Application of BERT-Free Pre-Training Model in Identifying Power Safety Risk Points

Siwu Yu[1], Yumin He[1], Guobang Ban[1], Teng Gui[2], Xiangquan Hu[2], Yang Mei[2], Jingjing Zhang[2]

[1]Electric Power Research Institute of Guizhou Power Grid Co., Guiyang 550002, Guizhou, China
[2]Guizhou Power Grid Co., Guiyang 550002, Guizhou, China

Corresponding author: Siwu Yu (e-mail: swyu2012@163.com)

ABSTRACT In intelligent power systems operating under increasingly complex electromagnetic environments, accurate identification of safety risk points is essential for ensuring reliable equipment operation and supporting electromagnetic compatibility assessment. Traditional rule-based methods suffer from limited semantic understanding and poor generalization, making them insufficient for processing complex operation and maintenance texts. To address this issue , this paper proposes a lightweight CNN-based pre-training model built on a BERT-Free architecture for efficient and accurate power safety risk identification. A professional dataset containing 58,000 power operation and maintenance texts is constructed, and a high-quality multi-label corpus covering four categories and 17 subcategories is established through dictionary-enhanced word segmentation and expert cross-annotation. The model is exported in ONNX format to facilitate flexible deployment in engineering applications. Experimental results demonstrate that the proposed model achieves an overall accuracy of 89.1% and an F1-score of 87.3%. Under class imbalance, the F1-score reaches 0.901 for majority classes and 0.784 for minority classes, exhibiting strong robustness. The average inference time per sample is only 8.3 ms, indicating high computational efficiency. These results demonstrate that the proposed model provides an effective and practical solution for intelligent power safety risk identification while offering valuable support for electromagnetic infrastructure monitoring and reliable operation in modern power systems.

INDEX TERMS BERT-Free architecture, cnn-based pre-training model, power safety risk identification, electromagnetic environment, intelligent power systems

I. INTRODUCTION

IN modern society, the power system is a key national infrastructure . The reliability of these advanced functional fabrics in safeguarding electrical equipment is essential for ensuring the continuous operation of power grids and protecting people's lives and property[1,2]. With the rapid development of smart grids and new energy technologies, the structure of the power system has become more complex, and the operating conditions have become increasingly diverse, which puts higher requirements on safe operation and maintenance [3,4]. As a key part of power safety management, the risk point identification task aims to timely discover potential faults and anomalies by analyzing unstructured text data, such as operation and maintenance records and hidden danger reports, to effectively warn and guide the adjustment of operation and maintenance measures to avoid the occurrence of safety accidents [5,6]. Therefore, existing methods still have obvious shortcomings in terms of intelligence, deployability, and industry adaptation. Based on the above problems, this paper proposes a lightweight CNN-based pre-trained language model that integrates the BERT-Free architecture, aiming to achieve efficient and accurate identification of power safety risk points. "BERT-Free" in this paper does not negate the ideas of self-supervised pre-training or mask modeling, but rather refers to the model's encoding structure not using the Transformer architecture based on self-attention mechanisms, but instead using convolutional neural networks for semantic modeling. This model inherits the advantages of CNN in capturing local features and parallel computing and replaces the self-attention mechanism with convolution operations, effectively reducing the model calculation complexity and improving the reasoning efficiency, making it more suitable for deployment in edge devices or real-time systems. At the same time, in view of the special expressions and implicit semantic features in the field of power text, the model design applies a task-driven training strategy. It performs customized optimization in terms of corpus construction and masking mechanism to improve the generalization ability in diverse expressions and complex contexts. This study

not only provides a more efficient intelligent method for identifying power safety risks associated with complex insulation materials and fiber-based electrical components, but also offers a new practical path for expanding BERT-Free language models into specialized power vertical industries. By accurately capturing the technical semantics of high-performance fabric performance, this approach enhances the cross-industry application of large-scale models in ensuring infrastructure safety. It should be noted that this model is primarily aimed at safety risk identification in operation and maintenance texts of medium- and high-voltage power transmission and transformation systems. It is applicable to typical equipment scenarios such as transformers, circuit breakers, switchgear, and transmission lines. Its role is to assist operation and maintenance personnel in text-level risk screening and early warning support, rather than directly participating in power system dispatching or macro-level safety decisions.

II. RELATED WORK

As an important component of intelligent power system operation and maintenance, the identification of power safety risk points mainly relied on rule-based and expert system methods in the early stage of research [7,8]. This method extracts text information and identifies risks by manually formulating keywords, pattern-matching rules, or relying on experts to build a knowledge base, which has the advantages of simple implementation and clear logic. However, with the rapid growth of power system data and the increasing complexity of text expression, these methods have gradually exposed problems such as reliance on manual experience, poor adaptability, and difficulty in generalization [9-11]. On the one hand, the construction and updating of rules are highly dependent on domain experts, which results in heavy workload and high maintenance costs.

To address the above issues, the BERT-Free architecture pre-trained language model that has gradually emerged in recent years has provided new ideas for the application of deep learning methods in vertical industries[12]. The BERT-Free model aims to replace or simplify the traditional self-attention mechanism, mainly using convolution, frequency domain transformation, and other methods to construct semantic representation, significantly reducing computing resource consumption. Among them, the CNN-based pre-trained model has shown good performance in semantic understanding tasks due to its natural advantages in local context modeling and efficient reasoning, becoming a representative path for lightweight semantic modeling[13,14]. At the same time, FNet further reduces the amount of multiplication and addition operations of the model by applying the Fourier transform mechanism[15,16], while ConvBERT combines convolution and attention mechanisms to balance modeling capabilities and execution efficiency[17,18]. In power safety risk identification, although some studies have attempted to apply deep learning models for improvement, most are still focused on rule optimization or improvement of shallow network structures. There is still a lack of research on power text semantic modeling based on the CNN-based BERT-Free architecture [19,20]. Based on this research gap, this paper attempts to apply a CNN-based pre-training model to explore its application potential and actual effect in power risk identification tasks.

