IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 20, 2019, accepted December 1, 2019, date of publication December 4, 2019,
date of current version December 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957572

Investigation of Different CNN-Based Models for Improved Bird Sound Classification

JIE XIE[1,2,3], (Member, IEEE), KAI HU[2,3], MINGYING ZHU[4], JINGHU YU[1,5], AND QIBING ZHU[2,3]

[1]Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Jiangnan University, Wuxi 214122, China
[2]Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China
[3]School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China
[4]Department of Economics, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[5]School of Mechanical Engineering, Jiangnan University, Wuxi 214122, China

Corresponding author: Jie Xie (xiej8734@gmail.com)

This work was supported in part by the 111 Project, in part by the Fundamental Research Funds for the Central Universities under Grant JUSRP11924, in part by the Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology under Grant FM-2019-06 and Grant FMZ201901, in part by the National Natural Science Foundation of China under Grant 61902154, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK2019043526.

ABSTRACT Automatic bird sound classification plays an important role in monitoring and further protecting biodiversity. Recent advances in acoustic sensor networks and deep learning techniques provide a novel way for continuously monitoring birds. Previous studies have proposed various deep learning based classification frameworks for recognizing and classifying birds. In this study, we compare different classification models and selectively fuse them to further improve bird sound classification performance. Specifically, we not only use the same deep learning architecture with different inputs but also employ two different deep learning architectures for constructing the fused model. Three types of time-frequency representations (TFRs) of bird sounds are investigated aiming to characterize different acoustic components of birds: Mel-spectrogram, harmonic-component based spectrogram, and percussive-component based spectrogram. In addition to different TFRs, a different deep learning architecture, SubSpectralNet, is employed to classify bird sounds. Experimental results on classifying 43 bird species show that fusing selected deep learning models can effectively increase the classification performance. Our best fused model can achieve a balanced accuracy of 86.31% and a weighted F1-score of 93.31%.

INDEX TERMS Bird sound classification, deep learning, class-based late fusion, time-frequency representation.

I. INTRODUCTION

In the past decade, bird sound classification has received increasingly attention due to its worldwide population decline. Therefore, it is becoming ever more necessary to protect bird biodiversity, where monitoring bird population is the first step for the protection. Traditional methods for monitoring birds are time-consuming and costly [32]. Recent advances in wireless acoustic sensor networks and deep learning techniques provide a novel way for monitoring animal populations [3], [34]. Relying on the wireless sensor network, bird sounds can be continuously collected in an open environment, which can then be used for monitoring bird's population [31], [33]. However, various sound sources and low signal-to-noise ratio of those collected recordings become a crucial issue, especially when building an automated robust bird sound classification system [14], [26], [29].

Recently, deep learning models have drawn much attention in constructing the automatic bird sound classification system owning to its high performance [1], [7], [13], [16], [16], [25]. Convolutional Neural Networks (CNNs) [1], [13], [16], [16], Binarized Neural Networks [25], and Convolutional Recurrent Neural Networks [7] have been widely explored for bird sound detection and classification. In addition, data augmentation and preprocessing techniques have been selectively used for further improving bird sound classification performance [2], [15], [23].

Since different deep learning based classification frameworks have been proposed for classifying bird sounds, a direct research question to be asked is whether the

The associate editor coordinating the review of this manuscript and approving it for publication was Jingchang Huang.

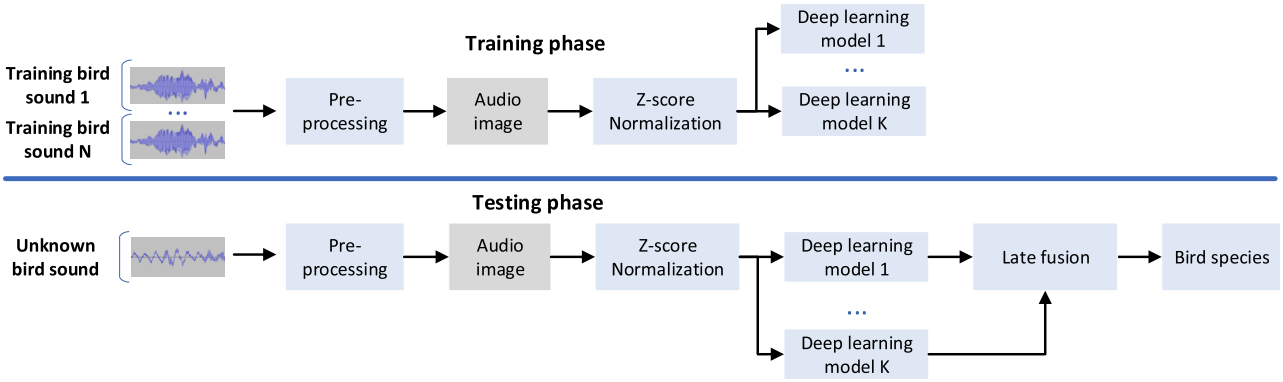


FIGURE 1. [Color online] The flow diagram of our proposed approach. The value *K* denotes the number of selected CNN-based models for the fusion, where it is set to 2, 3, or 4 in this study.

overall classification performance can be improved after fusing those frameworks. Here, the difference among those CNN-based classification frameworks is defined mainly based on (1) the input to CNNs; (2) the architecture of CNNs. Previous studies have demonstrated that fusing different CNNs can improve the classification performance of acoustic events [21], [28], [35]. Yin et al. used three CNN-based models for acoustic classification, where the input of those CNNs were one-dimensional raw waveform modeling, two-dimensional time-frequency image modeling, and three-dimensional spatial-temporal dynamics modeling, respectively [35]. Here, the ensemble focus of the model was the feature. Skashita et al. proposed to use Mel-spectrogram from binaural audio, mono audio, Harmonic-percussive source separation audio, adaptively divided the spectrogram using multiple ways, and learned nine neural networks. Then, those nine neural networks were ensembled for obtaining the final classification result. It is worth noting that the result of [21] was ranked the first in *DCASE 2018 Task 1A*, which demonstrated the effectiveness of the fusion of different CNN-based models. Here, the fusion part of the model was the output probability of each CNN-based model. Su et al. developed a two-stream CNN system for environment sound classification based on decision-level fusion [28]. For those two CNN-based models, the input feature to both CNNs were different. Compared to existing models, the proposed model achieved the highest taxonomic accuracy on the *UrbanSound8K* dataset.

