scientific reports

Check for updates

OPEN

Research on multi-branch residual connection spectrum image classification based on attention mechanism

Zhong Xiaohui[1], Dong Sheng[1], Zhang Yiyi[1], Lu Wei[1] & Jiang Lincen[2]✉

The acoustic spectrogram arranges the frequencies in the sound along the frequency spread, and translates the spectral changes into the intensity, wavelength and frequency of the electrical signals. Currently, the extensive use of convolutional neural networks for spectral image classification can extract signal features in the spectrogram, but the redundancy of noisy data generated by a large number of bands of the spectrum affects the feature information at different levels of the image. In order to optimize this problem, this paper proposes a multi-branch residual-connected Efficient Global Attention (EGA) acoustic spectral image classification network based on the attention mechanism, which firstly separates the components with their respective acoustic features from the spectral noise, so as to achieve the purpose of noise reduction, and then extracts the Phase Resolved Partial Discharge (PRPD) Spectrum of the Intermediate Frequency (IF) cycle for the original signals that have undergone noise reduction, which is based on the attention mechanism through the Improved Global Attention Mechanism (IGAM) in the EGA of the backbone network. mechanism pays more attention to the channel and spatial features of the spectrogram, then improves the feature extraction ability by residual connection, and finally performs feature fusion with the mask branch. The results show that a more accurate detection of abnormal partial discharge type of carbon brushes in gantry cranes is made, and the feasibility and innovativeness of the method is verified through experiments and production use.

Keywords Image classification, Attention mechanism, Residual connection, EGA, Feature fusion

Voice print is the general name of the speech features contained in speech, which can represent and identify the sound source, and the speech model based on these features. The voiceprint can be broken down into different wavelengths, frequencies, intensives and rhythms of different sounds, and the characteristic information of these frequencies and intensives can be integrated to obtain the speech spectrum of a sound. Speech spectrum is a kind of acoustic information that can represent the characteristics of voiceprint, and the spectrum parameters can reflect the distribution of energy in different frequency bands.

A power supply device for sliding touch line of gantry crane comprises a power supply device body, which comprises a U-shaped roller fixed support with a central cavity, an insulating layer, a roller and a carbon brush assembly. The carbon brush plays a very important role in the motor, the carbon brush conducts the current between the moving parts of the motor, and the carbon brush also changes the direction of the current, that is, the role of reversing. However, while meeting the ease of use, the carbon brush of this structure will adsorb a lot of tiny particles in the air under certain specific climate conditions. When the carbon brush passes by, attachments will be formed on the surface of the carbon brush to form carbon deposits, as shown in Fig. 1, resulting in a smaller contact surface between the carbon brush and the sliding contact line, and a larger instantaneous current passing through, resulting in safety hazards, easy breakdown of wire equipment, and circuit fire.

In recent years, with the rapid development of artificial intelligence and acoustic signal recognition technology, carbon brush fault detection methods, mainly based on deep learning abnormal voiceprint detection, have made some new progress in model optimization and data processing. The authors of reference[1] converted the pre-processed voice print signals into Mel spectrum, and used convolutional neural networks to classify various fault features, which can effectively identify the abnormal state of power transformers. However, the method

[1]Ningbo Beilun Third Container Terminal Co., LTD, Ningbo 315000, Zhejiang, China. [2]School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. ✉email: 2022010301@njupt.edu.cn

Fig. 1. Carbon brush with heavy carbon accumulation.

heavily depends on labeled data and does not incorporate adaptive mechanisms to handle varying background noise, which limits its applicability in real-world noisy scenarios. Based on gamma-frequency cepstrum (GFCC) voice print spectrum, the authors of reference[2] classified the voice print signals of 10 kV dry transformers under different faults. This method integrates a capsule attention network to enhance spatial feature extraction and address class separability in hyperspectral image classification. While effective, this method is highly sensitive to the choice of feature representation and lacks robustness when the spectral characteristics are degraded due to external noise interference. Literature[3] proposes a highly generalized fan blade anomaly detection method based on voice print, which is based on the clustering and median convergence of periodic audio cutting methods to effectively cut the voice print, and uses the steady-state difference method between the three blades of the wind turbine to detect anomalies, avoiding the problem of algorithm migration failure caused by changes in objects to be detected and channels. The method also utilizes wavelet pooling and graph-based attention to enhance classification robustness under small object tracking conditions. Nonetheless, its reliance on handcrafted signal segmentation rules and assumptions about signal periodicity make it less flexible for non-periodic or transient signals such as carbon brush voiceprints. In literature[4], hierarchical thresholds are used to denoise acoustic signals, and an algorithm for wavelet packet extraction of acoustic signal feature vectors measured by electrical equipment is presented. The tracking framework leverages query-guided attention to handle occlusion, combining detection and re-identification. The method provides interpretable signal components, but its feature extraction is shallow and not adaptive to the hierarchical semantic features that deep models can learn, leading to limited classification capacity. Literature[5] analyzed the noise characteristics generated by the UAV itself and its impact on the sound source. A consistent representation mining strategy is applied using cross-drone graph representations to enhance long-term object consistency. The signal-to-noise ratio could be improved through noise separation technology and wavelet packet decomposition and reconstruction, and the rationality of partial discharge detection technology based on acoustic diagnosis was verified through the detection of abnormal noise of the line tower. Li, Y et al. proposed a data enhancement method for short-term voltage stability assessment[6], which achieved good results in processing small sample data sets. Although successful in improving performance on limited data, this method was designed for voltage signals and lacks validation on complex multi-class voiceprint data. Sofia, M.d.A.L et al. adopted the over-sampling technique to balance the DGA data of power transformers, and the effect was significantly improved compared with the original data set[7]. The above methods have achieved good expansion effect on unbalanced data. However, due to the changeable background environment, the noise data generated by a large number of bands of the voiceprint spectrum is redundant, which affects the feature information of different levels of the image, and the recognition rate is not high, which can not accurately classify the voiceprint spectrum image and accurately predict the fault.

