Evaluating the Effectiveness of Transfer Learning with BirdNET for the Supervised Classification and Automated Detection of Non-Bird Acoustic Recordings

Humberto Martinez

April 30, 2024

Abstract

Passive Acoustic Monitoring (PAM) is a powerful approach for monitoring vocal animals and their habitats. Spectrograms generated from the recordings of autonomous recording units (ARUs), the main instrument used in PAM, can be extracted to interpret unique acoustic data, which can help in the conservation efforts of species and ecosystems. However, the use of PAM often leads to terabytes of data that need to be processed, and this is often not feasible when relying on trained analysts and observers alone. Recently, deep learning models have been employed to manually detect and classify species based on their sound. BirdNET, a deep learning model, was developed to identify species based on their sound with high accuracy. Previous studies have used deep learning models trained on other tasks for classifying and detecting species using sound, leaving it unknown whether or not BirdNET can be trained to detect non-bird acoustic recordings. In this study, we employed BirdNET for the automated detection and supervised classification of non-bird acoustic recordings. Supervised classification uses labeled data for the classification task while unsupervised classification data does not have labels. Meerkat calls were used for the automated task and cat meows, which were split into three different classes based on stimuli performed, were used for the supervised classification task. The datasets both originated from open datasets, highlighting the feasibility of performing bioacoustics research. A previous deep learning model approach conducted by Morfi et al. (2021) was used as a baseline for comparison in this study. Overall, BirdNET outperformed the previous deep learning approach in the automated detection and classification of non-bird acoustic recordings. This study shares important real-world applications and encourages researchers to pursue opportunities in bioacoustics.

Evaluating the Effectiveness of Transfer Learning with BirdNET for the Supervised Classification and Automated Detection of Non-Bird Acoustic Recordings

Monitoring wildlife through bioacoustics—the study of animal sound—yields essential insights into animal behavior and ecosystem biodiversity, contributing to a deeper understanding of natural environments. Large amounts of acoustic data can be acquired autonomously and remotely with little human intervention (Teixeira et al., 2018). In recent years, computational analysis methods, such as deep learning, have provided researchers with new ways to analyze bioacoustic data efficiently. Bioacoustics has benefited from deep learning with the development of deep learning methods to automatically detect and classify acoustic recordings. Automated detection and classification approaches have significantly reduced the time required compared to manual bioacoustic analysis. Automated detection refers to the process of detecting sound from recordings using machine learning algorithms.

Despite recent advances in deep learning for bioacoustics research, these methods are rooted in the availability of well-annotated training data, which can be difficult to obtain. Additionally, many species have diverse acoustic repertoires and acoustic cues produced in various settings. Passive acoustic surveys frequently rely on the identification of singular, high-amplitude vocalizations to indicate a species' presence (McGinn et al., 2023). However, given the heterogeneous nature of signal types and acoustic environments, no single detection and classification algorithm demonstrates consistent performance across all signal categories and recording settings.

In this study, BirdNET, a newly developed deep convolutional neural network trained using bird vocalizations, was employed for the supervised classification of cat meows and automated detection of meerkat calls. Generally speaking, when dealing with machine learning,

unsupervised and supervised classification are the main branches used. The biggest difference in supervised classification and unsupervised classification is that supervised uses labeled training data while unsupervised classification does not. The use of BirdNET in this study underscores the importance of transfer learning—the reuse of a pre-trained machine learning model on a new dataset. BirdNET's capabilities of detecting and classifying avian vocalizations can be used to improve the classification accuracy and detection performance of non-bird vocalizations and generalize BirdNET's capabilities of detecting and classifying diverse acoustic signals. Given BirdNET's ability to learn complex patterns from data and performance at detecting and classifying bird acoustic signals, existing literature suggests that BirdNET will be effective in the automated detection and supervised classification of non-bird acoustic recordings by using the performance metrics F1, recall, and precision to evaluate its effectiveness (Kahl et al., 2021).

Passive Acoustic Monitoring

The majority of animals actively produce sound for communication, navigation, prey-seeking, and other skills needed for survival (Bradbury & Vehrencamp, 1998). Acoustic signals are very suitable for investigating the development of animal communication due to the convenient ability to capture, analyze, synthesize, and efficiently play back sounds (Laiolo, 2010; Gerhardt and Huber, 2002). In the past, skilled surveyors would identify species in the field using well-established manual auditory survey techniques, such as point counts (Gregory et al., 2004). With the emergence of new technology, passive acoustic monitoring has become an increasingly popular tool used in bioacoustics. PAM involves the process of deploying battery-operated autonomous recording units (ARUs) in a targeted area to record wildlife in their natural environment. Furthermore, ARUs are small, non-invasive, and inexpensive, and may be left in the deployment area for lengthy periods of time (Browning et al., 2017).

PAM is more efficient than manual techniques because the use of stored acoustic data allows several analysts to independently verify and authenticate the detections and classifications of animal vocalizations at any given point in time (Clink et al., 2023). PAM is also more effective for monitoring species that are situated in inaccessible areas where the visibility of the animal is low (Deichmann et al., 2018). PAM can provide data on the deployment area that may subsequently be used to estimate species occupancy, abundance, and population density, track geographical and temporal changes in animal behavior, and calculate auditory proxies for biodiversity indices (Browning et al., 2017). Clink et al. (2023) adds that manually examining spectrograms to identify the desired vocalizations has become a traditional method in PAM for retrieving vocalizations of interest, but this method requires the use of trained analysts. Additionally, PAM often generates terabytes of data, making it impractical for human observers alone. With growing worldwide interest in conservation, the use of trained analysts to retrieve vocalizations of interest is increasingly impractical.

