IMPROVING GENERALIZATION CAPABILITIES OF FEW-SHOT LEARNING MODELS FOR AVIAN VOCALIZATIONS

GIULIANO SERGIO

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

799233

COMMITTEE

Dr. Dan Stowell

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 17, 2025

WORD COUNT

8029 words
(exluding the title page, acknowledgements, Data Source, Ethics, Code, and Technology statement, references, self-reflection, appendices, tables, and figures)

ACKNOWLEDGMENTS

I am very proud to share the research I have been working on over the past few months. This project could not have been optimized like it is without the help of certain people. Firstly, I would like to thank Dr. Dan Stowell, who supervised me throughout this process and helped me with finding a fitting project. Furthermore, I would like to thank Céline Angonin and Ben McEwen for their help to overcome challenges that were faced during this project. Lastly, I wanted to thank the authors of the datasets, which are used and credited in this paper.

IMPROVING GENERALIZATION CAPABILITIES OF FEW-SHOT LEARNING MODELS FOR AVIAN VOCALIZATIONS

GIULIANO SERGIO

Abstract

Animal monitoring can be automated using AI models. However, a lot of AI models require lots of labeled data to perform well. Furthermore, in the field of animal data, not a lot of data is labeled and there are many endangered species, which make it difficult to gather large amounts of data. A solution for this problem is to use few-shot learning (FSL) models. These models learn to learn optimally with very limited data. Currently, there is no consensus on how to optimize the generalization of FSL models in the field of bioacoustics. The current study investigated this using bird vocalization datasets with the following research question: "How can the generalization of few-shot learning models be improved for birdsong classification?" How many samples are needed for these models to generalize well, whether training on more datasets will enhance generalization, how different FSL models generalize, and whether the FSL models are able to outperform a supervised trained CNN baseline model was examined. The results demonstrate that using ten samples per class significantly enhances generalization compared to using five samples. Furthermore, the performance difference between Prototypical Networks and Optimization-Based Meta-Learning models differed based on the other parameters used for these models. Additionally, using more data complexity only enhances generalization when the used model is capable of working with more complex data. Lastly, the FSL models did not outperform a model that trained like a supervised learning model. These results contribute to enhancing the generalization of FSL models in the field of bioacoustics.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The current study has used multiple open source datasets from "Zenodo.org". All authors are credited and referred to in this paper. Furthermore, all images used in the current paper are made by the author himself. Claims made by other studies also all contain references to the original papers. Additionally, there are some tools used to enhance the end product of the current paper. Firstly, "WordVice" is used to detect grammatical and spelling errors. Furthermore, when encountering errors or bugs in the code used for the current paper, "ChatGPT 4.0" is used to detect these errors and to help with fixing them. Lastly, to develop the code, there are multiple libraries used. All these libraries can be found in the README of the code used for this paper.

2 INTRODUCTION

2.1 *Context*

Animal monitoring is an important process for ensuring the well-being of animals. Not only could this process capture where different animals live, it could also estimate how many species live in a specific environment (Congdon et al., 2022). There are multiple ways of monitoring animals. Examples of these methods are: human observations or using wearable devices, such as a tracking GPS (Congdon et al., 2022; Michielon et al., 2024). However, these methods have a major limitation: they require manual observations. This requires a lot of time, and using wearable devices could affect the natural behavior of animals (Congdon et al., 2022; Michielon et al., 2024). Automating this process with machines, such as microphones and microcontrollers, has shown to be a promising method for monitoring animals (Segura-Garcia et al., 2024).

Additionally, monitoring animals with Artificial Intelligence (AI) systems is an automated machine based method that has demonstrated promising results (Segura-Garcia et al., 2024). The field of research has shown that monitoring animals with AI models reduces processing time and enhances the precision of identifying animals compared to human annotators (Samiappan et al., 2024). Furthermore, the advantage of using AI models instead of other machines is their ability to make independent decisions without the constant need for human assistance (Dupuis-Desormeaux, Davidson, Mwololo, Kisio, & MacDonald, 2016; Samiappan et al., 2024). Since this differentiates AI models from other machines, this is the aspect that is focused on in this study to optimize animal monitoring.

Previous studies have investigated classifying animals with different AI models. Supervised learning models, for example, are used to classify animals using their sounds. The study by Samiappan et al. (2024) demonstrated that AI models are capable of classifying animals by their sounds. However, these supervised learning models require lots of labeled data to perform well (X. Li, Jia, Islam, Yu, & Xing, 2020). This can be considered as a bottleneck, since a lot of data available on animals is not labeled (Agilandeeswari & Meena, 2023; Samiappan et al., 2024). Additionally, there are many species with small populations, which makes it very challenging to gather a lot of information (McEwen et al., 2024). These aspects make it relevant to consider using different AI methods for animal classification.

These limitations of supervised learning have been investigated before and a potential solution for this problem of a lack of labeled data is few-shot learning (FSL). These models learn to learn efficiently, which allows them to find patterns in data using very little samples (Wang, Yao, Kwok,

& Ni, 2020). A FSL model is very functional for datasets that do not have many labeled data like a lot of animal datasets (Agilandeeswari & Meena, 2023; Samiappan et al., 2024). This is shown in a previous study that investigated using FSL to detect animals in bioacoustics data (Nolasco et al., 2023). Not only did these models use very few samples, but most of the FSL models even outperformed supervised learning models that trained on more data (Nolasco et al., 2023). This demonstrates the potential of using FSL models for bioacoustics tasks, like animal classification.

However, FSL models are still not optimized for these tasks. The study by Nolasco et al. (2023) found that the performance of these models declined when the used data used for learning was not representative of the entire class. This entails that when the generalization of animal species was not captured fully, it decreased the performance of the models. Furthermore, research has demonstrated that the generalizability of FSL models enhances when the data it learns from is more diverse, which can be realized by using different datasets for a single model (Roy, 2024). Thus, these studies emphasize the importance of enhancing generalization for FSL models to improve them. Investigating how generalizability can be improved in the field of bioacoustics, currently has no conclusive answer. That is why investigating this will provide meaningful information about how to improve generalization for FSL models.

2.2 *Research goal*

The current study investigates the knowledge gap of improving generalization from FSL models in the field of bioacoustics. This is done by classifying birdsongs based on different audio datasets with FSL models, because previous research has demonstrated that AI models are able to classify birdsongs and there is a lot of data available on birdsongs (Rauch et al., 2024; Segura-Garcia et al., 2024).

Since current FSL models may struggle with generalizing classes, it is relevant to investigate how many samples should be used for these models to perform optimal. Additionally, different FSL models are investigated in the current research. Since a gap in the field of research is investigated regarding FSL models, testing on multiple FSL models will capture the impact on these models better. The current study also investigates whether adding more diverse training data will lead to improvements in generalization. Furthermore, FSL models will be compared to the performance of a baseline model that trains like a supervised learning model to investigate if training episodically improves generalizability of AI models.

2.3 *Research question*

How can the generalization of few-shot learning models be improved for birdsong classification?

2.4 *Sub-questions*

a. *How does the performance of few-shot learning models differ to supervised learning models in bird classification?*

b. *How many samples are required for few-shot learning models to classify birds?*

c. *How does training on more datasets impact the generalization of few-shot learning models for bird classification?*

d. *How do different few-shot learning methods compare in their performance of classifying birds?*

Investigating these sub-questions provides an answer to the research question of the current research. The findings of this study will provide evidence on how to enhance the generalization of FSL models. This can aid in solving the problem of the absence of labeled data for animals, since FSL models do not require much labeled data. This can lead to improving animal monitoring using AI models in general.

2.5 *Findings*

The results of the current study demonstrate that the generalization of FSL models can be improved using more samples per class to train and test on. Furthermore, adding more data complexity can enhance generalization, depending on the model that is used. Additionally, the performances of Prototypical Networks (ProtoNets) and Optimization-Based Meta-Learning (MAML) models are dependent on the other parameters of these models, which impact which of these models is better in generalizing bird vocalizations. Lastly, the baseline model that used the training method cross-entropy, outperformed all FSL models when they were all validated and tested on unseen classes. This indicates that training like a supervised learning model is beneficial to improve generalization.

3 RELATED WORK

As mentioned, the current study investigates the impact of different methods to improve generalizability of birdsong classification using FSL models.

To achieve this, the field of research relating to this topic is explored to gain insights into the state of the art and the gaps of this field.

3.1 *FSL models compared to supervised learning models*

To get a better understanding of the impact of good performing FSL models in the field of AI for bioacoustics, it is important to investigate the differences in performances between different types of AI models. The field of research has already examined different types of AI models to classify animals. The study by Segura-Garcia et al. (2024) created a CNN model that was able to predict 72.95% of the times correctly which audio fragment belonged to which bird. Since there were 41 specific species the model could choose, this model performed significantly better than randomly guessing. This study demonstrates that supervised learning models can successfully learn patterns from audio data from birds.

Additionally, there are also FSL models that are successful in this task. The study by Moummad, Farrugia, and Serizel (2024a) shows that their meta-learning FSL model (FO-Meta-Curvature) was able to classify birds using their audio in a 5-way, 1-shot setting with an accuracy of 61.34%. This study demonstrates that FSL models are also able to learn differences between birds using their vocals. Since there are multiple types of AI models that are successful for this task, it is relevant to compare FSL models with a supervised learning model to demonstrate the impact of FSL for bird classification for Deep Learning (DL).

The main difference between supervised learning models and FSL models is the way these models are trained. Supervised learning models train on entire datasets using batches to update the model and cross-entropy as the loss function (Segura-Garcia et al., 2024). On the other hand, FSL models train episodically, meaning that they train with mini datasets of a few classes containing a few images for it to train on, called shots, and to test on, called query samples (Wang et al., 2020). This is done with N-way-K-shot classification, where the N-way stands for the number of classes that are present in the dataset, and the K-shot stands for the number of samples that will be used to train on per class. Furthermore, FSL models also use feature extractors, called their backbone (Y. Hu, Pateux, & Gripon, 2022). These backbones use the raw input and transform it into valuable data for the models. In the current research, an episode can be seen as a small classification task of five different bird species using their sounds. Thus, while supervised learning models train on entire datasets, FSL models train with a lot of mini datasets to try to enhance generalization.

The differences in performance between these models on a task where data is limited has been investigated before. These studies reveal that when

data is limited, FSL models will outperform supervised learning models (Snell, Swersky, & Zemel, 2017; Wolters, Sizemore, Daw, Hutchinson, & Phillips, 2021). The study by Snell et al. (2017) revealed that when classical models were trained with batches, but tested on 5-way 5-shot tasks, they barely performed higher than chance level (25%). In contrast, ProtoNets performed significantly better at this task (50%). However, the field of research demonstrates that FSL models do not always outperform baseline models. When backbones of FSL models are deep, and they train and test with five or more shots, then baseline models perform competitively with FSL models (Chen, Liu, Kira, Wang, & Huang, 2019).

The differences between baseline and FSL models have not been explored thoroughly in the field of bioacoustics. Furthermore, research demonstrated that baseline models perform differently depending on the way they are designed (Chen et al., 2019; Snell et al., 2017). Thus, investigating the differences in generalization abilities between baseline and FSL models will provide more insights in the field of bioacoustics.

3.2 *The right number of samples for FSL models*

In order to create good performing FSL models, it is important to optimize the hyperparameters of these models. FSL models train on one or multiple datasets to optimize their ability to generalize (Wang et al., 2020). They use this to quickly adapt to patterns from unseen tasks without needing a lot of data. Different FSL models use various techniques, but in essence all FSL models aim to generalize optimally without requiring much data. For classification tasks, such as the current paper, this is done using episodically learning with N-way-K-shot classification, as previously mentioned (Wang et al., 2020).

