Acoustic Species Classification

Atul Ahire
Meghana Dayathri
Raghad Alsulami
Raviteja T

1. Abstract

Birds can serve as effective indicators of changes in biodiversity due to their mobility and diverse habitat needs. The alterations in the species composition and bird population can reflect the progress or shortcomings of restoration projects. However, conventional bird biodiversity surveys relying on observer-based methods are often difficult and expensive to conduct over extensive areas. Alternatively, passive acoustic monitoring (PAM) combined with modern analytical techniques that employ machine learning can enable conservationists to study the correlation between restoration interventions and biodiversity in greater depth by sampling larger spatial scales with higher temporal resolution. Ultimately, the optimal objective would be to create a pipeline capable of precisely detecting a wide range of species vocalizations within a specified location where audio equipment is installed.

2. Introduction

Climate change has significant and wide-ranging effects on biodiversity. Rising temperatures, altered precipitation patterns, and changing climatic conditions disrupt ecosystems, leading to shifts in species distributions and habitat suitability. Many species are experiencing range contractions or are being forced to migrate to more suitable habitats, but some may face barriers and struggle to adapt. Climate change also intensifies other threats to biodiversity, such as habitat loss, pollution, and invasive species. The loss of key species can disrupt ecological interactions and cascading effects throughout the food web. These impacts underscore the urgent need for climate action to safeguard the intricate web of life on Earth.

Studying bird populations is vital for measuring and monitoring biodiversity. Birds are sensitive to climate change and serve as indicators of ecosystem health. Monitoring their populations helps assess the effectiveness of restoration projects and conservation efforts.

Traditional methods employed by researchers for studying bird populations are on-site surveys and using cameras to capture bird images. On site survey involves researchers physically going to specific locations and conducting field surveys to observe and record bird species present in those areas. However, this method can be challenging and labor

intensive. Researchers need to spend significant amounts of time and effort traveling to various locations, setting up observation points, and conducting surveys. In recent years, there has been a shift from using cameras to utilizing audio recordings for capturing bird data. The advantages of audio recordings, such as affordability, robustness in challenging conditions, higher coverage, and non-intrusiveness, have led to their increased popularity. Researchers and bird enthusiasts alike have recognized the potential of audio devices in providing a more cost-effective and comprehensive approach to studying bird populations and their behaviors. By leveraging the distinct vocalizations of birds, we can gather reliable data over larger areas and minimize any disturbance to their natural habitat. As a result, audio recordings have emerged as a superior choice in the field of bird data collection and have gained traction as the preferred method in recent years. Developing machine learning models for bird classification can enhance monitoring by automating species identification and improving data analysis, facilitating more efficient biodiversity assessments and informing conservation strategies.

2.1 Dataset

The train data contains 264 species from Kenya, Africa, and the test set consists of approximately 200 10-minute soundscapes. Xeno-canto provided 16,900 audio recordings which can be used to train a classifier. This data is part of the BirdClef 2023 Kaggle competition [1][2].

2.2 Project Statement

In this work, we train a machine learning model to identify birds from 5 second audio recordings by converting them to Mel spectrograms. Our machine learning model for bird species identification using audio recordings provides a solution for assessing and understanding biodiversity. By accurately identifying bird species remotely, it enables comprehensive monitoring, informs conservation strategies, and contributes to the preservation of ecosystems and their inhabitants.

3. Related Work

Recently, deep learning has become increasingly popular for identifying bird species based on their sounds [7][8]. However, there are ongoing challenges when it comes to accurately classifying birds in natural environments. These challenges include not having enough labeled data and dealing with an uneven distribution of different bird species [9].

To deal with the lack of labeled bird data, two common strategies are used: transfer learning and building a custom lightweight model from scratch. Transfer learning helps overcome

the issue of not having enough labeled training data [10]. Many research studies in the field of bird sound classification have adopted transfer learning techniques.

In past BirdCLEF competitions, different challenges focused on identifying birds in various sound environments. One approach centered on learning from less detailed labels and successfully classifying bird sounds in the wild while ignoring background noises [11]. Interestingly, the initial experiments found that smaller models like ResNet-50 and EfficientNet-B0 performed better than larger ones. Another solution used a "nocall" detector [12] with ResNeSt-26. The "nocall" label was added when other predictions had low confidence. This may result in an imperfect prediction list, but it's a trade-off that helps ensure that at least one correct label is predicted when multiple labels are present.

