Bird Audio Diarization with Faster R-CNN

Roman Shrestha[1][0000−0002−7420−1329], Cornelius Glackin[1][0000−0001−5114−6403], Julie Wall[2][0000−0001−6714−4867], and Nigel Cannings[1]

[1] Intelligent Voice Ltd., London, UK
[2] University of East London, UK
roman.shrestha, neil.glackin, nigel.cannings@intelligentvoice.com
j.wall@uel.ac.uk

Abstract. Birds embody particular phonic and visual traits that distinguish them from 10,000 distinct bird species worldwide. Birds are also perceived to be indicators of biodiversity due to their propensity for responding to changes in their environment. An effective, automatic wildlife monitoring system based on bird bioacoustics, which can support manual classification, can be pivotal for the protection of the environment and endangered species. In modern machine learning, real-life bird audio classification is still considered as an esoteric challenge owing to the convoluted patterns present in bird song, and the complications that arise when numerous bird species are present in a common setting. Existing avian bioacoustic monitoring systems struggle when multiple bird species are present in an audio segment. To overcome these challenges, we propose a novel Faster Region-Based Convolutional Neural Network bird audio diarization system that incorporates object detection in the spectral domain and performs diarization of 50 bird species to effectively tackle the 'which bird spoke when?' problem. Benchmark results are presented using the Bird Songs from Europe dataset achieving a Diarization Error Rate of 21.81, Jaccard Error Rate of 20.94 and F1, precision and recall values of 0.85, 0.83 and 0.87 respectively.

Keywords: Deep Neural Networks · Audio Classification · Diarization · Automatic Wildlife Monitoring.

1 Introduction

Bioacoustics, a blend of biology and acoustics, has facilitated several pioneering biodiversity monitoring systems resulting in major advances towards the conservation of species prone to extinction [1, 2]. Most of these systems are based on monitoring avian phonetics since bird songs are acknowledged to be the most prominent, reliable, and consistent indicators of biodiversity, capable of providing invaluable insights on the state of the ecology [2]. Unfortunately, tracking birds manually can be an onerous task [3].

Recent advances in machine and deep learning have made possible the automation of biodiversity monitoring systems. However, the precision of these systems has been severely undermined due to the presence of numerous bird

species vocalising in an environment, which can also be further occluded by other environmental sounds [4]. Consequently, this research aims to improve upon traditional bird audio classification approaches by adopting an object detection approach to bird audio diarization, in which objects are in the form of bird audio vocalisations in the spectral domain. This will group an input audio stream into homogeneous segments based on a bird species identity, hence revealing 'which bird sang when' along with the number of distinct bird species singing within a specified time-frame in an ecosystem [5].

This research uses the Bird Songs from Europe corpus, a subset of the Xeno-canto database containing intrinsic audio recordings of the 50 most common bird species in Europe [6]. A Faster Region-Based Convolutional Neural Network (R-CNN) model with a pre-trained ResNet50 Feature Pyramid Network (FPN) backbone was trained with spectrograms and their corresponding annotations obtained from pre-processed bird audio segments. The Faster R-CNN classifier [7] performs object detection based on features extracted to locate bird specific spectral patterns for effective bird species recognition. The rest of this paper is structured into four sections. Section 2 examines the background of this research and details existing approaches in bird audio-based wildlife monitoring, followed by the methodology, experiments and results in Section 3. Discussions are provided in Section 4 and Section 5 provides conclusions and future work.

2 Literature Review

Global concern of ecological deterioration has led to much research on automated bioacoustics monitoring. Accordingly, several annual challenges, such as Conference and Labs of the Evaluation Forum (CLEF), Detection and Classification of Acoustic Scenes and Events (DCASE), Neural Information Processing Scaled for Bioacoustics (NIPS4B), and Machine Learning for Signal Processing (MLSP) have led to the development of some ground-breaking architectures for acoustic wildlife monitoring [1, 3]. Even though, modern systems can identify the majority of the species present in a natural setting, a highly accurate automatic bird audio-based wildlife monitoring system capable of identifying all the vocalising species is still missing [3].