III. CONSTRUCTION OF CNN-BASED POWER SAFETY RISK POINT IDENTIFICATION MODEL

DATA PREPARATION

(1) Data collection

The data used in this study comes from an internal database provided by a power company in a province in East China. The database contains power system operation data and operation and maintenance records from January 2018 to December 2023. Specific data types include, regular operation logs, fault event reports, hidden danger investigation records, team safety records, on-site inspection feedback, and other text materials. Data collection covers 45 major substations, more than 600 transmission lines, and supporting switchgear, protection devices, and other facilities in the province, which is highly representative of the industry and has practical application value. It should be noted that dynamic numerical information such as voltage, current, and temperature in power operation and maintenance texts typically appears in natural language. This paper does not directly introduce the original time series data. Instead, it encodes key numerical change information into textual semantic features through methods such as numerical-unit joint lexical units, anomaly description normalization, and trend modifier retention, thereby serving the risk point identification task. Time-related information is also used in modeling in discrete language form and is subsequently extracted by the convolutional coding module.

A total of about 58,000 original text records are collected, with a total word count of more than 12 million words, of which about 35% are records that clearly marked safety risk events. Data collection is mainly achieved through interface docking with the enterprise information center, automatic extraction of relevant fields, and data cleaning and screening combined with a manual review mechanism. In addition, some special scenario data, such as major failures and typical accident cases, are supplemented by operation and maintenance experts to ensure the diversity and integrity of the data.

(2) Data preprocessing

To improve the effectiveness of model training and the accurate understanding of the semantics in the power field, this paper carries out systematic data preprocessing on the original text data, covering three core links: text cleaning, word segmentation, and risk labeling. In the text cleaning stage, the UTF-8 encoding format is uniformly used to filter redundant characters, special symbols, and control characters in the original records, and HTML tags and non-critical information fields automatically generated by the log

system are removed to ensure the structural standardization and semantic purity of the input text. In order to achieve the uniform expression of terms, a terminology standard table is constructed based on power operation and maintenance regulations and typical cases, and entries with similar semantics but different expressions are normalized, such as "circuit breaker tripping" and "tripping protection action" are uniformly marked as standard terms.

In view of the characteristics of dense professional terms and non-standard language structure in power texts, the word segmentation process adopts a custom dictionary method based on the jieba framework. During the research process, a "Power Equipment and Risk Terminology Vocabulary" containing more than 1,800 entries is constructed, covering multiple dimensions such as equipment type, risk behavior, and operation terms. At the same time, in order to ensure semantic integrity, a mandatory segmentation rule is designed to achieve accurate recognition of high-frequency phrases in the fields of "busbar fault", "three-phase imbalance", and "switch cabinet abnormality".

In the risk point labeling phase, relying on the experienced expert resources in the power industry, this study forms a labeling team consisting of 5 power safety experts with more than 10 years of work experience. The experts jointly develop a risk labeling system covering four categories, namely equipment failure, electrical anomaly, operation violation, and environmental risk, with a total of 17 subcategories, as shown in Table 1.

All text data are cross-labeled in multiple rounds. Each text is processed independently by at least two annotators, and the consistency of annotation is evaluated using the Fleiss' Kappa coefficient. The parts with disagreements are finally confirmed through a collective arbitration mechanism to ensure the high reliability and professionalism of the annotation quality.

MODEL CONSTRUCTION

This study uses a CNN-based architecture to build a BERT-Free pre-trained model, aiming to significantly reduce the parameter size and computing overhead while ensuring semantic modeling capabilities, improving training efficiency, and adapting to the massive text data processing needs of the power industry. It should be noted that fragment mask reconstruction and semantic consistency judgment are self-supervised signals and are independent of the specific encoder form. This paper introduces them into the CNN architecture to enhance the model's domain semantic representation ability, rather than to reproduce BERT's Transformer training mechanism.

(1) Overall model architecture design

The overall architecture of the BERT-Free pre-trained model constructed in this study consists of an embedding layer, a convolutional coding module, a feature fusion layer, and an output layer. It aims to fully retain the modeling ability of text semantics while controlling the computational complexity and model parameter scale to adapt to large-scale, short text-intensive actual application scenarios in the power industry. The specific structure is shown in the figure 1.

Figure 1 shows the overall architecture of the model. The gray ellipse in the figure is the output layer, which receives the preprocessed power operation and maintenance text sequence. Its dimension label Batch×SeqLen represents the tensor structure of batch size and sequence length. The embedding layer (blue box) converts discrete vocabulary into a continuous vector space representation. The d-dimensional dimension marked in the figure needs to be determined according to the actual parameters. The text features then enter the multi-scale convolutional encoding module. Three sets of parallel 1D convolutional layers (dark blue boxes) use three convolution kernel sizes of 3/5/7, respectively, aiming to capture short, medium, and long-distance local semantic patterns in the equipment status description. The maximum pooling layer (green box) following each convolutional layer compresses the feature dimension through a downsampling operation with a step size of 2. The feature fusion layer (purple component) concatenates the feature tensors output by multi-scale convolution along the channel dimension to form a comprehensive semantic representation. The self-supervised pre-training module marked with orange diamonds in the figure performs the tasks of fragment masking reconstruction and semantic consistency judgment, which is a strategic design in the model training stage. Finally, the output layer (red ellipse) maps the fused features to the label space through the fully connected layer. The Sigmoid activation function and the dynamic threshold judgment mechanism jointly handle the multi-risk co-occurrence scenario and output the risk classification results that meet the power safety regulations. The arrows between the layers clearly represent the topological relationship between the feature transfer path and the model components.

(2) Convolutional coding module design

In the convolutional encoding module designed in this study, the input is the representation matrix $X \in R^{L\times d}$ generated by the embedding layer, where L represents the length of the token sequence and d is the embedding dimension (set to 256 in this experiment). To enhance the model's ability to capture local semantics at different scales, the convolution module sets three sets of one-dimensional convolution kernels in parallel, with sizes of 3, 5, and 7, respectively. The number of channels is set to 128, the step size is 1, and the padding strategy uses "same" to ensure that the number of time steps of the convolution output remains unchanged.

