Archives of Acoustics Vol. 50, No. 2, pp. 223–233 (2025), https://doi.org/10.24425/aoa.2025.154812

Research Paper

Multi-label Bird Species Classification Using Transfer Learning Network

Xue HAN, Jianxin PENG*

School of Physics and Optoelectronics, South China University of Technology
Guangzhou, China

*Corresponding Author e-mail: phjxpeng@163.com

Received September 20, 2024; revised April 3, 2025; accepted April 28, 2025; published online June 9, 2025.

Bird sounds collected in the field usually include multiple birds of different species vocalizing at the same time, and the overlapping bird sounds pose challenges for species recognition. Extracting effective acoustic features is critical to multi-label bird species classification task. This work has extended an efficient transfer learning technique for labelling and classifying multiple bird species from audio recordings, further laying the foundation for conservation plans. A synthetic dataset was created by randomly mixing original single-species bird audio recordings from the Cornell Macaulay Library. The final dataset consists of 28 000 audio clips, each 5 s long, containing overlapping vocalizations of two or three bird species among 11 different species. Several pre-trained convolutional neural networks (CNNs), including InceptionV3, ResNet50, VGG16, and VGG19, were evaluated for extracting deep features from audio signals represented as mel spectrograms. The long short-term memory network (LSTM) was further employed to extract temporal features. A multi-label bird species classification was investigated. The absolute matching rate, accuracy, recall, precision, and $F1$-score of the InceptionV3+LSTM model for multi-label bird species classification are 98.25 %, 99.32 %, 99.41 %, 99.90 %, and 99.57 %, respectively, with the minimum Hamming loss of 0.0062. The results show that the proposed method has excellent performance and can be used for multi-label bird species classification.

Keywords: transfer learning; multi-label bird species classification; InceptionV3; LSTM.

1. Introduction

Field recordings of bird sounds typically contain vocalizations from multiple bird species occurring simultaneously, known as the 'dawn chorus', a phenomenon common in natural habitats. However, relatively few studies have addressed the challenge of multi-label bird species classification in these realistic acoustic environments. Early studies primarily relied on classical acoustic features and traditional machine learning approaches. For example, Briggs *et al.* (2012) manually segmented overlapping bird sounds recorded from the H.J. Andrews Experimental Forest (548 audio clips, each containing 1–5 species) and utilized multi-instance multi-label K-nearest neighbor (MIML-KNN), achieving an accuracy of 96.1 %. Leng and Dat Tran (2014) combined spectral features, MFCC, and linear predictive coding (LPC) extracted from NIPS4B dataset (687 audio clips, containing multiple bird species per clip) and trained ensemble classifiers, obtaining an AUC of 91.74 %. Liu (2016) introduced a transfer learning feature mapping method based on MFCC and Gaussian mixture models (GMM) for multi-label bird sound classification. The method was evaluated on NIPS4B and an artificial dataset (constructed by mixing xeno-canto bird audio), achieving the Hamming loss of 0.1024.

With the advancement of deep learning, recent approaches have increasingly utilized convolutional neural networks (CNNs) to automatically learn acoustic features. Sprengel *et al.* (2016) proposed a CNN approach trained on the BirdCLEF 2016 dataset, achieving a mean average precision (MAP) score of 0.686 for identifying the dominant bird species in audio recordings, surpassing previous state-of-the-art results. Bravo Sanchez *et al.* (2021) used the CNN framework SincNet on NIPS4Bplus bird recordings, achieving an accuracy of 73.56 %. Noumida and Rajan (2022) proposed a hierarchical attention-based bidirectional gated recurrent unit (BiGRU) model with

MFCC, trained on the xeno-canto dataset, achieving an $F1$-score of 0.85. Abdul Kareem and Rajan (2023) fused MFCC-RNN and mel spectrogram-CNN methods, obtaining an $F1$-score of 0.75 on the xeno-canto dataset.

Although CNNs effectively extract local features from spectrograms, they often neglect long-term temporal dependencies in acoustic data. Integrating CNNs with recurrent neural networks (RNNs), such as long short-term memory (LSTM), addresses this limitation by capturing sequential acoustic patterns (Sainath *et al.*, 2015; Nishikimi *et al.*, 2021; Liu *et al.*, 2021).

Transfer learning allows models to leverage the knowledge learned from large-scale datasets and tasks, significantly reducing training parameters and accelerating learning processes (Weiss *et al.*, 2016). Transfer learning is very helpful when there is insufficient data to fully train a model, such as recognizing uncommon bird species (Huang, Basanta, 2021). Gunawan *et al.* (2021) applied a transfer learning technique to avoid overfitting when classifying endangered species, such as the small footed owl in Indonesia. Deep CNN models, including VGG (Simonyan, Zisserman, 2014), ResNet (He *et al.*, 2016), and Inception networks (Szegedy *et al.*, 2016; 2017), have shown superior performance on image classification tasks, making them ideal candidates for transfer learning. Sevilla and Glotin (2017) successfully adapted the Inception-v4 network to bird sound classification, achieving the highest accuracy 71.4 % on the BirdCLEF 2017 dataset. Transfer learning has also proven effective in various multi-label classification tasks, including autonomous driving (Li *et al.*, 2021), natural language sentiment analysis (Tao, Fang, 2020), and transformer-based models across multiple domains (Gómez-Gómez *et al.*, 2023).

In this study, inspired by previous works, we utilize transfer learning models including VGG16, VGG19, InceptionV3, and ResNet50 to extract deep acoustic features from mel spectrograms of bird sounds. An LSTM network is integrated to capture temporal dependencies across frames. We specifically focus on multi-label classification tasks involving simultaneous vocalizations of two or three bird species, using synthetic datasets created from Cornell's Macaulay Library recordings (28 000 audio clips, each lasting 5 s). The feature extraction capability, classification accuracy, and generalization performance of these integrated models are comprehensively analyzed.

2. Method

2.1. Transfer learning models

The core idea of transfer learning is to leverage knowledge from the source domain to improve performance in a related target task. In this study, we utilize four pre-trained convolutional neural network architectures – VGG16, VGG19, InceptionV3, and ResNet50 – originally trained on the ImageNet dataset (Deng *et al.*, 2009). Each architecture has distinct characteristics that influence its performance on multi-label bird species classification tasks.