The majority of approaches, which contemplate passive wildlife monitoring centred on avian phonetics, share three identical pre-processing measures: a) Noise Filtering, b) Bird audio detection, and c) Feature extraction. Initially, the audio segments are filtered from environmental noise followed by bird audio detection on the filtered chunks that enables the system to identify segments with bird audio, which leads to the extraction of the features relevant for bird species recognition [1]. Early systems of passive bioacoustic monitoring used traditional speech recognition-based techniques, such as template-matching (Dynamic Time Warping) and Hidden Markov Models (HMMs), as they were the most effective audio processing systems of the time [8]. Algorithms that demonstrated success with speech recognition struggled when it came to bird species recognition, as avian phonetics are composed of complex patterns unlike those found in the

human voice [8]. Substantial approaches have been developed since this early work, among these systems employing Support Vector Machines (SVMs), Machine Learning and CNN based approaches, have demonstrated gradual progress comparatively [1, 2].

Initially, SVMs were not able to achieve much success with bird audio classification, while classifiers utilising decision trees as a base demonstrated better results [9]. Later, it was discovered that SVMs based on syllable segmentation algorithms outperformed the avian phonetics classification models of that time when feature selection computed from combined Mel Frequency Cepstral Coefficients (MFCCs) [10]. The segmentation algorithm was successfully able to filter environmental noise and extract bird audio syllables through the application of a pre-emphasis filter, which focused on high frequencies that were most likely to represent avian phonetics [11]. SVMs also achieved success with multi-class bird audio classification, demonstrating an average accuracy of 98.7% while categorizing 7 distinct bird species from the Xeno-canto database [12] using a Gaussian radial basis function kernel [10].

Further work described how the traditional SVM model was extended to classify 11 species extracted from the Xeno-canto database [12] with 92.8% accuracy [13]. The approach was centred on MFCC-based feature extraction from an acoustic event-based-sifting approach combined with a Gaussian Mixture Model (GMM)-based frame selection for distinguishing specific spectral patterns from the songs of the 11 bird species [13]. In 2018 [14], a two-windows method was adapted to minimise processing time by 24% and node-level space requirements by 43% as a speed boost for the SVM classifier. This approach was evaluated on 214 wildlife recordings from the Xeno-canto database[12], based on 5 species, and achieved a maximum accuracy of 93.85% [14]. Despite this strong performance of the SVM with fewer bird species, the accuracy of the system decreased rapidly as more bird species were included in the classification [14].

In recent years, machine learning-based classifiers have exhibited major improvements for recognising bird species from audio recordings and dominated the leader boards in major competitions [2]. It was found that machine learning classifiers worked relatively well with spectrograms for bioacoustic monitoring. The second-place team for MLSP 2013 employed Extremely Randomized Trees and obtained an area under the curve (AUC) score of 95.05% while categorizing 19 bird species[11]. This work, updated with randomized decision trees [15], proved to be the winning system for the NIPS4B 2013 competition with an AUC score of 91.7% while classifying 87 bird species [15].

Artificial Neural Network-based approaches to bird audio classification began in 1997 when a neural network trained with back propagation on manually collected open source wildlife recordings was evaluated against 6 audio clips corresponding to 1 recording per species and demonstrated 82% accuracy for the task [16]. Deep Neural Networks (DNN), Recurrent Neural Networks (RNN) and CNNs have shown improvements with their ability to extract features and classify images with higher accuracy [16]. Currently, CNNs are mostly preferred for effective feature extraction for spectrogram-based approaches [17].

In BirdCLEF 2016, the winning solution incorporated a simple CNN architecture with five convolutions and one dense layer for the classification of 999 bird species, achieving an official mean Average Precision (mAP) of 0.686 and 0.555 for the foreground species and foreground species mixed with background species, respectively [18]. However, when this system was evaluated for a soundscape with arbitrary bird species singing in the background, it obtained a mAP score of 0.078 [18].