Considering the input features to CNNs for bird sound classification, Duan et al. has claimed that bird calls typically included five categories of acoustic component: lines (at any angle), blocks, warbles, oscillations and stacked harmonics [6]. Here, acoustic component denotes the basic element of audible events that are attributable to a particular source (a bird call). Therefore, we assume that accurately discriminating those acoustic components can help improve the bird call classification performance. Following this assumption, Dong et al. extracted spectral ridge features for similarity-based bird call retrieval [4]. The performance using the spectral ridge method was better than the structure tensor method and the histogram of gradients. This result indicated that carefully capturing the acoustic component of bird sounds can be successfully applied for recognizing bird species. In addition, previous studies have demonstrated that characterizing the target acoustic component using well designed features can significantly improve the classification performance [8]–[11].

In this study, we aim to fuse different CNN-based models for improved bird sound classification. Specifically, we use three different TFRs to characterize different components of bird calls: Mel-spectrogram, harmonic-component based spectrogram, and percussive-component based spectrogram. In addition to different features to the CNN, we incorporate a different CNN architecture for improving the final classification performance. Our final classification results are reported by the fusion of selected CNN-based models.

This paper is organized as follows: In section II, we describe the proposed approach for bird sound recognition, which includes data description, feature extraction, and recognition. Section III gives the description of the class-based late fusion method. Section IV reports the experimental results. Section VI presents conclusions and directions for future work.

II. DATA AND METHODS

Our bird sound classification system consists of three modules: preprocessing, feature extraction, and model construction and ensemble (see Fig. 1). The detail of those modules are described as follows.

A. DATASET DESCRIPTION

In this study, we use a public dataset (CLO-43DS), provided by Salamon *et al.* [23] to evaluate our propose method. This dataset includes flight calls of 43 different North American wood-warblers. The number of instance for all bird species is shown in Fig. 2. All audio clips were recorded in different conditions using different recording devices including highly

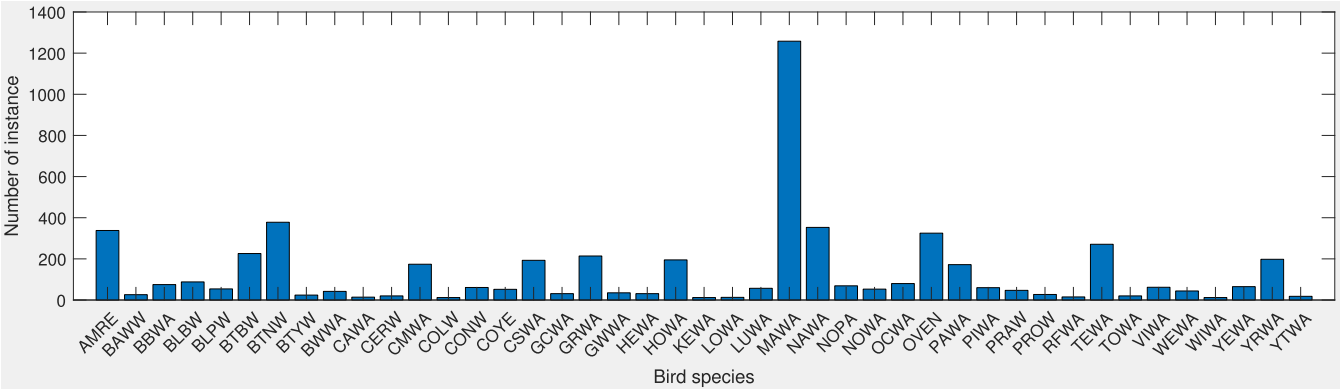


FIGURE 2. [Color online] The number of instance for all bird species in the CLO-43DS dataset. Here, x-axis denotes the abbreviations of common names of those 43 bird species, which can be found in [27].

