Computers and Electronics in Agriculture 197 (2022) 106994

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Fusion of acoustic and deep features for pig cough sound recognition

Weizheng Shen [a], Nan Ji [a], Yanling Yin [a,*], Baisheng Dai [a], Ding Tu [b], Baihui Sun [a,c], Handan Hou [d], Shengli Kou [e], Yize Zhao [f]

[a] School of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China
[b] Tus College of Digit, Guangxi University of Science and Technology, Liuzhou 545000, China
[c] Mudanjiang Branch of Heilongjiang Academy of Agricultural Machinery Sciences, Mudanjiang 157000, China
[d] School of Computer Science, Harbin Finance University, Harbin 150030, China
[e] School of Electrical Engineering and Information, Northeast Agricultural University, China
[f] Department of Computer Science, Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA, USA

ARTICLE INFO

Keywords:
Pig cough
Feature fusion
Time-frequency representations
Convolutional neural networks

ABSTRACT

The recognition of pig cough sound is a prerequisite for early warning of respiratory diseases in pig houses, which is essential for detecting animal welfare and predicting productivity. With respect to pig cough recognition, it is a highly crucial step to create representative pig sound characteristics. To this end, this paper proposed a feature fusion method by combining acoustic and deep features from audio segments. First, a set of acoustic features from different domains were extracted from sound signals, and recursive feature elimination based on random forest (RF-RFE) was adopted to conduct feature selection. Second, time-frequency representations (TFRs) involving constant-Q transform (CQT) and short-time Fourier transform (STFT) were employed to extract visual features from a fine-tuned convolutional neural network (CNN) model. Finally, the ensemble of the two kinds of features was fed into support vector machine (SVM) by early fusion to identify pig cough sounds. This work investigated the performance of the proposed acoustic and deep features fusion, which achieved 97.35% accuracy for pig cough recognition. The results provide further evidence for the effectiveness of combining acoustic and deep spectrum features as a robust feature representation for pig cough recognition.

1. Introduction

Cough is an early sign of respiratory disease in pig houses (Racewicz et al., 2021). Typically, it is monitored by resident specialists at each site. However, it heavily relies on manual experience and the continuity of operations to ensure accuracy and timeliness. Therefore, it is highly beneficial to build a monitoring system for continuous and automatic pig cough detection. However, the task of identifying pig coughs is particularly challenging. The acoustics environment is quite complicated in a pig house, with other interference sounds such as screams and sneezes that have similar acoustic properties to coughs (Benjamin and Yik, 2019). To overcome these challenges, many researchers have focused on investigating acoustic features for providing good discrimination between coughs and non-coughs.

Traditionally, acoustic features have been manually extracted from audio waveforms to distinguish pig cough among various sound categories in pig houses. In particular, Mel-frequency cepstral coefficient (MFCC) was frequently used in pig cough classification and early disease detection (Chung et al., 2013). Besides, frequency and time domain features, such as power spectral density (PSD) and root mean square (RMS) were also considered in the classification (Exadaktylos et al., 2008; Ferrari et al., 2008). However, the performance of a single feature was not satisfactory, especially under field conditions. Although the overall classification accuracy was low, it still inspired researchers to robust the model by enhancing representative features (Guarino et al., 2008). Given by its non-stationary characteristic, sound data have poor robustness when the signal-to-noise ratio is low. In addition, sound content is associated with each sound feature, which leads to restrictions on certain kinds of acoustic features. Consequently, it is desirable to conduct an ensemble of various features based on sound properties in pig houses.

Recently, convolutional neural networks (CNNs) have been successfully applied in the field of sound classification in two ways. One is to complete the classification task in an end-to-end manner. For instance, Yin et al. (2021) employed fine-tuned AlexNet to recognize pig cough by transforming sound signals to time-frequency representation

* Corresponding author.
E-mail address: yinyanling@neau.edu.cn (Y. Yin).

https://doi.org/10.1016/j.compag.2022.106994
Received 28 January 2022; Received in revised form 29 March 2022; Accepted 17 April 2022
Available online 25 April 2022

spectrograms. Nevertheless, this approach has a high cost in terms of time and hardware configuration resources. The other trend is extracting deep features from CNN and feeding the feature vectors to a light classifier. In other words, CNN is regarded as a feature extractor to reduce the extensive computational cost and accelerate classification efficiency. Additionally, STFT spectrograms as typical TFRs were frequently used in existing works related to the realm of animal sound recognition (Ko et al., 2018). Recently, constant-Q transform (CQT) has been widely exploited in speech analysis and environmental sound classification (Pham et al., 2019), giving a promising direction of investigating various TFRs to achieve more valuable features for pig cough recognition.

In this context, this work aimed to provide a robust and effective feature representation of sounds in pig houses to improve pig cough recognition performance. First, acoustic features from different domains were extracted from sound segments. To reduce feature redundancy, a feature selection strategy was adopted to construct a representative feature set. Then, we built a shallow CNN network to extract deep features. The early fusion analysis of layer and TFRs was investigated during the process to select the optimal deep feature representations. Finally, acoustic and deep features were combined and put into SVM for classification. Overall, the contributions of this work are summarized as follows:

(1) A novel acoustic and deep feature fusion framework for pig cough recognition is proposed.