In BirdCLEF 2019 [19, 20], the Inception v3 model provided better classification results for biodiversity monitoring, possibly due to the increased number of parameters, which allowed the model to represent the mappings more accurately [19]. The classification mAP (cmAP) is the standard evaluation metric considered for this challenge [20]. The Inception v3 model winning submission [20] was trained with sophisticated data augmentation techniques, such as filtering audio chunks with random transfer functions and applying local time stretching and pitch shifting in time domain identification, along with the use of validation data for fine-tuning the pre-trained network [20]. This result surpassed state-of-the-art model performances by 20%, achieving a cmAP score of 35.6% while classifying 659 bird species from intrinsic recordings belonging to the BirdCLEF 2019 evaluation set[20].

The winning submission of BirdCLEF 2020 [21] achieved a cmAP score of 13.1% while classifying 960 species. Even though the system outperformed other competing systems, most of the species that were present in the test recordings could not be recognised [21]. In this approach, a 1D convolution/Gabor wavelet transformation first layer accepts augmented spectrograms and the remaining layers of the network were determined by performing a Neural Architecture Search (NAS).

The CNN has also performed well with a simultaneous segmentation and classification approach using a five-layer encoder-decoder model [4]. The encoder layers in the network encode high-dimensional features from the spectrograms and the decoder layers decode the encoded features and their location in the spectrogram, allowing the network to execute segmentation and classification simultaneously [4]. The network was able to predict the classes for 19 species with a True Positive Rate (TPR) of 98% on the MLSP 2013 dataset.

Existing systems and architectures are still struggling to perform highly accurate classification of bird species in the wild [21]. The concept of speaker diarization could be applied to perform diarization on bird audio to identify the birds present in an environment and recognise which bird sang when [22]. In the only diarisation-based research on bird audio, an accuracy of 53% was achieved while identifying 10 bird species from the H.J. Andrews Long-Term Experimental Research Forest (HJA) dataset [22]. The performance of the model was satisfactory compared to the standard deep learning-based bird audio classification approaches [11, 15] undertaken at that time. Research involving a bird activity detector [22], which detected segments voiced by birds followed by a change point detector, detected a change in speaker/bird turns through the application of Bayesian Information Criterion (BIC) with Agglomerative Clustering.

Our proposed Faster R-CNN model approaches bird audio diarization by performing object detection in the spectral domain and shows significant refinement to the diarization based bio-acoustic monitoring approach.

3 Methodology

3.1 Data Acquisition and Pre-Processing

The acquisition of real-life bird audio datasets with sufficient recordings per bird species extracted from a naturally occurring habitat is challenging. Several datasets such as BirdCLEF and RefSys have insufficient recording samples per bird species [23]. Thus, the balanced, medium-sized subset Bird Songs from Europe, consisting of 50 discrete European bird species with 43 high-quality natural recordings per species [6] was used in this work. Pre-processing of the raw input audio consisted of downsampling, conversion to the wav file format, segmentation, overlapping, bird audio detection, merging of audio segments, generation of spectrograms, accurate data annotation and data partitioning for training, validation, and evaluation. Firstly, the 16 kHz audio was downsampled to 8 kHz and converted from mp3 to wav. The wav files were then segmented into uniform 1-second chunks with 50% overlap as depicted in Fig. 1, which played a significant role in increasing the volume of training data.

A Pydub-based [24] bird audio detector operates as the filtering layer that processes the incoming audio stream and combines 1-second segments with bird songs from a certain species with a random 1-second segment representing a different bird species to simulate a complex natural audio recording with multiple bird species in an audio segment, see Fig. 2. Spectrograms facilitate the visualisation of the magnitude of the raw frequencies and signals in an audio chunk as a spectrum of sound over contrasting time-frames [17]. Subsequently, a Short-time Fourier Transfer (STFT)-based spectrogram of size 256×256 pixels was generated using Librosa [25].