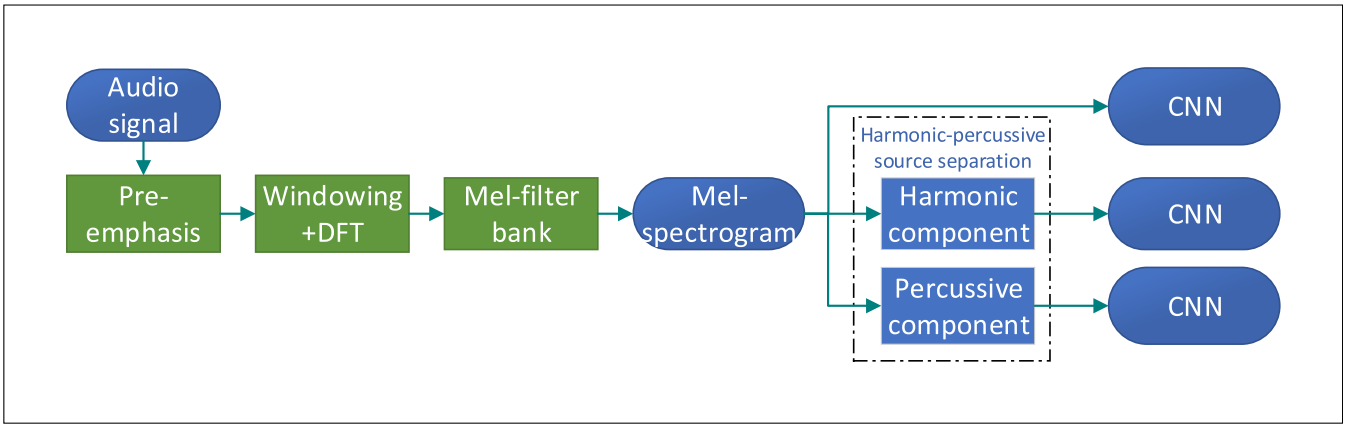


FIGURE 3. [Color online] The time-frequency representation generation procedure. Here, Mel-spectrogram is divided into harmonic and percussive components using Harmonic-percussive source separation. DFT denotes discrete Fourier transform.

directional microphones and omni-directional microphones. In addition, the signal-to-noise ratio varies among different recordings. The detailed information of those wood-warbler species involved in this dataset can be found in [23].

B. PREPROCESSING

For the preprocessing, each audio clip is processed and clipped to only contain a single flight call. The audio clips are sampled at 22.05 kHz. The provided Mel-spectrograms by dataset authors are obtained using 11.6 ms frame size with an overlap of 1.25 ms and 40 Mel bands. It must be noted that 11.6 ms frame size is optimum to analyze flight calls as suggested by [22]. For the feature extraction, we first use three different TFRs to characterize the patterns of bird sounds. Then, z-score normalization is applied as follows.

$$\hat{v}_i = \frac{v_i - \mu_i}{\sigma_i} \tag{1}$$

where v_i denotes the feature vector i, μ_i and σ_i are the mean and standard deviation of each feature vector i, $\hat{v}_i$ is the normalized feature.

C. TIME-FREQUENCY REPRESENTATION

Previous studies have demonstrated that aggregated features or models can achieve better classification performance of bird sounds that single feature or model [17], [34]. However, multiple models can significantly increase the training time for constructing the final classification system. Here, we first ensemble three CNN-based models to classify bird sounds. The only difference among those models is the input feature. Mel-spectrogram have been widely used for classifying bird calls [18], [27], [34]. In this study, we further use harmonic-percussive source separation to split Mel-spectrogram into harmonic-component based spectrogram and percussive-component based spectrogram, which are used to characterize harmonics and oscillations structure of Mel-spectrogram.

In summary, harmonics and oscillations are characterized using harmonic-component and percussive - component based spectrograms, respectively. For blocks, Mel-spectrogram is more suitable for the characterization than harmonic and percussive components. The procedure of calculating input features is shown in Fig. 3.

TABLE 1. Our proposed model specifications. *BN*: Batch Normalization, *ReLU*: Rectified Linear Unit, *K* denotes the number of frame per recording.

Input 40×K×1
3×3 Conv(pad-1, stride-2)-32-BN-ReLU
3×3 Conv(pad-1, stride-2)-32-BN-ReLU
2×2 Max-Pooling + Drop-Out(0.2)
3×3 Conv(pad-1, stride-2)-64-BN-ReLU
3×3 Conv(pad-1, stride-2)-64-BN-ReLU
2×2 Max-Pooling + Drop-Out(0.2)
3×3 Conv(pad-1, stride-2)-128-BN-ReLU
3×3 Conv(pad-1, stride-2)-128-BN-ReLU
2×2 Max-Pooling + Drop-Out(0.2)
Global-Average-Pooling
Dense(512) + Drop-Out(0.2)
43-way Soft-Max

The duration of audio files in CLO-43DS data is different, which cannot be directly used as the input to the CNN. The first method for dealing with the multi-variate varying length audio data is that the signal is repeated from the beginning to force the fixed duration of 2s, which has been used in [30]. The second method is to directly resize the audio image to a fixed size.

D. DEEP LEARNING ARCHITECTURE

The feature learning part of our proposed model follows a *VGG style network* [24], which has been previously used for classifying acoustic scenes [5]. The overall architecture is illustrated in Table 1. This network is trained using Adam optimizer with a learning rate of 10^{-4}. The categorical cross entropy is utilized as the loss function. The batch size is 64 samples and the network is trained with 200 epochs.

In addition to the *VGG style network*, we employ a SubSpectralNet for classifying bird sounds. The architecture is shown in Table 2, which has been used in [19] for classifying acoustic scenes. Here, the SubSpectralNet used is with 20 sub-spectrogram size and 10 Mel-bin hop-size.

III. CLASS-BASED LATE FUSION

To further improve the classification performance over various CNN-based models, we apply a class-based late fusion method. Let's assume, the fusion of decisions from n models for a m-class problem. The sets of models and classes can be presented as $M = M_1, M_2, \ldots, M_n$ and $C = C_1, C_2, \ldots, C_m$. When classifying a test instance x, each model provides a predicted class label along with a posterior probability of the predicted label, which is a measure of the confidence of the decision from that model for that test instance. Let the predicted vector for that instance be $V(x) = V_1(x), V_2(x), \ldots, Vn(x)$, where each $Vi(x) \in C$, and the posterior probabilities be $W_2(x) = W_{21}(x), W_{22}(x), \ldots, W_{2n}(x)$. A decision fusion technique provides a final prediction for x by combining individual predictions $V(x)$.