Therefore, this paper proposes an Efficient Global Attention (EGA) voire spectrum image classification network based on multi-branch residual connection of attention mechanism, which can detect carbon deposition in advance, and can detect carbon deposition in advance to the extent that there is slight carbon deposition but it has not caused production faults, so as to eliminate hidden dangers. The main contributions of the proposed method are described as follows:

(1) In order to better capture image context information, a multi-branch residual connection EGA classification network is proposed.

(2) In order to pay more attention to the waveform features of the whole spectral graph while paying more attention to the local features, an Improved Global Attention Mechanism (IGAM) is proposed.

(3) The proposed model can identify early-stage carbon deposition—before it leads to production faults—thus supporting proactive maintenance and safety assurance in power equipment.

Voiceprint detection

Voiceprint detection is a kind of authentication technology based on voice characteristics to verify the identity of an individual by analyzing and comparing his voice. Through the use of various features of the sound signal, such as frequency, amplitude and tone characteristics, is the principle of voicing detection. The main step is to convert these features into digital signals and then carry out comparative analysis. In recent years, a lot of research has been devoted to tasks such as audio classification and speech recognition.

The feature extraction of voiceprint based on deep learning includes input, network structure, time pool strategy and objective function, which are the basic components of the voiceprint recognition subtask[8]. The realization of voiceprint detection first needs to collect personal voice samples, and then carry out feature extraction on the collected voice samples to extract features such as frequency, amplitude and tone. Then, the extracted features are modeled to generate a voiceprint model. Finally, the voice sample to be verified is compared with the voiceprint model to determine whether it matches. CNN-based models have been used for a variety of tasks, including music genre classification[9] and environmental sound classification[10]. In addition, processing raw audio waveforms using the EnvNET model is a rare example of using raw audio as input[11]. Most studies obtain SOTA results from CNNS on spectrographs, complicating designs with multiple models that take different inputs and then aggregate their outputs to make predictions. For example, three networks are used to manipulate raw audio, spectral delta STF coefficients[12]. It can be seen from the existing tasks that audio transmission learning is mainly focused on audio data sets. The features used by the model are very large, and the features used are becoming more and more complex, so a task of audio classification for a pre-trained model on ImageNet is proposed[13].

Transfer learning for voiceprint detection

Voiceprint detection refers to the analysis and processing of audio signals to extract the feature information. Among them, beat detection is an important application of audio detection, it can detect the rhythm and beat in audio, and convert them into digital signals, so as to facilitate subsequent processing and application. Common beat detection algorithms include time-domain algorithms, frequency-based algorithms and mixed-domain algorithms. There are also machine learning-based algorithms that can be used for tasks such as audio event detection and classification.

Transfer learning is a method that extends a model trained on a particular task with a large amount of data to another task and extracts useful features for the new task based on the prior knowledge of the new task. In recent years, depth models trained in classification on large corpora such as ImageNet have been widely used in transfer learning for tasks such as image segmentation[14]. In the video model, pre-training on ImageNet and dynamics data sets achieved 98% performance.

Transfer learning in audio detection can be implemented in many ways, and it mainly focuses on pre-training the model on a large number of audio datasets, such as audio datasets, million song datasets. Some studies pre-train a simple CNN network on a dataset of millions of songs and find that they can perform various tasks on these networks, such as audio event classification and emotion prediction[15]. There are also researchers trying to use large models such as VGG and ResNet for audio classification on audio sets[16] and the models they trained are also used in many audio transmission learning applications[17].

