Bird Calls Identification in Soundscape Recordings Using Deep Convolutional Neural Network

Muhammad Azeem, Ghulam Ali, Riaz Ul Amin, and Zaheer Ud Din Babar

Abstract Bird species diversity assumes a significant part in giving essential dimensions to people. Many bird species are threatened by climate change. To safeguard them, we must first determine the species to which they belong and then necessary precautions must be taken to ensure their existence. A conventional way to deal with the inspection of bird diversity is through manual review. This methodology depends on experts ecologists to accomplish exact outcomes. Another methodology is bioacoustics observing, which utilizes computerized recorders to gather wildlife vocalizations for helping researchers in bird studies. For researchers, conservation biologists, and birders, identifying bird species accurately in recorded audio files would be a game-changing tool. So an automatic bird detection system by their calls is the need of the hour. In this sense, artificial neural networks have better the recognition excellence of machine learning and deep learning methods for bird species recognition expressively in recent years. Convolutional neural networks (CNNs) are machine learning algorithms that are effective in image processing and sound detection. A CNN system for classifying bird calls is proposed and evaluated in this work using various setups and hyperparameters. The proposed model has achieved 96% of the testing accuracy of the BirdCLEF Kaggle competition 2021 dataset.

Keywords Convolutional Neural Network (CNN) · Artificial Neural Networks (ANN) · Machine Learning (ML) · Deep Learning (DL) · Bird calls · Kaggle BirdCLEF 2021 · Bird species classification · Bio acoustical monitoring · Mel spectrograms · Precision · Recall · F1-Measure · Accuracy · Evaluation matrices · Librosa · Audio detection

M. Azeem (✉) · G. Ali · R. U. Amin · Z. U. D. Babar
Department of Computer Science, University of Okara, Okara, Pakistan
e-mail: azeemchaudharyg@gmail.com

G. Ali
e-mail: ghulamali@uo.edu.pk

R. U. Amin
e-mail: dr.riazulamin@uo.edu.pk

M. Ouaissa et al. (eds.), *Big Data Analytics and Computational Intelligence for Cybersecurity*, Studies in Big Data 111,
https://doi.org/10.1007/978-3-031-05752-6_20

1 Introduction

Birds contribute an important part to ecology. Several birds, however, have been endangered as a result of human activities and environmental changes. Therefore, there is a vital need to keep track of species diversity. Bird species diversity monitoring can take two forms: field observation and acoustical monitoring. In contrast to field observation, sound monitoring provides significant benefits in terms of research scalability by gathering massive volumes of bird data. There are around 10,000 bird species in the world. Therefore, there is an urgent need to automatically classify bird species based on their calls. Taking this into account machine learning techniques are playing an important part to classify birds from their voices even from dense and noisy audios. Many state-of-the-art machine learning models have been used for this purpose like support vector machine, decision tree, resnet50, etc. Although, these models perform brilliantly in-plane datasets but losses their accuracies with datasets having noise and dense environment audios. Because birds have resilient connections with species of different kinds in ecology, they are usually recognized as a reliable indicator of animal species diversity. Around two key methods are exist to keep track of bird variety: (1) physical reviews established on explanations on the ground, and (2) audio checking utilizing independent recording units. The physical techniques, for example, the five-minute bird include utilized in New Zealand, depending on expert information on specialists and can accomplish solid outcomes. However, because the majority of species of birds are movable and counted on a spot basis, there is a chance that certain species would be missed. Manual approaches are also limited in their sustainability due to the cost of keeping professionals in remote areas. Acoustical monitoring uses acoustic sensors to support ecologists in bird studies.

Sensors can work consistently for a longer period and the gathered sounds can give an industrious and unquestionable record of the acoustic soundscape. Moreover, sensing technology is these days a modest method to assist ecologists with considering singing species when joined with automatic exploration methods. Over many years and in various locations, audio sensors were used to collect singing animal sounds. It is a difficult assignment for environmentalists to snoop to entire accounts or else to evaluate the conforming audio signals produced from sound data from the outside. There is a critical need for automated tools to manage the recordings that were obtained. Character recognition technologies have recently been used to computerize bird sound detection in acoustic records. Many example acknowledgment approaches have been investigated for programmed bird species acknowledgment. Commercial software, for example, Raven, and Song Scope is presently accessible to section and describe bird calls. While completely automated analysis techniques can be increased to deal with huge volumes of sound information, in practice, their unwavering quality and accuracy stay tricky. Building precise bird call recognizers on these actual birds singing accounts is a difficult task because various ecological rushes and calls vary geologically, occasionally, and then over the entire lifespan of animal classifications.