2.1.1. Classification model based on pre-trained network VGG16/VGG19

VGG is a classic image classification network based on the ImageNet database. Its characteristic is to use a convolutional layer with a smaller kernel (3×3) instead of a convolutional layer with a larger kernel. On the one hand, it can reduce parameters, and on the other hand, it is equivalent to performing more nonlinear mapping, increasing the network's expressive power (Simonyan, Zisserman, 2014). The VGG16 pre-trained network framework is shown in Fig. 1. The VGG19 model adds three additional convolutional layers on top of VGG16: one $3 \times 3 \times 256$ and two $3 \times 3 \times 512$. VGG19 has a deeper network than VGG16, and increasing the depth can effectively improve performance. The part before the fully connected layer is commonly referred to as the feature extraction layer.

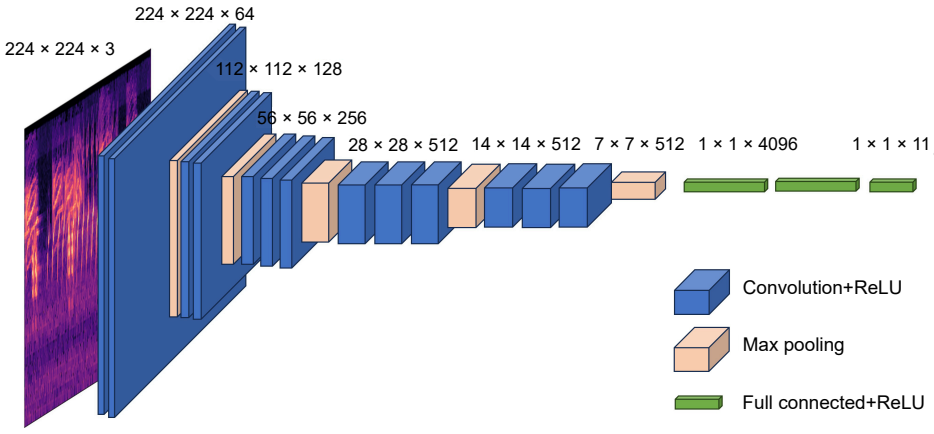

Fig. 1. VGG16 network structure diagram (based on (Simonyan, Zisserman, 2014)).

2.1.2. Classification model based on pre-trained network InceptionV3

The method of increasing the number of convolutional layers to enhance the learning ability of the network is not always feasible, because after the network reaches a certain depth, increasing the number of network layers will cause the problem of random gradient disappearance and explosion, and also lead to a decrease in accuracy. Moreover, complex networks can also bring high computational costs. The Inception module decomposes large convolutions into multiple small convolutions, where multiple small convolution kernels simultaneously convolve the image and aggregate information at different scales, as shown in Fig. 2 (Szegedy *et al.* 2016). This can significantly reduce network parameters without losing features. The key to the InceptionV3 network is to use Inception modules and two asymmetric decomposition structures to construct different types of Inception module groups.

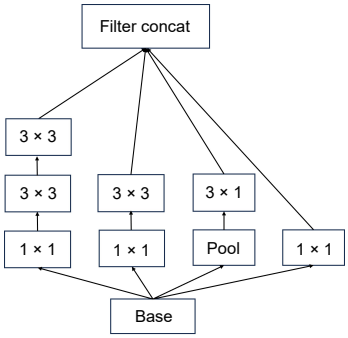

Fig. 2. Each 5×5 convolution in the Inception module is replaced by two 3 × 3 convolutions (based on (Szegedy *et al.*, 2016)).

2.1.3. Classification model based on pre-trained network ResNet50

In response to the problem of gradient disappearance, He *et al.* (2016) proposed a residual structure that not only solves the gradient problem, but also improves its feature expression ability with the increase of network layers, thereby improving classification performance. Figure 3 shows a residual structure in ResNet50, which includes cross layer connections that allow input to be directly passed across layers and then added to the convolutional result. This helps the model converge towards the equal mapping

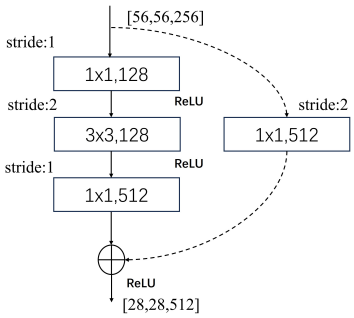

Fig. 3. Residual structure diagram (He *et al.*, 2016).

direction, ensuring that the final accuracy is not affected by the depth of the model. Table 1 shows the ResNet50 model structure and network parameters.

Table 1. Model training parameter statistics.

Models	LSTM input size	LSTM trainable parameters	FC parameters	Total trainable parameters
VGG16/VGG19	3 584	3 933 184	2 827	3 936 011
InceptionV3	16 384	17 040 384	2 827	17 043 211
ResNet50	14 336	14 943 232	2 827	14 946 059

We chose these models due to their distinct advantages: VGG models provide strong representational power, InceptionV3 excels at multi-scale feature extraction with fewer parameters, and ResNet50 effectively manages training of very deep networks. A comparative analysis of these models is summarized in Table 2.

2.2. Pre-trained convolutional neural network fused with LSTM

The feature sequence extracted by the pre-trained CNN cannot be directly fed into the LSTM network. Taking the VGG16 model as an example, the final extracted feature map has dimensions of $7 \times 7 \times 512$ ($H \times W \times C$). The original mel spectrogram of size 224×224 is compressed spatially to 7×7, and the number of channels deepens from 3 (original RGB channels) to 512. Through convolution and pooling operations, the spatial structural information in the mel spectrogram is transformed into deep feature representations, where each channel corresponds to a particular response pattern, such as edges, textures, and colors.

Table 2. CNN networks comparison.

