Applied Intelligence (2023) 53:4907–4923
https://doi.org/10.1007/s10489-022-03581-6

Domestic pig sound classification based on TransformerCNN

Jie Liao[1,2] · Hongxiang Li[1,2] · Ao Feng[1,2] · Xuan Wu[2] · Yuanjiang Luo[2] · Xuliang Duan[1,2] · Ming Ni[1,2] · Jun Li[1,2]

Accepted: 3 April 2022 / Published online: 16 June 2022

Abstract

Excellent performance has been demonstrated in implementing challenging agricultural production processes using modern information technology, especially in the use of artificial intelligence methods to improve modern production environments. However, most of the existing work uses visual methods to train models that extract image features of organisms to analyze their behavior, and it may not be truly intelligent. Because vocal animals transmit information through grunts, the information obtained directly from the grunts of pigs is more useful to understand their behavior and emotional state, which is important for monitoring and predicting the health conditions and abnormal behavior of pigs. We propose a sound classification model called TransformerCNN, which combines the advantages of CNN spatial feature representation and the Transformer sequence coding to form a powerful global feature perception and local feature extraction capability. Through detailed qualitative and quantitative evaluations and by comparing state-of-the-art traditional animal sound recognition methods with deep learning methods, we demonstrate the advantages of our approach for classifying domestic pig sounds. The scores for domestic pig sound recognition accuracy, AUC and recall were 96.05%, 98.37% and 90.52%, respectively, all higher than the comparison model. In addition, it has good robustness and generalization capability with low variation in performance for different input features.

Keywords Audio recognition · Transformer · Neural networks · Domestic pig · Animal behavior

1 Introduction

Domestic pigs have considerable economic value and have been used as an important source of nutritional intake for people. Pig rearing has been closely related to the production and lifestyle of people around the world. On the one hand, people as breeders are often concerned about pig nutritional health status and its changes, which directly affect the quality and yield of pork. On the other hand, people as consumers are more concerned about pork health. The outbreak of swine fever in recent years has directly affected pork production and quality in affected regions and countries, leading to a sharp rise in pork prices, which has seriously affected socioeconomic development and people's production and lifestyle [1]. The development of science and technology has led to rapid development in the feeding industry, and computerized intelligent systems have brought many benefits to agriculture, such as increased cost efficiency, improved animal welfare, and better production monitoring [2]. The intelligent ecological feeding mode provides convenience for feeding development on the one hand and makes it possible for the breeder to regulate the feeding process intuitively on the other hand.

Through the precise control of the feeding process, it is possible to directly understand the growth and living conditions of the animals, which is an important part of the feeding process. The introduction of intelligent technology allows remote monitoring of animal feeding and real-time access to data, which in turn improves production efficiency [3, 4].

Intelligent systems are now widely used in animal behavior and condition research. S. Hua et al. used image processing techniques to solve monitoring and management problems in smart pig farming to improve productivity and management efficiency [5]. Tian et al. used deep learning for pig density calculation and showed good accuracy [6]. Cowton et al. and Alameer et al. used detection technology for pig movement tracking for behavior monitoring [7,

Jie Liao, Hongxiang Li and Ao Feng contributed equally to this work.

✉ Jun Li
lijun@sicau.edu.cn

Extended author information available on the last page of the article.

8]. D Li et al. used a nuclear extreme learning machine (KELM) to detect injury behavior in pigs [9]. Zhang et al. used the mel inverse spectral coefficient of pig coughs to construct a recognition system for automated management in intensive feedlots [10]. Leliveld et al. studied sow calls through emotional responses in context and found encoding potential in the emotional state of sow calls [11]. The application of intelligent detection systems tremendously simplifies the management process and helps greatly in feeding.

In recent years, researchers have become increasingly concerned with the health status of captive animals, both for animal conservation purposes and for meat health monitoring purposes. One of the most significant studies focuses on the expression of emotional states in animals, and understanding nonverbal animal emotions plays a crucial role in people's understanding of animal habits and states [12, 13]. Animal emotions can be used to explain the motivations that produce certain behaviors [14]. The judgment ability of pigs can be influenced by emotional changes and personality differences [15]. However, in the existing studies, there are still problems such as lack of standard datasets, low accuracy of model recognition, and weak generalization and robustness. To solve the above problems, this paper proposes a method to discriminate behavior and emotion by pig grunts, and the main contributions of this paper are as follows.

1. We used recording equipment to collect domestic pig calls in breeding plants, and by characterizing domestic pig calls, we unified them in dimensionality. The collected call data were constructed as a standardized dataset for other scholars to study. No publicly available dataset has been provided for reference in this research direction in past studies. Our work will be an important foundation for subsequent related studies.
2. We use a total of eight audio features for extensive comparison experiments by extracting multiple sound audio features and their combinations to select the best audio features.
3. We design a novel network structure (Transformer-CNN) that has excellent global feature perception and local feature extraction capability. By parallel computing of the Transformer and CNN, we effectively compensate for the RNN feature, which cannot be parallelized in processing speech sequences, and on this basis, we combine the excellent CNN local feature extraction ability. By structural parallelism, richer feature information can be learned in different call signals than in a single network structure. In addition, this work is the first study of domestic pig sounds using the Transformer and CNN parallel models.
4. The proposed TransformerCNN for animal call classification of four animal (cat(4), bird(16), bird(8), pig(4), Cetaceans(4)) call datasets all exhibit advanced classification results with little effect on the difference of animals.

The main structure of this paper is as follows. In Section 2, we discuss the related research work. In Section 3, the main research methods of this paper are provided. In Section 4, we conduct qualitative and quantitative experiments. In Section 5, the work of this paper is discussed. In Section 6, we conclude the paper and propose directions for future work.

2 Related work

The Gaussian mixture model (GMM) and hidden Markov model (HMM) are widely used to construct acoustic models in speech recognition tasks. B. O. Kang proposed using the Gaussian mixture model (GMM), which maximizes the shared context information and location of the state of the logarithmic likelihood to implement robust speech recognition caused by environmental factors. To solve the identification error, this method can effectively solve the influence of different noise and improve the stability of recognition [16]. Marek B. et al. used hidden Markov models (HMMs) for automatic classification and vocalization detection of whales [17],and with the continuous development of artificial intelligence, deep learning is widely used in various fields. Dias Issa et al. used four spectral contrast features to train deep convolutional neural networks for performing speech emotion recognition tasks with success [18]. Saon et al. demonstrated impressive results using recurrent neural networks in speech recognition tasks [19]; M. Nasef Mohammed et al. proposed an end-to-end system based on a self-attention mechanism for speech recognition tasks in unconstrained environments [20]. A. Orhan et al. proposed an end-to-end 3D CNN-LSTM model with attention guidance for solving speech-based emotion recognition [21].An encoder-decoder with an attention mechanism was introduced, and the new model Transformer architecture was also applied to machine translation tasks [22].

The unlimited potential of deep learning has been verified in recent studies, and many researchers have continued to attempt to apply deep learning to animal sound classification tasks. Lbrahim et al. proposed the use of deep neural network LSTM networks for sound classification of four species of grouper, showing that LSTM has better performance than WMFCC [23]. Zhang et al. used the LSTM-RNN method for automatic detection and classification of marmoset calls and found that the method outperformed DNN and SVM [24]. With the excellent results of convolutional neural networks in the image field, some scholars have attempted to use CNNs for speech classification [25, 26]. Tao et al. converted three marine animal sounds into spectral images and obtained 99% accuracy for classification using AlexNet, demonstrating the excellent performance of CNNs for sound classification tasks [27].Yanling Yin et al. used AlexNet to extract spectrograms for cough sound classification and obtained 96.8% accuracy [52]. Weizheng Shen et al. used MFCC-CNN to obtain 97.72% accuracy in a pig cough sound classification task [53]. Mustaqeem et al. designed a lightweight architecture of a one-dimensional dilation convolutional neural network (DCNN) for efficient real-time SER recognition [28]. In the latest research, attentional mechanisms have demonstrated excellent performance in both computer vision and natural language processing. The attention mechanism can be understood as the full perception of the surrounding environment by humans. Yang et al. used multiple attention mechanisms in a multiscale convolutional model for image classification [29]. L. Dongdong et al. used a bidirectional long short-term memory network with directed self-attention (BLSTM-DSA) to mine diversity in audio signals [30].Kumar Pandey et al. used tensor decomposition to capture information in speech emotions and obtained an emotional speech recognition model that outperforms LSTM and CNN by combining a tensor decomposition network (TFNN) with an attention-gated tensor decomposition neural network (AG-TFNN) [31]. Ziping Zhao and Jinsong Su et al. were able to effectively improve the performance of the model by introducing self-attention by using a self-attention mechanism for the sentiment classification task [32, 33].