Consequently, semi-automated methods, where specialists included are needed to work on the accuracy of mechanized methodologies, have been investigated.

In the recent past, much research was undertaken to dissect potential methods to the stated challenge. The yearly BirdCLEF recognition challenge: a biodiversity data review campaign, may be to blame for the increased interest. The BirdCLEF 2017 [1] training dataset includes approximately 36,000 audio files from 1500 distinct species collected from Xeno-canto, with classes having varied numbers of sound samples. Infield recordings, the problem centers on recognizing single auditory species as well as separating several overlaid sounds. In 2017 [2], take up the task, with 60% accuracy for overlaid sounds and 68% for distinguishing the dominating species in their trials. Pick [3] takes a similar method in 2016, with 41.2% accuracy for multi-labeled data and 52.9% for single-labeled data. Networks trained from scratch, melscaled power spectrograms, an upper-frequency cap, and noise filtering are all used in his research. The winner of the 2016 BirdCLEF competition [4], with a single labelling score of 68.6% and a multiple labelling score of 55.5%, discusses categorization using a CNN with five convolutional and one deep layer. After isolating the noise from the genuine bird sound, spectrograms are generated using the audio files as input. Unsupervised learning [5], decision tree-based feature selection [6], recurrent CNNs [7], and Hidden Markov models [8] are all proposed as alternatives or complements to CNNs in the literature. This study looks towards generic feature extraction for a large range of species of birds. The paper makes key contributions to propose a state-of-the-art deep convolutional neural network model for bird calls identification from soundscape recordings having noises in the background. The rest of the chapter is organized as follows in section II related work will be discussed, in section III methodology is described, in section IV experiments results are discussed, and in last the conclusion a future directions will be discussed.

2 Related Work

Convolutional neural network (CNN) based models have become the most well-known methodology in birdcall recognition. All in all, the spectrogram of bird sound is viewed as the info and the model would treat the bird-call recognizable proof assignment as a picture grouping issue. This is natural, since highlights of birds call one of a kind to every animal group, for example, the pitch and the tone, which can be seen in the spectrograms by experienced natural eyes. In [9] Normalization has been applied, the mean and the fluctuation for change were determined from the whole preparing dataset. Tracking down the best CNN design is a tedious assignment and is regularly done simply by instinct. Present status of-the-art approaches attempt to handle this issue with mechanized hyperparameters search. In [10] To lessen the measure of conceivable plan choices and dependent on currently accepted procedures for CNN formats have been chosen. Every weighted layer (aside from input and output layers) use Batch Normalization, exponential linear units for unit initiation. It has not been utilized any of the extra metadata except for the class id of closer view species. The presence of various background species misshapes the training data and makes a single mark preparing especially testing. In [11] the ResNet-50 (a 50 layer deep-CNN architecture), is the preliminary deep CNN design that used residual learning in 2015. ResNet-50 has been fruitful in expanding precision in computer vision seat stamping difficulties, winning first prize in the ImageNet Large Scale Visual Recognition Challenge 2015.

Given the achievement of ResNet-50 architecture in the computer vision space, in this examination, the ResNet-50 design for mechanized bird call recognition has been used. Keras ResNet-50 is applied to other bird sounds, the outcomes might measure up more genuinely. This improved the accuracy to about 72% training precision and 65% validation precision. The accuracy started to level after 400 epochs. The utilization of more limited info spectrograms to improve the precision of the ResNet-50 model in the future. In [12] The research describes a convolutional neural network-based deep learning approach for bird melody order that was utilized in a sound record-based bird ID challenge, called BirdCLEF 2016. The preparation and test set contained about 24k and 8.5k chronicles, having a place with 999 bird species. The recorded waveforms were different regarding length and substance. We changed over the waveforms into a recurrence area and split them into equivalent portions.