In conclusion, transfer learning can help models in audio detection tasks to make better use of existing data and knowledge, thus improving the performance of models.

Attention mechanism

The core idea of attention mechanism is from focus to focus, attention mechanism was first used in computer vision, and later in the field of Natural Language Processing (NLP) gradually began to apply. At present, attention mechanism is a widely used mechanism in the field of deep learning, most of which operate under the framework of Encoder-Decoder. On the left side of Fig. 2 is the encoder, which converts the input text into a vector representation of the text; On the right side of the diagram is the decoder, which converts the vector representation of the decoder's input into an output text sequence.

In order to consider the interrelation between visual features and audio features, multi-head attention mechanism is adopted, combined with Conv-TasNet, a convolution time-domain separation model, and DPRNN, a dual-path recurrent neural network, to propose a multi-head time-domain audiovisual speech separation model MHATD-AVSS[18]. In order to solve the problem of low recognition accuracy due to underutilization of speech features, an attention mechanism is introduced in the framework of deep neural networks, and a residual network constrains source features and target features is added, thus minimizing the cepstrum distortion of target features[19]. Some researchers have proposed a Global Attention Mechanism (GAM) to improve the performance of deep neural networks by reducing information reduction and amplifying global interactive representations[20].

Datasets and methods

In this section, we first introduce Power private data set and the noise reduction preprocessing method for gantry carbon brush voiceprint data through wavelet packet decomposition, and then introduce our two-branch

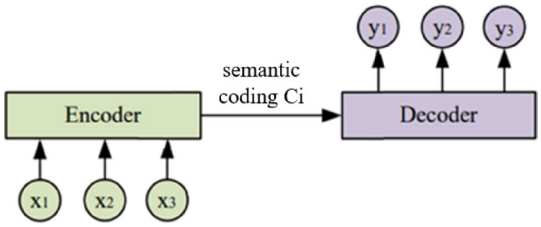

Fig. 2. Encoder-Decoder framework for introducing attention mechanisms.

Efficient Global Attention (EGA) network in detail. Finally, the Improved Global Attention Mechanism (IGAM), which focuses on both channels and Spaces, is introduced.

Power private data set

In this paper, the experiment is carried out on the private data set of power system, from which the sound characteristics of abnormal noise are extracted and processed, and the good performance is shown.

In the experiment, the wavelet packet decomposition method is introduced to process the data, which is a signal decomposition method based on wavelet analysis, and can be used for signal compression, denoising, feature extraction and so on. The main step is to break the signal down into multiple components, each of which contains information over a certain frequency range. Then, through the analysis and processing of different components, the signal energy and characteristics in different frequency ranges can be obtained, and the signal components satisfying the correlation of their sound characteristics can be extracted. After processing, the spectrum diagram of the related sound is obtained, and the components of the sound can be seen directly.

In order to improve the model effect, this paper adds CClotho data set for audio detection. The CClotho data set consists of three parts: development set, verification set and evaluation set, among which there are 3839, 1045 and 1045 audio signals respectively. Audio signals in the Clotho dataset range in length from 30 s to less, and each audio signal is labeled with five labeled natural language text summaries[21]. The ESC-50 dataset consists of 2000 clips belonging to 50 classes, each 5s in length. These segments were sampled at a uniform rate of 44. 1 kHz. The data set is formally divided into five fold and the accuracy is calculated by cross-validation of all folds. This ESC-50 is made up of ambient sounds, from birds to car horns[22]. The City Sounds 8k dataset contains 8,732 fragments belonging to 10 different categories of city sounds. The length of each audio clip is <=4s, and the sampling rate varies from 16 kHz to 44.1 kHz[23].

Data noise reduction preprocessing

Carbon brush anomalies occur in various scenes and are easily disturbed by environmental sounds. Airborne solutions also receive strong interference from drone noise. The carbon accumulation, grinding and abnormal discharge of carbon brush will produce their own sound characteristics, and it is necessary to separate the components with their own sound characteristics from the noise of drones and other noise, so as to achieve the purpose of noise reduction.

As shown in Fig. 3, wavelet packet decomposition is an important tool for component analysis. The signal is decomposed into multiple components through the wavelet packet decomposition method, and the decomposed results are shown in Fig. 4.

The optimal wavelet packet tree search method is introduced here. By calculating the feature-related components of the signal and the continuous entropy of the discrete information source, the signal components satisfying the correlation of the respective sound features are extracted to carry out signal reconstruction. The Phase Resolved Partial Discharge (PRPD) spectrum containing the respective sound characteristics can be obtained.

The original signal after transformation is highly aggregated in the feature space, which reduces the dependence on large-scale neural network, has good interpretability, controllable model performance, and reduces the possibility of Bias in the model.