Our class-based fusion scheme considers the class-based weights $W_1(x)$ and current prediction vector $V(x)$ to make a final prediction for a test instance (x). This method calculates score for each class using following formula.

$$Score_k = \sum_{V_i(x)=C_k} (W_{1ik}), 1 \leq k \leq m, 1 \leq i \leq n \qquad (2)$$

Finally, it selects the class label as final prediction, which has maximum score using the following equation.

$$Label_{final} = C_{argmax_{k=1}^{m}} Score_k \qquad (3)$$

IV. EVALUATION RULE

In this experiment, the dataset was first randomly divided into two parts (85%-15%), where 15% was used as the testing data. Then, we further split the 85% part into 60%-40% for tuning the parameters of CNNs. This process is repeated five times and an averaged classification result is reported. Since the dataset we used is highly imbalanced (see Fig. 2), the performance of our proposed bird call classification system is evaluated using balanced accuracy, weighted precision, weighted sensitivity, weighted specificity, and weighted

TABLE 2. The SubSpectralNet architecture. BN: Batch Normalization, ReLU: Rectified Linear Unit, K denotes the number of frames per recording.

Input 20×K×1	Input 20×K×1	Input 20×K×1
3×3 Conv(pad-1,stride-2)-32-BN-ReLU	3×3 Conv(pad-1,stride-2)-32-BN-ReLU	3×3 Conv(pad-1,stride-2)-32-BN-ReLU
3×3 Conv(pad-1,stride-2)-32-BN-ReLU	3×3 Conv(pad-1,stride-2)-32-BN-ReLU	3×3 Conv(pad-1,stride-2)-32-BN-ReLU
2×2 Max-Pooling+Drop-Out(0.3)	2×2 Max-Pooling+Drop-Out(0.3)	2×2 Max-Pooling+Drop-Out(0.3)
3×3 Conv(pad-1,stride-2)-64-BN-ReLU	3×3 Conv(pad-1,stride-2)-64-BN-ReLU	3×3 Conv(pad-1,stride-2)-64-BN-ReLU
3×3 Conv(pad-1,stride-2)-64-BN-ReLU	3×3 Conv(pad-1,stride-2)-64-BN-ReLU	3×3 Conv(pad-1,stride-2)-64-BN-ReLU
2×2 Max-Pooling+Drop-Out(0.3)	2×2 Max-Pooling+Drop-Out(0.3)	2×2 Max-Pooling+Drop-Out(0.3)
Concatenate		
Global-Average-Pooling		
Dense(128)+Drop-Out(0.3)		
Dense(64)+Drop-Out(0.3)		
43-way Soft-Max		

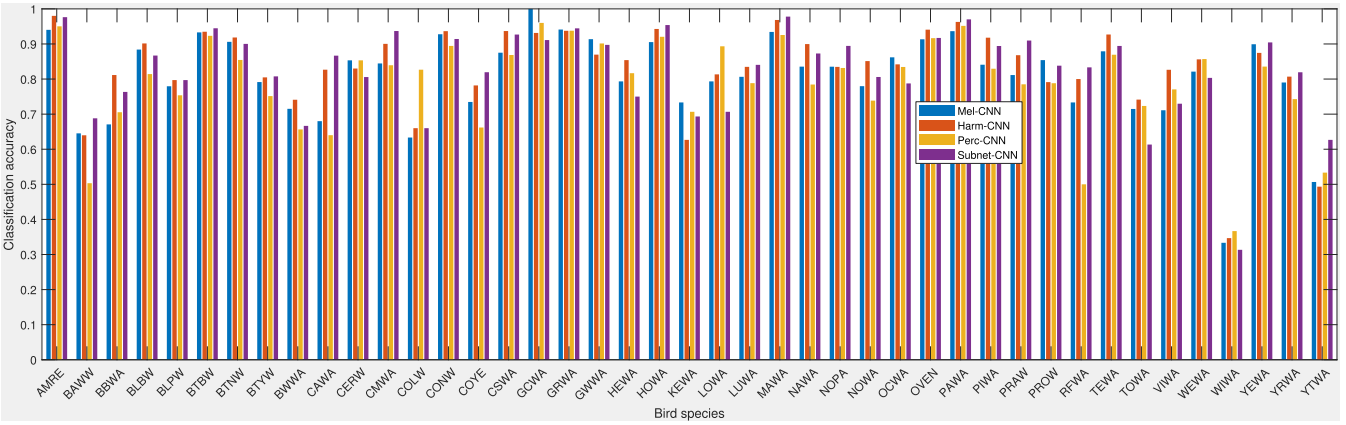


FIGURE 4. [Color online] Class-wise f1-score comparison of 43 bird species using four single CNN-based classification systems. Here, x-axis denotes the abbreviations of common names of those 43 bird species, which can be found in [27]; y-axis denotes the classification accuracy.