In [10] The portions were taken care of into a convolutional neural organization for including realizing, which was trailed by completely associated layers for grouping. In the authority scores, our answer arrived at a MAP score of more than 40% for primary species, and a MAP score of more than 33% for primary species blended in with foundation species. In [13] Deep convolutional neural networks, at first expected for picture portrayal, are changed and adjusted to perceive the presence of birds in strong accounts. Diverse data extension systems are applied to extend model execution and further develop hypotheses to cloud account conditions and new regular environmental elements. In [12] the proposed approach is surveyed on the dataset of the bird sound location task which is fundamental for the IEEE AASP Challenge on Detection and Classification of acoustic scenes and occasions 2018. It beats past bleeding-edge achieving a region under the bend more than 95 % on the public test leaderboard. A pre-prepared Inception-v3 convolutional neural organization has been used in this exploration. The association was tweaked on 36,492 sound records tending to 1,500 bird species in the particular situation of the BirdCLEF 2017 endeavor. In [14] sound records were changed into spectrograms moreover, further took care of by applying bandpass adjusting, clatter modifying, besides, and calm area departure. For data development purposes, time moving, time broadening, pitch moving, and pitch expanding were applied. This paper [15] shows that no-tuning a pre-arranged convolutional neural organization performs better contrasted with setting up a neural association without any planning. Region variety from picture to sound space could be viably applied. The associations' results were surveyed in the BirdCLEF 2017 endeavor besides, cultivated an authority mean typical precision (MAP) score of 0.567 for standard records and a MAP score of 0.496 for records with establishment species on the test dataset.

3 Methodology

The strategies utilized to the appliance and calculate the bird species models are described in the chapter. The research design that was adopted for this study is discussed here and the procedures for data collecting are also outlined. The methodologies utilized to analyze the data are highlighted, as well as the procedures that were followed to carry out this study, are described. Furthermore, a novel CNN model is designed and trained for bird species classification.

3.1 Dataset

The dataset used in this work has been acquired from Kaggle Official BirdCLEF challenge 2021. Dataset comprises of total 6,900 audio files with Ogg format. In this research due to some system limitations, only 1500 files are considered to train and test the proposed model. The training data that has been acquired comprises 6,900 audio files for 397 species. This is too much for this research due to some system and environment limitations, dataset has been limited to species that have at least 200 recordings with a rating of 4 or better. For this purpose train_metadata.csv file is used which plants with 27 species from 8,548 audio files. As the audio file cannot pass directly to the CNN model. Therefore, spectrograms for the audio are generated which are about 4,157 in total for 27 bird species. After that spectrograms are divided into training and testing segments. A total of 1,247 spectrograms are used for testing and 2,883 spectrograms are used for training purposes which are about 70% training and 30% testing of ratio. The species list comprises very communal species such as the House Sparrow (houspa), Blue Jay (blujay), or Song Sparrow (sonspa). This is not a wicked choice at all to start experimenting.

3.2 Mel-Spectrograms

A spectrogram is a visual representation of a signal's frequency spectrum, where the frequency spectrum of a signal refers to the frequency range that the signal covers. Underdone acoustic data is not suitable for neural network input, and consequently, the acoustic pointer is typically converted interested in a time-spectral depiction. A function that extracts spectrograms for a given audio file has been defined. That function requires to contents a file with Librosa (we just use the primary 15 seconds in this work), get Mel spectrograms, and save each spectrogram as a PNG image in an operational directory for future access. Total 1500 audio files that comprise 27 species have been considered final and extracted the spectrograms. A total of 4,157 training spectrograms have been extracted. That's approximately 150 for every species which is not too wicked. To make definite the spectrograms appearance right and display

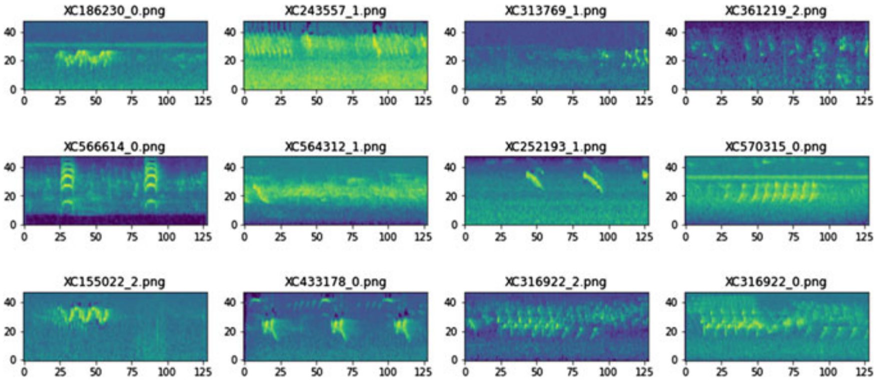

Fig. 1 Extracted spectrograms for training

the first 12. Parse audio files from train_short_audio data source and extract training samples (Fig. 1 shows the first 12 extracted spectrograms for training).

