Analysis of Sounds of Seattle Birds

Tileshwar Narayan
tnarayan@seattleu.edu

Abstract:

The analysis revolves around voice recognition with the use of deep learning methods. In this analysis, the neural network models have been utilized for the binary and multiclass classification of bird species based on the provided data and evaluated their performance on three test voices. The binary classification, utilizing neural networks, achieved accuracies of 90.32% and 80.64% with batch sizes of 32 and 64, respectively. In this analysis, the focus was on segregating between two bird species, 'American Crow' (amecro) and 'Blue Jay' (blujay). This high level of accuracy underscores the efficiency of neural network architectures in differentiating between two classes of bird calls. The next model addressed multiclass classification and provided some good results with accuracies of 59.77% and 64.94% with batch sizes of 32 and 64, respectively, spread among the 12 bird species. This analysis highlights the ability of machine learning algorithms to correctly recognize different types of bird songs. Besides evaluating model performance on internal datasets, we also examined the classification of real test data consisting of raw sound clips. Some of these clips contained sounds from multiple bird species, making classification challenging for automatic systems. In the analysis of three test clips, our model demonstrated its capability to predict bird species with a top-5 classification output. For the first test, the 'American Crow' (amecro) was identified as the most probable species with a confidence score of 51.68%, followed by the 'Western Meadowlark' (wesmea) at 44.45%. In the second test, the 'Western Meadowlark' predicted as the primary bird with a confidence score of 86.06%, while the 'American Crow' followed with 13.51%. Similarly, in the third test, the 'Western Meadowlark' and 'American Crow' were the top contenders, with a score of 49.77% and 47.55% respectively. These outcomes suggest the model's consistent performance in identifying dominant bird species across various audio samples, underscoring its efficiency in multiclass bird classification tasks.

Introduction:

Voice recognition is the most important results of deep learning techniques in the context of machine learning. The analysis includes the classification of bird species by means of neural network models and deep learning on birds' voices. With the help of three unprocessed test datasets alongside preprocessed data in this analysis, and for both binary and multiclass calls, this analysis is aimed to distinguish the bird species. The analysis targets include an experimental set of two bird types—the dataset involves the American Crow (amecro) and the Blue Jay (blujay) so that binary classification can be effectively applied and 12 bird types where the birds are American Crow (amecro), Barn Swallow (barswa), Black-capped Chickadee (bkcchi), Blue Jay (blujay), Dark-eyed Junco (daejun), House Finch (houfin), Mallard (mallar3), Northern Flicker (norfli), Red-winged Blackbird (rewbla), Steller's Jay (stejay), Western Meadowlark (wesmea), and White-crowned Sparrow (whcspa) for multiclass classification. This information is obtained from Xeno-Canto (**1**), a world's database of bird sounds. Additionally, there are three raw sound test clips (mp3) that have been provided and that must be processed to know about the kinds of bird's sound in the test clips, which may contain voices of more than one kind of bird. The data must be transformed of these clips by using a model procedure and then utilize our 12-species neural network, which will help us to predict the calling birds. The analysis also highlights the specificity of the machine learning system in distinguishing one bird sound from the rest and the prospect of using the deep learning methods for addressing the classification processes in voice and natural language.

Overview:

The analysis utilizes the use of deep learning approaches and methods for the classification of bird species that are based on their voices. With the help of neural network models and linear regression models, we are trying to get some results by using both binary and multiclass classification tasks, targeting to separate the bird voices of specific species in a more accurate manner. The analysis provided the efficiency of these models, which obtain very high accuracy rates in the differentiation of distinct bird species. After this, next model also worked on the difficult part of having real-world test data and have tried to get the results as much as accurate as the model can. The results are the basis of the development of the monitoring and maintaining birds' data which can be helpful in further studies on the birds' voices and their ecological environment.

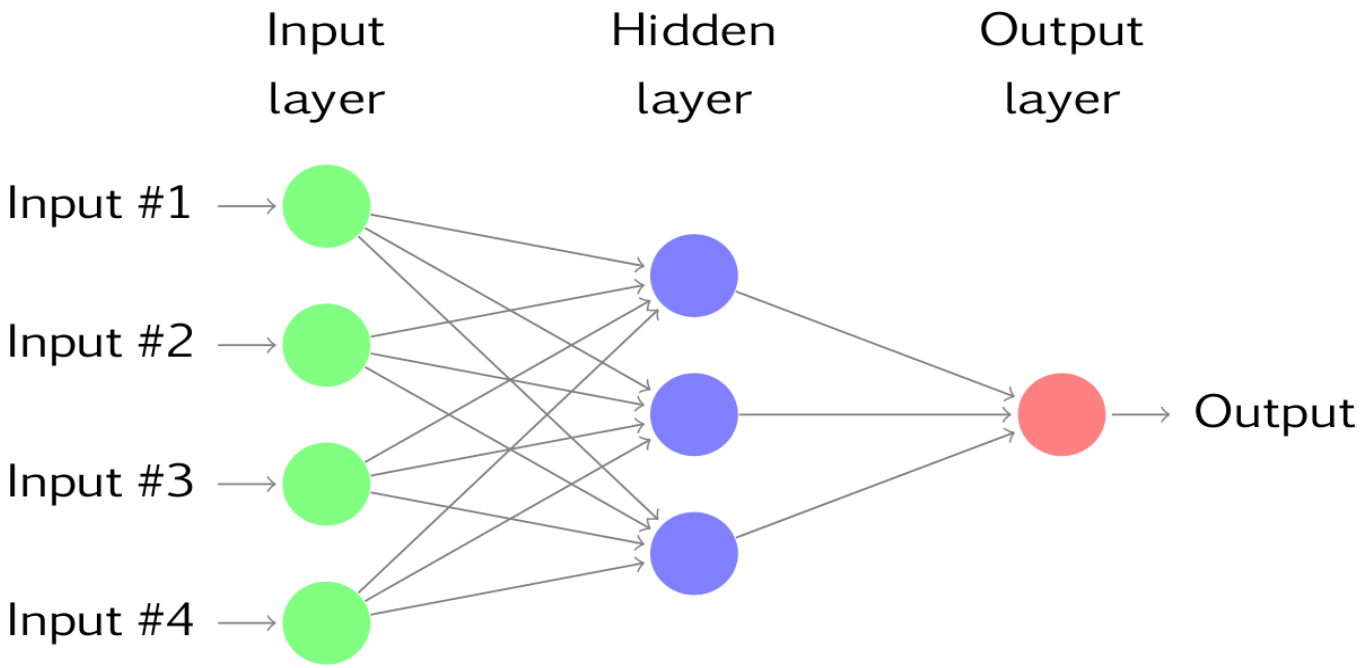
Description of the Dataset:

The provided dataset has 2 parts one is the precleaned data where we have been 12 birds' voices and the names of the birds are also well detailed and another set of 3 test data where there are 3 test voices of birds namely test1, test2 and test3. The first dataset is already cleaned and has been processed to directly work on. The data has been transformed into spectrograms which are again used for prediction for our data. The second dataset where there are 3 test data must be processed to start working on that. The data has been processed with 343 (time) and 256 (frequency) and has been flattened the data so that it should match the shape as we have the data as per the multiclass model. So, we have transformed the data into the size where it is easier to implement the model directly from the multiclass.

Theoretical Background:

Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions (**2**). Deep learning drives many applications and services that improve automation, performing analytical and physical tasks without human intervention. It lies behind everyday products and services—e.g., digital assistants, voice-enabled TV remotes, credit card fraud detection—as well as still emerging technologies such as self-driving cars and generative AI. Deep learning distinguishes itself from classical machine learning by the type of data that it works with and the methods in which it learns. Machine learning algorithms leverage structured, labeled data to make predictions—meaning that specific features are defined from the input data for the model and organized into tables. This doesn't necessarily mean that it doesn't use unstructured data; it just means that if it does, it generally goes through some pre-processing to organize it into a structured format. Deep learning eliminates some of data pre-processing that is typically involved with machine learning. These algorithms can ingest and process unstructured data, like text and images, and it automates feature extraction, removing some of the dependency on human experts. For example, let's say that we had a set of photos of different pets, and we wanted to categorize by "cat", "dog", "hamster", et cetera. Deep learning algorithms can determine which features (e.g. ears) are most important to distinguish each animal from another. In machine learning, this hierarchy of features is established manually by a human expert **(3).**

At the core of deep learning are neural networks, computational models composed of interconnected layers of nodes (neurons) that process input data to produce output predictions. These networks consist of an input layer, where data is received, hidden layers for computation, and an output layer for generating predictions (**6**). During training, weights between neurons are adjusted repetitively to minimize prediction errors. Neural networks have various designs specially made for different tasks and data types, including convolutional neural networks (CNNs) for image recognition, and recurrent neural networks (RNNs) for sequential data processing. Activation functions introduce nonlinearity to the network, enabling it to learn complex mappings between inputs and outputs. Common activation functions include sigmoid, ReLU (Rectified Linear Unit) and sigmoid for Binary Classifications and softmax for multiclass classification. These components and techniques make neural networks to learn special patterns and relationships in data, making them very important tools for tasks such as voice recognition, natural language processing, and predictive analytics. Some other models that must discussed are the CNN models which are very useful for transfer learning in the image classification tasks, using the features that were learned from the big datasets to help the knowledge transfer to the new tasks with the little labeled data. Some challenges of applying the models to the new domains is the careful fine-tuning and regularization which should be done to make the balance between the task-specific adaptation and knowledge transfer. The recurrent neural networks (RNNs), which are the neural networks designed for the sequential data processing such as time series prediction and natural language processing, are the best for the modeling of the sequential data with the variable-length inputs since they can capture the temporal dependencies and context. Training RNNs has its own set of problems such as long short-term memory (LSTM) and at the same time taking a lot of computational resources and a lot of parameter tuning is also required.



Some of the hypertunning parameters also include Convolutional Layer (Conv2D), MaxPooling Layer (MaxPooling2D), Flatten Layer, Dense Layer (Fully Connected), and Output Layer (Dense). In neural networks Conv2D layers are used for filtering the inputs, where basic features like edges and textures are induced. These layers, being convolutional, employ the ReLU activation function to enable the identification of nonlinear features of the input data. They extract feature maps from the large data set reducing its computational power and increases translation invariance and able to perform max pooling layers (MaxPooling2D). The Flatten layer changes high level feature maps into a one-dimensional vector so that Dense layers can analyze the obtained features. Both dense layers learn nonlinear transforms of inputs features and produce the mapping from features to target outputs and ReLU activation function helps in feature learning. The network layers are composed of the Convolutional and ReLU layers, which are followed by the pooling layers; the final layer, known as Dense, provides the final predictions and probabilities for classification through the softmax activation function. Some of the benefits include the ability of convolutional layers to learn features on its own, spatial parameters sharing, ability to provide translation invariance, the models capacity to have large nonlinear space, and general flexibility to undertake multiple tasks.