(2) Deep features extracted from shallow CNN architecture are proven to be a feasible approach to enrich the acoustic features.

(3) The proposed method is evidenced to be a representative feature for pig cough recognition, and it outperforms the results of the existing CNN models.

The remainder of this paper is organized as follows. Section 2 describes the relevant works related to our dataset. The methods involved in the experiment are illustrated in Section 3. The experimental results are shown in Section 4. A discussion of the results is provided in Section 5. Finally, the conclusions are presented in Section 6.

2. Datasets

2.1. Animals and housing

The data used in this study were collected in a large commercial pig house in Harbin, Heilongjiang Province, China. One hundred and twenty-eight pigs from the crossbred fattening stage (120d, ~60 kg) of the Northeast Folk and the Great White breed were reared in the barn. Fig. 1 shows the layout of the pig house in our experiment. The barn had a size of 27.5 m × 12.8 m × 3.2 m (length × width × height), and it was subdivided into 21 pens, 12 of which were 4.15 m × 3.6 m (length × width) in two adjacent columns, and 9 of which were 3.6 m × 2.75 m (length × width) in one column, with a half-slatted floor. As shown in Fig. 1, the larger pens (Pen 1–12) contained an average of ten pigs per pen. Two sick pigs with heavy cough were separated in a smaller pen (Pen 13) near the door.

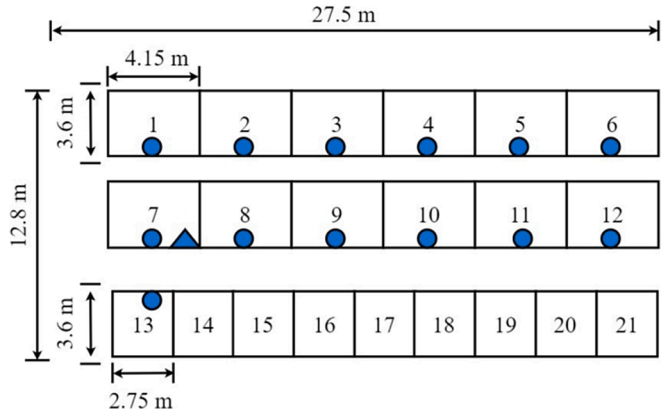

Fig. 1. Layout of the experimental pig house. The blue triangle represents the location of microphone, and the blue circles represent the position of the pigs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Data collection and preprocessing

The sound data were recorded using a microphone (LIQI LM320E, Cardioid electret microphone) connected to the sound card (Conexant Smart Audio HD) of a laptop. The microphone was fixed in Pen 7 adjacent to the door at a height of 1.4 m (approximately 0.8 m from the back of pigs). The recordings were made at a sampling rate of 44.1 kHz with a resolution of 16 bits. The recordings including coughs, healthy pig sounds and other sounds were extracted and labeled with the assistance of a veterinarian. In addition to pig coughs and normal pig sounds (screams, sneezes, and eating), other sounds were also picked up, such as clearing sounds produced by shovels, water flows and human speech. Then, the sounds were passed through a 10th-order Butterworth filter with a cutoff frequency of 100–16000 Hz. In total, 2546 individual sounds were extracted from the recordings, including 1273 coughs and 1273 non-cough segments.

3. Methods

In this work, we aimed to propose distinguishable features to complete pig cough recognition tasks effectively. The flowchart of the proposed method is illustrated in Fig. 2. First, acoustic features were obtained from the pre-processed sound segments. Second, one-dimensional sound signals were transformed into two-dimensional TFRs, and then deep features were extracted from the new CNN architecture based on various TFRs. Subsequently, acoustic and deep features were concatenated by early fusion to provide a set of distinguishable features. Finally, we employed SVM to complete the task of pig cough recognition. The details of each part are presented below.

3.1. Acoustic features

A set of acoustic features was selected, which were inspired by the description presented in Sharma et al. (2020). Specifically, RMS and zero crossing rate (ZCR) were extracted from the time domains (Er, 2020; Toffa and Mignotte, 2021). The RMS measures the volume level of an audio signal, and ZCR is the number of times that the amplitude of a signal crosses zero in a given time interval. We also extracted thirteen MFCCs from the cepstral domain (Chowdhury and Ross, 2020). It is based on human auditory perception to obtain distinctive sound values. The MFCCs are generated by implementing the orthonormal discrete cosine transform for a subband mel-frequency spectrum within a short time period. In addition, the other six features were adopted to capture the characteristics in the frequency domain, including spectral centroid (Flores-Fuentes et al., 2014), spectral flatness (Supradeepa and Weiner, 2012), spectral bandwidth (Xie and Zhu, 2019), spectral rolloff (Luz et al., 2021),spectral contrast (Stilp, 2019), and spectral flux (Fu et al., 2011).

3.2. Deep features

This section described the following main components: construction of TFRs from sound segments and deep feature extraction based on CNN architecture.