The choice of data features has a significant impact on the model accuracy. Husam Ali Abdulmohsin et al. explored the impact of feature extraction methods on speech recognition tasks by modifying the number of standard deviations (SD) on both sides of the mean for feature extraction experiments, and the new extraction method achieves excellent performance in a variety of datasets [34, 35]. Different extraction methods can have different degrees of impact on prediction results. Paul et al. used the GAEF feature extraction method to demonstrate the differences between different methods [36].Arumugam et al. proposed audio feature extraction using enhanced the mel frequency cepstrum coefficient (EMFCC) and enhanced power normalized cepstrum coefficient (EPNCC) methods, which were tested in pairs of music audio and achieved excellent performance in terms of accuracy performance [37]. Garima Sharma et al. discussed audio signal analysis and classification tasks and found that feature extraction methods using modern machine learning methods combined with speech signal processing can provide solutions to many modern problems, showing the importance of feature extraction [38].

3 The proposed approach

3.1 Noise handling

In the sound samples, noise can interfere with the recognition accuracy to a certain extent. By analyzing the domestic pig breeding environment, it is found that the noise mainly comes from the sound generated by the exhaust fan or machine operation, which is manifested as a slight fluctuation of the signal outside the vocal period. It is worth noting that domestic pig vocalization will cover this type of noise when it reaches a certain decibel, and we only need to focus on noise treatment for noise outside the vocal cycle. In our work, we use the Hamming window to attenuate the paraflop signal, which was verified in the study of Veerendra et al [39]. In processing the squawk signal, the squawk signal is divided into multiple short-time frames of length 60,000 by framing the squawk signal. We find that the data length may be lower than the split-frame signal length; then, in this case, the number of frames is set to 1, and for insufficient length, it is filled by zero. Considering the possible discontinuity of multiple short-time frames, the continuity of the voice signal is ensured by panning the adjacent short-time frame signals by a certain length and allowing the signals to overlap with each other. In our work, we set the offset length to 128, the length of the overlapping part of the signal is the length of the short-time frame minus the offset length, and the specific formula for the frame-splitting process is shown below.

$$fn = (N - overlap)/inc = (N - wlen)/inc + 1 \tag{1}$$

where N is the length of the voice signal, overlap is the overlap part, inc is the data frame shift length, wlen is the short-time frame length, and the starting position of each frame signal is shown in the following equation.

$$stratindex = (0 : (fn - 1)) * inc + 1 \tag{2}$$

In the preprocessing, the data were framed to make the signal interval by multiplying the speech signal with the Hamming window, which in turn ensured the global continuity and periodicity and weakened the data at both ends of the frame signal. One of the window functions is shown as follows.

$$f(x) = \begin{cases} 0.54 - 0.46cos(2\pi n/(L-1)), & 0 \le n < L-1 \\ 0, & else \end{cases} \tag{3}$$

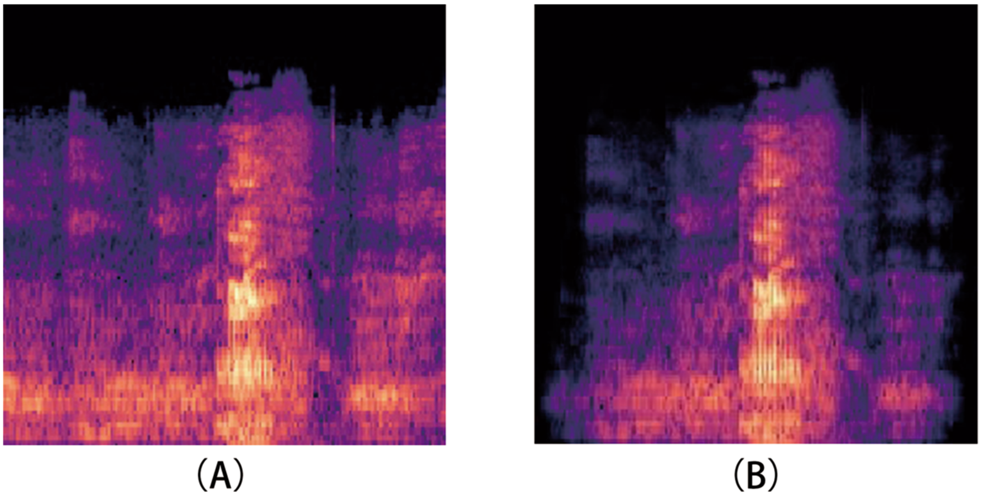


Fig. 1 Comparison of noise reduction data for domestic pig calls

By taking the original domestic pig call as input, the noise data were effectively processed after the framing and windowing process, as shown in Fig. 1.

3.2 Feature extraction

We used eight features for our experiments, which are log-mel spectrogram (LM), MFCC, chroma, spectral contrast, and tonnetz (for the description of subsequent articles, we refer to chroma, spectral contrast, and tonnetz together as CST), MC, LMC, and MLMC, as shown in Fig. 2. In this paper, we adopt a feature extraction method similar to that of [40] to extract chroma, spectral contrast, and tonnetz using the preset parameters of the Librosa library [41]. The channels of MFCC and mel spectrogram are set to 60. MC is aggregated from MFCC and CST. LM

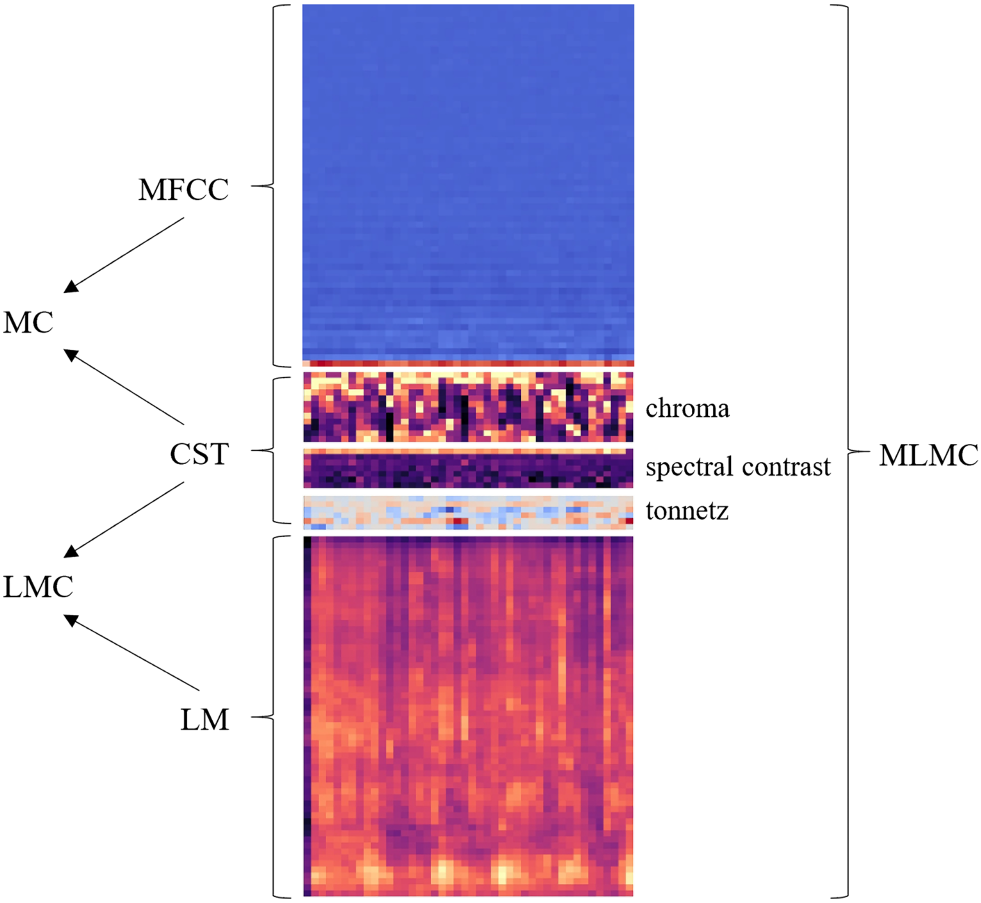


Fig. 2 Schematic diagram of spectrogram features

combines CST to form LMC. MLMC is combined by MC and LMC.

3.3 TransformerCNN

In this paper, we propose a two-stream parallel network structure incorporating the Transformer and CNN and using MLMC as the input to the model, as shown in Fig. 3. It consists of two parts. One is the Transformer module. Because sound classification does not require a decoder language model such as machine translation, we only use the encoder part of the Transformer. The other part is the CNN module, where we use a convolutional neural network with 4 convolutional layers.