To avoid overfitting and to improve the generalization capabilities of the neural network model to various recording conditions, another solution employed a bunch of data augmentation methods [13] to the audio segments like pitch augmentation, mask augmentation, noise augmentation, loudness augmentation, etc. Another approach took inspiration from PANNs (Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition) and created a backbone network using ResNet and DenseNet [14]. They performed a bunch of augmentation techniques like spectrogram augmentation, primary and secondary background noise addition, etc. to improve performance of the model. In this approach, after log-mel feature extraction, the inputs are passed to ResNets/DenseNets by removing the last fully connected layers and extracting only features. Then, a modified 1D attention based fully connected layer is attached to ResNet. The output of this network is a dictionary which contains clipwise and framewise outputs. Jie Xie [15] and Zhu found that deep cascade features being extracted from VGG16 and MobileNetV2 achieved best performance using repeat-based spectrogram. AMResNet [16], inspired by the integration of attentional mechanism and residual networks, has been proposed as a robust solution for achieving accurate classification. By leveraging the attentional mechanism, this approach effectively extracts and selects high-dimensional features, leading to enhanced classification accuracy. The attentional mechanism optimizes recognition efficiency by assigning appropriate weights to channels and spatial components, thereby improving the overall performance. Additionally, the incorporation of residual networks within AMResNet addresses the challenge of gradient disappearance and facilitates increased information flow through skip connections. This integration enables the model to effectively leverage the advantages of residual networks, leading to improved performance and better representation of complex bird sounds. Another novel approach has been proposed, combining transfer learning of a pre-trained deep convolutional neural network (CNN) model with a semi-supervised pseudo-labeling method and a customized loss function [17]. This innovative methodology aims to overcome the limitations of limited labeled data by leveraging unlabeled data effectively.

By simultaneously utilizing labeled and unlabeled data, the proposed approach enables the network to be trained in a supervised fashion, augmenting the size of the training set, and improving the model's performance. This combination of transfer learning, semi supervised learning, and a custom loss function results in a more robust and accurate classification model.

4. Data

Our dataset comprises audio recordings from diverse habitats, capturing the vocalizations of numerous bird species. The data preprocessing will involve rigorous steps, including noise removal, normalization, and feature extraction. Major callouts from the dataset include variations in recording quality due to environmental factors, seasonal changes affecting bird behavior, and the presence of multiple bird species in a single recording. Feature construction will involve extracting Mel-frequency cepstral coefficients (MFCCs) and spectrogram representations, enabling the models to discern unique patterns in bird vocalizations.

4.1. Constraints and Risks

- The data contains a lot of interference, which could include background noise from other birds or natural environmental sounds. Moreover, there is a lack of clear timestamps that accurately indicate when the bird calls are happening. Our plan of action is to use a binary classifier to identify if a bird exists or not.
- A multi-label classification problem arises when multiple bird species are singing simultaneously, and there are various categories or types of bird songs.
- The dataset may exhibit a high level of imbalance due to the greater prevalence of certain bird species over others. Furthermore, the dataset may consist of numerous species, and the recordings may vary in terms of their length and quality.
- As a result of the high level of imbalance and variability in the dataset, it may be challenging to train a model that accurately recognizes all bird species present in the recordings. Therefore, it is important to pre-process the data and explore different modeling techniques to achieve best possible results.

4.2 Additional information on Dataset -

1. train_audio/: The training data consists of short recordings of individual bird calls generously uploaded by users of xenocanto.org. These files have been downsampled to 32 kHz where applicable to match the test set audio and converted to the ogg format. The training data should have nearly all relevant files; we expect there is no benefit to looking for more on xenocanto.org.

2. test_soundscapes/: When you submit a notebook, the test_soundscapes directory will be populated with approximately 200 recordings to be used for scoring. They are 10 minutes long and in ogg audio format. The file names are randomized. It should take your submission notebook approximately five minutes to load all of the test soundscapes.

3. train_metadata.csv: A wide range of metadata is provided for the training data. The most directly relevant fields are:

    a. primary label - a code for the bird species. You can review detailed information about the bird codes by appending the code to https://ebird.org/species/, such as https://ebird.org/species/amecro for the American Crow.
    b. latitude & longitude: coordinates for where the recording was taken. Some bird species may have local call 'dialects,' so you may want to seek geographic diversity in your training data.
    c.
    d. author - The user who provided the recording.
    e. filename: the name of the associated audio file.

.

4. eBird_Taxonomy_v2021.csv - Data on the relationships between different species.

