L'INSTITUT agro Rennes Angers

COCHRANE ECOLOGICAL INSTITUTE CEI COCHRANE WILDLIFE RESERVE

Cochrane Ecological Institute – Cochrane Research Institute

Creation of a sound analysis algorithm capable of identifying bird species from the CEI

September – December 2024

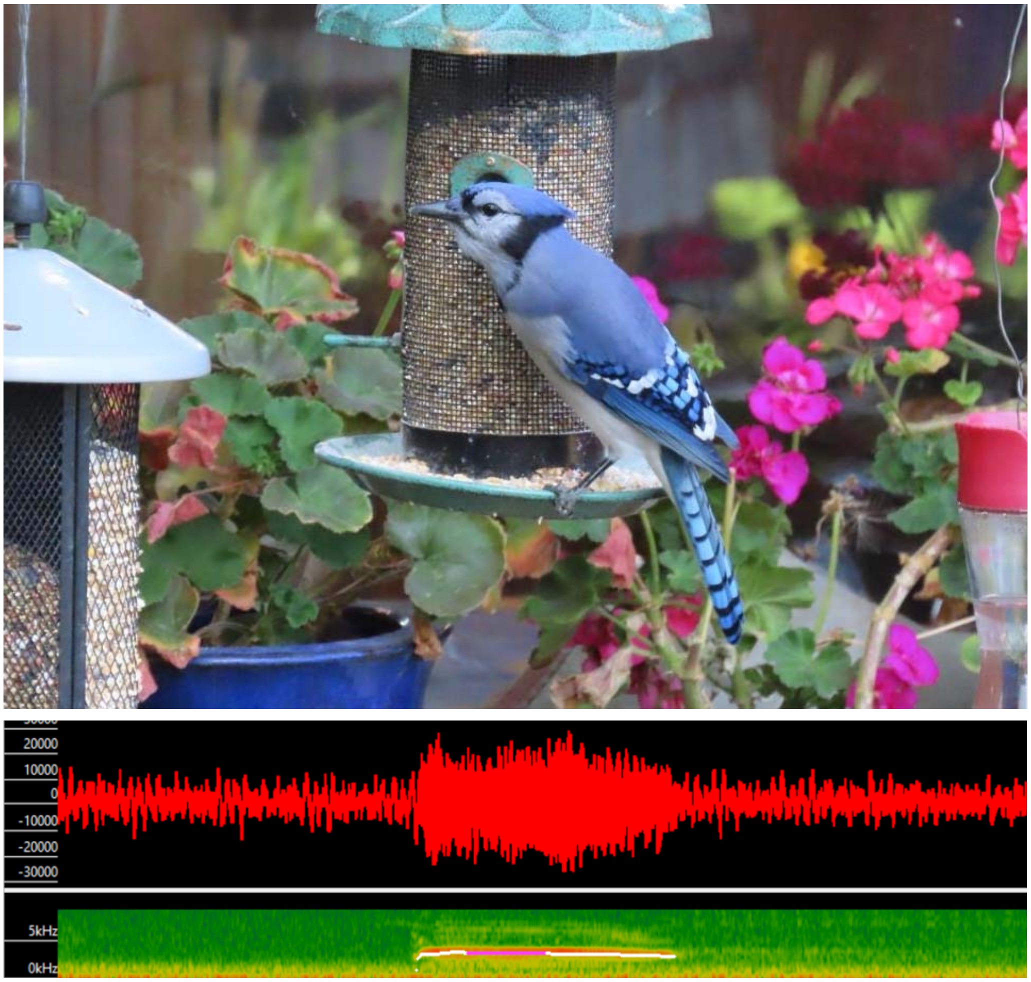


Figure 0:Photo of a Blue Jay and an excerpt from Kaleidoscope Pro for the Blue Jay's call

By Marine CAMUS – M1 Agronomy

Work placement mentor:Kenneth WEAGLE

I. Abstract:

This report presents a study conducted within the Cochrane Ecological Institute (CEI) reserve, aimed at developing acoustic analysis algorithms capable of identifying bird species from audio recordings. The context highlights the importance of monitoring bird populations as sensitive environmental indicators. Three main approaches were explored: the extraction of Mel-Frequency Cepstral Coefficients (MFCC), Short-Time Fourier Transform (STFT), and log-Mel spectrograms combined with the EfficientNetV2 model. These algorithms were trained on a dataset of 20 bird species present within the reserve. Results show that the MFCC-based model achieved the highest overall accuracy (85%), while the EfficientNetV2 model stood out for its robustness in handling complex and imbalanced data. Class imbalance management, data augmentation, and challenges related to multi-species detection are discussed. This work opens up new perspectives for automated bird song recognition in the context of biodiversity conservation.

Résumé:

Ce rapport présente une étude menée au sein de la réserve du Cochrane Ecological Institute (CEI), visant à développer des algorithmes d'analyse acoustique capables d'identifier les espèces d'oiseaux à partir d'enregistrements sonores. Le contexte souligne l'importance du suivi des populations d'oiseaux en tant qu'indicateurs environnementaux sensibles. Trois approches principales ont été explorées : l'extraction des coefficients cepstraux de fréquences Mel (MFCC), la transformation de Fourier à court terme (STFT) et les spectrogrammes log-Mel combinés au modèle EfficientNetV2. Ces algorithmes ont été entraînés sur un ensemble de données de 20 espèces d'oiseaux présents au sein de la réserve. Les résultats montrent que le modèle basé sur les MFCC offre la meilleure précision globale (85 %), tandis que le modèle EfficientNetV2 s'est distingué par sa robustesse dans un contexte de données complexes et déséquilibrées. La gestion des déséquilibres entre classes, l'enrichissement des données et les défis liés à l'identification de plusieurs espèces simultanément sont discutés. Ces travaux ouvrent des perspectives pour la reconnaissance automatisée des chants d'oiseaux dans un cadre de conservation de la biodiversité.

II. Acknowledgements

A heartfelt thank you to Ken Weagle and Clio Smeeton for welcoming me to the CEI reserve. You provided me with the best possible working conditions, and I hope to have the opportunity to collaborate with you again in the future.
Thank you, Ken, for all the time you devoted to helping me advance my research.

Thank you to Barry Stark, Lisa Campbell, Kelsey Baldwin, and Cat Matheson, with whom I greatly enjoyed working.

Finally, my deepest gratitude to Rebecca MacArthur and Anne Rulff, who supported me throughout these three months in completing my project. It was a pleasure working alongside you, and I look forward to renewing this experience in the future.

Remerciements

Un grand merci à vous, KenWeagleet ClioSmeeton, pour m'avoir accueilli au sein de la réserve du CEI. Vous m'avez permis de travailler dans les meilleures conditions possibles, et j'espère avoir l'opportunité de collaborer de nouveau avec vous à l'avenir.
Merci à Ken pour tout le temps consacré à m'aider dans l'avancement de mes recherches.

Merci à Barry Stark, Lisa Campbell, Kelsey Baldwin et Cat Matheson, avec qui j'ai beaucoup apprécié travailler.

Enfin, merci infiniment à Rebecca MacArthur et Anne Rulff, qui m'ont épaulée au cours de ces trois mois dans la réalisation de mon projet. Ce fut un plaisir de travailler à vos côtés, et j'espère pouvoir renouveler cette expérience dans le futur.

III. GLOSSARY

AI: Artificial Intelligence

ARU: Audio Recording Unit

CEI: Cochrane Ecological Institute

HMM: Hidden Markov Model

MFCC: Mel-Frequency Cepstral Coefficients

ML: Machine Learning

STFT:Short-Time Fourier Transform

WAV: Waveform Audio File Format

IV. CONTENTS

V. Table of Contents

I

VI. INTRODUCTION

A. Context

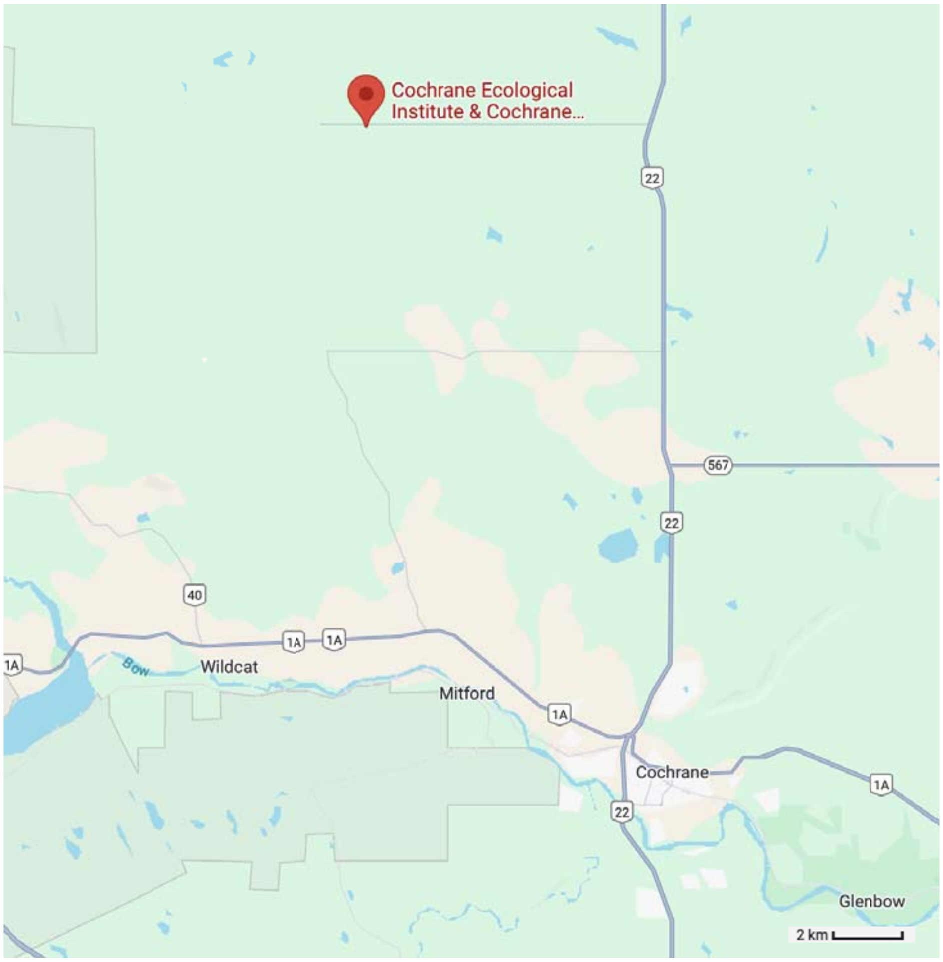
Birds are essential indicators of environmental change. Their rapid response to habitat loss and climate shifts enables early detection of ecosystem threats. By studying bird populations, scientists can assess ecosystem health and evaluate conservation efforts.Beyond their ecological value, birds contribute to key processes such as pollination, seed dispersal, and insect control, making their protection vital for biodiversity. However, human-induced pressures like habitat modification and climate change threaten bird populations(Canada, 2012).

This challenge highlights the importance of modern technologies for biodiversity monitoring. This is especially relevant in Alberta, Canada, where diverse bird species serve as critical bioindicators. Understanding their presence and behavior helps assess ecosystem health, reinforcing the need for efficient monitoring tools (Kaggle, 2023; Kaggle, 2024; Tang et al.2024).

Distinguishing between bird songs and calls is often challenging, especially for those unfamiliar with avian wildlife. However, audio recordings serve as essential tools for species identification, particularly for biologists studying, managing, and protecting bird populations.Yet, the vast diversity of vocalizations makes species recognition difficult. Manual classification methods require significant effort and are limited by an expert's ability to process large datasets. Additionally, these approaches are prone to human errors, complicating and slowing down conservation efforts.

To address these challenges, deep learning plays a key role in bird monitoring and demonstrates high performance in the field of vocal recognition. By combining autonomous recording units (ARUs) with recognition software, it is possible to analyze complex acoustic environments(Brooker et al., 2020). These automated tools enable more accurate and scalable bird species recognition, thereby contributing to the conservation of avian biodiversity.

Nevertheless, challenges remain, particularly in the precise extraction of relevant information from acoustic data.Moreover, these algorithms, although highly effective, require a significant amount of unique data, which can be a challenge for endangered species.