In our work, we use MLMC (145x55) as the input features of the model. We use the CNN module and Transformer module for high-dimensional feature extraction, merge the high-dimensional features from these two modules, reorganize them into new feature signals into a linear layer, and merge them to obtain 657x2 dimensional features. The CNN module contains four convolutional layers to obtain the 2x512 dimensional features, and the Transformer module contains multiple attention and encoding layers to obtain the 2x145 dimensional data. By such extraction, the model can obtain two different features, and the combination of features allows the feature signal to contain more information. With the above approach, TransformerCNN can effectively combine the advantages of both, i.e., maintaining the advantages of CNN in local feature extraction and retaining the powerful capability of the Transformer in processing sequential data. In addition, such a network structure brings the advantage of parallel computation that cannot be achieved by RNNs.

3.4 The CNN component of TransformerCNN

CNNs have powerful feature extraction and parallelism, and CNNs with 2D convolutional layers are the gold standard for image processing. CNNs benefit from globally shared weights and pool operations and have translation invariance. When we put an input and its translation into a CNN, we both obtain the same output. Thanks to local connectivity and weight sharing, the convolutional layer has translation equivalence. Since the convolutional kernel of a convolutional layer has a larger activation value for only a specific feature, the convolutional kernel will find the feature and present a larger activation value regardless of where a feature in the feature map of the previous layer has been translated. Due to the above CNN properties, we can imagine MLMC as a 145x55 black and white image with one signal strength channel for feature extraction using CNN, which does not destroy the temporal characteristics of the audio sequence and improves the local feature extraction capability of the model. Spectrograms do not have three channels and large scales like ImageNet, and 2 s of audio data are not suitable for generating 224x224 spectrograms. For the convenience of the narrative, we introduce the network structure of the model with MLMC as the input, and the inputs of other dimensions only need to modify some parameters. The CNN network structure we use is shown in Fig. 4 and Table 1.

As shown in Fig. 4, in the CNN module, we use four 3x3 convolutional kernels to extract the data features, normalize the input features using batch normalization, add the ReLU activation function to solve the possible gradient disappearance problem, promote the sparsity of the network, and alleviate the overfitting problem. The pooling

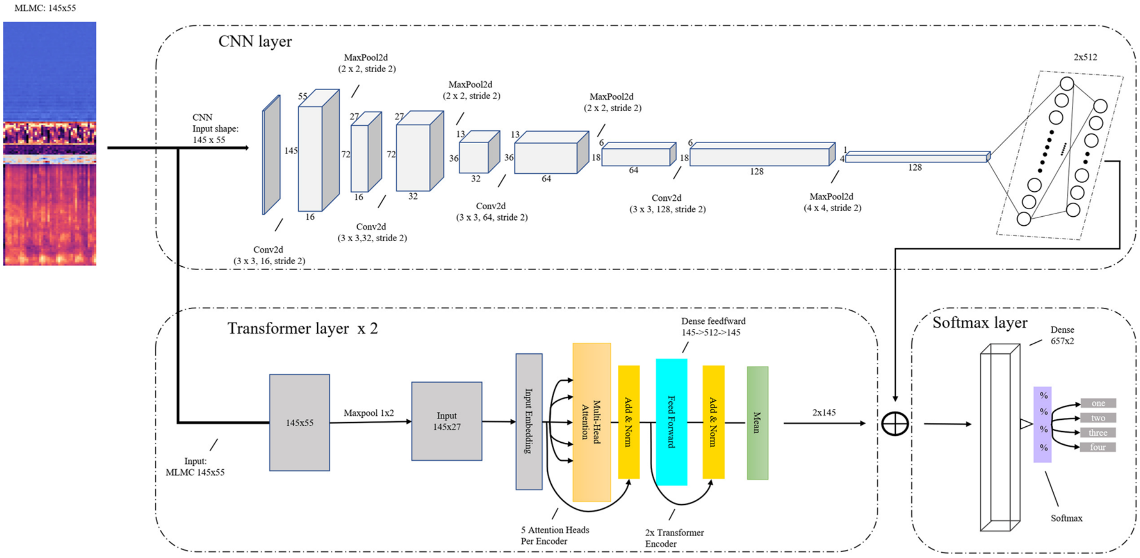


Fig. 3 TransformerCNN structure diagram

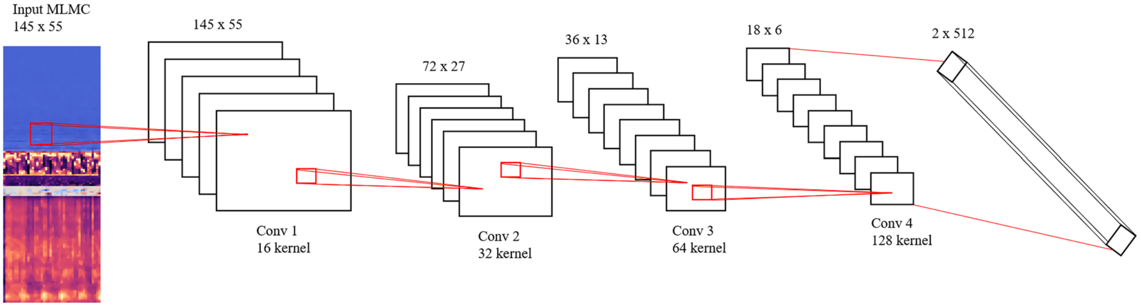


Fig. 4 The architecture and size of feature maps in each convolutional layer

layer is used to reduce the dimensionality of the data, remove the redundant information in the sound features, and make the network nodes work according to the probability by adding dropout. In this work, the node probability is set as 0.5. Through this treatment, the joint adaptation between neuron nodes is effectively weakened, and then the generalization ability of the model is improved.

3.5 The transformer component of TransformerCNN

In the speech classification task, we do not require the model to have linguistic expressiveness in machine translation, so we only use the encoder in the Transformer. We wanted the model to learn to predict the frequency distribution of different categories based on the overall structure of each voice. We could have used an RNN to learn the sequence of sound spectrograms for each emotion, but it would only learn to predict frequency changes based on adjacent time steps. Instead, the Transformer's multihead attention allows the network to look at multiple previous time steps when predicting the next step. This is very appropriate for pig grunt classification since the sound renders the entire frequency sequence, not just at one time step. The Transformer breaks through the shortcomings of RNNs that cannot be parallelized because of their inherent sequential structure. In contrast to CNNs, the attention mechanism dictates that the number of operations required by the Transformer to compute the association between two locations does not grow with distance. Self-attention can produce more interpretable models. We can examine the attention distribution from the model. Each attention head can learn to perform different tasks.

Table 1 CNN network structure with MLMC as input

Input	MLMC (145, 55)
Conv1	Conv2d (3 x 3, 16, stride 2)
	BatchNorm2d (16)
	RELU
	MaxPool2d (2 x 2, stride 2)
	Dropout (0.5)
Conv2	Conv2d (3 x 3, 32, stride 2)
	BatchNorm2d (32)
	RELU
	MaxPool2d (2 x 2, stride 2)
	Dropout,0.5
Conv3	Conv2d (3 x 3, 64, stride 2)
	BatchNorm2d (64)
	RELU
	MaxPool2d (2 x 2, stride 2)
	Dropout (0.5)
Conv4	Conv2d (3 x 3, 128, stride 2)
	BatchNorm2d (128)
	RELU
	MaxPool2d (4 x 4, stride 2)
	Dropout (0.5)

We use a 2-layer encoder layer stack to construct the Transformer encoder, and the main parameters of the encoder layer with MLMC (145x55) as input are shown in Table 2.

As shown in Table 2, in our work, 2 encoder layers are used to build the Transformer encoder, and in each encoder, a multihead attention layer is included, which takes the input MLMC vector data and first passes it through a linear layer, then divides the data equally into 5 heads and computes the attention. By adding multihead attention, the Transformer

Table 2 Transformer encoding layer structure with MLMC as input

Layer	Output shape
Multihead Attention	(-1,2,145), (-1,27,27)
Dropout	(-1,2,145)
Layer Norm	(-1,2,145)
Linear	(-1,2,512)
Dropout	(-1,2,512)
Linear	(-1,2,145)
Dropout	(-1,2,145)
Layer Norm	(-1,2,145)
Transformer EncoderLayer	(-1,2,145)

can capture richer information about the speech features of domestic pigs.

4 Experiment and results

4.1 Dataset

The experimental data were collected from June 2020 to September 2020 at a domestic pig farm in Asbestos County, Ya'an City, Sichuan Province, using farmed pigs as experimental animals. The length and width of the breeding pen were 3 m x 2.4 m, and the height was 1.2 m. The pig house was constructed using cement and bricks. The temperature in the pig house was kept at 23 degrees Celsius, and the maximum temperature was 26 degrees Celsius by supplemental lighting during the experiment. The experimental recording equipment was a Lenovo B610, which had a sound collection range of 10 meters. To avoid discomfort caused by direct contact with the pigs, the recording device was placed at a height of approximately 30 cm from the experimental barn, with a vertical height of 1.5 m. The specific placement is shown in Fig. 5 below. Sound data were recorded at a bit rate of 512 kbps and stored losslessly in wav format. The experimental data were collected during the normal resting time of the pigs, usually from 7:30 a.m. to 10:00 p.m. The feeders performed intensive feeding at 9:00 a.m., 12:00 noon and 6:00 p.m.