Timings for the ground-truth labels were provided by the bird audio detection algorithm, which calculates the bird audio start and end time parameters for the corresponding chunks associated with the bird species. These parameters

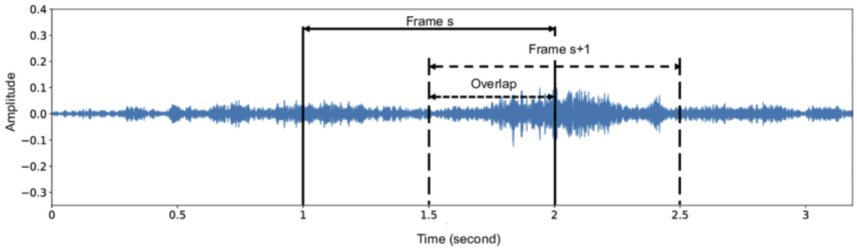

Fig. 1. 50% overlap for each 1-second audio segment

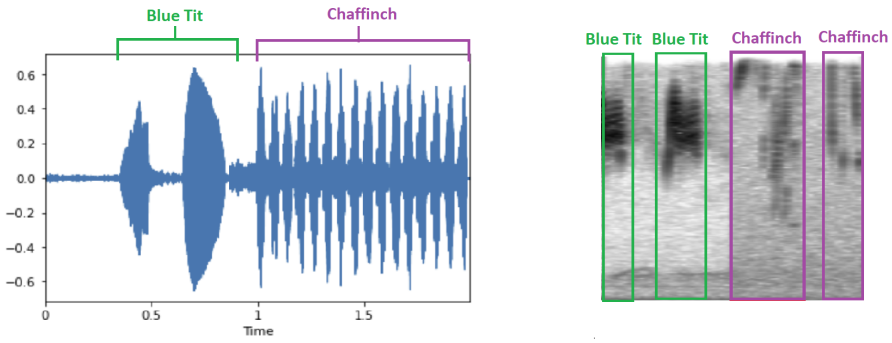

Fig. 2. New waveform (left) and spectrogram (right) generated after merging 1-second audio segments from two random species

in turn specify the start and end coordinates of the bounding boxes along the x-axis, which display 2 seconds of audio across a visual stretch of 256 × 256 pixels containing the bounding boxes, see Fig. 2. To represent a time frame of 2 seconds, 1 unit along the axis of the spectrogram should correspond to 0.128 ms as the ratio between 256 and 2,000 yields 1:0.128. Hence, the time frame can be represented in pixels by multiplying the time with 0.128 i.e. $bbox = time(ms) \times 0.128$ where *bbox* represents the corresponding bounding box coordinate for specific time represented as *time(ms)*. However, the coordinates along the y-axis for the bounding box remain as the default, i.e. the minimum value is set to '0' and the maximum value is set to '256'.

The final step in the pre-processing phase deals with accurate annotation in the Pascal Visual Object Classes (VOC) format, which stores bounding box coordinates along with essential information for object detection [7]. Fig. 2 depicts a spectrogram sample and visualises the bounding box and labels based on the Pascal VOC format. A total of 297,075 spectrograms were obtained from 91.71 hours of intrinsic audio recordings, out of which 247,479 (80%) of the data was used for training, 24,798 (10%) for validation, and the evaluation (test) set was comprised of the remaining 24,798 (10%) spectrogram images.

3.2 Model Training

A Faster R-CNN model with ResNet50 FPN backbone pre-trained on the COCO dataset with a Region Proposal Generator was used to train the model. Fig. 3 shows the Faster R-CNN object detection model's functionality pipeline. When a spectrogram is inputted, the Region Proposal Network acts as a selective search layer that generates anchor boxes for all the spectrogram regions. Based on feature maps generated by the ResNet50 architecture, the Region Proposal Layer computes the Region of Interest (ROI) proposals for specific regions in the spectrogram and selects the anchor box and segments that correspond with the ex-