1

Figure 1:Satellite view of the CEI (CEI)

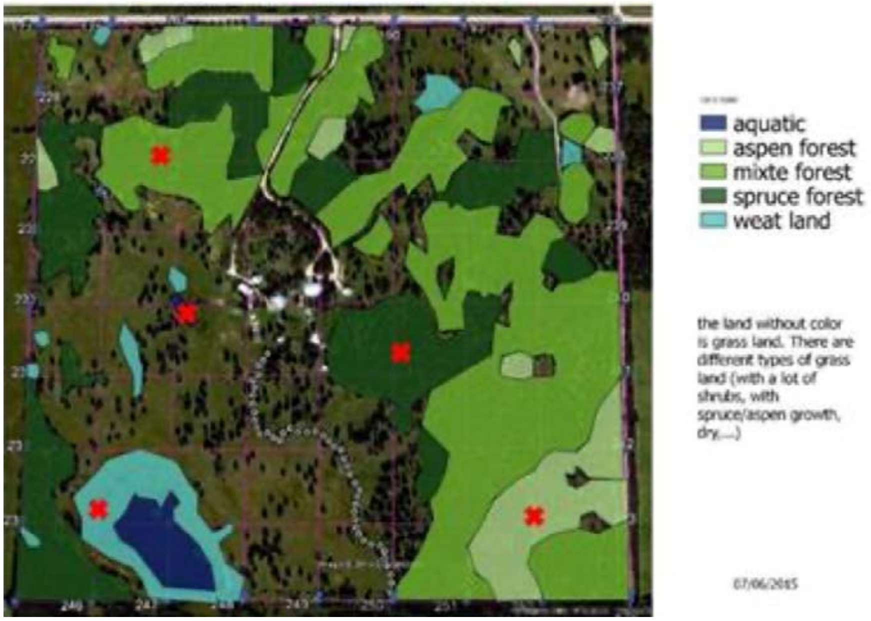

aquatic
aspen forest
mixte forest
spruce forest
weat land
the land without color is grass land. There are different types of grass land (with a lot of shrubs, with spruce/aspen growth, dry,...)
07/06/2015

Figure 2:Map of the different biomes that make up the CEI reserve in 2015 (CEI)

In recent years, significant research has focused on developing new algorithmic models to enhance bird sound recognition. Among the most effective techniques are log-Mel spectrograms and deep neural networks, which have improved classification accuracy.Convolutional neural networks (CNNs) trained on Mel spectrograms or Mel-frequency cepstral coefficients (MFCCs) have shown remarkable efficiency (Das et al., 2023). These methods have been validated in multiple studies, including those by Stowell and Plumbley (2014), as well as in major competitions like BirdCLEF 2024.

Regarding alternatives to short-time Fourier transform (STFT), other approaches, such as linear prediction (Fox, 2008), have been explored. However, Mel spectrograms and their variants generally dominate classification applications due to their ability to effectively represent the perceptual features of sounds. The BirdCLEF 2024 competition particularly highlighted advances in this field. The winning model achieved an impressive 96% accuracy by using an architecture based on EfficientNetV2 (Kaggle, 2024).

This study aims **to develop an artificial intelligence model for bird species recognition based on a dataset collected within the CEI reserve**.

A total of 20 bird species were identified from 4,698 audio recordings gathered in June, September, and October. One of the key challenges was addressing class imbalance to ensure reliable model training. To achieve this, we designed three deep learning models using different feature extraction techniques: **MFCC, STFT,** and **log-Mel spectrograms**combined with a **pre-trained EfficientNetV2-B0 model**.

These models were then evaluated based on accuracy, execution speed, and F1-score. Additional analyses included accuracy and loss evolution curves, as well as confusion matrices to assess classification performance.

The report is structured as follows: in the first section, we will present the materials and methodology used, followed by the results obtained and their analysis in the third section. Finally, we will discuss the implications of our results and potential directions for future research in the conclusion.

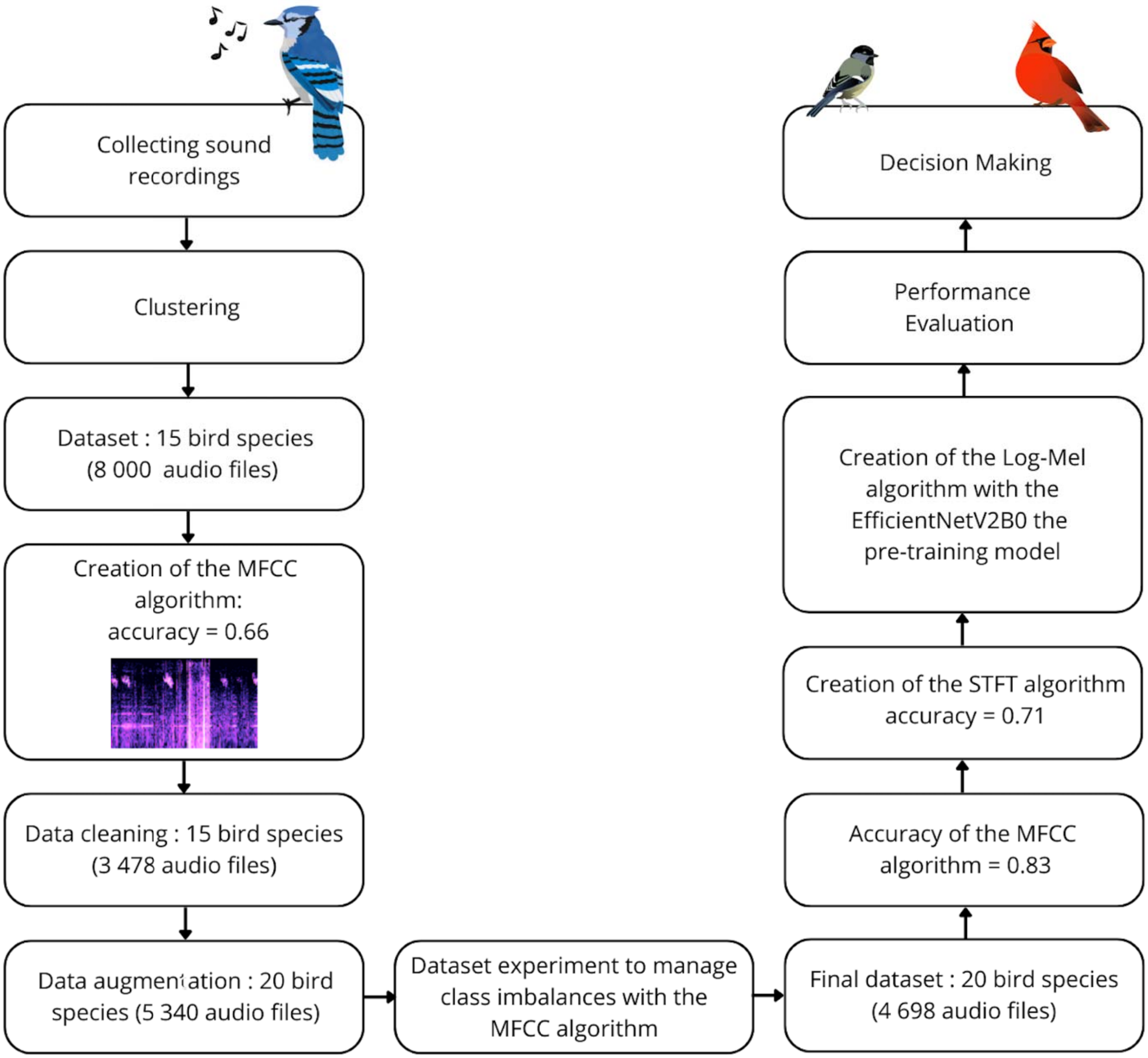

Collecting sound recordings
Clustering
Dataset : 15 bird species (8 000 audio files)
Creation of the MFCC algorithm: accuracy = 0.66
Data cleaning : 15 bird species (3 478 audio files)
Data augmentation : 20 bird species (5 340 audio files)
Dataset experiment to manage class imbalances with the MFCC algorithm
Final dataset : 20 bird species (4 698 audio files)
Accuracy of the MFCC algorithm = 0.83
Creation of the STFT algorithm accuracy = 0.71
Creation of the Log-Mel algorithm with the EfficientNetV2B0 the pre-training model
Performance Evaluation
Decision Making

Figure 3:Key steps throughout the project

B. Presentation of the Reserve

The Cochrane Ecological Institute (CEI) is located in Alberta, Western Canada, just north of the town of Cochrane (Figure 1). Covering over 60 hectares, the reserve consists of diverse biomes, including forests, grasslands, and wetlands (Figure 2)(Cochrane Ecological Institute, n.d.).

Dedicated to wildlife reintroduction and conservation, the CEI plays a crucial role in multiple research projects. It has deployed drones equipped with AI-powered cameras to monitor and identify mammals. Additionally, it oversees projects leveraging acoustic recordings to track mammal and bird populations.

The region's rich biodiversity makes it an ideal location for avian acoustic research. According to the Merlin Bird ID app from the Cornell Lab, 169 bird species have been recorded in Rocky View No. 44, AB, particularly during September and October (Cornell Lab of Ornithology, n.d.). The variety of habitats and the high species diversityprovide a unique environment for studying bird recognition through audio.

VII. MATERIALS AND METHODS

A. Collectingsoundrecordings

In this study, the dataset was created from acoustic data collected using acoustic recording units (ARUs): **Song Meter SM4** Acoustic Recorder and **Song Meter Mini Bat 2** from Wildlife Acoustics Inc. These devices were installed at approximately 2 meters in height on tree trunks at each of the four designated sites (Figure 4 and 5) within the reserve. These sites were selected as they represent different biomes, thereby increasing the chances of detecting a diverse range of species attracted to various habitats.

Each ARU was equipped with an omnidirectional microphone, allowing ambient sounds to be recorded in all directions. Recordings were made with a sampling rate of 16000 Hz and 16-bit encoding. No high-pass or band-pass filters were applied to preserve the full range of sound frequencies present in the environment.

The ARUs were programmed to record sounds for 5 minutes every 30 minutes over a period of 4 consecutive days, at multiple intervals: on September 21 and October 3, 2024. These dates correspond to a period of heightened bird activity (e.g., migration or breeding). The audio files were subsequently retrieved in WAV format for analysis.

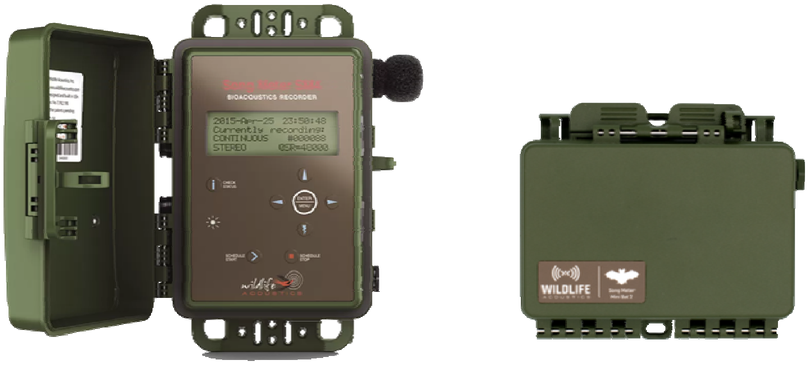


Figure 4:On the left, SONG METER SM4 ACOUSTIC RECORDER, and on the right, SONG METER MINI BAT 2 AA (Wildlife Acoustics)

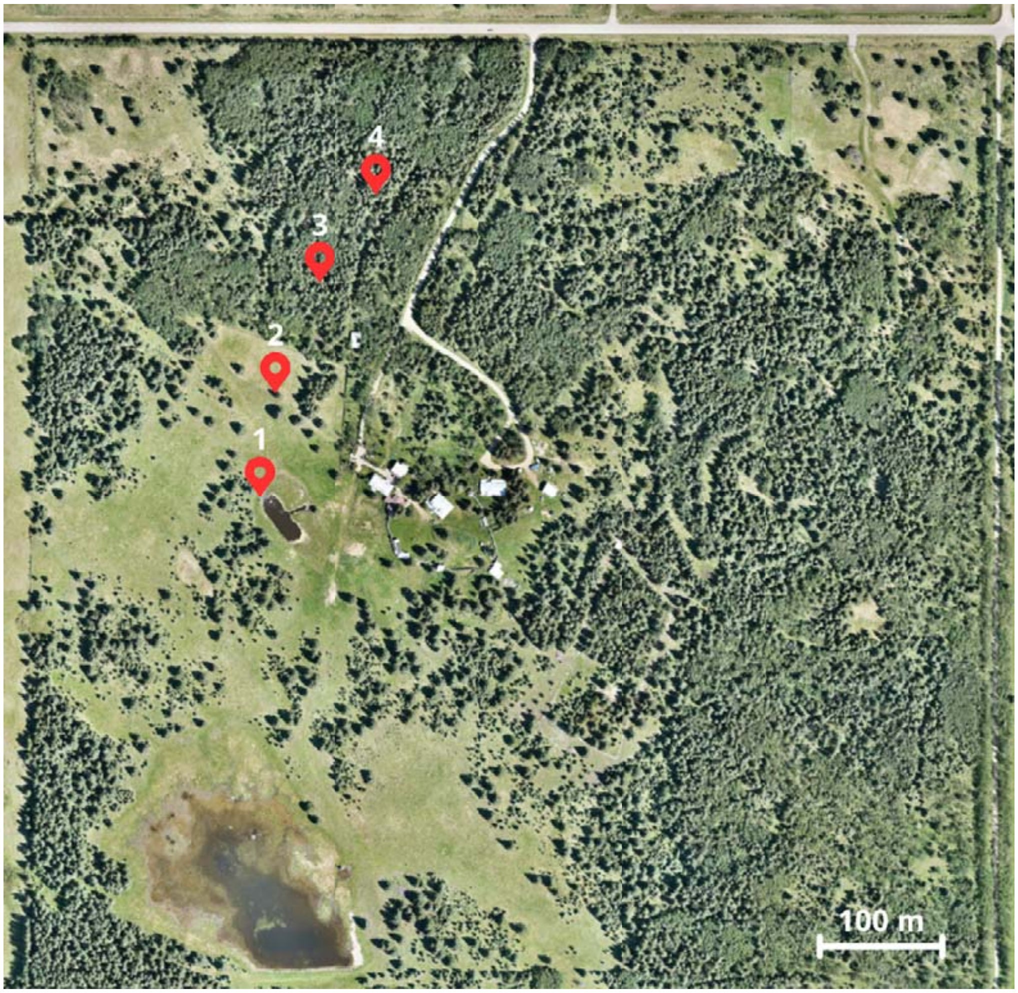


Figure 5:Positions of ARUs within the CEI reserve (Google map)

B. Clustering

1. Kaleidoscope pro

Kaleidoscope Pro (Wildlife Acoustics Inc, Maynard, USA) is an application that uses Hidden Markov Models (HMM) to group syllables into clusters based on their spatiotemporal features and similarity (Brooker et al., 2020). Each detection is assigned to a cluster with a score assessing its relevance to the cluster's center (Wildlife Acoustics Inc, n.d.).