The field of research demonstrates the importance of choosing the right number of samples, called shots, for each class. The study by Cao, Law, and Fidler (2019) found that FSL models perform the best when the number of shots used for training matches with the number of shots used for testing on a new dataset. Additionally, differences in performance and generalizability are also caused by selecting a specific number of samples for training and testing. The study by Laenen and Bertinetto (2021) found that FSL models that train and test with five shots per given class outperform models that use a single shot. Training with five shots leads to improvements in generalizability for FSL models including ProtoNets. Thus, adding more shots seems to improve the generalizability of FSL models. The study by S. X. Hu, Li, Stühmer, Kim, and Hospedales (2022) supports this claim by demonstrating that a 1-shot FSL model had an accuracy of 95.3%, while a 5-shot FSL model had an accuracy of 98.4%.

However, the performance improvements seem to become smaller when the number of shots added increases (Song, Wang, Cai, Mondal, & Sahoo, 2023). This demonstrates the importance of choosing the right number of shots.

To the extent of the researcher's knowledge, there are no studies that have experimented with using different shots for FSL on bioacoustics data at the time of writing. This gap makes it interesting to investigate different shots in the current study. Using five shots has proven to be an effective choice according to previous studies (S. X. Hu et al., 2022; Laenen & Bertinetto, 2021). Additionally, since adding more shots seems to improve the generalizability, using ten shots might be a good choice when the number of used data is limited and the generalizability might improve significantly.

3.3 *Data complexity to enhance generalizability of FSL models*

In the current field of research, there are several studies that investigated FSL models on bioacoustics data (Moummad et al., 2024a; Nolasco et al., 2023). These studies demonstrate that FSL models are capable to work with and effective for bioacoustics data. However, the generalizability needs to be optimized to create good performing FSL models (Moummad et al., 2024a). Nolasco et al. (2023) acknowledge the importance of generalizability in their study as well, and recommend using multiple datasets to improve this.

The idea of enhancing the generalizability by using multiple datasets is supported by the study by Zhang et al. (2024). They investigated this by testing the performances of deep learning (DL) models that were trained on different data. This study demonstrated that the models that used more complex and diverse data to train on were significantly better at generalizing compared to models that trained on simpler data. The findings of this study show that training on diverse data will lead to improvements in generalization, which is beneficial for FSL models.

However, the study by Sendra-Balcells et al. (2022) argues that combining datasets may not be the best way to increase generalizability for DL models. This study found that, even though using multiple datasets does improve generalizability, using transfer learning and/or data-augmentation is more efficient, and improves the generalizability even more. While this study demonstrates that methods like transfer learning and data-augmentation are very effective methods that should be investigated, other studies prioritize adding additional datasets (Nolasco et al., 2023; Sendra-Balcells et al., 2022; Zhang et al., 2024). Since there is no consensus among

studies on how to improve the generalization of DL models, it is relevant to investigate this in the current study.

Furthermore, the study by Van Merriënboer, Hamer, Dumoulin, Triantafillou, and Denton (2024) has analyzed the importance of generalizing in the field of bioacoustics. They found that AI models struggle working with bioacoustics data when these models are not capable of generalizing. This is due to the bioacoustics data containing differences, such as geographical influences and differences in recording devices. Additionally, this paper also mentions that single species can also produce different sounds, which makes it important to create an AI model that is able to generalize well. This way, the model can adapt to these differences.

Thus, the current field of research highlights the importance of creating generalizable AI models when analyzing bioacoustics data (Van Merriënboer et al., 2024). Adding more data diversity has proven to be an effective method to enhance generalizability (Zhang et al., 2024). However, not all studies fully agree with this claim (Sendra-Balcells et al., 2022). That is why investigating this in the current study will provide useful insights for this field of research.

3.4 *Different FSL methods*

Using FSL models on bioacoustics data has proven to be an effective method. FSL models are a great option when there is limited data available (McEwen et al., 2024). Additionally, they are also able to outperform supervised learning models that use more data to train on (Nolasco et al., 2023). Thus, the field of research has demonstrated that FSL is a great option when it is trained on bioacoustics data. However, there are a lot of different FSL models that can differ in their performances based on the given task (Chen et al., 2019; Parnami & Lee, 2022). The study by Chen et al. (2019) tested different FSL models on a classification task and found differences in performances based on the model used and the dataset domain. Thus, it is important to investigate which FSL methods are suitable for bird classification.

As there are many FSL methods, the current study focuses on two types. Metric-Based Meta-Learning models and Optimization-Based Meta-Learning models were selected, because they have effectively been used to classify birds using their vocalizations (Anderson & Harte, 2021; Moon, Kim, Hwang, & Hwang, 2023). Firstly, Metric-Based Meta-Learning models learn by creating a distance function that identifies new classes with examples that are used for learning (Parnami & Lee, 2022). An example of a Metric-Based Meta-Learning model is a Prototypical Network. Prototypical Networks learn to create embeddings to separate different

classes of the dataset (Parnami & Lee, 2022). The study by Anderson and Harte (2021) demonstrated that Prototypical Networks (ProtoNets) are an effective method to use to classify birds using their vocalizations. That is why ProtoNets seem to be promising models to use for the current study.

Additionally, Optimization-Based Meta-Learning models are FSL models that use a different method to learn to adapt to different tasks quickly. Optimization-Based Meta-Learning Models realize this by optimizing their parameters so that they can be changed very quickly when they are trained for new tasks (Parnami & Lee, 2022). Instead of creating optimal embeddings like Prototypical Networks do, these models optimize their initial parameters to adapt them quickly for unseen tasks. A Model-Agnostic Meta-Learning (MAML) model is an example of a model like this. MAML models also have demonstrated to be an effective method for bird classification. The study by Moon et al. (2023) classified birds using their sounds. Even though MAML models struggle with classifying sounds that differ a lot from each other, this study reveals that MAML models are very effective for bird classification, and they even outperformed traditional CNN models. Thus, the current field of research demonstrates that ProtoNets and MAML models are effective models for bird classification.

Furthermore, previous studies have demonstrated that using transfer learning combined with FSL models improves their performance and generalizability (Lu et al., 2022; Sendra-Balcells et al., 2022). The study by Lu et al. (2022) demonstrates that using transfer learning combined with outperformed FSL models without transfer learning and traditional supervised learning models, by 5% to 30%. Thus, implementing transfer learning also seems to be a very effective method.

The current field of research demonstrates that ProtoNets and MAML models are effective for birdsong classification (Anderson & Harte, 2021; Moon et al., 2023). Additionally, using transfer learning with these FSL models also has proven to be effective in enhancing their performance (Lu et al., 2022; Sendra-Balcells et al., 2022). Thus, using ProtoNets and MAML models with transfer learning seem to be good options for birdsong classification. This is why using these models will provide valuable insights when the impact of their generalizability is investigated. Furthermore, at the time of writing, there are not any studies that have investigated the performance differences of ProtoNets and MAML models for bird species classification using their vocalizations.

4 METHOD

4.1 *Datasets*

The entire process of the methodology of the current study can be seen in Figure 1, and all the code used to realize this study are available through this link. Firstly, there were different datasets collected to train and test the models on. Since FSL models are used to generalize on limited unseen data of unseen classes, it was important to gather datasets that differed in the bird species that were included in the datasets. To achieve this, five datasets had been gathered that include annotated bird vocalizations (Clapp et al., 2023; Kahl, Charif, & Klinck, 2022; Kahl, Wood, Chaon, Peery, & Klinck, 2022; Vega-Hidalgo et al., 2023; Weldy et al., 2024). The details of these datasets can be seen in Table 1. In these datasets, there were no duplicate classes between the train, validation, and test set. The specifications of the species that were included in the train, validation and test set can be found in Appendix A. Additionally, classes that contained less than 20 annotations were excluded for this experiment. When these models were trained with ten training shots and ten query samples, it was necessary to include at least 20 samples per class. With these adjustments on the datasets, the classes could be properly used for the FSL models. The datasets created for this study are available here.

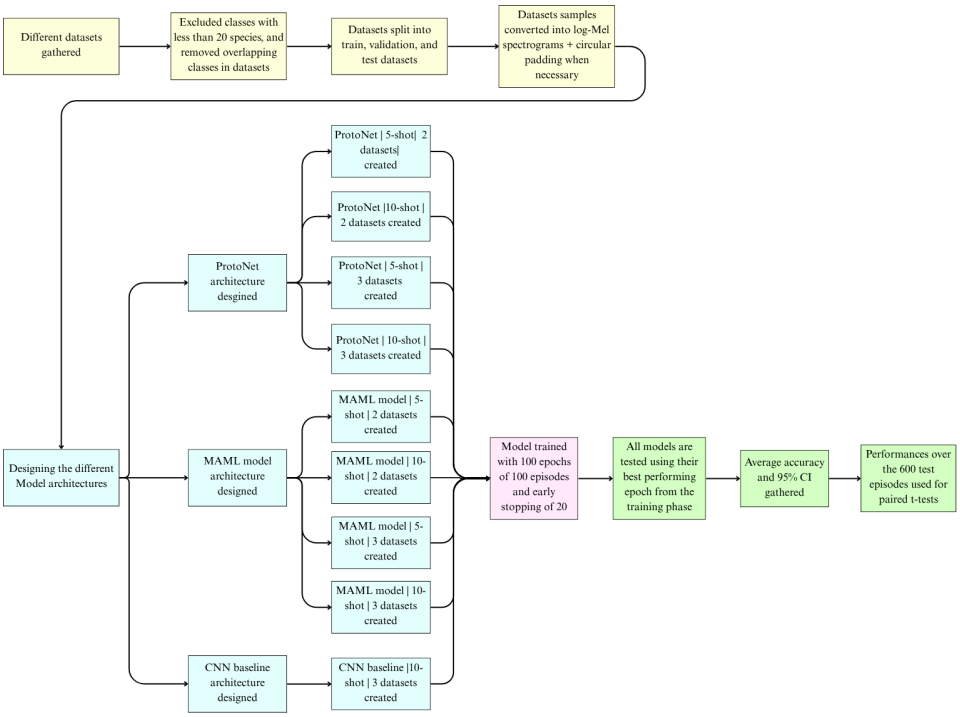

Figure 1: The flowchart of the methods of the current research. The yellow blocks represent the preprocessing, the blue blocks represent the model architecture design process, the purple block represent the training process, and the green blocks represent the evaluation process.

Table 1: Datasets specifications

Dataset	Classes	Total used annotations	Used as
Clapp et al. (2023)	17	10057	Training dataset
Kahl, Wood, et al. (2022)	44	14963	Training dataset
Kahl, Charif, and Klinck (2022)	74	44640	Training dataset
Vega-Hidalgo et al. (2023)	45	6489	Validation dataset
Hopping, Kahl, and Klink (2022)	70	13836	Test dataset

After the datasets were gathered, fitting data for the FSL models was created. To make the audio segments readable for AI models, they were converted into log-Mel spectrograms. In the field of FSL, these spectrograms and have proven to be an effective way of converting audio segments into interpretable data (Won & Kim, 2024). In the field of avian research, it is common to sample bird sounds between two and five seconds that all have the same length to contain consistency in the data patterns (Hexeberg, Chitre, Hoffmann-Kuhnt, & Low, 2025; Moummad, Farrugia, & Serizel, 2024b; Xie & Zhu, 2023). That is why all log-Mel Spectrograms convert

audio segments of 3.2 seconds. When segments were longer than 3.2 seconds, then the first part of this clip is used. When clips were shorter than 3.2 seconds, then circular padding was used. This is a method that was used in the study by Moummad et al. (2024b) in which, shorter clips are repeated until it fully filled the fixed time of 3.2 seconds. This ensured that each spectrogram contained insightful information of the class it represented. Furthermore, the colormap 'viridis' was used, since this colormap has proven to be effective for birdsong classification in previous research (Gibbons, King, Donohue, & Parnell, 2024). An example of a log-Mel spectrogram used in the current study can be seen in Figure 2.

