Applied Acoustics 197 (2022) 108901

Contents lists available at ScienceDirect

Applied Acoustics

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

Multi-label bird species classification from audio recordings using attention framework

Noumida A. *, Rajeev Rajan

College of Engineering Trivandrum, APJ Abdul kalam Technological University, Thiruvananthapuram, India

ARTICLE INFO

Article history:
Received 17 January 2022
Received in revised form 15 June 2022
Accepted 22 June 2022
Available online 6 July 2022

Keywords:
Multi-label
Attention mechanism
Bidirectional gated recurrent units
Sequential
Augmentation

ABSTRACT

For the conservation of avian biodiversity, bird detection is vital since it allows ornithologists to quantify which species exist in a particular area. Analyzing their acoustic signals enables the efficient identification of multiple bird species from overlapping recordings. This paper addresses classifying bird vocalizations in real-time audio recording using acoustic analysis. Schemes based on recurrent neural networks (RNN) are presented in the proposed work. Gated-recurrent units (GRU) are a particular type of RNN that has shown remarkable performance in acoustic classification. We propose a hierarchical Attention-based bidirectional gated recurrent unit (BiGRU) model for classifying acoustic signals of birds by using Mel-frequency cepstral coefficients (MFCC). The attention mechanism has proved its superior efficacy in many acoustic, speech and music processing applications. The attention mechanism is employed to give a different focus to the information outputted from the hidden layers of BiGRU. We adopted a short-time sliding-aggregation approach to decide on the test data, in which probability outcomes are species-wise summed and normalized. Species with the highest probability scores are assumed to be the dominant species in the recording. Our Attention-BiGRU classifier achieves pretty high performance in the Xeno-Canto dataset, with an F1-score of 0.84, with competing performance to the state-of-the-art multi-label classifiers.

1. Introduction

Bio-acoustics has a wide spectrum of applications such as automated species identification, monitoring of habitats and environmental education. Acoustic data can be collected by deploying the recorders in the forest for several months for periodic assessment. Recognition of different species from recorded acoustic signals is an active research field that has received much attention in the last few decenniums. Paramount research has been done in acoustic classification to develop intelligent and automatic systems that can correctly identify and classify these species. A particularly challenging task cognate to bio-acoustics classification is detecting multiple overlapping audio events present in an acoustic scene. With the expeditious growth of deep learning, automatic species analysis and recognition have surmounted challenges such as detecting rare and inconspicuous species, monitoring wildlife populations, and studying animal demeanour to a great extent.

In psychology, attention is the cognitive process of focusing on one or a few domains while ignoring others. The same concept is implemented in deep learning architectures to mimic human brain actions in a simplified manner by selectively concentrating on a few pertinent things while ignoring others in the feature space [1]. The network model selectively focuses on relevant segments of the input sequence and learns the association between them. The attention mechanism has already been implemented in diverse scientific and technological domains and was first developed in the field of image recognition, primarily to increase the performance of RNN-based encoders and decoders [2]. In recent years, research-based attention mechanisms have gained popularity, focusing attention on distinct parts of input to make precise predictions, and they have proven to be an expert mechanism in the deep learning community [1,3,4]. The proposed study uses a deep learning technique that employs an Attention-BiGRU model using sequential aggregation to detect multiple bird species from overlapping real-field recordings.

The parts and structure of the avian sound-producing mechanism are depicted in Fig. 1. The lungs, bronchi, syrinx, trachea, larynx, mouth, and beak are all part of the sound-producing mechanism in birds [5]. The syrinx is the vocal organ of birds. As air travels from the lungs through the bronchi to the syrinx, it is modulated by the vocal tract. Complete cartilage rings make up the trachea. The bronchial elements are made of complete C-shaped cartilage rings with open ends facing each other. During

* Corresponding author.
E-mail address: noumidaa@gmail.com (A. Noumida).

https://doi.org/10.1016/j.apacoust.2022.108901
0003-682X/

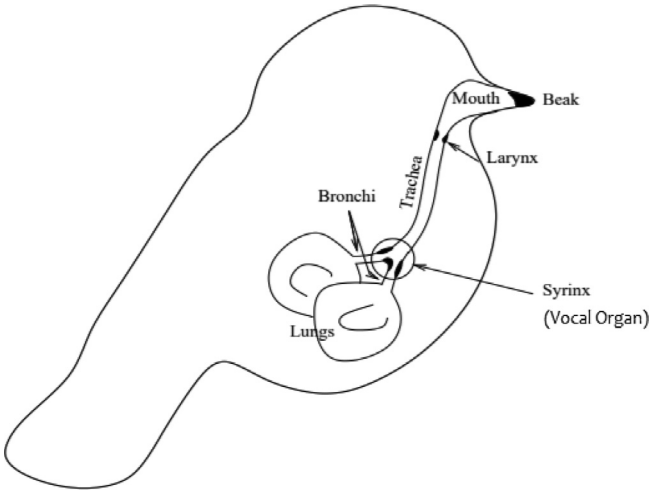

Fig. 1. Parts and organization of the avian sound-producing mechanism [5]. Syrinx acts as the vocal organ of birds.

a bird's song, airflow causes the syringeal medial tympaniform membrane (MTM) in each bronchus to vibrate non-linearly in the opposite direction of the cartilage wall. The mouth of birds, like that of humans, serves as a cavity resonator, but it is less flexible. Bird vocalizations include calls and songs. Bird songs are more complex vocalizations than bird calls, which are considered simple vocalizations.[1] The male is usually the one who composes songs on his own. Phrases, syllables, and elements make up the hierarchical levels of a bird's song. When a bird rearranges the phrases in its song, it can produce various singing styles. Bird calls are short and are made by stringing together a series of sounds. Fig. 2 illustrates the vocalization patterns of the W. Wood Pewee and Red. Lapwing using Mel-spectrograms.

2. Related work

Olden techniques for classifying bird species have necessitated a significant amount of human labour. Because of the size of the recording and storage devices, advanced fieldwork was not possible. The advancement of nanotechnology expands the capabilities of electronic devices while reducing the size of equipment, opening up new avenues for bird species classification. Deep learning techniques have achieved excellent performances in many classification tasks. Acoustic-based bird monitoring is an efficacious strategy since most birds communicate primarily through vocalizations [6]. Stefan Kahl et al. use various convolutional neural networks (CNNs) to create features retrieved from visual representations of field recordings to tackle large-scale bird sound classification [7]. However, the accurate prediction of all the species from overlapping recordings remains a challenge. Deep CNNs have achieved breakthrough performance in bird species identification using spectrograms of bird vocalizations. Two multi-channel fusion methods [8] were built with three feature identification models to enhance classification accuracy. The attention mechanism has been one of the most valuable breakthroughs in deep learning research in the last decade. Antoine Sevilla et al. presents the adaptation of the attention-based deep CNN Inception-v4 to solve bioacoustic classification problems, especially bird activity detection [9]. This attention-based model reports a mean average precision of 0.714 for classification. Attention mechanism can withal be integrated into CNNs to improve representation power by focusing on relevant features and suppressing those that are not [10,11]. Self-attention, also known as intra-attention, can be used to link different positions in a single sequence to create a more effective sequence representation [12–14]. One of the natural language processing (NLP) tasks, named entity recognition (NER), is proposed in [15]. Since traditional character representation is limited, and the neural network method fails to capture important sequence information, For NER, a self-attention-based BiGRU and capsule network (CapsNet) are utilized.

The use of transfer learning algorithms in CNN could be beneficial in bird call identification due to the difficulty of obtaining annotated training calls. In [16], a data-efficient bird call recognition strategy was created using the CNN transfer learning approach. RNN is a state-of-the-art sequential data algorithm that can solve variable-length sequence input, improve the feature representation, and convert to text vectors, resulting in a matrix with feature vector and sequence dimensions [17]. RNN networks, on the other hand, have issues with large-scale parallel computing. Long short-term memory (LSTM) is an RNN version that uses a series of gates to control the learning flow at each time step to capture long-range dependencies in long sentences. The BiLSTM model, in comparison to the LSTM model, can analyse a substantial amount of contextual information [18,19]. The amended structure of BiLSTM and BiGRU [20] preserves the original effect while standardizing the network. In [12], a self-attention mechanism is proposed for learning the dependency relationship among arbitrary characters and overcoming the problem of BiLSTM capturing information between long-distance characters. The information shared between characters can avail to understand the sentence structure and improve entity recognition performance. Multi-label bird classification is difficult because of the time–frequency overlapping in the audio recordings.

Forrest Briggs et al. proposes a bag generator algorithm [21] to transform input audio into a bag-of-instances suitable for use with multi-instance multi-label (MIML) classifiers, to solve the problem of classifying the set of species present in audio using the MIML framework for machine learning. A multi-label classification method for detecting simultaneous acoustic patterns in long-duration audio recordings is proposed in [22]. In a co-occurrence scenario, acoustic indices are appropriate indicators of broad ecological processes. In field recordings, multiple overlapping vocalizing birds of various species are common. Based on the features captured by Attention-BiGRU in the proposed task, we efficiently tackle the challenge of detecting the set of all species from overlapping vocalizations in a given audio recording. Fig. 3 shows the block diagram of the proposed attention-based multi-label bird species classification.