3.3 Proposed CNN Model

The proposed CNN model has comprised of four CNN blocks. Each block has the sequence like a convolutional layer (Conv), rectified linear unit (ReLU) as activation function, batch normalization layer used to overcome the gradient vanishing problem and, in the last max-pooling layer has been used to extract the more prominent features from the data. After CNN blocks global average pooling has been performed and the additional two dense layers. The last layer of the proposed model is the classification layer which is acts as an output layer as well with softmax activation function has been used for this particular layer. The convolutional layer in the first block has comprised of 16 filters with a 3 x 3 filter size a relu activation function and input size of data as hyperparameters. Max pooling layer with a pooling size of 2 x 2 has been used.

The convolutional layer in the second block has comprised of 32 filters with a 3 x 3 filter size a relu activation function and input size of data as hyperparameters. Max pooling layer with a pooling size of 2 x 2 has been used. The convolutional layer in the third block has comprised of 64 filters with a 3 x 3 filter size a relu activation function and input size of data as hyperparameters. Max pooling layer with a pooling size of 2 x 2 has been used. The convolutional layer in the fourth block has comprised of 128 filters with a 3 x 3 filter size a relu activation function and input size of data as hyperparameters. Max pooling layer with a pooling size of 2 x 2 has been used. The dense layers in the dense block have comprised 256 input channels and a relu activation function followed by a dropout layer with a rate of 0.5. Below is the summary of the model. All the detail of the results will be discussed in the next chapter results and analysis (Fig. 2 shows the model diagram).

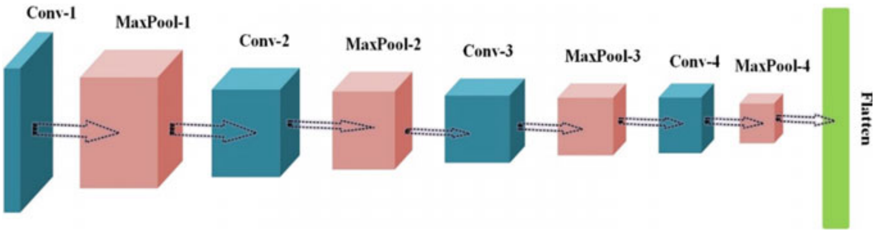

Fig. 2 Proposed CNN model

Table 1 Model training parameters

Sr. No	Parameter	Values
1	Learning rate	0.001
2	Batch size	32
3	Epochs	100
4	Optimizer	Adam
5	Activation	ReLU

3.4 Model Training

Before training the model has been compiled using Adam optimizer with an initial learning rate of 0.001 and accuracy metrics have been used for performance measures. Callbacks have been used to monitor the checkpoints as to where the model starts overfitting training will stop automatically so it makes things so easy. To train the model dataset has been divided into 70% for training and 30% for testing purposes. To fit the model for training batch size of 32, epochs size 100, and validation split of 0.2 has been used. Below is the training verbose of the model (Table 1 depicts important parameters for model training).

4 Experiments and Results

The experiment was done to examine if the system functioned appropriately with the default configuration. For the training, a preliminary LR of 0.001, Adam as an optimizer, and Cross-entropy Loss. BirdCLEF 2021 dataset is used with a partition of 70% & 30% for training and testing respectively (Table 2 shows the experimental results with accuracy).

Table 2 Experimental results

Batch size	Epochs	Mean square error	Val. accuracy
32	100	4.01	96%

It is noticed that validation accuracy does not increase significantly rather than a slight change in it. So an early stopping has occurred after 32 epochs which unloaded the computational burden from the system. The model has been trained on focal recording and soundscape recordings simultaneously. Focal audio recordings as training data can be deceptive and soundscapes have abundant greater noise levels and also comprise very unclear bird calls. The accuracy has been observed at the end of the training about 96%, which is relatively good for novel CNN-based model experiments.