Machine Learning Applications in Bioacoustics

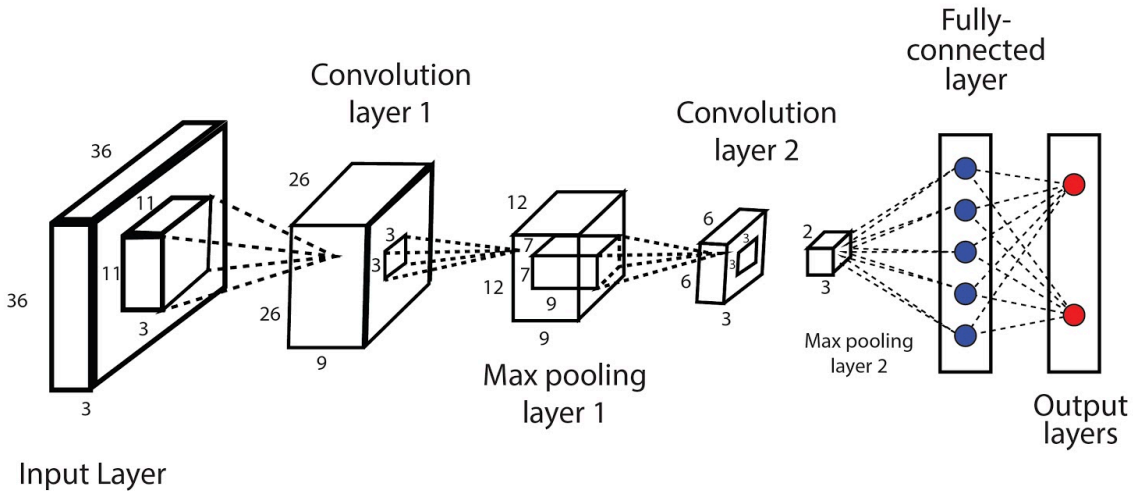
Deep learning, a branch of machine learning, denotes a class of algorithms that can automatically detect discriminative patterns within datasets of considerable complexity. (Chollet, 2018). Recently, deep learning has emerged as a powerful tool in bioacoustics with the development of automated detection and classification algorithms. Convolutional neural networks (CNNs) are commonly employed in deep learning to extract features from complex datasets. Feature extraction and selection were completed manually before deep learning was developed (LeCun et al., 2015). The organization of processing layers that perform linear and non-linear transformations in deep learning allows CNNs, the underlying technology within deep learning, to learn patterns of data across varying levels of abstraction. The lower layers of the

neural network capture low-level features such as edges, while higher layers combine these features to represent more complex classes (Yamashita et al., 2018). These convolutional layers allow for these algorithms to learn unique acoustic features from audio spectrograms to then be able to detect and classify calls. A brief explanation of this process is below.

> A convolutional layer consists of a collection of kernels, which allow the network to detect arbitrary patterns in the spectrographic image that correspond to arbitrary patterns of sound in the audio recording. When such a pattern is detected, its location within the image is recorded by activating corresponding elements in the output of the convolutional layer. Next, the max-pooling layers aggregate the activations in a convolutional layer into a smaller image, which helps make pattern detection less location-sensitive. Eventually, a visual pattern detected by a kernel anywhere in the spectrographic image can carry its activation all the way to the appropriate output of the CNN and predict the classification of the input species. (Knight et al., 2019, p. 4).

Figure 1

Schematic representation of a basic CNN (Dwivedi, 2021)



Existing Literature

In a recent study, Kahl et al. (2021) developed BirdNET to make the detection and identification of avian species from acoustic recordings more efficient. BirdNET is a convolutional neural network (CNN), more specifically a deep convolutional neural network (DCNN) since it has 157 convolutional layers, that are trained to detect and classify bird species based on their vocalizations. BirdNET is different from other DCNNs because it's trained using bird vocalizations and is specifically trained to detect and classify bird species based on their vocalizations.

Convolutional neural networks (CNNs), such as AlexNET, have been used for bioacoustics before (Colonna et al., 2016). AlexNET is trained using the ImageNet database and is designed for tasks such as image classification and object detection (Krizhevsky et al., 2017). Given that AlexNET is 8 convolutional layers deep, it may not be referred to as deep learning in the context of other intricate deep neural networks even though it employs feature extraction—the process of extracting relevant features from raw data, a key principle in deep learning. In a study conducted by Knight et al. (2019), AlexNET was used for the automated detection of 19 bird species. While this study suggested that AlexNET is effective at detecting the bird species part of their study, Knight et al. used a sheer amount of training data. The effectiveness of CNNs and deep learning models depends on the availability of well-annotated training data. The majority of training data available is often generalized to common bird species. As a result, endangered bird and non-bird species, which are often at the forefront of conservation efforts, do not benefit from the advances in deep learning due to a lack of sufficient training data (McGinn et al., 2023).