A summary of the system is given in Section 3. The performance evaluation is described in Section 4, followed by the analysis of results in Section 5. Finally, the paper is concluded in Section 6.

3. System description

The proposed model is shown in Fig. 3. We performed attention mechanisms on visual as well as acoustic features. While the top branch shows the methodology for the Attention-BiGRU, the bottom branch shows the steps in the Attention-VGG approach. The subsystems are explained in the following sections.

3.1. Feature extraction

Mel-frequency cepstral coefficients (MFCCs) and Mel-spectrograms have been used for the proposed experiments. MFCCs are widely employed in numerous perceptually motivated audio classification tasks [23], as predictors of perceived similarity

[1] https://en.wikipedia.org/wiki/Bird-vocalization.

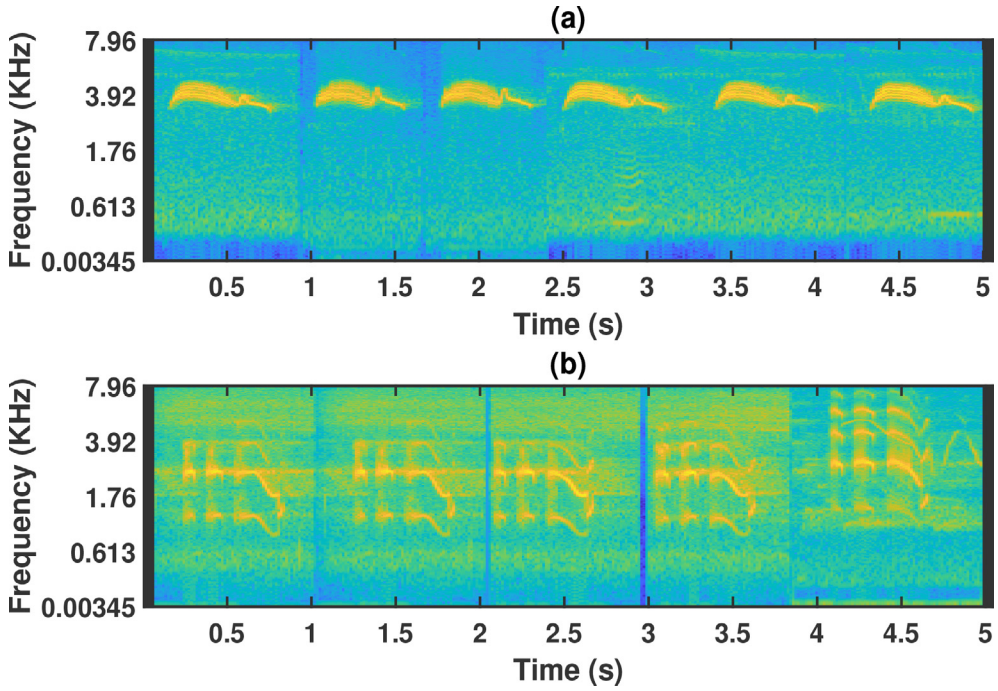

Fig. 2. Mel-spectrogram of bird calls of (a) W. Wood Pewee, (b) Red. Lapwing.

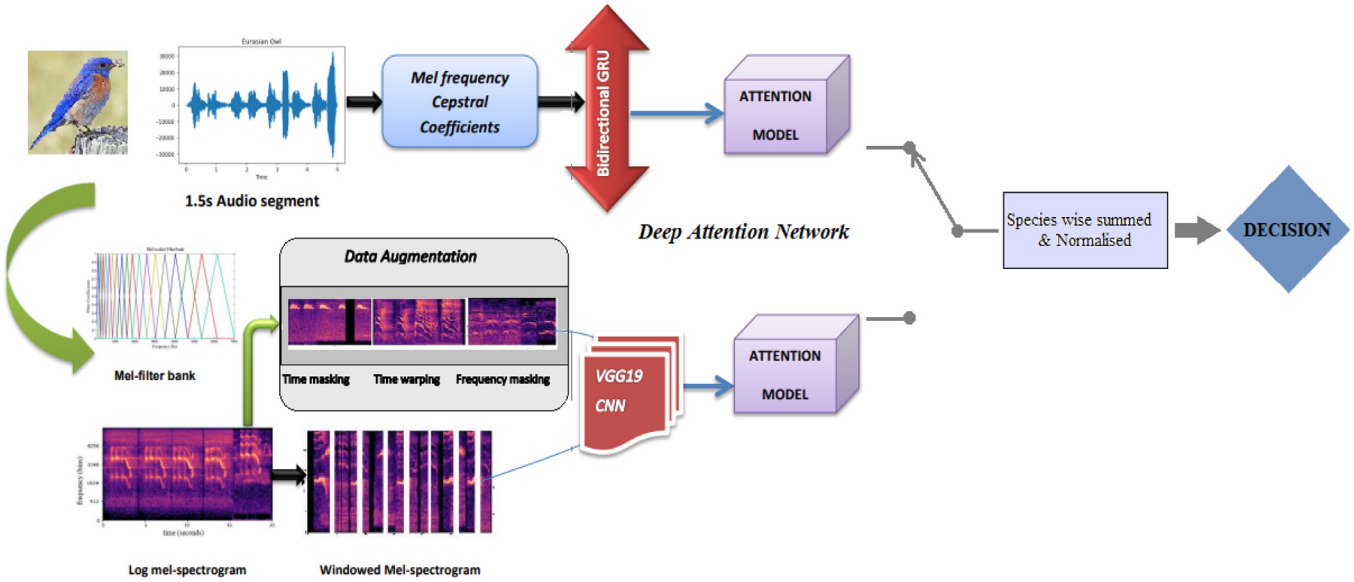

Fig. 3. Block diagram of the proposed methods for multi-label bird classification. Scheme 1(Top branch): Attention-BiGRU, Scheme 2(Bottom branch): Attention-VGG model.

of timbre [24]. The librosa python package is used to compute 40-dimensional MFCCs in the front-end. The formula below [25] is used to calculate Mel-frequency.

$$Mel(f) = 2595log(1 + \frac{f}{700}) \quad (1)$$

where Mel(f) is the subjective pitch(Mels) corresponding to a frequency(Hz).

Mel-spectrogram is widely used in speech and music processing applications [26,27]. Mel-spectrogram approximates how the human auditory system works and can be seen as the spectrogram smoothed, with high precision in the low frequencies and low precision in the high frequencies [28]. The time-domain waveform is converted to a time–frequency representation using a short-time Fourier transform (STFT) with a frame size of 30 ms and a hop size of 10 ms. Then the linear frequency spectrogram is converted to a mel-scale, and we use 128 Mel-frequency bins, which preserves harmonic characteristics while reducing the dimensionality of the input.

3.2. Network architecture

Network architectures for the two schemes are discussed below.

3.2.1. Hierarchical attention-based sequential BiGRU

GRU is the newest entrant in sequence modelling techniques, following RNN and LSTM, and it promises to outperform the other two in a variety of sequential processing applications [29,30]. GRU can process memories of sequential data by storing previous inputs in the internal state and planning target vectors based on the history of previous inputs. The ability of GRU to capture long-range dependency in learning temporal patterns is investigated using MFCC features in the proposed research. Since RNN work with sequential data, they've been widely used for language understanding [31]. There are different cells that improve neural networks when tracking long-term dependencies for deep sentence understanding.

The primary difference between GRU and LSTM is that GRU's bag has two gates, one for reset and one for update, whereas LSTM's bag has three gates. GRU cells have the advantage of being just as powerful as LSTM cells [29] for small data sets while requiring less computing power. The LSTM is more complicated than the two-gated GRU cell because it has three gates. The equations that govern fully gated recurrent units are as follows [32];

$$z_t = \sigma_g(W_z.x_t + U_z.h_{t-1} + b_z) \quad (2)$$

$$r_t = \sigma_g(W_r.x_t + U_r.h_{t-1} + b_r) \quad (3)$$

$$\hat{h}_t = \phi_h(W_h.x_t + U_h.(r_t \odot h_{t-1}) + b_h) \quad (4)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \quad (5)$$

where variables $x_t, h_t, \hat{h}_t, z_t, r_t$ represent input, new hidden state (output), candidate activation, update gate and reset gate vectors, respectively. W, U, and b represent parameter matrices. σ_g and ϕ_h are activation functions. The operation $\odot$ denotes Hadamard product and '.' represents elementwise multiplication. Fig. 4 shows the schematic of the proposed GRU model.

The Sequential GRU architecture is shown in Table 1.

The proposed work aims to develop a novel deep learning architecture for audio classification that incorporates different layers. Specifically, BiGRU networks are used in conjunction with an attention mechanism for multi-label bird species classification. BiGRU consists of two GRUs that access both the preceding and succeeding features by combining a forward-hidden layer, $\overrightarrow{h_t}$ and a backward-hidden layer, $\overleftarrow{h_t}$.