Methodologies:

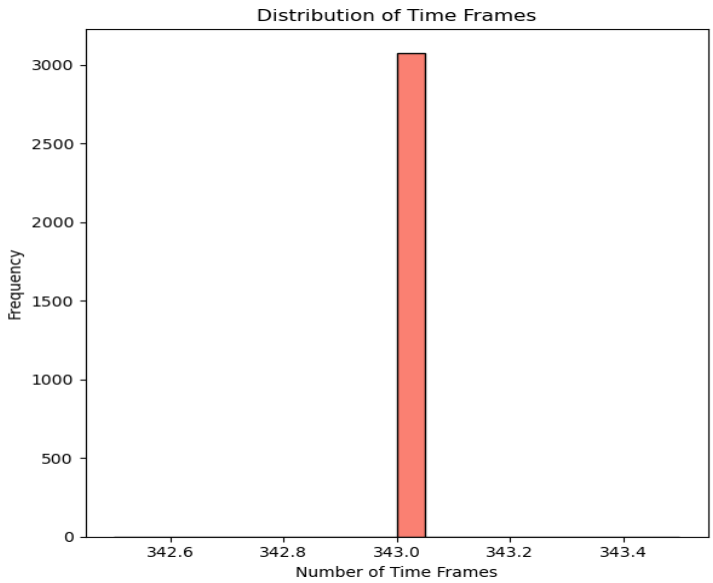
The analysis commences by loading the preprocessed dataset using the specified file path. This study focuses on a binary classification task aimed at distinguishing between the American Crow ('amecro') and the Blue Jay ('blujay'). Initial data preprocessing involves extracting sound samples of the bird's sound along with their corresponding labels from the dataset. The spectrogram lengths are normalized, reducing them to the maximum length found in the samples. The dataset is then split into training and testing sets, with 70% of the data allocated for training and the remaining 30% for testing the trained model. The data is structured to serve as input for a convolutional neural network (CNN) model. The CNN architecture comprises multiple layers, including convolutional layers followed by pooling layers and fully connected layers. These layers are designed to extract hierarchical features from the spectrogram inputs, down sample the feature maps, and learn important patterns and relationships between the features and the target classes. In the binary model, four convolutional layers with 256, 128, 64, and 32 filters, are used to extract features from the spectrogram inputs. Each convolutional layer is followed by a max-pooling layer, which reduces the spatial dimensions of the feature maps. The flattened feature maps are then used into two fully connected layers with 128 neurons. The output layer consists of a single neuron with a sigmoid activation function, producing a binary classification decision. For training the model, the Adam optimizer and binary cross-entropy loss function are utilized during the compilation of the model. Additionally, the model is trained twice, once with a batch size of 32 and again with a batch size of 64, to analyze how the batch size affects the model's performance. Training involves iterating over the data for a specified number of epochs, with the model's performance on the validation set monitored during training. At last, the trained model is used classifying bird species using the test set.

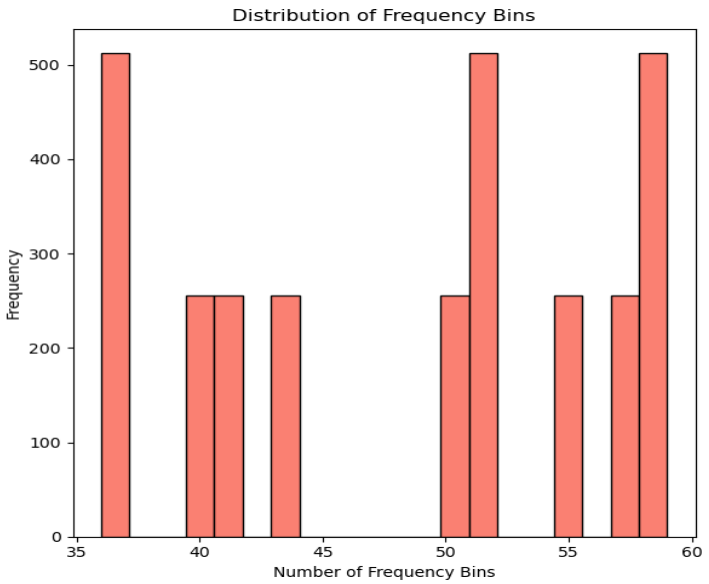
In the multiclass classification model, it involves the classification of data into 12 distinct bird species based on their voices. The audio data is provided with one-hot encoded labels, and spectrogram lengths are standardized. The dataset is partitioned into training and testing sets at a 70:30 ratio, with normalization by using StandardScaler. The multiclass neural network architecture comprises convolutional layers for feature extraction and max-pooling layers for down sampling. This design includes sequential convolutional layers with filter sizes of 256, 128, 64, and 32, followed by ReLU activation. Fully connected layers with 128 neurons each utilize ReLU activation for feature mapping, while the output layer employs softmax activation for multi-class classification. Model training has used Adam optimizer and categorical cross-entropy loss function, iterating over the dataset for multiple epochs, with performance assessed on the validation set. Additionally, the impact of batch size (32 & 64) on model performance is explored. Evaluation of the model's classification ability is conducted using a separate test set post-training.