CNN networks	Depth (layers)	Parameters, complexity	Feature extraction strategy
VGG16/19	16/19	High parameters, high computational complexity	Small 3 × 3 convolutions, captures fine-grained local features
InceptionV3	48	Moderate parameters, efficient computation due to parallel modules	Multi-scale feature extraction via parallel convolutions (Inception modules)
ResNet50	50	High parameters but efficient training	Deep residual structure enabling hierarchical feature abstraction

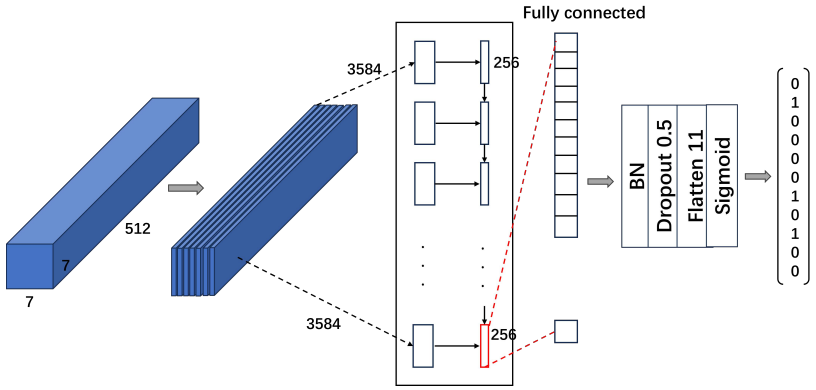

Fig. 4. Feature sequence obtained from pre-trained VGG16 input with LSTM.

To make these features suitable for temporal modeling by LSTM, the feature maps are reshaped based on the spatial dimension (e.g., width dimension) into a sequential format. Specifically, we expand the feature maps along the width dimension (W) into a feature sequence of size 7×3584, where 7 corresponds to the number of time steps, and 3584 (i.e., 7×512) represents the features at each time step. Figure 4 illustrates the detailed procedure of feeding these feature sequences input into the LSTM network. Similarly, for other pre-trained CNN architectures, the extracted feature map dimensions differ slightly: InceptionV3 produces a feature sequence of $8 \times 8 \times 2048$, and ResNet50 yields a feature sequence of $7 \times 7 \times 2048$. These feature maps are processed in the same way as described above to prepare sequential data suitable for input into the LSTM network.

The number of neurons in the output layer of the LSTM network is 11 for classifying bird species. Sigmoid is used as the activation function for the output layer, ensuring that each neuron outputs a probability of 1. When the probability value is greater than 0.5, the predicted value output by the neuron is 1, indicating that the bird species corresponding to the neuron exists in the audio sample. The fusion model also adds a batch normalization (BN) layer to avoid gradient disappearance. Dropout layers are introduced to reduce overfitting by randomly dropping units during training, thereby improving the generalization performance of the network. The parameters in the feature extraction modules of VGG16, VGG19, InceptionV3, and ResNet50 are frozen, and the remaining parts of the network are trained using our dataset in this work. To help assess model complexity, we provide the number of trainable parameters (excluding frozen CNN layers) for each configuration, as shown in Table 1.

3. Experimental settings

3.1. Dataset construction

Bird sound recordings of 11 bird species used for this work are collected from the Macaulay Library at Cornell University[1]. Table 3 shows the selected original audio of each bird. These recordings are in MP3 format, with a sampling frequency of 44 100 Hz and a bit rate of 128 000 bps. Each recording is annotated as a single species of bird vocalization. To create multi-label bird species samples, we randomly selected audio segments from the original single-species recordings and combined them digitally using audio mixing software 'Adobe Audition'. Before mixing, the audio segments were normalized to the same volume level to prevent any single recording from dominating due to volume differences. Each resulting 5-second segment contains clearly annotated overlapping vocalizations from either 2 or 3 different bird species. The detailed information is shown in Table 4.

Figures 5 and 6 represent mel spectrograms of syllable overlapping for 2 and 3 bird species, respectively. The mel spectrograms were constructed using the Librosa library with a sample rate of 44 100 Hz, the FFT window length (n_fft) of 1024 samples, a hop length of 512 samples (50 % overlap), and 128 mel filter banks. A visual inspection reveals that as the number of overlapping bird species increases, the spectral complexity and signal interference also become more pronounced. For instance, in Figs. 5a and 5c, the individual vocal patterns of each species are relatively separable, often occurring in distinct frequency bands or time intervals. However, in Figs. 6b and 6c, the spectrograms show significantly denser and more continuous activity across both frequency and time, making it more difficult to visually or algorithmically disentangle individual species. This suggests that classification tasks involving three or more overlapping bird species are inherently more challenging due to increased spectral overlap, which can obscure characteristic frequency patterns and temporal features.

The labels of each audio segment are manually reviewed and verified to ensure the presence of corresponding bird sounds. The ratio of training set to testing set is divided into 3:1.

[1]Specifically retrieved via the search interface at: https://search.macaulaylibrary.org/catalog

Table 3. Audio file information of 11 bird species.

Bird species	Number of downloaded audio files	Audio file name
Downy Woodpecker	12	ML107289, ML433684551, ML320270011, ML216529601, ML288560951, ML89889581, ML259178751, ML94232, ML282354581, ML249048571, ML218533941, ML539363
Northern Flicker	9	ML60535251, ML47981841, ML6891, ML84808, ML224667, ML176938031, ML6802, ML63072, ML299493831
Black-capped Chickadee	10	ML381756441, ML202239, ML227931651, ML228999, ML359860121, ML244530591, ML442275881, ML315584611, ML9334271, ML217850561
White-breasted Nuthatch	9	ML51757711, ML196990751, ML304498191, ML105313481, ML313785451, ML120214, ML169318341, ML88195851, ML245567141
Northern Cardinal	7	ML101113031, ML94284, ML94286, ML94285, ML325248201, ML434987071, ML24184651
House Finch	9	ML369617771, ML44967, ML110958961, ML331732541, ML161496541, ML22938, ML56843, ML22941, ML12932
Pine Siskin	5	ML156434831, ML22902731, ML176160, ML89549511, ML219631251
Western Backyard Birds	6	ML481585181, ML203884811, ML425203981, ML279795071, ML2425203911, ML168880461
Steller's Jay	8	ML35291431, ML202130641, ML44859, ML410551461, ML42204, ML119017701, ML192457, ML90747421
Evening Grosbeak	5	ML148939381, ML160442941, ML129191951, ML77259, ML227584
Blue Jay	13	ML166281501, ML177463211, ML345934681, ML107392, ML264268971, ML260458751, ML421603721, ML539887, ML219634, ML13448, ML359246651, ML223790721, ML20432

Table 4. Detailed information of multi-label dataset.

