



# LBP-based bird sound classification using improved feature selection algorithm

Xunsheng Ji<sup>1</sup> · Kun Jiang<sup>1</sup> · Jie Xie<sup>1,2,3</sup>

Received: 11 October 2020 / Accepted: 13 May 2021 / Published online: 5 July 2021  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

Local binary pattern (LBP)-based features for bird sound classification were investigated in this study, including both one-dimensional (LBP-1D) and two-dimensional (LBP-2D) local binary patterns. Specifically, the discrete wavelet transform was first used as a pooling method to generate multi-level features in both time (LBP-1D-T) and frequency domain (LBP-1D-F) signals. To obtain richer time–frequency information of bird sounds, uniform patterns (LBP-2D) were extracted from the log-scaled Mel spectrogram. To fully exploit the complementarity of different LBP features, a hybrid fusion method was implemented. Next, neighborhood component analysis (NCA) was employed as a feature selection method to remove redundant information in the fused feature set. In order to reduce the running time of NCA and improve the classification accuracy, an improved feature selection method (DSNCA) was proposed. Finally, two machine learning algorithms: K-nearest neighbor and support vector machine were used for classification. Experimental results on 43 bird species of North American wood-warblers indicated that LBP-2D achieved a higher balanced-accuracy than LBP-1D-T and LBP-1D-F (86.33%, 81.05% and 70.02%, respectively). In addition, the highest classification accuracy was up to 88.70%, using hybrid fusion.

**Keywords** Bird sound classification · Local binary pattern · Wavelet decomposition · Hybrid fusion

## 1 Introduction

The bird population worldwide is declining in recent years, due to ecological destruction. As an important indicator of biodiversity, enforcing protection measures for the birds is necessary. To this end, the first step is monitoring bird activity. In nature, many organisms, especially birds, often make sounds. These signals may be an effective way to monitor

individuals or populations. Nowadays, the monitoring of bird biodiversity is mainly carried out by experts, who usually use auditory signals to obtain bird species information (Salamon et al., 2016). However, this process is both time-consuming and costly, and there are many limitations about human monitoring, such as the variability of the monitoring skills of the observer (Emlen & DeJong, 1992; Hutto & Stutzman, 2009; Rosenstock et al., 2002).

The development of acoustic sensor technology has enabled recording of the vocalization of bird species through the autonomous deployment of wireless acoustic sensor networks, which can be analyzed by machine listening techniques to gain information such as composition and number of bird populations (Salamon et al., 2016). The ability to monitor bird biodiversity can thus be greatly improved and this information can be applied to several fields including ecology and conservation biology (Bairlein, 2016; Loss et al., 2015). A unique advantage of sound monitoring is that the composition of the species can be known, regardless of the time and location of the sensor network deployment. A special application of this automated system involves monitoring the night migration of birds.

✉ Jie Xie  
xiej8734@gmail.com

Xunsheng Ji  
jixunsheng@163.com

Kun Jiang  
6191918017@stu.jiangnan.edu.cn

<sup>1</sup> School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, People's Republic of China

<sup>2</sup> Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, People's Republic of China

<sup>3</sup> Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Jiangnan University, Wuxi 214122, People's Republic of China

Some vocalizations from migratory birds are called flight calls, which are specific calls produced by birds during their continuous flight (i.e., nocturnal migration). Different from general bird calls, flight calls are mainly single note sounds, generally less than 200 ms in length (Salamon et al., 2016). By detecting and classifying bird calls during their night migration in the Americas, information about bird species and populations in the migration can be obtained, which is a potential tool for studying the migration movement of birds (Marcarini et al., 2008).

At present, the research on night migration of birds has mainly used weather surveillance radars and isotope tracking. Among them, weather surveillance radars provide the density, speed and direction of bird movement. However, there is no information on species composition actually migration (Farnsworth et al., 2014). Stable isotope tracking of the fluctuations in key elements to define movements of birds (Kays et al., 2015; Stanley et al., 2015) is generally carried out only during the daytime, but its effect at night is very limited (Salamon et al., 2016). Therefore, an automatic bird sound classification system can be used as a supplementary solution for monitoring the night migration of birds, producing information on specific species, unable to be provided by the previous methods.

Significant differences in number and sound activity exist for different bird species, which complicate collection of sounds from certain birds, such as the Kentucky Warbler, Colima Warbler and Wilson's Warbler (Salamon et al., 2016). Therefore, the dataset of flight calls is generally imbalanced. For flight call recognition, the presence of background noise will affect the ability of identify species vocalizations. Even experts in animal acoustics cannot guarantee the correct distinction, leading to label uncertainty (Lostanlen et al., 2019). Meanwhile, for fully automated bird sound monitoring, the system needs to be able to distinguish between flight calls and natural sounds, such as geological sounds (e.g. wind, rain), biological sounds (e.g. cricket, frog) and human sounds (e.g. speech, transportation). For these reasons, flight call recognition remains a problem for further study.

In this study, an end-to-end approach for classifying bird species in continuous recordings using local binary pattern (LBP)-based features is proposed, which consists of three steps: LBP-based feature generation, including LBP-1D and LBP-2D, feature selection using improved neighborhood component analysis (DSNCA) and classification using two machine learning algorithms: K-nearest neighbor (KNN) and support vector machine (SVM). Finally, a hybrid fusion method is performed, to fully exploit the complementarity of different LBP features. To the best of the authors' knowledge, this is the first time a LBP-based audio classification technique has been applied to flight call classification.

The contributions of this work are as follows: (1) A shallow LBP-based approach for bird species classification is proposed. (2) We concatenate multi-level LBP features from both time and frequency domains, which can describe bird sounds in different ways. (3) An improved feature selection method (DSNCA) is proposed, which reduces feature selection time, while improving the classification performance of the model.

The remainder of this article is structured as follows: the related work for bird sound classification is described in Sect. 2. The LBP-based method for flight call recognition, which includes data description, feature extraction, feature selection and classification, is presented in Sect. 3. The class-based late fusion is explained in Sect. 4. Experimental results and discussion are reported in Sect. 5 and 6, respectively. Finally, the conclusion and future work are presented in Sect. 7.

## 2 Related work

Over recent years, feature extraction techniques and classification models for bird sound classification have been a popular research topic. Schrama et al. (2007) proposed extraction of seven acoustic features and searching for the optimal match by Euclidian Distance to classify 12 bird species. The investigated features consisted of call duration, highest frequency, lowest frequency, loudest frequency, average bandwidth, maximum bandwidth and average frequency slope. Marcarini et al. (2008) investigated spectrogram correlation and Gaussian mixture models of Mel frequency cepstral coefficient distributions for flight call recognition. Salamon et al. (2016) introduced the spherical k-means unsupervised learning algorithm to obtain effective time frequency features from the log-scaled Mel spectrogram. The extracted features were used as the input in a SVM classifier and achieved an accuracy of 93.96% on 43 bird species of North American wood-warblers. Following this study, (Salamon et al., 2017) explored the class-based late fusion of spherical k-means model and deep learning. The best classification accuracy obtained was 96.00%.