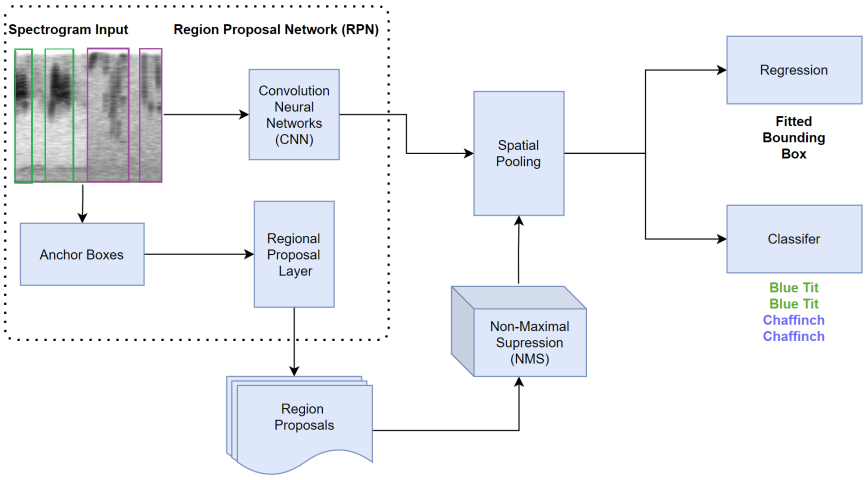

Fig. 3. Faster R-CNN Object Detection Pipeline

tracted features [7]. An Intersection over Union (IoU) score between '0' and '1' is used to compute the magnitude of intersection between the generated proposals and the ground truth labels where an IoU score closer to '1' represents a stronger intersection with the ground truth boxes for the spectrograms. Hence, the proposal regions would undergo a procedure known as Non-Maximal Suppression (NMS) that suppresses all the proposals with an IoU score less than 0.3, such that only boxes with a strong association with the ground truth would be used for training. Spatial pooling is used to select only the most important features from the feature map extracted by the FPN. Finally, bounding box co-ordinates are made more precise by performing regression and the Faster R-CNN classifier predicts the labels for the corresponding bounding boxes based on features extracted from that specific region [7].

To ensure that the model is optimally trained, the Fastai library [26] has been implemented utilising functionalities from the IceVision package [27]. To ensure optimal model training, Smith [28] suggests performing a cycle with two steps of equal length where the model is trained by cycling the learning rate (LR) between the maximum LR and the minimum LR, computed as one-tenth of the maximum LR. In the end, the LR can be reduced lower than the minimum LR i.e. to one-hundredth of the minimum LR, which has been deemed crucial for optimal model training [28].

For transfer learning, the latest version of Fastai makes use of several fit one cycle iterations to fine-tune modules with pre-trained weights more efficiently. Fine-tuning in Fastai allows the model to freeze the backbone by stopping gradient calculations and train only the head accompanied by randomly initialized parameters for the first few epochs [28]. Then, the model can be unfrozen and

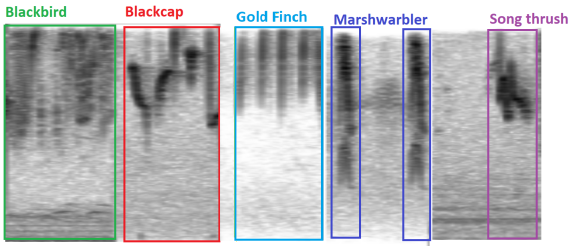

Fig. 4. Predictions generated for test set spectrogram simulating multiple bird species

trained with all the layers, allowing gradient calculations for the parameters to be adjusted until the model is optimally trained. Weights & Biases (W&B) callbacks were used for visualising and tracking the model training [29]. The weights of the model instance exhibiting minimum validation loss were saved and used for generating predictions on the unseen test set for inference. During inferencing, this trained Faster R-CNN classifier [7] was used to generate predictions for the spectrograms from the unseen test set which also contained additional 5 spectrograms obtained after combining audio-segments from more than two bird species to test if the system could cope with the presence of multiple birds in a common setting as shown in Fig. 4.

Using an NVIDIA GeForce GTX 1080 Ti GPU, the total time to train the model was 8 days and 12 hours, with an average training time of 3 hours and 13 minutes per epoch. The model was trained for a total of 60 epochs, during the first 5 epochs the backbone was frozen and only the model head was trained. This was followed by the remaining 55 epochs to train all the layers and adjust the parameters accordingly. For the first 5 epochs, the minimum validation loss achieved was 6.89 with a minimum training loss of 4.09. During the remaining 55 epochs, the validation loss of the model started gradually decreasing from 0.793 to 0.478 until the loss plateaued in the 52^{nd} epoch. Hence, the parameters of the model at the 52^{nd} epoch were saved for inferencing.