Now, we have a separate dataset which consists of 3 raw data. The external test data, comprising three raw audio clips ('test1, 'test2', and 'test3'), which undergoes preprocessing to extract bird call segments. Utilizing the librosa library (4), audio files are loaded and segmented based on loudness levels. Segments longer than 0.5 seconds are processed to generate spectrograms using the mel-spectrogram method. Spectrograms are standardized to a fixed size, ensuring consistency across samples (5). The processed spectrograms are flattened and reshaped to match the input dimensions expected by the neural network model. Predictions are made using the trained multiclass neural network model, and the top 5 predicted bird species are identified along with their associated confidence scores. These predictions provide insights into the bird species present in each test audio clip. Finally, the predicted bird species and their corresponding confidence scores are printed for each test file, offering a comprehensive assessment of the classification outcomes.

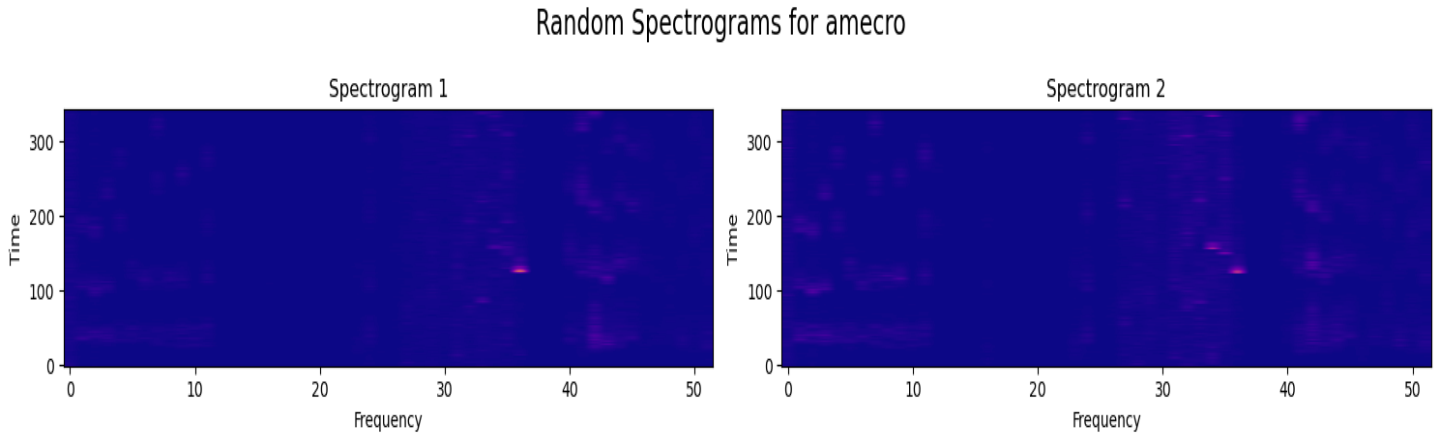
Computational Results:

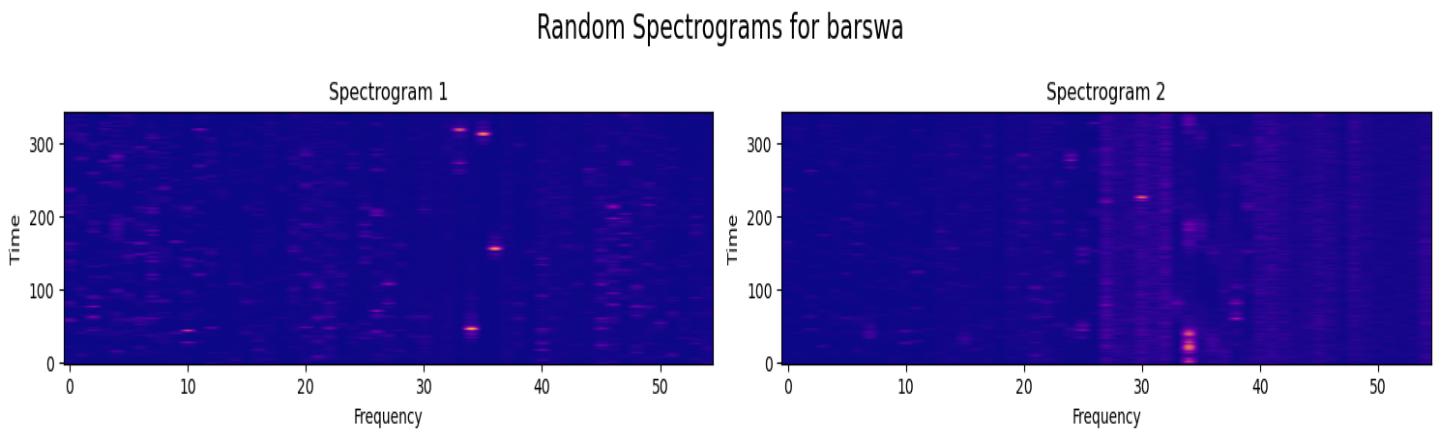
Data Exploration:

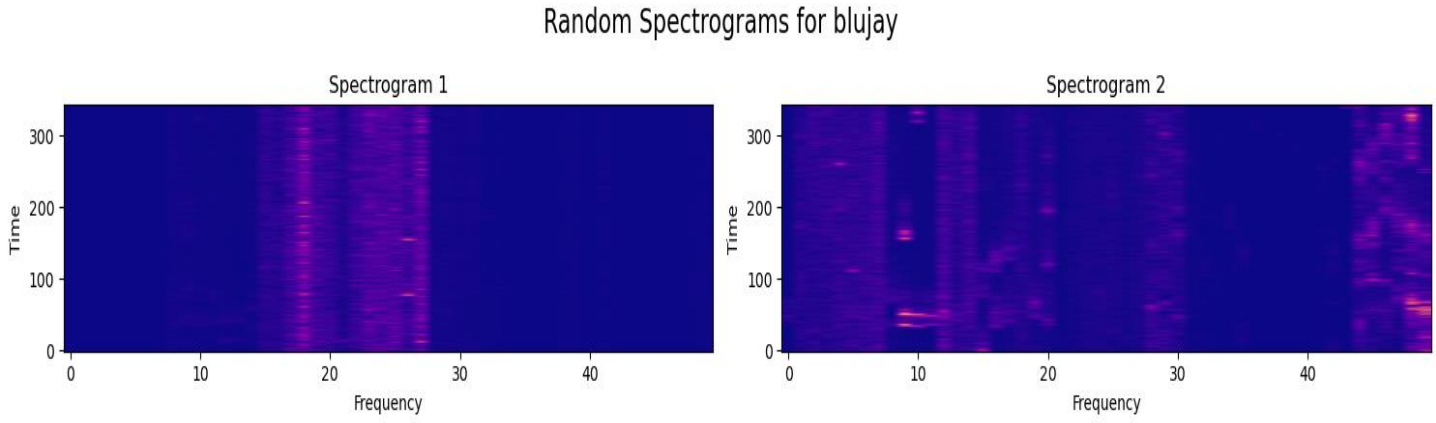


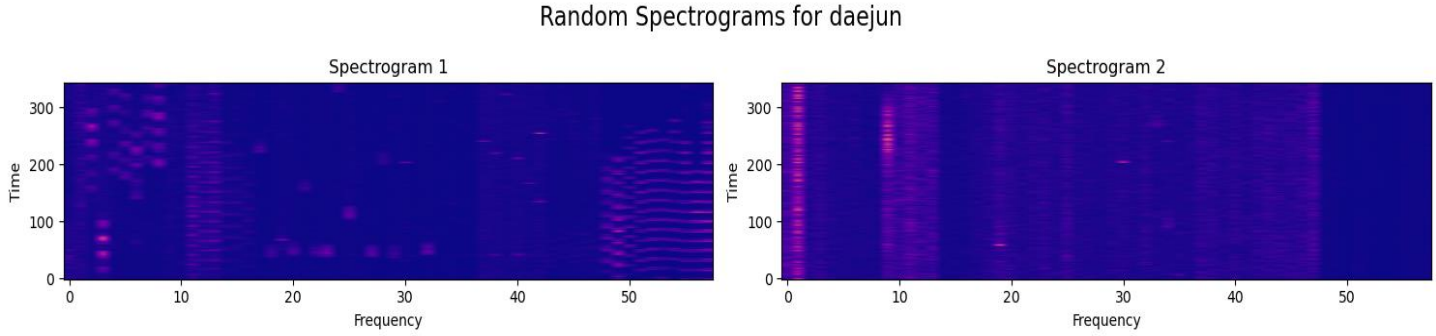


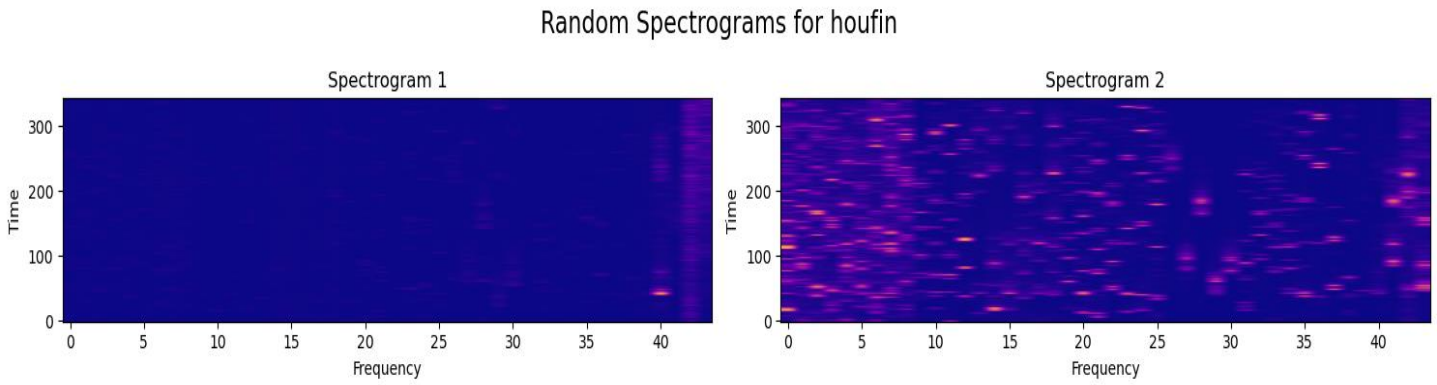
Exploring the Spectrograms for new Model:

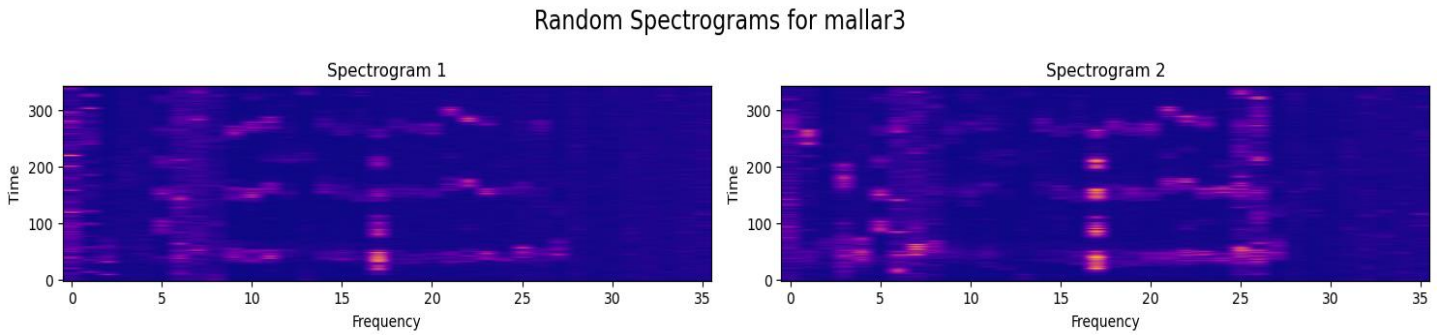


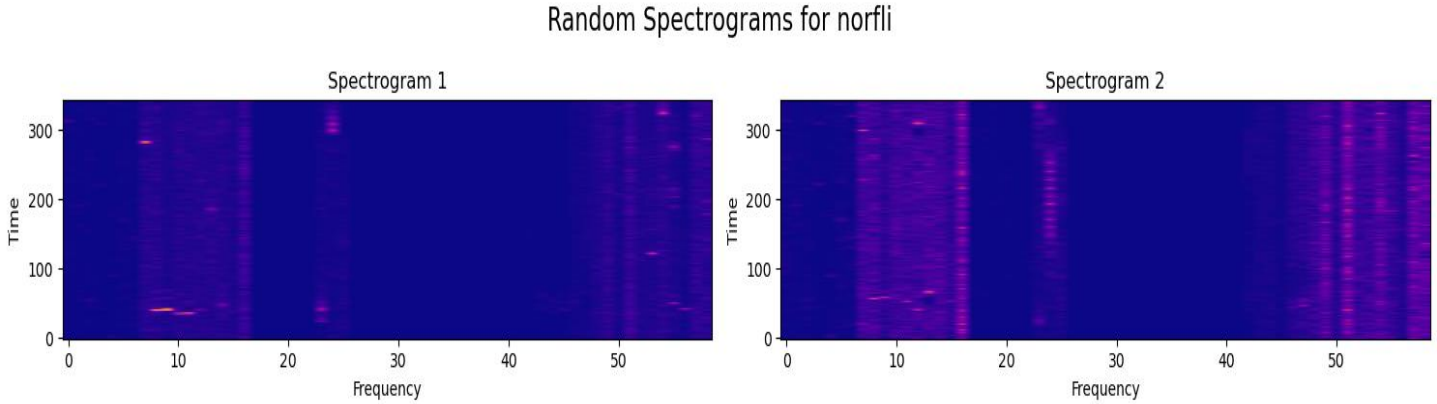


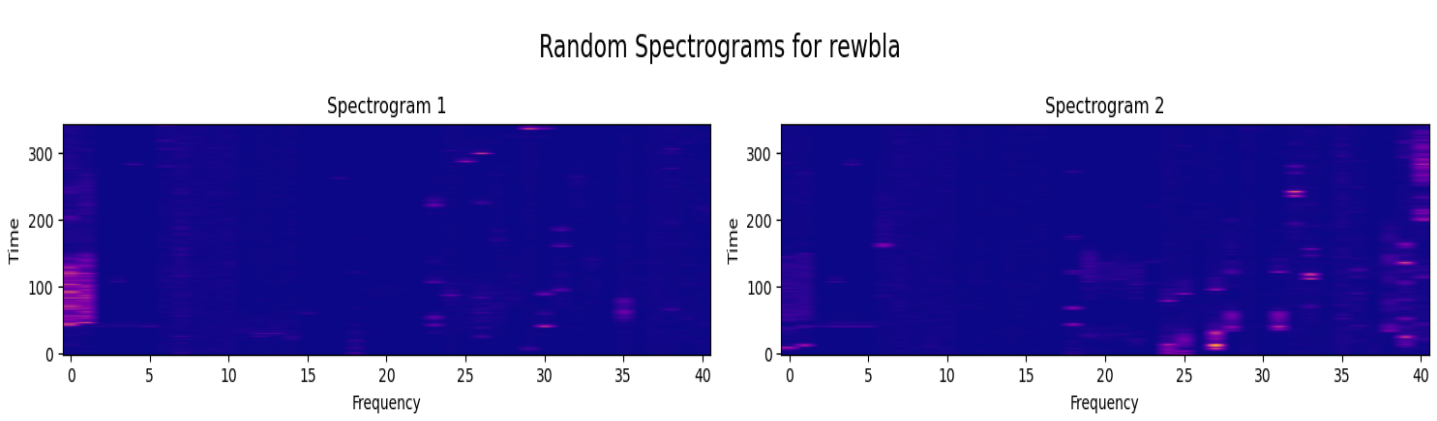


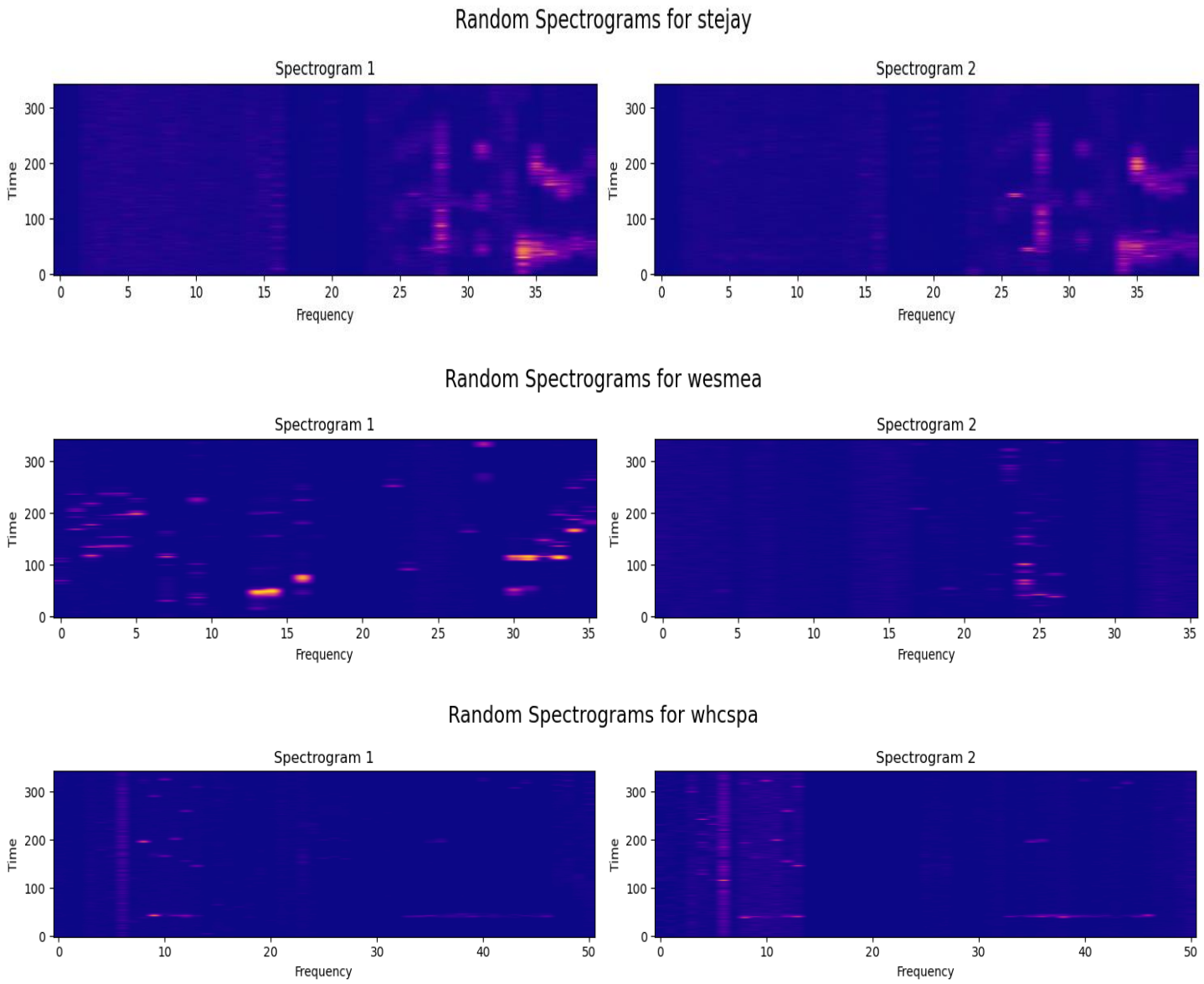



Random Spectrograms for houfin