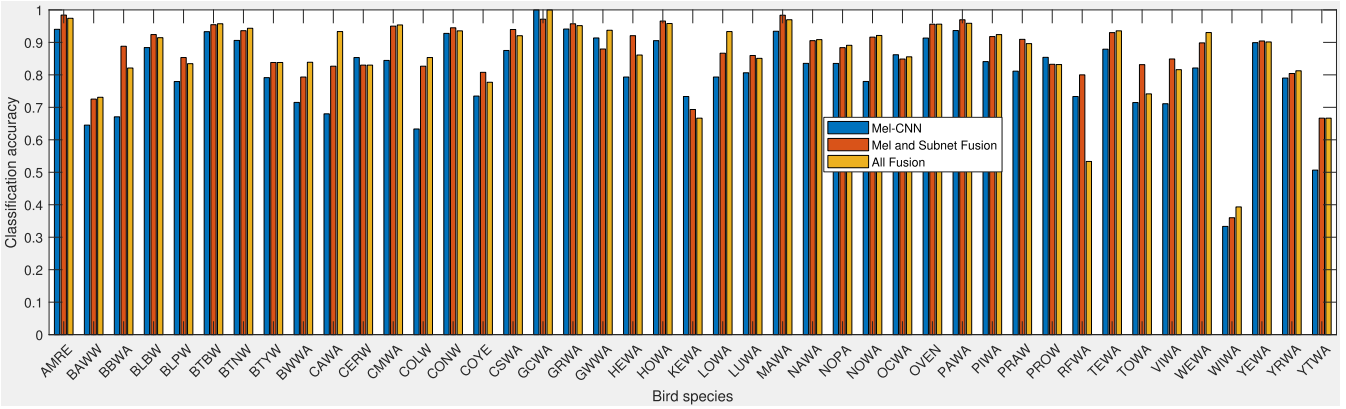


FIGURE 5. [Color online] Class-wise f1-score comparison of 43 bird species with three different classification systems. Here, x-axis denotes the abbreviations of common names of those 43 bird species, which can be found in [23]; y-axis denotes the classification accuracy.

F1-score which are defined as follows:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{TP(i)}{S_i} \tag{4}$$

$$Precision = \sum_{i=1}^{n}\frac{TP(i)}{TP(i)+FP(i)} * r_i \tag{5}$$

$$Specificity = \sum_{i=1}^{n}\frac{TN(i)}{FP(i)+TN(i)} * r_i \tag{6}$$

$$Sensitivity = \sum_{i=1}^{n}\frac{TP(i)}{TP(i)+FN(i)} * r_i \tag{7}$$

$$F1\text{-}score = \sum_{i=1}^{n} 2 \cdot \frac{precision(i) \cdot recall(i)}{precision(i)+recall(i)} * r_i \tag{8}$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative; i is the class index, r_i is the ratio between the number of samples of one class and total number of samples in all classes, S_i is the sample size of class i.

TABLE 3. Classification performance of different methods using single CNN-based model. Here, Mel-CNN, Harm-CNN, and Perc-CNN denote that the input to those CNNs are Mel-spectrogram, harmonic-component based spectrogram, and percussive-component based spectrogram. Subnet-CNN denotes that a SubSpectralNet architecture is used with the Mel-spectrogram as the input.

	Accuracy	Specificity	Sensitivity	Precision	F1-score
Mel-CNN	**83.27%**	**91.63%**	**99.50%**	**92.07%**	**91.42%**
Harm-CNN	80.05%	88.47%	99.07%	88.94%	88.21%
Perc-CNN	78.78%	86.95%	98.98%	87.53%	86.72%
Subnet-CNN	81.60%	91.16%	99.47%	91.62%	90.94%

V. RESULTS AND DISCUSSION

Previous study in audio classification has demonstrated that more mel-bin spectrograms achieve better performance than using lesser mel-bins [20]. However, a recent study using sub-spectrogram for acoustic scene classification indicates that 40 mel-bin spectrogram can achieve comparably superior accuracy when compared to 120 mel-bin spectrogram. In addition, two methods for dealing with multi-variate varying length acoustic data are investigated. A preliminary experiment is first employed using different mel-bin numbers and different methods for addressing multi-variate varying length acoustic data. It is found that using 120 mel-bin

TABLE 4. Classification performance using different fusion strategies.

Mel-CNN	Harm-CNN	Perc-CNN	Subnet-CNN	Accuracy	Specificity	Sensitivity	Precision	F1-score
√	√			84.39%	92.24%	99.42%	92.42%	92.01%
√		√		84.41%	92.44%	99.41%	92.61%	92.15%
√			√	**86.31%**	**93.49%**	**99.63%**	**93.76%**	**93.31%**
√	√		√	85.48%	93.12%	99.53%	93.33%	92.91%
√	√	√		84.76%	92.56%	99.41%	92.60%	92.26%
√		√	√	85.24%	93.17%	99.49%	93.33%	92.89%
√	√	√	√	85.24%	93.05%	99.47%	93.15%	92.77%