EGA network module

When the spectrum diagram is entered into EGA network, the backbone network passes through the initial residual unit, and then enters the main branch and mask branch respectively. The basic modules of the main branch are Fused-MBG Conv and MBGConv, and the residual unit is used for cross-layer connection. The mask branch is mainly composed of several residual units, as shown in Fig. 5. Among them, the initial residual unit is helpful to process the spectrum diagram; The main module of the main branch is a specially designed convolution operation, which has the characteristics of high efficiency and lightweight. Both modules also combine the ideas of deep separable convolution and residual connection to realize effective feature extraction and information fusion. Multiple residual units of the mask branch are used to further refine and process specific information to improve the characteristics of the main branch.

In the EGA network, the mask branch adopts a bottom-up structure to learn the mask M(x) of the same size, and the main branch outputs the feature T(x). The bottom-up structure mimics the attention process of fast feedforward and feedback. The output mask is used as the control gate of the main branch neuron, and the

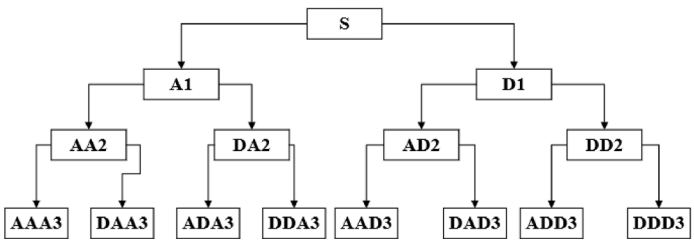

Fig. 3. Wavelet packet decomposition components.

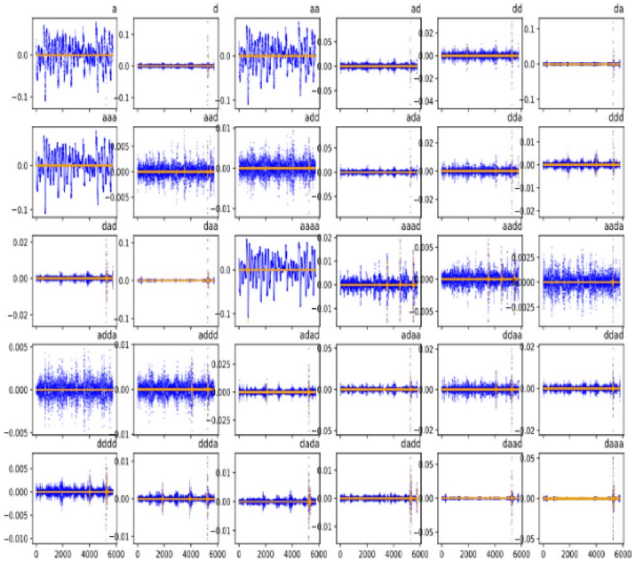

Fig. 4. Voicing PRPD spectrum.

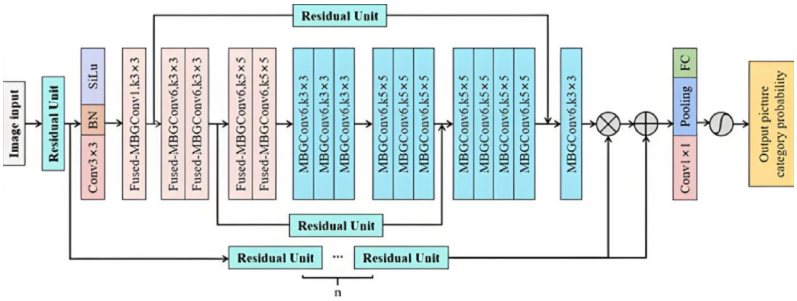

Fig. 5. EGA network.

output result of the mask branch is matrix multiplied with the output result of the main branch respectively. The output H is:

$$H_{(i,c)}(x) = M_{(i,c)}(x) * T_{(i,c)}(x)$$

Where I is valued at all spatial locations, c∈{1,… C} is the index of the channel. The entire structure can be trained end to end. However, masking the point generation from 0 to 1 reduces the value of deep features. Second, the mask may destroy the good characteristics of the trunk branch, and if the mask branch can be constructed as the same mapping, then the performance should not be worse than the corresponding unit without attention. Therefore, we modify the output H of the attention module to:

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * F_{i,c}(x)$$

The main branch in the EGA network proposed in this paper can be used as a feature selector in the forward inference process and as a gradient update filter in the backpropagation process. In the mask branch, the mask gradient of the input feature is:

$$\frac{\partial M(x,\theta)\,T(x,\phi)}{\partial \phi} = M(x,\theta)\,\frac{\partial T(x,\phi)}{\partial \phi}$$

Where θ is the parameter of the mask branch and φ is the main parameter of the dry branch. This attribute network is robust to noise labels. Inspired by Efficient network[24], to optimize and improve Efficient network, we replaced the first three layers of MBGConv module with Fused-MBConv module as shown in Fig. 6, which significantly improved the training speed.