However, due to differences in number and sound activity among bird species, the flight call dataset is generally imbalanced. Model performance evaluation for each species is challenging using accuracy. Xie et al. (2019) proposed the use of Mel spectrogram as a feature, investigated the effectiveness of fusion of different CNN-based models and achieved the best balanced-accuracy of 86.31%, which is a more suitable performance indicator for imbalanced datasets. However, CNN-based models needed high computational power and are difficult to use in low-power devices, such as mobile phones.

For LBP features, Xie and Zhu (2019) incorporated acoustic features, visual features, and deep learning for bird sound classification, which achieved the best F1-score of 95.95% on 14 bird species. The acoustic features consisted of spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, spectral roll-off, zero-crossing rate, root-mean-square energy and Mel-frequency cepstral coefficients, while visual features contained uniform LBP and histogram of oriented gradients. Zottesso et al. (2018) proposed the use of uniform LBP, robust local binary pattern and local phase quantization as textural features and classification system decrease in sensitivity to the increase in the number of classes by dissimilarity framework, which obtained an identification rate of 71% in the hardest scenario considering 915 bird species. Nanni et al. (2020) presented ensembles of deep learning and handcrafted approaches (LBP, Local Phase Quantization, etc.) for automated audio classification and obtained an accuracy of 94.7% and 99.0% on 46 and 11 bird species, respectively. According to the above mentioned works, a variety of datasets, features and classifiers have been investigated.

### 3 Data and methods

The bird sound classification system presented in this paper consists of five modules: preprocessing, LBP-based feature extraction, feature selection, classifier description and late fusion (Fig. 1). The details of these modules are described in the following paragraphs.

#### 3.1 Data description

A public dataset (CLO-43SD) was used in this work to evaluate the proposed method (Salamon et al., 2016). This dataset consists of 43 different North American wood-warbler species (in the family Parulidae), containing 5,428 flight calls (22.05 kHz sampling rate, mono, 16-bit depth, wav format). The audio clips were recorded in different recording conditions, namely using highly-directional shotgun microphones, using omnidirectional microphones and from captive birds (Lanzone et al., 2009) and are either clean or noisier. Every clip contained a single flight call from one of the 43 bird species. Due to differences in number and sound activity of different bird species, the CLO-43SD dataset is imbalanced, the data distribution of which is shown in Fig. 2.

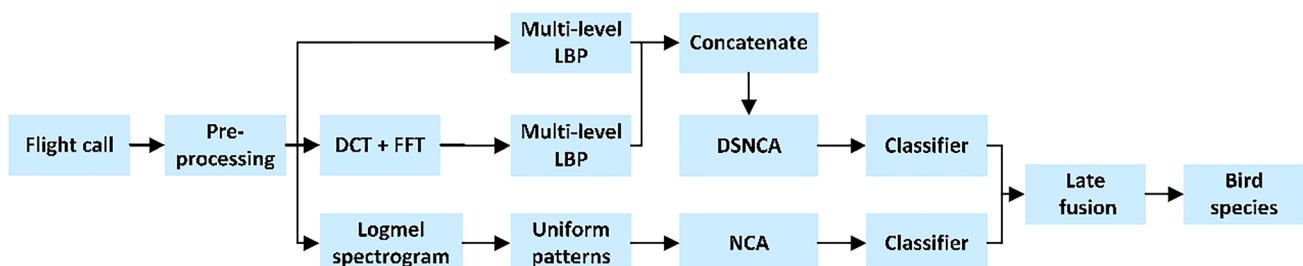


Fig. 1 The flow diagram of our proposed approach

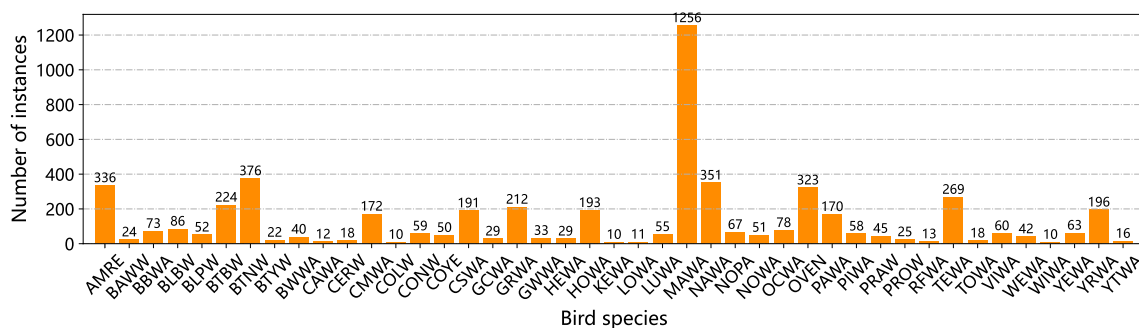


Fig. 2 The data distribution of bird species in the CLO-43SD dataset. The X-axis denotes the abbreviation for the bird names; the y-axis denotes the number of instances

### 3.2 Preprocessing

In the CLO-43SD dataset, the duration of the call of different bird species varies (Fig. 3), resulting in inconsistent spectrogram shapes. To make each audio clip the same length, we repeated the signal from the beginning and imposed a fixed duration of 2 s, which has been used in Thakur et al., (2019). Furthermore, notable differences exist in the audio amplitude of the same bird species, due to different recording conditions (see Fig. 4). The audio was normalized by subtracting the mean and scaling the amplitude to the interval [-1, 1], which is given in Eq. (1).

$$s = \frac{s - \text{mean}(s)}{\text{max}(\text{abs}(s))} \tag{1}$$

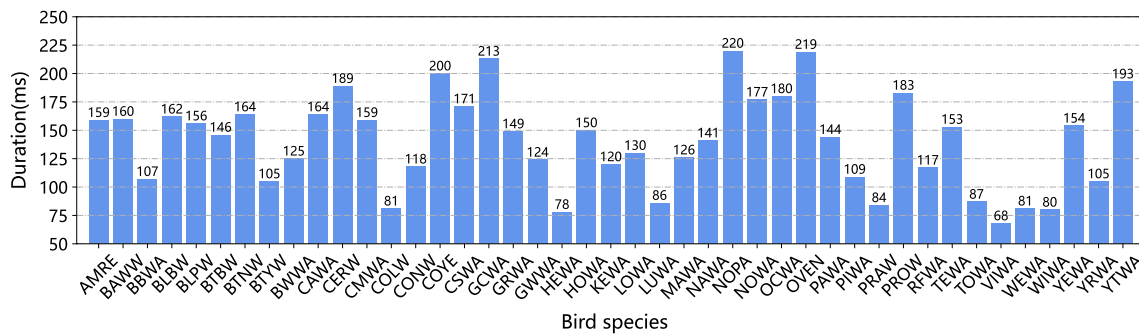
where  $S$  denotes the audio clip,  $\text{mean}(\cdot)$ ,  $\text{max}(\cdot)$  and  $\text{abs}(\cdot)$  denote the mean, maximum and absolute value, respectively.