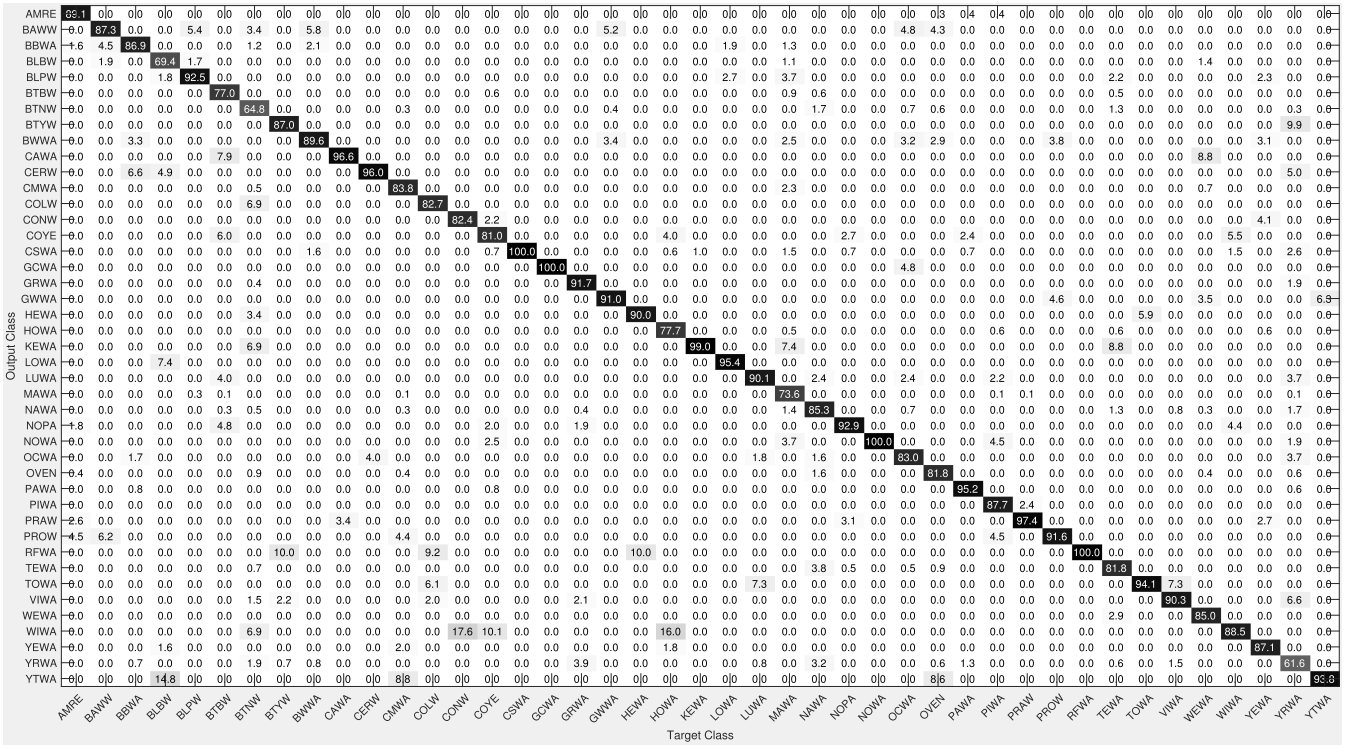


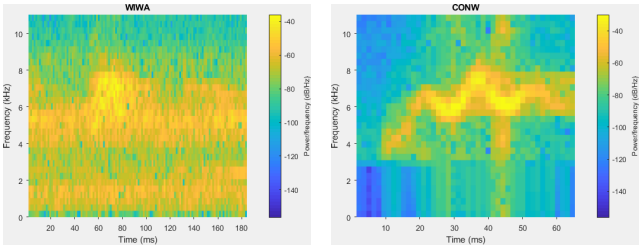
FIGURE 6. Confusion matrix (%) of the best result using the fusion of selected CNN-based models. Here, the *x* and *y* axes denote the code of each bird species to be classified.

(95.5%) can achieve a significantly higher classification accuracy than 40 mel-bin(91.2%), but with more computational load. In addition, since multiple CNN-based models need to be trained in this study, such a high computational load will significantly increase the training time. Furthermore, this study aims to investigate the effectiveness of fusion of different CNN-based models rather than obtaining a single CNN-based model with the best performance, Therefore, we select 40 mel-bin for the subsequent analysis.

For addressing multi-variate varying length acoustic data, repeating signals achieves higher accuracy (91.2%) than resizing spectrograms (88.5%). One reason might be that resizing audio images will lead to the loss of some acoustic patterns. Therefore, the combination of 40 mel-bin spectrogram and repeating signals is selected by balancing the performance and the efficiency.

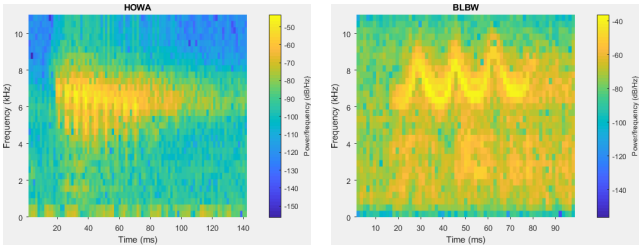
In the work of [30], a triplet sampling was used for generating the triplet spectrograms as the input to CNNs: full spectrogram, harmonic-component based spectrogram, and percussive-component spectrogram. Then, a dynamic triplet loss was used for the classification using multi-scale analysis module.

In our study, rather than combing different spectrograms as the triplet representation, we separately train different CNN-based models using different spectrograms. With those different types of spectrograms, we assume that acoustic components of different bird species can be well characterized, which will conversely increase the classification performance of different bird species. Finally, a class-based late fusion of those different CNN-based models can further improve the classification performance.



(a) NAWA (b) TEWA

FIGURE 7. [Color online] Visual representations of *WIWA* and *CONW*.



(a) HOWA (b) NAWA

FIGURE 8. [Color online] Visual representations of *HOWA* and *BLBW*.

The classification results using each single CNN-based model are shown in Table 3. It is found that Mel-CNN achieves the best overall performance in terms of all evaluation metrics. This result indicates that Mel-spectrogram can well characterize various types of acoustic components. The classification F1-score is shown in Fig. 4 for each bird species using four single CNN-based models. From the figure, we can observe that different models achieve different classification performance for different bird species. Therefore, it is worthwhile investigating the fusion of those different CNN-based models.

The classification results using different fusion strategies are shown in Table 4. From the table, we can have following observations: (1) the fusion of different CNN-based models often leads to the performance improvement; (2) the fusion of more CNN-based models architecture does not always achieve better performance than fewer CNN-based models; (3) different CNN architectures might have better discriminability in recognizing bird calls than different inputs to CNNs.

The best classification balanced accuracy and F1-score using a class-based late fusion of different inputs to the same CNNs are 84.76% and 92.26%, which is higher than 83.27% and 91.42%. Here, the improvement is mainly due to the discriminability of different inputs to the CNNs. The classification F1-score of each bird species is shown in Fig. 5. It is found that the classification performance of *WIWA* is the worst, which is in accordance with [23].