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \quad (6)$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t-1}}) \quad (7)$$

The output of BiGRU, h_t=[$\overrightarrow{h_t}$ $\overleftarrow{h_t}$] is fed through a one-layer MLP to obtain a hidden representation of h_t, then the normalized weights, $b_t$can be calculated as,

$$b_t = \frac{exp(W * h_t)}{\sum_{t=1}^{T} exp(W * h_t)} \quad (8)$$

where W is the weight matrix of the MLP and T is the length of input. The feature vector v is obtained as the weighted sum (weights, b_t) of all the features as,

$$v = \sum_{t=1}^{T} b_t * h_t \quad (9)$$

Table 2 describes the architecture of the proposed BiGRU model with the hierarchical attention mechanism. The input to the model is of size 64 × 40. These 40-dimensional MFCCs are then batch normalized before the activation of input. The model's first layer was chosen to be the aforementioned BiGRU layer, with 128 hidden neurons. The number of hidden units is then doubled to 256 due to the bidirectional nature of the layer. The number of hidden units was determined through rigorous testing that included different numbers of neurons. Following the attention layer, the most effective layer is a regular dense layer with the ReLU activation function. The four final layers are the dense layers, the first of which is 256 neurons, the second and third of which are 64 neurons, and the fully connected layer, which is 10 units that correspond to the number of classes outputted.

Hierarchical attention networks are widely used for document classification [33]. The hierarchical attention-based classification scheme has a parallel approach in the context of feature selection. A genre-dependent feature selection or hierarchical feature selection was used instead of using the entire training database to obtain a single list of selected features. Fig. 5 shows the schematic of the proposed Attention-BiGRU model.

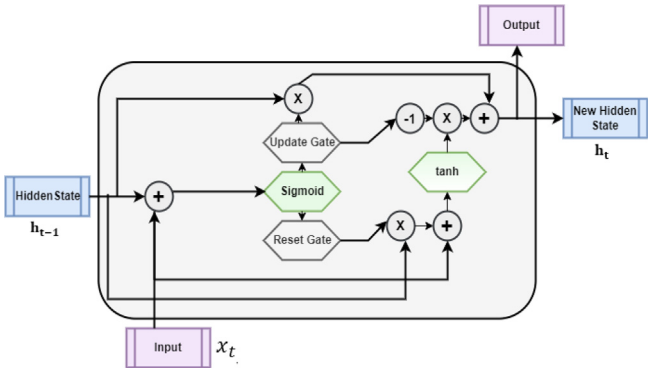

Fig. 4. Schematic of GRU.

Table 1
GRU architecture used for the experiment.

Sl no.	Layer	Ouput shape
1	GRU	(None, 64, 1048)
2	GRU	(None, 768)
3	Dense	(None, 10)

Table 2
Proposed Hierarchical Attention-based Sequential BiGRU.

Sl no.	Layer	Ouput shape
1	Input layer	(None, 64, 40)
2	Batch-Normalization	(None, 64, 40)
3	Bidirectional GRU	(None, 64, 256)
4	Attention	(None, 256)
5	Dense	(None, 256)
6	Dense	(None, 64)
7	Dense	(None, 64)
8	Dense	(None, 10)

3.2.2. Sequential transfer learning

Spectrograms are a visual representation of audio sequences that can be used to detect music genres [26]. We also evaluated the Mel-spectrogram-VGG CNN framework with and without attention for the proposed task. The frequencies have been converted to mel scale in a Mel-spectrogram. The mel scale was created in an attempt to scale frequency data in a way that is more similar to how humans perceive sound. Fig. 6 depicts a representation of audio files containing two (Grey Go-away & Red. Lapwing) and three (Asiankoel, Red. Lapwing & W. Wood Pewee) bird species, respectively. A sliding window is used to analyse the Mel-spectrogram over a short period. Sliced Mel-spectrograms are used

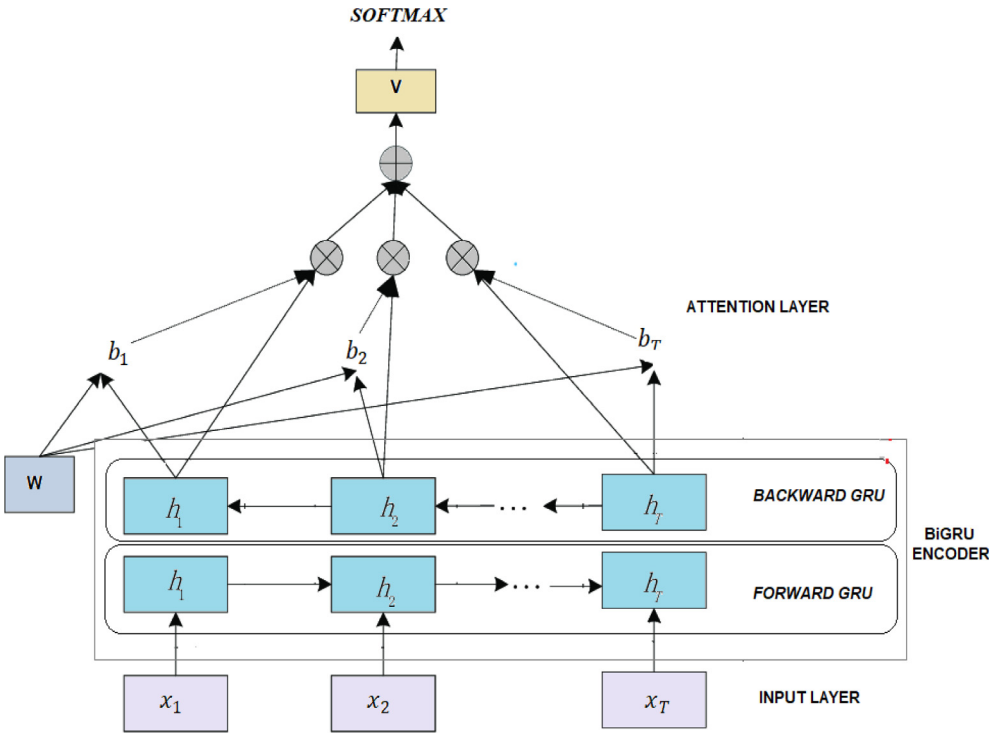

Fig. 5. Schematic of Attention-BiGRU.

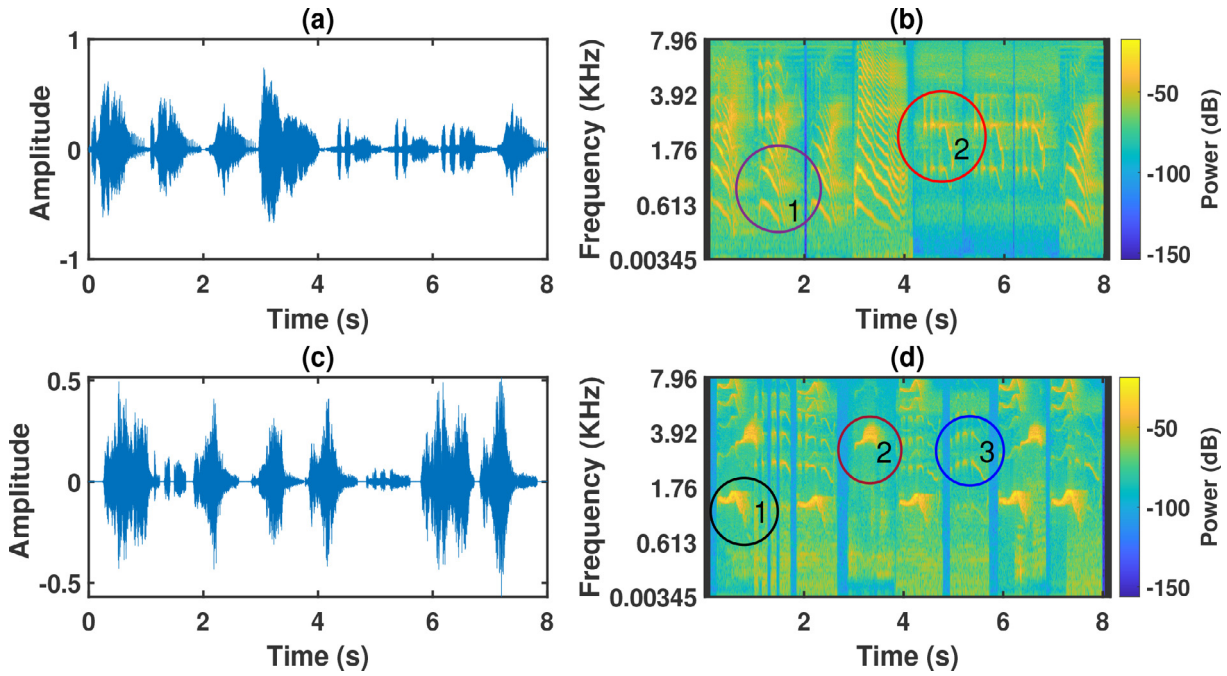

Fig. 6. Representation of bird's vocalization with multiple bird sounds in a single audio recording. Repetitive patterns in the vocalization are shown in circles for two species (top), and three species (bottom).

for training and testing. Data augmentation techniques such as time-warping, frequency masking, and time-masking are used to create additional Mel-spectrograms.