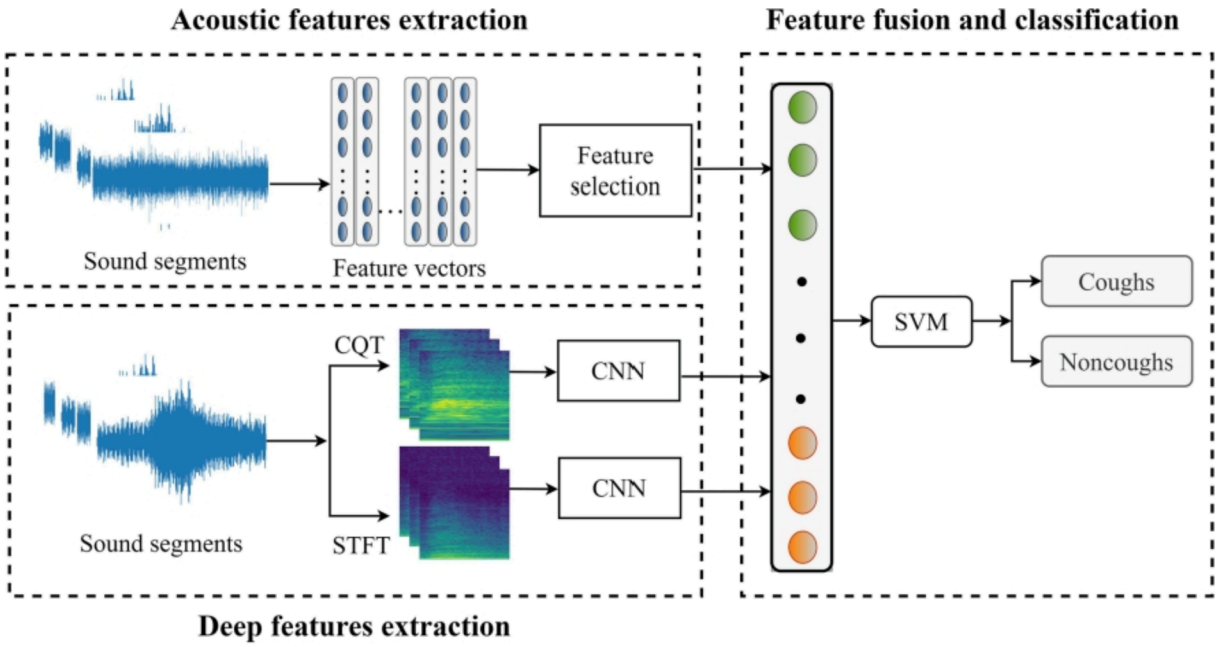

Fig. 2. Illustration of pig cough recognition by the proposed method.

3.2.1. Time-frequency representations

3.2.1.1. STFT. Fourier transform is applied to extract the frequency content of the local part of input signal over a short period of time when generating STFT (Nisar et al., 2016). As shown in Eq. (1), $s[n]$ represents audio signals with window length of L, and $w[t]$ denotes Hanning window function. In general, L is a tradeoff between temporal and frequency resolution and needs to be adjusted to various characteristics of the input signals. In this work, L was set to 2048. The hop size was fixed as $L/2$.

$$STFT[f,t] = \sum_{n=0}^{L-1} s[n] \cdot w[t] e^{-j2\pi fn} \tag{1}$$

3.2.1.2. CQT. The process of constant-Q transform (CQT) is able to be calculated by the following equations. First, constant Q is defined as the center frequency to bandwidth ratio in Eq. (2). Here, f_k is the central frequency of the kth filter. The calculation is given by $f_k = f_1 2^{\frac{k-1}{B}}$, where f_1 is the center frequency of the lowest-frequency bin, and B determines the number of bins per octave, which is associated with the quality factor (Q). In this paper, B was set to 32, and f_1 was set to 22.05 Hz.

$$Q = \frac{f_k}{\delta f_k} = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/B} - 1} \tag{2}$$

Similar to STFT, the CQT spectrograms are extracted using Fourier-based transformation, as presented in Eq. (3). Here, $X^{CQT}(k)$ represents the k component of the constant Q transforms, $x(n)$ represents the input signal, and $w(n,k)$ is the window function with a length of N_k. Given by $N_k = Q\frac{f_s}{f_k}$, N_k denotes the window length that changes with frequency, and f_s represents the sampling frequency.

$$X^{CQT}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x(n) w(n,k) e^{-j\frac{2\pi Q n}{N_k}} \tag{3}$$

A large number of studies have attempted to tackle the issue of sound recognition by employing two-dimensional TFRs attributed to the non-stationary nature of sound. Specifically, STFT spectrograms are the most frequently used in bioacoustic research (Knight et al., 2020). Compared to STFT, CQT provides frequency analysis on a logarithmic scale. The comparison of STFT and CQT spectrograms is shown in Fig. 3. Based on the diversity of characteristics, we assume it is feasible to provide distinct features for classification. Additionally, CQT spectrograms have not been employed in pig cough recognition as part of a CNN model. The above two reasons motivated us to dig into both of TFRs.