At the same time, in order to further improve the representation ability of the model, we add a skip layer connection to the main branch network, and add a residual unit to the skip layer connection. The network increases the ability of multi-layer feature fusion and improves the representation ability of the network. The

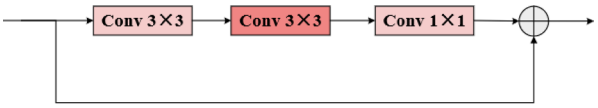

Fig. 6. Fusion-MB Conv module.

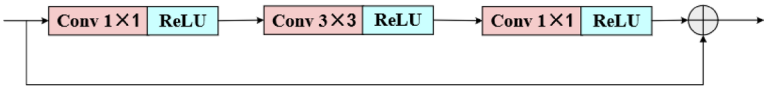

Fig. 7. Residual cell.

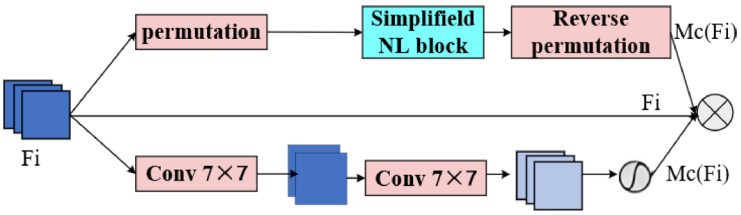

Fig. 8. IGAM attention mechanism.

mask branch includes fast feedforward scanning and top-down feedback steps, the former quickly collects the global information of the whole image, the latter combines the global information with the original feature map. In a convolutional neural network, these two steps unfold into a bottom-up and top-down full convolutional structure, which generates a soft weight mask by superimposing multiple residual units, as shown in Fig. 7.

Starting from the input, max-pooling is performed several times to rapidly increase the receptive field after a small number of residual units. After the minimum resolution is reached, the global information is extended through a symmetric top-down architecture to guide the input features at each location. Linear interpolation outputs some residual units after upsampling. The amount of bilinear interpolation is the same as the maximum pooling to keep the output size the same as the input feature map. Then, after two successive 1×1 convolution layers, an S-shaped layer normalizes the output range to [0,1]. On the power private dataset, we set up a stack of $n = 3$ residual cells while using a single mask branch would require an exponential number of channels to capture all combinations of different factors.

IGAM attention mechanism module

Considering that carbon deposition, grinding and local emission anomalies of the carbon brush will produce their own sound characteristics, we need to pay attention to the local characteristics and the waveform of the entire spectrum. Therefore, we propose the IGAM attention mechanism in the MBGConv module. When the feature graph F1 is input, the two modules are passed simultaneously, and the output is defined as:

$$F_i = \alpha_i M_c(F_i) \otimes \beta_i M_c(F_i) \otimes F_i$$

Among them, the weight α_i and β_i are set for the channel attention submodule Mc and the spatial attention submodule respectively, and i represents the network layer where the attention mechanism is located. After this improvement, the reasoning speed of the attention mechanism is improved, and the attention degree of different channel attention submodules and spatial attention subblocks are set for different network layers, which helps to better improve the utilization rate of the attention mechanism. To further improve the performance of the attention mechanism on the channel, we replaced the MLP of the channel attention submodule with a Simplified non-local NL block, as shown in Fig. 8.

In the channel attention submodule, the permutation modifies the channel order by the input feature Fi, which is then input to a simplified non-local block, and then the output is inverse permutation back to output Mc(Fi), which is then weighted by α_i. In the spatial attention submodule, the input feature Fi is modified by convolution operation to reduce the number of channels and reduce the calculation amount, then a convolution operation is performed to increase the number of channels, and finally Sigmod is performed to output Ms(Fi). As shown in Fig. 9, the IGAM attention mechanism is introduced into the MBConv module and the Fissed-MBConv module, the MBGConv module.

The MBGConv module that introduces IGAM attention mechanism not only simplifies non-local blocks to aggregate information from other locations, but also improves the ability of global information aggregation. Moreover, by enhancing the features of the query location to screen out the most informative features, we have implemented a more efficient feature extractor and improved the overall network performance.

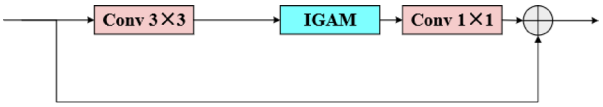

Fig. 9. MBGConv module.