### 3.3 Multi-level local binary pattern (multi-level LBP)

In general, bird sounds are characterized by short duration and high frequency. Using texture descriptors, the signal can be divided into small segments and the rapid changes in bird calls can be described by integrating the texture features in the small segments. In this work, multi-level LBP (Akbal, 2020; Tuncer et al., 2020a) were chosen to describe bird sounds, which include one dimensional binary pattern (1D-BP) and ternary pattern (1D-TP). Among them, 1D-BP uses the signum function as the kernel function to extract size information between neighborhood ( $x$ ) and center ( $c$ ), given in Eq. (2) (Kaya et al., 2014).

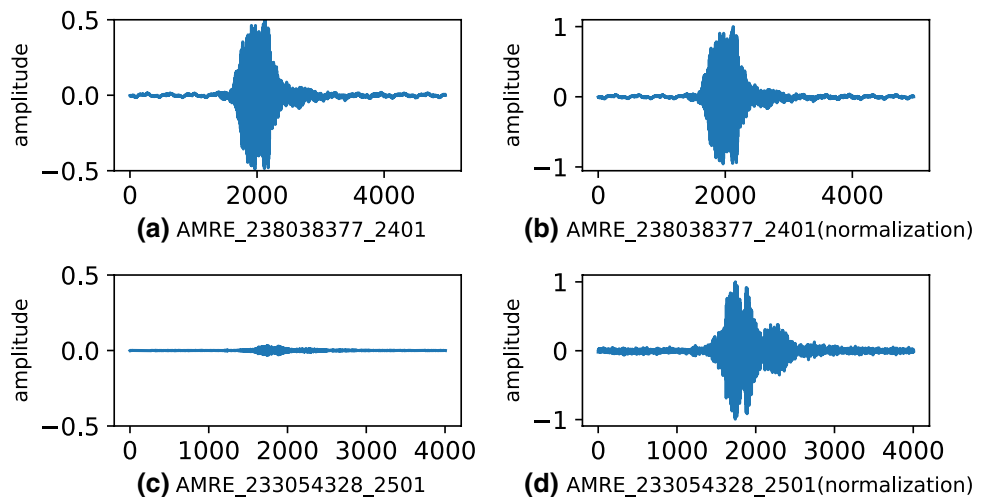
$$\text{signum}(x, c) = \begin{cases} 0 & x < c \\ 1 & x \geq c \end{cases} \tag{2}$$

1D-TP uses the ternary function as the kernel function to divide the difference between the neighborhood ( $x$ ) and



**Fig. 3** The duration of bird calls by species in the CLO-43SD dataset. The X-axis denotes the abbreviation for bird name; the y-axis denotes the call duration

**Fig. 4** Comparison of audio normalization. **a, c** The raw signal, **b, d** the normalized waveform



center ( $c$ ) into three intervals:  $(-\infty, -t)$ ,  $[-t, t]$  and  $[t, +\infty)$  (Kuncan et al., 2020), given in Eq. (3).

$$ternary(x, c, t) = \begin{cases} -1 & x - c < -t \\ 0 & -t \leq x - c \leq t \quad t \geq 0 \\ 1 & x - c > t \end{cases} \quad (3)$$

where  $t$  denotes the threshold and is determined by the signal standard deviation, given in Eq. (4) (Tuncer et al., 2020b).

$$t = \frac{std(signal)}{2} \quad (4)$$

TH and TL were used to calculate the upper and lower bits, to generate binary features.

$$TH = \begin{cases} 1 & ternary(x, c, t) = 1 \\ 0 & else \end{cases} \quad (5)$$

$$TL = \begin{cases} 1 & ternary(x, c, t) = -1 \\ 0 & else \end{cases} \quad (6)$$

Therefore, 1D-BP and 1D-TP can describe the signal in different ways and are often combined (1D-LBP) to represent signals (Akbal, 2020). The block diagram of 1D-LBP feature generation is shown in Fig. 5.

For rectangular window, a window length of 9 and hop size of 1 were used. To generate multi-level features, Akbal (2020) used the discrete wavelet transform (DWT) as a pooling method and extracted 1D-LBP features from input and low-frequency coefficients (Fig. 6).

However, multi-level LBP is good at describing waveform changes and (Akbal, 2020) only work within the time domain. To obtain richer time–frequency information of bird calls, we extracted multi-level LBP from the time (LBP-1D-T) and frequency domains (LBP-1D-F) respectively, which can describe signal in different ways. For LBP-1D-F, discrete cosine transform (DCT) and fast Fourier transform (FFT) are two classic time–frequency transformation methods. In this work, we combine them to analyze the frequency components of bird sounds (Tuncer et al., 2020a).

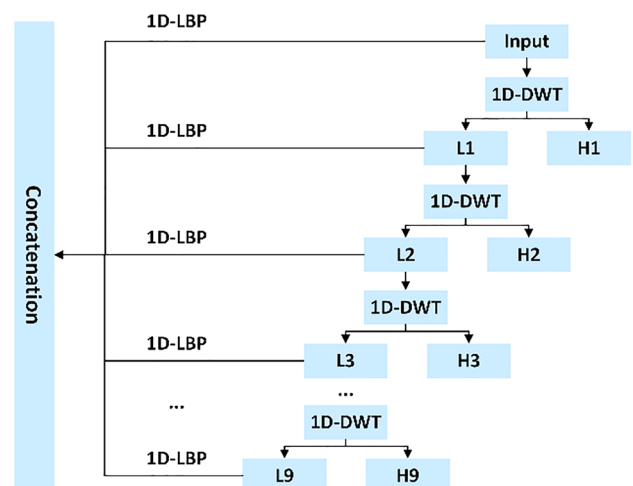


Fig. 6 Block diagram of multi-level LBP feature generation

Specifically, we first performed DCT and FFT on the input signal to obtain the frequency coefficients  $f_{DF}$ . The amplitude of  $f_{DF}$  was retained, phase information was discarded. Next, LBP-1D-F were generated from  $f_{DF}$ . Finally, LBP-1D-T and LBP-1D-F were concatenated to represent bird sounds.