Spectrogram 1
Spectrogram 2
Time
Frequency


Random Spectrograms for mallar3
Spectrogram 1
Spectrogram 2
Time
Frequency

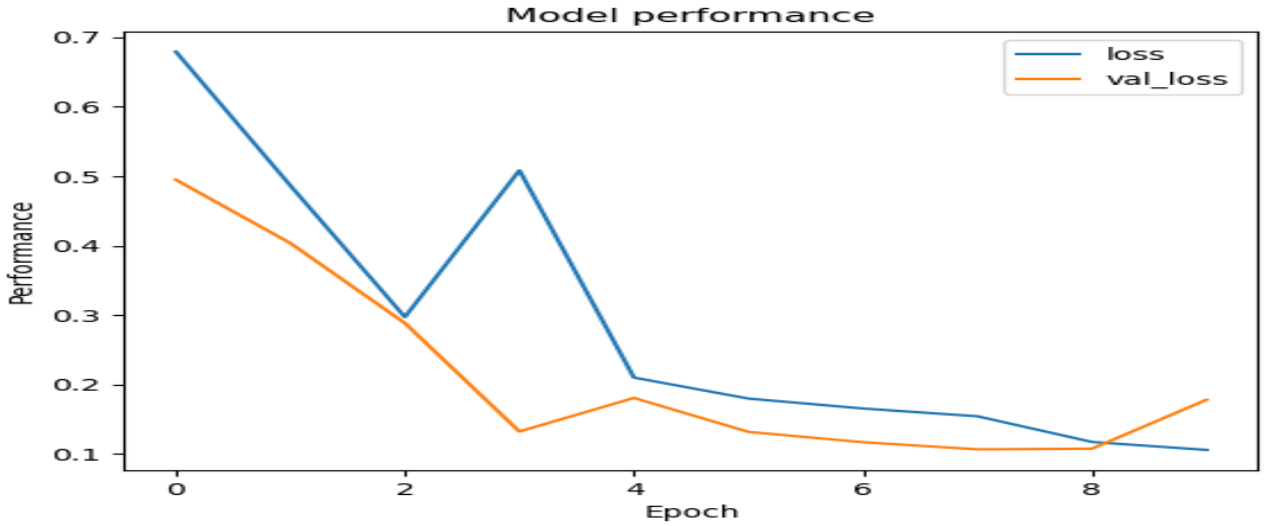

Random Spectrograms for norfli