To further improve the classification performance, we not only fuse CNN-based models with different input features, but also fuse another CNN-based model with a different architecture (see Table 2). The best classification F1-score can be up to 93.31%. Different from the *VGG style network*, the SubSpectralNet can be regarded as a filter with different resolutions, which can increase the discriminability of each part of the mel-spectrogram.

To fully understand the classification results, we plot the sum percentage of confusion matrix of the best classification framework (Figure 6). It is observed that 17.6% samples of *WIWA* are confused with *CONW*. In addition, 16% and 14.8% samples of *WIWA* and *YTWA* are confused with *HOWA* and *BLBW*, respectively. The visual patterns of those species are shown in Fig. 7, where we can find that the spectral components of those two species are very similar which make them difficult to be recognized. We also plot the visual patterns of *HOWA* and *BLBW*, both species have similar oscillation components.

VI. CONCLUSION AND FUTURE WORK

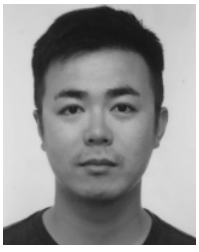
In this study, we investigate the fusion of CNN-based models using three types of TFRs for classifying 43 bird species. For those three TFRs, Mel-spectrogram, Harmonic-component based spectrogram, and Percussive-component based spectrogram are used to capture different acoustic patterns for the same audio file. Then, a VGG style network is used to classify bird species. In addition, we include another SubSpectralNet for classifying bird calls. To further improve the classification performance, a class-based late-fusion is used to selectively combine the output of four individual CNN-based models. The final best balanced accuracy, weighted specificity, weighted sensitivity, weighted precision and weighted F1-score of classifying 43 bird species are 86.31%, 93.49%, 99.63%, 93.76%, and 93.31%. However, since a fusion of different CNN-based models is used, we need to train four different CNN-based models which makes the whole framework less efficient.

Future work aims to build a more efficient classification framework to classify bird species. In addition, only 43 bird species are used in this study. More bird species from different countries will be included to test the robustness of our proposed framework in the future. The number of species to be classified is highly imbalanced (see Fig. 2), which makes imbalance learning worth being investigated. Previously, image data has been investigated for recognizing bird species [12], it it worthwhile fusing both audio and image data for classifying bird species. Another research direction is to build an efficient classification framework for recognizing bird species by intelligently fusing CNN models.

REFERENCES

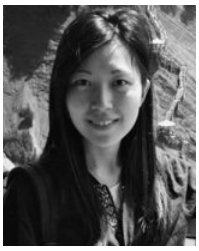
[1] S. Adavanne, K. Drossos, E. Çakir, and T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 1729–1733.
[2] A. Brown, S. Garg, and J. Montgomery, "Automatic and efficient denoising of bioacoustics recordings using MMSE stsa," *IEEE Access*, vol. 6, pp. 5010–5022, 2018.
[3] H. Chen, H. Sun, N. U. R. Junejo, G. Yang, and J. Qi, "Whale vocalization classification using feature extraction with resonance sparse signal decomposition and ridge extraction," *IEEE Access*, vol. 7, pp. 136358–136368, 2019.
[4] X. Dong, M. Towsey, A. Truskinger, M. Cottman-Fields, J. Zhang, and P. Roe, "Similarity-based birdcall retrieval from environmental audio," *Ecolog. Informat.*, vol. 29, pp. 66–76, Sep. 2015.