### 3.4 Two-dimensional local binary pattern (LBP-2D)

Similar to LBP-1D, LBP-2D also extracts the difference between the neighborhood and center in the window. The window size of  $1 \times 9$  is generally used in LBP-1D, the 5th value is used as the center, the rest are used as neighborhoods. Any radius and number of neighborhoods can be used in the LBP-2D, by using circular neighborhoods. Here, we define  $R$  as the radius and  $P$  as the number of neighborhoods. Until now, several classic LBP-2D patterns have been proposed, such as uniform patterns,  $LBP_{P,R}^{u2}$ , rotation invariant,  $LBP_{P,R}^{riu2}$  and rotation invariant uniform patterns,  $LBP_{P,R}^{riu2}$  (Ojala et al., 2002). Among them, rotation invariant,  $LBP_{P,R}^{ri}$  rotates the binary features and takes the feature pattern with the smallest value as the LBP feature, which is suitable to describe the texture, where the pattern value depends on

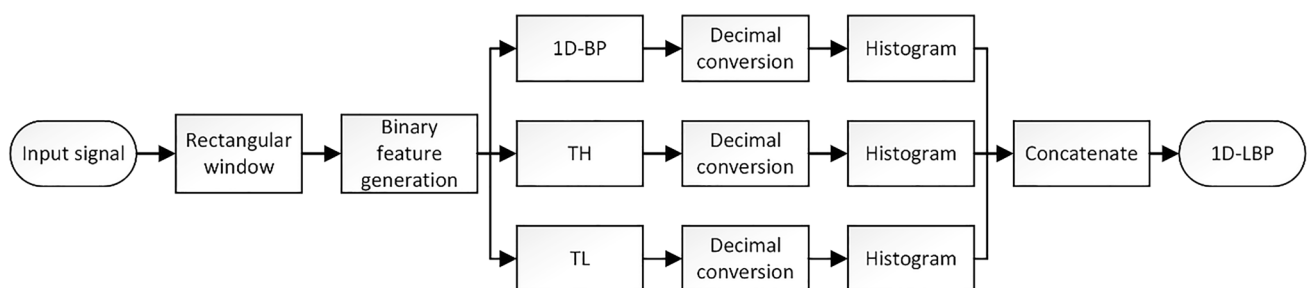


Fig. 5 The block diagram of 1D-LBP feature generation



the illumination, translation and rotational variance of the texture (Selin et al., 2006). However, it is unsuitable for bird call classification without rotation variations or translations. In this work, we first extracted the log-scaled Mel-spectrogram of each audio clip after pre-processing with 120 Melbin and 1.45 ms hop size. The frequency components were different among bird species and the peak frequency of all species of CLO-43SD is shown in Fig. 7.

Meanwhile, the spectrograms among bird species also differ. For example, the peak frequency of *LUWA* and *WEWA* are 31 Hz and 8025 Hz respectively, and their Mel-spectrogram are shown in Fig. 8. We can find that the energy of *LUWA* is concentrated in the low frequency, while the energy of *WEWA* is mainly concentrated in the high frequency, which is consistent with the peak frequency. Therefore, we selected uniform patterns,  $LBP_{P,R}^{u2}$  for our analysis, because Mel-spectrogram is characterized by the intensity along frequency and time instances, which provides a uniform pattern representation.

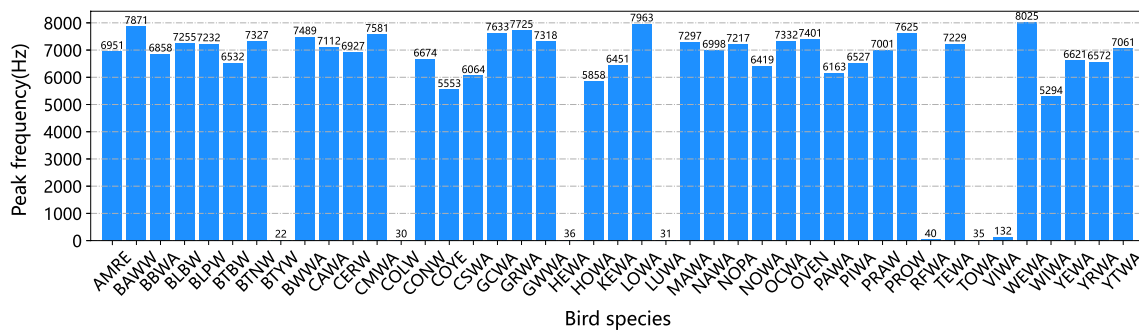
To generate more detailed features, the log-scaled Mel-spectrogram was divided into *K* linear zones of equal size by frequency, where  $LBP_{P,R}^{u2}$  was calculated for each zone and concatenated. Preliminary experiments indicated that the optimal (*P*, *R*) combination was (16, 5) and *K* was 10.

The block diagram of LBP-2D feature generation is shown in Fig. 9.

### 3.5 Feature selection

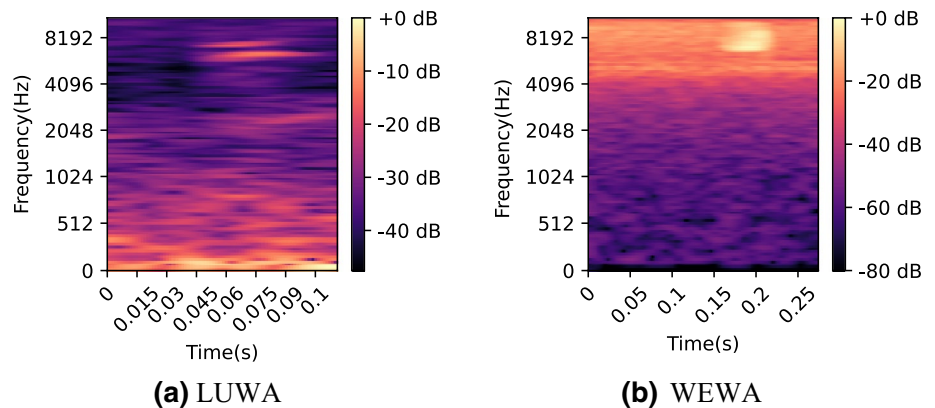
Following feature extraction, the dimensions of LBP-1D-T, LBP-1D-F and LBP-2D were 7680, 7680 and 2430, respectively. Large feature dimensions adversely affect the model, increasing the complexity of the classifier and training time. To select more discriminative features, feature selection was used. In this work, we propose an improved feature selection method named DSNCA. Compared with NCA, it reduces feature selection time, while improving the model performance. The following paragraphs illustrate the details.

For multi-level LBP features, the input signal was decomposed 9 times through discrete wavelet transform, and the 1D-LBP features were extracted from the input signal and low frequency coefficients of each level, resulting in similarity between the features. Therefore, multi-level LBP features are redundant. Neighborhood component analysis (Tuncer & Ertam, 2020) first randomly generates a weight value for each feature and calculates distances of features by Manhattan distance, given in Eq. (7).