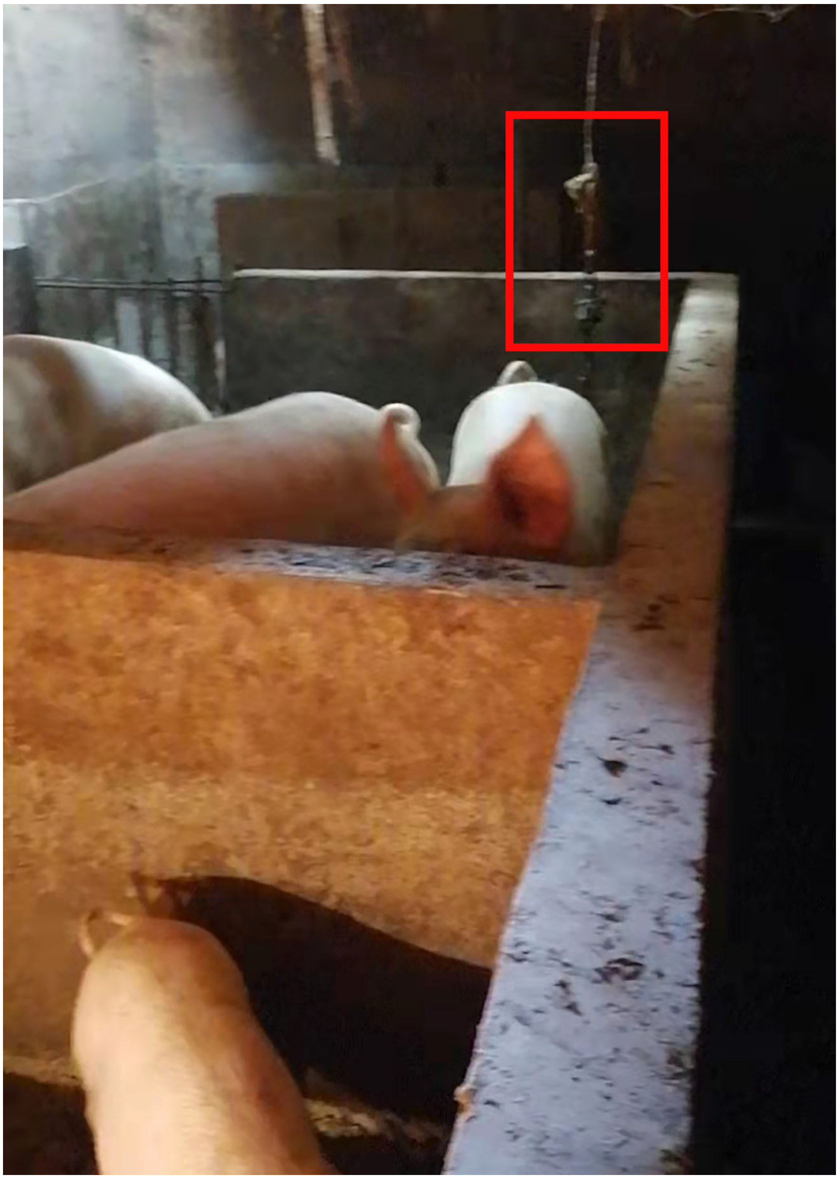


Fig. 5 Schematic diagram of the placement of the collection equipment

Table 3 Distribution of each class and dataset splitting

Class	Train	Test	Total
Calm	800	200	1,000
Feeding	800	200	1,000
Frightened	800	200	1,000
Anxious	800	200	1,000
Total	3200	800	4,000

By consulting with breeding experts[1], we classify the basic behavior of domestic pigs, which includes four categories: calm, feeding, frightened, and anxious. Calm is the normal grunting sound of domestic pigs in a nonstimulated state. The chewing sound made by feeding domestic pigs is defined as feeding. Frightened is the sound produced by intensifying the stimulation of domestic pigs, such as repelling by sticks or injecting vaccination to domestic pigs. Anxious is defined as the grunting sound of domestic pigs in an agitated state when they are not feeding and are hungry. To ensure the correctness of data annotation, we performed simultaneous video acquisition of domestic pigs' behaviors during the recording. In the data classification process, the state information of domestic pigs was confirmed by both video and voice for annotating and classifying grunting data. In the process of data labeling, we found that the duration of each domestic pig vocalization was 0.3 s 1.8 s, so we determined the duration of each data as 2 s. The dataset contains a total of 4,000 2 s of domestic pig grunting data, divided into 4 categories with 1,000 data points in each category, as shown in Table 3 and Fig. 6. Since our data are from real scenes and the data volume is large enough, it is not easy to overfit in the highly parameterized deep neural network model. Therefore, we do not need to perform data augmentation to enhance robustness, such as additive white Gaussian noise (AWGN), and excessive addition of noise will make the model difficult to fit.

4.2 Evaluation metrics

To fully demonstrate the model performance, we evaluated the model using the ACC, AUC, recall, precision and F1 score. The ACC formula is shown below:

$$ACC = \frac{TP + TN}{\text{TP+TN+FP+FN}} \tag{4}$$

AUC represents the area under the ROC curve, and the formula is shown below:

$$AUC = \frac{\sum_{ins_i} \in positiveclass^{rank_{ins_i}} - \frac{M\times(M+1)}{2}}{M \times N} \tag{5}$$

where M is the number of positive samples and N is the number of negative samples, where the negative sample book = total number of samples - number of positive samples.

Recall reflects the proportion of positive cases whose data were correctly determined to the total number of positive cases, and the formula is shown below:

$$TPR = \frac{TP}{\text{TP+FN}} \tag{6}$$

Precision indicates the proportion of samples classified as positive cases that are actually positive cases, and the formula is shown below:

$$Precision = \frac{TP}{\text{TP+FP}} \tag{7}$$

The F1 score is defined as the summed average of precision and recall, and is calculated as shown below:

$$F1\ score = \frac{2 \times Recall \times Precision}{\text{Recall+Precision}} \tag{8}$$

where TP denotes the number of positive classes predicted as positive, TN denotes the number of negative classes predicted as negative, FP denotes the number of false positives predicted as positive, and FN denotes the number of positive classes predicted as negative.

