Membrane Technology
ISSN (online): 1873-4049

An Analytical Method for Recognizing Cat Meowing States Utilizing Short-Time Fourier Transform and Vision Transformers

Jian Huang[1], Yuxia Tang[2], Peixuan Zhang[1*], Yanhua Liu[3]

[1]School of Intelligent Manufacturing and Electrical Engineering, Guangzhou Institute of Science and Technology, Guangzhou 510640, China;

[2]School of Computer Science and Engineering, Guangzhou Institute of Science and Technology, Guangzhou 510640, China;

[3]School of Art and Design, Guangzhou Institute of Science and Technology, Guangzhou 510640, China;

Correspondence：Zhang Peixuan, E-mail: zhangpeixuan@gist.edu.cn

Abstract:

Sound Event Detection (SED) technology is currently one of the research hotspots in the field of audio signal processing. Its goal is to identify the event categories present in an audio segment and label the start and end times of each event. Using sound detection technology to analyze and identify animal sound signals is important for understanding animal behavior patterns and detecting animal status. In view of the complex noise environment and low detection accuracy in practical application scenarios, this paper takes complex audio as the analysis object and explores the animal sound event detection method combining short-time Fourier transform technology and deep learning, which is an exploratory work for further developing practical animal sound recognition systems. The main work and innovations are as follows: (1) extracting the characteristics of animal sound events by analyzing the spectrogram imaging parameters, and (2) proposing a method for detecting animal state sound events based on short-time Fourier transform and deep learning.

Keywords: Sound Event Detection (SED), Spectrogram, Short-Time Fourier Transform, Complex Audio, Vision Transformers

INTRODUCTION

Sound serves as a pivotal means for animals to convey emotions to the external environment. By extracting features from animal sounds, the perception and understanding of animal emotions can be achieved, playing a crucial role in animal health assessments and ethological studies. However, traditional methods for detecting animal sound events have gradually revealed their limitations, unable to meet the urgent needs of current scientific research and animal health diagnostics.

Traditional approaches primarily rely on manual field recordings and subsequent audio analysis, demanding extensive acoustic knowledge and endurance for prolonged fieldwork from staff. They are susceptible to environmental noise interference, posing significant challenges, inefficiencies, and a high degree of subjectivity in data collection and analysis processes, thereby compromising the objectivity and accuracy of detection results. The difficulties in animal sound event detection lie in the acquisition of sound signals, the filtering of background noise, and the automatic identification of valid sound events.

Early research on animal sound event detection relied on manually deployed portable recording devices in wild environments or monitoring systems at ground-based fixed stations, followed by subsequent audio analysis. Buchan S. J., et al. [2] (2020) utilized Hidden Markov Model technology to achieve automatic detection and classification of blue whale vocalizations from PAM data, a method of great significance for advancing the monitoring of endangered whale populations, yet limited to specific datasets and species, necessitating additional optimization and adjustment for animal sound event detection. Premoli M., et al. [4] (2021) proposed and evaluated a series of supervised learning methods for automatic classification of ultrasonic vocalizations (USVs) to deeply analyze animal communication, significantly improving classification performance using Convolutional Neural Networks and other supervised learning algorithms to process spectrogram images. Romero-Mujalli D., et al. [5] (2021) validated the application potential of DeepSqueak software in the detection, clustering, and classification of high-frequency/ultrasonic

vocalizations in primates, capable of handling different call types, individual differences, and recording quality with high correct detection rates, successfully applied to species with close evolutionary relationships.

Abbasi R. L., et al. [6] (2022) introduced BootSnap, a classification method based on ensemble deep learning, successfully categorizing house mouse vocalizations into 12 classes with high generalization ability, albeit constrained to specific research contexts and animal species. Pessoa D., et al. [7] (2022) developed a system for classifying mouse USVs, employing a new segmentation algorithm based on spectral entropy analysis and a new classification method based on contour features, enabling efficient and accurate identification and classification of a broader range of USV categories. Bru E., et al. [8] (2023) combined acoustic localization with high-resolution land cover classification, offering potential for large-scale monitoring of vocally active animals, which can be used to infer animal habitat and landscape utilization. Stoumpou V., et al. [9] (2023) developed AMVOC, a free and open-source software incorporating unsupervised deep learning methods for detecting and analyzing mouse USVs.

In recent years, with the rapid development of the Internet of Things technology, deep learning algorithms, and audio signal processing techniques, the application of deep learning algorithms for animal sound event detection has emerged as a research hotspot.

Gómez-Armenta J. R., et al. [10] (2024) proposed a method using deep neural networks to analyze bark sounds for classifying dog identity, breed, age, gender, and barking context; Kucukkulahli E., et al. [11] (2024) classified cat sounds using deep learning models based on the Vision Transformer (ViT) and Convolutional Neural Network (CNN) architectures, finding that the ViT model based on BEiT outperformed the current best models; Manriquez P. R., et al. [12] (2024) investigated the application of Convolutional Neural Network architectures in bioacoustic classification tasks of emotional mammal vocalizations with small datasets, discussing the ability of networks to generalize emotional features of vocalizations across taxonomic groups; Pagliani B., et al. [13] (2024) classified the click sounds of common short-beaked dolphins based on clustering and discriminant analysis, finding that temporal parameters had the highest accuracy in comparisons, with time-frequency datasets being the best classification method; Pann V., et al. [14] (2024) proposed the use of Deep Convolutional Neural Networks (DCNN) and a novel feature extraction method, Mixed-MMCT, to automatically and accurately distinguish pig vocalizations from non-vocalizations, experimentally demonstrating the superiority of this method in real pig farming environments; Prakash R. V., et al. [15] (2024) introduced an automatic detection and emotional state classification method for wildlife based on a stacked Long Short-Term Memory (LSTM) network and hybrid features, achieving high classification accuracy; Salem, S. I., et al. [16] (2024) proposed a segmentation method based on acoustic anomaly detection and an integrated framework of machine learning models for extracting and classifying deer calls from long recordings, showing that all models performed well during validation and testing stages, with the ensemble method significantly improving classification accuracy; Shorten P. R., et al. [17] (2024) developed an algorithm using acoustic sensors worn on cow collars to distinguish cow vocalizations from other noises, validating the feasibility of identifying cow vocalization features, finding significant differences in vocalization features among cows, and enabling the identification of cows with abnormal vocalization patterns.

Automatic detection of bioacoustic events is crucial for monitoring wildlife. Given the cumbersome annotation process, limited annotated events, and large volumes of recordings, few-shot learning based on a small number of instances is of vital importance (2021) [18]. Researchers such as Liwen You [19] (2023) and Jinhua Liang [20] (2024) have attempted to train multifunctional animal sound detectors using small sets of audio samples. Few-shot bioacoustic sound event detection is also an important task in the field of animal vocalizations in nature [21] (2022).