A soundscape from the training data has been picked, however, the whole procedure can simply be automated and then applied to all soundscape files. The file has been loaded with Librosa, spectrograms have been extracted for 5-second chunks, and each chunk has been passed through the model and ultimately consign a label to the 5-second audio chunk. A soundscape is required that contains some of the species that have been trained the proposed model for. The "28933_SSW_20170408.ogg" file has been picked up from the dataset which looks to comprise a plethora of Song Sparrow (sonspa) vocalizations. After analyzing soundscape recording from the selected soundscape audio about 42 birds species have been identified that was encouraging results for a simple model.

Song Sparrow (sonspa) vocalizations have been identified significantly, however missed some others. The Northern Cardinal (norcar) and Red-winged Blackbird (rewbla) have not been detected successfully all though they had been in the training data. This would be a great example of the difficulties faced while analyzing soundscapes. Focal recordings as training data can be deceptive and soundscapes have much higher noise levels and also comprise very unclear bird calls. Therefore, these challenges can be mitigated by further improvement in the model organization and by preprocessing the dataset properly (Table 3 shows the soundscape analyses results of audios).

A predictive model's performance is measured using different evaluation metrics. This usually entails training a model on a dataset, then using the model to generate

Table 3 Soundscape analysis results

Row ID	Site	Audio ID	Seconds	Birds	Predictions
28933_SSW_5	SSW	28,933	5	Sonspa	Sonspa
28933_SSW_10	SSW	28,933	10	Rewbla	Rewbla
28933_SSW_15	SSW	28,933	15	Sonspa	Nocall
28933_SSW_20	SSW	28,933	20	Sonspa	Sonspa
28933_SSW_25	SSW	28,933	25	Houspa	Houspa
28933_SSW_30	SSW	28,933	30	Blujay	Blujay
28933_SSW_35	SSW	28,933	35	Sonspa	Nocall
28933_SSW_40	SSW	28,933	40	Sonspa	Sonspa
28933_SSW_45	SSW	28,933	45	Rewbla	Rewbla
28933_SSW_50	SSW	28,933	50	Houspa	Nocall

Table 4 Model evaluation matrices

Accuracy	Precision	Recall	F1-Measure
96%	100%	100%	100%

predictions on a holdout dataset that was not used during training, and comparing the predictions to the predicted values in the holdout dataset. Metrics used to evaluate the proposed model in this study are accuracy, precision, recall, and F1-measure (Table 4 shows the evaluation metrics to observe the efficiency of the proposed model).

The validation loss (blue) is shown on the right y-axis about 5% pieces of audio categories ordered by the number of training sections within every 5% piece in Fig. 3. On the left y-axis, also it displays the training loss (green) for every piece (green).

To summarize, the proposed model is trained on 4,152 spectrograms for 27 birds species. Model is trained using 32 batch sizes and 100 epochs. To observe the efficiency of the proposed model different evaluation metrics are used such as accuracy, precision, recall, and F1-measure. To make the proposed model much suitable for unseen data and minimize its error rate to enhance its accuracy and efficiency different optimizers are used such as Adam, Adamax, Nadam, Adaleta, Adagrad, SGD, and RMSprop. Adam, SGD, and Adelta show promising results. It is observed that the proposed model has shown 96% of accuracy which is comparatively good with other state-of-the-art models.