3.3 Model Testing

The trained model was evaluated with the 24,798 test set spectrograms, Fig. 5 outlines the Faster R-CNN Inference Pipeline. This model generates predictions on the test set, and a confidence score between '0' and '1' is provided for each prediction where a detection threshold of 0.5 is defined such that predictions with confidence score less than 0.5 are discarded. The predicted outputs were compared with the ground truth reference labels and Diarization Error Rate (DER), Jaccard Error Rate (JER), F1, recall and precision were calculated as evaluation metrics. A sample of predictions for the evaluation set can be seen in Fig. 6, which simulates vocalisation of multiple bird species in a single audio segment obtained by merging the audio segments from random species. The proposed Faster R-CNN model is able to perform bird audio diarization with

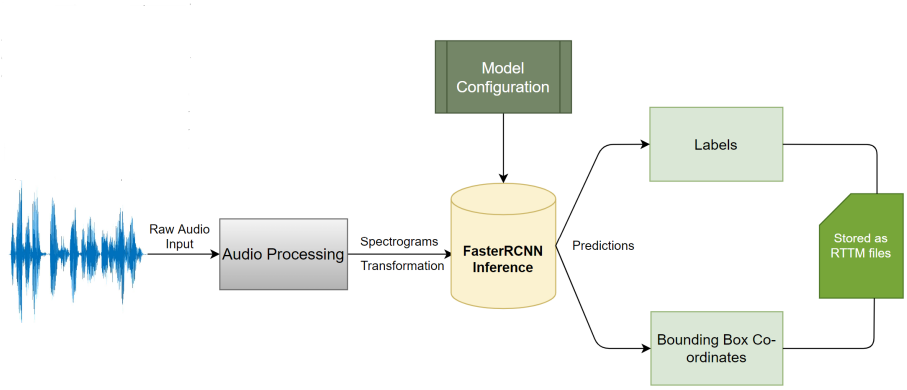

Fig. 5. Faster R-CNN Inferencing Pipeline

minimal DER and JER of 21.81 and 20.94, respectively, even under the complex circumstances simulated by combining multiple species in a single audio segment. The model achieved an F1 score of 0.85, with 0.83 precision and 0.87 recall value.

4 Discussion

From the obtained evaluation metrics it is evident that bird audio diarization implemented using the Faster R-CNN model and centered on object detection in the spectral domain is an improvement over previous diarization approaches [22]. This approach has also been shown to cope with the separation of 50 bird species from intrinsic audio recordings compared to the pioneering work with diarization on the HJA dataset that considered only 10 classes. There has been numerous research in the literature focused on bird audio classification, which have used the Xeno-canto database or one of its subsets, such as BirdCLEF, NIPS4B and DCASE. We have chosen three of these models, which have used a similar number of species, in order to compare and validate the performance of our model. Silla Jr. and Kaestner [30] approached acoustic bird species classification with 48 classes extracted from a different subset of the Xeno-canto database, using the Global Model Naive Bayes (GMNB) algorithm. This approach was able to yield an F1 score of 0.50, which outperformed other heirarchial-based classification approaches. Incze et al. [31] used a pre-trained MobileNet-based CNN architecture to classify bird species from another subset of the Xeno-canto database [12]. This approach initially showed promising results for audio classification of two bird species with an accuracy of over 80%, which reduced to below 40% when the number of classes was increased to 10. Finally, the model demonstrated an accuracy of 20% when trying to classify 50 bird species. The authors discussed the

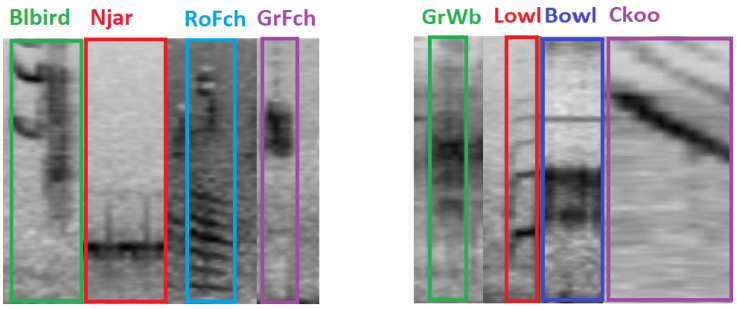

Fig. 6. Sample Prediction (Left), Blbird = Blackbird; Njar = NightJar; RoFch = RoseFinch; GrFch = GreenFinch, Sample Prediction (Right), GrWb = Great Warbler; Lowl = Little Owl; Bowl = Boreal Owl; Ckoo = Cuckoo