The detection parameters were set for a frequency range of 0 to 20 kHz, a minimum duration of 0.1 seconds, a maximum duration of 15 seconds, and a maximum interval between syllables of 0.5 seconds. The other settings remained as default. The program generates a cluster.kcs file and an Excel file listing the audio files grouped by cluster with their characteristics (frequency, duration, etc.).

Kaleidoscope Pro sorts files into clusters without identifying them. Therefore, manual verification is required to associate each cluster with a species. The validated files are exported in WAV format to ensure consistency in the algorithm's input data.

During the analysis with Kaleidoscope Pro, some files had a modified frequency from 16000 Hz to 97000 Hz. A Python algorithm was developed to convert them back to 16000 Hz—a rate suitable for bird sounds and more efficient in terms of storage space.

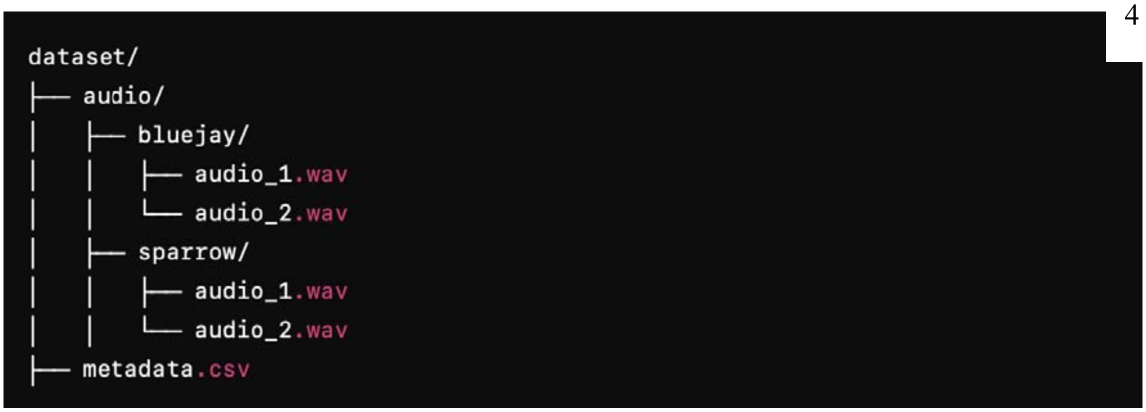
2. Merlin Bird ID

Merlin Bird ID (Cornell Lab of Ornithology, Ithaca, USA) is a mobile bird recognition application that uses convolutional neural networks to identify species from photos or sounds. Relying on a database of over 6000 species, it assigns a match score based on the similarity between the submitted recording and the training data. It also provides information on species' habitats, behaviors, and geographical distributions(Cornell Lab of Ornithology, n.d.).

In this study, the assignment of clusters obtained with Kaleidoscope Pro was performed using Merlin Bird ID.

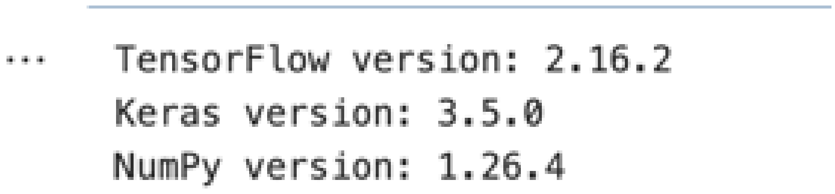
However, this identification may be limited when recordings are too short, have low sound intensity, or contain excessive background noise, making the analysis less accurate.

4



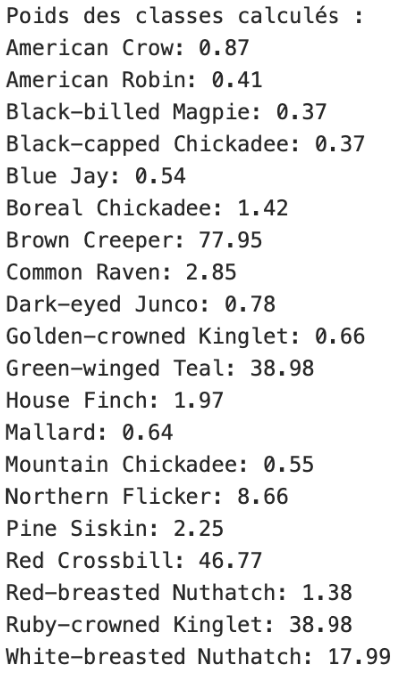
```
dataset/
├── audio/
│   ├── bluejay/
│   │   ├── audio_1.wav
│   │   └── audio_2.wav
│   ├── sparrow/
│   │   ├── audio_1.wav
│   │   └── audio_2.wav
├── metadata.csv
```

Figure 6: Example of the algorithm's database structure



```
...   TensorFlow version: 2.16.2
      Keras version: 3.5.0
      NumPy version: 1.26.4
```

Figure 7: Excerpt from the Jupyter Notebook showing the versions of the imported libraries



```
Poids des classes calculés :
American Crow: 0.87
American Robin: 0.41
Black-billed Magpie: 0.37
Black-capped Chickadee: 0.37
Blue Jay: 0.54
Boreal Chickadee: 1.42
Brown Creeper: 77.95
Common Raven: 2.85
Dark-eyed Junco: 0.78
Golden-crowned Kinglet: 0.66
Green-winged Teal: 38.98
House Finch: 1.97
Mallard: 0.64
Mountain Chickadee: 0.55
Northern Flicker: 8.66
Pine Siskin: 2.25
Red Crossbill: 46.77
Red-breasted Nuthatch: 1.38
Ruby-crowned Kinglet: 38.98
White-breasted Nuthatch: 17.99
```

Figure 8: Class weights in the MFCC algorithm

3. Database Format

The database was organized into two main directories for a clear structure (Figure 6). The first directory, titled "**audio**," contains individual subfolders for each identified species. These subfolders group all the audio files associated with the corresponding species, facilitating their access and management.

The second directory, named "**Metadata**," centralizes descriptive information extracted from the audio files using Kaleidoscope Pro. These metadata include details such as frequency, duration, and call characteristics, enabling a systematic and comprehensive analysis of the recordings.

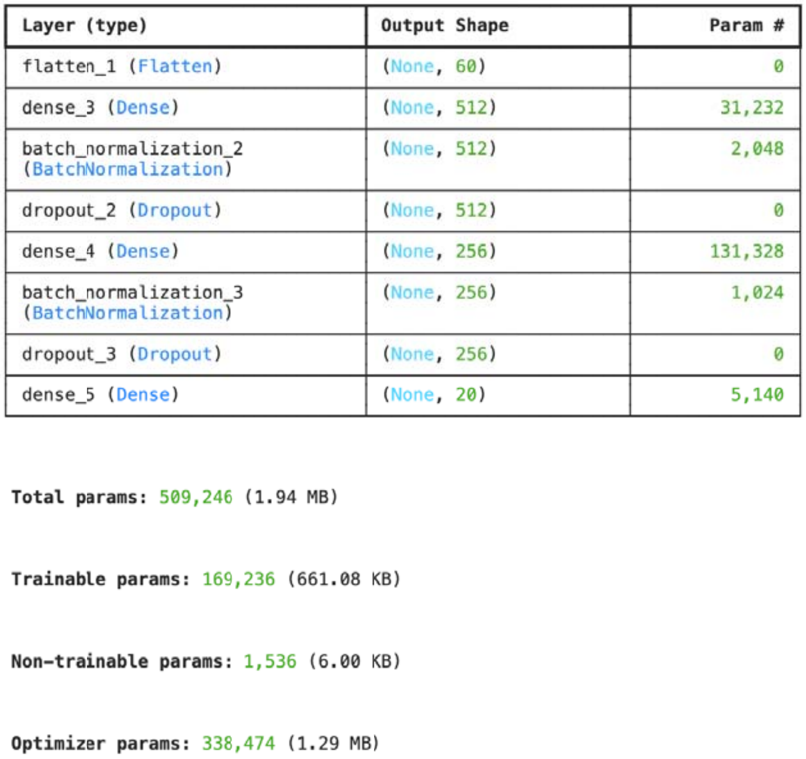
4. MFCC Algorithm

The first classification algorithm is based on the extraction of Mel-Frequency Cepstral Coefficients (MFCC), a set of features widely used in speech recognition and acoustic analysis tasks. These coefficients represent the spectral and perceptual information of audio signals, inspired by the human perception of sound frequencies(Stowell &Plumbley,2014; Das et al., 2023; Rezaul, 2024; Deng et al., 2020).

The algorithm was developed in Python (version 3.12.6) using the Visual Studio Code (VSCode) editor (Figure 7). The bird song audio files, in **WAV format**, were first preprocessed with the 'Librosa' library to extract 60 MFCCs per sample. These coefficients were then averaged over the time axis to reduce variability between recordings of the same species and produce a more stable and compact feature vector.

Normalization of the MFCCs was performed using 'StandardScaler', ensuring that all variables are on a similar scale, which improves the model's convergence during training.

The bird species labels were encoded using a One-Hot encoder, converting categorical data into numerical representations suitable for multiclass classification(Scikit-Learn Developers, n.d.). The dataset was split into two sets: **80% for training and 20% for validation**, to evaluate the model's performance on unseen data during learning.



```
┌──────────────────────────────────┬──────────────────────────┬───────────────┐
│ Layer (type)                     │ Output Shape             │       Param # │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ flatten_1 (Flatten)              │ (None, 60)               │             0 │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ dense_3 (Dense)                  │ (None, 512)              │        31,232 │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ batch_normalization_2            │ (None, 512)              │         2,048 │
│ (BatchNormalization)             │                          │               │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ dropout_2 (Dropout)              │ (None, 512)              │             0 │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ dense_4 (Dense)                  │ (None, 256)              │       131,328 │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ batch_normalization_3            │ (None, 256)              │         1,024 │
│ (BatchNormalization)             │                          │               │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ dropout_3 (Dropout)              │ (None, 256)              │             0 │
├──────────────────────────────────┼──────────────────────────┼───────────────┤
│ dense_5 (Dense)                  │ (None, 20)               │         5,140 │
└──────────────────────────────────┴──────────────────────────┴───────────────┘

 Total params: 509,246 (1.94 MB)

 Trainable params: 169,236 (661.08 KB)

 Non-trainable params: 1,536 (6.00 KB)

 Optimizer params: 338,474 (1.29 MB)
```

Figure 9:Summary of the MFCC algorithm and neural network layers

The classification model is based on a deep neural network with three main layers (Figure 9):

1. **Flatten Layer:** This layer transforms the input MFCCs, initially multidimensional, into a one-dimensional vector that can be processed by the subsequent dense layers.
2. **Dense Layers:** Two dense layers with 512 and 256 neurons, respectively, are used to capture complex relationships in the data. Each dense layer is followed by batch normalization to stabilize learning and accelerate convergence, as well as a dropout mechanism (with a rate of 0.3) to reduce the risk of overfitting.
3. **Output Layer:** The output layer, also a dense layer, applies a Softmax activation function to predict the probability of belonging to each bird species class.

The model training is optimized using the Adam algorithm with a learning rate of 0.0005. The categorical ‘crossentropy loss’ function is used to maximize the probability of correctly predicting species labels.

To address the class imbalance issue, class weights are calculated and integrated to compensate for the underrepresentation of rare species in the training data. Additionally, early stopping and best model saving mechanisms are employed to halt training as soon as the performance on the validation set begins to degrade.