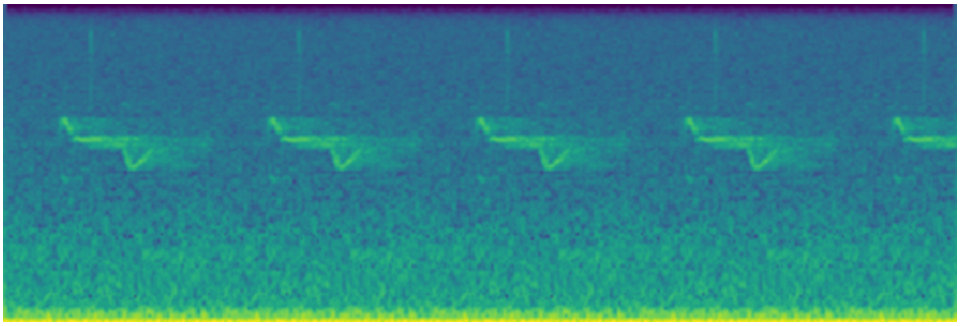

Figure 2: An example of the used log-Mel Spectrograms

4.2 *Fixed parameters*

After the data was successfully preprocessed, the models were built. For the current study, there were two types of FSL models built: ProtoNets and MAML models. Both networks use a pretrained ResNet18 model as their backbone. ResNet18 was trained on more than a million real-world images from ImageNet, containing animals (He, Zhang, Ren, & Sun, 2016). The last two layers of this backbone have been unfrozen to finetune during training. Unfreezing layers is a common method used in the field of FSL to quickly learn patterns without needing to train a backbone from scratch (Abbas, 2023). A transfer learning backbone has been chosen, because previous research has shown that this would enhance generalization (Sendra-Balcells et al., 2022). Furthermore, when ResNet18 is used as a transfer learning model, it has demonstrated to be successful in classifying animals using their sounds or visual features (Al Dawasari, Bilal, Moinuddin, Arshad, & Assaleh, 2023; Wei, Hossain, & Ahmed, 2022). This is why ResNet18 is a suitable transfer learning backbone for classifying birds using their sounds.

Furthermore, these models all contain Adam as their optimizer, since previous FSL studies have proven that this is a suitable optimizer for these models (Boudiaf et al., 2020; Snell et al., 2017). Additionally, research has

demonstrated that using a learning rate of 1-e4 leads to the best performances when using Adam with transfer learning (Herath, Meedeniya, Marasingha, & Weerasinghe, 2022). That is why this has been applied for all the models in the current study. All models train on 100 epochs that contain 100 episodes, since these parameters have been effective in previous studies in the field of FSL (Gao, Fei, Liu, Lu, & Xiang, 2021; Joshi, Mundra, & Mundra, 2025; Walsh, Abdelpakey, Shehata, & Mohamed, 2022). Furthermore, when the model did not improve after 20 epochs, it is considered that the model plateaued, and the model then stopped training earlier to avoid unnecessary computation. This is also an effective method used in the field of FSL to enhance the training process (Ye, Ming, Zhan, & Chao, 2022). With these parameters and methods set, the different models were created.

4.3 *Architectures from the models*

After the fixed parameters were chosen, it was possible to build the different types of FSL models. Firstly, the ProtoNets use the same architecture as the original ProtoNet from the study by Snell et al. (2017). The architecture used in the current study can be found in Figure 3. The ProtoNets of the current study train episodically: per epoch 100 episodes, each with five classes with ten query samples per class in each episode. Additionally, these models are trained to separate these classes with either five or ten samples per class, and these samples are embedded using the backbone of the model. Following this, the embeddings of each class were averaged to create one prototype of each class. Then, the query samples were used to test the effectiveness of these prototypes by embedding them as well and assigning them to the class which the embedding is the closest to using Euclidean distance. A visualization of this process can be seen in Figure 4. Additionally, for all the models created in the current study, the parameters of the best performing epochs are used, since this is the common way of training FSL models effectively (Chen et al., 2019).

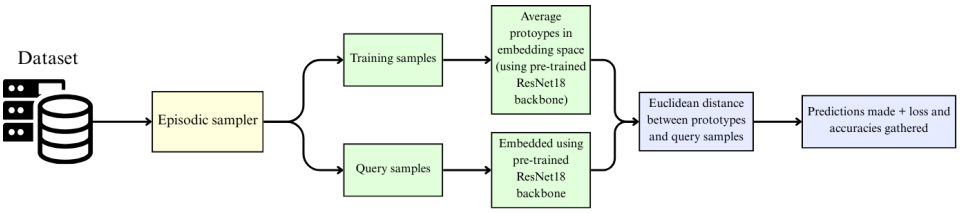

Figure 3: The architecture of a single training, validating, or testing episode of the ProtoNet used in the current study. The yellow block represents the preprocessing phase, the green blocks represent the training phase, and the blue blocks represent the evaluation phase of the episode.

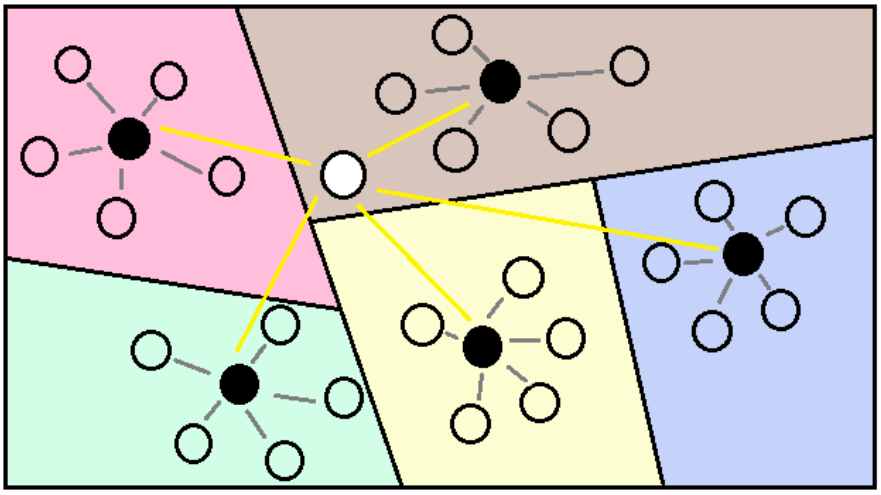

Figure 4: A visualization of how ProtoNets work inspired by Figure 1 of the study by Snell et al. (2017). The black circles represent the prototypes for each class, while the circles around them represent the training samples used. Furthermore, the white circle represents the query sample used to test these protoypes, and the yellow lines represent the distance between the prototypes and this sample.

Furthermore, the MAML model used in the current paper was inspired by and corresponded to the architecture of the study by Finn, Abbeel, and Levine (2017). The architecture used in the current paper can be seen in Figure 5. The goal of this FSL model is to set initial parameters that are optimal to use to quickly adapt to tasks it has not seen before. For this, it uses an inner loop, an outer loop and episodic training with five classes with ten query samples per class, and five or ten samples to learn from just like the ProtoNets. Each episode, the parameters of the inner loop will be adapted and used to perform well on the classification of the specific episodes it trains on. The inner loop uses the pretrained ResNet18 backbone alongside a classification head and cross-entropy to train on the

given examples. The loss of the inner loop adaptations of all the episodes of one epoch will then be used by the outer loop to slightly change the initial parameters to quickly adapt to all the seen episodes. A visualization of this process can be seen in Figure 6. After this process, the inner loop parameters were reset and this process was repeated.

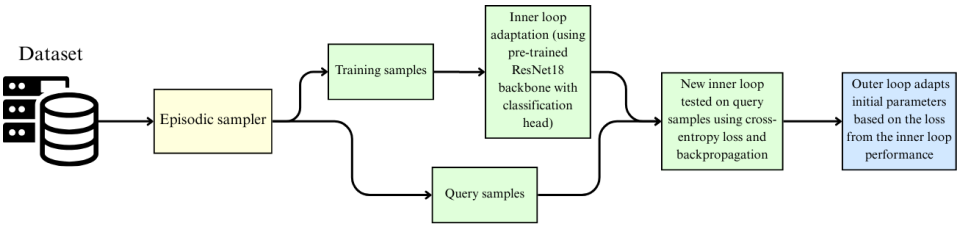

Figure 5: The architecture a single training, validating, or testing episode of the MAML model used in the current study. The yellow block represents the preprocessing phase, the green blocks represent the training phase, and the blue blocks represent the evaluation phase of the episode.

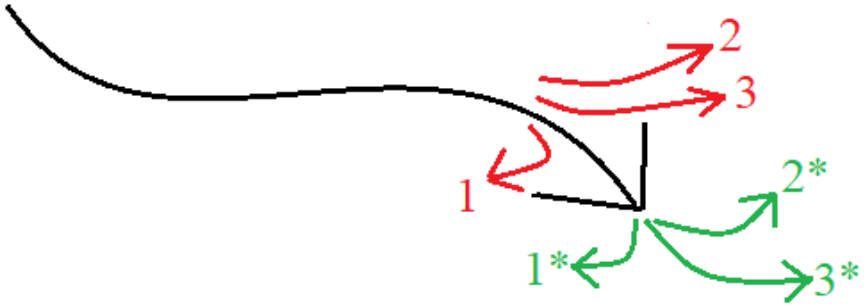

Figure 6: A visualization of how MAML models work inspired by Figure 1 of the study by Finn et al. (2017). The black arrow represents the initial parameters that are set with the outer loop. The red arrows can be seen as guidelines for updating the model based on the loss from the specific episode. Furthermore, the green arrows represent the updates of the initial parameters to minimize the losses from the specific episodes. After one epoch, the initial parameter was set to adapt to all these episodes equally.

Lastly, for the baseline model, a CNN is used. The architecture of this model can be seen in Figure 7. This model trains using updates from batches of 32 images of the dataset and cross-entropy like a supervised learning model. Each epoch consists of all the batches of the entire dataset.

This is the common way to train supervised learning models, and this is also used in the field of bioacoustics (Segura-Garcia et al., 2024). Furthermore, after each epoch, the model is validated and tested episodically with five unseen classes containing ten query samples per class. The study by Chen et al. (2019) demonstrated that CNN backbone models like this train better when given more samples to train from alongside with more diverse data. Thus, it was validated and tested on ten samples per class using three datasets. During validating and testing, the model creates embeddings and forms prototypes of the classes. Then, it uses Euclidean distance to investigate which prototype is the closest to the embeddings of the query samples. This mirrors the way how ProtoNets validate and test. Using the same method ensures a fair comparison between the FSL models, which has proven to be effective in the study by Chen et al. (2019). However, since this model trains like a supervised learning model, it is considered as a baseline model.

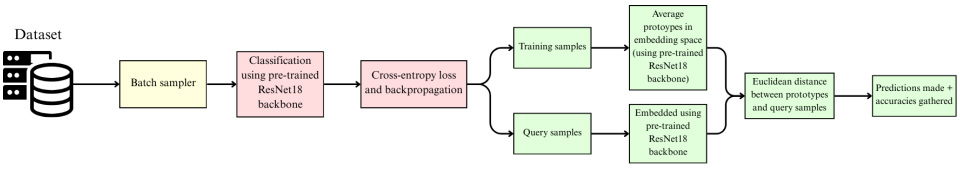

Figure 7: The architecture of the CNN baseline model used in the current study. The yellow block represents the preprocessing phase, the red blocks represent the training phase, and the green blocks represent the testing and validation phase.