On a similar note, BirdNET is able to identify over 3000 bird species in small 3-second segments with a quantitative confidence score (Wood et al., 2022). BirdNET is trained using approximately 1.5 million bird spectrograms extracted from the training dataset with a maximum of 3500 samples per class, whereas AlexNET is trained using the ImageNET dataset (Kahl et al. 2021; Krizhevsky et al., 2017). Having a large training dataset in BirdNET can lead to higher model performance, reduce overfitting, and help generalize diverse datasets. This is particularly important considering how environmental factors vary across different ecosystems. Moreover, the confidence score in BirdNET varies from 0 to 1. Users can apply a threshold value to filter the output at their preferred confidence level (Granados, 2023). The threshold value in BirdNET is a dynamic metric rather than a static one because the threshold value fluctuates based on the quality of the training and test data. Adjusting the threshold value allows research to maximize the model's performance based on various constraints. The study conducted by Kahl et al. (2021) employs a rigorous methodology involving diverse and massive amounts of training data used for the development of BirdNET. However, similar to the study conducted by Knight et al. (2019), they both fail to consider how BirdNET and AlexNET could be applied for the automated detection and classification (supervised and unsupervised) of diverse non-bird acoustic recordings.

In a different study, Allen et al. (2021) used a ResNet-50 convolutional network for the automated detection of humpback whale songs and used a relatively small training dataset to do so (He et al., 2016). ResNET-50 proved efficient at detecting the calls of humpback whales and Allen et al. (2021) highlight the ability of a trained CNN to learn from a small dataset and detect calls with high variability among the signal types. ResNET-50 is deeper (50 convolutional layers) and more sophisticated than AlexNET, but even then, Allen et al. (2021) generalized their

study to humpback whales. On a similar note, Dufourq et al. (2022) performed a study comparing twelve modern pre-trained CNNs for the automated detection of gibbon calls. Of the twelve pre-trained CNNs used in Dufourq's study, ResNET-50 and AlexNET were used. Dufourq discovered that the CNNs could be successfully trained with as little as 25 verified calls while still delivering promising results, suggesting similar results to those of Allen et al. (2021) and Knight et al. (2019). Dufourq also shared that transfer learning approaches do not require an enormous amount of training data to be successful.

After reviewing multiple studies surrounding using pre-trained CNNs for the automated detection and classification of non-bird vocalizations, the application of BirdNET to diverse non-bird acoustic recordings still needs to be discovered. While other CNNs like ResNet-50 and AlexNET have shown promising results in detecting non-bird acoustic recordings with high variability, it's uncertain whether or not BirdNET will yield similar results. Despite numerous studies using transfer learning with pre-trained CNNs for non-bird species, BirdNET stands out because of its distinct training dataset using bird vocalizations (Kahl et al., 2021; He et al., 2016; Krizhevsky et al., 2017). Thus leaving the question, "how effective is transfer learning with BirdNET in the classification and automated detection of non-bird acoustic recordings?" Open datasets of cat meows and meerkat calls were chosen to assess this question because of the diverse nature of the datasets.

Method

Two different open datasets consisting of meerkat calls and cat meows were used to evaluate BirdNET's performance in the supervised classification and automated detection of non-bird acoustic recordings. Both open datasets used PAM devices for capturing the acoustic signals, however, the duration and start time for each recording differed for each dataset. The use

of open datasets in this study offers the opportunity to easily use diverse acoustic recordings while also eliminating the need to spend resources on data acquisition. Additionally, using open datasets allows others to access the same datasets easily to replicate the analysis or conduct additional investigations.

The performance metrics F1, precision, and recall were then used to evaluate BirdNET's performance in the classification and detection task. An F1 score is the harmonic mean of precision and recall. Precision shows how often a machine learning model is correct when predicting the target class. Recall shows whether a machine learning model can find all objects of the target class (Sharma, 2023). The F1 score is measured from 0 to 1, and the closer it is to 1, the more accurate the algorithm is. A high F1 score is essential because it reflects a good balance between precision and recall. This is particularly important when dealing with imbalanced datasets or when both false positives and false negatives are critical.

Data Collection

In 2020, Ludovico et al. published an open dataset of cat vocalizations, CatMeows. The CatMeows dataset contains 440 sounds of the meows from 21 cats (10 Maine Coon and 11 European Shorthair) that were repeatedly exposed to three different stimuli. Each stimulus denotes a specific class of cat meows.

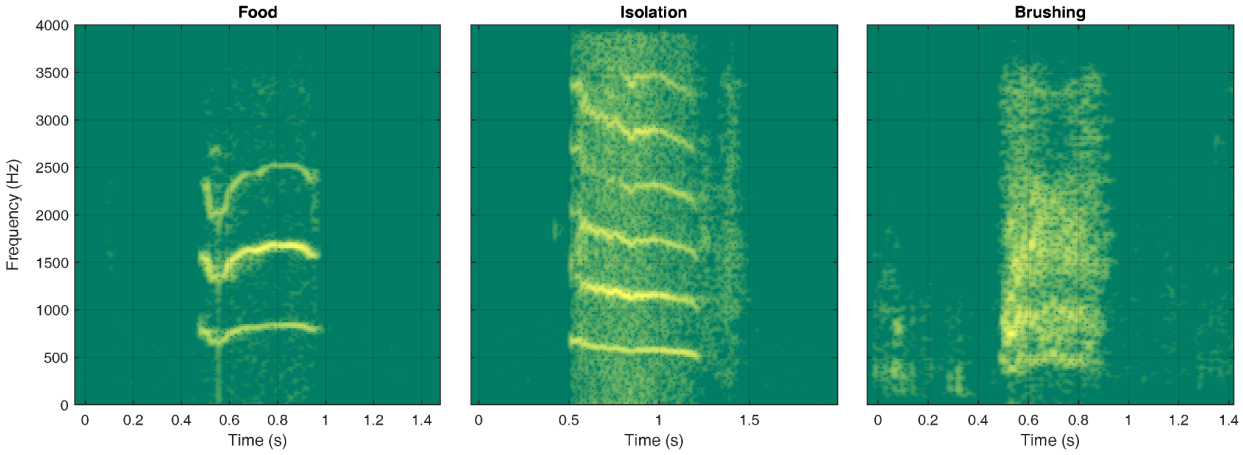
Table 1

Details for each stimulus the cats were exposed to (Ludovico et al., 2020).