Assuming the input sequence is $X = [x_1, x_2, \ldots, x_L]$, the k-th convolution kernel $W^{(k)} \in R^{n_k\times d}$ of size n_k performs a sliding operation on the input sequence, and the convolution result of the i-th time step is expressed as:

$$h_i^{(k)} = \mathrm{ReLU}\left(\sum_{j=0}^{n_k-1} W_j^{(k)} \cdot x_{i+j-\lfloor n_k/2\rfloor} + b^{(k)}\right) \quad (1)$$

TABLE 1. Risk labeling system classification structure

Risk category	Subcategory tags
Equipment failure	① Circuit breaker failure; ② Busbar failure; ③ Switch cabinet abnormality; ④ Transformer overheating; ⑤ Transformer damage;
Electrical abnormality	① Abnormal grounding; ② Voltage fluctuation; ③ Three-phase imbalance; ④ Harmonic interference; ⑤ Abnormal current
Operation violation	① Unauthorized operation; ② Unauthorized live work; ③ Illegal operation;
Environmental risks	① High temperature and high humidity; ② Dust pollution; ③ Water immersion corrosion; ④ Thunderstorm weather

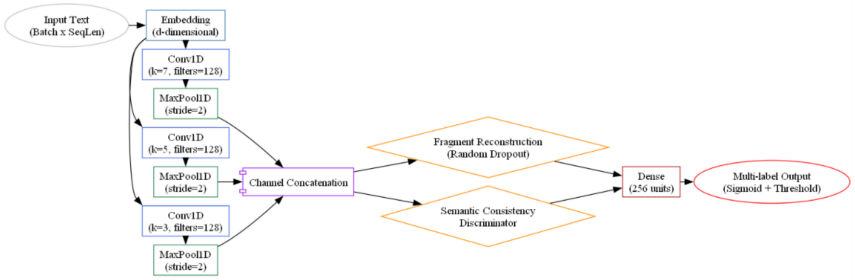

FIGURE 1. Overall architecture of the model

ReLU(·) is the activation function, and b(k) is the bias term. Through this operation, the model can extract local context features centered on the current word with left and right window sizes of 1, 2, and 3, respectively.

After each set of convolution outputs, the maximum pooling operation is connected to compress the sequence length and retain the significant response. The pooling process can be expressed as:

$$p_i^{(k)} = \max\left(h_{2i}^{(k)}, h_{2i+1}^{(k)}\right) \tag{2}$$

After completing three different sets of convolution paths and pooling, all feature maps are concatenated in the channel dimension, and the final feature representation is:

$$H = [P^{(1)}; P^{(2)}, P^{(3)}] \in R^{\frac{L}{2}\times 384} \tag{3}$$

Among them, 384 come from the concatenation of 128 channels of each of the three groups of convolution paths. In order to further enhance the model's deep expression ability, the convolution module can stack the second layer structure, use a convolution kernel size of 3, set the number of channels to 64, and keep padding as "same" so that it can build context aggregation across time steps on the local receptive field. Compared with the Transformer architecture that relies on the global attention mechanism, this module only relies on local convolution to complete feature modeling, which greatly reduces the computational complexity and memory consumption, avoids the problems of attention diffusion and redundant calculation under long text conditions, and is more suitable for deployment in resource-constrained environments in the power industry, such as edge terminals or real-time monitoring equipment.

MODELING METHOD FOR IDENTIFYING POWER SAFETY RISK POINTS

(1) Modeling method for risk point identification tasks

In this study, the task of identifying power safety risk points is modeled as a multi-label prediction problem in the context of sequence classification, aiming to cope with complex scenarios where a single operation and maintenance text may contain multiple risk factors at the same time. The model takes the power operation and maintenance text that has been processed by word segmentation, denoising, and vectorization embedding as input and outputs a confidence distribution vector corresponding to the label space, which is used to characterize the various safety risk labels corresponding to the text. The label system is built based on the actual needs of safety management in the power industry, and the design logic integrates the two dimensions of "risk level" and "risk type". Among them, the risk level is divided into "red warning", "orange warning" and "yellow warning", corresponding to serious, medium, and slight risk levels, respectively; the risk type covers typical scenarios such as equipment failure, environmental abnormalities, and human

operational errors. Through the combination of the above two dimensions, more than 30 composite labels are formed, which fully cover the most representative risk situations in the research data set. The label data is manually annotated by domain experts based on historical operation and maintenance records and fault samples. It is cross-reviewed multiple times to ensure its accuracy and consistency.

To support the actual needs of multi-label output, the model output layer uses the Sigmoid activation function to generate confidence scores for each label independently. Unlike the traditional Softmax mechanism, Sigmoid allows the model to identify multiple labels in the same text, which is more in line with the actual characteristics of "complex faults" and "risk linkage" in the power system. For the prediction confidence pj of the j-th label, the calculation formula is:

$$p_j = \sigma(z_j) = \frac{1}{1 + e^{-z_j}}, \quad j = 1, 2, \ldots, C \tag{4}$$

C is the total number of labels, and zj is the original output score corresponding to the label.

(2) Model output interpretation and result construction

To improve the operability and business interpretability of the model prediction results, this study applies a standardized threshold judgment mechanism in the result construction stage. The default judgment threshold is set to 0.6. If the prediction confidence of a label is greater than or equal to this threshold, the label is considered to be valid in the text; otherwise, it is considered invalid. This threshold can be dynamically adjusted on the validation set according to actual business needs, and the optimal decision point can be searched through the F1-score curve to achieve the optimal balance between precision and recall in risk identification.

The output results are displayed in a structured form, including all the valid labels corresponding to each text and their corresponding confidence. For example, for a text describing abnormal cable temperature, the system may output "Red Warning-Equipment Failure" (confidence 0.87) and "Yellow Warning-Environmental Risk" (confidence 0.63), which supports users to quickly understand the potential risk sources and levels. In the case of multi-label output conflicts, this study further designs a label conflict handling mechanism to ensure the consistency and credibility of the output. Specifically, when there are multiple level labels or multiple severity labels of the same type, the system first prioritizes the labels according to their confidence. Then, it cuts them based on the preset label mutual exclusion matrix and risk level priority rules. The label mutual exclusion matrix is used to characterize the objectively existing logical exclusive relationships in power business. Its rules originate from operating procedures and expert experience, not from automatic model learning. For example, when indicating that equipment is in a 'circuit breaker fault' state, it cannot simultaneously be labeled as a normal operation category such as 'routine inspection feedback'. If the confidence levels are close, or the labels are reasonable to coexist in business logic, a heuristic weighted average and business rule engine are applied to make judgments and select the label combination that best matches the actual semantics for output. The overall mechanism is designed to ensure the consistency of the risk label, which results in logical structure, representativeness in semantic expression, and interpretability in business applications.