Category	Number of training samples	Number of test samples	In total
2	11 550	3 850	15 400
3	9 450	3 150	12 600
Total	21 000	7 000	28 000

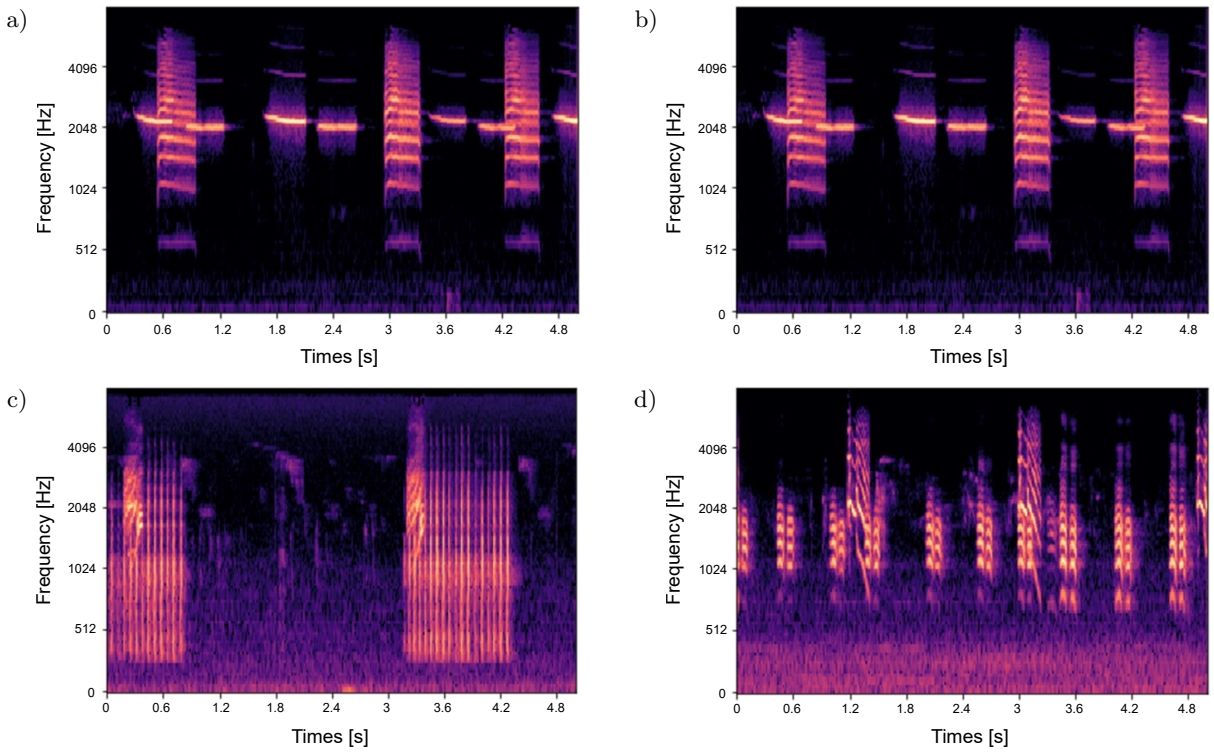

Fig. 5. Mel spectrogram of mixed audio of syllables overlapping between two species of birds: a) Black capped Chickadee–Blue Jay; b) Pine Siskin–Steller's Jay; c) Downy Woodpecker–House Finch; d) Northern Flicker–White-breasted Nuthatch.

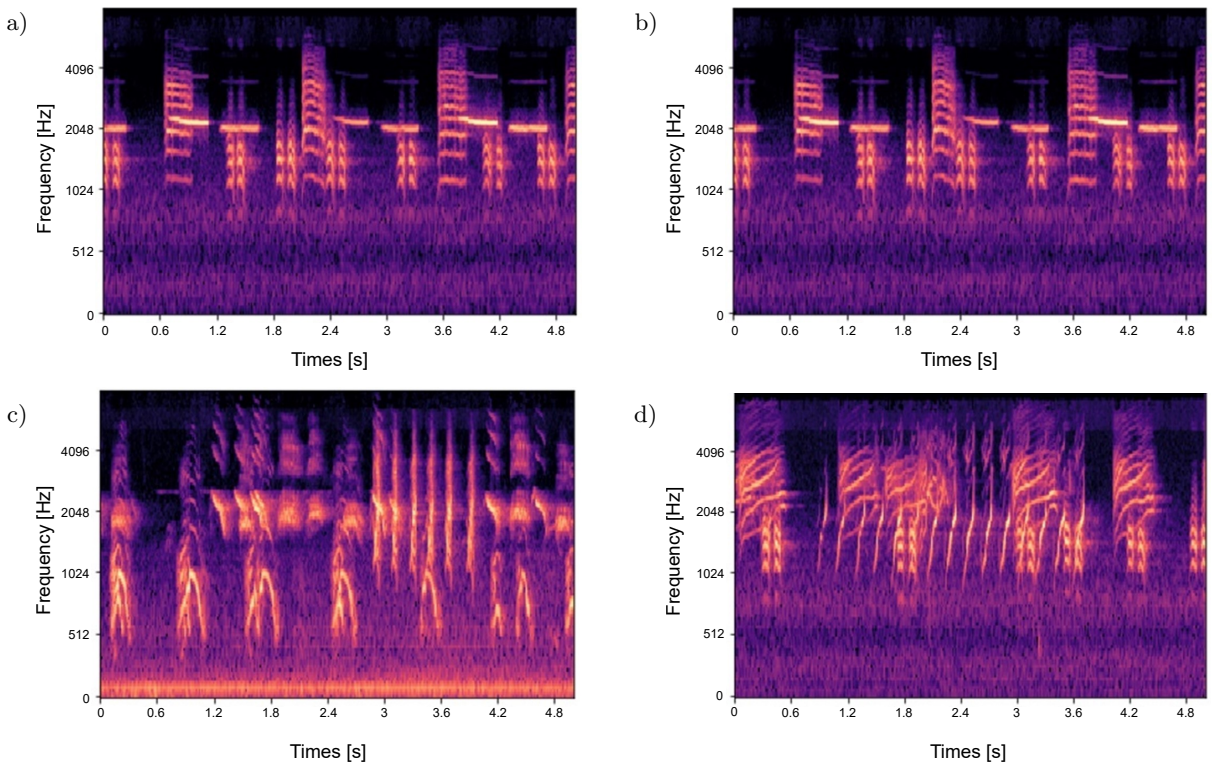

Fig. 6. Mel spectrogram of mixed audio of syllables overlapping between three species of birds: a) Black capped Chickadee–White-breasted Nuthatch–Blue Jay; b) Downy Woodpecker–Northern Flicker–White-breasted Nuthatch; c) Western Backyard Birds–Steller's Jay–Evening Grosbeak; d) White-breasted Nuthatch–Northern Cardinal–Pine Siskin.