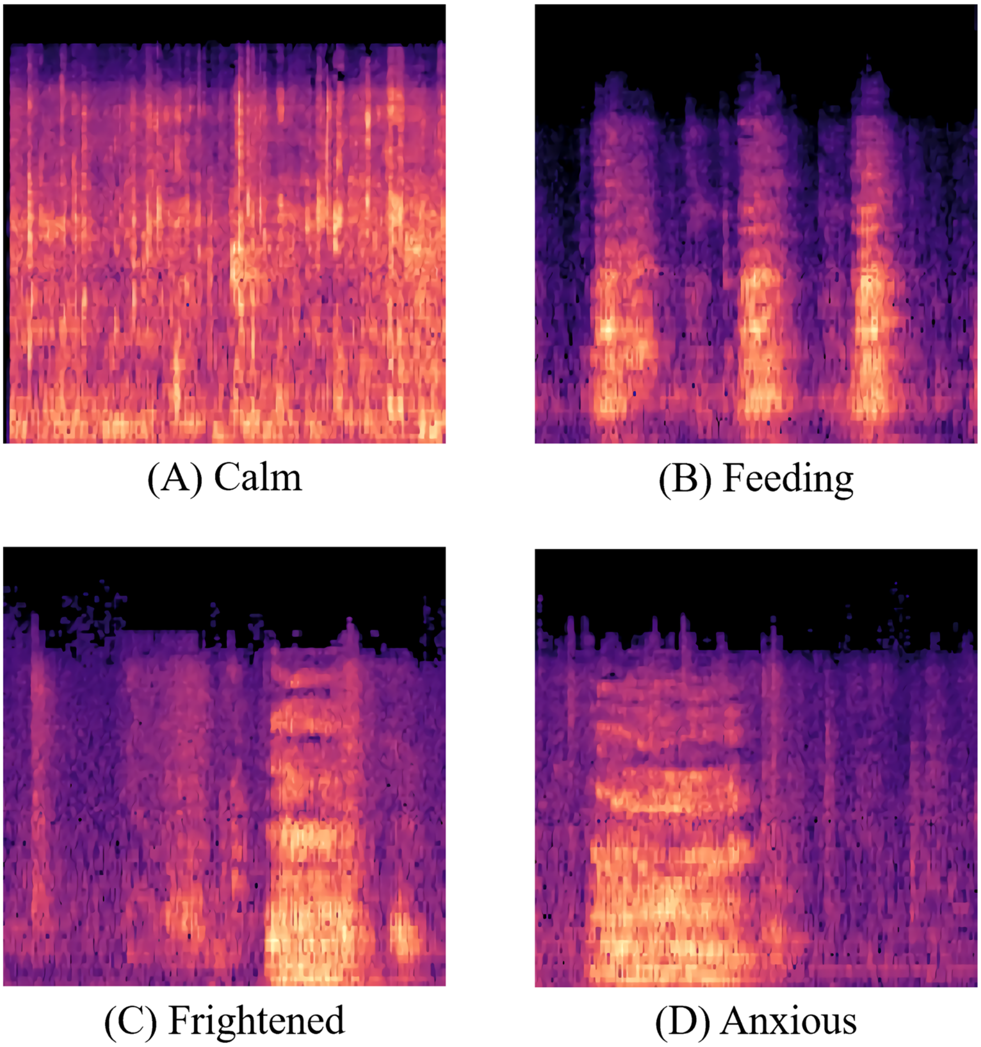
4.3 Experiment

We found that the accuracy of chroma, spectral contrast and tonnetz was low and that the difference between different models was large, as shown in Table 4. This is largely due to the small feature dimension, which is not conducive to feature extraction with a narrow input. Since it is difficult to achieve satisfactory performance with CST features alone, we decided to abandon their use alone. The combination of features achieves excellent performance on RNN, GRU, LSTM, CNN and the Transformer, which indicates that by combining audio features, the input dimension can be expanded to facilitate feature extraction and to include more useful information.

We achieved excellent performance in 8 features using only the Transformer, which demonstrates the power and applicability of the Transformer. In addition, the CNN also shows excellent performance, although not compared to the Transformer, which proves that it is feasible to use a CNN for domestic pig call classification. A simple stack of convolutional layers also has excellent performance and a much smaller computational complexity and number of parameters. As shown in Table 5, TransformerCNN achieves optimal results on all five features. It is worth noting that the accuracy variance in TransformerCNN on different features is small, which proves the excellent robustness and generalization of the proposed model. This

[1] Shuai Surong. Director, Animal Genetic Breeding Branch, Chinese Society of Animal Husbandry and Veterinary Medicine.

Fig. 6 Spectrogram of domestic pig sounds



(A) Calm

(B) Feeding

(C) Frightened

(D) Anxious

Table 4 Comparison of ACC obtained for the eight features on each model

		CST						
	MFCC	C	S	T	LM	MC	LMC	MLMC
Shape	(60, 55)	(12, 55)	(7, 55)	(6, 55)	(60,55)	(85, 55)	(85, 55)	(145,55)
RNN[1]	0.6704	0.5284	0.6079	0.6022	0.7329	0.7506	0.7102	0.7386
GRU[2]	0.7510	0.6466	**0.7672**	0.5113	0.7727	0.7841	0.7272	0.8011
LSTM[2]	0.7727	0.4545	0.5681	0.5573	0.7954	0.7954	0.7727	0.8011
CNN[3]	0.8068	0.6704	0.4147	0.6136	0.8125	0.8522	0.8409	0.8411
Transf[4]	**0.9089**	**0.6931**	0.7509	**0.6170**	**0.9203**	**0.9280**	**0.9312**	**0.9286**

We simply define the Transformer as Transf in this table only

[1] The model consists of 12 blocks of bidirectional cyclic RNN

[2] The model consists of 8 bidirectional circular layers

[3] The model consists of 3 hidden layers of convolutional blocks

[4] For this model, please refer to Table 2

The highlighted text is the optimal value for a more visual presentation of the comparison of the results

Table 5 Comparison of ACC obtained for the eight features on each model

	ACC	AUC	Recall	Precision	F1
MFCC	0.9382	0.9754	0.8304	0.8993	0.8478
LM	0.9554	0.9819	0.8965	0.9023	0.8983
MC	0.9541	0.9810	0.9051	**0.9083**	0.9016
LMC	0.9420	0.9748	0.8505	0.8919	0.8614
MLMC	**0.9605**	**0.9836**	**0.9051**	0.9080	**0.9064**

The highlighted text is the optimal value for a more visual presentation of the comparison of the results

reasonable model structure and efficient featurew extraction capability can adapt to different feature inputs.

The MLMC feature achieves satisfactory performance on all types of models, as shown in Tables 4 and 5. Therefore, we consider it to be the optimal feature for this task. During the experiment, we introduced Grad-CAM to uncover the importance of process features for the results [42]. The results are shown in Fig. 7 below. We visualized the speech signal features of four categories [A,B,C,D], extracted the MLMC (MC, LMC combined features) of the call signal as the model input, and tested the accuracy of the TransformerCNN model; the heatmap for the test set was approximately 95.13% - 96.01%. Through the results, it is found that the combined feature signal CST is important for the judgment of the model in the process of home pig squeak classification, and LM(MFCC, CST combined feature) and LMC(LM, CST combined feature) combined CST also has some importance, and the TransformerCNN reflects excellent feature perception ability in this task.

We used MLMC features as input and compared the most common speech classification models currently used, as shown in Fig. 8 and Table 6, and the experimental results show that the proposed model significantly outperforms the existing models in all metrics.

We used t-SNE to verify the effectiveness of the algorithm, and we limited the number of test data to make the display more appreciable [43]. During the test, we used a model with an accuracy of 96.01% for testing by extracting the dense layer (657x2) data to generate visualization results. Among them, 3D and 2D t-SNE are shown in Fig. 9 below.

In this task, feeding sounds are difficult to distinguish, as they are sometimes similar to frightened sounds and sometimes similar to anxious sounds in the spectrogram, but they convey very different meanings. The attention mechanism of the Transformer can capture the important distinction points of the spectrogram well, and this, together with the local feature extraction function of the CNN, is the main reason why the proposed model can achieve excellent performance in all metrics.

In later work, we explored the performance of the method on other datasets, and given the sparse open dataset of domestic pig calls in this domain, we decided to use other vocal animals for model testing. In that testing effort, five different animal datasets were included,