Brushing (Class B)	**Isolation in an unfamiliar environment (Class I)**	**Waiting for food (Class F)**
Their owners brushed cats for a maximum of 5 minutes.	The owners transferred the cats into an unfamiliar environment for 5 minutes.	The owner began familiar pre-food routines, delivering food within 5 minutes.

Figure 2

Time-frequency spectrograms of meows for each stimulus (Ntalampiras et al., 2019).



To minimize the discomfort of the cats in Class I, the distance to the unfamiliar environment was kept at less than 30 minutes. The cats in Class I were also allowed to be with their owner for 30 minutes after they arrived to recover from the transportation. The cat meows were recorded using the QCY Q26 Pro Mini Wireless Bluetooth Music Headset which had a microphone dynamic range of 98 ± 3 dB (Ntalampiras et al., 2019). The recording device was placed on the cat's collar and pointed upwards. The placement of the recording device was consistent with all cats which minimized any changes in sound and pitch intensity by movement.

The CatMeows dataset was best suited for the supervised classification task because the different classes were already identified in the dataset. Since cat meows are widely recognizable, this ensures that the dataset contains easily identifiable sound files which helps facilitate the training and evaluation of the BirdNET transfer learning classification model.

On a different note, in the 2023 Detection and Classification of Acoustic Scenes and Events (DCASE) challenge, Morfi et al. (2021) performed a study that focused on sound event detection in a few-shot learning setting for animal vocalizations. The main goal of the study was to train prototypical (deep learning) networks with only a few examples of training data for

detecting different animal vocalizations in recordings. The study by Morfi et al. was used as a baseline for comparison in this study to see if BirdNET transfer learning outperformed deep learning approaches.

Table 2

Baseline results from Morfi et al. (2021).

System	F-measure	Precision	Recall
Template Matching	0.04	0.02	0.18
Prototypical Network	0.30	0.36	0.25

The development dataset used in the study by Morfi et al. was split into training and validation datasets. The training dataset consisted of 4 different sub-folders which included BV (BirdVox-DCASE-10h), HT (spotted hyenas), JD (jackdaws), MT (meerkat), and WMW (Western Mediterranean Wetlands Bird) datasets. For the automated detection task, the MT (meerkat) dataset from the training dataset was the best because the cats and meerkats are both terrestrial mammals. Moreover, the high sampling rate and high event-to-duration ratio of the MT dataset makes it suitable for a transfer learning approach. A high sampling rate ensures sufficient temporal resolution for capturing unique acoustic information and a high event-to-duration ratio suggests more target events occurring at a higher frequency, which can aid the automated detection task in learning and detecting patterns. Details for the MT dataset are shown below.

Table 3

Statistics of MT (meerkat) sub-folder in the training dataset (Morfi et al., 2021).

Statistics	Values
Number of audio recordings	2
Total duration	1 hour and 10 mins
Total classes (excl. UNK)	4
Total events (excl. UNK)	1294
Ratio event/duration	0.04
Sampling rate	8,000 Hz

The meerkat vocalization data were recorded at the Kalahari Meerkat Project in the Kuruman River Reserve, South Africa using TS Market Edic Mini Tiny+ A77 audio devices with a Frequency band of 100 Hz - 10 kHz. The recorders were placed in collars which also included GPS monitors to monitor the meerkat movements. Recordings were carried out during the daytime. Additionally, a tiny percentage of unclear or unknown labels designated as UNK in the supplied annotation files are present in real-world datasets. To avoid penalizing algorithms that outperform a human annotator, these assessment metrics handle them independently during evaluation (Morfi et al., 2021).

Supervised Classification with BirdNET using Cat Meows

The cat meow recordings were provided in PCM streams (.wav) and stored in a directory structure. The directory included filenames following a specific convention, C_NNNNN_BB_SS_OOOOO_RXX, indicating the category of each recording. An explanation of the naming convention is listed below (Ludovico et al., CITE YR).

1. C = emission context (values: B = brushing; F = waiting for food; I: isolation in an unfamiliar environment)
2. NNNNN = cat's unique ID
3. BB = breed (values: MC = Maine Coon; EU: European Shorthair)
4. SS = sex (values: FI = female, intact; FN: female, neutered; MI: male, intact; MN: male, neutered)
5. OOOOO = cat owner's unique ID
6. R = recording session (values: 1, 2 or 3)
7. XX = vocalization counter (values: 01..99)

The programming language R was used to preprocess the dataset for the supervised classification task. The 'list.files' function was utilized to obtain the file paths of the audio recordings 'TempWavs' and their associated categories. Categories were then extracted from the filenames using string manipulation techniques. Full file paths for the audio recordings 'TempWavsLong' were obtained. An output directory 'OutputDir' was specified to organize the processed data.