IV. MODEL EFFECT VERIFICATION

In order to comprehensively evaluate the performance of the constructed CNN-based pre-trained model in the task of power text safety risk identification, this study evaluates its performance through experiments based on the data collected in the previous paper.

A. EXPERIMENTAL SETUP

(1) Experimental hardware and software environment description

All experiments are completed on a local server with an operating system of Ubuntu 20.04 LTS (64-bit), which provides a stable Linux execution environment and good Python support. The hardware configuration includes an Intel Xeon Silver 4210 processor (10 cores, 2.2GHz), 64GB DDR4 ECC memory, an NVIDIA RTX 3060 graphics card (12GB video memory, CUDA 8.6), and a 2TB SATA SSD, which can meet the training and inference requirements of medium-scale deep learning tasks. The Gigabit Ethernet interface supports remote management and multi-host data transfer. The software environment is managed by Conda, and the main dependencies include Python 3.9.13, PyTorch 2.0.1 (CUDA 11.7), Transformers 4.34.0, scikit-learn 1.3.0, torchmetrics 1.2.0, and matplotlib and seaborn for result visualization.

To enhance the reproducibility of the experiment, the random seed is uniformly set, and the model initialization strategy is fixed before training. The training and evaluation process supports command line parameter calls, and key hyperparameters such as model structure, batch size, and learning rate can be flexibly adjusted with good portability and deployment adaptability.

(2) Comparison method

In order to comprehensively evaluate the effectiveness of the constructed CNN-based model in the task of identifying power safety risk points, this paper selects four representative benchmark methods as Control Group, covering traditional feature engineering methods, mainstream pre-trained models, and lightweight structures to ensure the systematic and objective performance comparison.

Control Group 1 is a model based on traditional feature methods, which uses TF-IDF (Term Frequency– Inverse Document Frequency) to vectorize text data and combines it with a support vector machine (SVM) for classification. Control Group 2 is a Transformer-based pre-trained model, using BERT-Base (Bidirectional Encoder Representations

from Transformers) as a representative. It should be noted that control group 2 uses the general BERT-Base model without further pre-training in the power sector to reflect its typical performance under actual engineering deployment conditions. Control Group three and Control Group four are representatives of BERT-Free models, namely FNet and ConvBERT. Among them, FNet significantly reduces the computational complexity of the model by applying Fourier transform to replace the standard self-attention module, which is suitable for fast reasoning tasks in resource-constrained scenarios. ConvBERT combines dynamic convolution and self-attention mechanisms to enhance the expression of local context while maintaining modeling efficiency, taking into account both accuracy and speed.

EXPERIMENTAL PROCESS AND RESULT ANALYSIS

(1) Overall classification performance evaluation

To systematically evaluate the recognition performance and stability of the constructed model in the power system text classification task, this paper selects four common indicators: classification accuracy, precision, recall rate, and F1-Score to measure the overall performance of each model. Among them, the recall rate metric is used to measure the model's ability to detect texts with potential security risks, which is particularly important in power safety scenarios. Its results have been incorporated into the comprehensive evaluation of the F1 score. The experiment is based on the data set collected and cleaned in the previous paper, and the classification capabilities of the five models are compared and evaluated. In each round of experiments, a confusion matrix is generated to calculate various indicators. All experiments are repeated five times to reduce the volatility caused by parameter initialization. The final result is taken as the average value as the basis for evaluation. The average performance results of each model on the four indicators are shown in Figure 2.

The overall classification performance comparison results in Figure 2 show that the experimental group outperforms all Control Groups in all indicators, showing higher recognition accuracy and classification stability. This shows that the proposed model can more effectively capture the potential risk evolution characteristics in the power system text, thereby achieving more accurate discrimination in the actual classification process. The accuracy of the experimental group reaches 89.1%, and the F1 value is 87.3%, which has obvious advantages in balancing precision and recall.

Control Group 1, as a traditional method, has the lowest performance, with an accuracy rate of only 74.6%, mainly due to its insufficient feature expression ability and difficulty in adapting to the complex and changing text semantic structure. Control Group 2 applies a basic time series modeling mechanism in the model structure, and its performance is slightly improved compared with Control Group 1, but because its context dependency modeling is still shallow, the improvement in accuracy and F1 value is limited. Control Group 3 and Control Group 4 apply deeper neural network structures or graph neural modules, which enhance the model's ability to model semantic relationships to a certain extent, so the classification performance is significantly improved. Among them, Control Group 4 is better than Control Group 2 and Control Group 3 in accuracy and F1 value, and its performance is more stable. However, compared with the experimental group, there are still some shortcomings.

(2) Recognition robustness test under unbalanced samples

To further evaluate the robustness of each model under the condition of severely uneven sample category distribution, an unbalanced test set simulating the distribution characteristics of actual power text is constructed. This dataset is downsampled based on the original test set, and the proportion of main class samples is increased to more than 85% to simulate the typical scenario of the scarcity of small class events in power risk identification tasks. From an engineering safety perspective, minority class samples correspond to potentially high-risk events. Their higher F1 scores indicate that the model can effectively reduce the risk of false negatives while maintaining an acceptable false alarm rate, meeting the practical requirement of "better to overreport than underreport" in power safety applications.

In the experimental evaluation, Macro-F1 is selected as the main indicator, and the category sensitivity performance of each model on the main class and minority class samples is counted respectively to measure its stability and generalization ability in low-frequency category recognition tasks. The experimental process is consistent with the above-mentioned overall classification performance evaluation. All models are trained and tested independently for five rounds on the imbalanced dataset, and the average results are finally taken to reduce the impact of the initial parameter disturbance. The experimental results are shown in Figure 3.