4.4 *Creating the different models*

After the architectures of the models were built, the different models were created to investigate the research question and its sub-questions. For this, nine different models were built. The specifications of these models can be seen in Table 2. As this table shows, there were four models built that use the ProtoNet architecture and four that use the MAML architecture. Furthermore, there were four models that used five shots to train and test on and four models that trained and tested on ten shots. Additionally, there were four models that train on the two datasets (Clapp et al., 2023; Kahl, Charif, & Klinck, 2022). The other four models trained on three datasets (Clapp et al., 2023; Kahl, Charif, & Klinck, 2022; Kahl, Wood, et al., 2022). This meant that the models trained on two datasets contained 44 fewer species with 14963 samples. However, since these models train episodically, the number of episodes they trained on was the same for all FSL models. Lastly, the chosen parameters of the CNN baseline model were inspired by the study by Chen et al. (2019), which claims that a baseline model like this works optimally when it is trained on more and diverse data. When these

models were built, it was possible to test if adding more shots and datasets would enhance the generalizability of ProtoNets and MAML models, and if these models could outperform the baseline model.

Table 2: Model specifications

Model	Model type	Number of shots used	Number of datasets used
1	ProtoNet	5	2
2	ProtoNet	5	3
3	ProtoNet	10	2
4	ProtoNet	10	3
5	MAML	5	2
6	MAML	5	3
7	MAML	10	2
8	MAML	10	3
9	CNN	10	3

4.5 *Testing the models*

After the models were built and trained, they were tested using the test set. This dataset contained of species that the models had not seen before to optimally test the generalizability of the models. Furthermore, the performances of them were tested using 600 episodes each containing five classes it could choose from. Each model was tested on the same seed, thus each model was tested with the same episodes to make the comparisons fair. The models had five or ten shots per species to learn from, depending on how many shots they were trained, and they had to classify ten query samples of each class, thus 50 classifications tasks per episode. Then, the accuracies of the episodes are measured with a 95% confidence interval (CI), which is used to compare the models with each other. This testing method corresponds with the study by Snell et al. (2017). Additionally, F-scores were not used in the current study. Although it is a common evaluation method in the field of AI, since episodes of random classes with fixed samples were used, there was no imbalance in the classes.

After the accuracies from all the episodes with the 95% confidence intervals had been gathered, the models were compared using paired t-tests. These examined whether using five more shots and adding another dataset would significantly improve the generalization of MAML models and ProtoNets. Furthermore, it also investigated whether the baseline CNN and the FSL model significantly differed in performances. When these paired t-tests scored a p value (probability value) below or equal to .05, it was considered as a significant difference in the current study. Using this paired t-test method was based on the paper by T. Li et al. (2025).

5 RESULTS

5.1 *Accuracies and confidence intervals*

After the models were trained, their accuracies and their confidence intervals (CI's) were measured. These results can be found in Figure 8. In these tests randomly guessing would have resulted in an accuracy of 20%. Firstly, the baseline model achieved the highest accuracy of all the models (61.74%), with its CI not overlapping with any other model. Furthermore, all models that used ten shots had higher accuracies than the models with the same hyperparameters that used five shots. Additionally, the ProtoNets trained on three datasets performed better than the same models that were trained using two datasets. However, the MAML models that was trained on two datasets had a higher accuracy than the MAML models with the same parameters that was trained on three datasets. Lastly, the ProtoNets had higher accuracies than the MAML models three out of four times when the other hyperparameters were identical. However, when these models were trained on three datasets and used ten shots, the CIs did overlap. Furthermore, the MAML model using ten shots and two datasets had a higher accuracy than the ProtoNet model with the same hyperparameters. These interpretations were further investigated using paired t-tests.

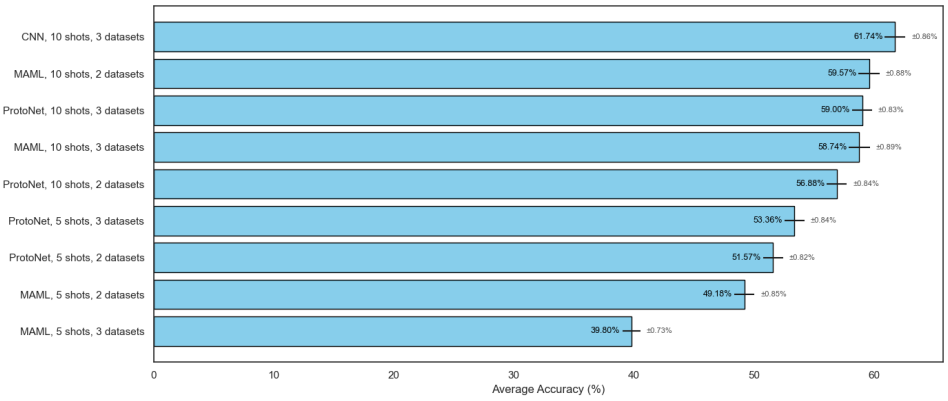

Figure 8: Average accuracies and confidence intervals of the models

5.2 *Paired t-tests*

5.2.1 *Differences between the baseline and the FSL models*

Firstly, the differences between the baseline model and the FSL models were tested using two t-tests. The performances of the FSL models with the same parameters were compared with the baseline model. This can be

found in Table 3. This table shows that the baseline model significantly outperforms both FSL models (with p values of $p < .0001$). Thus, when the parameters of the FSL models and the baseline model are the same, the baseline model outperforms the FSL models significantly.

Table 3: T-tests to test the differences between the baseline and the FSL models

Model 1	Model 2	Significance	Which model performed better?
ProtoNet, 10-shot, 3 datasets	CNN, 10-shot, 3 datasets	$p < .0001$	Model 2
MAML, 10-shot, 3 datasets	CNN, 10-shot, 3 datasets	$p < .0001$	Model 2

5.2.2 *Differences between the number of shots used*

The impact of the usage of five or ten shots is also investigated using four t-tests. These can be found in Table 4. These tests revealed that when the models used ten shots per class, they perform significantly better than the same model that used five shots (all the tests had a p value of $p < .0001$).

Table 4: T-tests to test the differences between the number of shots used

Model 1	Model 2	Significance	Which model performed better?
ProtoNet, 5-shot, 2 datasets	ProtoNet, 10-shot, 2 datasets	$p < .0001$	Model 2
ProtoNet, 5-shot, 3 datasets	ProtoNet, 10-shot, 3 datasets	$p < .0001$	Model 2
MAML, 5-shot, 2 datasets	MAML, 10-shot, 2 datasets	$p < .0001$	Model 2
MAML, 5-shot, 3 datasets	MAML, 10-shot, 3 datasets	$p < .0001$	Model 2

5.2.3 *Differences between used number of datasets*

Additionally, the impact of adding an extra dataset with different species is also investigated with four paired t-tests. The results of these tests can be seen in Table 5. These tests gave mixed results. Firstly, when ProtoNets were trained on three datasets, they perform significantly better compared to when they were trained on two datasets (with p values of $p < .0001$). However, MAML models performed significantly better when they were trained on two datasets compared to when they were trained on three datasets (with p values of $p < .0001$).

Table 5: T-tests to test the differences between used number of datasets

Model 1	Model 2	Significance	Which model performed better?
ProtoNet, 5-shot, 2 datasets	ProtoNet, 5-shot, 3 datasets	$p < .0001$	Model 2
MAML, 5-shot, 2 datasets	MAML, 5-shot, 3 datasets	$p < .0001$	Model 1
ProtoNet, 10-shot, 2 datasets	ProtoNet, 10-shot, 3 datasets	$p < .0001$	Model 2
MAML, 10-shot, 2 datasets	MAML, 10-shot, 3 datasets	$p < .0001$	Model 1

5.2.4 *Differences between FSL architectures*

Lastly, there are four t-tests done to investigate whether ProtoNet and MAML models significantly differ when they use the same hyperparameters. Table 6 demonstrates that ProtoNets perform significantly better when they are trained using five shots (with p values of $p < .0001$). However, when these models were trained using ten shots, the MAML model performs significantly better than the ProtoNet when they were trained on two datasets ($p < .0001$). Furthermore, when they were trained on three datasets and used ten shots, the differences in their performance does not differ significantly ($p = .4597$).

Table 6: T-tests to test the differences between FSL model architectures

Model 1	Model 2	Significance	Which model performed better?
MAML, 5-shot, 2 datasets	ProtoNet, 5-shot, 2 datasets	$p < .0001$	Model 2
MAML, 5-shot, 3 datasets	ProtoNet, 5-shot, 3 datasets	$p < .0001$	Model 2
MAML, 10-shot, 2 datasets	ProtoNet, 10-shot, 2 datasets	$p < .0001$	Model 1
MAML, 10-shot, 3 datasets	ProtoNet, 10-shot, 3 datasets	$p = .4597$	No significant difference

6 DISCUSSION

6.1 *Interpretation of Results*

The results of the tests of the current study give valuable insights for methods to improve generalization for FSL models. Firstly, the performances

ProtoNet and MAML model were compared with a baseline model. The results demonstrate that the baseline model is the best performing model, and it outperforms the other models significantly. This reinforces the claim made in the study by Chen et al. (2019) that baseline models with strong backbones and five or more shots to train on perform competitively with FSL models. In fact, the current study found that a baseline like this even outperform FSL models that train episodically. Thus, having a strong baseline model that validates and tests like ProtoNets, are good performing models that should not be overlooked in the field of avian bioacoustics. These results provide an answer to sub-question a: "How does the performance of few-shot learning models differ to supervised learning models in bird classification?".

A possible reason why this baseline model performs better than the FSL models is that all the datasets contain spectrograms of bird vocalizations. This can be considered a within-domain task, since there are no large differences in the types of sounds used in the datasets, next to the fact that the training validation and test sets contain different species. The study by Chen et al. (2019) found that baseline models, like the one used in the current study, perform comparable or even better than FSL models when they need to perform within-domain tasks. Thus, training episodically when training a model that only has to generalize across bird species using their audio does seem to perform better when trained with batches and cross-entropy, like a supervised learning model.

Furthermore, it was investigated whether using ten shots significantly improves different FSL models compared to using five shots per class in the field of bioacoustics. The results demonstrated that both ProtoNets and MAML models perform significantly better when they are trained on ten shots per class. These results support the claims made by Laenen and Bertinetto (2021) and S. X. Hu et al. (2022) that using more shots leads to better generalization, which results in better performances of the models. Additionally, the decrease of performance enhancement when adding more shots, what is stated by the study by Song et al. (2023), does not seem to be the case in the field of bioacoustics for the differences between the number of shots used in the current study. 10-shot models significantly enhance the performance of these FSL models compared to 5-shot models. These findings provide an answer to sub-question b: "How many samples are required for few-shot learning models to classify birds?".

A possible explanation for these findings, is because of the differences in the used bioacoustics data. The study by Jablonszky et al. (2022) found that birds differ in their vocalizations. This could imply that more data is needed to capture the entirety of the sounds a bird could make. That is a

possibility why using ten shots for these FSL models significantly enhances the generalizability of these models compared to using five shots.