The model is evaluated on a validation set, where its ability to generalize to unseen data is measured. The prediction accuracy is compared to the actual labels to estimate performance in a real-world context.

5. Data cleaning

One of the factors that most significantly influences the accuracy of an algorithm is the quality of the dataset. In our study, the first step to improve accuracy was to manually review the audio files one by one.

We first verified whether the bird sound corresponded correctly to its assigned class. Some recordings were removed due to poor quality or the presence of background noise. Many audio files were also discarded following the incorrect detection of sound events by Kaleidoscope Pro.

Moreover, some classes had a significantly higher number of detections compared to our target of approximately 500 audio samples per class (Figure 10).

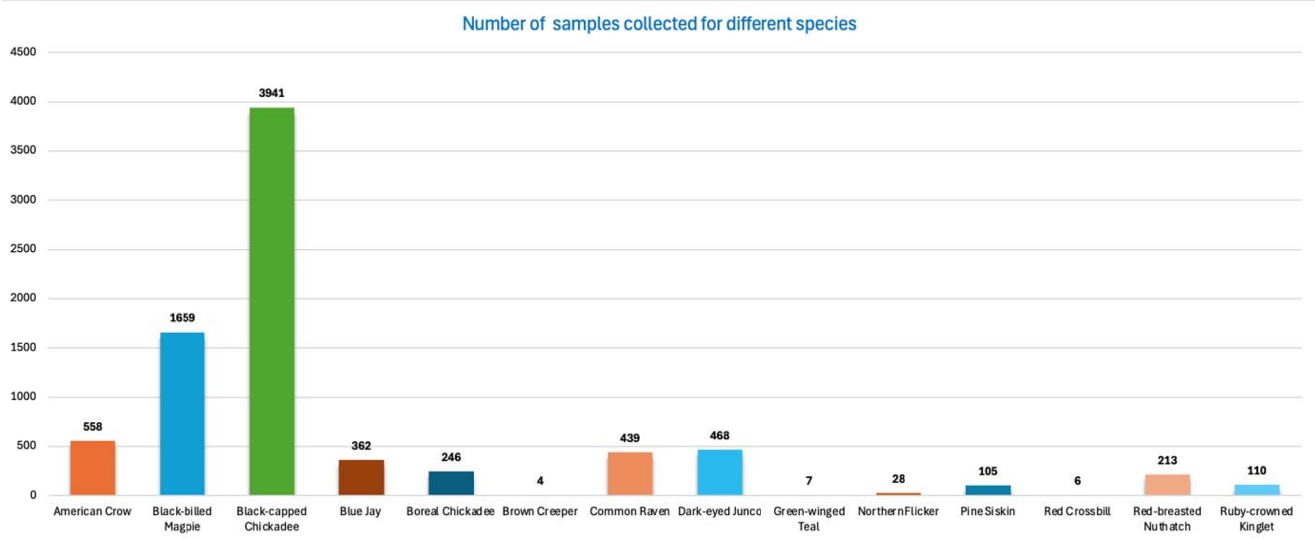


Figure 10: Uncleaned and incomplete dataset consisting of 15 species and approximately 8146 audio samples

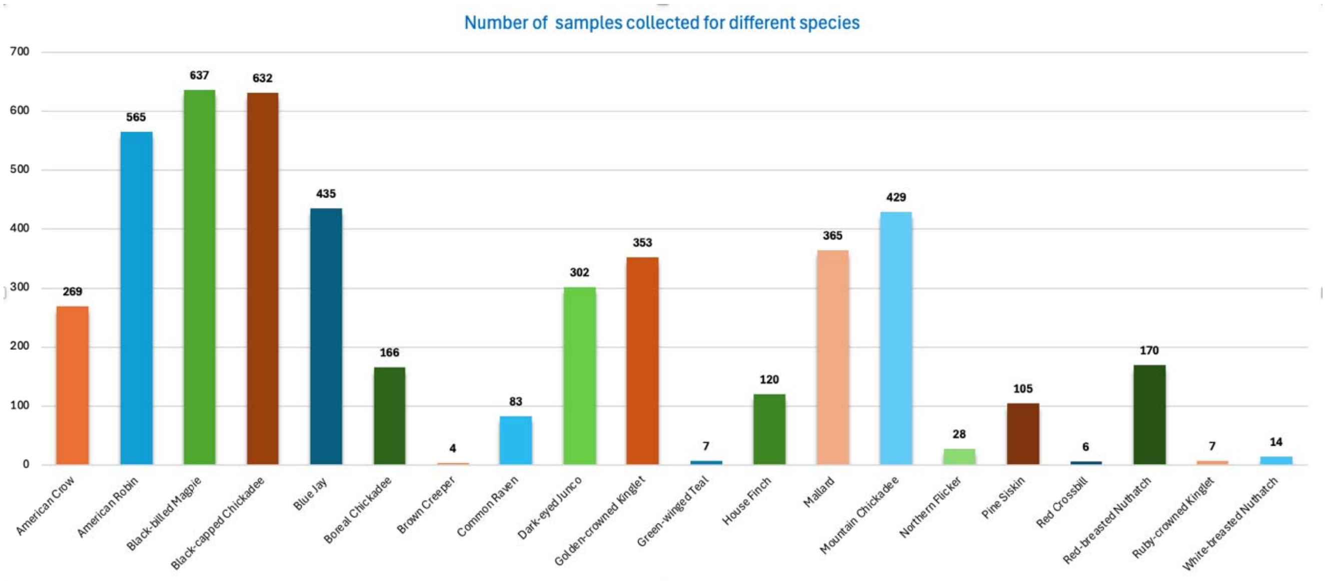


Figure 11:Final dataset with 4 698 audio samples and 20 classes

These classes contained up to 2000 recordings, mainly due to the high abundance of these species in the reserve and their particularly noisy behavior. To avoid excessive class imbalance, we reduced the number of audio samples in these categories.

The initial dataset consisted of 15 species, totaling 8 146 recordings. After this initial cleaning process, the number of audio samples was reduced to 3 478. These cleaning steps improved the overall quality of the data.

6. Data augmentation

Another essential factor for audio classification is the number of samples available for each species (Rezaul, 2024). To further enrich our dataset, recordings made on July 10 and 25, with the same settings as before, were integrated.

These new recordings underwent the same processing steps as the initial data: clustering by Kaleidoscope Pro, species identification via Merlin Bird ID, and manual verification of the audio files. Additionally, the July period allowed for the detection and inclusion of new species present during that time.

Following this step, the dataset consisted of 20 bird species, with a total of 5340 recordings. The integration of these additional data aimed to maximize the quantity and diversity of information available to train the algorithm, thereby improving its performance.

The updated accuracy of the MFCC algorithm after data cleaning and data augmentation was 83%, compared to the initial accuracy of 66% with the previous dataset.

7. Manage class imbalances with MFCC algorithm

In classification algorithms, class imbalance can affect overall performance by biasing the model in favor of overrepresented classes. During the development of our dataset, some species were overrepresented due to their natural abundance and noisy vocal behavior, while others were underrepresented because they were less common within the reserve or were rare species.

A study on the performance of the MFCC algorithm was conducted by modifying the dataset to determine the best configuration to use.

	UnmodifiedDataset	200 fewer audio samples for each overrepresented species	300 fewer audio samples for each overrepresented species	400 fewer audio samples for each overrepresented species	By removingunderrepresentedspecies	By removing underrepresented species and 200 audio samples from overrepresented species
Accuracy of the MFCC Algorithm	0.83	↘□ 0.82	↘□ 0.80	↘□ 0.79	↖□ 0 .85	↖□ 0.84

Table 1: Comparison of MFCC algorithm accuracy based on the dataset

First, tests were carried out by removing audio samples from overrepresented species, namely the American Robin, Black-billed Magpie, and Black-capped Chickadee. Further tests were performed by excluding underrepresented species, including the Red Crossbill, Northern Flicker, White-breasted Nuthatch, Brown Creeper, and Ruby-crowned Kinglet, as well as by combining both approaches (Table 1).

Ultimately, it was decided to remove 200 audio samples from the overrepresented species to limit their influence during training without losing a significant amount of information. The underrepresented species were retained, despite slightly reducing the model's accuracy, as they accurately reflect the diversity of the studied environment.

The final dataset comprises 20 species, with a total of 4698 audio samples (Figure 11).

8. STFT algorithm

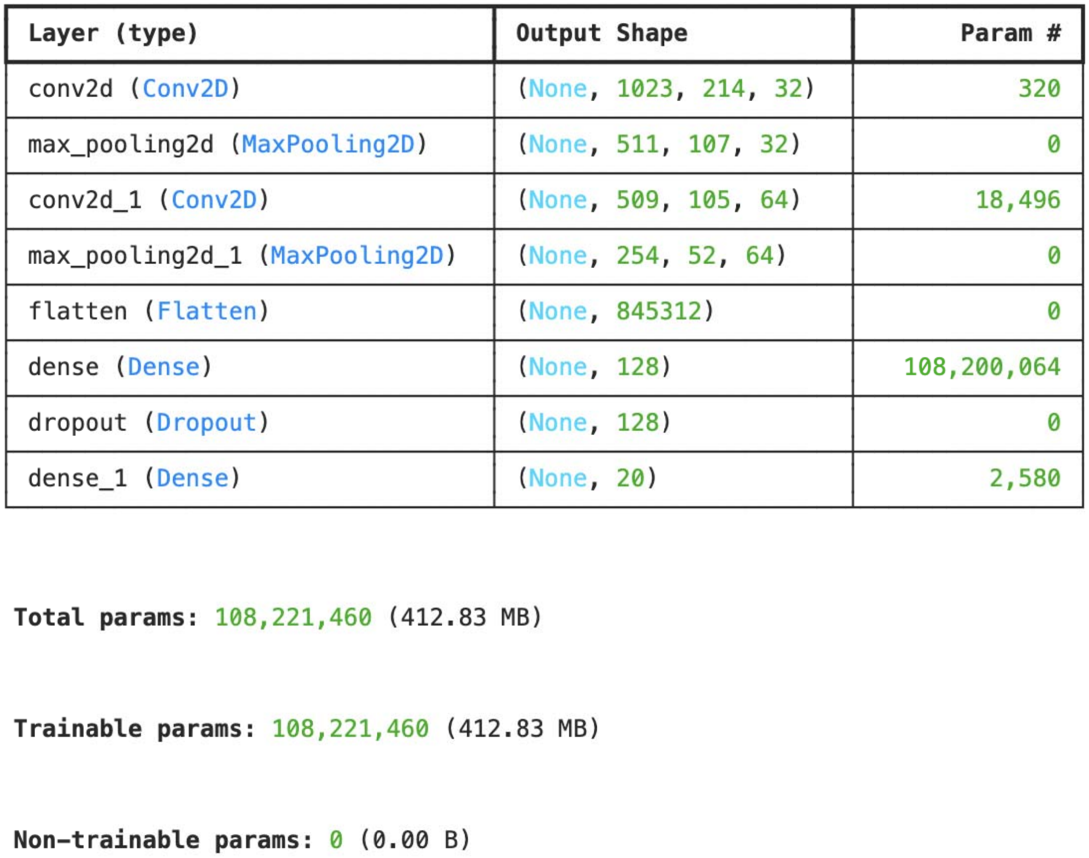
The classification algorithm based on Short-Time Fourier Transform (STFT) relies on the analysis of spectrograms, which represent the frequency variations of the audio signal over time.

The STFT is particularly suited for analyzing complex sounds, such as bird songs, as it provides a detailed spectral view of the signals. Unlike MFCCs, which focus on perceptual aspects by reducing dimensionality, spectrograms retain fine resolution in both frequency and time. This makes them ideal for capturing the spatiotemporal variations specific to each species(Xie et al., 2022; Puget, 2021).

For each audio file, spectrograms are generated and then converted to a logarithmic scale using the ‘amplitude_to_db’function from ‘Librosa’. This conversion enhances the representation of amplitude variations. The spectrograms are then resized to ensure uniform dimensions, facilitating their use as input in a convolutional neural network (CNN).

The CNN model is composed of the following elements (Figure 12):

1. **Two convolutional layers (Conv2D):** These extract characteristic patterns from the spectrograms, capturing local relationships in the spatiotemporal data.
2. **Two max-pooling layers:** These reduce the dimensionality while preserving essential information, making the model more robust and less prone to overfitting.



Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 1023, 214, 32)	320
max_pooling2d (MaxPooling2D)	(None, 511, 107, 32)	0
conv2d_1 (Conv2D)	(None, 509, 105, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 254, 52, 64)	0
flatten (Flatten)	(None, 845312)	0
dense (Dense)	(None, 128)	108,200,064
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 20)	2,580

```
Total params: 108,221,460 (412.83 MB)

Trainable params: 108,221,460 (412.83 MB)

Non-trainable params: 0 (0.00 B)
```

Figure 12: Summary of the STFT algorithm and neural network layers

3. **A fully connected Dense layer with 128 neurons:** This layer interprets the features extracted by the convolutional layers.
4. **A Dropout regularization layer:** This layer, with an appropriate dropout rate, helps prevent overfitting.
5. **An output Dense layer with Softmax activation:** This layer generates the probabilities for each bird species class.