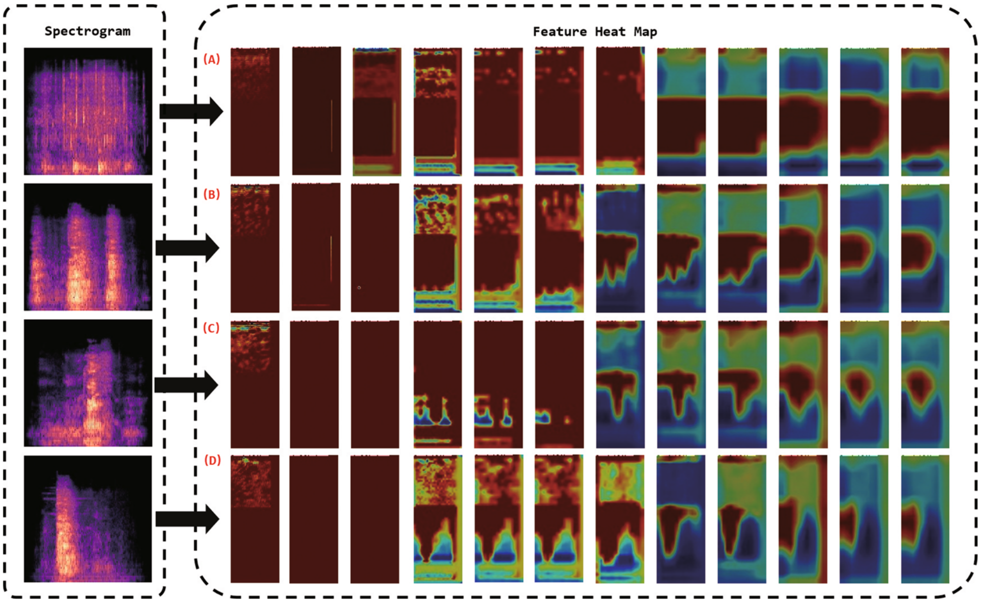


Fig. 7 Visualization of four call type characteristics

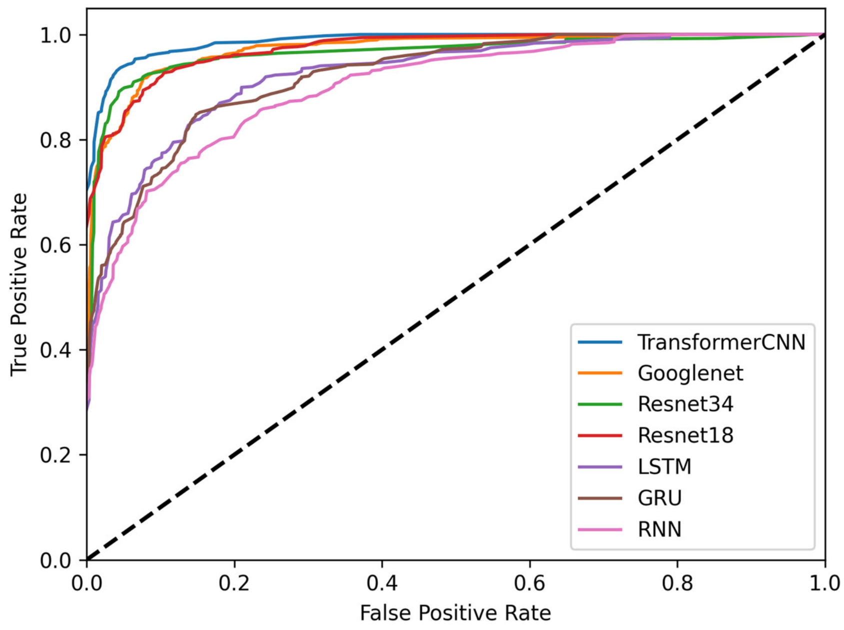


Fig. 8 ROC curves for each model

Table 6 Performance comparison of various models with MLMC as input

	ACC	AUC	Recall	Precision	F1
TransformCNN	**0.9605**	**0.9837**	**0.9052**	**0.9080**	**0.9047**
GoogLeNet	0.8352	0.9716	0.8224	0.8599	0.8214
ResNet34	0.8419	0.9744	0.8419	0.8665	0.7924
ResNet18	0.7471	0.9602	0.7471	0.7908	0.6971
LSTM	0.8017	0.9051	0.8017	0.7998	0.7959
GRU	0.7902	0.9197	0.7902	0.7946	0.7914
RNN	0.7528	0.9020	0.7528	0.7544	0.7529

The highlighted text is the optimal value for a more visual presentation of the comparison of the results

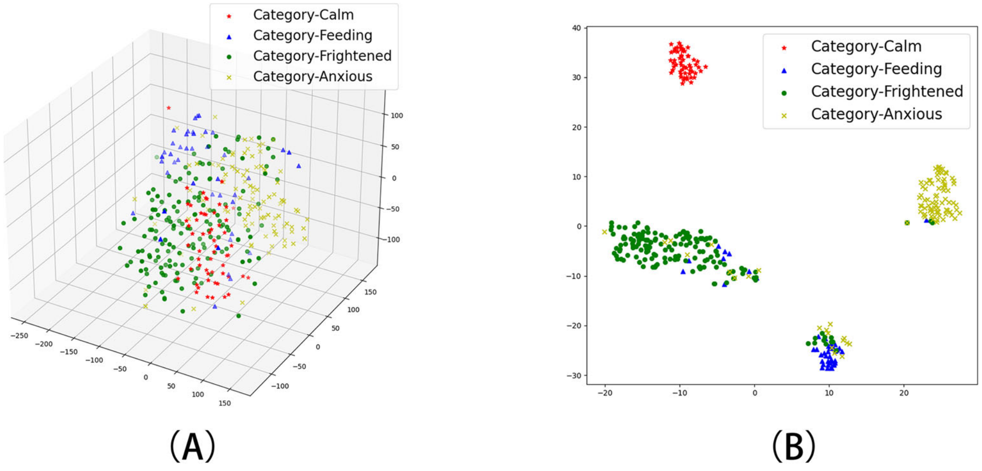


Fig. 9 Visualization of t-SNE in the dense layer

involving species such as birds, cats, marine animals, and our experimental data. One of them, cat vocal data, was collected by Stavros Ntalampiras et al. in 2019 and contains the call data of twenty-one cats in three states, and we extracted 145x80-dimensional MLMC features for the experiments [44]. Birdcall data were collected through the Xeno-Canto website (https://www.xeno-canto.org/) in two datasets, recording eight species (Macropygia phasianella, Burhinus grallarius, Eopsaltria australis, Trichoglossus moluccanus, Chrysococcyx lucidus, Pardalotu striatus, Cacatua galerita, Melithreptus albogularis) and 16 species (Macropygia phasianella, Burhinus grallarius, Eopsaltria australis, Trichoglossus moluccanus, Chrysococcyx lucidus, Pardalotus striatus, Cacatua galerita, Melithreptus albogularis, Phasianus colchicus, Tachybaptus ruficollis, Coturnix coturnix, Anas crecca, Botaurus stellaris, Ardea cinerea, Rallus aquaticus, Porzana pusilla) different birds, by selecting eight types of birds in this dataset to form a sub-dataset for comparing the effect of category on the network. The duration of each birdcall data ranged from 12 s to 100 s, and we split each data into 40 s of data. Each call extracted 225x400 dimensions of call data for the experiment. We also collected calls from five types of marine animals in the Watkins Marine Mammal Sound Database (https://whoicf2.whoi.edu/science/B/whalesounds/), which contains beluga whales, dolphins, false killer whales, sperm whales, and walruses, with each individual data The duration is approximately 2 s - 9 s, and we expand the data to 10 s by padding, which has made it retain all the feature data. We used these four dataset effects to compare with the domestic pig data. The results are shown in Table 7 below.

As shown in Table 7 above, the TransformerCNN still has excellent performance on other datasets, especially in the classification tasks of bird speech and whale speech with few categories, and the recognition accuracy of 95.50% and 92.12% for the two types of data, respectively, showing excellent performance. By adding 8 categories to the bird data, we still have a reliable performance of 90.64%, which has a small impact on the data. Through different experimental demonstrations, the method can effectively extract the feature signals of different animal speech signals, and through the effective combination of the Transformer and CNN, the accuracy of the animal speech classification task is significantly improved, and it has a strong tolerance to different data.

5 Discussion

In our work, we innovatively propose a parallel network of the CNN and Transformer for the task of classifying domestic pig calls. In the process of classifying domestic pig calls, it was found that there were similarities between call features. For the similarity between features, we propose using five different extraction methods (log-mel spectrogram (LM), MFCC, chroma, spectral contrast, and tonnetz) to obtain more information. Three types of combined features, LM, LMC, and MLMC, are generated by combining features (MC is formed by aggregating MFCC and CST). LM combines with CST to form LMC. MLMC is formed by combining MC and LMC). It has been demonstrated that the information obtained by using different extraction methods is not the same. By combining features, the potential features can be found better, which helps the TransformerCNN obtain a better feature signal and better deal with the similarity between information. This is very important for the classification task. In both modules of the TransformerCNN, we use the CNN for local feature extraction. The efficient feature extraction capability of CNN networks is widely supported in classification tasks. Introducing a CNN for local perception between call features is successful and effective for high-dimensional feature extraction. In another Transformer module, we use attention mechanism extraction and global feature extraction to compensate for the CNN shortcomings in time-series data and the inability of RNNs to be parallelized. The Transformer can compensate exactly for this lack of work in time series. The Transformer's multiheaded attention can view the information between previous time steps when making predictions for the next step. This is exactly what the CNN lacks. A reasonable combination of the two methods can more efficiently distinguish the subtle differences between the calls of domestic pigs. In previous work, the exploration of only one dimension or one method has often been focused on, which is not enough in our opinion. By introducing two methods with

Table 7 Performance of the TransformerCNN on different datasets

Dataset	ACC	AUC	Recall	Precision	F1
Pig Sound(4)	0.9605	0.9837	0.9052	0.9080	0.9047
Cat Sound(3)	0.8527	0.8927	0.8523	0.8220	0.8320
Bird Sound(8)	0.9550	0.9376	0.8692	0.8995	0.8704
Bird Sound(16)	0.9064	0.8716	0.8389	0.8528	0.8494
Marine Animals(4)	0.9212	0.8720	0.8512	0.8689	0.8586

different characteristics, the model is allowed to obtain more comprehensive information, achieving the most advanced results in the task of classifying domestic pig calls. We compare the use of the domestic pig call signal in other methods and demonstrate by comparison the importance of the method for the call classification task. State-of-the-art results were obtained in the domestic pig call classification task.

We validated our method with other datasets containing three other animals (bird, whale, and cat), and the TransformerCNN also performed reliably on the other datasets, with 85.27% for cat (3), 95.5% for bird (8), 90.64% for bird (16), and 92.12% for whale (4), which demonstrates the method's state-of-the-art performance in animal call classification task. We also found that the classification process using catcalls was not very accurate, and we compared the catcall dataset with other datasets to attempt to find the reason for this result. We found that the cat sound dataset used simpler traditional methods (HMM, SVM) in the original experiments, while our model had a more complex computational process. We also found that there was a large difference in the total number of datasets, and the training set of the cat sound dataset was small compared to the other datasets. We believe that this is the main reason for the poor results.

We compared the proposed method with the current studies by other scholars. We illustrate this in Table 8, which contains the study team, the number of animals and classifications studied, the method used and the classification results.