**Fig. 7** The peak frequency of bird species in the CLO-43SD dataset. The X-axis denotes the abbreviation for bird name; y-axis denotes the peak frequency

**Fig. 8** Mel-spectrogram of *LUWA* and *WEWA* bird species



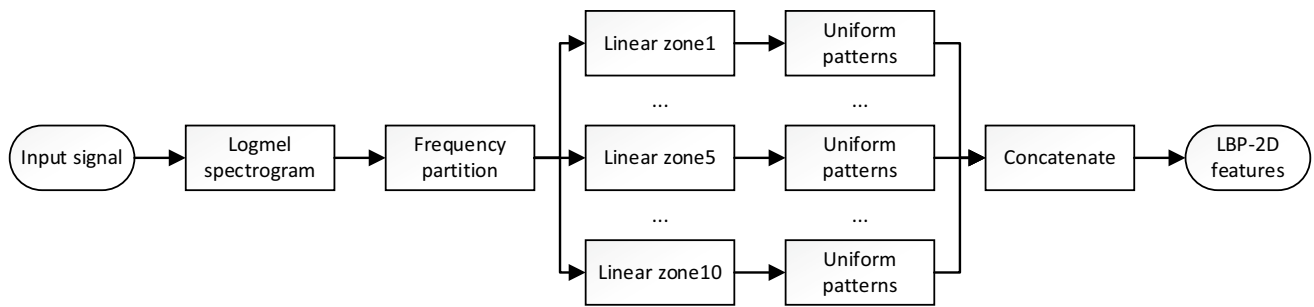


Fig. 9 The block diagram of LBP-2D feature generation

$$Dist(x_i, x_j) = |x_i - x_j| \tag{7}$$

Next, the weights are optimized by ADAM (Kingma & Ba, 2014) or stochastic gradient descent (Bottou, 2010) and the quality of the features is correlated to the weight. Finally, the most informative features are selected in order. However, as the dimensionality of the input features increases, the running time of NCA will increase considerably due to the deep learning optimizer. When multi-level LBP features are directly input to NCA, the process will be time-consuming and the selected features will still be redundant. For this reason, we propose an improved neighborhood component analysis (DSNCA) to remove redundant features before NCA by down sampling, which will be more suitable for highly redundant features. The block diagram of DSNCA is shown in Fig. 10.

Features were first divided into L sized non-overlapping groups, the first of each group was retained and concatenated in order. Next, the features were scaled to the interval [0, 1] by min–max normalization, given in Eq. (8).

$$feat^N = \frac{feat - feat_{min}}{feat_{max} - feat_{min}} \tag{8}$$

where  $feat^N$ ,  $feat_{min}$  and  $feat_{max}$  denote the normalized, minimum and maximum of  $feat$ , respectively. Finally, the most informative features were selected by NCA. Preliminary experiments indicated that the optimal value of L was 2.

### 3.6 Classifier description

Bird sound classification is a multi-class classification task. In this paper, two machine learning algorithms: K-nearest neighbor (KNN) and support vector machine (SVM) were selected and evaluated. For the KNN classifier, the distance between the test audio clip and training data was first calculated and the

predicted label was determined by the majority class of its k nearest neighbors (Han et al., 2011). SVM is a classifier based on statistical learning theory. For linearly separable problems, different categories are distinguished by finding the optimal partition surface. For linearly inseparable problems, the optimal partition surface is determined in a higher-dimensional space by non-linear mapping to distinguish different categories (Salamon et al., 2016).

## 4 Class-based late fusion

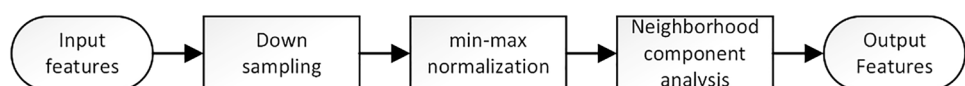
To further improve model performance, a class-based late fusion method (Xie et al., 2019) was used to take advantage of the complementarity between different LBP-based models. Assume the fusion of decisions from  $n$  models ( $M_1, M_2, \dots, M_n$ ) for a  $m$ -class problem ( $C_1, C_2, \dots, C_m$ ). When using model  $M_1$  to classify a test instance  $x$ , its predicted probability for each class can be obtained as  $M_{11}, M_{12}, \dots, M_{1m}$ . When using model  $M_2$  to classify instance  $x$ , its predicted probability for each class can be obtained as  $M_{21}, M_{22}, \dots, M_{2m}$ . In a similar fashion, the predicted probability of model  $M_n$  is obtained. For instance  $x$ , the final predicted probability  $P$  for each class can be calculated using Eq. (9).

$$P_k = \sum M_{kj}, \quad 1 \leq k \leq m, \quad 1 \leq j \leq n \tag{9}$$

where  $k$  denotes class,  $j$  denotes model and the final prediction is determined by the maximum  $P$  using Eq. (10).

$$Label_{final} = C_{\text{argmax}_{k=1}^m P_k} \tag{10}$$

Fig. 10 The block diagram of DSNCA



## 5 Experimental results

In this experiment, the dataset was first randomly divided into 85% and 15% to be used as training and test sets, respectively. The training set was further randomly divided into 60% and 40%, to be used as the training data and validation set, respectively. The divided datasets retained the same species proportion as the raw dataset. This process was repeated five times and the averaged result is reported. Since the CLO-43SD is an imbalanced dataset, we select balanced accuracy to evaluate the performance of our proposed model, given in Eq. (11):

$$\text{Balanced accuracy} = \frac{1}{n} \sum_{i=1}^n \frac{TP(i)}{S_i} \quad (11)$$

where  $n$  is the number of categories,  $TP$  is true positive and  $S_i$  is the sample size of class  $i$ .

### 5.1 Baseline

In order to benchmark our proposed model, 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) were combined with machine learning models as baseline (Logan, 2000). First, the bird sound was preprocessed to obtain the signal  $f^{2N}$  with length 2 s, mean 0 and amplitude of  $[-1, 1]$ . Subsequently, 13-dimensional MFCC and its delta and delta-delta were extracted from  $f^{2N}$ , with a total of 39 features where the number of filters was 40, the window size was 11.6 ms and the hop size was 1.45 ms. Finally, the features were used as input to KNN and SVM classifiers for bird sound classification. The experimental results show that the MFCCs cannot represent bird sound well (Table 1).