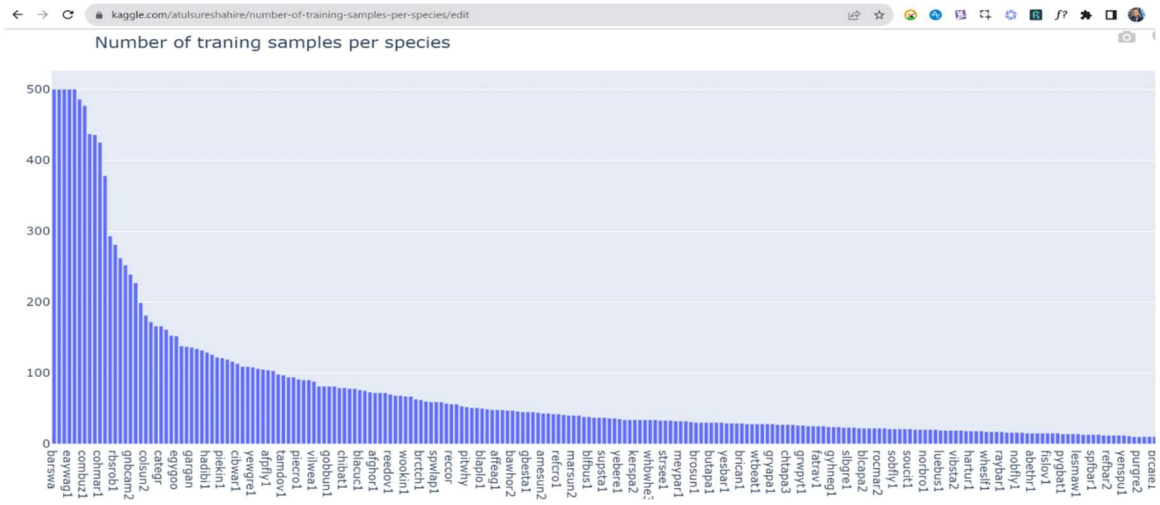

Figure 1: Training samples for each species

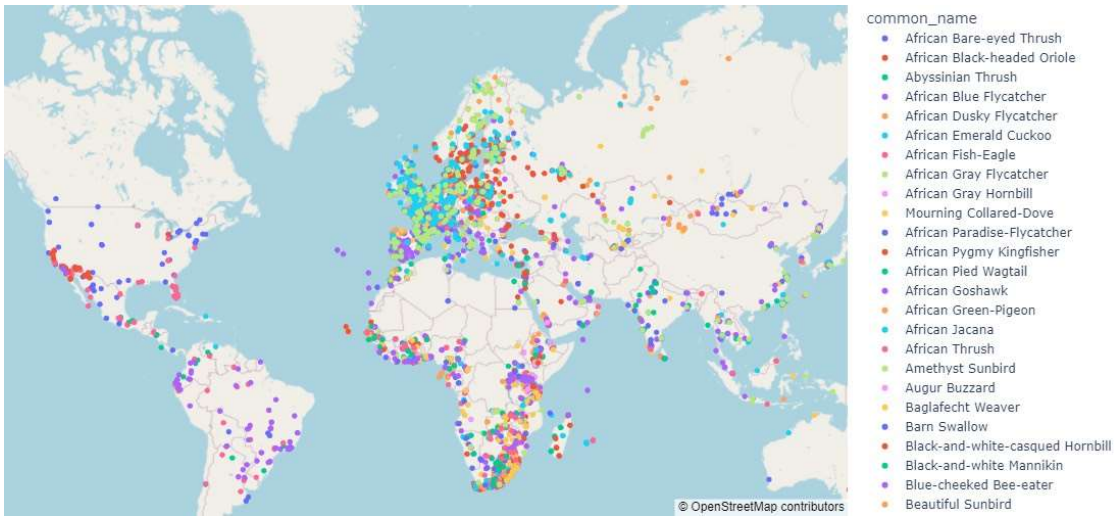

Figure 2: Location of birds

Data analysis notebook is attached in source code.

5. Methodology

This comprehensive report provides a detailed overview of our approach to acoustic bird classification, emphasizing the significance of our findings in the context of biodiversity conservation and ecological research. Through meticulous data preprocessing, and advanced machine learning techniques, our project will contribute valuable insights to the scientific community while acknowledging the challenges posed by real-world environmental variability.

To achieve the objective of identifying vocalizing birds within a specified 5-second segment of audio recorded in various soundscapes, the project will be divided into two parts -

- The first part will involve converting the weakly labeled data into strongly labeled time-based data. This process will help to improve the accuracy of the classification model by providing more specific and accurate labels for each segment of audio.
- In the second part of the project, a multi-species classification model will be developed to predict the bird species present in the 5-second segment of audio. The model will be trained on 16,900 audio recordings supplied by Xeno-canto, which will help to improve the accuracy of the model.