By comparing other field studies, our proposed TransformerCNN model has an excellent performance in animal call classification tasks. Most of the studies on pig grunting have focused only on the detection of coughing diseases,

Table 8 Comparison between the approach using the TransformerCNN and other studies

Study	Number of Classes	Method or Classifier	Performance
Yanling Yin et al.[52]	2 classes (pigs)	Spectrogram, AlexNet	Accuracy: 96.8%
Weizheng Shen et al.[53]	2 classes (pigs)	MFCC-CNN	Accuracy: 97.72%
Xie and Zhu[45]	14 classes (bird species)	CNN	F1-Score: 95.95%
Kücüktopcu et al.[46]	21 classes (bird species)	MFCCs, minimum distance classifier	Accuracy: 72%
Zottesso et al.[47]	8 classes (bird species)	Textural features	Accuracy rate: 71%
Zhang et al.[48]	4 classes (bird species)	Spectrogram-frame linear network	F1-Score: 96.9%
Jiang et al.[49]	2 classes (Cetaceans)	CNN	ACC: 95%
Marek et al.[50]	11 classes (Cetaceans)	MFCC,HMM,GMM	ACC: 84.11%
Tao Lu et al. [51]	3 classes (Cetaceans)	AlexNet	ACC: 99.96%
Stavros Ntalampiras et al. [44]	3 classes (cat)	MFCC,temporal modulation features DAG-HMM,HMM,SVM,ESN	ACC: 95.94%
Our Method	4 classes (pigs)	TransformerCNN	ACC: 96.05%
Our Method	3 classes (cat)	TransformerCNN	ACC: 85.27%
Our Method	8 classes (bird species)	TransformerCNN	ACC: 95.50%
Our Method	16 classes (bird species)	TransformerCNN	ACC: 90.64%
Our Method	4 classes (Cetaceans)	TransformerCNN	ACC: 92.12%

often containing only two categories. In our model, we analyze four cases, including two similar cases, which are more computationally complex in comparison but only 0.03% and 1.67% lower in accuracy performance than existing studies, but our work doubles the number of data categories. In the whale call classification task, we also achieved 92.12% accuracy. For the bird call classification task, we used more recognition categories than existing researchers and still had almost consistent accuracy. The TransformerCNN has excellent generalization ability compared to existing studies, which is reflected by the fact that the method remains advanced when applied to different datasets and different features.

In addition, we made the experimental dataset open access and available for researchers to view at (https://figshare.com/articles/dataset/Sow_call_dataset/16940389). We believe that a reasonable division is the basis for normalization. In our work, we divided each collected grunt signal into 2 s according to the vocal characteristics of domestic pigs to make each grunt data contain the complete grunt signal. During our exploration, we found that there are no normalized datasets on domestic pig grunting that have been published by other researchers. To promote research on domestic pig calls, we decided to grant open access to the dataset. Open access to the dataset fills the lack of a standardized dataset in this field and provides a reference for other researchers. We also encourage other researchers to build on this research and propose more novel methods. This study has important implications for behavioral monitoring of domestic pigs, and the excellent recognition performance provides new ideas for the development of smart farming. The study also remedies the problems in the visual method. Farm managers often cannot directly grasp the emotional behavior of domestic pigs through visual methods, which often requires long-time analysis of the behavior, which consumes much time, while the anxious and fearful grunts are the most intuitive expression of the pig's emotions.

The development of this study is groundbreaking for the study of domestic pig call classification tasks, and our study is the first to use a combined Transformer and CNN model for domestic pig sound research. Additionally, the open access to the dataset makes it possible to support the research on domestic pig calls more favorably and to solve the problems that cannot be solved by traditional methods in a groundbreaking way.

6 Conclusions

In this work, we collected and demonstrated the applicability of a standard sound classification dataset for domestic pigs. The dataset is publicly available to provide data support for future research on domestic pig grunts. In this paper, we proposed a parallel neural network based on the Transformer and CNN and compared it with existing grunt classification models. By comparing the experimental results, our proposed model significantly outperforms the existing models in all metrics. In addition, we validated the proposed method on other datasets and compared it with other methods. Through the experimental results, it was proven that the TransformerCNN achieves excellent performance for different data, with low impact on the data and strong generalization ability. The use of the TransformerCNN for animal call classification tasks is an advanced experiment in existing research. It also shows that this novel network structure can effectively improve the feature extraction ability of the model and can better improve the prediction accuracy and robustness by complementing each other between different extraction methods. This work can provide a reference for realistic behavioral and emotional analysis of domestic pigs, as well as experience for future research work.

For the future organization of the work, we have fully considered the existing scholars who have contributed to the visual and acoustic directions. We believe that there is still room for exploration in this work, and we hope to add visual methods for multimodal studies. Through the addition of visual methods, more biological information can be obtained to further explore the behavior and state performance of animals.

Acknowledgements We would like to thank Junliang Zhu for his help in carrying out our work, and we sincerely thank Professor Surong Shuai for his support and guidance of our work.

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Jie Liao, HongXiang Li and Ao Feng. The first draft of the manuscript was written by Jie Liao, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This project was funded by Sichuan Agricultural University Research Grant 040/2121996417.

Availability of data and materials All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval The experimental protocol was developed in accordance with the ethical guidelines of the Declaration of Helsinki and approved by the Animal Welfare Ethics Committee of Sichuan Agricultural University. Written informed consent was obtained from the participants' personal or guardian review number (20200040, 2020/5/23).

Conflict of Interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or nonfinancial interest in the subject matter or materials discussed in this manuscript.

References

1. Ma M, Wang HH, Hua Y, Qin F, Yang J (2021) African swine fever in China: Impacts, responses, and policy implications. Food Policy 102:102065. https://doi.org/10.1016/j.foodpol.2021.102065
2. Gncü S, Güngr C. (2018) The Innovative Techniques in Animal Husbandry, Animal Husbandry and Nutrition, https://www.intechopen.com/chapters/58095
3. Seo J, Sa J, Choi Y, Chung Y, Park D, Kim H (2019) A YOLO-based Separation of Touching-Pigs for Smart Pig Farm Applications. Int Conf Adv Commun Technol 102065:395–401
4. Lee S, Ahn H, Seo J, Chung Y, Park D, Pan S (2019) Practical Monitoring of Undergrown Pigs for IoT-Based Large-Scale Smart Farm, IEEE Access, vol. 7, pp 173796–173810. https://doi.org/10.1109/ACCESS.2019.2955761
5. Hua S, Han K, Xu Z, Xu M, Ye H, Zhou CQ (2021) Image Processing Technology Based on Internet of Things in Intelligent Pig Breeding Mathematical Problems in Engineering
6. Tian M, H Guo H, Chen Q, Wang Y (2019) Ma, Automated pig counting using deep learning. Comput Electron Agric, vol. 163:104840. https://doi.org/10.1016/j.compag.2019.05.049
7. Cowton J, Kyriazakis I, Bacardit J (2019) Automated Individual Pig Localisation, Tracking and Behaviour Metric Extraction Using Deep Learning, IEEE Access, vol. 7, pp 108049–108060. https://doi.org/10.1109/ACCESS.2019.2933060
8. Alameer A, Kyriazakis I, Dalton HA, Miller AL, Bacardit J (2020) Automatic recognition of feeding and foraging behaviour in pigs using deep learning. Biosyst Eng 197:91–104. https://doi.org/10.1016/j.biosystemseng.2020.06.013
9. Li D, Chen Y, Zhang K, Li Z (2019) Mounting behaviour recognition for pigs based on deep learning. Sensors 19(22):1–15. https://doi.org/10.3390/s19224924
10. Zhang Z, Tian J, Wang F, Zhang C (2017) The study on characteristic parameters extraction and recognition of pig cough sound. Heilongjiang Anim Sci Vet Sci 23:1–5
11. Leliveld LMC, Düpjan S., Tuchscherer A, Puppe B (2017) Vocal correlates of emotional reactivity within and across contexts in domestic pigs (Sus scrofa). Physiol Behav 181:117–126. https://doi.org/10.1016/j.physbeh.2017.09.010
12. Vere AJDe, Kuczaj SA (2016) Where are we in the study of animal emotions? Wiley Interdiscip Rev:, Cogn Sci 7(5):354–362. https://doi.org/10.1002/wcs.1399
13. Perry CJ, Baciadonna L (2017) Studying emotion in invertebrates: what has been done, what can be measured and what they can provide. J Exp Biol 220(21):3856–3868. https://doi.org/10.1242/jeb.151308
14. Fanselow MS (2018) Emotion, motivation and function. Curr Opin Behav Sci 19:105–109. https://doi.org/10.1016/j.cobeha.2017.12.013
15. Asher L, Friel M, K Grirrffin LM (2016) Collins, Mood and personality interact to determine cognitive biases in pigs. Biol Lett, vol. 11:12
16. Kang BO, Kwon OW (2016) Combining multiple acoustic models in GMM spaces for robust speech recognition. IEICE Trans Inf Syst 99(3):724–730
17. Marek B. (2021) Trawicki, Multispecies discrimination of whales (cetaceans) using Hidden Markov Models (HMMS), Ecological Informatics, vol. 61. https://www.sciencedirect.com/science/article/pii/S1574954121000145
18. Dias I, Fatih Demirci M, Adnan Y (2020) Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control, vol. 59. https://www.sciencedirect.com/science/article/pii/S1746809420300501
19. Saon G, Picheny M (2017) Recent advances in conversational speech recognition using convolutional and recurrent neural networks. IBM J Res Dev 61(4/5):11–110. https://doi.org/10.1147/JRD.2017.2701178
20. Nasef Mohammed M, Sauber Amr M, Nabil Mohammed M (2021) Voice gender recognition under unconstrained environments using self-attention. Applied Acoustics, p 175
21. Orhan A, Abdulkadir Ş (2021) Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition, Applied Acoustics, vol. 182. https://www.sciencedirect.com/science/article/pii/S0003682X21003546
22. Ashish V, Noam S, Niki P, Jakob U, Llion J, Gomez AN, Łukasz K, Illia P (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, pp 6000–6010
23. Ibrahim AK, Zhuang H, Chérubin LM, Schärer-Umpierre MT, Erdol N (2018) Automatic classification of grouper species by their sounds using deep neural networks. The Journal of the Acoustical Society of America 3:144. https://doi.org/10.1121/1.5054911
24. Zhang YJ, Huang JF, Gong N, Ling ZH, Yu H (2018) Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks. J Acoust Soc Am 144(1):478–487. https://doi.org/10.1121/1.5047743
25. Boddapati V, Petef A, Rasmusson J, Lundberg L (2017) Classifying environmental sounds using image recognition networks. Procedia Comput Sci 112:2048–2056. https://doi.org/10.1016/j.procs.2017.08.250
26. Dian Handy Permana S, Saputra G, Arifitama B, Yaddarabullah, Caesarendra W, Rahim R (2021) Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm, Journal of King Saud University - Computer and Information Sciences https://www.sciencedirect.com/science/article/pii/S1319157821000999
27. LU T, HAN B, YU F (2021) Detection and classification of marine mammal sounds using AlexNet with transfer learning. Ecol Inf 62:1–8. https://doi.org/10.1016/j.ecoinf.2021.101277
28. Mustaqeem K (2021) Soonil, MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach, Expert Systems with Applications, vol. 167 https://www.sciencedirect.com/science/article/pii/S0957417420309131
29. Yang Y, Xu C, Dong F, Wang X (2020) A new multi-scale convolutional model based on multiple attention for image classification. Appl Sci 10(1):1–18. https://doi.org/10.3390/app10010101
30. Dongdong L, Jinlin L, Zhuo Y, Linyu S, Zhe W (2021) Speech emotion recognition using recurrent neural networks