3.2. Objective evaluation

Unlike single label classification tasks, a sample in a multi-label classification task can have multiple labels. Firstly, without considering partially correct evaluation metrics, the sample can only be predicted correctly if the predicted label is exactly the same as the true label (Paniri *et al.*, 2020). This evaluation metric is called the exact match ratio (Zhang *et al.*, 2016) and the calculation formula is as follows:

$$\text{Exact match ratio} = \frac{1}{n}\sum_{i=1}^{n} I(\widehat{Y}_i = Y_i), \tag{1}$$

where I is the indicator function. When Y_i is completely equivalent to $\widehat{Y}_i$, I is 1, otherwise it is 0; $\widehat{Y}_i$ is the predicted label set for sample i, Y_i is the ground truth label set; n represents the total number of samples. It can be seen that this evaluation metric is very strict for the classification model. In addition, only some labels that are correctly predicted can also be used to evaluate the performance of classification models. The commonly used performance metrics include accuracy, recall, precision, and $F1$-score (Godbole, Sarawagi, 2004). Accuracy is defined as the proportion of correctly predicted labels to the union of predicted and true labels for each sample, averaged across all samples:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cup \widehat{Y}_i|}{|Y_i \cap \widehat{Y}_i|}, \tag{2}$$

where $|Y_i \cup \widehat{Y}_i|$ is number of correctly predicted labels for sample i, $|Y_i \cap \widehat{Y}_i|$ is total number of unique labels in the prediction and true labels for sample i. Recall measures the proportion of correctly predicted labels out of all true labels for each sample, averaged over all samples:

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap \widehat{Y}_i|}{|Y_i|}, \tag{3}$$

where $|Y_i|$ is total number of true labels for sample i. Precision is defined as the proportion of correctly predicted labels out of all predicted labels for each sample, averaged over all samples:

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap \widehat{Y}_i|}{|\widehat{Y}_i|}, \tag{4}$$

where $|\widehat{Y}_i|$ is total number of labels predicted for sample i. The $F1$-score for each sample is the harmonic mean of precision and recall, averaged across all samples:

$$F1 - \text{score} = \frac{1}{n}\sum_{i=1}^{n} 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \tag{5}$$

In addition, performance metrics also include the Hamming loss (Sorower, 2010). The Hamming loss evaluates the fraction of labels that are incorrectly predicted across all samples:

$$\text{Hamming loss} = \frac{1}{kn}\sum_{i=1}^{n}\sum_{j=1}^{k} I(y_{ij} \neq \widehat{y}_{ij}), \tag{6}$$

where k is the total number of labels, y_{ij} is the true value of the j-th label for sample i, and $\widehat{y}_{ij}$ is the predicted value of the j-th label for a sample i. The smaller the value of the Hamming loss, the better the performance of the classification model.

3.3. Implementation details

The hardware environment for experiments is a server with Inter I9-7920X CPU and NVIDIA GTX RTX1080Ti GPU, and the operating system is Ubuntu 16.04. All experimental models are built based on the PyTorch deep learning framework, with the PyTorch version number 1.9.1. During the training process, MultiLabelSoftMarginLoss (Cheng *et al.*, 2021) is used as the loss function, and the stochastic gradient descent (SDG) is used to update the network parameters. Momentum is set to 0.9, the learning rate is set to e-4, epoch is set to 300, and batch size is set to 32.

4. Result

4.1. Classification results of different transfer learning models

The results of multi-label bird species classification under different transfer learning models are shown in Table 5. According to Table 5, the InceptionV3 model has the best classification performance, with exact match ratio, accuracy, recall, precision, and $F1$-score of 93.04 %, 97.30 %, 97.50 %, 99.75 %, and 98.30 %, respectively, and the Hamming loss of 0.026. The InceptionV3 model uses decomposition convolution, which decomposes large convolution factors into small convolutions and asymmetric convolutions, effectively reducing parameters and avoiding overfitting. The Inception modules use multiple branches to extract high-order features with different levels of abstraction, enriching the network's expressive power (Szegedy *et al.*, 2016). The absolute matching rate of the VGG19 model is 4.23 % lower than that of the VGG16 model, which proves that blindly adding convolutional layers will not improve the classification performance and will lead to overfitting of the model. The exact match ratio, accuracy, recall, precision, and $F1$-score of the ResNet50 model are 85.62 %, 94.40 %, 95.22 %, 99.02 %, and 96.46 %, respectively, with the Hamming loss of 0.051. The classification performance of the ResNet50 model is better than VGG19, but inferior to VGG16, indicating that the residual structure has to some extent alleviated the overfitting phenomenon of the model. The CNN structure affects the results of multi-label bird sounds classification.

Table 5. Multi-label bird species classification results of four transfer learning models.

	VGG16	VGG19	InceptionV3	ResNet50
Exact match ratio [%]	87.87	83.64	**93.04**	85.62
Accuracy [%]	95.20	93.63	**97.30**	94.40
Recall [%]	95.69	94.41	**97.50**	95.22
Precision [%]	99.41	99.00	**99.75**	99.02
$F1$-socre [%]	96.97	95.97	**98.30**	96.46
Hamming loss	0.045	0.059	**0.026**	0.051

Figures 7 and 8, respectively, depict the variation curves of exact match ratio and the Hamming loss for four pre-trained models. As shown in Fig. 7, the exact match ratio of ResNet50 did not significantly improve after the 50th epoch, while the exact match ratio of InceptionV3 continued to increase, approaching saturation approximately after the 100th epoch. From Fig. 8, it can be seen that after the 50th epoch, the Hamming loss of InceptionV3 is significantly lower than VGG16, VGG19, and ResNet50. Overall, the VGG16 and VGG19 models are not suitable for multi-label bird species classification.