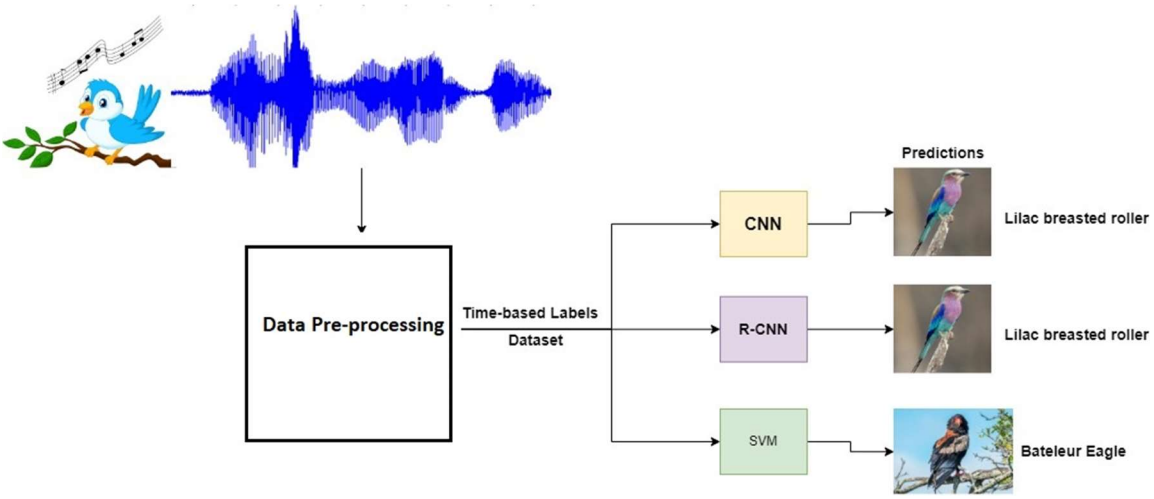

Figure 2: Multi-species classifier to identify the bird from its calling

5.1. Data Preprocessing

The training dataset contains weakly labeled data, where the labels indicate the bird species whose sounds are present in the audio, but not the precise time frame in which the audio occurs. The multi-species bird classifier can be trained by breaking the audio into 5-sec chunks and using weak labels. Training the model in this way may be less effective, as it is possible that some of the 5-second audio chunks used do not include the bird sounds. On the other hand, using strongly labeled data can help to improve the accuracy of the classification model by providing more specific and accurate labels for each segment of audio.

While examining a discussion thread on Kaggle related to the competition, we came to the realization that the 5-second segments in the test dataset might not contain any bird vocalizations. In such instances, our trained model would erroneously predict one of the 264 bird species even in the absence of bird calls, creating a ”no bird” scenario. To tackle this problem, we made the decision to introduce a new class specifically for ”no bird” cases. We are acheiving this using open source library called “PyHa” [17]. A tool designed to convert audio-based "weak" labels to "strong" moment-to-moment labels. Provides a pipeline to compare automated moment-to-moment labels to human labels. Current proof of concept work being fulfilled on Bird Audio clips using Microfaune predictions. Script of converting weak labels to strong labels using “PyHa” are attached in source code.

Finally, the 5 second audio chunk will be converted into Melspectogram using the library like librosa, which is utilized to train the classifiers.

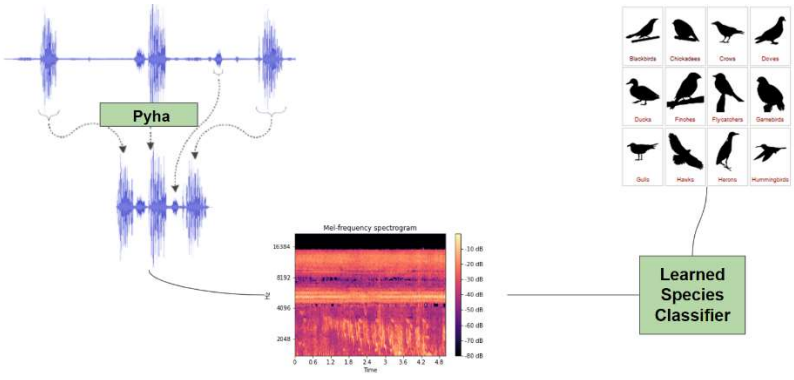

Figure 3: Data Preprocessing using PyHA

5.2. Model

To ensure the success of the project, it will be important to carefully design the data preprocessing pipeline, feature extraction, and model training. The resulting model should be accurate, efficient, and able to generalize well to unseen data. The machine learning models that we are planning to consider building our multi-species classification model are Convolutional Neural Networks (CNNs) VGG16, ResNET50, Efficient B0. Also, the Professor has suggested using Unsupervised learning to address this problem. We are going to implement that and evaluate the accuracy. After initial research, we discovered that these models perform well for image classification. Upon delving into research papers [3][4], it was found that a significant portion of the research had been prompted by different AI challenges, including BirdCLEF [1]. Winners of these competitions had some interesting solutions using CNNs and R-CNNs. We also intend to conduct a thorough literature survey to explore more models that will help us build a good multi-species classification model. All these models can be used for image classification with each having its own advantages and disadvantages. After testing all these models, we will be selecting the model that provides the highest accuracy. Once we have chosen a model that best suits the given dataset, we will perform hyper-parameter tuning and data augmentation techniques to improve the model.