As shown in Figure 3, in the unbalanced sample scenario, the recognition capabilities of each model on the main class and the minority class are significantly different, reflecting the strength of their robustness and generalization capabilities. The experimental group model achieves the highest Macro-F1 on the main class, reaching 0.901, showing excellent main class recognition performance. More importantly, its F1-Score on the minority class still maintains a high level of 0.784, showing good category sensitivity and robustness. In contrast, the performance of Control Group 1 on the main class is significantly behind, only 0.721, and it even drops to 0.493 on the recognition of the minority class, indicating that its adaptability to low-frequency events is weak. Although Control Group 2 has slightly improved in recognition of the main class, reaching 0.733, its Macro-F1 on the minority class is still low, only 0.537, indicating that it is still difficult to overcome the training bias under the condition of class imbalance. The performance degradation of BERT-Base in multilabel and imbalanced class scenarios is related to its high sensitivity to threshold settings and

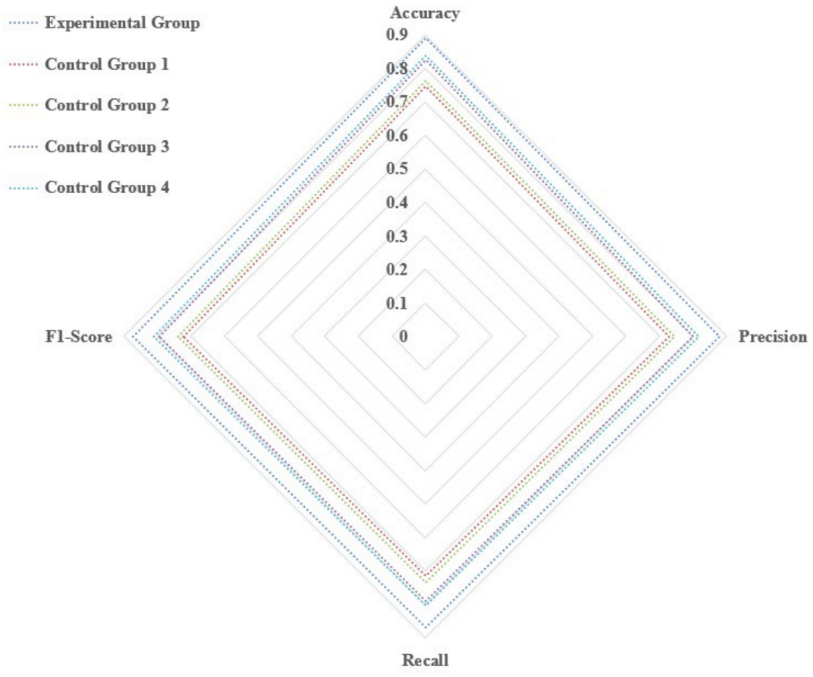

FIGURE 2. Overall classification results

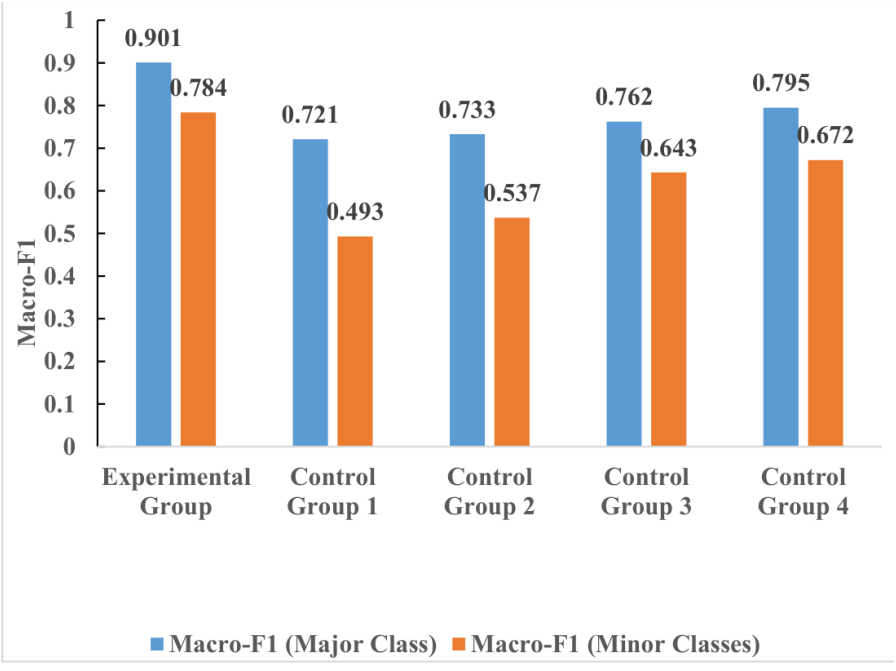

FIGURE 3. Macro-F1 comparison of each model for the main class and minority class in the imbalanced sample scenario

sample distribution, which is representative in engineering applications. The performance of Control Groups 3 and 4 has improved, especially in the recognition of minority classes, which has increased to 0.643 and 0.672, respectively, reflecting that they may have applied an additional attention mechanism for minority class information in model design or training strategy.

(3) Model inference efficiency test

To evaluate the response speed and resource usage of each model in the actual deployment scenario and further verify its reasoning efficiency and application feasibility, this experiment conducts a comparative test on the running performance of each model in the reasoning stage under the same hardware platform and data conditions. The evaluation indicators include the average reasoning time of a single sample and the peak value of GPU memory usage, which

respectively reflect the performance of the model in terms of real-time performance and resource consumption. The statistical results of the reasoning performance of each model under the standard configuration are shown in Figure 4.

As shown in Figure 4, the performance differences of each model in terms of response speed and video memory consumption are closely related to the algorithm architecture. The experimental group model uses a convolutional neural network to replace the self-attention mechanism and has an efficiency advantage with an average inference time of 8.3 milliseconds. Its lightweight design effectively avoids the high complexity of global attention calculation through local feature extraction. Although the traditional machine learning model Control Group 1 has become the solution with the lowest resource requirements due to its 498 megabytes of video memory usage, it is limited by the serial processing flow of feature engineering, and its inference delay of 15.7 milliseconds is significantly higher than that of the experimental group model, highlighting the necessity of end-to-end learning in real-time scenarios. The Transformer baseline model Control Group 2 reaches a peak inference time of 43.5 milliseconds and 3,412 megabytes of video memory due to the computational load of the multi-layer self-attention mechanism, becoming the most resource-intensive solution in the test, revealing the deployment bottleneck of complex architectures. Control Group 3, which uses Fourier transform, optimizes the inference time to 11.9 milliseconds by reducing the computational intensity of spatial interactions, but the video memory requirement of 1,756 megabytes is still higher than the experimental group, indicating the limitations of the frequency domain analysis method in terms of memory efficiency. Control Group 4, which combines convolution and attention, has an inference time of 18.6 milliseconds and occupies 2,683 megabytes of video memory, verifying the trade-off between accuracy and speed in the hybrid architecture. Its parameter redundancy problem further proves the key role of structural simplification in improving deployment efficiency. On this basis, to further analyze the responsiveness of the model under different load conditions, the experiment examines the average inference time change trend under five batch processing scales (batch size = 16, 32, 64, 128, 256). All tests are repeated five times to reduce the impact of chance, and the final result is taken as the average value as the evaluation basis. The change in inference time under different batch sizes is shown in Figure 5.