3.2.2. CNN architecture

Two reasons prompt us to adopt CNN architecture for the extraction of deep features. First, CNN descriptors are prominent in image processing (Kim et al., 2019; Pham et al., 2021). Second, spectrograms have proved to be feasible for pig cough identification (Hong et al., 2020; Yin

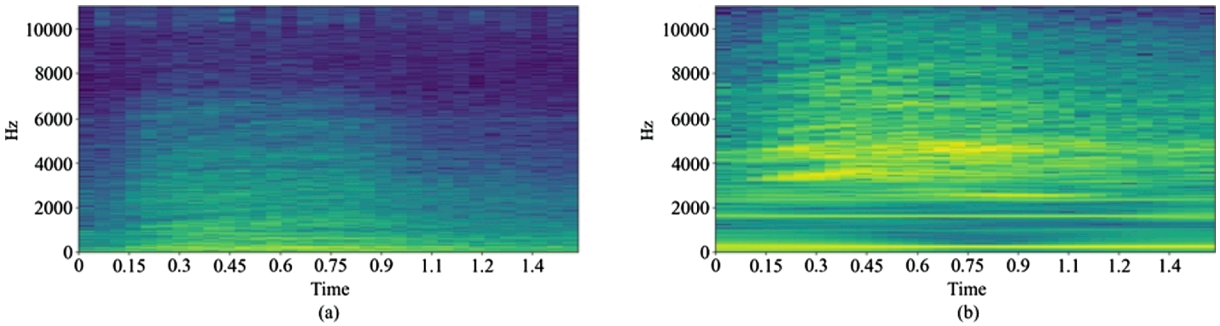

Fig. 3. Two-dimensional spectrograms of pig cough. (a) STFT spectrogram; (b) CQT spectrogram.

et al., 2021).

Lenet-5 is the inspiration for building our model, which has yielded outstanding performance on the MNIST handwritten digits (Lecun et al., 1998). Compared to other state-of-the-art architectures such as VGG (Simonyan and Zisserman, 2015) and Resnet (He et al., 2016), Lenet-5 is a shallow network, which is more suitable for smaller datasets. Besides, our experiments were evaluated on a private and limited dataset. Thus, we proposed the CNN architecture, as shown in Fig. 4.

The first convolutional layer takes the 100 × 100 pixels TFRs segments as inputs, and it has 32 filters with a size of 5 × 5. The second convolutional layer has 48 filters with a size of 5 × 5. Convolutional layers are performed with a stride of 1, followed by a max-pooling layer with 2 × 2 size. Both of the convolutional layers contain rectified linear unit (ReLU) activations. Moreover, two fully connected (FC) layers with 512 neurons are followed compactly. As different domains can be discriminative with various information, 2D convolutions are chosen to convolve both frequency and temporal domain. ReLU is chosen because it facilitates faster convergence and presents no vanishing gradient problem.

For deep feature extraction, we evaluated the performance of two fully connected layers with CQT and STFT separately in the first step. Then, feature vectors extracted from two FCs were fused for classification. Last, contributing to its own characteristics and complementarity, both STFT and CQT were adopted to elaborate the effect of their combination in the task of pig cough recognition. Here, TFRs were fed into the two parallel networks to extract deep features and then fused together to be sent to a classifier.

3.3. Feature selection

Recursive feature elimination (RFE) is a feature selection method that fits a model and systematically eliminates the weakest feature (Demarchi et al., 2020). First, all the selected acoustic features were put into the classifier for calculation of feature importance. Random forest (RF) was chosen as the predictive model in this work due to its unbiased and stable results in different domains. To obtain the best prediction accuracy, the critical part of RF-RFE is to select the least number of features by calculating the importance of each feature. In this work, permutation feature importance (PFI) was used to calculate the importance of features (Wei et al., 2015), as performed by Eq. (4).

$$i_x = a - \frac{1}{K}\sum_{k=1}^{K} a_{k,x} \tag{4}$$

where x represents each feature in the dataset, a represents the accuracy of the RF, K represents the number of repetitions employed to permute a feature, and $a_{k,x}$ is computed by the RF on the data generated by the randomly shuffled column x of the dataset. Here, K was set to 10. Then, a sub feature set was used to predict the accuracy according to the feature importance. The least important feature was removed during the process until no features remain. Finally, optimal combination of features was achieved by the highest predicted accuracy to generate a new feature set.

3.4. Classification and evaluation metrics

For the classifier, we used the open-source LIBSVM implementation provided in the scikit-learn machine learning library. Both the acoustic features and the deep features were fed to SVM for comparison. We prefer to apply SVM over CNN as the classifier for two reasons: first, the datasets are slightly small for a CNN, and SVM is robust for datasets with few samples available for training (Amiriparian et al., 2017). Second, linear and non-linear kernel functions are employed for flexibly fine-tuning the model.

The whole dataset was split into training set and testing set with a training testing split ratio of 8:2. Then, we employed a ten-fold cross validation of grid search to evaluate the proposed approach on the training set to identify the best parameters. Specifically, once the initial values of the hyperparameters were set to train the SVM algorithm, the corresponding models were generated based on the pre-processed training set. The testing set was computed and classified under this best pre-trained model. In grid search, C was set at 1,10 and 100 for a linear model. For a non-linear model, RBF was specified as the kernel type. Gamma was set at 0.1,0.01, and 0.001. The values of C were set as the same with the linear kernel.