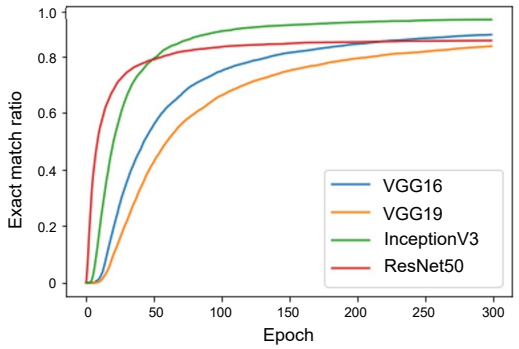

Fig. 7. Exact match ratio curves of four transfer learning models.

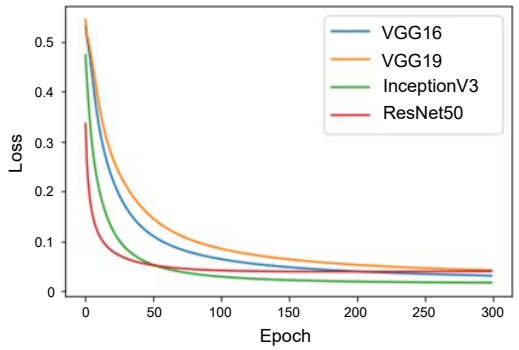

Fig. 8. Hamming loss curves of four transfer learning models.

4.2. Classification results of four transfer learning models fused with LSTM

Table 6 shows the classification results of four transfer learning models fused with LSTM. Comparing Tables 5 and 6, it can be seen that after fusing LSTM, the classification performance of VGG16, VGG19, InceptionV3, and ResNet50 has all improved significantly. The exact match ratio of ResNet50 in-

creased the most, by 12.89 %, VGG19 by 6.42 %, InceptionV3 by 5.21 %, and VGG16 by 3.73 %. The LSTM network can learn the time series characteristics in feature sequences, and the time series of vocalizations of different bird species vary. InceptionV3+LSTM has the best classification performance among all models, with exact match ratio, accuracy, recall, precision, and $F1$-score of 98.25 %, 99.32 %, 99.42 %, 99.90 %, and 99.57 %, respectively. Moreover, the Hamming loss also reaches a minimum of 0.0062. The Hamming loss of InceptionV3+LSTM is reduced by 0.0198 compared to InceptionV3, indicating that the prediction error and missing error of multi labels are minimized. This is because the CNN-LSTM model can learn complicated patterns from data more rapidly and correctly than the CNN model alone. However, the precision of InceptionV3+LSTM is slightly lower by 0.05 % than ResNet50+LSTM. The exact match ratio, accuracy, recall, precision, and $F1$-score of the ResNet50+LSTM model are 6.55 %, 2.46 %, 1.99 %, 0.52 %, and 1.52 % higher than those of VGG16+LSTM, respectively. This indicates that the feature sequence obtained by ResNet50 contains more time series characteristics. The classification performance of VGG19 fusion LSTM network has been improved, but it is still lower than the other three transfer learning models.

Table 6. Multi-label bird species classification results of four transfer learning models fused with LSTM.

	VGG16 + LSTM	VGG19 + LSTM	InceptionV3 + LSTM	ResNet50 + LSTM
Exact match ratio [%]	91.60	90.06	**98.25**	98.15
Accuracy [%]	96.78	96.21	**99.32**	99.24
Recall [%]	97.28	96.79	**99.41**	99.27
Precision [%]	99.43	99.32	99.90	**99.95**
$F1$-score [%]	97.99	97.62	**99.57**	99.51
Hamming loss	0.029	0.034	**0.0062**	0.0073

Figures 9 and 10, respectively, depict the exact match ratio and the Hamming loss variation curves of four transfer learning models fused with LSTM. As shown in Fig. 9, the exact match ratio of the VGG16+LSTM and VGG19+LSTM classification models is significantly lower than InceptionV3+LSTM and ResNet50+LSTM. Before the 70th epoch, the exact match ratio of ResNet50+LSTM was slightly higher than InceptionV3+LSTM. After the 70th epoch, there was no significant difference between the two fusion models. As shown in Fig. 10, before the 200th epoch, InceptionV3+LSTM had the higher Hamming loss than ResNet50+LSTM. After the 200th epoch, there was no significant difference between the two fusion models. Considering the performance of the model and computing resources, we chose InceptionV3+LSTM as the multi-label bird species classification model.

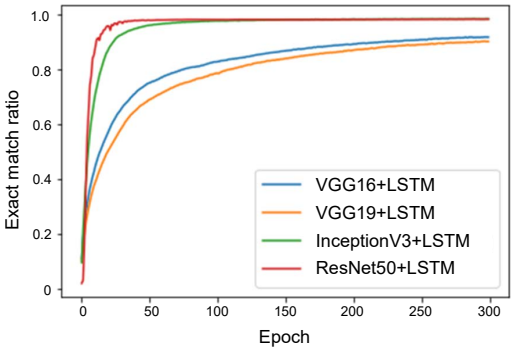

Fig. 9. Exact match ratio curves of four transfer learning models fused with LSTM.

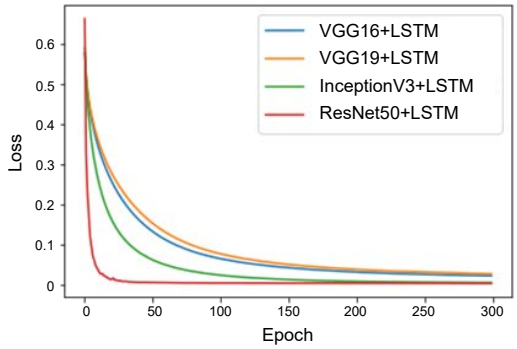

Fig. 10. Hamming loss curves of four transfer learning models fused with LSTM.

4.3. Classification confusion matrices of InceptionV3+LSTM

To further analyze the performance of the InceptionV3+LSTM model in multi-label bird species classification, we present the confusion matrices for each label in Table 7. These confusion matrices provide a detailed view of the model's prediction accuracy for each individual label. For label 0, the high number of true positives (1332) and true negatives (5661) indicates that the model performs well in identifying this label. However, label 2 has a relatively higher number of false positives (8) and false negatives (11) compared to other labels, which may suggest that this label is more challenging for the model to classify accurately. Label 4 has a moderate number of false positives (4) but a higher number of false negatives (23), indicating that the model may have difficulty in correctly identifying this label.