During the audio recognition process for animal sound event detection, it is inevitable to encounter noise interference from the environment, necessitating algorithms with strong environmental adaptability and anti-interference capabilities [26] (2023) to effectively extract animal sound features in complex and variable natural environments.

This paper proposes an animal sound event detection method based on artificial intelligence and deep learning technology. Building on previous animal sound event detection methods, it innovatively determines three imaging parameters of the spectrogram—audio frequency, audio duration, and audio signal intensity range—through statistical analysis to achieve feature

extraction of animal sound events. It then applies Short-Time Fourier Transform (STFT) to the data and finally inputs the generated spectrogram into a pre-trained Vision Transformers (ViT) model to achieve animal sound event detection.

STFT-VIT-BASED ANIMAL CHIRPING SOUND ANALYSIS AND HEALTH STATUS IDENTIFICATION METHOD

Animal chirping sound analysis and health status recognition method based on animal chirping sound spectrogram analysis of large model STFT-ViT process using statistical analysis, audio signal strength, audio duration data visualization and analysis, to determine the audio frequency, audio duration and audio signal strength range of the three imaging parameters of the sound spectrogram and short time distance Fourier transform of the data; to generate the sound spectrogram image input into the pre-trained Vision Transformer model to realize the detection of animal sound events.

The STFT-ViT based animal chirping sound analysis and health state recognition method contains 2 parts, one is the training part and the other is the online health state recognition application part. The detailed steps of training among them include:

① reading an audio file and obtaining data of audio frequency, audio signal strength, and audio duration in the audio file;

② Data visualization to analyze the data and determine the two imaging parameters of audio signal strength range and audio duration;

③ Process the audio data with short time-distance Fourier transform, and draw an acoustic spectrogram based on the three imaging parameters of audio frequency, audio signal intensity range and audio duration;

④ Construct the classification dataset of acoustic spectrogram, and train the Vision Transformer model for animal chirping sound analysis and health status recognition.

The detailed steps of the online animal chirping sound analysis and health state recognition application part include:

① Collecting animal chirping sounds online and obtaining data on frequency and signal strength in the sounds;

② Using a window with a timing length of 3s, the sound data is processed by a short time-distance Fourier transform to determine the final imaging parameters and draw a sound spectrogram based on the parameter constraints.

③ Input the image into the pre-trained Vision Transformer model for animal chirping sound analysis and health status recognition.

1. 1 Introduction and Analysis of Animal Chirping Sound Data Dataset

The data used in this paper comes from audio files generated by 10 adult Maine cats kept under the same conditions (same owner) and 11 adult European Shorthair cats kept under different conditions (different owners), totaling 21 cats in three different situations: being petted, before feeding, and being alone in an unfamiliar environment, respectively. The dataset contained 93 samples in the waiting-for-food situation, 135 samples when being petted, and 220 samples when alone in an unfamiliar environment.

Before constructing the dataset, the influence of the cat's breed and sex (neutered male/female) on the purring was considered, the cat's acclimatization period (in the presence of at least one veterinarian to avoid overstimulation of the cat), and the placement of the recording device (distance and angle) were taken into account to ensure the quality of the audio.

By inputting the audio file and obtaining the audio signal strength and audio duration data therein, a histogram of the audio signal strength distribution and audio duration distribution was plotted as shown in Figure 1. Data visualization and analysis are performed for Figure 3 to obtain the statistical features of audio signal strength and audio duration, and two key features of the animal sound audio data are shown through two audio data distribution graphs: audio signal strength and audio duration.

Among them, the figure on the left is the distribution graph of audio signal intensity, which takes audio signal intensity as the horizontal axis and frequency as the vertical axis, and statistically demonstrates the distribution of the maximum audio signal intensity (Max intensity) and the minimum audio signal intensity (Min intensity) of all audio files, respectively. As can be seen from the figure, the range of audio intensity varies from -125dB to 100dB, while most of the audio signal intensity is concentrated between [-50,75]dB, which indicates that the audio signal intensity within the range occurs more often; the right-hand side of the figure is the distribution of audio duration, which is statistically displayed with audio duration as the horizontal axis and

frequency as the vertical axis. From the figure, it can be seen that the range of audio duration varies from 0s to 4s, and most of the audio duration is within the range of 3s, which indicates that most of the animal sound durations are in the first 3s of the audio duration.It has been shown by Ntalampiras S [4] and other studies that the use of devices with a frequency reception range of [0,4000] Hz can be better realized to obtain the cat's purring audio data, and then to judge its sound events [1]. Based on the above analysis, it can be finally determined that the audio signal strength range is between [-50, 75] dB; the audio duration is limited to 3s, and the audio frequency range is between [0, 4000] Hz.

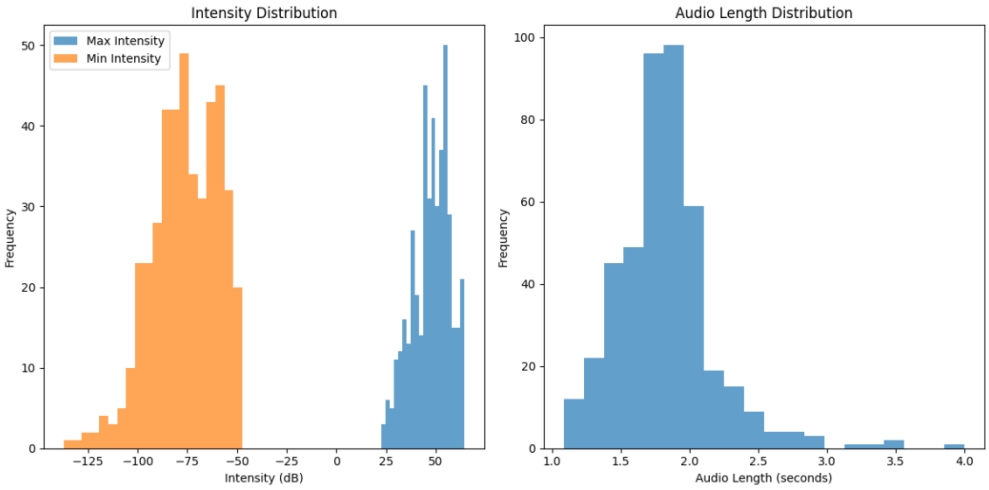

Fig. 1 Audio signal strength distribution and audio duration distribution of cat purring data

1.2 STFT Acoustic Spectrogram Classification Dataset Construction

(1) Data preprocessing

For the sampling frequency and audio data obtained by reading audio files, firstly, the audio data channels are uniformly processed as unidirectional channels. If the data is multi-channel, only the data of the first channel is taken, which simplifies the audio processing process and improves the processing efficiency without losing the key information. Then, based on the conclusions drawn from the statistical analysis, the audio duration is uniformly processed as 3s length. If the audio duration is less than 3s, then the audio is looped and extended to reach the target length; if the audio duration is longer than 3s, then it is truncated to 3s. Finally, the short time-distance Fourier transform is performed on the completed data preprocessing to obtain the audio data, and the output spectrograms are plotted based on the three parameters of the audio signal intensity range, the audio duration and the audio frequency.