with directional self-attention, Expert Systems with Applications, vol. 173 https://www.sciencedirect.com/science/article/pii/S095741742100124X

31. Sandeep KP, Hanumant SS, Prasanna SRM (2022) Attention gated tensor neural network architectures for speech emotion recognition, Biomedical Signal Processing and Control, vol. 71, Part A https://www.sciencedirect.com/science/article/pii/S1746809421007709
32. Ziping Z, Qifei L, Zixing Z, Nicholas C, Haishuai W, Jianhua T, Björn Schuller W (2021) Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition. Neural Netw 141:52–60. https://www.sciencedirect.com/science/article/pii/S0893608021000939
33. Jinsong S, Jialong T, Hui J, Ziyao L, Yubin G, Linfeng S, Deyi X, Le S, Jiebo L (2021) Enhanced aspect-based sentiment analysis models with progressive self-supervised attention learning, Artificial Intelligence, vol. 296 https://www.sciencedirect.com/science/article/pii/S000437022100028X
34. Ali Abdulmohsin H, Bahjat Abdul wahab H, Mohssen Jaber Abdul hossen A (2021) A new proposed statistical feature extraction method in speech emotion recognition, Computers & Electrical Engineering, vol. 93 https://www.sciencedirect.com/science/article/pii/S0045790621001749
35. Langari S, Marvi H, Zahedi M (2020) Efficient speech emotion recognition using modified feature extraction, Informatics in Medicine Unlocked, vol. 20 https://www.sciencedirect.com/science/article/pii/S2352914820305748
36. Paul D, Su R, Romain M, Sébastien V., Pierre V, Isabelle G (2017) Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. Comput Med Imaging Graph 60:42–49
37. Arumugam M, Kaliappan M (2016) An efficient approach for segmentation, feature extraction and classification of audio signals. Circ Syst 7(4):1–25. https://doi.org/10.4236/cs.2016.74024
38. Sharma G, Umapathy K, Krishnan S (2020) Trends in audio signal feature extraction methods, Applied Acoustics, Vol. 158, https://www.sciencedirect.com/science/article/pii/S0003682X19308795
39. Veerendra M, Bakhar RM (2016) Vani, Robust Blind Beam Formers for Smart Antenna System Using Window Techniques. Procedia Comput Sci 93:713–720. https://www.sciencedirect.com/science/article/pii/S1877050916315204
40. Su Y, Zhang K, Wang J, Madani K (2019) Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors 19(7):1–15. https://doi.org/10.3390/s19071733
41. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, pp 18–25
42. Selvaraju R, Cogswell RM, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM:, Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE Int Conf Comput Vision IEEE 1:618–626. https://doi.org/10.1109/ICCV.2017.74
43. Binu Melit D, Sony G (2020) Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE. Forensic Science International, vol 311. https://doi.org/10.1016/j.forsciint.2020.110194
44. Ntalampiras S, Ludovico LA, Presti G et al (2019) Automatic classification of cat vocalizations emitted in different contexts. Animals 9(8):543. https://doi.org/10.3390/ani9080543
45. Jie X, Mingying Z (2019) Handcrafted features and late fusion with deep learning for bird sound classification. Ecol Inf 52:74–81. https://www.sciencedirect.com/science/article/pii/S1574954118302991
46. Kücüktopcu O, Masazade E, Ünsalan C, Varshney PK (2019) A real-time bird sound recognition system using a low-cost microcontroller. Appl Acoust 148:194–201. https://doi.org/10.1016/j.apacoust.2018.12.028
47. Zottesso RH, Costa YM, Bertolini D, Oliveira LE (2018) Bird species identification using spectrogram and dissimilarity approach. Ecol Inf 48:187–197. https://doi.org/10.1016/j.ecoinf.2018.08.007
48. Xin Z, Aibin C, Guoxiong Z, Zhiqiang Z, Xibei H, Xiaohu Q (2019) Spectrogram-frame linear network and continuous frame sequence for bird sound classification. Ecological Informatics, vol 54. https://doi.org/10.1016/j.ecoinf.2019.101009.
49. Jiang JJ, Bu L, Duan F, Wang X, Liu W, Sun Z, Li C (2019) Whistle detection and classification for whales based on convolutional neural networks. Appl Acoust 150:169–178. https://doi.org/10.1016/j.apacoust.2019.02.007.
50. Trawicki MB (2021) Multispecies discrimination of whales (cetaceans) using Hidden Markov Models (HMMS). Ecological Informatics, vol. 61. https://doi.org/10.1016/j.ecoinf.2021.101223
51. Lu T, Han B, Yu F (2021) Detection and classification of marine mammal sounds using AlexNet with transfer learning. Ecological Informatics, vol 62. https://doi.org/10.1016/j.ecoinf.2021.101277.
52. Yanling Y, Ding T, Weizheng S, Jun B (2021) Recognition of sick pig cough sounds based on convolutional neural network in field situations. Inf Process Agric 8(3):369–379. https://doi.org/10.1016/j.inpa.2020.11.001
53. Weizheng S, Ding T, Yanling Y, Jun B (2020) A new fusion feature based on convolutional neural network for pig cough recognition in field situations Information Processing in Agriculture. https://doi.org/10.1016/j.inpa.2020.11.003

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

Affiliations

Jie Liao[1,2] · Hongxiang Li[1,2] · Ao Feng[1,2] · Xuan Wu[2] · Yuanjiang Luo[2] · Xuliang Duan[1,2] · Ming Ni[1,2] · Jun Li[1,2]

Jie Liao
liaojie@stu.sicau.edu.cn

Hongxiang Li
lhx@stu.sicau.edu.cn

Ao Feng
fengao@stu.sicau.edu.cn

Xuan Wu
wuxuan@stu.sicau.edu.cn

Yuanjiang Luo
201902233@stu.sicau.edu.cn

Xuliang Duan
duanxuliang@sicau.edu.cn

Ming Ni
nm@sicau.edu.cn

[1] College of Information Engineering, Sichuan Agricultural University, 46 Xinkang Road, Yucheng District, Ya'an, 625014, Sichuan province, China

[2] Sichuan Key Laboratory of Agricultural Information Engineering, Sichuan Agricultural University, 46 Xinkang Road, Yucheng District, Ya'an, 10587, Sichuan province, China