need for a deeper network being employed in future. It was observed that transfer learning on the pre-trained VGG16 CNN architecture achieved a bird audio classification accuracy of 73.5% on the evaluation set, on the same Bird Songs From Europe dataset consisting of 50 classes [23]. This demonstrated an improved accuracy on the existing systems for bird species classification, with this number of classes. Table 1 outlines the performance of these three approaches against the performance achieved in this work, based on an evaluation of bird species obtained from the Xeno-canto database.

Table 1 clearly shows that our proposed Faster R-CNN model outperforms standard classification approaches and has the potential to cope with the challenges associated with automated biodiversity monitoring. It was seen that segmentation of bird audio with 50% overlap plays a vital role in increasing the training data. Spectrograms generate distinct patterns based on the energies possessed by avian vocalisation and these patterns differ for every bird species. Functionalities from the Fastai library [26] support model training [28] to achieve minimal validation loss. This work, in performing object detection in the spectral domain for effective spectral pattern recognition could provide a breakthrough for biodiversity monitoring systems through diarization.

Table 1. Model Performances

Model	Number of species	Metrics
GMNB [30]	48	0.50 (F1)
MobileNet [31]	50	20% (Accuracy)
VGG16 [23]	50	73.5% (Accuracy)
Faster R-CNN	**50**	**0.85 (F1)**

5 Conclusions

A huge amount of research has been invested to build a fully functional automated non-invasive biodiversity monitoring system. However, this research area has lacked an exploration of diarization-based techniques. In this research, we approached bird audio diarization through a Faster R-CNN model, performing object detection in the spectral domain. The results achieved with this novel approach to this challenging problem show promise. It was observed that the augmentation techniques used, such as segmentation with 50% overlap, was crucial for improving the model performance by increasing the training data by 50%. The functionalities adopted from the Fastai library [26] were also extremely useful for ensuring optimal model training. The inferencing pipeline presented in this approach can be used directly with the pre-trained model weights to generate predictions in a real life-scenario.

Bird audio diarization is able to separate intrinsic avian vocalisations into separate homogeneous segments according to their species, and determine the length of their songs alongside identifying the number of species vocalising in an ecosystem [5]. We believe that this system and its spectral object detection approach can play an important role in the monitoring of population dynamics of bird species within an ecosystem. Our research demonstrates promising results for the diarization of 50 bird species from a subset of the Xeno-canto database. In future work, we aim to tackle larger and more challenging bird audio classification problems presented by challenges such as BirdCLEF, MLSP and DCASE, which would enable us to test and enhance our system further in this domain.

References

1. X. Dong and J. Jia, "Advances in automatic bird species recognition from environmental audio," J Phys Conf Ser, vol. 1544, pp. 012110, 2020.
2. S. Kahl, M. Clapp, et al., "Overview of BirdCLEF 2020: Bird sound recognition in complex acoustic environments," Conf and Labs of the Evaluation Forum (CLEF) task overview, 2020.
3. S. Kahl, T. Wilhelm-Stein, et al., "Large scale bird sound classification using Convolutional Neural Networks," Working Notes of Conf and Labs of the Evaluation Forum (CLEF), 2017.
4. R. Narasimhan, X. Fern, R. Raich, "Simultaneous segmentation and classification of bird song using CNN," IEEE Int Conf Acoust Speech Signal Process (ICASSP), 2017.
5. Z. Huang, S. Watanabe, et al., "Speaker diarization with region proposal network," IEEE Int Conf Acoust Speech Signal Process (ICASSP), 2020.
6. F. Lima, "Bird songs from Europe (Xeno-canto)," 2020. [Online]. Available: https://doi.org/10.34740/kaggle/dsv/1029985.
7. S. Ren, K. He, et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," Adv Neural Inf Process Syst (NeurIPS), 2015.
8. S. Anderson, A. Dave, D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," J Acoust Soc Am, vol. 100(2), pp. 1209-1219, 1996.