### 5.2 LBP-based features

In order to describe the rapid change in bird sounds, multi-level LBP features were first extracted from both time-domain (LBP-1D-T) and frequency-domain (LBP-1D-F) signals, respectively. The frequency components vary among bird species (Fig. 7). To obtain richer time–frequency information of bird calls, the uniform patterns (LBP-2D) were obtained from the log-scaled Mel-spectrogram. Next, the different LBP-based features were combined with KNN and SVM classifier for bird sound

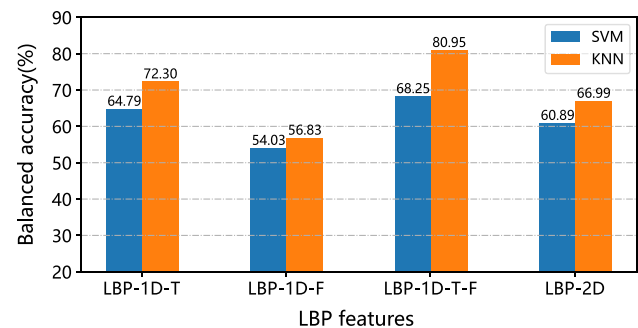
**Table 1** Balanced accuracy of 39-MFCCs combined with SVM and KNN respectively, which is 13-dimensional MFCCs and its delta and delta-delta

| Model       | Balanced accuracy (%) |
|-------------|-----------------------|
| MFCCs + SVM | 73.18                 |
| MFCCs + KNN | 74.11                 |

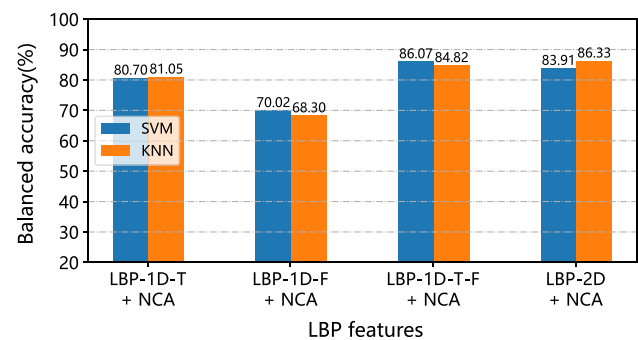
classification. When the LBP-based features are directly input into the classifier, KNN performs better than SVM and LBP-1D-T-F combined with KNN produced the highest balanced accuracy of 80.95% (Fig. 11). However, in general, these models obtained low accuracy.

### 5.3 LBP-based features with NCA

When the LBP-based features were directly used as input to classifiers, the model obtained low accuracy due to the redundant information in the LBP-based feature set. Therefore, NCA was selected for feature selection. The results show that the model performance improved by using NCA to select informative features before classification (Fig. 12). Among them, SVM produced similar accuracy to KNN and the highest classification accuracy of 86.33% was obtained using LBP-2D features.



**Fig. 11** Comparison of different LBP-based features combined with SVM and KNN classifiers for bird sound classification. LBP-1D-T and LBP-1D-F denote multi-level LBP features of time-domain and frequency-domain, respectively, LBP-1D-T-F denotes the concatenation of LBP-1D-T and LBP-1D-F and LBP-2D denotes the uniform patterns from the log-scaled Mel-spectrogram



**Fig. 12** Comparison of different LBP-based features with NCA. X-axis denotes the different LBP-based features combined with NCA; Y-axis denotes the balanced accuracy



### 5.4 Fusion of LBP-based models.

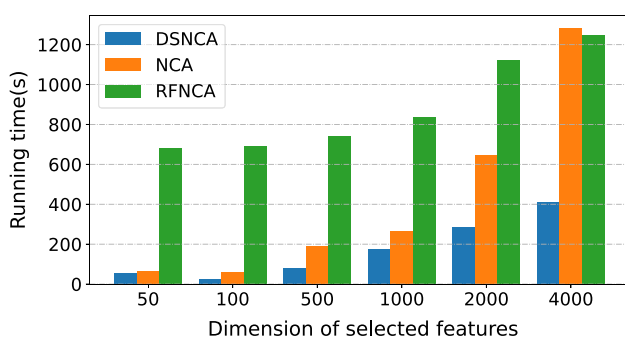
To fully exploit the complementarity of different LBP-based features and improve model performance, three fusion methods were implemented and compared, including early fusion, late fusion and hybrid fusion. The following paragraphs illustrate the details.

#### 5.4.1 Early fusion of LBP-based features

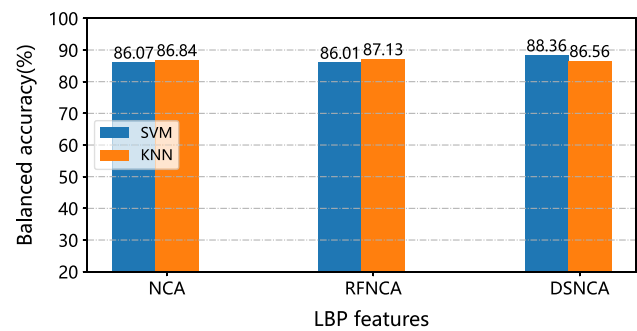
For the early fusion, the *LBP-1D-T*, *LBP-1D-F* and *LBP-2D* features were concatenated and a total of  $7680 + 7680 + 2430 = 17,790$  features (*LBP-T-F-2*) were obtained. However, directly using them as input to NCA for feature selection will be time-consuming and inefficient. To reduce feature selection time, while selecting more informative features, an improved feature selection method (DSNCA) was used. To test the effectiveness of this method, the time required for DSNCA, NCA and RFNCA (ReliefF combined with NCA) was tested (Demir et al., 2020). The results show that the running time of DSNCA was less than that of NCA and RFNCA (Fig. 13). As the dimension of selected features increases, the time complexity of NCA will increase significantly. Next, the selected features by the three feature selection methods were then combined with the classifier. The results show that DSNCA combined with SVM achieved the highest balanced accuracy of 88.36% (Fig. 14). Compared with NCA combined with SVM, the classification accuracy increased by 2.29%.

#### 5.4.2 Late fusion of LBP-based models

Ensemble learning has been proven to be an effective way to improve model performance in the case of imbalanced data, by combining results from several different models (Faris et al., 2020). To test the model performance of different



**Fig. 13** The running time of DSNCA (down-sampling combined with NCA), NCA and RFNCA (ReliefF combined with NCA), the input features are LBP-T-F-2 (feature concatenation of *LBP-1D-T*, *LBP-1D-F* and *LBP-2D*). X-axis denotes the dimension of selected features; Y-axis denotes the running time (s)



**Fig. 14** Comparison of model performance using NCA, DSNCA and RFNCA respectively, the input features are LBP-T-F-2. X-axis denotes the feature selection method; Y-axis denotes the classification accuracy

fusion methods, different class-based late fusion strategies of three LBP-based models were evaluated and the results are shown in Table 2. The highest balanced accuracy of 88.65% was obtained by the late fusion of three LBP-based models, which is lower than early fusion.

#### 5.4.3 Hybrid fusion of LBP-based models

Hybrid fusion, which includes early fusion and late fusion, was evaluated to further improve the model performance. First, the *LBP-1D-T* and *LBP-1D-F* features were concatenated and a total of 15,360 features (*LBP-1D-T-F*) were obtained. To select the most informative features, DSNCA was performed and the selected features were used as input in SVM for classification. Next, NCA was selected for *LBP-2D* and the obtained features were used as input in SVM for classification. Finally, a class-based late fusion was evaluated and the results show that the hybrid fusion achieved the highest balanced accuracy of 88.70%, which is superior to other fusion methods (Table 3).