[5] M. Dorfer, B. Lehner, H. Eghbal-Zadeh, H. Christop, P. Fabian, and W. Gerhard, "Acoustic scene classification with fully convolutional neural networks and I-vectors," in *Proc. DCASE Challenge*, Sep. 2018. [Online]. Available: http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a
[6] S. Duan, M. Towsey, J. Zhang, A. Truskinger, J. Wimmer, and P. Roe, "Acoustic component detection for automatic species recognition in environmental monitoring," in *Proc. 7th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process.*, Dec. 2011, pp. 514–519.
[7] I. Himawan, M. Towsey, and P. Roe, "3D convolutional recurrent neural networks for bird sound detection," in *Proc. DCASE Challenge*, Sep. 2018, pp. 108–112.
[8] J. Huang, X. Zhang, F. Guo, Q. Zhou, H. Liu, and B. Li, "Design of an acoustic target classification system based on small-aperture microphone array," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 7, pp. 2035–2043, Jul. 2015.
[9] J. Huang, X. Zhang, Q. Zhou, E. Song, and B. Li, "A practical fundamental frequency extraction algorithm for motion parameters estimation of moving targets," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 2, pp. 267–276, Feb. 2014.
[10] J. Huang, F. Guo, X. Zu, H. Li, H. Liu, and B. Li, "A novel multipitch measurement algorithm for acoustic signals of moving targets," *Mech. Syst. Signal Process.*, vol. 81, pp. 419–432, Dec. 2016.
[11] J. Huang, S. Xiao, Q. Zhou, F. Guo, X. You, H. Li, and B. Li, "A robust feature extraction algorithm for the classification of acoustic targets in wild environments," *Circuits, Syst., Signal Process.*, vol. 34, no. 7, pp. 2395–2406, Jul. 2015.
[12] Y. Huang and H. Basanta, "Bird image retrieval and recognition using a deep learning platform," *IEEE Access*, vol. 7, pp. 66980–66989, 2019.
[13] Á. Incze, H. Jancsó, Z. Szilágyi, A. Farkas, and C. Sulyok, "Bird sound recognition using a convolutional neural network," in *Proc. IEEE 16th Int. Symp. Intell. Syst. Inform. (SISY)*, Sep. 2018, pp. 000295–000300.
[14] P. Jancovic and M. Köküer, "Bird species recognition using unsupervised modeling of individual vocalization elements," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 5, pp. 932–947, May 2019.
[15] E. C. Knight, S. P. Hernandez, E. M. Bayne, V. Bulitko, and B. V. Tucker, "Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks," *Bioacoustics*, pp. 1–19, 2019, doi: 10.1080/09524622.2019.1606734.
[16] Q. Kong, Y. Xu, and M. D. Plumbley, "Joint detection and classification convolutional neural network on weakly labelled bird audio detection," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 1749–1753.
[17] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," in *Proc. DCASE Challenge*, Sep. 2018, pp. 143–147.
[18] C. H. Lee, C. C. Han, and C. C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1541–1550, Nov. 2008.
[19] S. S. R. Phaye, E. Benetos, and Y. Wang, "SubSpectralNet—Using subspectrogram based convolutional neural networks for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 825–829.
[20] J. K. Piczak, "The details that matter: Frequency resolution of spectrograms in acoustic scene classification," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, 2017, pp. 103–107.
[21] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," in *Proc. DCASE Challenge*, Sep. 2018. [Online]. Available: http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a
[22] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 141–145.
[23] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PLoS ONE*, vol. 11, no. 11, pp. 1–26, Nov. 2016.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations, (ICLR)*, San Diego, CA, USA, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556
[25] J. Song and S. Li, "Bird audio detection using convolutional neural networks and binary neural networks," in *Proc. DCASE Challenge*, Sep. 2018. [Online]. Available: http://dcase.community/challenge2018/task-acoustic-scene-classification-results-a
[26] D. Stowell, E. Benetos, and L. F. Gill, "On-bird sound recordings: Automatic acoustic recognition of activities and contexts," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1193–1206, Jun. 2017.
[27] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, Jul. 2014.
[28] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, 2019.
[29] S. Sumitani, R. Suzuki, N. Chiba, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, "An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques—Harkbird 2.0," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8246–8250.
[30] A. Thakur, D. Thapar, P. Rajan, and A. Nigam, "Deep metric learning for bioacoustic classification: Overcoming training data scarcity using dynamic triplet loss," *J. Acoust. Soc. Amer.*, vol. 146, no. 1, pp. 534–547, 2019.
[31] J. Wimmer, M. Towsey, B. Planitz, I. Williamson, and P. Roe, "Analysing environmental acoustic data through collaboration and automation," *Future Gener. Comput. Syst.*, vol. 29, no. 2, pp. 560–568, Feb. 2013.
[32] J. Wimmer, M. Towsey, P. Roe, and I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness," *Ecolog. Appl.*, vol. 23, no. 6, pp. 1419–1428, 2013.
[33] J. Xie, M. Towsey, J. Zhang, and P. Roe, "Adaptive frequency scaled wavelet packet decomposition for frog call classification," *Ecolog. Inform.*, vol. 32, pp. 134–144, Mar. 2016.
[34] J. Xie and M. Zhu, "Handcrafted features and late fusion with deep learning for bird sound classification," *Ecolog. Inform.*, vol. 52, pp. 74–81, Jul. 2019.
[35] Y. Yin, R. R. Shah, and R. Zimmermann, "Learning and fusing multimodal deep features for acoustic scene categorization," in *Proc. 26th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2018, pp. 1892–1900.



JIE XIE received the B.S. and M.Sc. degrees in communication and information engineering from Shanghai University, in 2009 and 2013, respectively, and the Ph.D. degree in electrical engineering and computer science from the Queensland University of Technology, in 2016. From 2016 to 2017, he was a Postdoctoral Fellow with the Electrical and Computer Engineering Department, University of Waterloo, Canada. He is currently an Associate Professor with Jiangnan University, China. His research interests include driving behavior modeling and bioacoustics monitoring.

KAI HU received the B.S., M.Sc., and Ph.D. degrees from Wuhan University. He is currently an Assistant Professor with the School of Internet of Things, Jiangnan University, China. His current research interests include scientometrics, semantic clustering, and domain knowledge.



MINGYING ZHU is currently pursuing the Ph.D. degree in economics with the University of Ottawa. Her main fields of interest are environmental economics, health economics, behavioral economics, and applied economics. She explores the causal link from short-term air pollution exposure to various health impacts using statistical analysis. For example, she has instrumented for air pollution using plausibly exogenous variations in wind pattern, local fire points, and temperature inversion. Her work focuses on the case of China and uses big data analytics via Python Scraping for people's Web search behaviors. Her work provides a previously unaccounted for benefit of more stringent air quality regulation.

JINGHU YU received the Ph.D. degree from the University of Science and Technology of China, in 2004. He was a Visiting Scholar with Purdue University. He is currently a Professor with the School of Mechanical of Engineering, Jiangnan University. He has published more than ten articles in the past three years and has applied for eight patents. His research interests include design and optimization of bionic robot, CAD, CAE, and CAM.

QIBING ZHU received the Ph.D. degree in mechanical and electronic engineering from Northeast University, in 2006. He was a Visiting Scholar with Michigan State University. He is currently a Professor and a Doctoral Supervisor with the School of Internet of Things, Jiangnan University. He mainly engaged in sensing and detection technology, information perception and intelligent processing, the Internet of Things system integration, and other fields of teaching and research. He is currently a member of the Special Committee on Intelligent Agriculture of the China Automation Society.