(2) Short time-distance Fourier transform

The short time-distance Fourier transform is applied to the preprocessed data for data feature extraction, and the expression is as follows:

$$\mathrm{STFT}_{\mathrm{X}}(\mathrm{t},\omega)=\int_{-\infty}^{\infty}\mathrm{x}(\tau)\mathrm{w}^{*}(\tau-\mathrm{t})\mathrm{e}^{-\mathrm{j}\mathrm{w}\tau}\mathrm{d}\tau \quad (1)$$

Table 1 Symbol Representation

symbol	definition
ω	Symbol Definition
$\mathrm{w}(\mathrm{t})$	Angular frequency
*	Conjugate

The short-time distance Fourier transform STFT_X (t,ω) of the signal is discretized by setting the discrete signal of the continuous time-domain signal x(t) to be x(n), and the window function to be x(m), which is shifted on the time axis, and the length of the window function to be N. Then, the discrete form of short-time distance Fourier transform is:

$$\mathrm{STFT_X}(n,k) = \sum_{m=0}^{N-1} x(n+m)w(m)e^{-j2\pi nk/N} \quad (2)$$

in Eq. (2) $x(n+m)w(m)$ is the short time series.

By calculating the result of the short time-distance Fourier transform for each time point n, the frequency-strength-time characteristic distribution of the signal can be obtained. Among them, the frequency characteristic distribution of the signal describes the distribution of the signal at different frequencies; the intensity characteristic of the signal describes the strength of the signal, and the time characteristic of the signal describes the law and characteristics of the signal change with time.

(3) STFT acoustic spectrogram plotting

After the short time distance Fourier transform, based on the audio signal intensity range, audio duration and audio frequency three parameters drawn out in Figure 4 ~ Figure 6 in the time-frequency diagram and the acoustic spectrogram, from the acoustic spectrogram can be derived from the horizontal axis of the time longitudinal axis of the frequency distribution of the frequency distribution of the frequency of the signal at the time of the existence of each position represented in the time.

The horizontal axis represents the audio duration of 3s, reflecting the change of the audio signal over time; the vertical axis represents the audio frequency range of [0,4000] Hz, which can be derived from the audio signal contains which frequency components, the frequency value of the fundamental tone, the frequency distribution of the overtones, and the change of frequency over time. The colors represent the audio signal strength range of [-50,75]dB, with higher color brightness representing stronger signal strength and darker representing weaker signal strength.

Figure 2 shows the sound spectrum of the audio generated when the cat is stroked. From the right side of the sound spectrum, it can be seen that the vocal signal has a certain continuity and regularity, the frequency range is more concentrated, the vocal duration is longer and continuous, and the vocal performance is stable. This indicates that when the cat is being petted, it will emit a calm purr because it is in a safe and stable environment, and express its specific emotions through a relatively uniform frequency.

Figure 3 shows the spectrogram of the audio produced by the cat in an unfamiliar environment. From the spectrogram on the right, it can be seen that the intensity of the vocal signal shows large fluctuations, and the duration of the vocalization is shorter and intermittent, with a wider frequency range. This indicates that cats in unfamiliar environments will emit tentative and intermittent sounds due to restlessness and panic, and express their fearfulness through purring at different frequencies.

Figure 4 shows the corresponding sound spectrogram of the audio generated when the cat waits for food. From the right side of the sound spectrogram, it can be seen that the intensity of the vocal signal is relatively low and stable, the duration of the vocalization is longer but intermittent, and the frequency range is mainly concentrated in the low-frequency band. This indicates that the sound emitted by the cat when waiting for food is weak and low and long due to hunger.

By comparing the acoustic spectrograms corresponding to the three different vocal events of being stroked, being in an unfamiliar environment and waiting for food, it can be found that there are significant differences in their corresponding audio signal strength, vocal duration and vocal frequency in the acoustic spectrograms, which can highlight the characteristics of the audio data of different animal vocal events.

By inputting the spectrograms into the pre-trained Vision Transformer model, the final output can be the situation of the cat under the corresponding sound event.

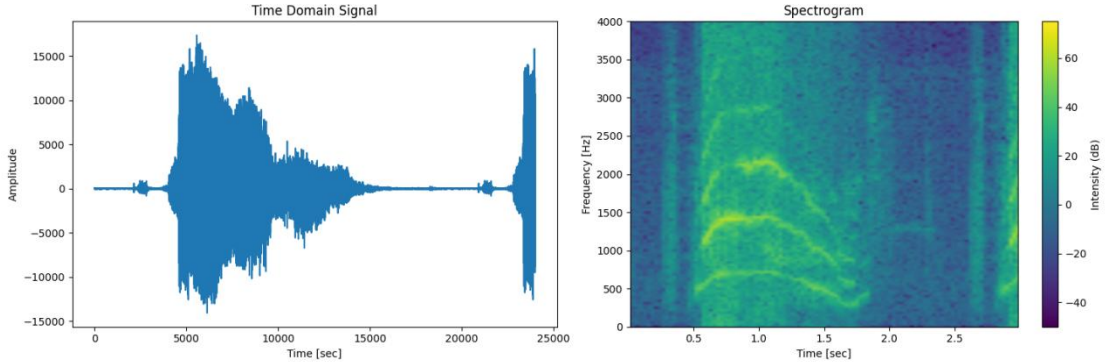

Fig. 2 Time-domain and spectrogram of a cat being petted.

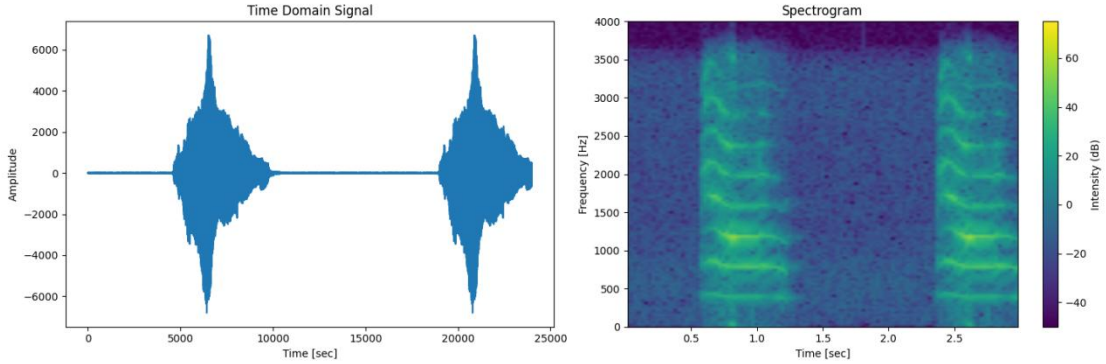

Figure 3 Time-domain map and spectrogram of a cat in an unfamiliar environment.