The results also demonstrate that mixed results when the FSL models were either trained on two or three datasets. Firstly, both the ProtoNets perform significantly better when they were trained on three datasets compared to when they were trained on two datasets. However, both of the MAML models perform significantly better when they were trained on two datasets. Thus, these results imply that adding more data complexity, with more datasets, could lead to generalization improvements, but is dependent on the model used. These mixed results correspond with the claim made in the study by Sendra-Balcells et al. (2022). They stated that adding more data complexity might not be the best way to improve generalization, which also seems to be the case for the MAML models. However, for the ProtoNets, increasing data complexity did significantly improve their performances, which aligns with the findings of the study by Zhang et al. (2024)that imply that adding data complexity enhances the ability to generalize. These findings provide an answer to sub-question c: "How does training on more datasets impact the generalization of few-shot learning models for bird classification?".

A possible reason for why increasing data complexity resulted in these mixed results is the models' capabilities to improve with more complex data. Firstly, the research by Stoff (2025) has porven that ProtoNets are effective FSL models when data complexity is high. Additionally, MAML models perform optimally when they have more data to learn from and when the data complexity is low (Kumar, Deleu, & Bengio, 2023). This could clarify why MAML models are able to generalize better with two datasets, while the generalization increases for ProtoNets when they are trained on three datasets. Thus, the impact of data complexity seems to be heavily dependent on the generalization capabilities of the models used.

Lastly, the results demonstrated that comparing the performances of ProtoNets and MAML models with the same parameters leads to mixed results. ProtoNets perform significantly better than MAML models when they used five shots. However, when both models used ten shots, the MAML model performed significantly better than the ProtoNet when they were trained on two datasets. Furthermore, when these models were trained on three datasets and used ten shots, they did not significantly differ in their performances. These results demonstrate that both models are able of successfully classifying birds using their sounds, corresponding with the claims made by Snell et al. (2017) and Moon et al. (2023). These results provide an answer to sub-question d: "How do different few-shot learning methods compare in their performance of classifying birds?".

A reason why the ProtoNets significantly outperform the MAML models when they use five shots, is because ProtoNets are FSL models that learn new patterns of unseen data quickly without needing a lot of data, while MAML models seem to perform significantly worse when training using very little data (Kumar et al., 2023; Parnami & Lee, 2022). Furthermore, as mentioned before, according to the paper by Kumar et al. (2023), the performance of MAML models significantly decrease when the complexity of the data it trains on increases. So, when MAML models train on less complex data and have more data to train on, it can generalize better. This effect can be seen in the results and that is a possibility why the MAML model with ten shots and two datasets outperforms the ProtoNet with these parameters. Lastly, a possibility why the performances of these models did not significantly differ when they were trained on three datasets and used ten shots, is because these models plateaued using these parameters. Using a certain amount of data could result in plateauing performances, which could explain why these models perform similarly (Liu et al., 2025).

6.2 *Points of improvement and future work suggestions*

It is important to take into consideration that the current study uses within-domain data to investigate the impact of different parameters and architectures of FSL models. However, research has demonstrated that the power of FSL models especially lie in being able to generalize cross-domain data (Chen et al., 2019). An example of a cross-domain scenario which can be used in the field of animal science, is training on bird vocalization spectrograms such as done in the current study, and then testing the generalizability of these models on spectrograms of marine mammals recorded underwater. The usage of cross-domain data may also worsen the performance of the baseline model used in the current research, as found in the study by Chen et al. (2019), where FSL models outperformed baseline models when using cross-domain datasets. That is why it is relevant to investigate whether the baseline model would also perform the best when it is trained on cross-domain scenarios.

However, using within-domain data is also commonly used for FSL models, since the purpose of these models is to generalize using little data (Moon et al., 2023). Thus, the findings of the current research are valuable for improving generalization for FSL models in the field of bioacoustics. Yet, it is relevant to investigate whether the current findings to improve generalization for FSL models also occur when they are trained and tested on a cross-domain scenario.

Furthermore, the current study trained its models using 100 epochs containing 100 episodes, with early stopping and a patience of 20 epochs.

As described in the methods section, these are parameter choices that previously have been used (Gao et al., 2021; Joshi et al., 2025; Walsh et al., 2022; Ye et al., 2022). In total, this could go up to 10,000 episodes the models train on. However, this seems like a very small number of episodes compared to the 150,000 episodes that were used in a model of the study by Bai, He, and Hu (2023). In their study, they found that learning with more episodes leads to significant improvements of the models. Thus, training on more episodes could affect the performances of these models.

Nevertheless, the backbone of FSL models could impact the number of episodes necessary to train on. The study by Chen et al. (2019) found that FSL models with a pretrained transfer learning backbone could outperform models that used more episodes to train on with a weaker backbone. Furthermore, the study by Bai et al. (2023) did not use a pretrained backbone for their FSL models. These results imply that increasing the number of episodes is not the only way to optimize FSL models. Since the current study uses a pretrained ResNet18 model as a backbone, it is presumable that the model converged if it did not improve after 2000 seen episodes when the early stopping is triggered. However, to strengthen this counter-argument, future research should investigate the performance differences of these models when they are trained on 10,000 episodes and 150,000 episodes.

Additionally, the study by Sendra-Balcells et al. (2022) states that using transfer learning and data augmentation will significantly improve the generalization performance of FSL models. In the current study, transfer learning is used for the backbone of the models. However, this study did not investigate the impact on the generalization when it was not present in the models. Furthermore, the impact of data augmentation is not implemented in the current study, thus not investigated as well. That is why, this is worthwhile to further investigate. Lastly, the current study has demonstrated that using 10-shot learning perform significantly better. Further research could investigate whether adding more shots would still significantly improve classification of avian bioacoustics, since research has stated that this improvement will decrease when you add more samples (Song et al., 2023).

7 CONCLUSION

To conclude, the current research has investigated different methods to improve generalization for FSL models. This study found that using ten shots to classify birds using their sounds leads to significantly better performance compared to using five shots for ProtoNets and MAML models. Furthermore, ProtoNets performed better than MAML models when they

were used five shots. Yet, when they used ten shots, MAML models performed better when these models were trained on two datasets, and the models performed similarly when they were trained on three datasets. Adding another dataset to train on leads to mixed results: ProtoNets benefitted from more data complexity, while MAML models did not. Thus, it is important to investigate whether the used FSL models are capable of training with more data complexity to generalize better. When that is the case, it could enhance generalization. Lastly, models that train like supervised learning models and test like ProtoNet models were the best at generalizing in the current study. This demonstrates that this method would be interesting to investigate more to enhance generalization in the field of FSL.

Thus, to answer the research question: "How can the generalization of few-shot learning models be improved for birdsong classification?", using the right architecture, the right amount of samples, and the right amount of data diversity can all significantly improve generalization for birdsong classification. These findings offer valuable contributions to the field of FSL research. FSL models have proven to be an effective automatic animal monitoring method, thus this paper provides impactful insights in enhancing the process of preserving the wellbeing of wildlife. Furthermore, these findings also contribute to enhancing classification performances of endangered species that do not have much data, which can help with conservation of these species. Thus, enhancing FSL models in the field of bioacoustics strengthens animal monitoring, resulting in better conservation of bird species, including the endangered species.

REFERENCES

Abbas, Q. (2023). An intelligent medical image classification system using few-shot learning. *Concurrency and Computation: Practice and Experience*, *35*(2), e7451. https://doi.org/10.1002/cpe.7451.

Agilandeeswari, L., & Meena, S. D. (2023). Swin transformer based contrastive self-supervised learning for animal detection and classification. *Multimedia Tools and Applications*, *82*(7), 10445–10470. https://doi.org/10.1007/s11042-022-13629-x.

Al Dawasari, H. J., Bilal, M., Moinuddin, M., Arshad, K., & Assaleh, K. (2023). Deepvision: Enhanced drone detection and recognition in visible imagery through deep learning networks. *Sensors*, *23*(21), 8711. https://doi.org/10.3390/s23218711.

Anderson, M., & Harte, N. (2021). Bioacoustic event detection with prototypical networks and data augmentation. *arXiv preprint arXiv:2112.09006*. https://doi.org/10.48550/arXiv.2112.09006.

Bai, Y., He, Z., & Hu, J. (2023). On the episodic difficulty of few-shot learning. In *Asian conference on machine learning* (pp. 48–63).

Boudiaf, M., Ziko, I., Rony, J., Dolz, J., Piantanida, P., & Ben Ayed, I. (2020). Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, *33*, 2445–2457. https://doi.org/10.48550/arXiv.2008.11297.

Cao, T., Law, M., & Fidler, S. (2019). A theoretical analysis of the number of shots in few-shot learning. *arXiv preprint arXiv:1909.11722*. https://doi.org/10.48550/arXiv.1909.11722.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., & Huang, J.-B. (2019). A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*. https://doi.org/10.48550/arXiv.1904.04232.

Clapp, M., Kahl, S., Meyer, E., McKenna, M., Klinck, H., & Patricelli, G. (2023). A collection of fully-annotated soundscape recordings from the southern sierra nevada mountain range. *Dataset on Zenodo, January*. https://doi.org/10.5281/zenodo.7525804.

Congdon, J. V., Hosseini, M., Gading, E. F., Masousi, M., Franke, M., & MacDonald, S. E. (2022). The future of artificial intelligence in monitoring animal identification, health, and behaviour. *Animals*, *12*(13), 1711. https://doi.org/10.3390/ani12131711.

Dupuis-Desormeaux, M., Davidson, Z., Mwololo, M., Kisio, E., & MacDonald, S. E. (2016). Comparing motion capture cameras versus human observer monitoring of mammal movement through fence gaps: a case study from kenya. *African Journal of Ecology*, *54*(2), 154–161. https://doi.org/10.1111/aje.12277.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for

fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126–1135). https://doi.org/10.48550/arXiv.1703.03400.

Gao, Y., Fei, N., Liu, G., Lu, Z., & Xiang, T. (2021). Contrastive prototype learning with augmented embeddings for few-shot learning. In *Uncertainty in artificial intelligence* (pp. 140–150). https://doi.org/10.48550/arXiv.2101.09499.

Gibbons, A., King, E., Donohue, I., & Parnell, A. (2024). Generative ai-based data augmentation for improved bioacoustic classification in noisy environments. *arXiv preprint arXiv:2412.01530*. https://doi.org/10.48550/arXiv.2412.01530.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.48550/arXiv.1512.03385.

Herath, L., Meedeniya, D., Marasingha, J., & Weerasinghe, V. (2022). Optimize transfer learning for autism spectrum disorder classification with neuroimaging: A comparative study. In *2022 2nd international conference on advanced research in computing (icarc)* (pp. 171–176). https://doi.org/10.1109/ICARC54489.2022.9753949.

Hexeberg, S., Chitre, M., Hoffmann-Kuhnt, M., & Low, B. W. (2025). Semi-supervised classification of bird vocalizations. *arXiv preprint arXiv:2502.13440*. https://doi.org/10.48550/arXiv.2502.13440.

Hopping, W. A., Kahl, S., & Klink, H. (2022). A collection of fully-annotated soundscape recordings from the southwestern amazon basin. 1. *Dataset on Zenodo*, *7079124*. https://doi.org/10.5281/zenodo.7079124.

Hu, S. X., Li, D., Stühmer, J., Kim, M., & Hospedales, T. M. (2022). Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. *arXiv preprint arXiv:2204.07305*. https://doi.org/10.48550/arXiv.2204.07305.

Hu, Y., Pateux, S., & Gripon, V. (2022). Squeezing backbone feature distributions to the max for efficient few-shot learning. *Algorithms*, *15*(5), 147. https://doi.org/10.3390/a15050147.