Transfer learning is a popular deep learning technique in which a model created for one task is used as the base model for a different task. We experimented with two variants of pre-trained VGG networks, namely VGG-16 and VGG-19 for the proposed task. The VGG-16 architecture is depicted in Table 3, which consists of 13 convolutional and three fully connected layers. The convolution layers are 3×3 layers, with the same padding and stride 1, and the pooling layers are 2×2 layers, with stride 2. After data preparation, the 432×1008 arrays of Mel-spectrograms are reduced to 256×256 pixels. The last feature map has 512 channels before the fully connected layers. Finally, using the dense layer with 4096 units, the fully connected layers are added, followed by a dropout layer with a value of 0.5. The output layer employs sig-

Table 3
VGG-16 architecture

Input shape	Description
3 × 256 × 256	Convolution 3 × 3 (2 times), Stride 1
64 × 256 × 256	Max-pooling 2 × 2, Stride 2
64 × 128 × 128	Convolution 3 × 3 (2 times), Stride 1
128 × 128 × 128	Max-pooling 2 × 2, Stride 2
128 × 64v64	Convolution 3 × 3 (3 times), Stride 1
256 × 64 × 64	Max-pooling 2 × 2, Stride 2
256 × 32 × 32	Convolution 3 × 3 (3 times), Stride 1
512 × 32 × 32	Max-pooling 2 × 2, Stride 2
512 × 16 × 16	Convolution 3 × 3 (3 times), Stride 1
512 × 16 × 16	Max-pooling 2 × 2, Stride 2
32768	Flattened and fully connected(FC)
4096	Dropout 0.5
4096	Dropout 0.5
10	sigmoid

moid activation to forecast the species probabilities for each file. The pre-processing model VGG-19 is used, and its architecture is very similar to that of VGG-16. The model has three additional convolutional layers. The VGG-19 architecture has 16 layers of CNNs, three fully connected layers, and a softmax function layer. The network depth has been improved over traditional CNNs. It has a better structure than a single convolutional layer because it alternates between multiple convolutional layers and non-linear activation layers. The layer structure allows for better image feature extraction, downsampling with max-pooling, and modification of the linear unit (ReLU) as the activation function, which selects the largest value in the image area as the pooled value of the area. Fig. 7 illustrates the VGG-19 architecture commonly used in computer vision and natural language processing tasks.

We customized VGG architectures to incorporate the attention mechanism for the proposed task. Table 4 shows the Attention-based Sequential VGG-19 architecture used for the experiment. The softmax and the last fully connected layers are replaced with the custom architecture to adapt VGG-19 to the proposed method. Features were extracted using the VGG-19 model, which was then batch normalized. The attention mechanism improves classification accuracy and makes the model more efficient. We use an attention mechanism to turn pixels on and off in the global average pooling. It essentially creates a spatial mask that specifies which feature map regions should be used. It's a fundamental feature map segmentation that ignores the nearby region. These deeper features are retrieved using convolutional layers.

256 filters with a 1 × 1 kernel size were used in the first convolutional layer followed by ReLU activation function. Each of the 64 filters in the second convolution layer had a 1 × 1 kernel size and a ReLU function. The third convolutional layer consisted of 16 1 × 1 size filters. The fourth convolutional layer is a locally connected layer similar to the Conv2d layer, except that weights are not shared, so each input patch gets its own set of filters. In this layer, a single kernel sigmoid activation is used. An attention layer with a linear activation function is added and is not trained during the training because it is only used for attention. The deeper features extracted after adding more convolutional layers are multiplied with the initial extracted features of the VGG-19 model to calculate mask features. Global average pooling features and GAP mask are generated from mask features and attention layers, respectively, using global average pooling. A lambda layer was used to rescale the features. Two dropout layers with a 0.2 drop rate are used, with the fully connected layer followed by a dense layer with 1024 neurons and ReLU activation. Another layer with 512 units and ReLU activation has been added. The final output layer was given a softmax activation with ten neurons to classify the ten labels.

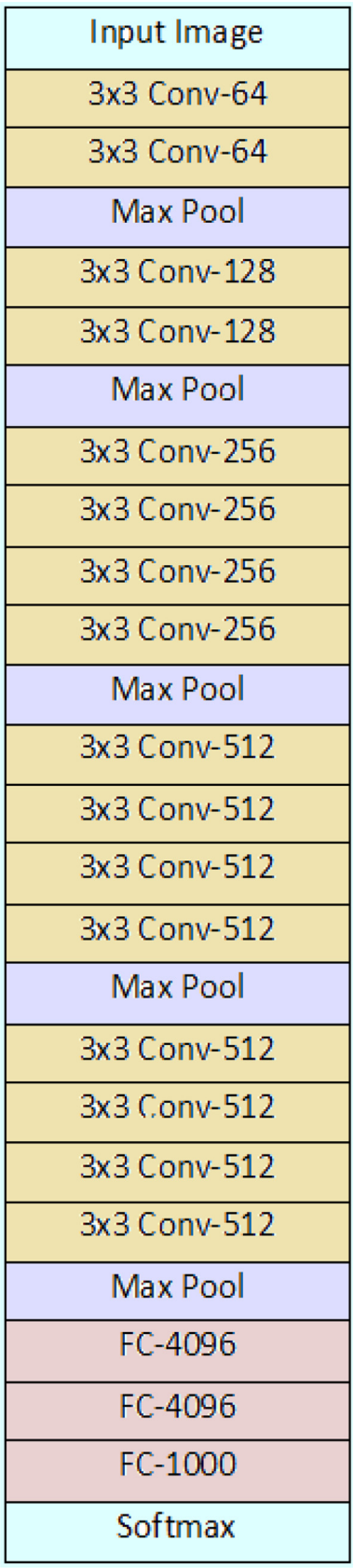

Fig. 7. VGG-19 architecture.

Table 4
Attention-based Sequential VGG-19 architecture.

Layer	Output Shape
Input_Layer	(None, 432, 432, 3)
vgg19	(None, 13, 13, 512)
batch_normalization	(None, 13, 13, 512)
conv2d	(None, 13, 13, 256)
conv2d_1	(None, 13, 13, 64)
conv2d_2	(None, 13, 13, 16)
locally_connected2d	(None, 13, 13, 1)
conv2d_3	(None, 13, 13, 512)
multiply	(None, 13, 13, 512)
global_average_pooling2d	(None, 512)
global_average_pooling2d_1	(None, 512)
RescaleGAP (Lambda)	(None, 512)
dropout	(None, 512)
dense	(None, 1024)
dropout_1	(None, 1024)
dense_1	(None, 512)
dropout_2	(None, 512)
dense_2	(None, 10)

3.3. Sequential aggregation & decision strategy

Audio recordings of longer duration are sliced into fixed-length segments and fed into the attention model. MFCC features and Mel-spectrograms of sliced segments are fed into the respective Attention-BiGRU and Attention-VGG-19 CNN as shown in Fig. 3. The trained neural network then predicts the probability of each species present in a segment. We used an aggregation strategy that aggregated and normalized the probability outputs to decide on the test file. The species with the highest average probability are those that are predicted for each test file. Fig. 8 illustrates the schematic of the sequential aggregation process used in all the proposed methods.

4. Performance evaluation

4.1. Data set

The audio recordings are collected from the widely used Xeno-canto bird sound database [34] of the Xeno-Canto Foundation (Xeno-canto).[2] The Xeno-canto foundation is an online bird audio database with recordings of wild birds from around the world. Xeno-canto aims to have a representation of all bird sounds, meaning all taxa, to subspecies level, all of the geographic variability, at all stages of development. It has collected over 697364 sound recordings from more than 10,000 species worldwide.

This study uses 250 bird call files (flight call, alarm call, intimidation call, nocturnal flight call etc.) varying in length from seconds to several minutes consisting of 10 species downloaded from the Xeno-Canto website. Our dataset is composed of Asian Koel (AK), Blue Jay (BJ), House Crow (HC), Mallard Duck (MD), Grey Go-away (GG), Red. Lapwing (RL), Eurasian Owl (EO), Indian Peafowl (IP), House Sparrow (HS), and W. Wood Pewee (WW). The stereo input files are converted to 16-bit mono waves by taking the mean of the left and right channels, and then it is downsampled to 16,000 Hz. For training the models, we split the dataset into training and testing. Due to the variable length of the audio samples, the files have been refined so that each audio file contains one 1.5-s vocalization. The train set contains 1078 isolated audio recordings of 10 species as described in Table 5. The test data consists of 434 audio files with time-overlapping vocalizations and multiple bird calls (2 or 3 distinct bird species vocalizations in each audio file). The test dataset used for the proposed experiment is described in Table 6.

[2] www.xeno-canto.org (Xeno-canto)

4.2. Data augmentation

In visual representation-based methodologies, the authors [35] highlight the importance of more training sets. CNN, according to [36], requires a large amount of data to achieve better results because it fails with small data sizes. Data augmentation methods are used to compensate for the lack of sufficient training data. The proposed scheme uses SpecAugment [37], which operates on the log Mel-spectrogram of the input audio rather than the raw audio. The augmentation policy includes warping the features, masking blocks of frequency channels and masking blocks of time steps.

In time warping, a deformation of the time series in the time direction is applied. We can visualize a log Mel-spectrogram with τ time steps as an image with the time axis horizontal and the frequency axis vertical. Within the time steps $(W, \tau - W)$, a random point along the horizontal line passing through the centre of the Mel-spectrogram is to be warped to the left or right by a distance w. w is chosen from a uniform distribution from 0 to the time warp parameter W along that line. In time and frequency masking, we mask a block of consecutive time steps or mel frequency channels. Time masking is applied to mask t consecutive time steps $[t_0; t_0 + t)$, where t is selected from a uniform distribution ranging from 0 to the time mask parameter T, and t_0 is selected from [0, $\tau - t$). Fig. 9 shows examples of data augmentations applied to the Mel-spectrogram of audio file input(W. Wood Pewee).