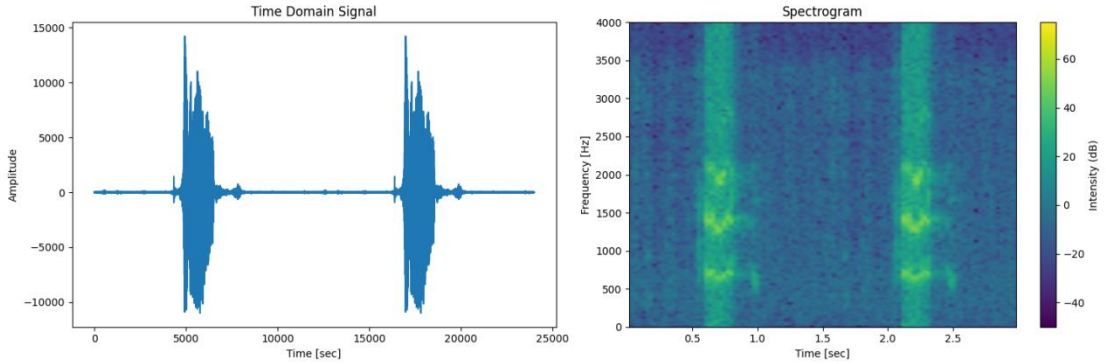

Fig. 4 Time-domain and acoustic spectrograms of cats waiting for food

(4) Construction of STFT spectrogram classification dataset

After the data preprocessing, short time distance Fourier transform, and acoustic spectrogram plotting are performed sequentially on all the animal sound audio files, the acoustic spectrograms with hidden axis labels are saved, and based on the three sound event types of brushing, isolation_in_an_unfamiliar_environment, and waiting_for_food, the corresponding animal sound events are categorized to construct the STFT acoustic spectrogram classification dataset. The acoustic spectrograms of the corresponding animal sound events were categorized to construct the STFT acoustic spectrogram classification dataset.

Among them, brushing represents the acoustic spectrogram folder corresponding to the audio data generated when the cat is being petted, and there are 127 samples under this folder; isolation_in_an_unfamiliar_environment represents the acoustic spectrogram folder corresponding to the audio data generated when the cat is in an unfamiliar environment; and there are 221 samples under this folder. The folder "waiting_for_food" represents the folder of sound spectrograms corresponding to the audio data generated by a cat waiting for food; there are 92 samples under this folder.

VISION TRANSFORMER MODEL FOR ANIMAL CHIRPING SOUND ANALYSIS AND HEALTH STATUS RECOGNITION

ViT divides the input picture into multiple sub-patches (16×16), and then projects each sub-patch into a fixed-length vector to be fed into the Transformer model, and the subsequent encoder operation is exactly the same as in the original Transformer model. However, because of the picture classification, a special classification token is added to the input sequence, and the output corresponding to this token is the final category prediction.

According to the flowchart in Fig. 5, a ViT block can be divided into the following steps:

(1) patch embedding: for example, if the input image size is 224 × 224, and the image is divided into fixed-size subgraphs with a subgraph size of 16 × 16, then 224 × 224/16 × 16 = 196 subgraphs will be generated for each image, i.e., the length of the input sequence is 196, and the dimensionality of each subgraph is 16 × 16 × 3 = 768, and the dimensionality of the linear projection layer is 768 × N (N = 768), so the dimension of the input after passing through the linear projection layer is still 196 × 768, i.e., there are a total of 196 tokens, each with a dimension of 768. a special character cls needs to be added here as well, so the final dimension is 197 × 768. so far a visual problem has been transformed by patch embedding into a seq2seq problem;

(2) positional encoding (standard learnable one-dimensional positional embedding): ViT also needs to incorporate positional encoding, which can be interpreted as a table with a total of N rows, the size of N is the same as the length of the input sequence, and each row represents a vector, the dimension of the vector is the same as the dimension of the input sequence embedding (768). Note that the operation of position encoding is summing, not splicing. The dimension remains 197 × 768 after adding the position encoding information;

(3) LN/multi-head attention/LN: The LN output dimension is still 197 × 768. for multi-head self-attention, the inputs are first mapped to q, k, and v. If there is only one head, the dimensions of qkv are all 197 × 768, and if there are twelve heads (768/12 = 64), the dimensions of qkv are 197 × 64, and there are a total of 12 groups of qkvs, and then finally the output of 12 groups of qkvs are spliced together, and the output dimension is 197×768, and then after another layer of LN, the dimension is still 197×768;

(4) MLP: the dimension is enlarged and then shrunk back, 197×768 is enlarged to 197×3072, and then shrunk again to 197×768 After block, the dimension is still the same as the input, which is 197×768, so multiple blocks can be stacked. Finally, the output z_L^0 corresponding to the special character cls will be used as the final output of the encoder to represent the final picture display (another approach is to leave the cls character out, and do an average of all the labeled outputs), as shown in the following figure in Eq. (6), followed by an MLP for picture classification;

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \cdots; x_p^N E\right] + E_{pos} E \in \mathbb{R}^{(P^2 \cdot C) \times D},\ E_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{3}$$

$$z' = MSA(LN(z_{\ell-1})) + z_{\ell-1}, \qquad \ell = 1 \ldots L \tag{4}$$

$$z_\ell = MLP\left(LN(z'_\ell)\right) + z'_\ell, \quad \ell = 1 \ldots L \tag{5}$$

$$y = LN(z_L^0) \tag{6}$$

where input image image $x \in \mathbb{R}^{H \times W \times C}$,2D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$,C is the number of channels, P is the size of the subgraphs, and there are a total of N patches, N=HW/P^2.

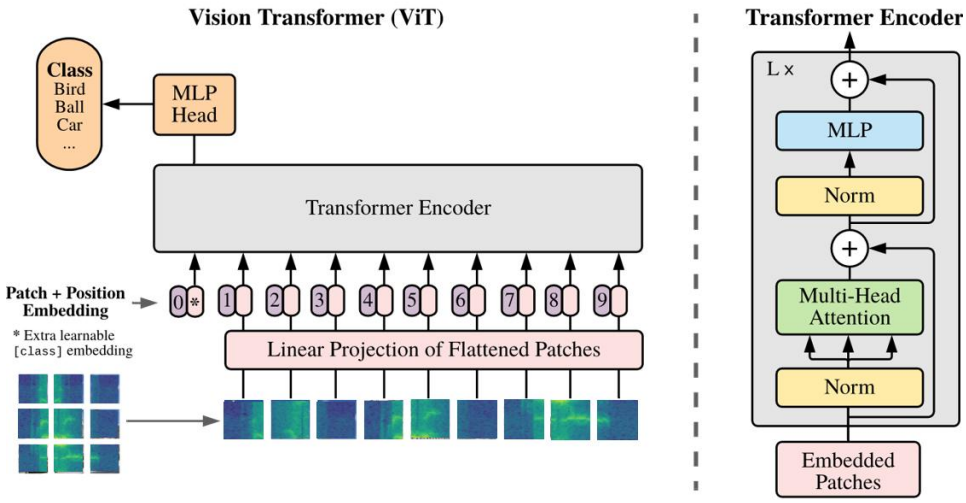

Fig. 5 Structure of Vision Transformer model

EXPERIMENTAL RESULTS AND ANALYSIS

Our model is a pre-trained fine-tuned version of google/vit-base-patch16-224-in21k on the imagenet 21k dataset.