Subsequently, the results of evaluation metrics were given by the testing set with the trained model. The metrics for evaluating the models are Accuracy, Recall, Precision, and F1-score. These metrics are calculated by using formulas (5)–(8). Here, true positive, true negative, false positive, and false negative are denoted by TP, TN, FP, and FN, respectively. We defined cough as a positive sample and non-cough as a negative sample.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{5}$$

$$Recall = TP/(TP + FN) \tag{6}$$

$$Precision = TP/(TP + FP) \tag{7}$$

$$F1 - score = 2 \times (Precision \times Recall)/(Precision + Recall) \tag{8}$$

4. Results

4.1. Deep feature extraction

In order to reduce processing time, we resampled the datasets to 22050 Hz. The toolbox of LibROSA was utilized to extract the manually selected features and to generate both STFT and CQT spectrograms as input representations for CNN. The experiments were conducted using a

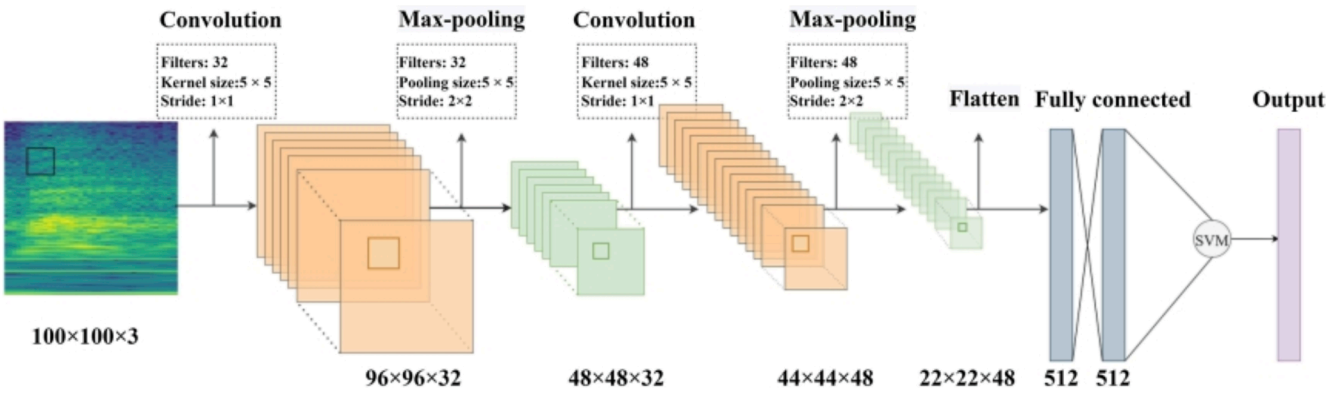

Fig. 4. CNN architecture for deep feature extraction.

configuration of an Intel(R) Core (TM) i7-10750H CPU running at 2.60 GHz with 16 GB of memory, and an NVIDIA GeForce GTX 1650 Ti GPU with 4 GB memory. The software used in this work was Python 3.7.

Table 1 shows the performance of the first fully connected layer (FC1) and the second fully connected layer (FC2) of the proposed CNN architecture, corresponding to the input features of CQT and STFT, respectively. From Table 1, it could be seen that FC1 achieved better recognition performance than FC2. In general, CQT outperformed STFT in terms of recognition.

According to the results, CQT spectrograms were selected to combine the feature vectors of FC1 and FC2 to investigate the effect of layer fusion. In addition, the output of FC1 was employed to fuse the feature vectors from CQT and STFT, which was denoted as TFRs fusion.

4.2. Acoustic feature selection

When using RF-RFE, the number of acoustic features in the model along with their cross-validated test score and variability is plotted in Fig. 5. Obviously, the curve jumped to a better accuracy when five informative features were captured. Then, it gradually grew in accuracy with the number of features added to the model. As seen in Fig. 5, the model with a number of twenty-one features provided the best prediction accuracy. Therefore, we considered all acoustic features as a whole feature set, which was denoted as Acoustic.

4.3. Proposed method

The results of pig cough sound classification achieved by different feature fusion strategies are shown in Table 2. It was shown that deep feature fusion strategies provided a better result compared to acoustic features, except for layer fusion. The fusion of TFRs outperformed single-spectrogram models. As expected, the combination of acoustic and deep features is an effective way to provide a significant performance boost in pig cough recognition. Moreover, TFRs fusion combined with Acoustic achieves 97.35% accuracy in classification.

In addition, two families of CNN architecture, including VGG and Resnet, were investigated for comparison. VGG learns deep information by stacking layers, and Resnet has a shortcut connection from one layer to another. For the two deep learning models, transfer learning was applied by using ImageNet dataset as the initial parameters. Meanwhile, features from pre-trained models were extracted and fed into SVM for two reasons. One was to further compare the impact on different feature extractors based on deep learning architectures, and the other was to evaluate the time cost.

The results of combination with Acoustic and different CNN architectures are shown in Table 3. As summarized in Table 3, VGG16 was better than VGG19 and comparable to ResNet152. Overall, our proposed approach was superior to all of the other model architectures. The results confirm that extracting features from different dimensions for fusion is preferable to extracting homogeneous features from complex networks.

Furthermore, each feature extraction technique was evaluated with respect to its time cost. This was performed because eventual technical deployment would be beneficial for time reduction in computation and operation acceleration in the recognition process (Bishop et al., 2019). A clear distinction between the four compared methods is provided in Fig. 6. Five times were calculated during the process. The points and diamond markers indicated the average and individual time requirement to extract features for both Acoustic features and deep features based on different CNN architectures respectively. The process time of proposed feature fusion method was markedly the fastest compared to the other three, which equated to a sharp decrease in execution time. The difference in maximum execution time was even more pronounced, with over 230.67 s discrepancy between the proposed fusion method and VGG19. Overall, compared to complex neural networks, the fusion of acoustic and deep features extracted from shallow neural network offers a definite advantage in terms of processing time in pig cough recognition.