As shown in Figure 5, the trend of the inference time consumption of each model under different batch sizes is significantly correlated with its algorithm complexity and hardware parallelization capability. When the batch size is gradually expanded from 16 to 256, the inference time of the experimental group model increases from 22.4 milliseconds to 210.3 milliseconds, showing the most gentle growth curve. This is due to the inherent local perception field characteristics of the convolutional neural network and the hardware parallel computing optimization. The linear relationship between its computational complexity and input size effectively controls the computational load increment brought by batch expansion. In contrast, the time consumed by Control Group 2 surges from 86.5 milliseconds to 1140.7 milliseconds under the same conditions, reflecting the restriction of the quadratic computational complexity of the attention matrix on the scalability of the system. Although Control Group 1 shows a lower latency of 18.9 milliseconds when processing small batches, it exposes the efficiency bottleneck of the CPU computing paradigm based on serial feature engineering in large-scale batch processing scenarios when processing large batches. The time consumption of Control Group 3 increases from 29.7 milliseconds to 352.8 milliseconds, indicating that although the frequency domain analysis method can alleviate the pressure of spatial interaction calculation, its global feature fusion mechanism still leads to lower calculation efficiency than the pure convolution architecture. It is worth noting that the time consumption of Control Group 4 during batch expansion increases from 44.8 milliseconds to 512.7 milliseconds, revealing the synergistic effect of combining convolutional modules with attention mechanisms. Although its parameter sharing mechanism partially reduces the computational complexity, the additional overhead caused by cross-module data interaction still limits the further improvement of batch processing capabilities.

(4) Comparison of training process stability

In order to comprehensively evaluate the convergence efficiency and stability of each model during the training phase and further analyze the controllability of its parameter tuning and the robustness of the training process, this experiment systematically records the loss function change curve of each model during the training process and the evolution trend of the F1-Score of the validation set. By comparing the average convergence rounds required by the model during the training process, the fluctuation range of the loss value, and the change trajectory of the validation performance, the training stability and the reliability of the optimization process are measured.

In the specific experimental setting, the Loss value and the corresponding verification F1 score of each round of training are recorded, respectively, and the average value of the key indicators is calculated based on 10 rounds of repeated experiments. The extracted convergence rounds and Loss standard deviation quantitative indicators are used to reflect whether the model is easy to converge during the training process and the degree of performance fluctuation. The training stability evaluation results of each model are shown in Table 2.

From the data in Table 2, it can be seen that the convergence efficiency and stability of each model in the training stage are closely related to its architecture complexity and optimization characteristics. The experimental group model, with its simplified structural design, reaches convergence in

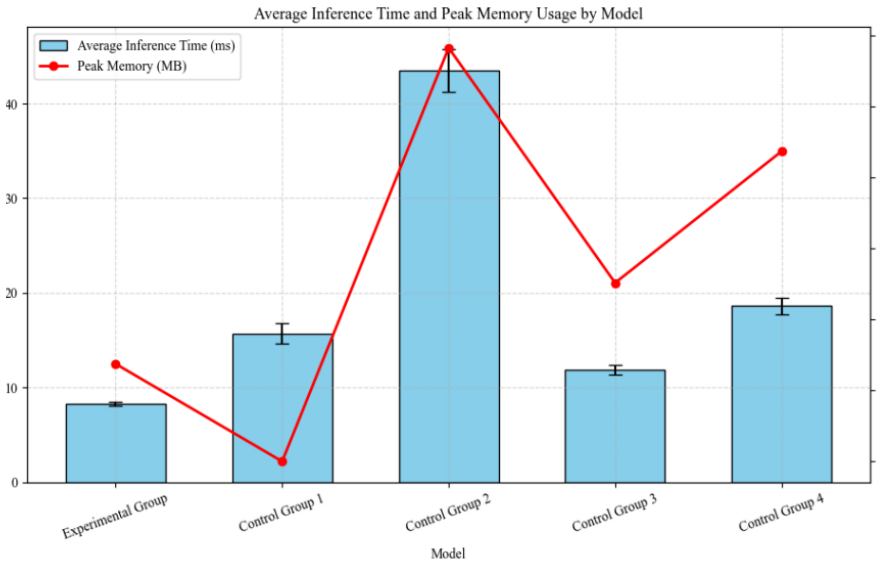

FIGURE 4. Inference performance results of each model under standard configuration

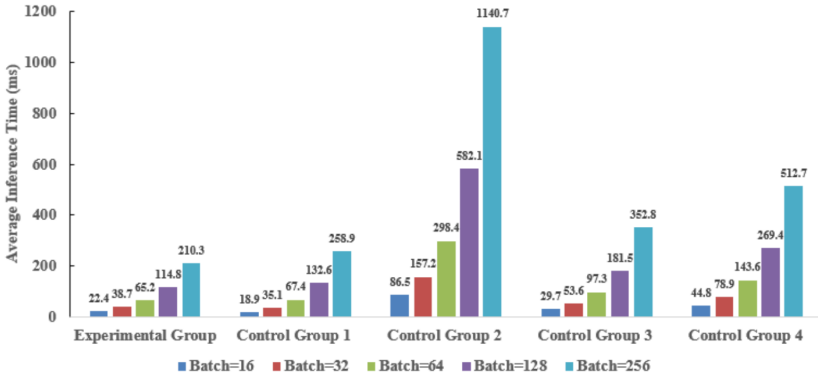

FIGURE 5. Average inference time changes of each model under different batch sizes

TABLE 2. Comparison of average convergence rounds and Loss standard deviation in each model training stage