5.3 Data Augmentation Techniques

By applying data augmentation techniques, we wanted to enhance the quality of the dataset and generalize our model. We discovered that audio and spectrogram augmentation can be used based on the literature review and past Kaggle competition solutions [2]. Audio augmentation involves manipulating the audio data directly, while spectrogram augmentation involves applying augmentations to the spectrogram representation of the audio. Pitch shift, time stretching, time masking, and frequency masking were the spectrogram augmentation methods used. On the other hand, for audio augmentation, we applied noise addition and noise reduction techniques [6]. Due to time constraints, we

couldn’t do much data preprocessing. We believe by doing this augmentation, we would have achieved better results.

6. Experimentation

6.1. Model :

We came to the realization that building a model from scratch and training it on the bird dataset would consume a significant amount of time, and it might not yield optimal results since it lacks pretraining. We decided to explore transfer learning approach. We proceeded to experiment with several established models, including EfficientNet B0, ResNet 50, and VGG16. These models have been pre- trained on large-scale image datasets, making them well-suited for our project. Please check the attached notebooks for all three models.

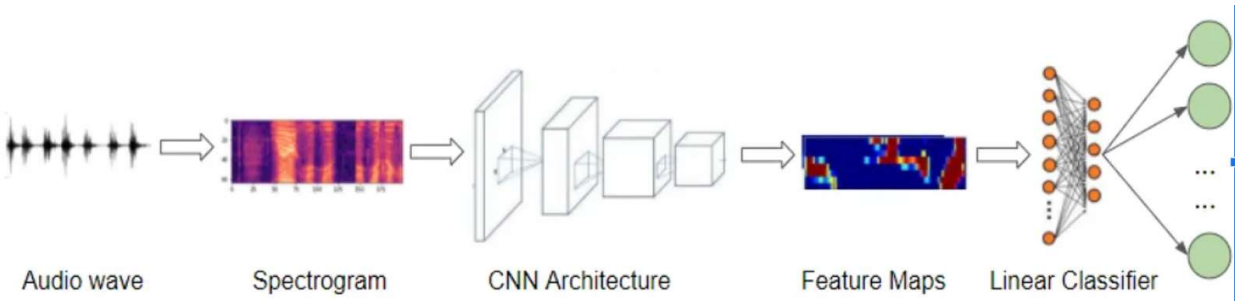

Figure 4: Model architecture

6.2 Unsupervised Learning:

Unsupervised learning techniques are employed to categorize bird species based on their unique sounds. The process starts with extracting Mel-Frequency Cepstral Coefficients (MFCCs) from audio files, encapsulating the distinct acoustic features of each sound. K-Means clustering is then applied to group the sounds into clusters, with evaluation metrics such as Silhouette Score and Calinski-Harabasz Score used to assess the effectiveness of the clustering. Subsequently, DBSCAN and Hierarchical Clustering techniques are implemented, each offering unique approaches to cluster identification. These unsupervised learning techniques showcase their ability to cluster sounds into groups, each of which internally contains multiple classes of bird species. Finally, Gaussian Mixture Models (GMM) are applied, assuming a mixture of Gaussian distributions in the data.