Table 1
Performance of FC1 and FC2 extracted from CQT and STFT.

Features		Evaluation Metrics			
		Accuracy	Precision	Recall	F1-score
CQT	FC1	92.94%	92.65%	94.03%	93.33%
	FC2	90.39%	92.61%	88.81%	90.67%
STFT	FC1	92.35%	93.21%	92.16%	92.68%
	FC2	89.02%	91.09%	87.69%	89.35%

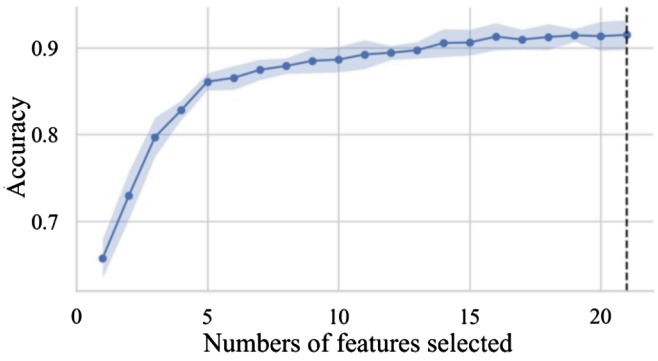

Fig. 5. RF-RFE diagram.

Table 2
The results of all the feature fusion strategies for pig cough recognition.

Features	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-score
Acoustic	93.33%	94.35%	96.84%	95.58%
Layer fusion	92.55%	92.91%	92.91%	92.91%
TFRs fusion	94.51%	95.45%	94.03%	94.74%
Acoustic + Layer fusion	96.27%	96.98%	95.90%	96.44%
Acoustic + TFRs fusion	97.35%	98.41%	96.51%	97.46%

Table 3
Recognition performance with different CNN architectures.

Features	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-score
VGG16	96.57%	97.24%	96.20%	96.72%
VGG19	95.78%	97.18%	94.70%	95.94%
ResNet152	96.18%	97.21%	95.45%	96.33%
Proposed method	97.35%	98.41%	96.51%	97.46%

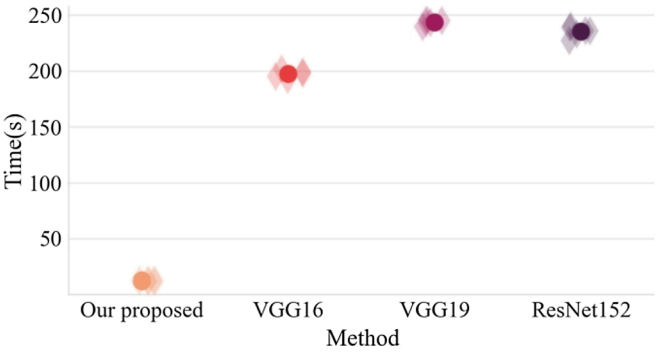

Fig. 6. Point plots of computational time (s) based on VGG, ResNet and our proposed method.

5. Discussion

From Table 1, we confirm that CQT spectrograms show considerable potential as a tool for identifying pig cough sounds under field conditions. Pig screams, as a typical non-cough sound, are common to occur in pig houses. They dominate in high frequencies (5–10 kHz), while pig coughs generally range from 2.5 kHz to 8 kHz. Other sounds such as that of waterflow have lower frequencies below 4 kHz. Based on the characteristics of the various sound components in pig houses, CQT exhibits its distinct advantages: it achieves high-frequency resolution at low frequencies and high temporal resolution at high frequencies. Therefore, we should concentrate on the sensitivity of CQT spectrograms to better convey sound information. For instance, the color of TFRs is an intuitive factor of experimental performance. It was indicated that different CNN models depend on various colors of TFRs (Amiriparian et al., 2018). The limitation of this study was that only one color (viridis) of TFRs was utilized. In the future, we will deep into color selection of TFRs to choose the optimal representation for our CNN model. In addition, it can be found that adding more FC layers to a shallow architecture does not result in performance gain. Instead, due to an incorrect increase in depth, it leads to reduced accuracy for classification. This conclusion is consistent with Basha et al. (2020), which benefits us to improve our model for better results in the future.

From Fig. 5, it can be inferred that although a satisfying result is given by the acoustic feature set, it still has a potential for performance improvement. The curve shows the trends in feature elimination, which means that adding more features may enhance the classification performance. On the other hand, it also implies an expansion of feature dimensionality in acoustics. Consequently, from the slope of the curve in Fig. 5, it prompts us to consider other different sound characteristics and attempts to select more representative features by RF-RFE. In other words, more expanded features could be explored, from which fewer and more suitable feature sets are obtained for pig cough recognition. In our experiment, the representative acoustic features were selected from three domains. In the future, other acoustic features from the three domains will be further extended for pig cough recognition, such as linear prediction cepstral coefficients (LPCCs) and signal peaks.