For model training, we choose to use the Adam optimizer and set its parameters betas to (0.9,0.999), epsilon to 1×10-8, set the learning rate to 2×10-5, and the type of learning rate scheduler to linear, which is convenient for the model to achieve fast convergence in the early stage of training while keeping the learning rate small in the late stage of training to stabilize the model performance; for the training process For the training process, we set the number of training rounds to 50.0, the size of the training batch to 8, and the size of the evaluation batch to 8 to maintain consistency with the size of the training batch, and set the random seed to 1337 to control the initial state of the random number generator to ensure the reproducibility of the experiment.

Next, the training data from the training set is input into the model for forward propagation calculation, the difference between the model prediction result and the actual label is calculated using the cross-cross entropy loss function to obtain the loss value, and the weights of the model are updated using the optimizer based on the gradient of the loss function. Then, the model is trained iteratively for a predetermined number of training cycles.

Finally, our network structure can be trained using the above parameters to obtain the large model STFT-ViT for animal chirping spectrogram analysis, whose final model validation loss is 0.0295 and model accuracy is 100%.

As Figure 6~Figure 7 shows the relevant curves of model training. Among them, Fig. 6 shows the model training loss curve and the validation loss curve, and it can be seen from the images that both the training loss curve and the validation loss curve show a decreasing trend with the increase of the number of iterations. This indicates that the model is gradually learning the features of the data during the training process and gradually reducing the prediction error. Since both curves show a decreasing trend, and there is no obvious fluctuation or rise, it can be initially judged that the model is stable during the training process, and there is no serious overfitting or underfitting phenomenon. Among them, the training loss and the validation loss are gradually close to each other over time, although there still exists a certain difference between the two, and the difference significantly decreases with the increase in the number of iterations, indicating that the model not only performs well on the training data, but also maintains a better performance on the unseen validation data, and has a better generalization ability.

Figure 7 shows the model accuracy curve, from the image can be seen that the model accuracy curve with the increase in the number of iterations shows an upward trend and tends to stabilize, indicating that the model gradually learns the characteristics of the data during the training process, and tends to stabilize in the later stage, which means that the model has converged to a relatively good state.

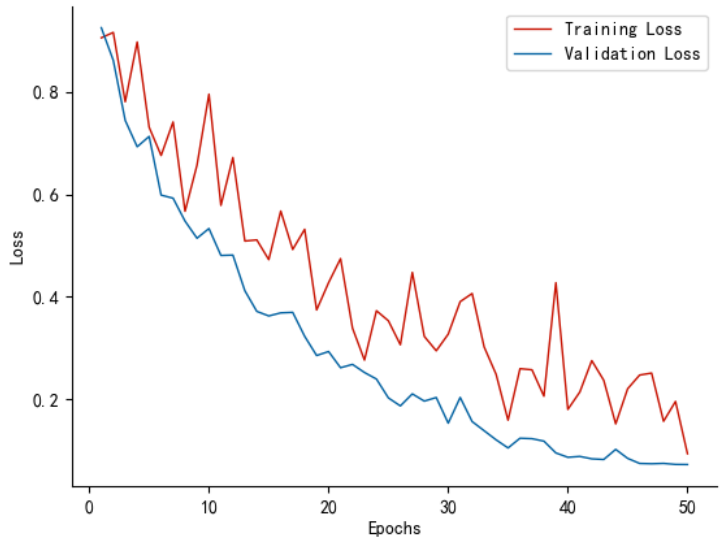

Fig. 6 Training Loss curve and Validation Loss curve of model training with adaptive acoustic spectrogram parameters.

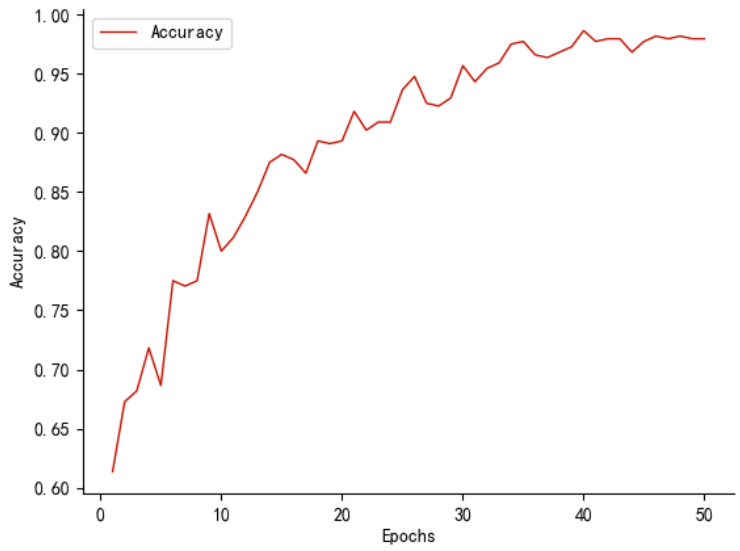

Fig. 7 Accuracy curve of model training for acoustic spectrogram parameter adaptation.

(1) Comparison experiment:

As Figure 8~Figure 10 shows the model training correlation curves when the audio duration, audio signal strength and audio frequency are all adaptive.

Among them, Fig. 8~Fig. 10 shows the comparison between the training loss curve and validation loss curve of our model and the model with parameter adaption. As can be seen from the images, the training loss shows a trend of rapid decline followed by leveling off with the increase in the number of iterations, but the validation loss starts to show a rising trend after it rapidly decreases to reach a lower point. This indicates that although the model learns faster at the beginning of training, and is able to quickly learn the features of the data and reduce the loss, as the training proceeds, the model fits the training data better and better, and the overfitting phenomenon may have occurred, resulting in a decrease in the generalization ability on the validation

set, and the final model validation loss is obtained to be 0.0295; Fig. 10 shows a comparison of the accuracy curves of the model, and it can be seen from the image that the model Accuracy curve with the increase in the number of iterations shows an upward trend, and the final model accuracy rate obtained is 100%. Combining the model loss curves with the accuracy curves, it can be seen that the model based on the three acoustic spectrogram imaging factors of audio duration, audio signal strength, and audio frequency are all adaptive has overfitting phenomenon during the training process, and performs too well on the training dataset, so much so that it learns noises or specific patterns in the training data, which do not apply to the new, unseen data.