We used the parameter setting chosen in [37] to generate masked/warped Mel-spectrograms. The authors masked multiple patches along the time and frequency axis. The multiple masks may overlap. But in our approach, we utilized single patch (patch width is empirically chosen) in the Mel-spectrogram for masking. Time warping parameter contributes but is not a major factor in improving performance as indicated in [37]. We used the basic approach, called LB with parameters W = 80, F = 27, T = 100, mF = 1, and mT = 1. W, F and T represents warping, frequency and time masking parameters respectively. mF and mT denote the number of frequency and time masks applied. We generated 3344 Mel-spectrograms for the transfer learning model as part of data augmentation based on the experimental set-up described in [37].

4.3. Experimental framework

This work uses an acoustic RNN-based method, Attention-BiGRU, for efficiently detecting multiple bird species. The Attention-BiGRU was trained for a maximum of 15 epochs with a batch size of 32. The GRU models' training process required far fewer epochs and a much higher learning rate to achieve optimal performance, reflecting the relatively simple nature of the networks themselves. Adaptive moment estimation (Adam) was used to train the network by optimizing the categorical cross-entropy between predictions and targets. The softmax activation function was used to train the model. During the experiment, 20% of the corpus was used for validation. Training accuracy and loss of the Attention-BiGRU model are shown in Fig. 10.

Audio files are converted to a time–frequency representation using STFT, with 480 samples for the window. The Mel-spectrogram is segmented into small duration chunks and fed to a pre-trained VGG model. 256 Mel-frequency bins were used to synthesize each bird call into its Mel-spectrogram representation. As the attention layer is employed for attention, it is not trained during the training process. The VGG-19 and the other layers in the model are completely trained. The model is compiled using an Adam optimizer with a batch size of 32, and the initial training

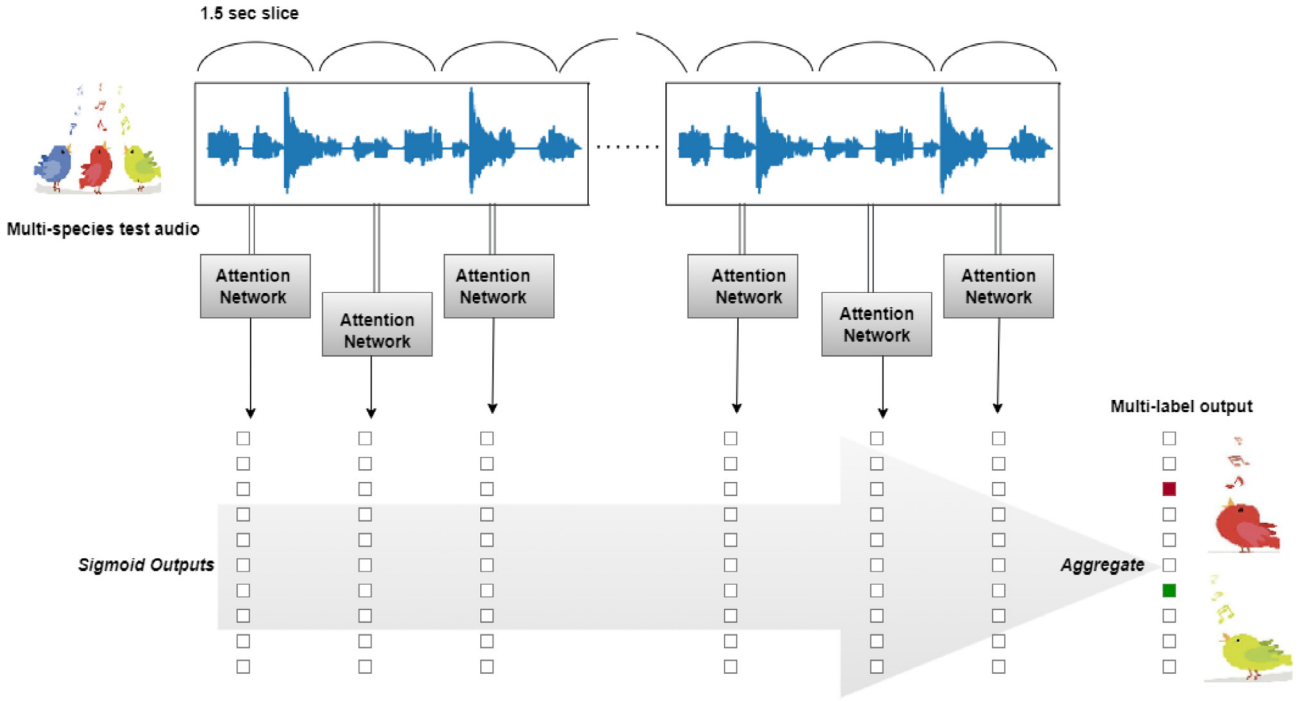

Fig. 8. Schematic of sliding window analysis on the input audio.

Table 5
Data statistics for training.

SL.No	Bird Species	Scientific name	Bird Id.	XC files	Calls
1	House Crow	Corvus splendens	HC	27	111
2	Mallard Duck	Anas platyrhynchos	MD	25	106
3	Asian Koel	Eudynamys scolopaceus chinensis	AK	26	121
4	Eurasian Owl	Bubo bubo	EO	25	107
5	House Sparrow	Passer domesticus	HS	24	100
6	Blue Jay	Cyanocitta cristata	BJ	27	109
7	Red. Lapwing	Vanellus vanellus	RL	24	104
8	Grey Go-away	Corythaixoides concolor	GG	19	109
9	Indian Peafowl	Pavo cristatus	IP	29	103
10	W. Wood Pewee	Contopus sordidulus	WW	24	108
	Total			250	1078

Table 6
Test Dataset specification.

Sl.No	Class	Number (Audio Files)	# Calls
1	Train Files	1078	1078
2	Test Files		
	Two Species	334	668
	Three Species	100	300
	Total	1512	**2046**

for 30 epochs is done with a learning rate of 0.0001; then, the learning rate is changed to 0.00001, and the model has trained again for more epochs, and the best results are saved. During the experiment, 10% of the corpus was used for validation. The output layer has ten perceptrons and the softmax activation function is used.

5. Analysis of results

Precision, recall, and F1-score are used to evaluate the performance of attention-based models, which are then compared to that of VGG-16 and the acoustic GRU approach. Overall classification scores for VGG-16, Attention-VGG-16, Attention-VGG-19, GRU and Attention-BiGRU are 0.65, 0.67, 0.70, 0.58, 0.84, respectively.

The confusion matrix for the GRU, Attention-BiGRU, VGG-16, and Attention-VGG-16 models for the target dataset comprising two and three species are given in Tables 7, 8, 9, and 11, respectively. It is found that the proposed Attention-BiGRU model outperforms the acoustic MFCC-GRU model by a remarkable spike of 26%. Our experiment demonstrated that the attention mechanism and the GRU model are best suited to increase model accuracy even with complex datasets. In our experiments, the best-performing Attention-BiGRU model outperforms all other architectures with an accuracy of 84%. By combining the attention mechanism with BiGRU layer, the output of the intermediate node of BiGRU neurons is retained, so the model can be trained selectively to learn the input audio file. It is worth noting that the proposed attention mechanism is capable of focusing on important temporal events while reducing the impact of background noise.

The class-wise accuracy of Blue Jay, Eurasian Owl, House Sparrow, and Indian Peafowl using the MFCC-GRU approach is less than 50%. In the proposed Attention-BiGRU approach, all classes report an accuracy of greater than 70%, and the model almost halved