Model	Average Convergence Rounds	Loss Standard Deviation
Experimental Group	15.3	0.021
Control Group 1	3.0	0.000
Control Group 2	28.7	0.156
Control Group 3	19.5	0.087
Control Group 4	22.1	0.103

an average of 15.3 training rounds, and the standard deviation of training loss is only 0.021, showing good training efficiency and stability, thanks to its convolutional architecture that reduces parameter coupling and gradient propagation path optimization. In contrast, due to the complexity of parameter interactions in the deep self-attention mechanism, Control Group 2 requires an average of 28.7 rounds to complete convergence, and its loss standard deviation of 0.156 further reveals the significant impact of gradient instability and local extreme value interference during training. Although Control Group 1 converges quickly within 3.0 rounds and the loss standard deviation is zero, its strictly convex optimization characteristics are only applicable to linearly separable scenarios and cannot reflect the typical training behavior of deep learning models. Control Group 3 shortens the average convergence rounds to 19.5 rounds by replacing attention calculations with Fourier transforms. The loss standard deviation of 0.087 shows that although frequency domain operations can improve gradient flow, there is still a defect in that the volatility of parameter updates is higher than that of pure convolutional architectures. The average convergence speed of 22.1 rounds and the loss standard deviation of 0.103 in Control Group 4 confirm the limited improvement of the stability of attention mechanism

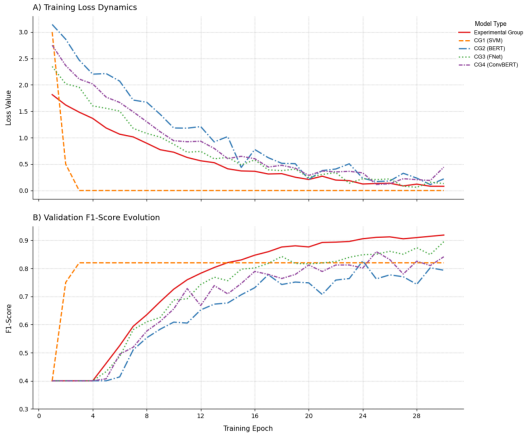

FIGURE 6. Comparison of Loss and Verification F1 Curves at Each Model Training Stage

training by the convolution module in the hybrid architecture and the complexity of the optimization path caused by its parameter redundancy still restricts the training efficiency.

In order to intuitively present the performance evolution trend of the model during the training process, this study plots the changes in the Loss curve and the verification F1 curve of each model, as shown in Figure 6.

As shown in sub-figure 6, the dynamic visualization features of the training of each model are significantly correlated with its algorithm architecture and optimization mechanism. The experimental group model shows a steep initial decline curve in the training loss sub-figure A. Its loss value drops rapidly from 1.815 in the first round and stabilizes after the 15th round. This rapid convergence feature is due to the smooth gradient flow generated by the local feature extraction mechanism of the convolution layer and the effective control of parameter coupling by the simplified structural design. In sharp contrast, the loss curve of Control Group 2 shows obvious high-level oscillations, which is due to the gradient vanishing problem of the self-attention mechanism in the deep Transformer architecture and the parameter interference effect between multi-head attention modules. It is worth noting that the loss value of Control Group 1 model is reset to zero in the third round, and its verification F1 value is simultaneously stabilized at 0.82, which verifies the deterministic convergence characteristics of the convex optimization model but also exposes the bottleneck of traditional methods in feature expression capabilities - although the training process is efficient and stable, its final performance ceiling is insufficient.

Observing sub-graph B, the F1 value of the experimental group model exceeds 0.84 in the 16th round and kept rising steadily, indicating that its optimization process has good generalization consistency, which is attributed to the translation invariance of the convolution kernel, which enhances the model's robustness in capturing local semantic patterns of text. However, the F1 curve of Control Group 2 continues to fluctuate after reaching 0.827, revealing the side effects of the attention mechanism's sensitivity to position, especially the gradient instability when dealing with long-distance dependencies. The F1 value of Control Group 3 increases faster than that of Control Group 2, but the plateau after the 20th round indicates that the frequency domain analysis method has an efficiency bottleneck in global feature fusion. Control Group 4 shows a unique dual-stage optimization feature: the F1 value rises rapidly in the first 15 rounds and then enters a slow improvement stage, which reflects the acceleration effect of the convolution module in the initial feature screening stage and the necessity of the attention mechanism in the fine-tuning stage.

V. CONCLUSION

To address the semantic understanding and generalization limitations of traditional methods, this study proposes a lightweight CNN-based pre-training model based on the BERT-Free architecture to enhance the intelligent identification of power safety risks involving specialized insulation and flame-retardant fiber materials. This approach accurately captures the technical characteristics of high-performance protective fabrics within power systems, ensuring a more robust and accurate safety assessment across these interconnected vertical fields. Experiments show that the model can effectively improve the recognition efficiency on large-scale power text data and show high practicality. It has made significant progress in semantic understanding and generalization capabilities compared with traditional methods, especially when dealing with class imbalance problems, showing good robustness and generalization performance and successfully breaking through the generalization bottleneck of traditional methods when facing new equipment and new hidden dangers. At the same time, the model also performs well in reasoning efficiency and deployment flexibility, which can meet the power industry's needs for lightweight and real-time systems and effectively improve the problems of high computing resource consumption and high reasoning latency in traditional Transformer and BERT pre-trained models. However, the model still has certain limitations. It mainly relies on a single text modality and has not yet fully utilized heterogeneous data such as images and sensors in the power industry, which, to some extent, limits the model's adaptability in more complex scenarios. In addition, although the model has lightweight features, real-time deployment on extreme edge devices integrated into smart protective wearable sensing fabrics still faces the dual challenges of inference latency and resource occupation. Optimizing these algorithms for fiber-embedded monitoring systems remains essential to ensure the immediate detection of power safety hazards .

Future research can be further expanded from the following three aspects: First, multi-modal data, including images, time series, sensors, etc., are applied to enhance the model's comprehensive perception of risk events; second, the model is applied to other power text intelligent analysis tasks, such as accident tracing, dispatch instruction parsing, and risk chain tracking; third, the model structure and

reasoning mechanism are continuously optimized to further improve its practicality and response speed in online early warning and edge deployment scenarios so as to better meet the growing needs of intelligent and safe operation and maintenance in the power industry.

