, which combines XGBoost decision tree and ProbSparse technology, has limited generation of logic rules, and for more complex logic relationships, it is still necessary to validate and correct the rules by combining the knowledge of domain experts to ensure their accuracy and credibility. Mechanisms and standards for rule validation can be established, and the rule base can be continuously improved and refined through expert review and feedback mechanisms.

Funding

This work was supported by SGTJXT00SJJS2400175 (Information and Communication - Research and Development 2024-02) Research on Key Technologies of Power Artificial Intelligence Speech Semantic Large Model.

References

[1] Omitaomu, O. A., & Niu, H. (2021). Artificial intelligence techniques in smart grid: A survey. Smart Cities, 4(2), 548-568.
[2] Khan, M. A., Saleh, A. M., Waseem, M., & Sajjad, I. A. (2022). Artificial intelligence enabled demand response: Prospects and challenges in smart grid environment. Ieee Access, 11, 1477-1505.
[3] Zhang, D., Jin, X., & Shi, P. (2023). Research on power system fault prediction based on GA-CNN-BiGRU. Frontiers in Energy Research, 11, 12454
[4] Zhang, S., Wang, Y., Liu, M., & Bao, Z. (2017). Data-based line trip fault prediction in power systems using LSTM networks and SVM. Ieee Access, 6, 7675-7686.
[5] Chen, X., Fu, W., Liu, C., Liu, Z., Li, J., Hu, Z., & Hu, D. (2025). Application of Artificial Intelligence Large Language Model in Power Equipment Operation and Maintenance. Strategic Study of Chinese Academy of Engineering, 27(1), 180-192.
[6] Liu, J., Duan, Z., & Liu, H. (2024). A grid fault diagnosis framework based on adaptive integrated decomposition and cross-modal attention fusion. Neural Networks, 178, 106400.
[7] Zhang, X., Liu, T., Zheng, H., Zhang, F., & Zhang, Q. (2024). A Knowledge Extraction Method of Power Equipment Operation and Inspection for Multimodal Data. In Applied Mathematics, Modeling and Computer Simulation (pp. 249-259). IOS Press.
[8] Zhou, F., Wen, G., Ma, Y., Geng, H., Huang, R., Pei, L., ... & Qiu, R. (2022). A comprehensive survey for deep-learning-based abnormality detection in smart grids with multimodal image data. Applied Sciences, 12(11), 5336.
[9] Cui, W., & Liu, H. (2024, December). Research on Transformer Condition Assessment Based on Multimodal Data. In International Conference on Artificial Intelligence and Autonomous Transportation (pp. 304-312). Singapore: Springer Nature Singapore.
[10] Kong, Z., Zhang, C., Lv, H., Xiong, F., & Fu, Z. (2020). Multimodal feature extraction and fusion deep neural networks for short-term load forecasting. IEEE access, 8, 185373-185383.
[11] Zhao, Y., Zhang, Y., Li, Z., Bu, L., & Han, S. (2023). AI-enabled and multimodal data driven smart health monitoring of wind power systems: A case study. Advanced Engineering Informatics, 56, 102018.
[12] Xu, B., Li, H., Ding, R., & Zhou, F. (2025). Fault diagnosis in electric motors using multi-mode time series and ensemble transformers network. Scientific Reports, 15(1), 7834.
[13] Xiaodong, Z., Runzhen, Y., Hui, L., & Erfei, J. (2023). Artificial Intelligence in power multimodal data analysis. Procedia Computer Science, 221, 1312-1320.
[14] Alqudah, M., Kezunovic, M., & Obradovic, Z. (2022). Automated power system fault prediction and precursor discovery using multi-modal data. IEEE Access, 11, 7283-7296.
[15] Xing, Z., & He, Y. (2023). Multi-modal information analysis for fault diagnosis with time-series data from power transformer. International Journal of Electrical Power & Energy Systems, 144, 108567.
[16] Afrasiabi, S., Allahmoradi, S., Afrasiabi, M., Liang, X., Chung, C. Y., & Aghaei, J. (2024). A Robust Multi-modal Deep Learning-Based Fault Diagnosis Method for PV Systems. IEEE Open Access Journal of Power and Energy.
[17] Alsaif, K. M., Albeshri, A. A., Khemakhem, M. A., & Eassa, F. E. (2024). Multimodal Large Language Model-Based Fault Detection and Diagnosis in Context of Industry 4.0. Electronics, 13(24), 4912.
[18] Ke, L., Hu, G., Yang, Y., & Liu, Y. (2023). Fault diagnosis for modular multilevel converter switching devices via multimodal attention fusion. IEEE Access, 11, 135035-135048.
[19] Imen Jarraya,Safa Ben Atitallah,Fatimah Alahmed,Mohamed Abdelkader,Maha Driss,Fatma Abdelhadi & Anis Koubaa. (2025). SOH-KLSTM: A hybrid Kolmogorov-Arnold Network and LSTM model for enhanced Lithium-ion battery Health Monitoring. Journal of Energy Storage,122,116541-116541.
[20] Jing Zhou,Zhanliang Ye,Sheng Zhang,Zhao Geng,Ning Han & Tao Yang. (2024). Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data. Heliyon,10(16),e35945-e35945.
[21] Xiaofang Liu,Guotian He,Shuge Li,Fan Yang,Songxiying He & Lin Chen. (2025). Multi-level feature decomposition and fusion model for video-based multimodal emotion recognition. Engineering Applications of Artificial Intelligence,152,110744-110744.
[22] Abhijeet Rathore,Priya Gupta,Raksha Sharma & Rhythm Singh. (2025). Day ahead solar forecast using long short term memory network augmented with Fast Fourier transform-assisted decomposition technique. Renewable Energy,247,123021-123021.
[23] Yuhan Xie,Gensuo Mi,Ding Tan & Chenning Liu. (2025). WOA-XGBoost Based Railway Accident Type Prediction and Cause Analysis. Engineering Letters,33(3),
[24] H. Li H, Y.X. Zhang, Y , S.J. Guo, et al. A comprehensive study on texture evolution and recrystallisation behaviour of Fe-50Co alloy, Electrical Materials and Applications, 1 (1) e12010, 2024.
[25] H. Sun, Y.L. Shang, Y. Han, et al. Research on magnetic field characteristics of amorphous alloy and grain-oriented silicon steel hybrid magnetic circuit iron core, Electrical Materials and Applications, 1 (2) e12014, 2024.
[26] X.Y. Wang, R.F. Xue, G Ma. Error analysis and correction strategy for measuring oriented silicon steel by SST method, 1 (2) e70002, 2024.