Method	F1-score
MLP-CW[25]	0.57
BGNet[26]	0.53
BGNet-SE3[27]	0.57
MSFF-Net[28] AudioFormer[29] EAViT[30]	0.62 0.62 0.64
EGA	0.66

Table 1. Comparative experimental results.

Method	F1-score
baseline	0.49
baseline + SE	0.57
baseline + IGAM	0.66
baseline + SE + IGAM	0.65

Table 2. Ablation results.

Results

Experimental setup

To validate the effectiveness of the proposed EGA network in the audio classification task, experiments were conducted on a private power system dataset. The dataset was randomly split into training, validation, and test sets in a ratio of 8:1:1. The F1-score, which considers both precision and recall, was selected as the primary evaluation metric due to its robustness in handling class imbalance. All experiments were implemented on a workstation equipped with Ubuntu 20.04, Python 3.9, PyTorch 2.0.1, CUDA 11.6, and an NVIDIA GeForce RTX 3090 GPU. The model was trained using the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 1e-4. The learning rate was dynamically adjusted using a ReduceLROnPlateau scheduler, with a patience of 5 epochs and a decay factor of 0.5. We trained the model for 100 epochs with a batch size of 128. Cross-entropy loss was used as the objective function to guide the classification task. In addition, data augmentation techniques such as random time masking and frequency masking were applied to improve generalization and reduce overfitting. During training, the model with the highest F1-score on the validation set was saved and used for final evaluation on the test set.

We use the cross-entropy loss function to train the proposed EGA network for the classification task. Given the true label y and the predicted probability distribution $\widehat{y}$, the loss is defined as:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{C} y_i \cdot \log\left(\widehat{y}_i\right)$$

This loss effectively measures the difference between the predicted class distribution and the ground truth, and is widely adopted in multi-class classification problems.

Experimental result

In order to verify the universality of the method proposed in this research scheme, four audio classification algorithms were selected as comparative experimental objects, namely MLP-CW, BGNet, BGNET-SE3, MSFF-Net, AudioFormer and EAViT as shown in Table 1. The experiment shows that the proposed EGA network can improve the accuracy of audio classification compared with the other four methods, which further proves the validity and rationality of the proposed network.

Ablation experiment

In order to verify the validity and rationality of the proposed method, the ablation experiment of EGA voiceprint detection network was carried out in Table 2. Firstly, the ablation experiment verified the effectiveness of adding SE module, which significantly improved the performance of the neural network with only a few additional parameters. Then, it is verified that the IGAM module enhances the information exchange between the feature

channels. The addition of IGAM significantly improves the performance of the audio classification network compared to the original Efficient network and the Efficient network with the addition of the SE module.

Visual experimental results

In the practical application scenario, EGA network is applied to the airborne acoustic detection equipment independently developed by electric power, mounted on the UAV for flight detection, which can detect the voice print generated by abnormal carbon brush of the gantry crane, and has the acoustic imaging function, which can visually locate the abnormal position of the carbon brush, detect whether there is a partial discharge type, and indicate the corona discharge or surface discharge type. As shown in Fig. 10.

Robustness analysis in multi-source environments

To further verify the robustness and generalization ability of the proposed EGA network in real-world scenarios, we conducted additional experiments under simulated multi-source acoustic environments. In practical industrial settings such as UAV-based inspections of gantry cranes, it is common to encounter overlapping sound signals from various sources, including background noise, drone rotor sound, and unrelated mechanical activity. These acoustic interferences may potentially impact recognition performance. To simulate such conditions, we constructed a mixed dataset by adding interfering sounds (e.g., ambient machine noise, wind, drone noise) to the original test set with varying signal-to-noise ratios (SNRs): 20 dB, 10 dB, and 0 dB.

As shown in Table 3, the performance of all models decreased as the level of noise interference increased. However, the proposed EGA network maintained a relatively stable F1-score, especially under 0 dB SNR conditions, outperforming the baseline by a margin of 25% points. The improvement demonstrates that the EGA architecture, particularly the IGAM attention mechanism, enables better extraction of discriminative features and suppresses irrelevant noise.

This result confirms the strong robustness of the EGA network in environments with multiple sound sources and validates its practical value in real-world airborne acoustic detection applications.

Conclusions

In this paper, a kind of EGA network based on the attention mechanism is proposed and the related research is carried out. In this method, wavelet packet decomposition is used to pre-process the private voice print data set, and the characteristic PRPD map of power frequency period phase resolution is extracted from the original signal after noise reduction. Through IGAM attention mechanism, EGA network pays more attention to the channel and spatial features of the spectral graph, realizes more efficient feature extraction, and makes more accurate detection of abnormal partial discharge types of carbon brush. In this paper, the feasibility and innovation of this method are verified by ablation experiment and comparison experiment. Experimental results show that this network model can effectively classify spectral images and improve the performance of voiceprint anomaly detection.