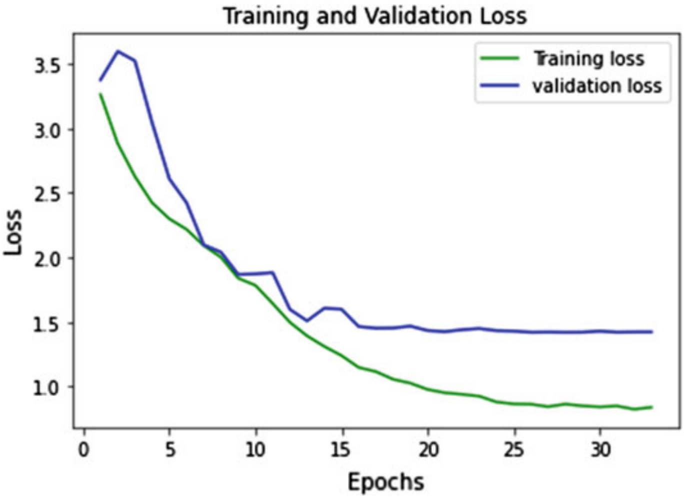

Fig. 3 Training and validation loss graph

5 Conclusion and Future Work

The field of bird calls identification utilizing AI strategies has seen a consistent expansion in recent years, with most works focusing on training different neural networks from the start. In this research a novel CNN model has been proposed by utilizing a visual portrayal of sound; in this case, spectrograms. In this thesis work, the objective of developing a CNN model for the bird calls identification has been achieved, and test it, using a dataset acquired from Kaggle BirdCLEF 2021 with different recording circumstances. Furthermore, a signal pre-processing system has been developed that permits fetching spectrograms, and also allows fetching measurements from the training model. In retrieving the specified bird species, the approach suggested in this thesis shows good results. There are, however, some limits to be aware of. First, the birdcall identification model is evaluated on a set of unique datasets obtained during the Kaggle BirdCLEF-2021 competition. The proposed model must train numerous parameters of the machine learning algorithms utilized to adopt this strategy to different datasets. This means that using this method in different recordings may necessitate more parameter adjustment research. Second, the generated characteristics were destined to be general for the designated classes (27 bird classes), covering previously discovered common birdcall structures. The general feature is insufficiently wide to encompass all of the bird species recorded in the audio records. There are over 390 different bird species in the universe. Due to a lack of annotation data, the model is only evaluated on 27 species in this thesis. The implications for other species must be investigated.

The research's shortcomings, as stated in the last subsection, inspire future research. To improve performance, the developed model might be modified and optimized further. At long last, notice some methods of performance improvement that have been accepted could be a continuation of this research work. More bird species will be studied in the future, and further animal classes might be explored as well. To evaluate the re-productiveness of the constructed procedures, additional tests on diverse datasets must be investigated. Other automatic birdsong analyses, like categorization, can benefit from the datasets utilized in the study. As a result, it is worthwhile to make the datasets available to public researchers. Concerning the signal preprocessing part: it has been seen that the dataset is altogether different among others available publically, only one dataset has been used in this work to observe performance. A potential improvement is to upgrade them for all other datasets as well. The point is to get a system equipped for extricating cleaner bird sound features for an assortment of non-bird or background sounds in the input data. Procedures, for example, commotion decrease, for instance, Wiener filtering, or data enhancement and correlation normalization. As to the training part: a basic model with a few layers and improvement of certain boundaries has been used. A possible improvement is to change the design of this model, mixing different layers and playing with different configurations. Experiments with different analyzers fluctuate their hyperparameters since they would be set just the most basic ones. Some important improvements can be made by using data augmentation and transfer learning techniques. It is assumed that these techniques would be very beneficial in terms of increasing the proposed model accuracy and efficiency.

References

1. A. Joly, et al., Lifeclef 2017 lab overview: multimedia species identification challenges. in International Conference of the Cross-Language Evaluation Forum for European Languages (Springer, 2017)
2. S. Kahl, et al., Large-Scale Bird Sound Classification using Convolutional Neural Networks, in CLEF (working notes) (2017)
3. K.J. Piczak, Recognizing Bird Species in Audio Recordings using Deep Convolutional Neural Networks. in CLEF (working notes) (2016)
4. E. Sprengel et al., Audio based bird species identification using deep learning techniques (2016)
5. D. Stowell, M.D. Plumbley, Audio-only bird classification using unsupervised feature learning (2014)
6. M. Lasseck, Improved Automatic Bird Identification through Decision Tree based Feature Selection and Bagging. CLEF (working notes), **1391** (2015)
7. E. Cakir, et al. Convolutional recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)* (IEEE, 2017)
8. P. Jančovic, et al., Bird species recognition using HMM-based unsupervised modelling of individual syllables with incorporated duration modelling. in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2016)
9. L. Solé Franquesa, Birds Sound Detection Using Convolutional Neural Networks. Universitat Politècnica de Catalunya (2019)
10. D. Stowell et al., Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods Ecol. Evol. **10**(3), 368–380 (2019)
11. M. Sankupellay, D. Konovalov, Bird call recognition using deep convolutional neural network, ResNet-50. in Proc. ACOUSTICS (2018)
12. M. Lasseck, Acoustic bird detection with deep convolutional neural networks. in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018) (2018)
13. J. Xie et al., Investigation of different CNN-based models for improved bird sound classification. IEEE Access **7**, 175353–175361 (2019)
14. A. Incze, et al., Bird sound recognition using a convolutional neural network. in 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY) (IEEE, 2018)
15. A. Joly, et al., LifeCLEF 2015: multimedia life species identification challenges. in International Conference of the Cross-Language Evaluation Forum for European Languages (Springer, 2015)