Jablonszky, M., Canal, D., Hegyi, G., Herényi, M., Laczi, M., Lao, O., ... others (2022). Estimating heritability of song considering within-individual variance in a wild songbird: The collared flycatcher. *Frontiers in Ecology and Evolution*, *10*, 975687. https://doi.org/10.3389/fevo.2022.975687.

Joshi, P., Mundra, S., & Mundra, A. (2025). Proto-att-fsl: enhanced prototypical network for cross-domain few-shot airline sentiment classification. *Social Network Analysis and Mining*, *15*(1), 2. https://doi.org/10.1007/s13278-025-01436-9.

Kahl, S., Charif, R., & Klinck, H. (2022). A collection of fully-annotated soundscape recordings from the northeastern united states. *Dataset on Zenodo, September 2022. URL https://doi.org/10.5281/zenodo.7079380, 7079380*. https://doi.org/10.5281/zenodo.7079380.

Kahl, S., Wood, C. M., Chaon, P., Peery, M. Z., & Klinck, H. (2022). A collection of fully-annotated soundscape recordings from the western united states. *Dataset on Zenodo, September 2022c. URL https://doi. org/10.5281/zenodo, 7050014*. https://doi.org/10.5281/zenodo.7050014.

Kumar, R., Deleu, T., & Bengio, Y. (2023). The effect of diversity in meta-learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 8396–8404). https://doi.org/10.1609/aaai.v37i7.26012.

Laenen, S., & Bertinetto, L. (2021). On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, *34*, 24581–24592. https://doi.org/10.5555/3540261.3542143.

Li, T., Zhang, Y., Su, D., Liu, M., Ge, M., Chen, L., … Tang, J. (2025). Knowledge graph-based few-shot learning for label of medical imaging reports. *Academic Radiology*. https://doi.org/10.1016/j.acra.2025.02.045.

Li, X., Jia, M., Islam, M. T., Yu, L., & Xing, L. (2020). Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, *39*(12), 4023–4033. https://doi.org/10.1109/TMI.2020.3008871.

Liu, M., Wu, F., Li, B., Lu, Z., Yu, Y., & Li, X. (2025). Envisioning class entity reasoning by large language models for few-shot learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 39, pp. 18906–18914). https://doi.org/10.1609/aaai.v39i18.34081.

Lu, Q., Liu, W., Zhuo, Z., Li, Y., Duan, Y., Yu, P., … Liu, Y. (2022). A transfer learning approach to few-shot segmentation of novel white matter tracts. *Medical Image Analysis*, *79*, 102454. https://doi.org/10.1609/aaai.v39i18.34081.

McEwen, B., Soltero, K., Gutschmidt, S., Bainbridge-Smith, A., Atlas, J., & Green, R. (2024). Active few-shot learning for rare bioacoustic feature annotation. *Ecological Informatics*, *82*, 102734. https://doi.org/10.1016/j.ecoinf.2024.102734.

Michielon, A., Litta, P., Bonelli, F., Don, G., Farisè, S., Giannuzzi, D., … others (2024). Mind the step: An artificial intelligence-based monitoring platform for animal welfare. *Sensors*, *24*(24), 8042. https://doi.org/10.3390/s24248042.

Moon, J., Kim, E., Hwang, J., & Hwang, E. (2023). Task-adaptive parameter transformation scheme for maml-based few-shot animal sound classification. *Available at SSRN 4516279*.

https://doi.org/10.2139/ssrn.4516279.

Moummad, I., Farrugia, N., & Serizel, R. (2024a). Regularized contrastive pre-training for few-shot bioacoustic sound detection. In *Icassp 2024-2024 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1436–1440). https://doi.org/10.1109/ICASSP48485.2024.10446409.

Moummad, I., Farrugia, N., & Serizel, R. (2024b). Self-supervised learning for few-shot bird sound classification. In *2024 ieee international conference on acoustics, speech, and signal processing workshops (icasspw)* (pp. 600–604). https://doi.org/10.1109/ICASSPW62465.2024.10627576.

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., ... others (2023). Learning to detect an animal sound from five examples. *Ecological informatics*, *77*, 102258. https://doi.org/10.1016/j.ecoinf.2023.102258.

Parnami, A., & Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*. https://doi.org/10.48550/arXiv.2203.04291.

Rauch, L., Schwinger, R., Wirth, M., Heinrich, R., Huseljic, D., Herde, M., ... others (2024). Birdset: A large-scale dataset for audio classification in avian bioacoustics. *arXiv preprint arXiv:2403.10380*. https://doi.org/10.48550/arXiv.2403.10380.

Roy, P. (2024). Enhancing real-world robustness in ai: Challenges and solutions. *Journal of Recent Trends in Computer Science and Engineering (JRTCSE)*, *12*(1), 34–49. https://doi.org/10.70589/JRTCSE.2024.1.6.

Samiappan, S., Krishnan, B. S., Dehart, D., Jones, L. R., Elmore, J. A., Evans, K. O., & Iglay, R. B. (2024). Aerial wildlife image repository for animal monitoring with drones in the age of artificial intelligence. *Database*, *2024*, baae070. https://doi.org/10.1093/database/baae070.

Segura-Garcia, J., Sturley, S., Arevalillo-Herraez, M., Alcaraz-Calero, J. M., Felici-Castell, S., & Navarro-Camba, E. A. (2024). 5g ai-iot system for bird species monitoring and song classification. *Sensors*, *24*(11), 3687. https://doi.org/10.3390/s24113687.

Sendra-Balcells, C., Campello, V. M., Martín-Isla, C., Viladés, D., Descalzo, M. L., Guala, A., ... Lekadir, K. (2022). Domain generalization in deep learning for contrast-enhanced imaging. *Computers in Biology and Medicine*, *149*, 106052. https://doi.org/10.1016/j.compbiomed.2022.106052.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1703.05175.

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A comprehensive survey of few-shot learning: Evolution, applications,

challenges, and opportunities. *ACM Computing Surveys*, *55*(13s), 1–40. https://doi.org/10.1145/3582688.

Stoff, M. (2025). *Prototypical visualization: Using prototypical networks for visualizing large unstructured data* (Doctoral dissertation, Technische Universität Wien). https://doi.org/10.34726/hss.2025.119321.

Van Merriënboer, B., Hamer, J., Dumoulin, V., Triantafillou, E., & Denton, T. (2024). Birds, bats and beyond: Evaluating generalization in bioacoustics models. *Frontiers in Bird Science*, *3*, 1369756. https://doi.org/10.3389/fbirs.2024.1369756.

Vega-Hidalgo, A., Kahl, S., Symes, L. B., Ruiz-Gutiérrez, V., Molina-Mora, I., Cediel, F., … Klinck, H. (2023). A collection of fully-annotated soundscape recordings from neotropical coffee farms in colombia and costa rica. *Dataset on Zenodo, January*. https://doi.org/10.5281/zenodo.7525348.

Walsh, R., Abdelpakey, M. H., Shehata, M. S., & Mohamed, M. M. (2022). Automated human cell classification in sparse datasets using few-shot learning. *Scientific Reports*, *12*(1), 2924. https://doi.org/10.1038/s41598-022-06718-2.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, *53*(3), 1–34. https://doi.org/10.1145/3386252.

Wei, X., Hossain, M. Z., & Ahmed, K. A. (2022). A resnet attention model for classifying mosquitoes from wing-beating sounds. *Scientific Reports*, *12*(1), 10334. https://doi.org/10.1038/s41598-022-14372-x.

Weldy, M. J., Denton, T., Fleishman, A. B., Tolchin, J., McKown, M., Spaan, R. S., … Lesmeister, D. B. (2024). Audio tagging of avian dawn chorus recordings in california, oregon and washington. *Biodiversity Data Journal*, *12*, e118315. https://doi.org/10.3897/BDJ.12.e118315.

Wolters, P., Sizemore, L., Daw, C., Hutchinson, B., & Phillips, L. (2021). Proposal-based few-shot sound event detection for speech and environmental sounds with perceivers. *arXiv preprint arXiv:2107.13616*. https://doi.org/10.48550/arXiv.2107.13616.

Won, J.-h., & Kim, D.-h. (2024). Metric-based few-shot transfer learning approach for voice pathology detection. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3480523.

Xie, J., & Zhu, M. (2023). Acoustic classification of bird species using an early fusion of deep features. *Birds*, *4*(1), 138–147. https://doi.org/10.3390/birds4010011.

Ye, H.-J., Ming, L., Zhan, D.-C., & Chao, W.-L. (2022). Few-shot learning with a strong teacher. *IEEE transactions on pattern analysis and machine intelligence*, *46*(3), 1425–1440. https://doi.org/10.1109/TPAMI.2022.3160362.

Zhang, X., Huang, H., Zhang, D., Zhuang, S., Han, S., Lai, P., & Liu, H. (2024). Cross-dataset generalization in deep learning. *arXiv preprint arXiv:2410.11207*. https://doi.org/10.48550/arXiv.2410.11207.

8 self-reflection

Working on my thesis gave me multiple new insights. Firstly, it was desired to learn more about FSL models and to work with them. Since this thesis is about enhancing generalization capabilities, a lot of insights were gained during the process of the thesis. Furthermore, there were also different FSL models built in this thesis, which enhanced my knowledge gained about these models even more.

Additionally, I wanted to learn more about using AI in the field of animal science. The current paper was the perfect fit for this goal. Using FSL models were capable of classifying birds using a few samples of their vocalizations. Furthermore, there were different methods investigated to enhance generalization. These are relevant insights that can be used in real world scenario's to monitor birds, thus this thesis has introduced me to the field of animal science and AI.

Lastly, I wanted to make as fair comparisons as possible between the created models so that the results are really meaningful. This has been done successfully by only changing a single parameters to compare the models. Furthermore, the models were tested on a fixed seed. This also ensures a fair comparison. In conclusion, this project has given me new meaningful insights that were desired at the start of this project.

9 APPENDICES

9.1 Appendix A

Dataset classes specifications

Train dataset containing three different datasets:

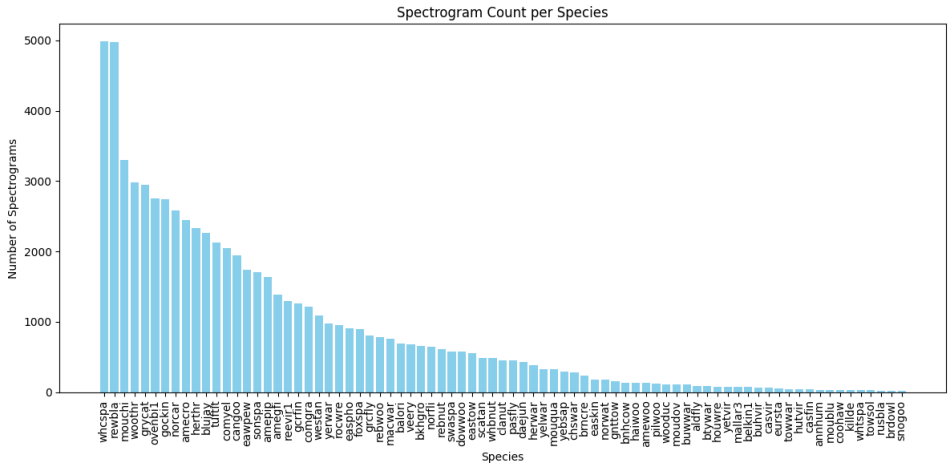

Figure A1: A visualization of the classes in the training dataset that contains three datasets

Table A1: The specifications of the birds in the training dataset that contains three datasets