A specified number of randomizations ('N.randomization') were performed to create multiple splits for evaluation. The dataset was randomly split into 70% training ('TrainWavs') and 30% testing sets ('TestWavs'), ensuring a balanced representation of categories in each set. The training set was used to train BirdNET and then the trained BirdNET model was run over 5 iterations. Each iteration denoted a different experiment in the context of this study (e.g. iteration 1 = experiment 1). The 70/30 split and 5 iterations are an effective way to evaluate model performance for supervised classification.

The training and testing sets were then organized into separate directories, ‘TrainOutPut’ and ‘TestOutPut’ respectively, for each category and randomization. The cat meow audio recordings were copied from their original locations to the corresponding train and test directories. This process ensured category-wise organization.

Automated Detection with BirdNET using Meerkat Calls

Annotations of the meerkat calls were provided in CSV format which detailed the start and end times of each meerkat vocalization. The recordings and annotations were processed using the R programming language with the ‘tuneR’ and ‘seewave’ packages.

A custom script was developed to segment the audio recordings based on the provided annotations. For each annotation in the CSV file, the script extracted the start and end times of the vocalization event. Using the ‘cutw’ function from the ‘seewave’ package, the corresponding segment of the audio waveform was extracted. The vocalization events were categorized into their respective classes (e.g., positive vocalizations) based on the annotation column names. Segmented audio segments were saved into separate directories according to their classes for further processing. The segmented audio segments, along with their corresponding annotations, were organized into a structured directory format suitable for training BirdNET. Each segmented audio segment was saved as a separate WAV file within directories named after their respective classes. The trained BirdNET model was then run over a 43-min meerkat vocalization file for detection.

Data Analysis Methods

In order to retrieve the performance metrics of the supervised classification task, a confusion matrix was calculated using the ‘caret’ package in R. The confusion matrix calculates the performance metrics and shows how many predictions are correct and incorrect per class

(Pandy et al., 2022). The performance metrics were then aggregated into a structured data frame, allowing easy representation of BirdNET's performance at classifying cat meows across the different experiments and classes.

To calculate the performance metrics of the automated detection task, the 'ohun' R package was employed. 'ohun' is a tool designed for evaluating the performance of bioacoustic detection algorithms (Araya-Salas et al., 2023). 'ohun' was able to calculate the F1 score as well as precision and recall of BirdNET for the automated detection task. One crucial aspect of using the 'ohun' package is the specification of an overlap parameter. However, determining the optimal overlap parameter was difficult because of the differences in segment lengths between BirdNET's 3-second output clips and the variable duration of meerkat vocalization clips. The wide array of performance metrics tailored for bioacoustic applications in 'ohun' make the package the most ideal for evaluating the performance of the BirdNET model at the detection of meerkat calls.

Results

In the context of this study, the closer the F1 score is to 1, the more effective BirdNET is in detecting or classifying calls. The table and graph below show the results of the supervised classification task.

Table 4

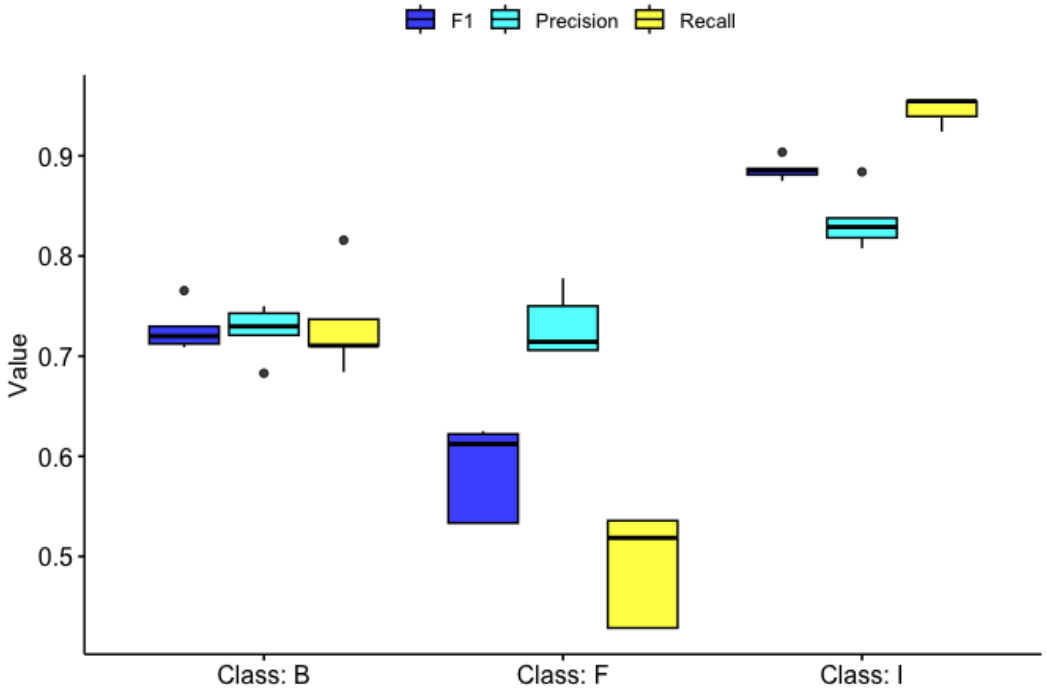
Classification Accuracy of BirdNET

Experiment	**Class**	**F1**	**Precision**	**Recall**
Experiment 1	Class: B	0.71	0.74	0.68
Experiment 1	Class: F	0.61	0.71	0.54
Experiment 1	Class: I	0.89	0.83	0.95
Experiment 2	Class: B	0.71	0.68	0.74