The algorithm is trained using the ‘Adam’ optimizer and the categorical ‘crossentropy loss’ function, with the dataset split into training (80%) and validation (20%) sets. To optimize the training process, callbacks such as early stopping and model checkpoint are used.

This STFT-based approach retains rich spectral information, which is particularly useful for distinguishing bird species with similar songs. It provides a complementary perspective to MFCCs by emphasizing detailed spatiotemporal characteristics.

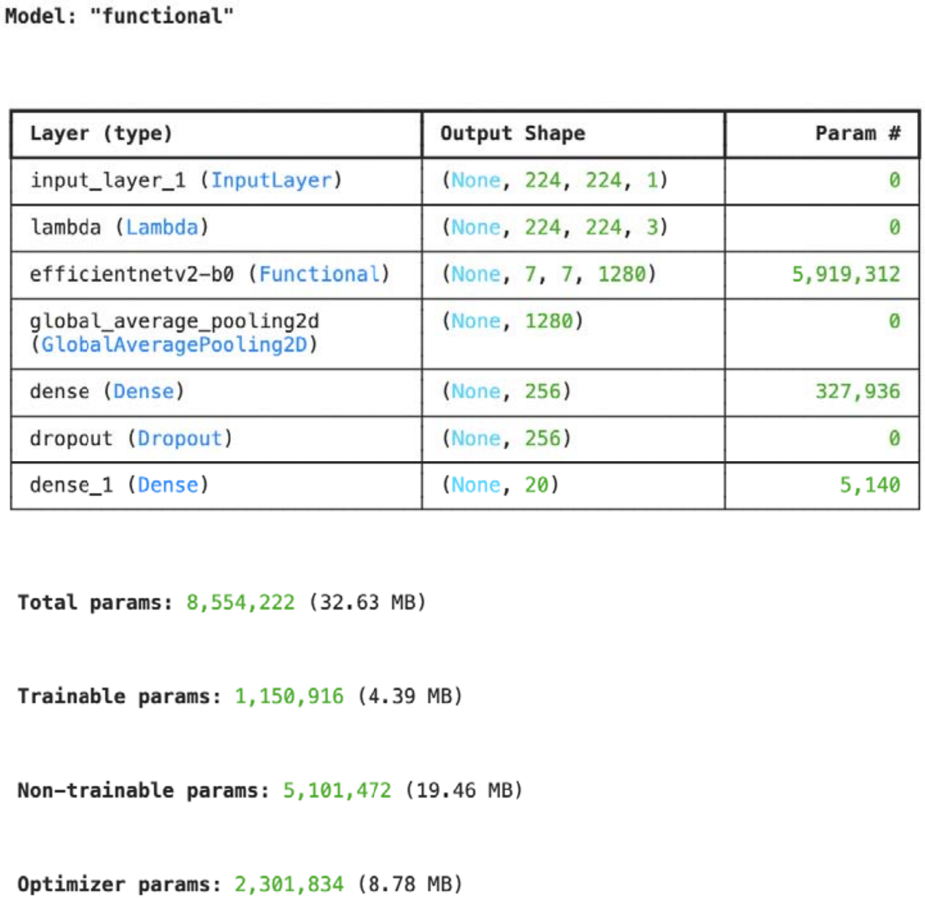
9. Log-Melspectrograms and EfficientNetV2 algorithm

Pre-trained models, such as EfficientNetV2, provide a powerful foundation for classification tasks by leveraging knowledge gained from large datasets. They reduce the need for task-specific data, accelerate training, and enhance performance, particularly for complex applications such as bird sound classification(Das et al., 2023).

These models have demonstrated their effectiveness in competitions such as BirdCLEF and in various audio classification algorithms (Kaggle, 2024).

In this algorithm, the bird song audio files are transformed into log-Mel spectrograms, a representation suited to the human perception of sounds. Unlike traditional spectrograms, log-Mel places more emphasis on lower frequencies—characteristic of bird songs—while minimizing the influence of very faint or loud sounds(Das et al., 2023).

These spectrograms are resized to 224x224 pixels to meet the input requirements of the EfficientNetV2 model, a convolutional neural network pre-trained on ImageNet, which is an image dataset. This resizing allows the model to leverage the complex visual features it has already learned.



```
Model: "functional"

┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                     ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ input_layer_1 (InputLayer)       │ (None, 224, 224, 1)    │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ lambda (Lambda)                  │ (None, 224, 224, 3)    │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ efficientnetv2-b0 (Functional)   │ (None, 7, 7, 1280)     │     5,919,312 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ global_average_pooling2d         │ (None, 1280)           │             0 │
│ (GlobalAveragePooling2D)         │                        │               │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ dense (Dense)                    │ (None, 256)            │       327,936 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ dropout (Dropout)                │ (None, 256)            │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                  │ (None, 20)             │         5,140 │
└──────────────────────────────────┴────────────────────────┴───────────────┘

 Total params: 8,554,222 (32.63 MB)

 Trainable params: 1,150,916 (4.39 MB)

 Non-trainable params: 5,101,472 (19.46 MB)

 Optimizer params: 2,301,834 (8.78 MB)
```

Figure 13:Summary of the Log-Mel spectrograms and EfficientNetV2 algorithm and neural network layers

To enhance the model’s robustness, data augmentation techniques inspired by ‘SpecAugment’ are applied(Das et al, 2023). These augmentations introduce artificial variations, such as time and frequency masking, simulating natural recording conditions.

During training, only the final layers of the EfficientNetV2 model are retrained (fine-tuning), while the deeper layers remain frozen to retain the knowledge acquired during pre-training.

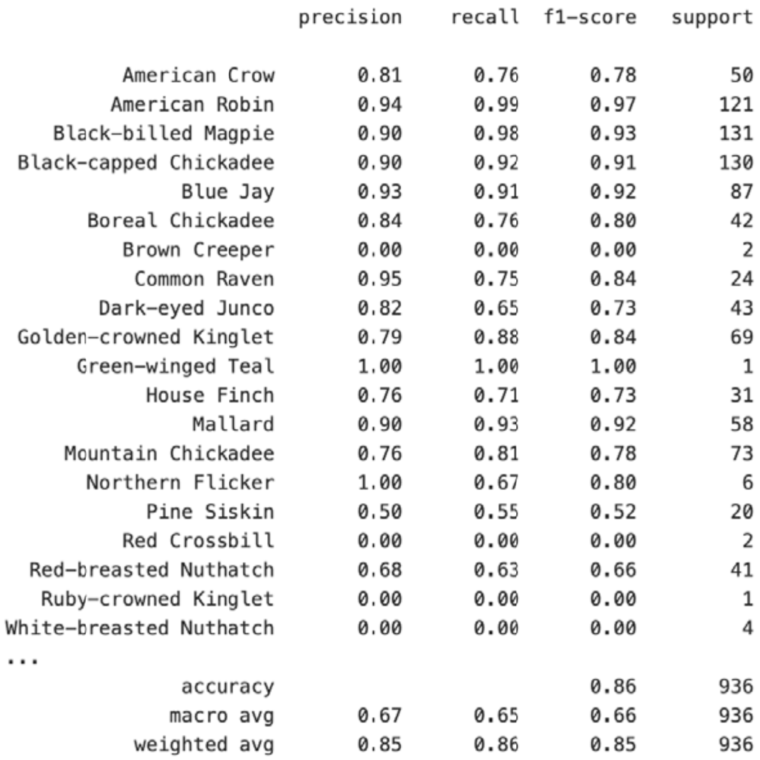
A custom head is added to the top of the model, consisting of the following elements:

1. **A Dense layer with 256 neurons,** followed by a Dropout mechanism to prevent overfitting.
2. **An output Dense layer with ‘Softmax’ activation,** which predicts the probabilities for the 20 bird species classes.

The training process is optimized using the categorical ‘crossentropy loss’ function, the ‘Adam’ optimizer, and a learning rate of 0.0005.

Mechanisms such as ‘Earlystopping’ and ‘ModelCheckpoint’ are implemented to prevent overfitting and ensure optimal generalization.

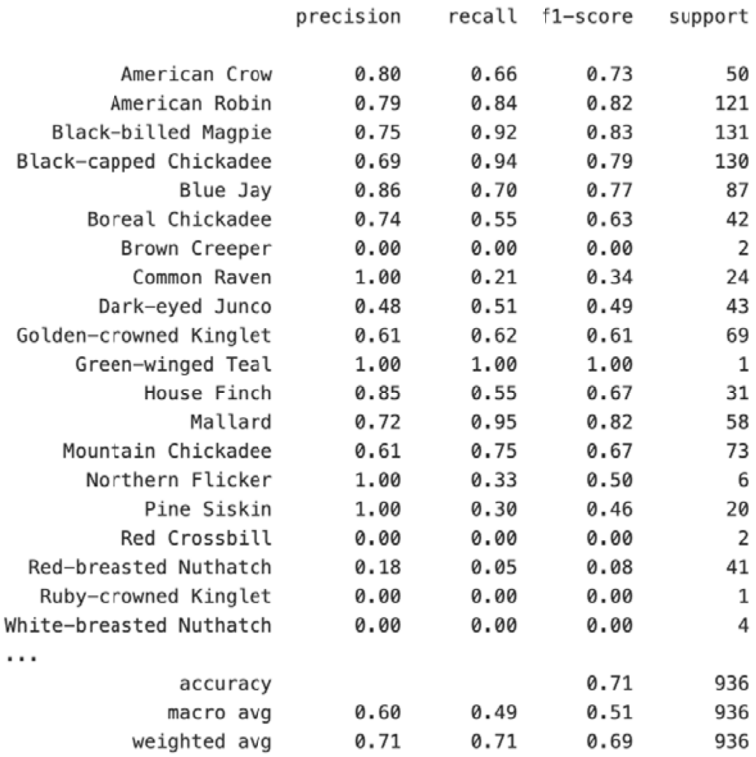
This model is evaluated on a validation set to assess its ability to generalize and classify bird species based on their songs. By combining log-Mel spectrograms with EfficientNetV2, this model provides a robust solution, even with a limited dataset.



```
                         precision    recall  f1-score   support

          American Crow       0.81      0.76      0.78        50
        American Robin       0.94      0.99      0.97       121
    Black-billed Magpie       0.90      0.98      0.93       131
 Black-capped Chickadee       0.90      0.92      0.91       130
               Blue Jay       0.93      0.91      0.92        87
       Boreal Chickadee       0.84      0.76      0.80        42
          Brown Creeper       0.00      0.00      0.00         2
          Common Raven       0.95      0.75      0.84        24
        Dark-eyed Junco       0.82      0.65      0.73        43
 Golden-crowned Kinglet       0.79      0.88      0.84        69
      Green-winged Teal       1.00      1.00      1.00         1
            House Finch       0.76      0.71      0.73        31
                Mallard       0.90      0.93      0.92        58
     Mountain Chickadee       0.76      0.81      0.78        73
       Northern Flicker       1.00      0.67      0.80         6
            Pine Siskin       0.50      0.55      0.52        20
          Red Crossbill       0.00      0.00      0.00         2
  Red-breasted Nuthatch       0.68      0.63      0.66        41
   Ruby-crowned Kinglet       0.00      0.00      0.00         1
White-breasted Nuthatch       0.00      0.00      0.00         4
...
               accuracy                           0.86       936
              macro avg       0.67      0.65      0.66       936
           weighted avg       0.85      0.86      0.85       936
```

Table 2: MFCC Model Classification Report for Bird Species Identification



```
                         precision    recall  f1-score   support

          American Crow       0.80      0.66      0.73        50
        American Robin       0.79      0.84      0.82       121
    Black-billed Magpie       0.75      0.92      0.83       131
 Black-capped Chickadee       0.69      0.94      0.79       130
               Blue Jay       0.86      0.70      0.77        87
       Boreal Chickadee       0.74      0.55      0.63        42
          Brown Creeper       0.00      0.00      0.00         2
          Common Raven       1.00      0.21      0.34        24
        Dark-eyed Junco       0.48      0.51      0.49        43
 Golden-crowned Kinglet       0.61      0.62      0.61        69
      Green-winged Teal       1.00      1.00      1.00         1
            House Finch       0.85      0.55      0.67        31
                Mallard       0.72      0.95      0.82        58
     Mountain Chickadee       0.61      0.75      0.67        73
       Northern Flicker       1.00      0.33      0.50         6
            Pine Siskin       1.00      0.30      0.46        20
          Red Crossbill       0.00      0.00      0.00         2
  Red-breasted Nuthatch       0.18      0.05      0.08        41
   Ruby-crowned Kinglet       0.00      0.00      0.00         1
White-breasted Nuthatch       0.00      0.00      0.00         4
...
               accuracy                           0.71       936
              macro avg       0.60      0.49      0.51       936
           weighted avg       0.71      0.71      0.69       936
```

Table 3: STFT Model Classification Report for Bird Species Identification

VIII. RESULTS

We evaluated three classification models: MFCC, STFT, and log-Mel spectrograms combined with a pre-trained EfficientNetV2 model. Their performances are detailed in Tables 2, 3, and 4, highlighting key differences in accuracy, recall, and F1-score.