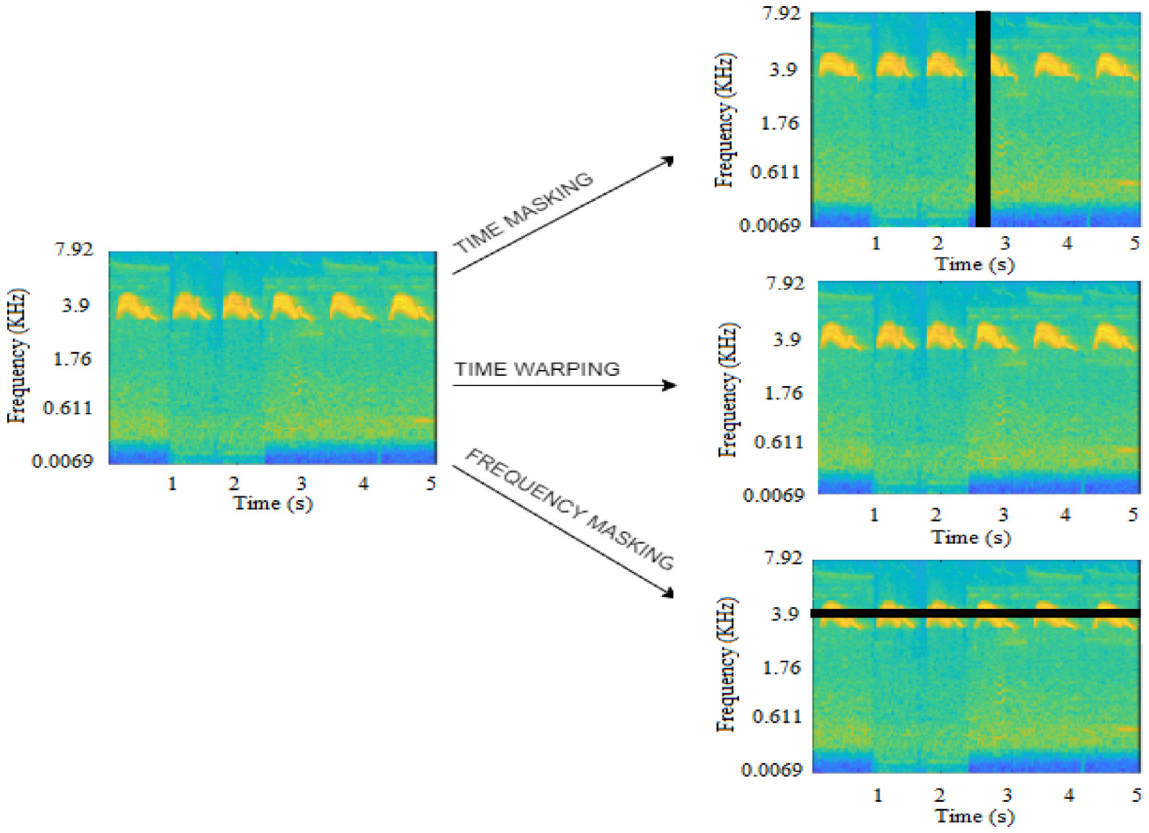

Fig. 9. Data Augmentation- SpecAugment-generated Mel-spectrograms of W. Wood Pewee.

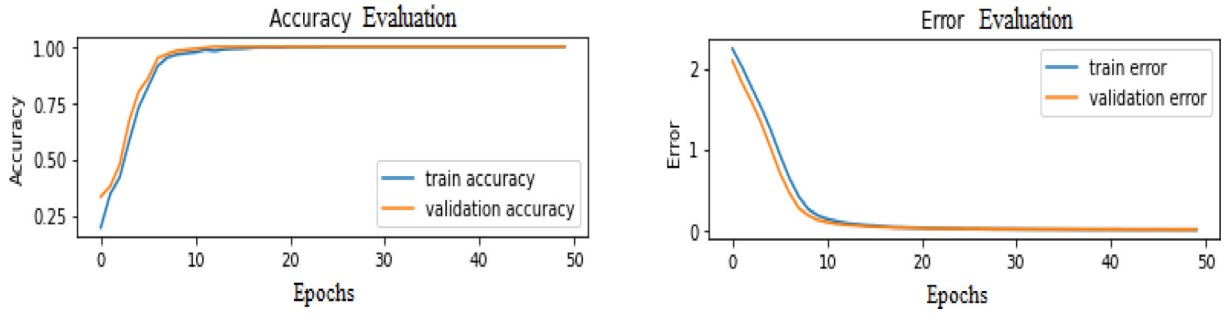

Fig. 10. Training accuracy and loss for Attention-BiGRU model.

Table 7
Confusion matrix for Attention-BiGRU model.

	AK	BJ	HC	MD	GG	RL	EO	IP	HS	WW
AK	**67**	0	0	0	1	0	1	0	0	0
BJ	4	**66**	0	2	18	6	1	0	3	1
HC	0	0	**91**	2	7	0	0	0	0	0
MD	0	4	0	**66**	15	1	1	0	1	0
GG	0	0	0	0	**145**	0	0	0	0	0
RL	1	1	0	1	0	**133**	0	0	0	0
EO	1	3	1	4	6	2	**36**	0	2	0
IP	13	0	1	0	10	1	0	**38**	0	0
HS	7	0	3	0	15	1	2	0	**91**	0
WW	0	0	0	0	0	1	2	0	2	**87**

the misclassification errors of Eurasian Owl and House Sparrow significantly. In particular, the Attention-BiGRU model yielded substantially better results for the species such as W. Wood Pewee (97%), Red. Lapwing (95%) and House Crow (93%), compared with the Attention-VGG-19 model, yielded 80%, 62%, and 82% for W. Wood Pewee, Red. Lapwing and House Crow, respectively. Unlike models without attention, which are incapable of learning some species, the attention mechanism concentrates on all species.

Table 8
Confusion matrix for Sequential GRU model.

	AK	BJ	HC	MD	GG	RL	EO	IP	HS	WW
AK	**53**	6	0	0	4	1	2	3	0	0
BJ	1	**72**	0	2	13	7	3	3	0	0
HC	0	1	**86**	5	5	0	0	3	0	0
MD	1	2	4	**67**	9	4	0	1	0	0
GG	0	30	0	0	**111**	0	0	0	4	0
RL	1	33	0	4	17	**71**	0	7	3	0
EO	7	10	3	2	2	9	**12**	10	0	0
IP	14	0	1	0	11	2	6	**29**	0	0
HS	6	19	2	5	47	8	0	4	**28**	0
WW	2	23	0	2	11	0	0	8	0	**46**

Table 9
Confusion matrix for VGG-16 model.

	AK	BJ	HC	MD	GG	RL	EO	IP	HS	WW
AK	**55**	5	2	2	3	2	0	0	0	0
BJ	5	**65**	1	4	8	10	1	3	3	1
HC	2	1	**72**	3	8	7	2	1	3	1
MD	7	2	7	**50**	5	16	0	0	0	1
GG	2	14	6	5	**94**	8	1	1	8	6
RL	12	18	1	10	6	**67**	7	6	4	5
EO	4	5	1	1	3	5	**29**	6	1	0
IP	0	2	2	0	0	0	2	**56**	1	0
HS	3	8	2	3	12	28	1	3	**59**	0
WW	0	6	1	2	3	10	0	0	4	**66**

Table 10
Precision (P), recall (R), and F1 score for the experiment

SL.No	Species name	VGG-16			Attention-VGG-16			Attention-VGG-19			GRU			Attention-GRU		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	Asian Koel	0.61	0.80	0.70	0.77	0.74	0.76	0.82	0.84	0.83	0.62	0.77	0.69	0.72	0.97	0.83
2	Blue Jay	0.51	0.64	0.57	0.66	0.60	0.63	0.61	0.75	0.67	0.37	0.71	0.48	0.90	0.65	0.76
3	House Crow	0.76	0.72	0.74	0.79	0.86	0.82	0.83	0.81	0.82	0.90	0.86	0.88	0.95	0.91	0.93
4	Mallard Duck	0.62	0.57	0.60	0.69	0.58	0.63	0.55	0.73	0.62	0.77	0.76	0.76	0.88	0.75	0.81
5	Grey Go-away	0.66	0.65	0.66	0.67	0.72	0.70	0.66	0.58	0.61	0.48	0.76	0.60	0.67	1.00	0.81
6	Red. Lapwing	0.44	0.50	0.46	0.74	0.54	0.63	0.73	0.54	0.62	0.70	0.52	0.60	0.92	0.98	0.95
7	Eurasian Owl	0.67	0.53	0.60	0.64	0.53	0.58	0.70	0.67	0.68	0.52	0.22	0.31	0.84	0.65	0.73
8	I. Peafowl	0.74	0.89	0.80	0.59	0.94	0.72	0.49	0.94	0.64	0.43	0.46	0.44	1.00	0.60	0.75
9	House Sparrow	0.71	0.50	0.58	0.67	0.60	0.63	0.78	0.53	0.63	0.80	0.23	0.36	0.91	0.76	0.83
10	W. Wood Pewee	0.82	0.72	0.77	0.57	0.75	0.65	0.90	0.73	0.80	1.00	0.50	0.67	0.99	0.94	0.97
	Macro Average	0.65	0.65	**0.65**	0.68	0.66	**0.67**	0.70	0.71	**0.70**	0.66	0.58	**0.58**	0.88	0.82	**0.84**

Table 11
Confusion matrix for Attention-VGG-16 model.

	AK	BJ	HC	MD	GG	RL	EO	IP	HS	WW
AK	**51**	1	5	2	4	0	2	0	1	3
BJ	2	**61**	2	2	13	4	2	11	2	2
HC	0	3	**86**	2	5	2	0	0	0	2
MD	3	1	7	**51**	12	8	0	5	0	1
GG	1	8	0	7	**105**	3	4	1	10	6
RL	1	8	3	4	7	**74**	2	15	11	11
EO	1	5	1	2	5	0	**29**	5	1	6
IP	0	1	2	0	0	0	0	**59**	0	1
HS	5	0	2	2	3	9	4	4	**71**	19
WW	2	4	1	2	2	0	2	0	10	**69**

Despite the promising results, nothing is known about how most of the complicated models work internally or how they attain such high performance. The development of improved models becomes a trial-and-error process without a clear understanding of how and why they work. The goal of visualizing a feature map for a certain input image is to figure out which of the input's features are detected or preserved in the feature maps [38]. The assumption is that feature maps near the input capture small or fine-grained data, but feature maps near the model's output detect more generic traits.