Experiment 2	Class: F	0.53	0.71	0.43
Experiment 2	Class: I	0.89	0.84	0.94
Experiment 3	Class: B	0.72	0.73	0.71
Experiment 3	Class: F	0.53	0.71	0.43
Experiment 3	Class: I	0.88	0.81	0.95
Experiment 4	Class: B	0.73	0.75	0.71
Experiment 4	Class: F	0.62	0.78	0.52
Experiment 4	Class: I	0.88	0.82	0.95
Experiment 5	Class: B	0.77	0.72	0.82
Experiment 5	Class: F	0.63	0.75	0.54
Experiment 5	Class: I	0.90	0.88	0.92

Figure 3

Graph showing BirdNET's performance with the supervised classification task with F1, Precision, and Recall values on the y-axis and the different classes on the x-axis.



As shown in Table 4, BirdNET was most effective in classifying cat meows in Class I with an achieved maximum F1 score of 0.90 in Experiment 5, a precision score of 0.88, and a recall score of 0.92. The F1 score of 0.90 demonstrates that BirdNET can classify cat meows with high accuracy through the use of transfer learning. BirdNET achieved a minimum F1 score of 0.53 with a precision score of 0.71 and a recall score of 0.43 in classifying cat meows in Class F during Experiment 3. BirdNET was most effective in classifying cat meows belonging to Class I, but demonstrated average accuracy in classifying cat meows belonging to Class B.

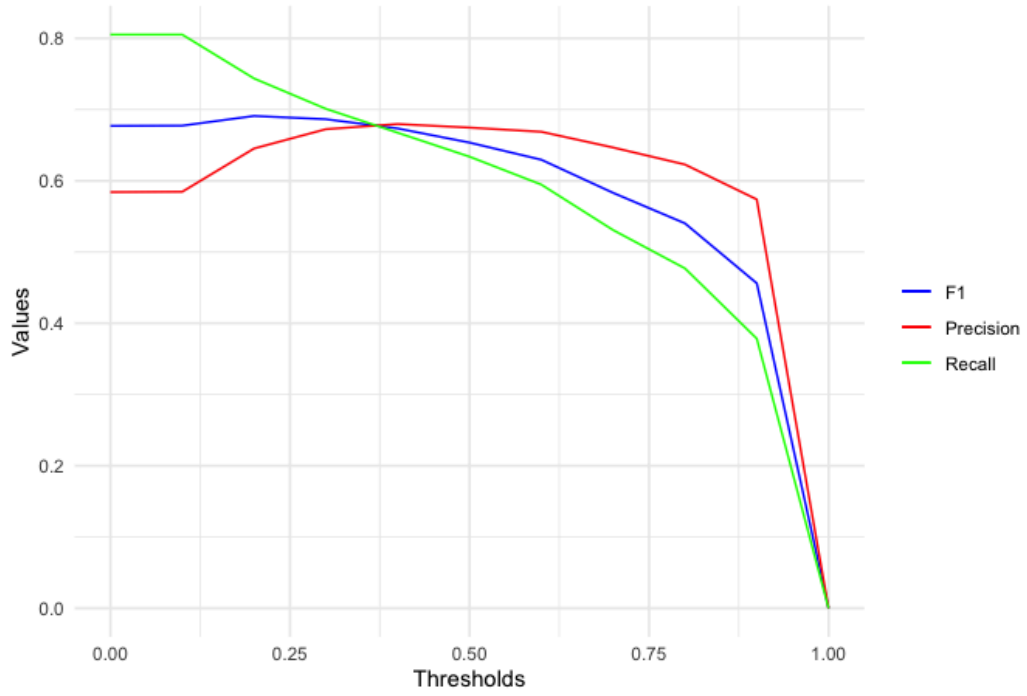
Table 5

Detection Accuracy of BirdNET

F1	Precision	Recall	Threshold
0.68	0.58	0.81	0
0.68	0.58	0.81	0.1
0.69	0.65	0.74	0.2
0.69	0.67	0.7	0.3
0.67	0.68	0.67	0.4
0.65	0.67	0.63	0.5
0.63	0.67	0.59	0.6
0.58	0.65	0.53	0.7
0.54	0.62	0.48	0.8
0.46	0.57	0.38	0.9
0	0	0	1

Figure 4

Graph showing BirdNET's performance on the automated detection task with F1, Precision, and Recall on the y-axis and the corresponding threshold values on the x-axis.



The threshold value is a dynamic metric rather than a static one because the threshold value fluctuates based on the quality of the training and test data, as previously mentioned. Adjusting the threshold value allows research to maximize the model's performance based on various constraints. Here, BirdNET achieved a maximum F1 score of 0.69 at threshold values of 0.2 and 0.3 at detecting meerkat vocalizations. The threshold value of 0.3 had a higher precision score compared to the precision score of threshold value 0.2. Similarly, the threshold value of 0.2 had a higher recall score compared to the precision score of threshold value 0.3. BirdNET yielded high F1, precision, and recall scores up to the threshold value of 0.3, but the performance metrics decreased dramatically after the 0.3 threshold value suggesting that the model became less effective at detecting meerkat vocalizations.