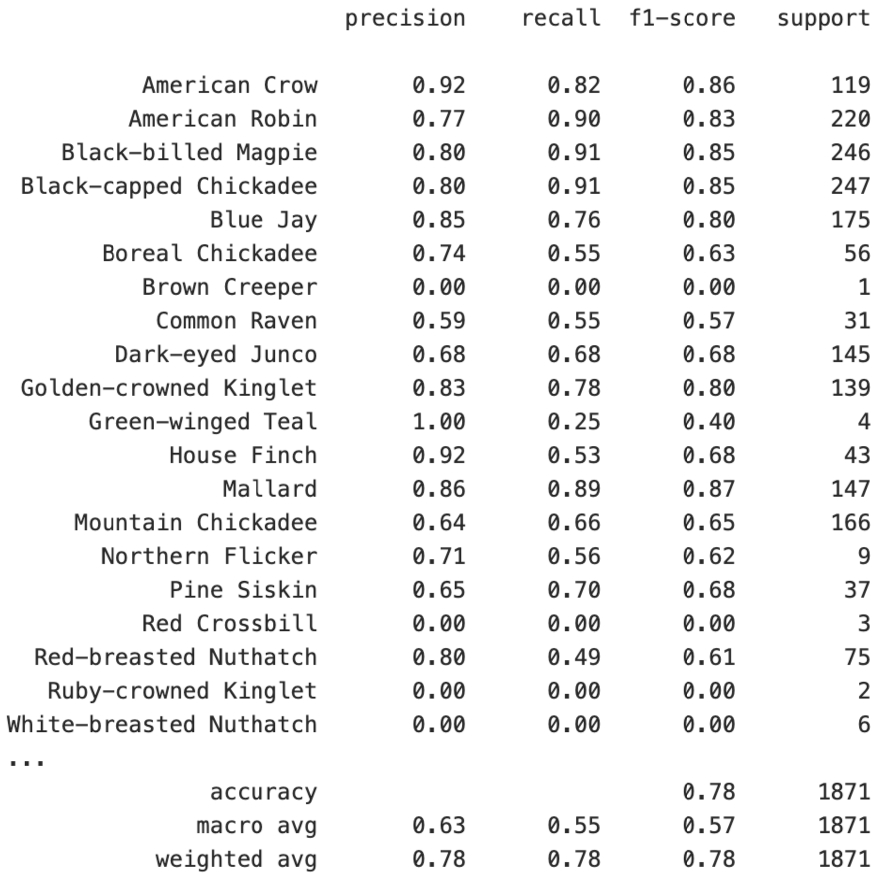
The **MFCC model** shows promising results, with an overall **accuracy of 85%, a recall of 86%, and an average F1-score of 85%**. However, difficulties persist in detecting certain species despite having enough samples, such as the Red-breasted Nuthatch, with 41 samples used for testing. This difficulty may be because the audio for this species often contains noisy environments with overlapping bird sounds. We can assume that the algorithm might prioritize other species due to their higher frequencies or dominance in the audio. The Brown Creeper, Red Crossbill, Ruby-crowned Kinglet, and White-breasted Nuthatch were not recognized at all, which negatively affects the overall performance.

The **STFT model**, on the other hand, achieves an average **accuracy of 71%, a recall of 60%, and an F1-score of 51%**. Although some species, such as the Green-winged Teal, are detected with perfect accuracy (100%), possibly because only one audio sample was used for testing, several other species, including the Brown Creeper and Ruby-crowned Kinglet, are not recognized at all, affecting the overall performance similarly to the MFCC model.

The **log-Mel model combined with EfficientNetV2** shows generally better performance, with an average **accuracy of 78%, a recall of 63%, and an F1-score of 57%**. Despite strong results for some species, such as the American Crow and Mallard, there are still difficulties in classifying species like the Brown Creeper and Ruby-crowned Kinglet, limiting the overall model performance once again.

These results highlight the importance of audio representations and model architectures in the classification of bird songs. In terms of overall performance, the MFCC model outperforms the others, followed by the log-Mel model with EfficientNetV2, while the STFT model shows more modest results, although it demonstrates high accuracy for some specific species.

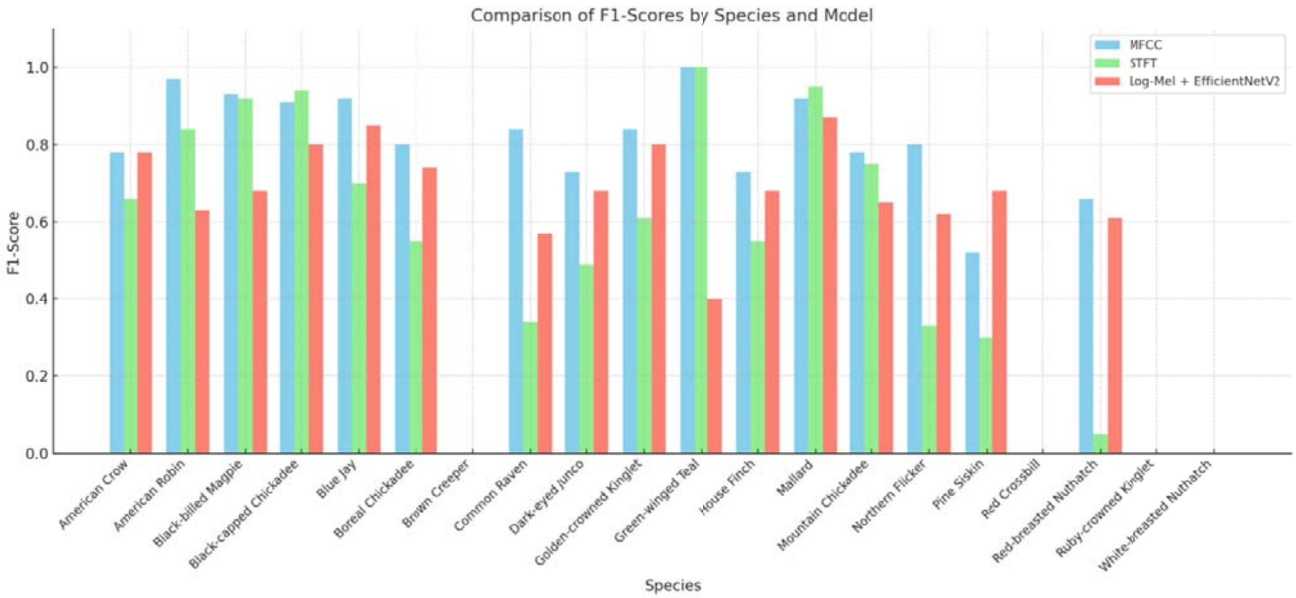
However, the results also show significant variations depending on the species: some have accuracies close to 1, while others achieve an accuracy of 0.



```
                        precision    recall  f1-score   support

           American Crow       0.92      0.82      0.86       119
          American Robin       0.77      0.90      0.83       220
     Black-billed Magpie       0.80      0.91      0.85       246
  Black-capped Chickadee       0.80      0.91      0.85       247
                Blue Jay       0.85      0.76      0.80       175
        Boreal Chickadee       0.74      0.55      0.63        56
           Brown Creeper       0.00      0.00      0.00         1
          Common Raven       0.59      0.55      0.57        31
          Dark-eyed Junco       0.68      0.68      0.68       145
  Golden-crowned Kinglet       0.83      0.78      0.80       139
       Green-winged Teal       1.00      0.25      0.40         4
            House Finch       0.92      0.53      0.68        43
                Mallard       0.86      0.89      0.87       147
      Mountain Chickadee       0.64      0.66      0.65       166
        Northern Flicker       0.71      0.56      0.62         9
             Pine Siskin       0.65      0.70      0.68        37
           Red Crossbill       0.00      0.00      0.00         3
   Red-breasted Nuthatch       0.80      0.49      0.61        75
    Ruby-crowned Kinglet       0.00      0.00      0.00         2
 White-breasted Nuthatch       0.00      0.00      0.00         6
...
                accuracy                           0.78      1871
               macro avg       0.63      0.55      0.57      1871
            weighted avg       0.78      0.78      0.78      1871
```

Table 4: Log-Melspectrograms and EfficientNetV2Model Classification Report for Bird



Species Identification

Figure 14: Comparison Graph of F1-Scores Across Different Models

The performance variations observed across models are largely explained by **data imbalance** in the training and testing sets. For instance, species such as the Green-winged Teal achieve perfect accuracy with the MFCC and STFT models, likely because only one sample was available for this species. In contrast, species like the Brown Creeper and Ruby-crowned Kinglet, which both had an accuracy of 0%, underscore the limitations of these algorithms in correctly identifying species with low sample counts or acoustic similarities to other birds.

Data imbalance significantly affects these results, as highly represented species tend to be better recognized. In the EfficientNetV2 model with log-Mel features, species such as the American Crow (0.92 accuracy) and Blue Jay (0.85 accuracy) exhibit strong recognition rates, reflecting their dominance in the dataset. Conversely, the Pine Siskin and Red Crossbill display low classification performance, likely due to a lack of training samples or complex acousticfeatures that the models struggle to differentiate.

Another important limitation in our model comparison is that the **log-Mel model** was **not trained using the same 80%-20% training-validation split** as the MFCC and STFT models. This inconsistency could influence the reported performance differences and should be considered when analyzing the results.

Overall, while some models perform well for majority or acoustically distinctive species, their effectiveness remains limited by dataset imbalance, highlighting the need for improved sampling strategies.

A. Accuracy

Model performance was assessed by analyzing training and validation accuracy curves.

The MFCC-based model demonstrated strong performance, with the training curve reaching around 90% after 15 epochs. However, the validation accuracy stagnates around 85%, and a gradual gap between the two curves is observed, suggesting the onset of overfitting. This phenomenon could be mitigated by adding regularization mechanisms such as dropout or by using early stopping.

In contrast, the STFT-based model shows training and validation curves converging around 70%, with no apparent overfitting.

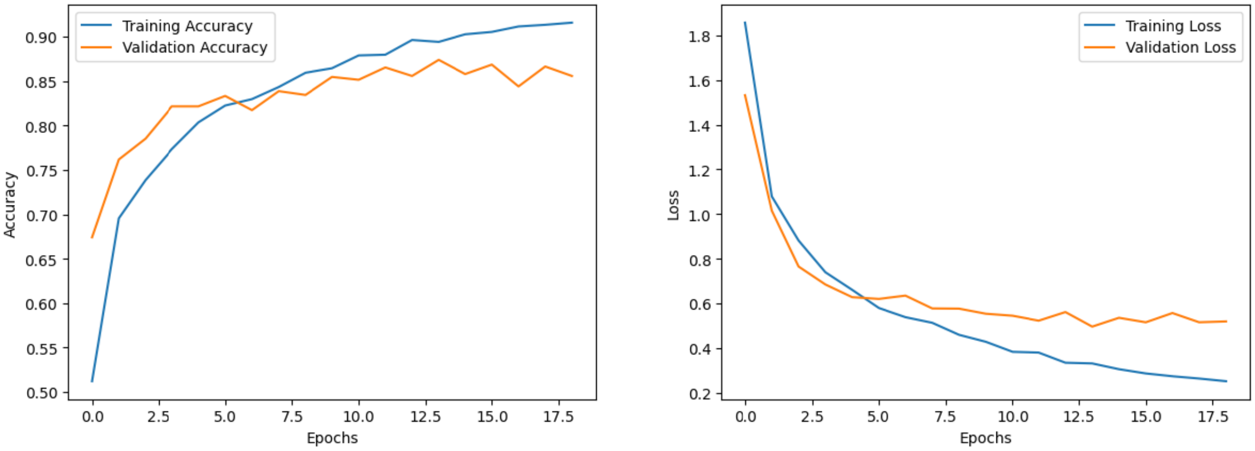


Figure 15:Accuracy Curve of the MFCC Algorithm (left),the gap observed after 10 epochs suggests overfitting.and Loss (right)