Overall, the confusion matrices demonstrate that the InceptionV3+LSTM model has a high accuracy in classifying most labels, with only a few exceptions where misclassifications occur. This suggests that the model is capable of effectively learning the features and temporal patterns in the bird sound data, leading to accurate multi-label classification results.

Table 7. Confusion matrix for different labels.

Confusion matrix for label 0		
Label 0: real	Predict	
	0	1
0	5661	5
1	2	1332
Confusion matrix for label 1		
Label 1: real	Predict	
	0	1
0	5082	3
1	2	1913
Confusion matrix for label 2		
Label 2: real	Predict	
	0	1
0	5154	8
1	11	1827
Confusion matrix for label 3		
Label 3: real	Predict	
	0	1
0	5284	0
1	4	1712
Confusion matrix for label 4		
Label 4: real	Predict	
	0	1
0	5318	4
1	23	1655
Confusion matrix for label 5		
Label 5: real	Predict	
	0	1
0	5379	0
1	16	1605
Confusion matrix for label 6		
Label 6: real	Predict	
	0	1
0	5440	0
1	6	1554
Confusion matrix for label 7		
Label 7: real	Predict	
	0	1
0	5543	0
1	0	1457
Confusion matrix for label 8		
Label 8: real	Predict	
	0	1
0	5620	0
1	13	1367
Confusion matrix for label 9		
Label 9: real	Predict	
	0	1
0	5647	0
1	20	1333

5. Discussion

Table 8 shows a comparison of the relevant studies with the present study in terms of method and performance. In multi-label bird species classification tasks, syllable overlap can limit manual feature extraction (Liu, 2016; Briggs *et al.*, 2012; Noumida, Rajan, 2022; Leng, Dan Tran, 2014; Abdul Kareem, Rajan, 2023), because syllable segmentation is a crucial step. The accuracy of any classifier that relies on segmentation is sensitive to the quality of the segmentation (Fagerlund, 2004). A recent study has shown that deep learning is an effective method for classifying birds based on their sounds, such as processing large amounts of audio data, which allows it to detect subtle differences between bird sounds (Michaud *et al.*, 2023). Researchers usually increase the number of convolutional layers to extract more detailed features from the audio raw waveform (Bravo Sanchez *et al.*, 2021) and mel spectrograms (Abdul Kareem, Rajan, 2023). These methods result in more training parameters and the need for sufficient data to train model parameters. To reduce the number of trainable parameters and address the issue of data availability for the deep convolutional network to be effectively trained, a transfer learning approach is adopted in this study. For multi-label bird species classification, we employed the ImageNet-trained InceptionV3 convolution network. However, the CNN model ignores the temporal dependence of bird sounds. The pre-trained InceptionV3 is further fused with LSTM to extract time series characteristics from the feature sequence. Table 7 demonstrates that the proposed multi-label bird species classification method based on pre-trained InceptionV3 fused with LSTM has excellent performance.

The experimental results demonstrate the superiority of the InceptionV3+LSTM model in multi-label bird species classification, particularly in challenging cases involving overlapping syllables from two or three bird species. This confirms that combining convolutional feature extraction with temporal modeling via LSTM yields significant benefits over CNNs alone.

However, despite the strong performance, the proposed method relies on a large number of parameters and pre-trained models trained on image datasets (ImageNet), which may not optimally capture the characteristics of audio spectrograms. Furthermore, although we simulate overlapping bird calls, the synthetic nature of the dataset may not fully capture the complexities of real-world soundscapes such as environmental noise or unpredictable call patterns.

Compared to previous studies listed in Table 8, our method achieves state-of-the-art performance, yet direct comparisons remain difficult due to the diversity of datasets, label types, and evaluation metrics. Future work could benefit from the establishment of standardized multi-label bird audio benchmarks and the integration of more audio-specialized architectures, such as attention-based transformers or audio foundation models.

Table 8. Comparative analysis with other methods.

Reference work	Method	Dataset	Performance
Liu (2016)	Based on MFCC feature transfer	NIPS4B and xeno-canto	Hamming loss 0.1024
Briggs *et al.* (2012)	Spectral features with MIML-KNN	H.J. Andrews Experimental Forest	Accuracy 96.1 %
Noumida, Rajan (2022)	MFCC with BiGRU	xeno-canto	$F1$-score 0.85
Leng, Dat Tran (2014)	Spectral features, MFCC and LPC with ensemble model	NIPS4B	AUC 91.74 %
Bravo Sanchez *et al.* (2021)	SincNet	NIPS4Bplus	Accuracy 73.56 %, AUC 74.85 %, Precision 74.81 %, Recall 73.56 %
Abdul Kareem, Rajan (2023)	Fused the MFCC-RNN and mel spectrogram-CNN	xeno-canto	$F1$-score 0.75
Proposed method in this work	Pre-trained InceptionV3 with LSTM	Cornell Macaulay Library	Exact match ratio 98.25 %, Accuracy 99.32 %, Recall 99.42 %, Precision 99.90 %, $F1$-score 99.57 %, Hamming loss 0.0062

6. Conclusions

In recent years, research on multi-label bird sound classification has been limited, particularly for realistic scenarios where two or three bird species vocalize simultaneously within the same audio segment. Moreover, most existing works rely on researcher-constructed datasets due to the lack of publicly available multi-label bird datasets, which makes performance comparison and method validation challenging. To address these issues, this study proposed a multi-label bird species classification model based on a transfer learning architecture fused with LSTM. Specifically, our method focuses on the realistic challenge of identifying two or three bird species vocalizing simultaneously within the same 5-second audio clip. Traditional syllable segmentation methods often struggle in such overlapping scenarios. By applying pre-trained convolutional neural networks to extract deep acoustic features from mel spectrograms, and further integrating LSTM to capture temporal dependencies, the proposed model effectively addresses this challenge. Experimental results demonstrate that the InceptionV3+LSTM fusion model achieves outstanding performance in multi-label classification, with an exact match ratio of 98.25 %, accuracy of 99.32 %, recall of 99.42 %, precision of 99.90 %, $F1$-score of 99.57 %, and the minimum Hamming loss of 0.0062.

Conflict of 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.