Species code	Common name	Scientific name	Sample count
whcspa	White-crowned Sparrow	Zonotrichia leucophrys	4986
rewbla	Red-winged Blackbird	Agelaius phoeniceus	4973
mouchi	Mountain Chickadee	Poecile gambeli	3299
woothr	Wood Thrush	Hylocichla mustelina	2981
grycat	Gray Catbird	Dumetella carolinensis	2943
ovenbi1	Ovenbird	Seiurus aurocapilla	2760

Species code	Common name	Scientific name	Sample count
gockin	Golden-crowned Kinglet	Regulus satrapa	2741
norcar	Northern Cardinal	Cardinalis cardinalis	2583
amecro	American Crow	Corvus brachyrhynchos	2444
herthr	Hermit Thrush	Catharus guttatus	2332
blujay	Blue Jay	Cyanocitta cristata	2268
tuftit	Tufted Titmouse	Baeolophus bicolor	2126
comyel	Common Yellowthroat	Geothlypis trichas	2049
cangoo	Canada Goose	Branta canadensis	1947
eawpew	Eastern Wood-Pewee	Contopus virens	1745
sonspa	Song Sparrow	Melospiza melodia	1705
amepip	American Pipit	Anthus rubescens	1634
amegfi	American Goldfinch	Spinus tristis	1386
reevir1	Red-eyed Vireo	Vireo olivaceus	1297
gcrfin	Gray-crowned Rosy-Finch	Leucosticte tephrocotis	1259
comgra	Common Grackle	Quiscalus quiscula	1217
westan	Western Tanager	Piranga ludoviciana	1087
yerwar	Yellow-rumped Warbler	Setophaga coronata	973
rocwre	Rock Wren	Salpinctes obsoletus	953
easpho	Eastern Phoebe	Sayornis phoebe	912

Species code	Common name	Scientific name	Sample count
foxspa	Fox Sparrow	Passerella iliaca	900
grcfly	Great Crested Flycatcher	Myiarchus crinitus	809
rebwoo	Red-bellied Woodpecker	Melanerpes carolinus	778
macwar	MacGillivray's Warbler	Geothlypis tolmiei	765
balori	Baltimore Oriole	Icterus galbula	687
veery	Veery	Catharus fuscescens	677
bkhgro	Black-headed Grosbeak	Pheucticus melanocephalus	660
norfli	Northern Flicker	Colaptes auratus	643
rebnut	Red-breasted Nuthatch	Sitta canadensis	610
swaspa	Swamp Sparrow	Melospiza georgiana	578
dowwoo	Downy Woodpecker	Dryobates pubescens	576
eastow	Eastern Towhee	Pipilo erythrophthalmus	557
scatan	Scarlet Tanager	Piranga olivacea	490
whbnut	White-breasted Nuthatch	Sitta carolinensis	490
clanut	Clark's Nutcracker	Nucifraga columbiana	456
pasfly	Pacific-slope Flycatcher	Empidonax difficilis	448
daejun	Dark-eyed Junco	Junco hyemalis	434
herwar	Hermit Warbler	Setophaga occidentalis	384
yelwar	Yellow Warbler	Setophaga petechia	333

Species code	Common name	Scientific name	Sample count
mouqua	Mountain Quail	Oreortyx pictus	325
yebsap	Yellow-bellied Sapsucker	Sphyrapicus varius	298
chswar	Chestnut-sided Warbler	Setophaga pensylvanica	284
brncre	Brown Creeper	Certhia americana	237
easkin	Eastern Kingbird	Tyrannus tyrannus	184
norwat	Northern Waterthrush	Parkesia noveboracensis	175
gnttow	Green-tailed Towhee	Pipilo chlorurus	158
bnhcow	Brown-headed Cowbird	Molothrus ater	138
haiwoo	Hairy Woodpecker	Dryobates villosus	138
amewoo	American Woodcock	Scolopax minor	134
pilwoo	Pileated Woodpecker	Dryocopus pileatus	127
wooduc	Wood Duck	Aix sponsa	113
moudov	Mourning Dove	Zenaida macroura	107
buwwar	Blue-winged Warbler	Vermivora cyanoptera	106
aldfly	Alder Flycatcher	Empidonax alnorum	93
btywar	Black-throated Gray Warbler	Setophaga nigrescens	89
houwre	House Wren	Troglodytes aedon	81
yetvir	Yellow-throated Vireo	Vireo flavifrons	78
mallar3	Mallard	Anas platyrhynchos	76

Species code	Common name	Scientific name	Sample count
belkin1	Belted Kingfisher	Megaceryle alcyon	73
buhvir	Blue-headed Vireo	Vireo solitarius	66
casvir	Cassin's Vireo	Vireo cassinii	60
eursta	European Starling	Sturnus vulgaris	53
towwar	Townsend's Warbler	Setophaga townsendi	43
hutvir	Hutton's Vireo	Vireo huttoni	39
casfin	Cassin's Finch	Haemorhous cassinii	37
annhum	Anna's Hummingbird	Calypte anna	34
moublu	Mountain Bluebird	Sialia currucoides	34
coohaw	Cooper's Hawk	Accipiter cooperii	33
whtspa	White-throated Sparrow	Zonotrichia albicollis	32
killde	Killdeer	Charadrius vociferus	32
towsol	Townsend's Solitaire	Myadestes townsendi	28
rusbla	Rusty Blackbird	Euphagus carolinus	21
brdowl	Barred Owl	Strix varia	20
snogoo	Snow Goose	Anser caerulescens	20

Train dataset containing two different datasets:

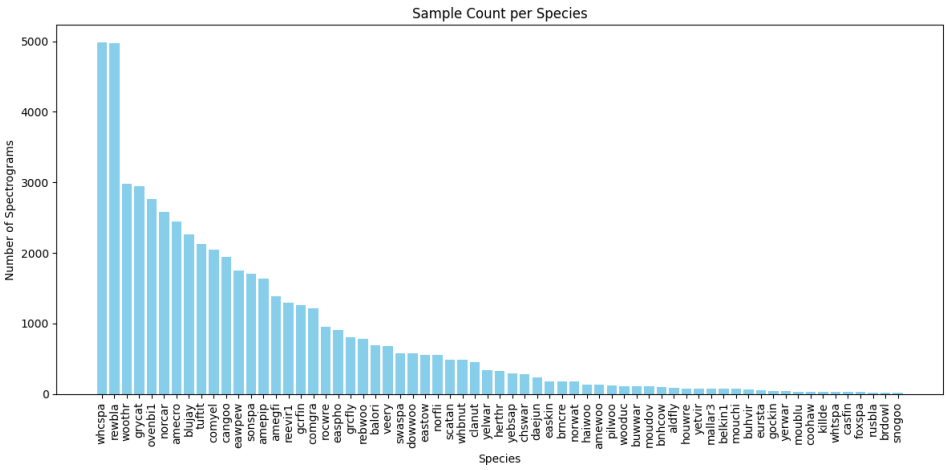

Figure A2: A visualization of the classes in the training dataset that contains two datasets

Table A2: The specifications of the birds in the training dataset that contains two datasets

Species code	Common name	Scientific name	Sample count
whcspa	White-crowned Sparrow	Zonotrichia leucophrys	4981
rewbla	Red-winged Blackbird	Agelaius phoeniceus	4973
woothr	Wood Thrush	Hylocichla mustelina	2981
grycat	Gray Catbird	Dumetella carolinensis	2943
ovenbi1	Ovenbird	Seiurus aurocapilla	2760
norcar	Northern Cardinal	Cardinalis cardinalis	2583
amecro	American Crow	Corvus brachyrhynchos	2444
blujay	Blue Jay	Cyanocitta cristata	2268
tuftit	Tufted Titmouse	Baeolophus bicolor	2126

Species code	Common name	Scientific name	Sample count
comyel	Common Yellowthroat	Geothlypis trichas	2049
cangoo	Canada Goose	Branta canadensis	1945
eawpew	Eastern Wood-Pewee	Contopus virens	1745
sonspa	Song Sparrow	Melospiza melodia	1705
amepip	American Pipit	Anthus rubescens	1634
amegfi	American Goldfinch	Spinus tristis	1385
reevir1	Red-eyed Vireo	Vireo olivaceus	1297
gcrfin	Gray-crowned Rosy-Finch	Leucosticte tephrocotis	1259
comgra	Common Grackle	Quiscalus quiscula	1217
rocwre	Rock Wren	Salpinctes obsoletus	953
easpho	Eastern Phoebe	Sayornis phoebe	912
grcfly	Great Crested Flycatcher	Myiarchus crinitus	809
rebwoo	Red-bellied Woodpecker	Melanerpes carolinus	778
balori	Baltimore Oriole	Icterus galbula	687
veery	Veery	Catharus fuscescens	677
swaspa	Swamp Sparrow	Melospiza georgiana	578
dowwoo	Downy Woodpecker	Dryobates pubescens	576
eastow	Eastern Towhee	Pipilo erythrophthalmus	557
norfli	Northern Flicker	Colaptes auratus	551

Species code	Common name	Scientific name	Sample count
scatan	Scarlet Tanager	Piranga olivacea	490
whbnut	White-breasted Nuthatch	Sitta carolinensis	490
clanut	Clark's Nutcracker	Nucifraga columbiana	456
yelwar	Yellow Warbler	Setophaga petechia	333
herthr	Hermit Thrush	Catharus guttatus	329
yebsap	Yellow-bellied Sapsucker	Sphyrapicus varius	298
chswar	Chestnut-sided Warbler	Setophaga pensylvanica	284
daejun	Dark-eyed Junco	Junco hyemalis	233
easkin	Eastern Kingbird	Tyrannus tyrannus	184
brncre	Brown Creeper	Certhia americana	179
norwat	Northern Waterthrush	Parkesia noveboracensis	175
haiwoo	Hairy Woodpecker	Dryobates villosus	138
amewoo	American Woodcock	Scolopax minor	134
pilwoo	Pileated Woodpecker	Dryocopus pileatus	127
wooduc	Wood Duck	Aix sponsa	113
buwwar	Blue-winged Warbler	Vermivora cyanoptera	106
moudov	Mourning Dove	Zenaida macroura	105
bnhcow	Brown-headed Cowbird	Molothrus ater	98
aldfly	Alder Flycatcher	Empidonax alnorum	93

Species code	Common name	Scientific name	Sample count
yetvir	Yellow-throated Vireo	Vireo flavifrons	78
houwre	House Wren	Troglodytes aedon	78
mallar3	Mallard	Anas platyrhynchos	76
belkin1	Belted Kingfisher	Megaceryle alcyon	73
mouchi	Mountain Chickadee	Poecile gambeli	71
buhvir	Blue-headed Vireo	Vireo solitarius	66
eursta	European Starling	Sturnus vulgaris	53
gockin	Golden-crowned Kinglet	Regulus satrapa	45
yerwar	Yellow-rumped Warbler	Setophaga coronata	40
moublu	Mountain Bluebird	Sialia currucoides	34
coohaw	Cooper's Hawk	Accipiter cooperii	33
killde	Killdeer	Charadrius vociferus	32
whtspa	White-throated Sparrow	Zonotrichia albicollis	32
foxspa	Fox Sparrow	Passerella iliaca	30
casfin	Cassin's Finch	Haemorhous cassinii	30
rusbla	Rusty Blackbird	Euphagus carolinus	21
snogoo	Snow Goose	Anser caerulescens	20
brdowl	Barred Owl	Strix varia	20

Validation dataset:

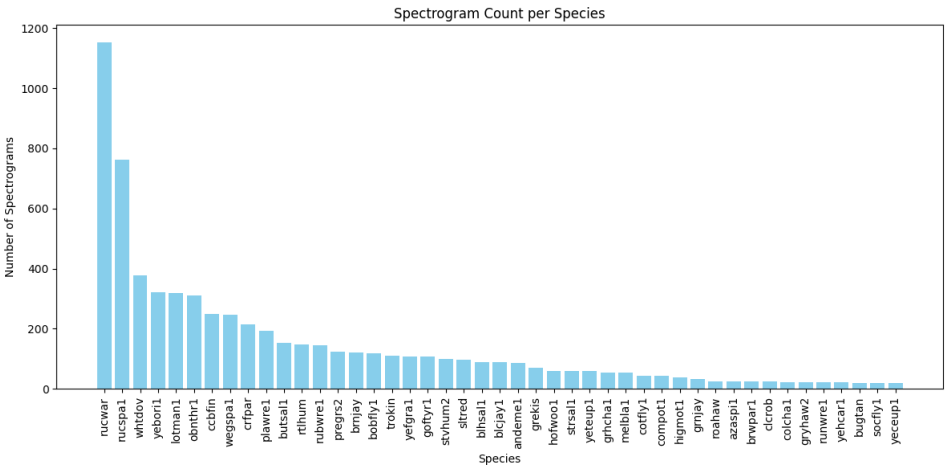

Figure A3: A visualization of the classes in the validation dataset

Table A3: The specifications of the birds in the validation dataset

Species code	Common name	Scientific name	Sample count
rucwar	Rufous-capped Warbler	Basileuterus rufifrons	1153
rucspa1	Rufous-collared Sparrow	Zonotrichia capensis	762
whtdov	White-tipped Dove	Leptotila verreauxi	377
yebori1	Yellow-backed Oriole	Icterus chrysater	321
lotman1	Long-tailed Manakin	Chiroxiphia linearis	318
obnthr1	Orange-billed Nightingale-Thrush	Catharus aurantiirostris	311
ccbfin	Chestnut-capped Brushfinch	Arremon brunneinucha	248
wegspa1	White-eared Ground-Sparrow	Melozone leucotis	247
crfpar	Crimson-fronted Parakeet	Psittacara finschi	215
plawre1	Cabanis's Wren	Cantorchilus modestus	192

Species code	Common name	Scientific name	Sample count
butsal1	Buff-throated Saltator	Saltator maximus	153
rtlhum	Rufous-tailed Hummingbird	Amazilia tzacatl	147
rubwre1	Rufous-breasted Wren	Pheugopedius rutilus	144
pregrs2	Cabanis's Ground-Sparrow	Melozone cabanisi	122
brnjay	Brown Jay	Psilorhinus morio	120
bobfly1	Boat-billed Flycatcher	Megarynchus pitangua	118
trokin	Tropical Kingbird	Tyrannus melancholicus	111
yefgra1	Yellow-faced Grassquit	Tiaris olivaceus	108
goftyr1	Golden-faced Tyrannulet	Zimmerius chrysops	106
stvhum2	Steely-vented Hummingbird	Saucerottia saucerottei	99
sltred	Slate-throated Redstart	Myioborus miniatus	97
blhsal1	Black-headed Saltator	Saltator atriceps	89
blcjay1	Black-chested Jay	Cyanocorax affinis	88
andeme1	Andean Emerald	Uranomitra franciae	86
grekis	Great Kiskadee	Pitangus sulphuratus	69
yeteup1	Yellow-throated Euphonia	Euphonia hirundinacea	60
hofwoo1	Hoffmann's Woodpecker	Melanerpes hoffmannii	60
strsal1	Streaked Saltator	Saltator striatipectus	60

Species code	Common name	Scientific name	Sample count
grhcha1	Gray-headed Chachalaca	Ortalis cinereiceps	55
melbla1	Melodious Blackbird	Dives dives	55
cotfly1	Common Tody-Flycatcher	Todirostrum cinereum	44
compot1	Common Potoo	Nyctibius griseus	42
higmot1	Andean Motmot	Momotus aequatorialis	39
grnjay	Green Jay	Cyanocorax yncas	33
roahaw	Roadside Hawk	Rupornis magnirostris	25
azaspi1	Azara's Spinetail	Synallaxis azarae	24
brwpar1	Bronze-winged Parrot	Pionus chalcopterus	24
clcrob	Clay-colored Thrush	Turdus grayi	23
gryhaw2	Gray Hawk	Buteo plagiatus	21
runwre1	Rufous-naped Wren	Campylorhynchus rufinucha	21
colcha1	Colombian Chachalaca	Ortalis columbiana	21
yehcar1	Yellow-headed Caracara	Milvago chimachima	21
bugtan	Blue-gray Tanager	Thraupis episcopus	20
yeceup1	Yellow-crowned Euphonia	Euphonia luteicapilla	20
socfly1	Social Flycatcher	Myiozetetes similis	20

Test dataset:

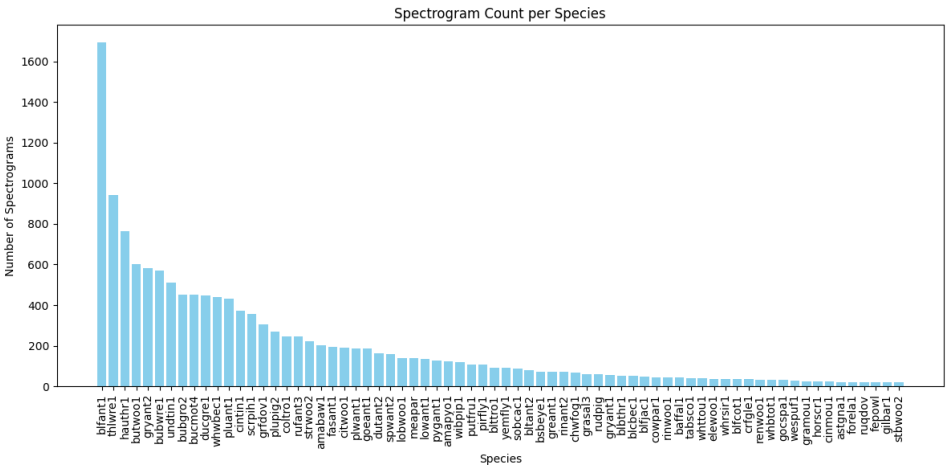

Figure A4: A visualization of the classes in the test dataset

Table A4: The specifications of the birds in the test dataset

Species code	Common name	Scientific name	Sample count
blfant1	Black-faced Antthrush	Formicarius analis	1694
thlwre1	Thrush-like Wren	Campylorhynchus turdinus	942
hauthr1	Hauxwell's Thrush	Turdus hauxwelli	763
butwoo1	Buff-throated Woodcreeper	Xiphorhynchus guttatus	602
gryant2	Gray Antbird	Cercomacra cinerascens	583
bubwre1	Buff-breasted Wren	Cantorchilus leucotis	570
undtin1	Undulated Tinamou	Crypturellus undulatus	509
bubgro2	Amazonian Grosbeak	Cyanoloxia rothschildii	453
bucmot4	Amazonian Motmot	Momotus momota	450
ducgre1	Dusky-capped Greenlet	Pachysylvia hypoxantha	446

Species code	Common name	Scientific name	Sample count
whwbec1	White-winged Becard	Pachyramphus polychopterus	439
pluant1	Plumbeous Antbird	Myrmelastes hyperythrus	432
cintin1	Cinereous Tinamou	Crypturellus cinereus	372
scrpih1	Screaming Piha	Lipaugus vociferans	355
grfdov1	Gray-fronted Dove	Leptotila rufaxilla	307
plupig2	Plumbeous Pigeon	Patagioenas plumbea	269
coltro1	Collared Trogon	Trogon collaris	247
rufant3	Rufous-fronted Antthrush	Formicarius rufifrons	247
strwoo2	Striped Woodcreeper	Xiphorhynchus obsoletus	221
amabaw1	Amazonian Barred-Woodcreeper	Dendrocolaptes certhia	203
fasant1	Fasciated Antshrike	Cymbilaimus lineatus	196
citwoo1	Cinnamon-throated Woodcreeper	Dendrexetastes rufigula	191
plwant1	Plain-winged Antshrike	Thamnophilus schistaceus	187
goeant1	Goeldi's Antbird	Akletos goeldii	185
dutant2	Dusky-throated Antshrike	Thamnomanes ardesiacus	162
spwant2	Spot-winged Antshrike	Pygiptila stellaris	158
meapar	Mealy Parrot	Amazona farinosa	139
lobwoo1	Long-billed Woodcreeper	Nasica longirostris	139

Species code	Common name	Scientific name	Sample count
lowant1	Long-winged Antwren	Myrmotherula longipennis	136
pygant1	Pygmy Antwren	Myrmotherula brachyura	128
amapyo1	Amazonian Pygmy-Owl	Glaucidium hardyi	125
wibpip1	Wing-barred Piprites	Piprites chloris	120
putfru1	Purple-throated Fruitcrow	Querula purpurata	107
pirfly1	Piratic Flycatcher	Legatus leucophaius	106
blttro1	Black-tailed Trogon	Trogon melanurus	93
yemfly1	Yellow-margined Flycatcher	Tolmomyias assimilis	92
sobcac1	Solitary Black Cacique	Cacicus solitarius	89
bltant2	Black-throated Antbird	Myrmophylax atrothorax	79
bsbeye1	Black-spotted Bare-eye	Phlegopsis nigromaculata	74
rinant2	Ringed Antpipit	Corythopis torquatus	71
greant1	Great Antshrike	Taraba major	71
chwfog1	Chestnut-winged Foliage-gleaner	Dendroma erythroptera	68
grasal3	Blue-gray Saltator	Saltator coerulescens	61
rudpig	Ruddy Pigeon	Patagioenas subvinacea	59
gryant1	Gray Antwren	Myrmotherula menetriesii	58
blbthr1	Black-billed Thrush	Turdus ignobilis	54

Species code	Common name	Scientific name	Sample count
blcbec1	Black-capped Becard	Pachyramphus marginatus	53
blfjac1	Bluish-fronted Jacamar	Galbula cyanescens	50
rinwoo1	Ringed Woodpecker	Celeus torquatus	44
cowpar1	Cobalt-winged Parakeet	Brotogeris cyanoptera	44
baffal1	Barred Forest-Falcon	Micrastur ruficollis	43
tabsco1	Tawny-bellied Screech-Owl	Megascops watsonii	42
whttou1	White-throated Toucan	Ramphastos tucanus	40
whrsir1	White-rumped Sirystes	Sirystes albocinereus	38
elewoo1	Elegant Woodcreeper	Xiphorhynchus elegans	38
crfgle1	Cinnamon-rumped Foliage-gleaner	Philydor pyrrhodes	36
blfcot1	Black-faced Cotinga	Conioptilon mcilhennyi	36
renwoo1	Red-necked Woodpecker	Campephilus rubricollis	34
whbtot1	White-bellied Tody-Tyrant	Hemitriccus griseipectus	32
gocspa1	Golden-crowned Spadebill	Platyrinchus coronatus	31
wespuf1	Western Striolated-Puffbird	Nystalus obamai	28
gramou1	Grayish Mourner	Rhytipterna simplex	24
horscr1	Horned Screamer	Anhima cornuta	24
cinmou1	Cinereous Mourner	Laniocera hypopyrra	23
astgna1	Ash-throated Gnateater	Conopophaga peruviana	22

Species code	Common name	Scientific name	Sample count
forela1	Forest Elaenia	Myiopagis gaimardii	21
ruqdov	Ruddy Quail-Dove	Geotrygon montana	21
fepowl	Ferruginous Pygmy-Owl	Glaucidium brasilianum	20
stbwoo2	Straight-billed Woodcreeper	Dendroplex picus	20
gilbar1	Gilded Barbet	Capito auratus	20