To further analyze the classification results, the confusion matrix of the hybrid fusion method is plotted in Fig. 15. Our model achieved an accuracy of 100% for *HEWA* and *KEWA*. For *BLPW*, the model performed the worst and obtained only 68% accuracy and 15% of samples were misclassified as *MAWA*. In addition, 20% of *BWWA* samples were misclassified into *MAWA*. The visual patterns of *BLPW*, *MAWA* and *BWWA* are shown in Fig. 16. The visual patterns of those bird species are similar, which explains the difficulty in their accurate classification.

## 6 Discussion

In previous work, the highest balanced accuracy of 86.31% was obtained by the late fusion of Mel-VGG and Mel-Subnet (*Fusion I*). In this work, LBP-based features for bird

**Table 2** Classification performance using different late fusion strategies

| LBP-1D-T+NCA+SVM | LBP-1D-F+NCA+SVM | LBP-2D+NCA+SVM | Accuracy (%) |
|------------------|------------------|----------------|--------------|
| √                | √                |                | 81.37        |
| √                |                  | √              | 86.12        |
|                  | √                | √              | 83.95        |
| √                | √                | √              | <b>86.65</b> |

The best result is displayed in bold

sound classification were investigated (Fig. 17). In order to describe the rapid changes in bird sounds, multi-level LBP features were extracted from both time-domain (LBP-1D-T) and frequency-domain (LBP-1D-F) signals, respectively. For LBP-2D, uniform patterns,  $LBP_{PR}^{u2}$  were selected for the analysis because Mel-spectrogram is characterized by the intensity along frequency and time instances, which provides a uniform pattern representation. To reduce the feature selection time, while retaining the information quality of the features, an improved feature selection method DSNCA was proposed.

To take full advantage of the complementarity of different LBP-based features and improve the model performance, two fusion methods of early fusion and late fusion were evaluated, which obtained the highest balanced accuracy of 88.36%, 86.65% respectively. Additionally, early fusion and late fusion (hybrid fusion) were combined for bird sound classification, to further improve classification accuracy and achieved a highest balanced accuracy of 88.70%. In previous work, CNN-based models demanded high requirements on hardware due to their algorithm complexity, making it difficult to be used in lightweight devices, such as mobile phones. To address this, the proposed method is based on shallow learning, which will be more suitable for lightweight devices.

## 7 Conclusion and future work

In this study, LBP-based features for bird sound classification were investigated, including both one-dimensional (LBP-1D) and two-dimensional (LBP-2D) local binary patterns. For LBP-1D, multi-level LBP features were extracted from both time-domain (LBP-1D-T) and

frequency-domain (LBP-1D-F) signals, respectively, to describe the rapid changes in bird sounds. To obtain richer time–frequency information, the uniform patterns (LBP-2D) were obtained from the log-scaled Mel-spectrogram. When the LBP-based features are used as input in the classifier, the model obtains low accuracy and this process is time consuming due to the redundant information in the LBP-based feature set. Next, an improved feature selection method named DSNCA was proposed, which reduces the feature selection time while improving the model performance, when compared with NCA. To take full advantage of the complementarity of different LBP-based features, early fusion and late fusion were evaluated. However, neither achieved the highest accuracy individually. When the two fusion methods were combined (hybrid fusion), the model achieved the best balanced accuracy of 88.70%. Meanwhile, the LBP-based system adopted shallow classifier to identify bird sounds, compared with deep learning, which will be more suitable for lightweight devices.

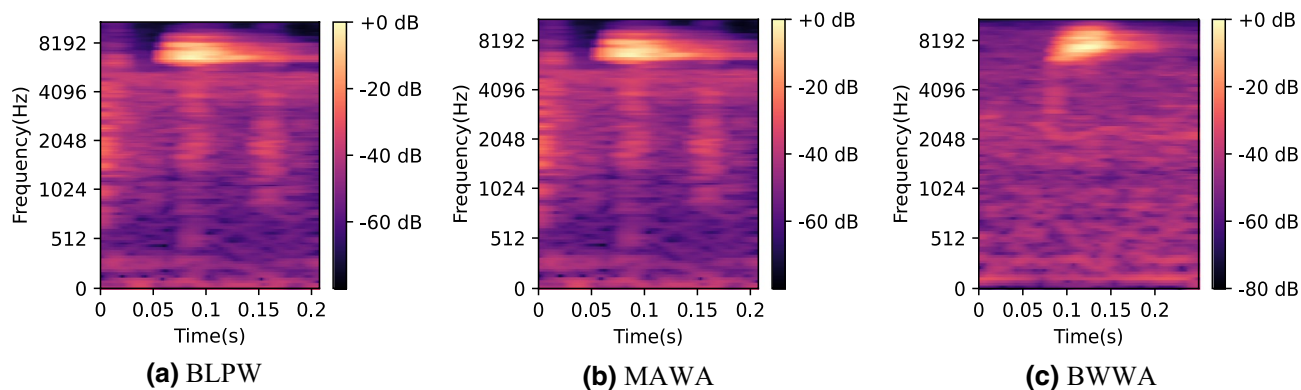
Future work will focus on improving the effectiveness of the bird sound classification system. First, a novel texture descriptor and light weight model that are more suitable for bird sound will be investigated. Next, in this work, only 43 bird species were studied. More bird species from different countries will be worth studying. Finally, given that CLO-43SD is an imbalanced dataset, a suitable imbalanced learning method for bird sound classification is equally important.

**Table 3** Balanced accuracy of hybrid fusion of LBP-based models, where LBP-1D-T-F denotes feature concatenation of the LBP-1D-T and LBP-1D-F

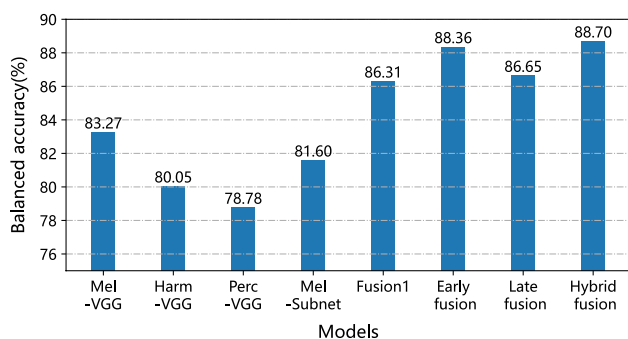
| Model                           | Accuracy (%) |
|---------------------------------|--------------|
| LBP-1D-T-F+DSNCA+SVM            | 86.25        |
| LBP-2D+NCA+SVM                  | 83.91        |
| Late fusion of above two models | <b>88.70</b> |

Bold value indicate the highest classification accuracy





**Fig. 16** Visual representations of *BLPW*, *MAWA* and *BWWA*



**Fig. 17** Classification results of the proposed system compared to previous work. X-axis denotes the different bird sound classification systems; Y-axis denotes the balanced accuracy. (Mel-VGG: Mel-spectrogram combined with *VGG style network*, Harm-VGG: harmonic-component based Mel-spectrogram combined with *VGG style network*, Perc-VGG: percussive-component based Mel-spectrogram combined with *VGG style network*, Mel-Subnet: Mel-spectrogram combined with *SubSpectralNet*)

**Acknowledgements** This work is supported by the 111 Project. This work is also supported by Fundamental Research Funds for the Central Universities (Grant No: JUSRP11924) and Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment & Technology (Grant No: FM-2019-06). This work is partially supported by National Natural Science Foundation of China (Grant No: 61902154). This work is also partially supported by Natural Science Foundation of Jiangsu Province (Grant No: BK2019043526). This work is also partially supported by Jiangsu province key research and development project—modern agriculture (Grant No: BE2018334).