As presented in Table 2, the fusion of different dimensions of sound features yields promising results, which means that different representations of sound reflect the distinct characteristics of sounds in various domains. Moreover, complementarity between features is demonstrated, which is beneficial for the improvement of pig cough sound recognition. Compared to Yin et al. (2021), a shallow CNN ensembled with acoustic features is potential and sufficient to extract features from pig cough under field conditions. Our findings also motivate us to consider enhancing the shallow CNN model with a machine learning method to achieve a further improved result and to reduce computational complexity in future work.

One of the primary challenges in training deep learning networks is the requirement of a large amount of labeled data, which is a weakness of our model. A logarithmic trend was observed between accuracy and the number of images in the training set for the automated classification of wildlife camera trap images (Tabak et al., 2019). In other words, with regard to limited resources and image processing time consumption, it is possible to effectively improve classification accuracy with an appropriate training set. This motivates us to improve the performance of the model by increasing the training set for further shrinking the gap between human and model accuracy in the future. Another limitation of this research is that the proposed method relies on hand-labeled private datasets. Automated vocalization detection is capable of reducing a certain selection bias associated with time-consuming manual segmentation. Undoubtedly, it is a crucial part in the whole detection system for pig cough recognition. Thus, we will focus on endpoint detection from sound recordings in pig houses and attempt to form an available dataset.

Limited by field conditions, only one microphone was placed in the experimental room. Inevitably, data quality collected by the microphone was impacted by the distance (Ferrari et al., 2013). Some pig sounds close to the microphone were louder, while the far-away data were sounded weaker. However, both coughs and non-coughs had been collected and labeled by an expert in strong-labelled way (Mesaros et al., 2021). The reason was to do our best to cover acoustic variability throughout the experiment in order to overcome the challenges posed by devices. Consequently, to some extent, classification accuracy may be expected to degrade in real-world applications, but classification performance would not be significantly affected in general. Overall, recent advancement in biosensors technology is beneficial for improving livestock health (Neethirajan, 2020). In the future, we will try to improve the audio quality collected in the pig houses by adding more microphones and upgrading the devices.

6. Conclusions

In this work, we extracted deep features from a shallow CNN to enrich acoustic features, in order to improve the recognition performance of pig cough sounds based on the complementary nature of various sounds. We conclude that CQT is more suitable for sound recognition in a pig housing environment than traditional linear STFT. A possible extension to our work may be the application of other samples of bioacoustics for sound classification under field environment, which is of great significance to improve animal welfare and to achieve higher precision livestock farming in the future.

CRediT authorship contribution statement

Weizheng Shen: Conceptualization, Methodology, Funding acquisition, Investigation. **Nan Ji:** Software, Methodology, Writing – original draft, Visualization, Formal analysis. **Yanling Yin:** Writing – review & editing, Funding acquisition, Resources. **Baisheng Dai:** Funding acquisition. **Ding Tu:** Data curation. **Baihui Sun:** Supervision. **Handan Hou:** Supervision. **Shengli Kou:** Supervision. **Yize Zhao:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding: This work was supported by the project of the National Natural Science Foundation of China [grant numbers 32172784, 31902210]; the National Key Research and Development Program of China [grant number 2019YFE0125600]; the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province [grant number UNPYSCT-2020092]; and the China Agriculture Research System of MOF and MARA (CARS-36,CARS-35).

References

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., Schuller, B., 2017. Snore sound classification using image-based deep spectrum features. Interspeech 2017, ISCA, pp. 3512–3516.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Pugachevskiy, S., Schuller, B., 2018. Bag-of-deep-features: noise-robust deep feature representations for audio analysis. International Joint Conference on Neural Networks (IJCNN), IEEE, Rio de Janeiro, pp. 1–7.

Basha, S.H.S., Dubey, S.R., Pulabaigari, V., Mukherjee, S., 2020. Impact of fully connected layers on performance of convolutional neural networks for image classification. Neurocomputing 378, 112–119.

Benjamin, M., Yik, S., 2019. Precision livestock farming in swine welfare: a review for swine practitioners. Animals 9, 133.

Bishop, J.C., Falzon, G., Trotter, M., Kwan, P., Meek, P.D., 2019. Livestock vocalisation classification in farm soundscapes. Comput. Electron. Agric. 162, 531–542.

Chowdhury, A., Ross, A., 2020. Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals. IEEE Trans. Inform. Forensic Secur. 15, 1616–1629.