References

1. Abdul Kareem N., Rajan R. (2023), Multi-label bird species classification using sequential aggregation strategy from audio recordings, *Computing and Informatics*, **42**(5): 1255–1280, https://doi.org/10.31577/cai_2023_5_1255.
2. Bravo Sanchez F.J., Hossain M.R., English N.B., Moore S.T. (2021), Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture, *Scientific Reports*, **11**: 15733, https://doi.org/10.1038/s41598-021-95076-6.
3. Briggs F. *et al.* (2012), Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach, *The Journal of the Acoustical Society of America*, **131**(6): 4640–4650, https://doi.org/10.1121/1.4707424.
4. Cheng Y., Ma M., Li X., Zhou Y. (2021), Multi-label classification of fundus images based on graph convolutional network, *BMC Medical Informatics and Decision Making*, **21**: 82, https://doi.org/10.1186/s12911-021-01424-x.
5. Deng J., Dong W., Socher R., Li L.J., Li K., Li F.F. (2009), ImageNet: A large-scale hierarchical image database, [in:] *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
6. Fagerlund S. (2004), *Automatic recognition of bird species by their sounds*, MSc. Thesis, Helsinki University of Technology.
7. Godbole S., Sarawagi S. (2004), Discriminative Methods for Multi-labeled Classification, [in:] *Advances in Knowledge Discovery and Data Mining. PAKDD 2004. Lecture Notes in Computer Science*, Dai H., Srikant R., Zhang C. [Eds.], **3056**: 22–30, https://doi.org/10.1007/978-3-540-24775-3_5.
8. Gómez-Gómez J., Vidaña-Vila E., Sevillano X. (2023), Western Mediterranean Wetland Birds dataset:

A new annotated dataset for acoustic bird species classification, *Ecological Informatics*, **75**: 102014, https://doi.org/10.1016/j.ecoinf.2023.102014.

9. Gunawan K.W., Hidayat A.A., Cenggoro T.W., Pardamean B. (2021), A transfer learning strategy for owl sound classification by using image classification model with audio spectrogram, *International Journal on Electrical Engineering and Informatics*, **13**(3): 546–553, https://doi.org/10.15676/ijeei.2021.13.3.3.
10. He K., Zhang X., Ren S., Sun J. (2016), Deep residual learning for image recognition, [in:] *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
11. Huang Y.-P., Basanta H. (2021), Recognition of endemic bird species using deep learning models, *IEEE Access*, **9**: 102975–102984, https://doi.org/10.1109/ACCESS.2021.3098532.
12. Leng Y.R., Dat Tran H. (2014), Multi-label bird classification using an ensemble classifier with simple features, *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1–5, https://doi.org/10.1109/APSIPA.2014.7041649.
13. Li G., Ji Z.F., Chang Y.L., Li S., Qu X.D., Cao D.P. (2021), ML-ANet: A transfer learning approach using adaptation network for multi-label image classification in autonomous driving, *Chinese Journal of Mechanical Engineering*, **34**: 78, https://doi.org/10.1186/s10033-021-00598-9.
14. Liu A. *et al*. (2021), Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores, arXiv, http://arxiv.org/abs/2010.13418.
15. Liu H.T. (2016), *A study on multi-label transfer learning algorithm and application in the bird sounds recognition*, Msc. Thesis, Nanjing Forestry University.
16. Michaud F., Sueur J., Le Cesne M., Haupert S. (2023), Unsupervised classification to improve the quality of a bird song recording dataset, *Ecological Informatics*, **74**: 101952, https://doi.org/10.1016/j.ecoinf.2022.101952.
17. Nishikimi R., Nakamura E., Goto M., Yoshii K. (2021), Audio-to-score singing transcription based on a CRNN-HSMM hybrid model, *APSIPA Transactions on Signal and Information Processing*, **10**(1): e7, https://doi.org/10.1017/ATSIP.2021.4.
18. Noumida A., Rajan R. (2022), Multi-label bird species classification from audio recordings using attention framework, *Applied Acoustics*, **197**: 108901, https://doi.org/10.1016/j.apacoust.2022.108901.
19. Paniri M., Dowlatshahi M.B., Nezamabadi-pour H. (2020), MLACO: A multi-label feature selection algorithm based on ant colony optimization, *Knowledge-Based Systems*, **192**: 105285, https://doi.org/10.1016/j.knosys.2019.105285.
20. Sainath T.N., Vinyals O., Senior A., Sak H. (2015), Convolutional, long short-term memory, fully connected deep neural networks, [in:] *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584, https://doi.org/10.1109/ICASSP.2015.7178838.
21. Sevilla A., Glotin H. (2017), Audio bird classification with Inception-v4 extended with time and time-frequency attention mechanisms, *Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum*, Cappellato L., Ferro N., Goeuriot L., Mandl T. [Eds.], **1866**, https://ceur-ws.org/Vol-1866/paper_177.pdf.
22. Simonyan K., Zisserman A. (2014), Very deep convolutional networks for large-scale image recognition, arXiv, http://arxiv.org/abs/1409.1556.
23. Sorower M.S. (2010), A literature survey on algorithms for multi-label learning.
24. Sprengel E., Jaggi M., Kilcher Y., Hofmann T. (2016), Audio based bird species identification using deep learning techniques, *Working Notes of CLEF 2016 – Conference and Labs of the Evaluation forum*, Balog K., Cappellato L., Ferro N., Macdonald C. [Eds.], **1609**, https://ceur-ws.org/Vol-1609/16090547.pdf.
25. Szegedy C., Ioffe S., Vanhoucke V., Alemi A. (2017), Inception-v4, Inception-ResNet and the impact of residual connections on learning, [in:] *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**(1), https://doi.org/10.1609/aaai.v31i1.11231.
26. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. (2016), Rethinking the Inception Architecture for Computer Vision, [in:] *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, https://doi.org/10.1109/CVPR.2016.308.
27. Tao J., Fang X. (2020), Toward multi-label sentiment analysis: a transfer learning based approach, *Journal of Big Data*, **7**: 1, https://doi.org/10.1186/s40537-019-0278-0.
28. Weiss K., Khoshgoftaar T.M., Wang D. (2016), A survey of transfer learning, *Journal of Big Data*, **3**: 9, https://doi.org/10.1186/s40537-016-0043-6.
29. Zhang L., Towsey M., Xie J., Zhang J., Roe P. (2016), Using multi-label classification for acoustic pattern detection and assisting bird species surveys, *Applied Acoustics*, **110**: 91–98, https://doi.org/10.1016/j.apacoust.2016.03.027.