Table 2 Training data for our model and the parameter adaptive model

STFT parameter setting method	Accuracy	Training set loss	Test set loss
Our Approach	100%	0.307	0.0295
Audio duration, audio signal strength, audio frequency adaption	100%	0.307	0.0295

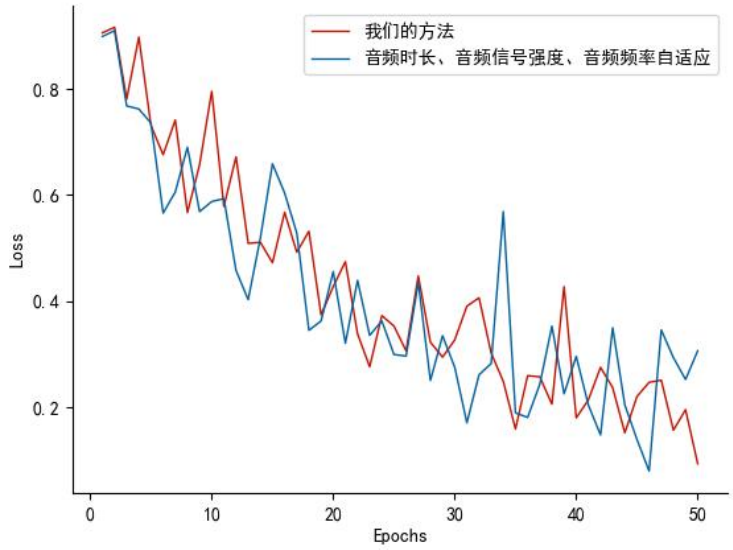

Fig. 8 Comparison of the Training Loss curves of our method with audio duration, audio signal strength, and audio frequency adaption.

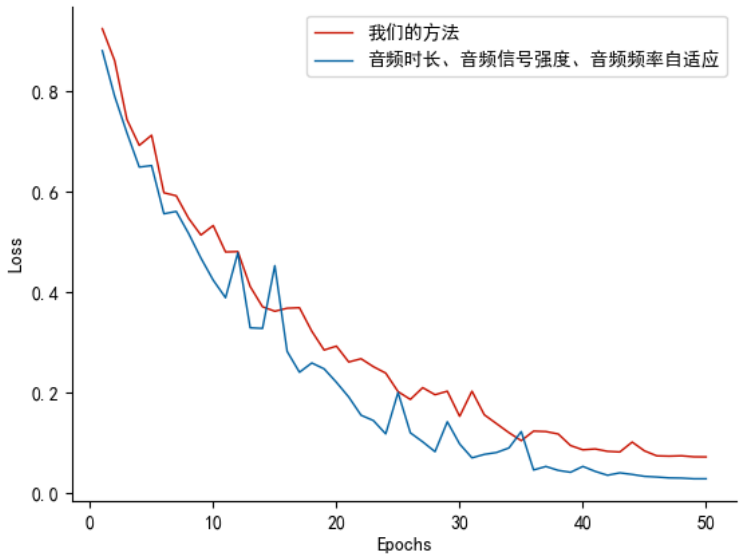

Fig. 9 Validation Loss curve of our method versus audio duration, audio signal strength, and audio frequency adaption.

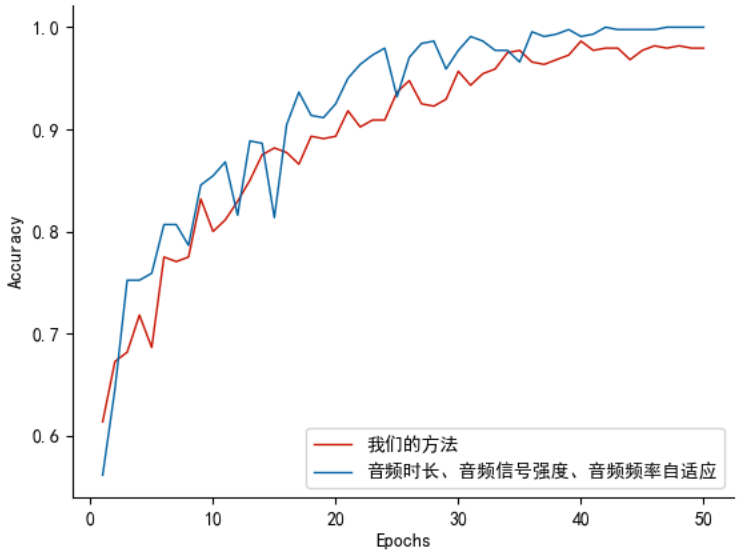

Fig. 10 Accuracy curve of our method versus audio duration, audio signal strength, and audio frequency adaption.

(2) Ablation experiment

As shown in Figures 11 to 13, the accuracy curves obtained by model training after making changes to the three acoustic spectrogram imaging parameters, namely, audio duration, audio signal strength, and audio frequency, respectively. Among them, Fig. 11 is the model training accuracy curve when the audio duration is adaptive, and the final model validation loss is 0.0301, with a model accuracy of 99.5%; Fig. 12 is the model training accuracy curve when the audio duration reaches 3s through the complementary 0 to fill in the topology, and the final model validation loss is 0.1606, with a model accuracy of 95.23%; Fig. 13 is the model training accuracy curve when the audio signal strength is adaptive; Fig. 13 is the model training accuracy curve when the audio signal strength is adaptive; Fig. 14 is the model training curve when the audio signal strength is adaptive. Figure

13 shows the model training accuracy curve when the audio signal strength is adaptive, and the final model validation loss obtained is 0.0444 and the model accuracy is 99.32%; Figure 19 shows the model training accuracy curve when the audio frequency is adaptive, and the final model validation loss obtained is 0.0527 and the model accuracy is 99.09%.

As can be seen in the four images from Figure 11 to Figure 13, the model gradually learns the characteristics of the data as the training progresses, gradually adapts to the training data, and begins to make more accurate predictions.

Combining the model training loss curve and the validation loss curve when the audio duration is adaptive, it can be seen that although the final accuracy of the model training in Fig. 11 reaches 99.5%, the failure to limit the range of audio durations leads to the model learning the noise or specific patterns in the training data, and overfitting phenomenon, which suggests that the model, although it is able to perform well on the training data, is unable to maintain better performance on the unseen validation data; Fig. 11~13 shows that the model gradually learns the features of the training data and starts to make better predictions. This shows that although the model can perform well on the training data, it cannot maintain good performance on the unseen validation data; Figure 12 combines the model training loss curves with the validation loss curves when the audio duration is top-filled to a length of 3s by complementary zeros in Figure 13, which shows that the simple complementary zeros top-filling method has a certain negative impact on the training of the model, and the model is difficult to learn more features of the data, which leads to an increase in the training loss and a decrease in the accuracy; Figure 14 combines the model training loss curves and the validation loss curves when the audio signal strength is adaptive model training loss curves and validation loss curves can be seen that the model performance on the training set is improved in the later stages, but the performance on the validation set is not significantly improved, indicating that the model training in the later stages of the overfitting phenomenon, due to the failure to limit the range of the audio signal strength leads to learning too much noise or details on the training set, rather than learning features with strong generalization ability; audio frequency adaptive The model training loss curve and validation loss curve at time can be seen, due to the failure to restrict the audio frequency, resulting in the model learning as the model training advances, the model learning is subject to the noise or specific patterns in the training data, and it is difficult to learn more features of the data that are conducive to classification.