**Data availability** Our proposed approach is implemented in Python, and the code will be publicly available on <https://github.com/jiang-rgb?tab=repositories> for future comparison and development by other researchers.

## References

- Akbal, E. (2020). An automated environmental sound classification methods based on statistical and textural feature. *Applied Acoustics*, 167, 107413.
- Bairlein, F. (2016). Migratory birds under threat. *Science*, 354(6312), 547–548.
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010* (pp. 177–186).
- Demir, S., Key, S., Tuncer, T., & Dogan, S. (2020). An exemplar pyramid feature extraction based humerus fracture classification method. *Medical Hypotheses*, 140, 109663.
- Emlen, J. T., & DeJong, M. J. (1992). Counting birds: The problem of variable hearing abilities (contando aves: El problema de la variabilidad en la capacidad auditiva). *Journal of Field Ornithology*, 63, 26–31.
- Faris, H., Abukhurma, R., Almanaseer, W., Saadeh, M., Mora, A. M., Castillo, P. A., & Aljarah, I. (2020). Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market. *Progress in Artificial Intelligence*, 9(1), 31–53.
- Farnsworth, A., Sheldon, D., Geevarghese, J., Irvine, J., Van Doren, B., Webb, K., et al. (2014). Reconstructing velocities of migrating birds from weather radar—A case study in computational sustainability. *AI Magazine*, 35(2), 31–48.
- Han, N. C., Muniandy, S. V., & Dayou, J. (2011). Acoustic classification of Australian anurans based on hybrid spectral-entropy approach. *Applied Acoustics*, 72(9), 639–645.
- Hutto, R. L., & Stutzman, R. J. (2009). Humans versus autonomous recording units: A comparison of point-count results. *Journal of Field Ornithology*, 80(4), 387–398.
- Kaya, Y., Uyar, M., Tekin, R., & Yıldırım, S. (2014). 1D-local binary pattern based feature extraction for classification of epileptic EEG signals. *Applied Mathematics and Computation*, 243, 209–219.
- Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, 348(6240), aad2478.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- Kuncan, M., Kaplan, K., Minaz, M. R., Kaya, Y., & Ertunc, H. M. (2020). A novel feature extraction method for bearing fault classification with one dimensional ternary patterns. *ISA Transactions*, 100, 346–357.

- Lanzone, M., Deleon, E., Grove, L., & Farnsworth, A. (2009). Revealing undocumented or poorly known flight calls of warblers (Parulidae) using a novel method of recording birds in captivity. *The Auk*, *126*(3), 511–519.
- Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. *Ismir*, *270*, 1–11.
- Loss, S. R., Will, T., & Marra, P. P. (2015). Direct mortality of birds from anthropogenic causes. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 99–120.
- Lostanlen, V., Palmer, K., Knight, E., Clark, C., Klinck, H., Farnsworth, A., et al. (2019). Long-distance detection of bioacoustic events with per-channel energy normalization. arXiv preprint. [arXiv:1911.00417](https://arxiv.org/abs/1911.00417).
- Marcarini, M., Williamson, G. A., & de Sisternes Garcia, L. (2008). Comparison of methods for automated recognition of avian nocturnal flight calls. In *2008 IEEE international conference on acoustics, speech and signal processing* (pp. 2029–2032).
- Nanni, L., Costa, Y. M., Aguiar, R. L., Mangolin, R. B., Brahnam, S., & Silla, C. N. (2020). Ensemble of convolutional neural networks to improve animal audio classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2020*, 1–14.
- Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.
- Rosenstock, S. S., Anderson, D. R., Giesen, K. M., Leukering, T., & Carter, M. F. (2002). Landbird counting techniques: Current practices and an alternative. *The Auk*, *119*(1), 46–53.
- Salamon, J., Bello, J. P., Farnsworth, A., & Kelling, S. (2017). Fusing shallow and deep learning for bioacoustic bird species classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 141–145).
- Salamon, J., Bello, J. P., Farnsworth, A., Robbins, M., Keen, S., Klinck, H., & Kelling, S. (2016). Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PLoS ONE*, *11*(11), e0166866.
- Schrama, T., Poot, M., Robb, M., & Slabbekoorn, H. (2007). Automated monitoring of avian flight calls during nocturnal migration. In *International Expert meeting on IT-based detection of bioacoustical patterns* (pp. 131–134).
- Selin, A., Turunen, J., & Tanntu, J. T. (2006). Wavelets in recognition of bird sounds. *EURASIP Journal on Advances in Signal Processing*, *2007*, 1–9.
- Stanley, C. Q., McKinnon, E. A., Fraser, K. C., Macpherson, M. P., Casbourn, G., Friesen, L., et al. (2015). Connectivity of wood thrush breeding, wintering, and migration sites based on range-wide tracking. *Conservation Biology*, *29*(1), 164–174.
- Thakur, A., Thapar, D., Rajan, P., & Nigam, A. (2019). Multiscale CNN based deep metric learning for bioacoustic classification: Overcoming training data scarcity using dynamic triplet loss. arXiv preprint. [arXiv:1903.10713](https://arxiv.org/abs/1903.10713).
- Tuncer, T., Dogan, S., Ertam, F., & Subasi, A. (2020a). A dynamic center and multi threshold point based stable feature extraction network for driver fatigue detection utilizing EEG signals. *Cognitive Neurodynamics*, *15*, 223–237.
- Tuncer, T., & Ertam, F. (2020). Neighborhood component analysis and reliefF based survival recognition methods for Hepatocellular carcinoma. *Physica A: Statistical Mechanics and Its Applications*, *540*, 123143.
- Tuncer, T., Subasi, A., Ertam, F., & Dogan, S. (2020b). A novel spiral pattern and 2D M4 pooling based environmental sound classification method. *Applied Acoustics*, *170*, 107508.
- Xie, J., Hu, K., Zhu, M., Yu, J., & Zhu, Q. (2019). Investigation of different CNN-based models for improved bird sound classification. *IEEE Access*, *7*, 175353–175361.
- Xie, J., & Zhu, M. (2019). Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics*, *52*, 74–81.
- Zottesso, R. H., Costa, Y. M., Bertolini, D., & Oliveira, L. E. (2018). Bird species identification using spectrogram and dissimilarity approach. *Ecological Informatics*, *48*, 187–197.

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