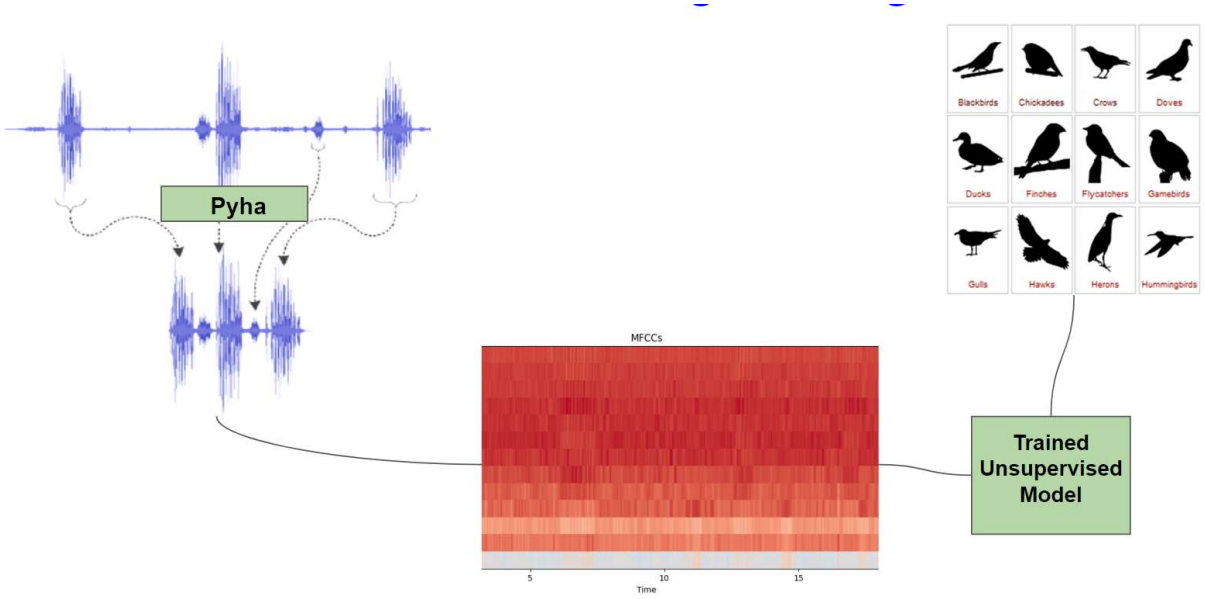

Figure 5. Unsupervised learning architecture

While these unsupervised methods provide insights into bird sound categorization, the code acknowledges potential limitations when compared to deep learning techniques. Deep learning, especially neural networks, excels in capturing intricate patterns from raw audio data, outperforming traditional clustering methods like K-Means or GMM. These traditional methods rely on predefined features, whereas deep learning models autonomously learn complex relationships present in the data. The choice between unsupervised techniques and deep learning depends on the intricacy of the data and the desired level of feature extraction sophistication.

Model	Silhouette Score	Calinski-Harabasz Scores	Davies-Bouldin
K-Means	0.256	407.51	1.46
DBSCAN	0.536	44.47	2.80
Agglomerative	0.688	1535.99	0.43
GMM	0.617	734.26	0.78

6.3 Training Details :

In all the experiments we train for 10 epochs, we use batch size 10 and Adam Optimizer. We use a Cross Entropy loss function implemented in PyTorch. Additionally, we used a learning rate of 0.0001. We have used 80% of the dataset for training and 20% of the dataset for validation. During training we track the loss function, F1 score, Precision, Accuracy, Recall, average precision score for both training and validation data.

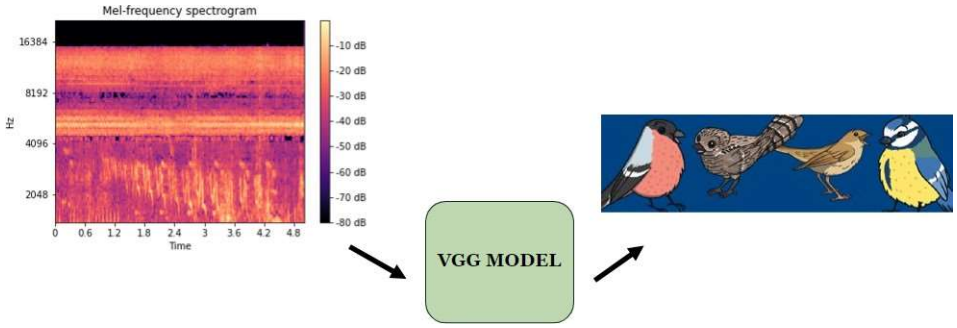

Figure 7: Training pipeline using VGG model

7. Results

To choose the best model among EfficientNet, ResNet, and VGG we trained and evaluated the performance of these models on a subset of the dataset. We adopted a random selection process to choose 30 classes from the available data. The purpose of this selection was to generate melspectrogram images for the corresponding data frames derived from the PyHa output. By selecting a diverse set of 30 classes, we aimed to obtain a representative sample of the dataset. We compared validation accuracy, macro average precision, F-1 score of these models to gain insights into the strengths and weaknesses of each model and identify the model best suited for our application.

Model	Initial Accuracy	Training Loss	Validation Loss	Training Accuracy	Validation Accuracy
Resnet50	73.02%	0.17	0.63	95%	85.94%
Vgg16	77.08%	0.20	0.77	94.55%	83.14%
EfficientNet B0	78.12%	0.10	0.47	97.65%	88.51%