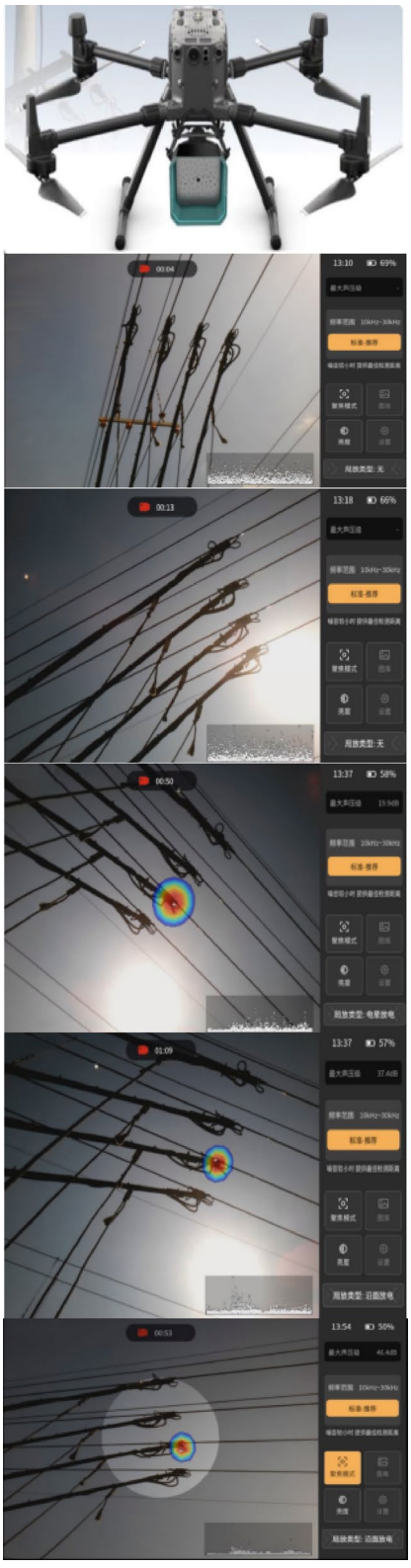

Fig. 10. Visualization of abnormal types of voiceprint detection.

Method	Clean (∞ dB)	SNR 20dB	SNR 10dB	SNR 0dB
baseline	0.49	0.43	0.34	0.21
baseline + IGAM	0.66	0.61	0.55	0.42
baseline + SE + IGAM	0.65	0.63	0.58	0.46

Table 3. F1-score under different SNR levels.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due our dataset comes from the application of electronic systems. Electric utilities are commercial for-profit companies, but are available from the corresponding author Jiang lincen (2022010301@njupt.edu.cn) on reasonable request.