In summary, restricting the audio duration to 3s intercepts the audio time range where the data features are located as much as possible, restricting the audio signal strength to [-50,75]dB retains the audio signal strength of the main sound as much as possible, and restricting the audio frequency to 4000Hz filters out noise as much as possible, which effectively improves the model's ability to learn the data features through the restriction of the three acoustic spectrogram imaging parameters, and avoids the occurrence of overfitting phenomenon as much as possible. avoid overfitting phenomenon.

Table 3 Comparison of our method with models of audio timbre, intensity, and frequency adaptive

STFT parameter setting method	Accuracy	Training set loss	Test set loss
Our Methods	97.95%	0.0941	0.0730
Audio Duration Adaptation	99.5%	0.0288	0.0301
Audio duration with 0 topology delay	95.23%	0.4276	0.1606
Audio Signal Strength Adaptive	99.32%	0.3043	0.0444
Audio Frequency Adaptive	99.09%	0.3265	0.0527

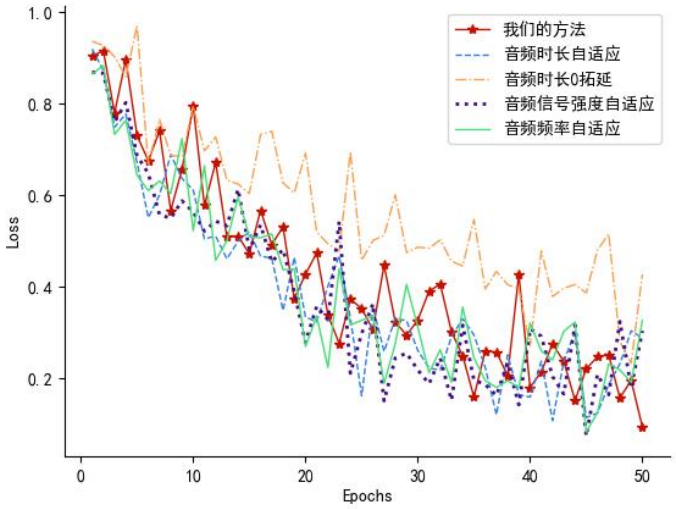

Fig. 11 Comparison of Training Loss curve of our method with other methods of model training

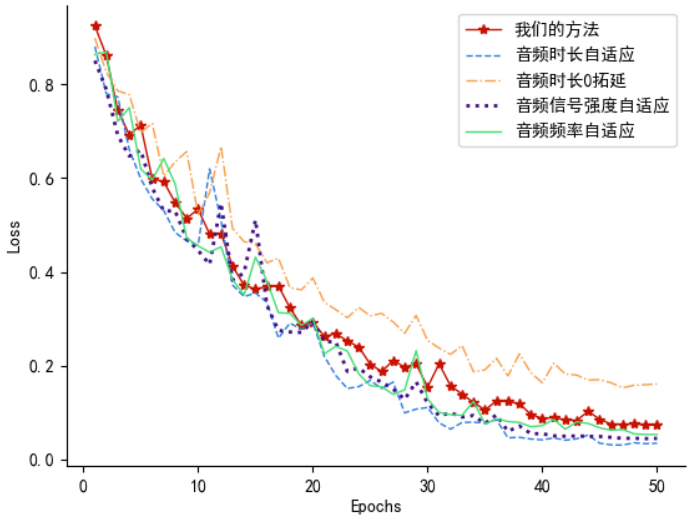

Fig. 12 Validation Loss curve of our method vs. other methods for model training.

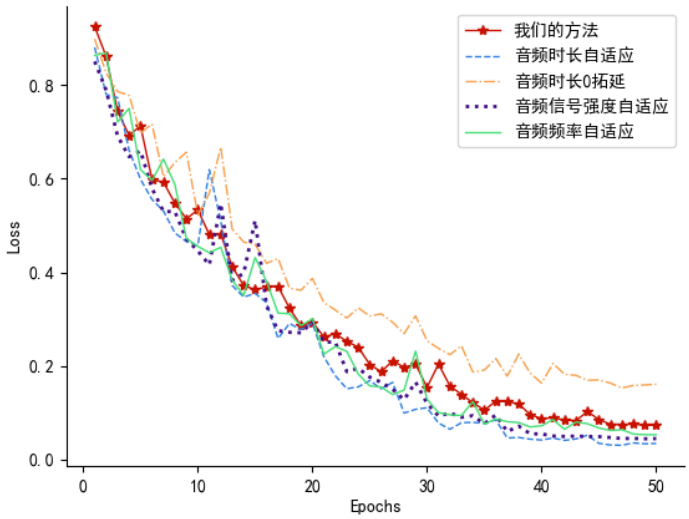

Fig. 13 Accuracy curve of our method vs. other methods for model training.

CONCLUSION

This paper combines statistical analysis and artificial intelligence deep learning technology to propose an animal sound event detection method that can achieve optimized signal analysis and improve measurement accuracy, effectively improving the accuracy of animal sound event detection. The innovation points are:

① The method of statistical analysis is used to limit the information of the three imaging parameters of the acoustic spectrogram, so as to realize the limitation of the resolution bandwidth of the acoustic spectrogram, enhance the differentiation ability of the signals of similar frequency, and improve the measurement accuracy of the signals.

② Use the pre-trained Vision Transformer model to provide a good initialization for sound event detection, and with the help of the pre-trained weights, realize that it can still effectively learn audio features and accurately detect sound events in the face of less labeled data.

③ The experimental results show that the animal sound event detection method based on artificial intelligence deep learning technology is accurate and effective, which can improve the model accuracy rate of 98.64% and effectively improve the accuracy of animal sound event detection.

However, the generalization ability of the model across species, the scarcity and diversity of data in animal sound event detection are all great challenges to animal sound event detection, and the next step is to combine audio processing technology with deep learning acoustic detection technology to achieve the enhancement of animal sound data.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

DATA SHARING AGREEMENT

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

FUNDING

The project was supported by National Key R&D Plan National Quality Infrastructure System Special Project (2022YFF0607805), Guangdong Provincial Market Supervision Administration Science and Technology Project - Key Laboratory Special Fund (2024CJ09), Guangzhou Institute of Technology and Research Startup Funded Project (2023KYQ177).