Chung, Y., Oh, S., Lee, J., Park, D., Chang, H.-H., Kim, S., 2013. Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. Sensors 13, 12929–12942.
Demarchi, L., Kania, A., Ciężkowski, W., Piórkowski, H., Oświecimska-Piasko, Z., Chormański, J., 2020. Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of poland based on airborne hyperspectral and lidar data fusion. Remote Sens. 12, 1842.
Er, M.B., 2020. A novel approach for classification of speech emotions based on deep and acoustic features. IEEE Access 8, 221640–221653.
Exadaktylos, V., Silva, M., Aerts, J.-M., Taylor, C.J., Berckmans, D., 2008. Real-time recognition of sick pig cough sounds. Comput. Electron. Agric. 63, 207–214.
Ferrari, S., Silva, M., Exadaktylos, V., Berckmans, D., Guarino, M., 2013. The sound makes the difference: the utility of real time sound analysis for health monitoring in pigs. In: Aland, A., Banhazi, T. (Eds.), Livestock Housing. Wageningen Academic Publishers, The Netherlands, pp. 407–418.
Ferrari, S., Silva, M., Guarino, M., Aerts, J.M., Berckmans, D., 2008. Cough sound analysis to identify respiratory infection in pigs. Comput. Electron. Agric. 64, 318–325.
Flores-Fuentes, W., Rivas-Lopez, M., Sergiyenko, O., Gonzalez-Navarro, F.F., Rivera-Castillo, J., Hernandez-Balbuena, D., Rodríguez-Quiñonez, J.C., 2014. Combined application of power spectrum centroid and support vector machines for measurement improvement in optical scanning systems. Signal Process. 98, 37–51.
Fu, Z., Lu, G., Ting, K.M., Zhang, D., 2011. A survey of audio-based music classification and annotation. IEEE Trans. Multimedia 13, 303–319.
Guarino, M., Jans, P., Costa, A., Aerts, J.-M., Berckmans, D., 2008. Field test of algorithm for automatic cough detection in pig houses. Comput. Electron. Agric. 62, 22–28.
He, K., Zhang, X., Ren, S., Sun, J., 2016. In: Deep residual learning for image recognition. IEEE, Las Vegas, NV, USA, pp. 770–778.
Hong, M., Ahn, H., Atif, O., Lee, J., Park, D., Chung, Y., 2020. Field-applicable pig anomaly detection system using vocalization for embedded board implementations. Appl. Sci. 10, 6991.
Kim, T., Lee, J., Nam, J., 2019. Comparison and analysis of sampleCNN architectures for audio classification. IEEE J. Sel. Top. Signal Process. 13, 285–297.
Knight, E.C., Poo Hernandez, S., Bayne, E.M., Bulitko, V., Tucker, B.V., 2020. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. Bioacoustics 29, 337–355.
Ko, K., Park, S., Ko, H., 2018. Convolutional feature vectors and support vector machine for animal sound classification. IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Honolulu, HI, pp. 376–379.
Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.
Luz, J.S., Oliveira, M.C., Araújo, F.H.D., Magalhães, D.M.V., 2021. Ensemble of handcrafted and deep features for urban sound classification. Appl. Acoust. 175, 107819.
Mesaros, A., Heittola, T., Virtanen, T., Plumbley, M.D., 2021. Sound event detection: a tutorial. IEEE Signal Process. Mag. 38, 67–83.
Neethirajan, S., 2020. The role of sensors, big data and machine learning in modern animal farming. Sens. Bio-Sens. Res. 29, 100367.
Nisar, S., Khan, O.U., Tariq, M., 2016. An efficient adaptive window size selection method for improving spectrogram visualization. Comput. Intell. Neurosci. 2016, 1–13.
Pham, L., McLoughlin, I., Phan, H., Palaniappan, R., 2019. A robust framework for acoustic scene classification. Interspeech 2019, ISCA, pp. 3634–3638.
Pham, L., Phan, H., Nguyen, T., Palaniappan, R., Mertins, A., McLoughlin, I., 2021. Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework. Digital Signal Process. 110, 102943.
Racewicz, P., Ludwiczak, A., Skrzypczak, E., Składanowska-Baryza, J., Biesiada, H., Nowak, T., Nowaczewski, S., Zaborowicz, M., Stanisz, M., Ślósarz, P., 2021. Welfare health and productivity in commercial pig herds. Animals 11, 1176.
Sharma, G., Umapathy, K., Krishnan, S., 2020. Trends in audio signal feature extraction methods. Appl. Acoust. 158, 107020.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs].
Stilp, C.E., 2019. Auditory enhancement and spectral contrast effects in speech perception. J. Acoust. Soc. America 146, 1503–1517.
Supradeepa, V.R., Weiner, A.M., 2012. Bandwidth scaling and spectral flatness enhancement of optical frequency combs from phase-modulated continuous-wave lasers using cascaded four-wave mixing. Opt. Lett. 37, 3066.
Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., Vercauteren, K.C., Snow, N.P., Halseth, J.M., Di Salvo, P.A., Lewis, J.S., White, M.D., Teton, B., Beasley, J.C., Schlichting, P.E., Boughton, R.K., Wight, B., Newkirk, E.S., Ivan, J.S., Odell, E.A., Brook, R.K., Lukacs, P.M., Moeller, A.K., Mandeville, E.G., Clune, J., Miller, R.S., 2019. Machine learning to classify animal species in camera trap images: applications in ecology. Methods Ecol. Evol. 10, 585–590.
Toffa, O.K., Mignotte, M., 2021. Environmental sound classification using local binary pattern and audio features collaboration. IEEE Trans. Multimedia 23, 3978–3985.
Wei, P., Lu, Z., Song, J., 2015. Variable importance analysis: a comprehensive review. Reliab. Eng. Syst. Saf. 142, 399–432.
Xie, J., Zhu, M., 2019. Investigation of acoustic and visual features for acoustic scene classification. Expert Syst. Appl. 126, 20–29.
Yin, Y., Tu, D., Shen, W., Bao, J., 2021. Recognition of sick pig cough sounds based on convolutional neural network in field situations. Inform. Process. Agric. 8, 369–379.