AUTHOR CONTRIBUTIONS

Conceptualization –Siwu Yu, Yumin He, Guobang Ban, Teng Gui, Xiangquan Hu, Yang Mei and Jingjing Zhang; methodology – Siwu Yu, Yumin He, Guobang Ban, Teng Gui, Xiangquan Hu, Yang Mei and Jingjing Zhang; investigation – Siwu Yu, Yumin He, Guobang Ban, Xiangquan Hu, Yang Mei and Jingjing Zhang; writing-original draft preparation – Siwu Yu, Yumin He, Guobang Ban, Teng Gui, Xiangquan Hu, Yang Mei and Jingjing Zhang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by the Science and technology project of China Southern Power Grid Co., Ltd. (No.GZKJXM20232503).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] N. D. Fîță, S. M. Radu, and D. Păsculescu, "The resilience of critical infrastructures within the national energy system in order to ensure energy and national security," Bulletin of" Carol I" National Defence University, vol. 11, no. 2, pp. 57-67, 2022, doi: 10.53477/2284-9378-22-75.

[2] A. Mohanty, A. K. Ramasamy, R. Verayiah, S. Bastia, S. S. Dash, E. Cuce, et al., "Power system resilience and strategies for a sustainable infrastructure: A review," Alexandria Engineering Journal, vol. 105, pp. 261-279, 2024, doi: 10.1016/j.aej.2024.06.092.

[3] J. J. Moreno Escobar, O. Morales Matamoros, R. Tejeida Padilla, I. Lina Reyes, and H. Quintana Espinosa, "A comprehensive review on smart grids: Challenges and opportunities," Sensors, vol. 21, no. 21, pp. 6978, 2021, doi: 10.3390/s21216978.

[4] M. Kiasari, M. Ghaffari, and H. H. Aly, "A comprehensive review of the current status of smart grid technologies for renewable energies integration and future trends: The role of machine learning and energy storage systems," Energies, vol. 17, no. 16, pp. 4128, 2024, doi: 10.3390/en17164128.

[5] S. Akhtar, M. Adeel, M. Iqbal, A. Namoun, A. Tufail, and K. H. Kim, "Deep learning methods utilization in electric power systems," Energy Reports, vol. 10, pp. 2138-2151, 2023, doi: 10.1016/j.egyr.2023.09.028.

[6] W. Weerakkody, B. Rathnayaka, and C. Siriwardana, "Identifying and Prioritizing Critical Risk Factors in the Context of a High-Voltage Power Transmission Line Construction Project: A Case Study from Sri Lanka," CivilEng, vol. 5, no. 4, pp. 1057-1088, 2024, doi: 10.3390/civileng5040052.

[7] Q. Liu, V. Hagenmeyer, and H. B. Keller, "A review of rule learning-based intrusion detection systems and their prospects in smart grids," IEEE Access, vol. 9, pp. 57542-57564, 2021, doi: 10.1109/ACCESS.2021.3071263.

[8] Z. Han, Y. Li, Z. Zhao, and B. Zhang, "An Online safety monitoring system of hydropower station based on expert system," Energy Reports, vol. 8, pp. 1552-1567, 2022, doi: 10.1016/j.egyr.2022.02.040.

[9] P. G. George and V. R. Renjith, "Evolution of safety and security risk assessment methodologies towards the use of bayesian networks in process industries," Process Safety and Environmental Protection, vol. 149, pp. 758-775, 2021, doi: 10.1016/j.psep.2021.03.031.

[10] A. M. Stanković, K. L. Tomsovic, F. De Caro, M. Braun, J. H. Chow, N. Čukalevski, et al., "Methods for analysis and quantification of power system resilience," IEEE Transactions on Power Systems, vol. 38, no. 5, pp. 4774-4787, 2022, doi: 10.1109/TPWRS.2022.3212688.

[11] S. Zhang, Z. Zhu, and Y. Li, "A critical review of data-driven transient stability assessment of power systems: principles, prospects and challenges," Energies, vol. 14, no. 21, pp. 7238, 2021, doi: 10.3390/en14217238.

[12] S. Gao, R. Takanobu, A. Bosselut, and M. Huang, "End-to-end task-oriented dialog modeling with semistructured knowledge management," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 2173-2187, 2022, doi: 10.1109/TASLP.2022.3153255.

[13] Q. Li, J. Li, L. Wang, C. Ji, Y. Hei, J. Sheng, et al., "Type information utilized event detection via multi-channel gnns in electrical power systems," ACM Transactions on the Web, vol. 17, no. 3, pp. 1-26, 2023, doi: 10.1145/3577031.

[14] Y. Shu and Y. Dai, "An effective link prediction method for industrial knowledge graphs by incorporating entity description and neighborhood structure information," The Journal of Supercomputing, vol. 80, no. 6, pp. 8297-8329, 2024, doi: 10.1007/s11227-023-05767-2.

[15] J. Ortiz-Bejar, M. R. A. Paternina, A. Zamora-Mendez, L. Lugnani, and E. Tellez, "Power system coherency assessment by the affinity propagation algorithm and distance correlation," Sustainable Energy, Grids and Networks, vol. 30, Art. no. 100658, 2022, doi: 10.1016/j.segan.2022.100658.

[16] S. Liu, S. You, Z. Lin, C. Zeng, H. Li, W. Wang, et al., "Data-driven event identification in the US power systems based on 2D-OLPP and RUSBoosted trees," IEEE Transactions on Power Systems, vol. 37, no. 1, pp. 94-105, 2021, doi: 10.1109/TPWRS.2021.3092037.

[17] G. Wu and Y. Zheng, "The construction of Bert fusion model of speech recognition and sensing for South China electricity charge service scenario," EURASIP Journal on Advances in Signal Processing, vol. 2023, no. 1, pp. 113, 2023, doi: 10.1186/s13634-023-01073-4.

[18] X. Tong, B. Jin, J. Wang, Y. Yang, Q. Suo, and Y. Wu, "MM-ConvBERT-LMS: detecting malicious web pages via multimodal learning and pre-trained model," Applied Sciences, vol. 13, no. 5, pp. 3327, 2023, doi: 10.3390/app13053327.

[19] Z. Li, J. Yan, Y. Liu, W. Liu, L. Li, and H. Qu, "Power system transient voltage vulnerability assessment based on knowledge visualization of CNN," International Journal of Electrical Power & Energy Systems, vol. 155, Art. no. 109576, 2024, doi: 10.1016/j.ijepes.2023.109576.

[20] A. Quawi, Y. M. Shuaib, and M. Manikandan, "CNN based fault event classification and power quality enhancement in hybrid power system," International Journal of Power Electronics and Drive Systems (IJPEDS), vol. 15, no. 3, pp. 1851-1870, 2024, doi: 10.11591/ijpeds.v15.i3.pp1851-1870.