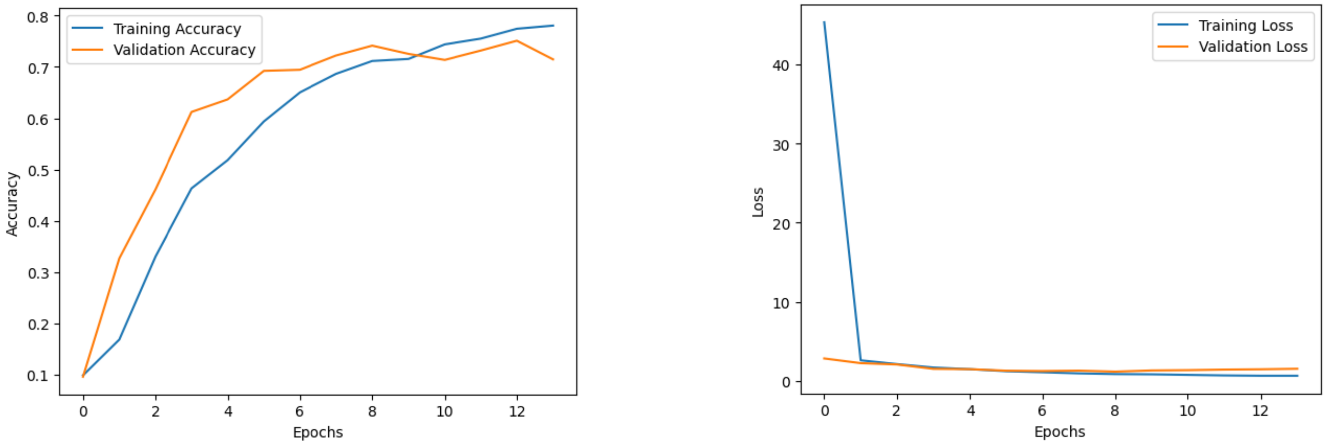
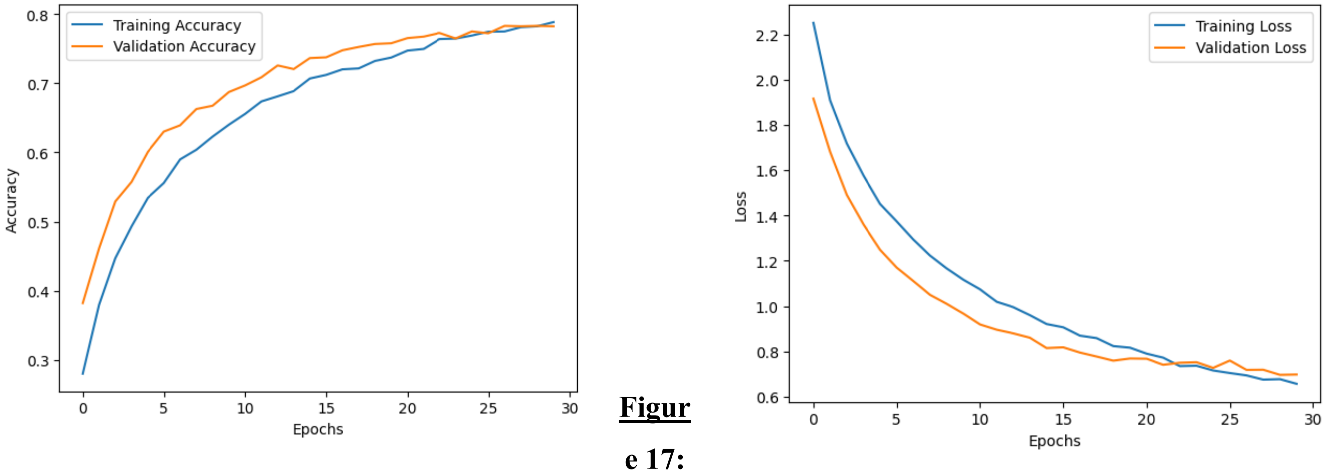


Figure 16: Accuracy Curve of the STFTAlgorithm (left) and Loss (right)



Figur e 17:

Accuracy Curve of the Log-Mel spectrograms and EfficientNetV2 Algorithm (left) and Loss (right)

However, the overall performance remains lower than that of the MFCC model, suggesting that this method captures the relevant features of bird sounds less effectively.

Finally, the log-Mel-based model using EfficientNetV2 stands out for its balanced performance: the training and validation curves almost perfectly converge around 80%.This result indicates optimal generalization and better robustness in capturing the complex features of the data, making this model particularly suitable in a context where the data is imbalanced. Thus, the log-Mel-based model with EfficientNetV2 proves to be the most promising for bird sound classification.

B. Loss

The analysis of the loss curves highlights differences in the learning performance of the three models. The MFCC-based model shows a training loss that steadily decreases to a very low value (~0.2), while the validation loss stabilizes around 0.6 after 10 epochs, with some fluctuations. This indicates the onset of overfitting, suggesting that regularization mechanisms, such as dropout or early stopping, could be beneficial.

The STFT-based model reaches near-zero training loss within the first few epochs, with a stable validation loss at a low level. This rapid drop may reflect overfitting to the training data and early saturation, possibly due to the limited complexity of features captured by the STFT.

In contrast, the model using log-Mel features with EfficientNetV2 shows both training and validation losses that decrease steadily and converge around 0.6, without significant gaps or oscillations. This convergence reflects excellent generalization, demonstrating that this model is particularly well-suited for bird sound classification.

These results highlight the robustness and effectiveness of the log-Mel approach combined with EfficientNetV2 in the context of complex and imbalanced data.

1. Confusion Matrix

The confusion matrices of the three models highlight significant differences in their ability to correctly classify the sounds of the 20 bird species. The MFCC-based model demonstrates good performance for majority classes, such as the Black-billed Magpie and Black-capped Chickadee, with 128 and 120 correct predictions along the diagonal, respectively.

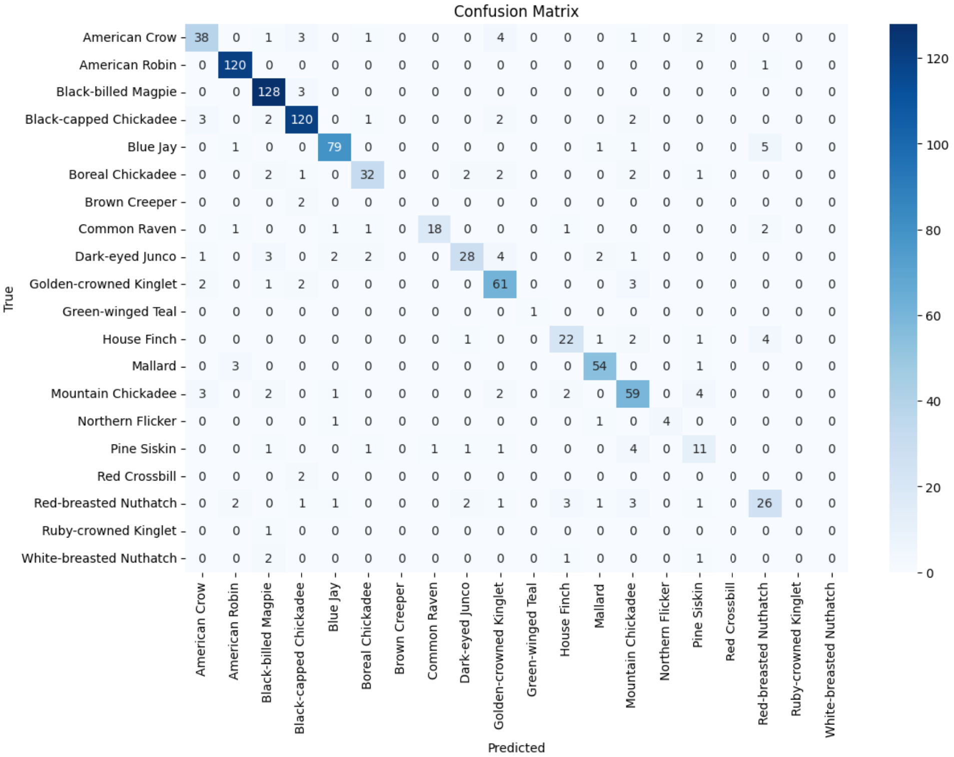
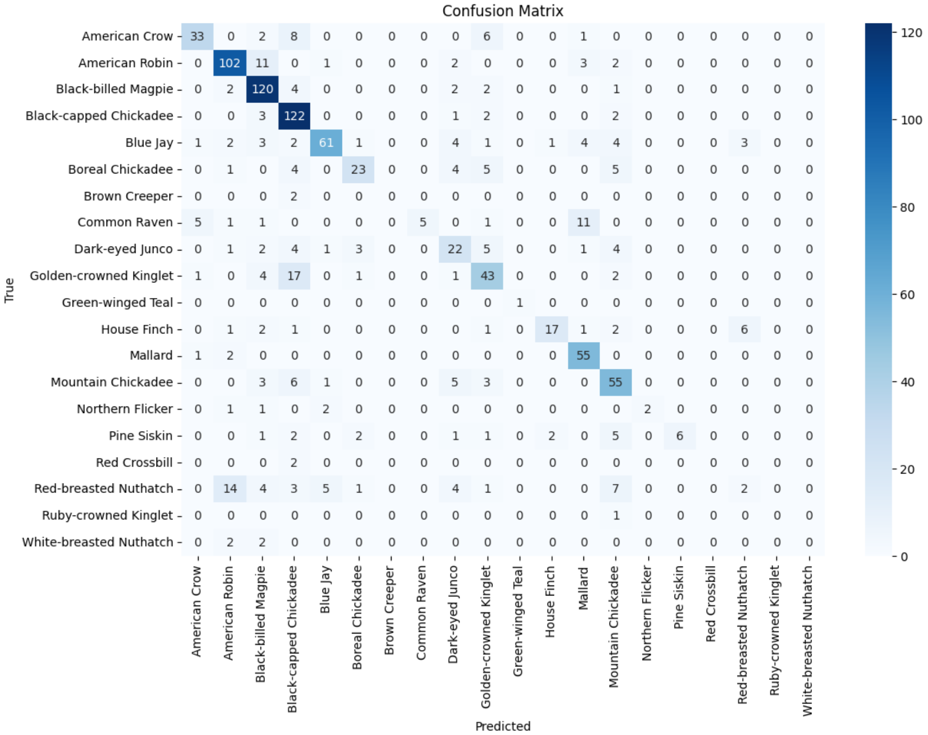


Figure 18:Confusion Matrix of the MFCC Algorithm



True \ Predicted	American Crow	American Robin	Black-billed Magpie	Black-capped Chickadee	Blue Jay	Boreal Chickadee	Brown Creeper	Common Raven	Dark-eyed Junco	Golden-crowned Kinglet	Green-winged Teal	House Finch	Mallard	Mountain Chickadee	Northern Flicker	Pine Siskin	Red Crossbill	Red-breasted Nuthatch	Ruby-crowned Kinglet	White-breasted Nuthatch
American Crow	33	0	2	8	0	0	0	0	0	6	0	0	1	0	0	0	0	0	0	0
American Robin	0	102	11	0	1	0	0	0	2	0	0	0	3	2	0	0	0	0	0	0
Black-billed Magpie	0	2	120	4	0	0	0	0	2	2	0	0	0	1	0	0	0	0	0	0
Black-capped Chickadee	0	0	3	122	0	0	0	0	1	2	0	0	0	2	0	0	0	0	0	0
Blue Jay	1	2	3	2	61	1	0	0	4	1	0	1	4	4	0	0	0	3	0	0
Boreal Chickadee	0	1	0	4	0	23	0	0	4	5	0	0	0	5	0	0	0	0	0	0
Brown Creeper	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Common Raven	5	1	1	0	0	0	0	5	0	1	0	0	11	0	0	0	0	0	0	0
Dark-eyed Junco	0	1	2	4	1	3	0	0	22	5	0	0	1	4	0	0	0	0	0	0
Golden-crowned Kinglet	1	0	4	17	0	1	0	0	1	43	0	0	0	2	0	0	0	0	0	0
Green-winged Teal	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
House Finch	0	1	2	1	0	0	0	0	0	1	0	17	1	2	0	0	0	6	0	0
Mallard	1	2	0	0	0	0	0	0	0	0	0	0	55	0	0	0	0	0	0	0
Mountain Chickadee	0	0	3	6	1	0	0	0	5	3	0	0	0	55	0	0	0	0	0	0
Northern Flicker	0	1	1	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
Pine Siskin	0	0	1	2	0	2	0	0	1	1	0	2	0	5	0	6	0	0	0	0
Red Crossbill	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Red-breasted Nuthatch	0	14	4	3	5	1	0	0	4	1	0	0	0	7	0	0	0	2	0	0
Ruby-crowned Kinglet	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
White-breasted Nuthatch	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 19: Confusion Matrix of the STFT Algorithm

However, this model struggles to generalize for underrepresented classes, such as the Common Raven and Pine Siskin, where off-diagonal confusions are significant.These errors indicate that the model favors majority classes, likely due to the data imbalance. While this model is effective for overrepresented species, it lacks robustness in handling rare species.