That being said, automated detection and supervised classification through transfer learning with BirdNET still outperformed Morfi's study which was used as a baseline for comparison in this study. BirdNET outperformed the traditional template matching technique, an

F1 score of 0.04, and the prototypical (deep learning) network, F1 score of 0.30, suggesting that automated detection and classification through transfer learning with BirdNET can be more effective than previous methods.

Discussion

BirdNET had a moderate performance in classifying cat meows in Class F, which could be explained for a few reasons. To begin with, the cat meows across all classes were very similar. The meows in Class F may have lacked unique acoustic features and it may have caused BirdNET to misclassify the meows in Class F more often than the other classes. Additionally, the meows in Class F could've had an imbalance distribution in samples, leading to biased training of the BirdNET. The limited exposure of Class F during training could explain the lower performance in classification.

The performance of BirdNET during the automated detection tasks highlights the importance of selecting an appropriate threshold value that fits the dataset for that specific task. BirdNET performed very well up until the threshold value of 0.3, where it then quickly lost its accuracy. This can suggest that BirdNET is effective at detecting meerkat calls at a relatively low confidence level, but struggles to detect meerkat calls at high confidence levels.

Research Goals

Overall, this research study attempts to answer the question, "how effective is transfer learning with BirdNET in the classification and automated detection of non-bird acoustic recordings?" The question rose to the topic due to the gap in the body of knowledge between BirdNET and transfer learning approaches in passive acoustic monitoring. BirdNET was developed to detect and classify bird species based on their vocalizations. Since BirdNET is trained on bird vocalizations, this limits what species it can detect and classify. Endangered

species or species that need to be researched may not benefit from advances like BirdNET if the model's training data is based on specific species. Dufourq et al. (2022) shared how they successfully used other machine learning models to detect gibbon calls through transfer learning with good accuracy, which set up the basis of this study. BirdNET was used in the supervised classification and automated detection of non-bird acoustic recordings to see whether or not transfer learning through BirdNET can outperform previous automated detection and classification tasks. The study by Morfi et al. was used as a baseline for comparison in this study because Morfi already compared previous deep learning and template matching models, offering the opportunity to see if BirdNET can outperform the methods that were already used before.

New Understandings

Automated detection and supervised classification through transfer learning with BirdNET is effective, as demonstrated by the high F1 scores in the detection and classification tasks, and outperformed previous deep learning and template matching techniques. BirdNET was most accurate at classifying cat meows in Class I and had the highest accuracy when detecting meerkat calls at threshold values of 0.2 and 0.3.

Leveraging BirdNET on detection and classification tasks of non-bird acoustic recordings can prove to be effective. The ability to detect and classify bird and non-bird species through BirdNET with high accuracy can lead to the encouragement of continuing to use PAM to monitor wildlife. The workflow of manually classifying species sounds can become more efficient through the implementation of machine learning algorithms, like BirdNET. Moreover, using efficient techniques to extract ecological data from vocalizations can lead to a greater understanding of different species and their populations.

Explanation of New Understandings

BirdNET most likely outperformed previous deep learning and template matching techniques in detecting and classifying non-bird acoustic recordings due to BirdNET's sheer amount of convolutional layers and vast training dataset. Thomas W. Malone, the founding director of the MIT Center for Collective Intelligence, shared that "The more layers you have, the more potential you have for doing complex things well" (Brown, 2021). Although BirdNET has many convolutional layers, it still manages to prevent overfitting, meaning that the model can handle vast amounts of training data and learn from it while still being effective at detecting and classifying calls in test datasets. With 157 convolutional layers, BirdNET can handle complex datasets and still obtain high accuracy in detecting and classifying calls. All the convolutional layers allow BirdNET to learn intricate and unique features from acoustic datasets and learn from them for detection and classification tasks. BirdNET did not need massive amounts of training data to detect and classify non-bird acoustic recordings, suggesting that BirdNET can be effective in detecting and classifying other non-bird acoustic recording datasets with little training data to do so.

Limitations and Implications

One limitation of this study is that the datasets that were used contained relatively noisy datasets which could've impacted the model performance. Similarly, if BirdNET had been trained on more acoustic events in the meerkat and cat meow datasets, there could have been an increase in accuracy in detection and classification. It's important to consider that greater amounts of training data can lead to more overfitting of the model, affecting the model's performance. Additionally, it cannot be fully concluded on whether or not BirdNET will yield

similar results for non-terrestrial species like dolphins or whales. The meerkat calls and cat meows datasets are of terrestrial species and both had similar frequencies.

Future studies should continue to evaluate whether or not complex machine learning models, like BirdNET, are effective at being trained and tested on different tasks. It's important for researchers to investigate if BirdNET can yield accurate results for other non-bird species. BirdNET's moderate performance in Class F highlights the importance of refining and improving detection and classification algorithms. Future research needs to include how to improve the detection and classification of diverse datasets.

Conclusion

The introduction of the BirdNET guided user interface (GUI) and BirdNET app can lead to the encouragement of using PAM to monitor wildlife (Wood et al., 2022). Through these recent developments, non-coders can easily use complex machine learning algorithms to automate detection and classification tasks in PAM research.