REFERENCE

[1] Ntalampiras S, Ludovico L A, Presti G, et al. Automatic Classification of Cat Vocalizations Emitted in Different Contexts[J]. Animals, 2019, 9(8).

[2] Buchan S J, Mahú R, Wuth J, et al. An unsupervised Hidden Markov Model-based system for the detection and classification of blue whale vocalizations off Chile[J]. Bioacoustics-the International Journal of Animal Sound and Its Recording, 2020, 29(2): 140-167.

[3] Prato-Previde E, Cannas S, Palestrini C, et al. What's in a Meow? A Study on Human Classification and Interpretation of Domestic Cat Vocalizations[J]. Animals, 2020, 10(12).

[4] Premoli M, Baggi D, Bianchetti M, et al. Automatic classification of mice vocalizations using Machine Learning techniques and Convolutional Neural Networks[J]. Plos One, 2021, 16(1).

[5] Romero-Mujalli D, Bergmann T, Zimmermann A, et al. Utilizing DeepSqueak for automatic detection and classification of mammalian vocalizations: a case study on primate vocalizations[J]. Scientific Reports, 2021, 11(1).

[6] Abbasi R L, Balazs P, Marconi M a L, et al. Capturing the songs of mice with an improved detection and classification method for ultrasonic vocalizations (BootSnap)[J]. Plos Computational Biology, 2022, 18(5).

[7] Pessoa D, Petrella L, Martins P, et al. Automatic segmentation and classification of mice ultrasonic vocalizations[J]. Journal of the Acoustical Society of America, 2022, 152(1): 266-280.

[8] Bru E, Smith B R, Butkiewicz H, et al. Combining acoustic localisation and high-resolution land cover classification to study predator vocalisation behaviour[J]. Wildlife Research, 2023, 50(12): 965-979.

[9] Stoumpou V, Vargas C D M, Schade P F, et al. Analysis of Mouse Vocal Communication (AMVOC): a deep, unsupervised method for rapid detection, analysis and classification of ultrasonic vocalisations[J]. Bioacoustics-the International Journal of Animal Sound and Its Recording, 2023, 32(2): 199-229.

[10] Gómez-Armenta J R, Pérez-Espinosa H, Fernández-Zepeda J A, et al. Automatic classification of dog barking using deep learning[J]. Behavioural Processes, 2024, 218.

[11] Kucukkulahli E, Kabakus A T. Towards Understanding Cat Vocalizations: A Novel Cat Sound Classification Model Based on Vision Transformers[J]. Applied Acoustics, 2024, 226.

[12] Manriquez P R, Kotz S A, Ravignani A, et al. Bioacoustic classification of a small dataset of mammalian vocalisations using deep learning[J]. Bioacoustics-the International Journal of Animal Sound and Its Recording, 2024, 33(4): 354-371.

[13] Pagliani B, Amorim T O S, De Castro F R, et al. Sounds in common: Time-frequency as the classification parameters for pulsed sounds produced by Delphinus delphis[J]. Behavioural Processes, 2024, 221.

[14] Pann V, Kwon K S, Kim B, et al. DCNN for Pig Vocalization and Non-Vocalization Classification: Evaluate Model Robustness with New Data[J]. Animals, 2024, 14(14).

[15] Prakash R V, Karthikeyan V, Vishali S, et al. Multi-level LSTM framework with hybrid sonic features for human-animal conflict evasion[J]. Visual Computer, 2024.

[16] Salem S I, Shirayama S, Shimazaki S, et al. Ensemble deep learning and anomaly detection framework for automatic audio classification: Insights into deer vocalizations[J]. Ecological Informatics, 2024, 84.

[17] Shorten P R, Hunter L B. Acoustic sensors to detect the rate of cow vocalization in a complex farm environment[J]. Applied Animal Behaviour Science, 2024, 278.

[18] Veronica Morfi, Inês Nolasco, Vincent Lostanlen, Shubhr Singh, Ariana Strandburg-Peshkin, Lisa F. Gill, et al., "Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge", Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021) 2021 Detection and Classification of Acoustic Scenes and Events 2021 Workshop DCASE2021, Conference date: 15-11-2021.

[19] L. You, E. P. Coyotl, S. Gunturu and M. Van Segbroeck, "Transformer-Based Bioacoustic Sound Event Detection on Few-Shot Learning Tasks," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10097081.

[20] J. Liang, I. Nolasco, B. Ghani, H. Phan, E. Benetos and D. Stowell, "Mind the Domain Gap: A Systematic Analysis on Bioacoustic Sound Event Detection," 2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 1257-1261, doi: 10.23919/EUSIPCO63174.2024.10714948.

[21] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. Van Langevelde, T. Burghardt et al., "Perspectives in machine learning for wildlife conservation", Nature communications, vol. 13, no. 1, pp. 1-15, 2022.

[22] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap", PeerJ, vol. 10, pp. e13152, 2022.

[23] J. Hamer, E. Triantafillou, B. Van Merrienboer, S. Kahl, H. Klinck, T. Denton, et al., "Birb: A generalization benchmark for information retrieval in bioacoustics", arXiv preprint, 2023.

[24] R. Li, J. Liang and H. Phan, "Few-shot bioacoustic event detection: Enhanced classifiers for prototypical networks", Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), November 2022.

[25] J. Liang, X. Liu, H. Liu, H. Phan, E. Benetos, M. D. Plumbley, et al., "Adapting Language-Audio Models as Few-Shot Audio Learners", Proc. INTERSPEECH 2023, pp. 276-280, 2023.

[26] I. Nolasco, B. Ghani, S. Singh, E. Vidaña-Vila, H. Whitehead, E. Grout, M. Emmerson, F. Jensen, I. Kiskin, J. Morford et al., "Few-shot bioacoustic event detection at the dcase 2023 challenge", arXiv preprint, 2023.

[27] J. Liang, H. Phan and E. Benetos, "Learning from taxonomy: Multi-label few-shot classification for everyday sound recognition", ICASSP 2024 – 2024 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 1-5, 2024.

[28] P. Mangal, A. Rajesh and R. Misra, "Big data in climate change research: opportunities and challenges", 2020 International Conference on Intelligent Engineering and Management (ICIEM), pp. 321-326, 2020.

[29] C. Poepel, K. Finger, N. Peters and B. Edler, "Exploring a long-term dataset of nature reserve ambisonics recordings", Proceedings of the 17th International Audio Mostly Conference, pp. 84-87, 2022.

[30] K. Darras, N. Pérez, Mauladi, L. Dilong, T. Hanf-Dressler, M. Markolf, et al., "ecosound-web: an open-source online platform for ecoacoustics [version 2; peer review: 2 approved]", F1000 Research, vol. 9, March 2023.