The application of the filters in the first convolutional layer results in numerous alternative versions of the bird call, each with different attributes highlighted. The most activated neurons in the first layer for all species strongly suggest that their primary

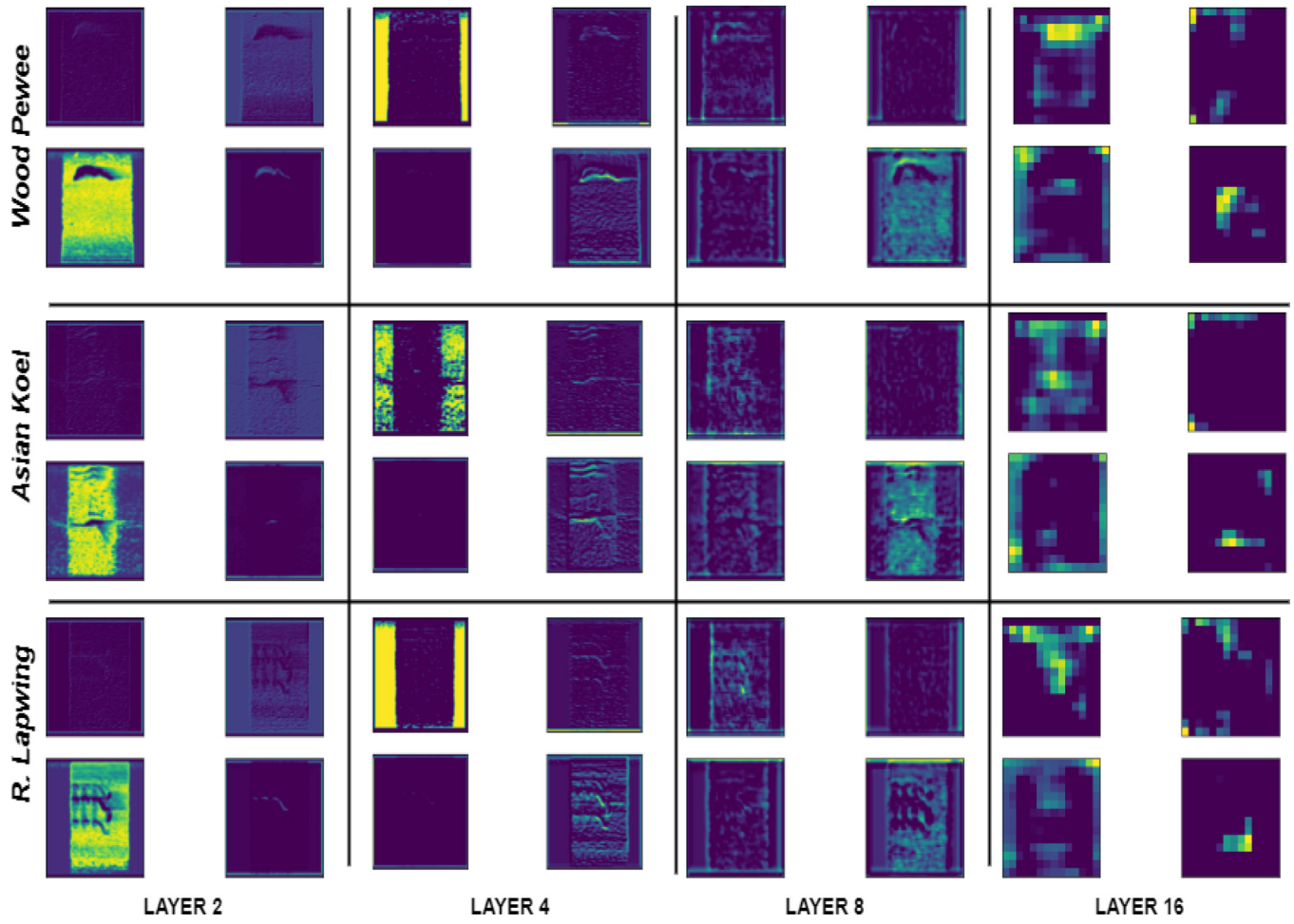

Fig. 11. Visualization of feature maps after the output of the VGG-16 convolutional layers; Layer 2 (column 1), Layer 4 (column 2), Layer 8 (column 3), Layer 16 (column 4) of species W. Wood Pewee (row 1), Asian Koel (row 2), Red. Lapwing (row 3) respectively.

function is to identify horizontal edges in the input Mel-spectrogram to detect a harmonic component, as illustrated in the first columns of Fig. 11. Small or inhibitory weights are indicated by the dark region, while large or excitatory weights are indicated by the light region. The result of applying the filters to an input, such as the input image or another feature map, is captured by activation maps, also known as feature maps. This is an intriguing result that broadly matches our expectations. The model is updated to plot feature maps based on the output of other convolutional layers such as layers 2, 4, 8 and 16 as shown in Fig. 11. We can observe that the feature maps near the model's input capture a lot of fine detail in the image, while the feature maps display less and less detail as we go deeper into the model. Although it is unclear from the final image that the model saw the patterns of the bird call in the Mel-spectrogram, we generally lose the ability to comprehend these deeper feature maps.

Table 10 shows the precision, recall, and F1 measure of the experiments. The GRU framework's macro average precision, recall, and F1 measures are 0.66, 0.58, and 0.58, respectively, while the Attention-VGG-19 model's macro average precision, recall, and F1 measures are 0.70, 0.71, and 0.70, respectively. The metrics reported by the proposed Attention-BiGRU approach are 0.88, 0.82, and 0.84, respectively.

Multi-label classification tasks are also evaluated using hamming loss and exact match ratio metrics [40]. Hamming loss takes into account the prediction error (an incorrect label is predicted) and missing error (a relevant label not predicted), normalized over

Table 12
Hamming loss and Exact match for classification models.

Sl.No	Method	Hamming loss	Exact match
1	Sequential VGG-16	0.365	0.290
2	**Attention-VGG-16**	0.320	0.350
3	**Attention-VGG-19**	**0.318**	**0.370**
4	Sequential GRU	0.408	0.244
5	**Attention-BiGRU**	**0.144**	**0.702**

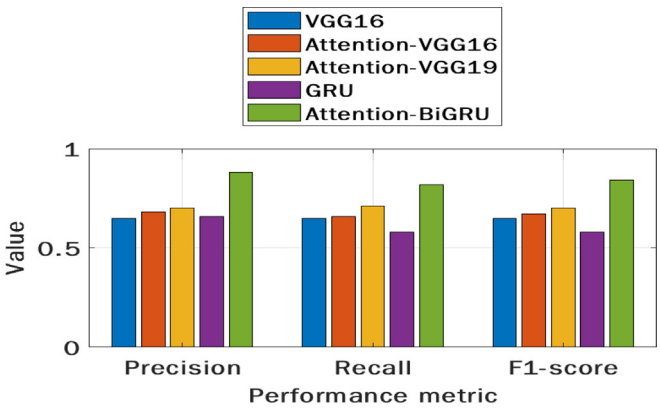

Fig. 12. Performance metrics for various schemes.

Table 13
Performance comparison with existing methods.

Sl.No	Method	Precision	Recall	F1-Score
1	T. Grill et al. [Model1] [39]	0.50	0.50	0.45
2	T. Grill et al. [Model2] [39]	0.51	0.48	0.48
3	D. B. Efremova et al. (Extended to multi-label) [16]	0.61	0.54	0.53
4	Multi-label- Transfer Learning Approach [41]	0.65	0.65	0.65
5	Bag-of-birds Approach [42]	0.70	0.87	0.78
6	**Proposed Attention-VGG-16**	0.68	0.66	0.67
7	**Proposed Attention-VGG-19**	0.70	0.71	0.70
8	**Proposed Attention-BiGRU**	0.88	0.82	**0.84**

a total number of classes and a total number of examples. The exact match evaluation metric extends the concept of the accuracy from the single-label classification problem to a multi-label classification problem. Files, where all ground truth labels are estimated correctly, are counted for computing the exact match metric. In Table 12, we present the hamming loss and exact match obtained for all the proposed classification algorithms. Our proposed Attention-BiGRU achieves a hamming loss of 0.144 and an exact match of 0.702. It is worth noting that the attention mechanism shows performance improvement for both schemes, with better performance for the Hierarchical Attention-BiGRU model. The performance metrics for the proposed acoustic RNN approach and transfer learning schemes with and without attention are compared in Fig. 12. The comparison of various algorithms using our dataset in terms of precision, recall, and F1 score is listed in Table 13. T. Grill et al. [39] compared two approaches to detect the presence of bird calls in audio recordings using feed-forward CNNs trained on Mel-scaled log-magnitude spectrograms. It reports precision, recall, and an F1 score of 0.50, 0.50, and 0.45 for model 1 (Global architecture) and 0.51, 0.48, and 0.48 for model 2 (Local architecture), respectively. D. B. Efremova et al. [16] used one of the transfer learning-based ResNet-50 models to measure the efficacy of bird call classification. As mentioned in the paper, we extended the approach for multi-label classification. This model reports an F1 score of 0.53 when using our multi-label dataset. A bag-of-birds approach based on a shallow artificial neural network trained on pre-computed sound features report F1-score of 0.78 for 10 species. Transfer learning-based deep convolutional architecture for multi-label classification, proposed in [41] reports F1-score of 0.65. The F1 metric for our best-performing Attention-BiGRU model using sequential aggregation is 0.84, which is 39%, 36%, and 31% superior to the existing models[16,39]. The results clearly show that the attention mechanism combined with sliding window analysis improved the detection performance of multiple overlapping bird species, a significant challenge.