This study's findings can help facilitate the investigation of BirdNET for detecting and classifying non-bird acoustic recordings through transfer learning. The use of open datasets in this study demonstrates how bioacoustics and machine learning research can be inexpensive while sharing important real-world applications and encouraging researchers to pursue opportunities in this field.

References

Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., Cattiau, J., & Oleson, E. M. (2021). A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset. *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.607321

Browning, E., Gibb, R., Glover-Kapfer, P., & Jones, K. E. (2017). *Passive acoustic monitoring in ecology and conservation*.

Clink, D. J., Kier, I., Ahmad, A. H., & Klinck, H. (2023). A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution*, *11*. https://doi.org/10.3389/fevo.2023.1071640

Colonna, J., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., & Gama, J. (2016). Automatic Classification of Anuran Sounds Using Convolutional Neural Networks. *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering - C3S2E '16*. https://doi.org/10.1145/2948992.2949016

Deichmann, J. L., Acevedo‐Charry, O., Barclay, L., Burivalova, Z., Campos‐Cerqueira, M., d'Horta, F., Game, E. T., Gottesman, B. L., Hart, P. J., Kalan, A. K., Linke, S., Nascimento, L. D., Pijanowski, B., Staaterman, E., & Mitchell Aide, T. (2018). It's time to listen: there is much to be learned from the sounds of tropical ecosystems. *Biotropica*, *50*(5), 713–718. https://doi.org/10.1111/btp.12593

Dufourq, E., Batist, C., Foquet, R., & Durbach, I. (2022). Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics*, *70*, 101688. https://doi.org/10.1016/j.ecoinf.2022.101688

Dwivedi, R. (2021). *5 Common Architectures in Convolution Neural Networks (CNN) | Analytics Steps*. Www.analyticssteps.com. https://www.analyticssteps.com/blogs/common-architectures-convolution-neural-networks

François Chollet. (2021). *Deep Learning with Python, Second Edition*. Shelter Island, Ny Manning Publications.

Gibb, R., Browning, E., Glover‐Kapfer, P., & Jones, K. E. (2018). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, *10*(2), 169–185. https://doi.org/10.1111/2041-210x.13101

H Carl Gerhardt, & Franz Huber. (2002). *Acoustic communication in insects and anurans : common problems and diverse solutions*. University Of Chicago Press, Cop.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. Openaccess.thecvf.com. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236

Knight, E. C., Poo Hernandez, S., Bayne, E. M., Bulitko, V., & Tucker, B. V. (2019). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, 1–19. https://doi.org/10.1080/09524622.2019.1606734

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Laiolo, P. (2010). The emerging significance of bioacoustics in animal species conservation. *Biological Conservation*, *143*(7), 1635–1645. https://doi.org/10.1016/j.biocon.2010.03.025

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Ludovico, L. A., Ntalampiras, S., Presti, G., Cannas, S., Battini, M., & Mattiello, S. (2021). CatMeows: A Publicly-Available Dataset of Cat Vocalizations. *MultiMedia Modeling*, 230–243. https://doi.org/10.1007/978-3-030-67835-7_20

McGinn, K., Kahl, S., Peery, M. Z., Klinck, H., & Wood, C. M. (2023). Feature embeddings from the BirdNET algorithm provide insights into avian ecology. *Ecological Informatics*, *74*, 101995. https://doi.org/10.1016/j.ecoinf.2023.101995

Morfi, V., Nolasco, I., Lostanlen, V., Singh, S., Strandburg-Peshkin, A., Gill, L., Pamuła, H., Benvent, D., & Stowell, D. (2021). *Detection and Classification of Acoustic Scenes and Events*. https://dcase.community/documents/workshop2021/proceedings/DCASE2021Workshop_Morfi_52.pdf

Ntalampiras, L. A., Ludovico, S., Presti, G., Prato Previde, E. P., Battini, M., Cannas, S., Palestrini, C., & Mattiello, S. (2019). Automatic Classification of Cat Vocalizations Emitted in Different Contexts. *Animals*, *9*(8), 543. https://doi.org/10.3390/ani9080543

Pandey, R., Khatri, S. K., Singh, N. K., & Verma, P. (Eds.). (2022). *Artificial Intelligence and Machine Learning for EDGE Computing*. https://doi.org/10.1016/c2020-0-01569-0

Pérez‐Granados, C. (2023). BirdNET: applications, performance, pitfalls and future opportunities. *Ibis*. https://doi.org/10.1111/ibi.13193

Sharma, N. (2023, June 6). *Understanding and Applying F1 Score: A Deep Dive with Hands-On Coding*. Arize AI. https://arize.com/blog-course/f1-score/#:~:text=F1%20score%20is%20a%20measure

Sutherland, W. J., Newton, I., & Green, R. E. (2004). *Bird Ecology and Conservation*. https://doi.org/10.1093/acprof:oso/9780198520863.001.0001

Teixeira, D., Maron, M., & Rensburg, B. J. (2019). Bioacoustic monitoring of animal vocal behavior for conservation. *Conservation Science and Practice*, *1*(8). https://doi.org/10.1111/csp2.72

Wood, C. M., Kahl, S., Chaon, P., Peery, M. Z., & Klinck, H. (2021). Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. *Methods in Ecology and Evolution*, *12*(5), 885–896. https://doi.org/10.1111/2041-210x.13571

Wood, T. (2019, May 17). *F-Score*. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/f-score

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional Neural networks: an Overview and Application in Radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9