Table 1: Comparison of validation accuracy of various models

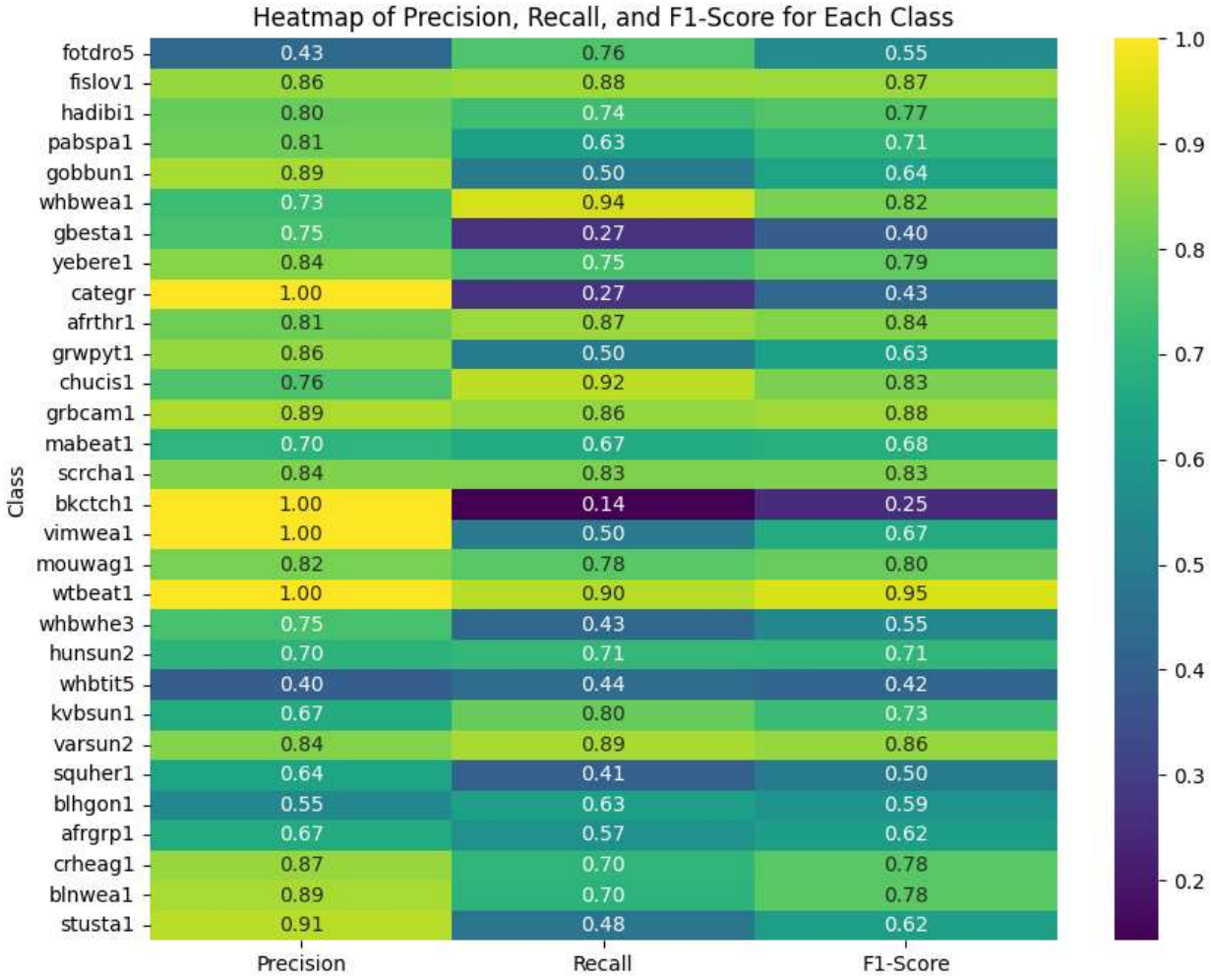

Figure 7: Evaluation metrics for VGG16

Analysis:
ResNet50 exhibits good initial accuracy, with a relatively low training loss indicating efficient learning during training. However, there's a substantial gap between the training and validation loss, suggesting potential overfitting. The model achieves a high training accuracy, but the validation accuracy lags, signifying some difficulty in generalizing to unseen data.

VGG16 starts with a slightly higher initial accuracy compared to ResNet50. It also shows a lower training loss but has a relatively higher validation loss, indicating a higher level of overfitting. The training and validation accuracies are close but still demonstrate a similar issue of overfitting, with the model struggling more to generalize.

EfficientNet B0 boasts the highest initial accuracy among the three models. It demonstrates the lowest training loss and validation loss, signifying efficient learning and a good balance between fitting the training data and generalizing to unseen data. The high training accuracy also aligns well with the validation accuracy, indicating better generalization compared to the other models.

Overall Assessment:

- **Efficiency:** EfficientNet B0 seems to be the most efficient in terms of training and generalization, showcasing the lowest losses and a well-balanced accuracy.
- **Overfitting:** ResNet50 and VGG16 both suffer from overfitting, indicated by wider gaps between training and validation losses, and slightly lower validation accuracies compared to their training accuracies.
- **Generalization:** EfficientNet B0 demonstrates better generalization ability, as its validation accuracy is closer to the training accuracy, suggesting it might perform better on unseen data.

In summary, based on these metrics, EfficientNet B0 appears to be the most promising model among the three, showcasing better overall performance and generalization capability.

8. Future work

Training multiple models to create an ensemble model for bird detection is the right step toward improving the accuracy of bird detection. In addition, the data augmentation

technique of mixing audio to create new audio can be very effective in enhancing the training data set. Another approach to try out is to pre-train the model on different bird datasets before fine-tuning it to the specific application. This method can help boost the performance of the model significantly. Adding noise to the audio can be used as an effective data augmentation technique to generalize the model. In our project, we have explored adding Gaussian noise to the audio for training the model. In the future, we can explore if training the model by adding different types of noises can be effective for bird detection. Currently, the model built to identify bird species based on the call is based on primary labels. If the audio has multiple bird sounds, our model uses only the primary bird label. To address this issue, the model can be converted into a multi-label classifier by training it on primary and secondary bird labels.