Received: 9 July 2024; Accepted: 9 July 2025
Published online: 15 July 2025

References

1. Dang, X., Wang, F. & Ma, W. Fault Diagnosis of Power Transformer by Acoustic Signals with Deep Learning. In Proceedings of the 2020 IEEE International Conference on High Voltage Engineering and Application (ICHVE), Beijing, China, 6–10 September ; pp. 1–4. (2020).
2. Zou Yijin, L. et al. Research on anomaly detection method of highly generalized fan blade based on voice print [J]. *J. Univ. Electron. Sci. Technol. China*. **50** (05), 795–800 (2019).
3. Pan, Liangliang, Zhao Shutao1 & Aug, L. B. Electrical equipment fault diagnosis based on acoustic wave signal analysis. *Electr. Power Autom. Equip.* **29** (08), 87–90 (2009).
4. Chen, Z. et al. Partial Discharge Detection Technology for Transmission Lines Based on Drone Acoustic Diagnosis, the 3rd International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT), Nanjing, China, 2023, 2013 Pp. 2023-59, doi: 65.10 / ICEEMT1109.59522.2023.
5. Li, Y., Zhang, M. & Chen, C. A Deep-Learning intelligent system incorporating data augmentation for Short-Term voltage stability assessment of power systems. *Appl. Energy*. **308**, 118347 (2022).
6. Sofia, M. A. L., Rogério, A. F. & Ruy, A. C. A. Incipient fault diagnosis in power Transformers by data-driven models with over-sampled dataset. *Electr. Power Syst. Res.* **201**, 107519 (2021).
7. Zhongxin Bai and Xiao-Lei Zhang. Speaker Recognition Based on Deep Learning:An Overview [EB/OL]. (2021) [2023-10-30]. https://arxiv.org/abs/2012.00931
8. Dong, M. Convolutional neural network achieves human-level accuracy in music genre classification [EB/OL]. (2018)[2023-10-30]. https://arxiv.org/abs/1802.09697
9. Guzhov, A., Raue, F., Hees, J., Dengel, A. & ESResNet Environmental Sound Classification Based on Visual Domain Models[C]//*2020 International Conference on Pattern Recognition (ICPR)*. Milan, Italy, pp. 4933–4940. (2021).
10. Tokozume, Y. & Harada, T. Learning environmental sounds with end-to-end convolutional neural network[C]. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2721–2725. (2017).
11. Li, X. & Zhou, L. Adaptive recognition of Open-Set voiceprint based on deep learning combined with MSM-SVDD Method[J]. SSRN electronic journal, (2022). https://doi.org/10.2139/ssrn.4203257
12. Gwardys, G. & Grzywczak, D. M. Deep image features in music information retrieval[J]. *Int. J. Electron. Telecommunications*. **60** (4), 321–326 (2014).
13. Iglovikov, V. & Shvets, A. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation[EB/OL]. (2018) [2023-10-30]. https://arxiv.org/abs/1801.05746
14. Choi, K., Fazekas, G., Sandler, M. & Cho, K. Transfer learning for music classification and regression tasks [EB/OL].(2017) [2023-10-30]. https://arxiv.org/abs/1703.09179
15. Hershey, S. et al. Cnn architectures for large-scale audio classification [EB/OL].(2016) [2023-10-30]. https://arxiv.org/abs/1609.09430
16. Xie, H. & Virtanen, T. *Zero-shot Audio Classification Based on Class Label embeddings[C]. in 2019 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics (WASPAA)*pp. 264–267 (IEEE, 2019).
17. Boles, A. & Rad, P. Voice biometrics: Deep learning-based voiceprint authentication system[C]. *System of Systems Engineering Conference. EEE*, 2017:1–6. https://doi.org/10.1109/SYSOSE.2017.7994971
18. Lan Chaofeng, J. & Pengwei, C. Huan, et al. Audio-visual speech separation based on multi-head attention mechanism based on dual-path recursive network and Conv-TasNet[J]. *J. Electron. Inform. Technol.* :1–8[2023-12-28.
19. Cui, Z., Zhong, Y. & Jiang, S. Research on Scheduling Phone Authentication Technology Based on Deep Learning Voiceprint Recognition[C]//2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT). https://doi.org/10.1109/ACAIT60137.2023.10528455
20. Yichao Liu and Zongru Shao and Nico Hoffmann.Global Attention Mechanism. Retain Information to Enhance Channel-Spatial Interactions[EB/OL].(2021)[2023-10-30]. https://arxiv.org/abs/2112.05561
21. Drossos, K., Lipping, S. & Virtanen, T. Clotho: An audio captioningdataset[C].in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*,Barcelona, Spain, pp. 736–740. (2020).
22. Piczak, K. J. ESC: Dataset for Environmental Sound Classification [EB/OL]. [2023-10-30]. http://dl.acm.org/citation.cfmdoid=2733373.2806390
23. Salamon, J., Jacoby, C. & Bello, J. P. A dataset and taxonomy for urbansound research[C]. in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. pp. 1041–1044. (2014).
24. Zhu, K. et al. Optimization Research on Abnormal Diagnosis of Transformer Voiceprint Recognition based on Improved Wasserstein GAN[J]. *Journal of Physics Conference Series*, **1746**, 012067. (2021). https://doi.org/10.1088/1742-6596/1746/1/012067
25. Chen, Y. et al. Effective Audio Classification Network Based on Paired Inverse Pyramid Structure and Dense MLP Block. arXiv. (2022).
26. Liu, J. et al. *BgNet: Classification of Benign and Malignant Tumors with MRI multi-plane Fusion* (Frontiers in Neuroscience, 2022).
27. Chen, Z. et al. *Partial Discharge Detection Technology for Transmission Lines Based on Drone Acoustic Diagnosis* (ICEEMT, 2023).
28. Peng, X., Zhou, X., Zhu, H., Ke, Z. & Pan, C. MSFF-Net: Multi-Stream Feature Fusion Network for surface electromyography gesture recognition. *PLOS ONE*. (2022).
29. Bhattacharya, Moinak and Prasanna, Prateek. Audio-visual feature fusion for improved thoracic disease classification.*Proceedings of SPIE*. (2023).

30. Iqbal, A. et al. Eavit: External attention vision transformer for audio classification[C]//*2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, : 1–6. (2024).

Author contributions

Conceptualization, ZHONG Xiaohui, DONG Sheng and ZHANG Yiyi; methodology, Jiang Lincen, LU Wei and DONG Sheng; software, ZHANG Yiyi; validation, DONG Sheng; writing—original draft preparation, ZHONG Xiaohui; writing—review and editing, Jiang Lincen, DONG Sheng andZHANG Yiyi; supervision, ZHANG Yiyi . All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.