Compared to the pre-trained VGG network based on Mel-spectrograms, there is a significant improvement in the acoustic cue model with attention. It is worth noting that the proposed Attention-BiGRU architecture outperforms the visual and acoustic frameworks. By learning and training, the proposed architecture can learn the areas where attention needs to be paid in each new spectrogram, thereby forming attention. It is also noted that the GRU framework without attention fails to differentiate overlapping partials in many simultaneous vocalizations. As a result of the simultaneous consideration of forward and backward states during the inference process, the 84% superiority of our experimental results suggests that bidirectional RNNs may be better suited to achieve higher performance in multi-label audio classification.

6. Conclusion

The goal of the paper was to identify multiple bird species, and the difficult task was to distinguish between species using overlapping audio recordings. An Attention-BiGRU architecture with a sequential aggregation model is used for the proposed classification task. It is desirable for conservation purposes to have highly accurate (high precision) identification of all target species from a set of many other species. A deep learning method is also implemented using the VGG-19 pre-trained network and attention mapping. Compared to the pre-trained VGG and acoustic GRU methods, the proposed method includes an attention mechanism to improve the detection performance of multiple species in audio. The Attention-BiGRU model achieves the best performance with an F1 score of 0.84.

Data availability statement

The datasets analyzed in this manuscript are publicly available. It can be accessed at the Xeno-canto bird sound database (www.xeno-canto.org).

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

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

References

[1] Tu G, Wen J, Liu C, Jiang D, Cambria E. Context- and Sentiment-Aware Networks for Emotion Recognition in Conversation. IEEE Trans 2022;AI.:1–9.
[2] V. Mnih, N. Heess, A. Graves and K. Kavukcuoglu, Recurrent Models of Visual Attention, CoRR abs/1406.6247, 2014, http://arxiv.org/abs/1406.6247.
[3] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben and B.W. Schuller, Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap, arXiv preprint arXiv:2203.07378, 2022.
[4] Al-Malla MA, Jafar A, Ghneim N. Image Captioning Model using Attention and Object Features to Mimic Human Image Understanding. J Big Data 2022;9 (20):1–16.
[5] S. Fagerlund, Automatic Recognition of Bird Species by their Sounds, Masters Thesis, Helsinki University of Technology, Finland, 2004.
[6] Jancovic P, Kokuer M. Bird Species Recognition using Unsupervised Modeling of Individual Vocalization Elements. IEEE/ACM Trans Audio Speech Lang Process 2019;27(5):932–47.
[7] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter and M. Eibl, Large-Scale Bird Sound Classification using Convolutional Neural Networks, in Proc. of CLEF, Dublin, Ireland, 2019.
[8] F. Zhang, L. Zhang, H. Chen and J. Xie, Bird Species Identification using Spectrogram based on Multi-channel Fusion of DCNNs, Entropy 2021, vol. 23, no. 11, pp. 1507, 2021.
[9] A. Sevilla and H. Glotin, Audio Bird Classification with Inception-v4 Extended with Time and Time-Frequency Attention Mechanisms, in Proc. of CLEF 2017, vol. 1866, 2017.
[10] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, Squeeze-and-Excitation Networks, in Proc. of cs.CV 2018, pp. 7132–7141, 2018.
[11] Y. Mahayossanunt, T. Thannamitsomboon and C. Keatmanee, Convolutional Neural Network and Attention Mechanism for Bone Age Prediction, in Proc. of IEEE APCCAS 2019, pp. 249–252, 2019.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is All you Need, in Proc. of cs.CL 2017, pp. 5998–6008, 2017.
[13] H. Zhao, J. Jia and V. Koltun, Exploring Self-Attention for Image Recognition, in Proc. of IEEE/CVF Conf.e on CVPR, pp. 10073–10082, 2020.
[14] Ramachandran P, Parmar N, Vaswani A, Bello I, Levskaya A, Shlens J. Stand-Alone Self-Attention in Vision Models. Adv Neural Inf Process Syst 2019(5).
[15] J. Deng, L. Cheng and Z. Wang, Self-Attention based BIGRU and Capsule Network for Named Entity Recognition, arXiv preprint arXiv:2002.00735, 2020.
[16] D.B. Efremova, M. Sankupellay and D.A. Konovalov, Data-Efficient Classification of Bird call through Convolutional Neural Networks Transfer Learning, in Proc. of IEEE DICTA, pp. 1–8, 2019.
[17] Liu F, Zheng J, Zheng L, Chen C. Combining Attention-based Bidirectional Gated Recurrent Neural Network and Two-Dimensional Convolutional Neural Network for Document-Level Sentiment Classification. IEEE Trans Neurocomput 2020:39–50.
[18] M.S. Islam, A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification, in Proc. of ICBSLP, pp. 1-4, 2018.
[19] N.T. Sima Siami-Namini and A.S. Namin, The Performance of LSTM and BiLSTM in Forecasting Time Series, in Proc. of Big Data, pp. 3285–3292, 2019.
[20] Zhang C, Wang D, Wang L, Song J, Liu S, Li J, Guan L, Liu Z, Zhang M. Temporal Data-Driven Failure Prognostics using BIGRU for Optical Networks. J Optical Comms Networking 2020;12(8):277–87.
[21] Briggs F, Lakshminarayanan B, Neal L, Fern XZ, Raich R, Hadley SJK, Hadley AS, Betts MG. Acoustic Classification of Multiple Simultaneous Bird Species: A Multi Instance Multi-Label Approach. J Acoust Soc Am 2012;131(6):4640–50.
[22] L. Zhang, M. Towsey, J. Xie, J. Zhang and P. Roe, Using Multi-Label Classification for Acoustic Pattern Detection and Assisting Bird Species Surveys, Applied Acoustics, pp. 91–98, 2016.
[23] J. Seppanan, Computational Models for Musical Meter Recognition, Masters Thesis, Tampere University of Technology, Department of Information Technology, 2015.
[24] R. Parncutt, A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms, Music Perception, pp. 409–464, 1994.
[25] M.A. Hossan, S. Memon and M.A. Gregory, A Novel Approach for MFCC Feature Extraction, in Proc. of ICSPCS, pp. 1–5, 2010.
[26] M. Sukhavasi and S. Adappa, Music Theme Recognition using CNN and Self-Attention, arXiv preprint arXiv:1911.07041, 2019.
[27] D. Ghosal and M. Kolekar, Music Genre Recognition using Deep Neural Networks and Transfer Learning, in Proc. of Interspeech, pp. 2087–2091, 2018.
[28] O'shaughnessy D. Speech Communication: Human and Machine. Universities press; 1987. pp. 1–5.
[29] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, Empirical Evaluation of Gated Recurrent Neuronal Networks on Sequence Modeling, cs.NE; arXiv:1412.3555, 2014
[30] J.X. Chen, D.M. Jiang, and Y.N. Zhang, A Hierarchical Bidirectional GRU Model With Attention for EEG-Based Emotion Classification, IEEE Access on Deep Learning Algorithms For Internet of Medical Things, pp. 118530–118540, 2019.
[31] A. Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, Tools, and Techniques to build intelligent systems, 2017.
[32] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, cs.CL, arXiv:1406.1078, 2014.
[33] Yang Z, Yanga D, Dyer C, He X, Smola A, Hovy E. Hierarchical Attention Networks for Document Classification. Proc. of Conf. North American Chapter of the Assoc. for Comput. Linguistics: Human Language Technologies. 2016:1480–9.
[34] Vellinga, Willem-Pier, Planqu'e and Robert, The Xeno-canto Collection and its Relation to Sound Recognition and Classification, in Proc. of Working Notes of CLEF, 2015.
[35] Liu C, Feng L, Liu G, Wang H, Liu S. Bottom-up Broadcast Neural Network For Music Genre Classification. Pattern Recog Lett, Multimedia Tools Appl 2021;80 (5):7313–31.
[36] M. Kaya and H.S. Bilge, Deep Metric Learning: A Survey, Symmetry 2019, vol. 11, no. 9, p. 1066, 2019.
[37] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk and Q.V. Le, SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, in Proc. of Interspeech, arXiv:1904.08779, 2019.
[38] M.D. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, in ECCV, Springer, Cham, pp. 818–833, 2014.
[39] T. Grill and J. Schluter, Two Convolutional Neural Networks for Bird Detection in Audio Signals, in EUSIPCO, pp. 1764–1768, 2017.
[40] Nazmi S, Yan X, Homaifar A, Doucette E. Evolving Multi-Label Classification Rules by Exploiting High-Order Label Correlations. Neurocomputing 2020;417:176–86.
[41] R. Rajan and N. Abdul kareem, Multi-label Bird Species Classification Using Transfer Learning, in International Conference on Communication, Control and Information Sciences (ICCISc), pp. 1–5, 2021, doi: 10.1109/ICCISc52257.2021.9484858.
[42] Ghani B, Hallerberg S. A Randomized Bag-of-Birds Approach to Study Robustness of Automated Audio Based Bird Species Classification. Appl Sci 2021;11(19):9226. https://doi.org/10.3390/app11199226.