The STFT-based model shows slightly lower performance compared to the MFCC-based model. Although it maintains a certain level of accuracy for majority classes, such as Black-capped Chickadee (122) and Black-billed Magpie (120), it exhibits more off-diagonal confusions, particularly for species like the American Crow and Red-breasted Nuthatch. This could be attributed to the model's difficulty in extracting complex temporal and spectral features from the audio signals. This limitation affects not only the model's ability to generalize but also its overall accuracy, making it less suitable for complex or imbalanced data.

Finally, the model using log-Mel features combined with EfficientNetV2 stands out significantly in terms of overall performance. The correct predictions along the diagonal are substantially higher for almost all classes, such as Black-capped Chickadee (224), Black-billed Magpie (223), and American Robin (199). Additionally, this model handles complex and underrepresented classes, such as Dark-eyed Junco and Mountain Chickadee, with fewer off-diagonal confusions. This can be explained by the powerful combination of log-Mel features, which capture fine spectral and perceptual information, and the advanced EfficientNetV2 architecture, optimized for classifying spectrogram-derived images.This combination enables the model to generalize better and handle data imbalances more effectively.

The MFCC-based model performs well for majority classes and the STFT model offers some stability, the log-Mel model with EfficientNetV2 stands out as the most suitable for this task. It balances accuracy and generalization while reducing confusions for underrepresented classes, making it particularly robust for scenarios involving complex and imbalanced data.

2. Decisionmaking

The choice of the algorithm depends on the specific classification objectives, data constraints, and available resources. Among the three tested models, the MFCC-based model offers the highest overall accuracy, particularly for majority classes, while requiring relatively modest computational resources. This model is ideal for contexts where simplicity and execution speed are paramount.

14

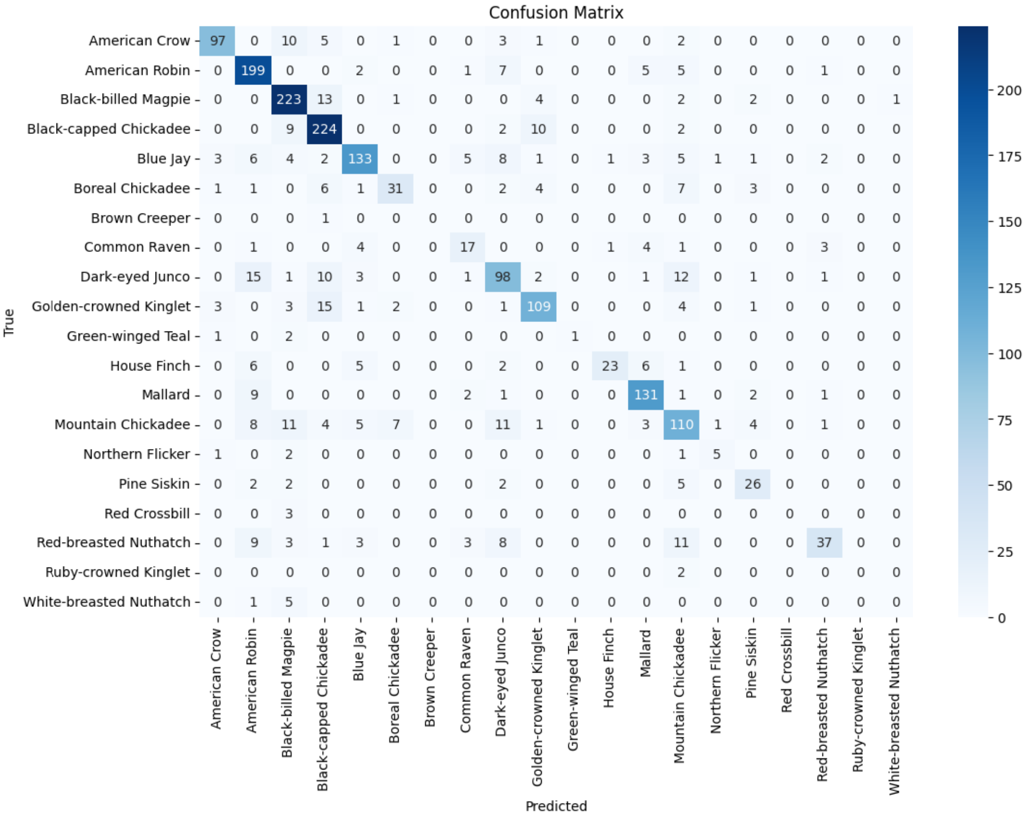


Figure 20: Confusion Matrix of the Log-Mel spectrograms and EfficientNetV2 Algorithm

The STFT-based model, despite providing a rich spatiotemporal representation, shows overall lower performance due to its increased sensitivity to class imbalances and higher computational demands.In contrast, the model combining log-Mel spectrograms and EfficientNetV2 stands out for its robustness and generalization capability, thanks to the use of a pre-trained feature extractor. This model is particularly suited for complex environments or limited data, where species diversity and accuracy are crucial.

Therefore, the choice of the algorithm should be guided by a balance between accuracy, robustness, and operational constraints, with EfficientNetV2 being recommended for applications requiring strong generalization and diverse recordings.

IX. DISCUSSION

The results obtained in this study show that deep learning methods applied to the classification of bird songs offer promising performance while highlighting several challenges inherent to this type of task. The combination of MFCCs, STFT, and log-Mel spectrograms with neural network architectures, such as EfficientNetV2, allowed us to evaluate the advantages and limitations of these approaches (Das et al, 2023).

Among the three tested algorithms, the MFCC-based model proved to be the most effective in terms of overall accuracy (85%), demonstrating a better ability to capture the essential acoustic features of bird songs. However, this method remains limited by a reduced ability to differentiate underrepresented species or species recorded in noisy environments, such as the Red-breasted Nuthatch or the Ruby-crowned Kinglet.

The STFT-based model, although effective for certain specific species, showed overall lower performance (average accuracy of 71% and F1-score of 51%). This result can be attributed to the difficulty of fully exploiting complex spatiotemporal information, combined with increased sensitivity to class imbalance.

In contrast, the model combining log-Mel spectrograms and EfficientNetV2 demonstrated greater robustness to spectral and temporal variations, largely thanks to the pre-trained feature extractor, as previously observed by Das et al (2023). Despite slightly lower performance compared to the MFCC model (average accuracy of 78%), this approach stands out for its ability to generalize across complex and diverse data.

Our analyses also highlighted the critical importance of data quality and balance (Reza 2024; Stowell &Plumbley, 2014). Data cleaning and augmentation steps significan improved the algorithms' performance, particularly for the MFCC model, whose accuracy increased from 66% to 83%. However, class imbalance remains a major challenge, especially for rare or underrepresented species (Rezaul, 2024). Adjustments such as reducing the number of recordings for overrepresented species while preserving rare species helped to limit biases without altering the ecological diversity of the dataset. These results underscore the importance of tailored strategies to manage these imbalances.

Moreover, although automated deep learning-based approaches have proven effective, they require a large volume of high-quality data to achieve optimal performance (Zhang et al., 2018). The algorithms used in this study showed limitations in contexts of low class representativity or significant noise in the recordings. Additional techniques, such as synthetic data augmentation or the use of alternative acoustic representations (alternatives to STFT), could be explored to further improve performance (Das et al, 2023).Furthermore, other improvements could have been implemented, such as adding class weights to the STFT and log-Mel models with EfficientNetV2, as well as using a common training set (80%-20%) to better compare our algorithms.

The main limitation of our algorithms is that they detect only one bird species at a time. It would be relevant to design a model capable of detecting multiple species simultaneously (Springer et al., 2013).When an audio recording contains multiple bird species, the algorithm arbitrarily selects one species without explicitly defining the underlying criteria. This choice could be influenced by factors such as the dominant frequency or the prevalence of certain species in the training set.

X. Conclusion

This study implemented deep learning approaches to classify the songs of 20 bird species within the Cochrane Ecological Institute reserve. By using three main techniques—MFCC, STFT, and log-Mel spectrograms combined with EfficientNetV2—we compared their performance in terms of accuracy, generalization, and the ability to handle class imbalances.

The MFCC-based model stood out for its overall accuracy, demonstrating its effectiveness in capturing essential acoustic features. The EfficientNetV2 model, on the other hand, showed better generalization due to its pre-trained architecture, while the STFT model presented more limited results. The data cleaning and augmentation steps played a key role in improving the quality and diversity of the dataset, thereby increasing the models' accuracy.

These results highlight the importance of data quality, managing class imbalances, and tailoring methods to the specificities of acoustic recordings. By combining robust approaches with appropriate data preparation techniques, this study paves the way for effective automated tools for bird song recognition.

Future work could focus on enabling multi-species recognition from single recordings, real-time audio processing, and integrating environmental data to enhance model performance. These advancements would further support biodiversity monitoring and conservation efforts globally.

XI. BIBLIOGRAPHY

Brooker, S. A., Stephens, P. A., Whittingham, M. J., & Willis, S. G. (2020).

Automated detection and classification of birdsong: An ensemble approach.

Ecological Indicators, *117*, 106609.

Retrieved September 27, 2024, fromhttps://doi.org/10.1016/j.ecolind.2020.106609

Canada, Environment and Climate Change. (June 27, 2012).

Trends in Canada's bird populations.

Retrieved September 9, 2024, fromhttps://www.canada.ca/fr/environnement-changement-climatique/services/indicateurs-environnementaux/tendances-populations-oiseaux.html

CochraneEcologicalInstitute. (n.d.).

CEI Wildlife: Evidence-based conservation solutions.

Retrieved September 27, 2024, from https://ceiwildlife.org

Cornell Lab of Ornithology. (n.d.).

Merlin Bird ID - Free, instant bird identification help and guide for thousands of birds.

Retrieved October10, 2024, fromhttps://merlin.allaboutbirds.org

Das, N., Padhy, N., Dey, N., Bhattacharya, S., & R.S. Tavares, J. M. (2023).

Deep Transfer Learning-Based Automated Identification of Bird Song.

International Journal of Interactive Multimedia and Artificial Intelligence, *8*(4), 33.

Retrieved November10, 2024, from https://doi.org/10.9781/ijimai.2023.01.003

Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., & Fan, H. (2020).

Heart sound classification based on improved MFCC features and convolutional recurrent neural networks.

Neural Networks, *130*, 22-32.

Retrieved October15, 2024, from https://doi.org/10.1016/j.neunet.2020.06.015

Fox, E. (2008).

Call-independent identification in birds.

Kaggle. (2023).

BirdCLEF 2023.

Retrieved September24, 2024, fromhttps://www.kaggle.com/competitions/birdclef-2023

Kaggle. (2024).

BirdCLEF24: KerasCV starter [Train].

Retrieved November25, 2024, fromhttps://www.kaggle.com/code/awsaf49/birdclef24-kerascv-starter-train/notebook

Puget, J. F. (2021).

STFT Transformers for Bird Song Recognition.

In *CLEF (Working Notes)* (pp. 1609-1616).

Retrieved October15, 2024, fromhttps://ceur-ws.org/Vol-2936/paper-137.pdf

Rezaul, K. M., Jewel, M., Islam, M. S., Siddiquee, K. N. E. A., Barua, N., Rahman, M. A., ... &Asha, U. F. T. (2024).

Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models.

International Journal of Advanced Computer Science and Applications, *15*(7), 37-53.

Retrieved September25, 2024, fromhttps://thesai.org/Downloads/Volume15No7/Paper_4-Enhancing_Audio_Classification_Through_MFCC.pdf

Scikit-Learn Developers. (n.d.).

OneHotEncoder.

Retrieved September22, 2024, fromhttps://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

Springer, J., Duan, Z., & Pardo, B. (2013).

Approaches to multiple concurrent species bird song recognition.

In *The 2nd International Workshop on Machine Listening in Multisource Environments*.

Retrieved November10, 2024, fromhttps://interactiveaudiolab.github.io/assets/papers/birdsong_recognition_pardo.pdf

Stowell, D., &Plumbley, M. D. (2014).

Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning.

PeerJ, *2*, e488.

Retrieved October15, 2024, fromhttps://doi.org/10.7717/peerj.488

Tang, Y., Liu, C., & Yuan, X. (2024).

Recognition of bird species with birdsong records using machine learning methods.

PLOS ONE, *19*(2), e0297988.

Retrieved October3, 2024, fromhttps://doi.org/10.1371/journal.pone.0297988

Wildlife Acoustics. (n.d.).

Kaleidoscope: Bioacoustics sound analysis software.

Retrieved September24, 2024, fromhttps://www.wildlifeacoustics.com/products/kaleidoscope

Xie, S., Lu, J., Liu, J., Zhang, Y., Lv, D., Chen, X., & Zhao, Y. (2022).

Multi-view features fusion for birdsong classification.

Ecological Informatics, *72*, 101893.

Retrieved October21, 2024, fromhttps://doi.org/10.1016/j.ecoinf.2022.101893

Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2018).

A survey on deep learning for big data.

Information Fusion, *42*, 146-157.

Retrieved September24, 2024, fromhttps://doi.org/10.1016/j.inffus.2017.10.006