9. Conclusion

In conclusion, this project has demonstrated the potential of leveraging Machine Learning and Con- volutional Neural Networks (CNNs) to identify bird species from audio recordings. By utilizing low-cost audio recording devices or sensors, researchers can now effectively monitor bird populations and biodiversity within various ecosystems. This approach overcomes the limitations of conventional surveys that are labor-intensive and difficult to conduct.

The pipeline developed in this project converts weakly labeled data to strongly labeled data using the PyHa binary classifier. This strongly labeled data is then used to train the VGG model, enabling accurate classification of bird species. To enhance the robustness and generalization capability of the models, various data augmentation techniques were employed.

The machine learning model developed in this project provides a valuable tool for evaluating biodiversity restoration projects and monitoring bird populations. By harnessing the power of technology, we can gain deeper insights into the changes in bird populations and their implications for ecosystem health.

Individual contribution to group work:

- EDA and data visualization - Raviteja
- Data Preprocessing - Atul
- Model Training and Classification:
    - VGG16 model - Atul
    - RESNET50 model - Raghad
    - EFFICIENTNET model - Meghana
- Unsupervised learning – Raviteja

References

[1] https://www.imageclef.org/BirdCLEF2023

[2] https://www.kaggle.com/competitions/birdclef-2023

[3] Hanguang Xiao, Daidai Liu, Kai Chen, and Mi Zhu. Amresnet: *An automatic recognition model of bird sounds in a real environment*. Applied Acoustics, 201:109121, 2022.

[4] Marcos V Conde, Kumar Shubham, Prateek Agnihotri, Nitin D Movva, and Szilard Bessenyei. *Weakly-supervised classification and detection of bird sounds in the wild.* a birdclef 2021 solution. arXiv preprint arXiv:2107.04878, 2021.

[5] Arunodhayan Sampathkumar and Danny Kowerko. Tuc media computing at birdclef 2021: *Noise augmentation strategies in bird sound classification in combination with densenets and resnets.* In CLEF, 2021

[6] Nanni, L., Maguolo, G., & Paci, M. (2020). *Data augmentation approaches for improving animal audio classification. Ecological Informatics*, 57, 101084. https://doi.org/10.1016/j.ecoinf.2020.101084

[7] Jie Xie and Mingying Zhu. *Handcrafted features and late fusion with deep learning for bird sound classification.* Ecological Informatics, 52:74–81, 2019.

[8] Xin Zhang, Aibin Chen, Guoxiong Zhou, Zhiqiang Zhang, Xibei Huang, and Xiaohu Qiang. *Spectrogram-frame linear network and continuous frame sequence for bird sound classification.* Ecological Informatics, 54:101009, 2019.

[9] Jie Xie, Kai Hu, Ya Guo, Qibin Zhu, and Jinghu Yu. *On loss functions and cnns for improved bioacoustic signal classification.* Ecological Informatics, 64:101331, 2021.

[10] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In Artificial Neural Networks and Machine Learning–ICANN
[11] Marcos V Conde, Kumar Shubham, Prateek Agnihotri, Nitin D Movva, and Szilard Bessenyei. *Weakly-supervised classification and detection of bird sounds in the wild*. a birdclef 2021 solution. arXiv preprint arXiv:2107.04878, 2021.

[12] Kyle Maclean and Isaac Triguero. *Identifying bird species by their calls in soundscapes.* Applied Intelligence, pages 1–15, 2023.

[13] Markus M¨uhling, Jakob Franz, Nikolaus Korfhage, and Bernd Freisleben. Bird species recognition via neural architecture search. In CLEF (Working Notes), pages 1–13, 2020.

[14] Arunodhayan Sampathkumar and Danny Kowerko. Tuc media computing at birdclef 2021: Noise augmentation strategies in bird sound classification in combination with densenets and resnets. In CLEF, 2021.

[15] Jie Xie and Mingying Zhu. Acoustic classification of bird species using an early fusion of deep features. Birds, 4(1):138–147, 2023.

[16] Hanguang Xiao, Daidai Liu, Kai Chen, and Mi Zhu. Amresnet: An automatic recognition model of bird sounds in real environment. Applied Acoustics, 201:109121, 2022.

[17] Ming Zhong, Jack LeBien, Marconi Campos-Cerqueira, Rahul Dodhia, Juan Lavista Ferres, Julian P Velev, and T Mitchell Aide. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Applied Acoustics, 166:107375, 2020.
[18] https://github.com/UCSD-E4E/PyHa