9. I. Mporas, T. Ganchev, et al., "Automated acoustic classification of bird species from real-field recordings," IEEE Int Conf on Tools with Artificial Intelligence, 2012.
10. S. Fagerlund, "Bird species recognition using support vector machines," Eur Assoc for Signal Process (EURASIP), J Advances Signal Process, vol. 1, pp. 64–64, 2007.
11. H.W. Ng, T.N.T. Nguyen, "The 9th Annual MLSP Competition: Second place," IEEE Int W on Machine Learning for Signal Process (MLSP), pp. 1-2, 2013.
12. W. Vellinga, "Xeno-canto - Bird sounds from around the world," Xeno-canto Foundation for Nature Sounds, [Online]. Available: https://doi.org/10.15468/qv0ksn
13. Z. Zhao, S. Zhang, et al., "Automated bird acoustic event detection and robust species classification," Ecological Informatics, vol. 39, pp. 99-108, 2017.
14. H. Weerasena, M. Jayawardhana, et al., "Continuous automatic bioacoustics monitoring of bird calls with local processing on node level," IEEE Region 10 Conf (TENCON), pp. 235-239, 2018.
15. M. Lassek, "Bird song classification in field recordings: Winning solution for NIPS4B 2013 competition," Neural Information Process Scaled for Bioacoustics (NIP4B): From Neurons to Big Data, pp. 176-181, 2013.
16. A.L. McIlraith, H.C. Card, "Bird song identification using artificial neural networks and statistical analysis," IEEE Canadian Conf Electrical Computer Engineering, Engineering Innovation: Voyage of Discovery, vol. 1, pp. 63-66, 1997.
17. B. Schuller, "Intelligent audio analysis," Signals Comm Technol, pp. 99-124, 2013.
18. E. Sprengel, M. Jaggi, et al., "Audio based bird species identification using deep learning techniques," Working Notes of Conf and Labs of the Evaluation Forum (CLEF), 2016.
19. C.Y. Koh, J.Y. Chang, et al., "Bird sound classification using Convolutional Neural Networks," Working Notes of Conf and Labs of the Evaluation Forum (CLEF), 2019
20. M. Lassek, "Bird species identification in soundscapes," Working Notes of Conf and Labs of the Evaluation Forum (CLEF), 2019.
21. Muhling, J. Franz, et al., "Bird species recognition via neural architecture search," Working Notes of Conf and Labs of the Evaluation Forum (CLEF), 2020.
22. C. Maina, "Audio diarization for biodiversity monitoring," IEEE Africon Int Conf Green Innovation for African Renaissance, pp. 1-5, 2015.
23. F. Lima, "Audio classification in R," poissonisfish, 2020. [Online]. Available: https://poissonisfish.com/2020/04/05/audio-classification-in-r/.
24. J. Robert, M. Webbie, et al., "Pydub," Github, 2018. [Online]. Available: http://pydub.com/
25. B. McFee, V. Lostanlen, et al., "Librosa/librosa: 0.8.0," Zenodo, 2020.
26. J. Howard, S. Gugger, "Fastai: A layered API for Deep Learning", Information, vol. 11, no. 2, p. 108, 2020.
27. L. Vazquez, F. Hassainia, "Icevision: An agnostic object detection framework," Github, 2020. [Online]. Available: https://github.com./airctic/icevision
28. L.N. Smith, "Cyclical learning rates for training neural networks," IEEE Conf on Applications of Computer Vision (WACV), pp. 464-472, 2017.
29. L. Biewald, "Experiment tracking with weights and biases," Weights and Biases, 2020. [Online]. Available: https://www.wandb.com/
30. C.N. Silla Jr., C.A.A. Kaestner, "Hierarchical classification of bird species using their audio recorded songs," IEEE Int Conf Systems, Man, and Cybernetics, 2013.
31. A. Incze, H. Jancso, et al., "Bird sound recognition using a Convolutional Neural Network," IEEE Int Symp Intelligent Systems and Informatics (SISY), 2018.