Spectrogram 1
Spectrogram 2
Time
Frequency

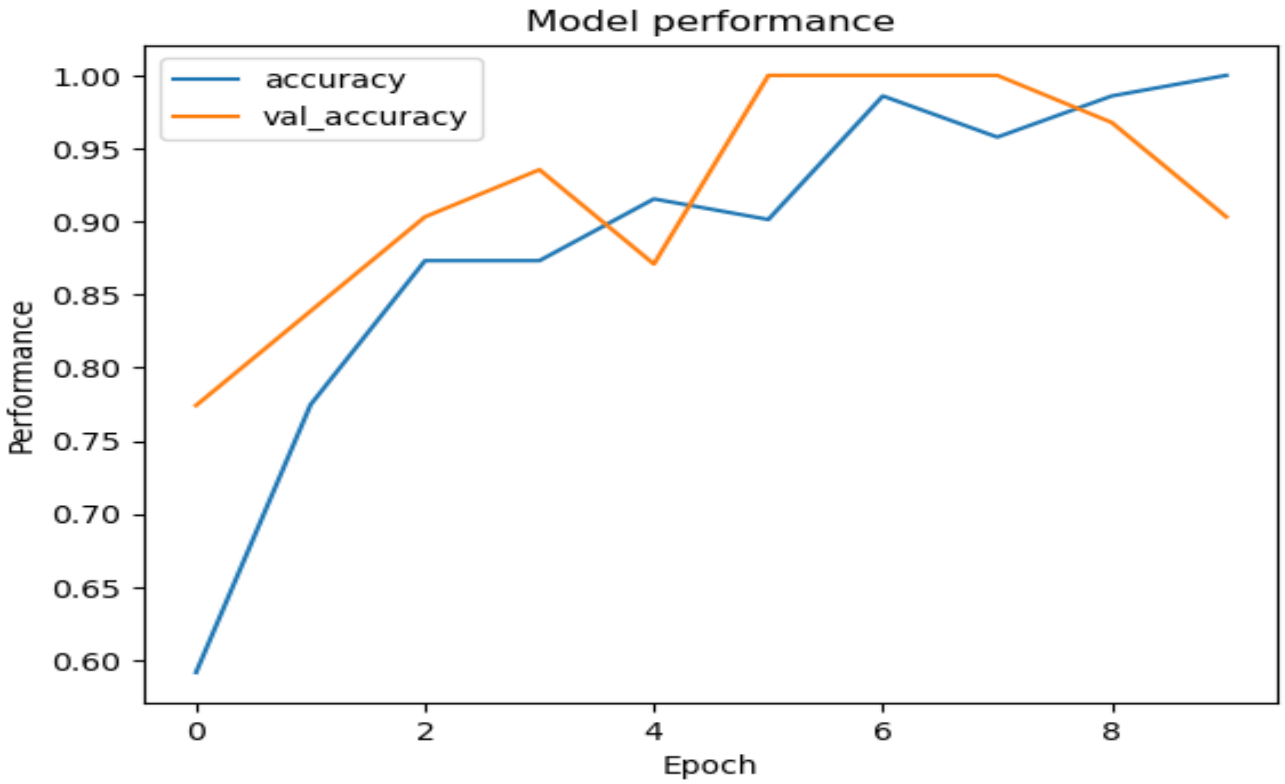

Random Spectrograms for rewbla
Spectrogram 1
Spectrogram 2
Time
Frequency



Binary Classification Models:

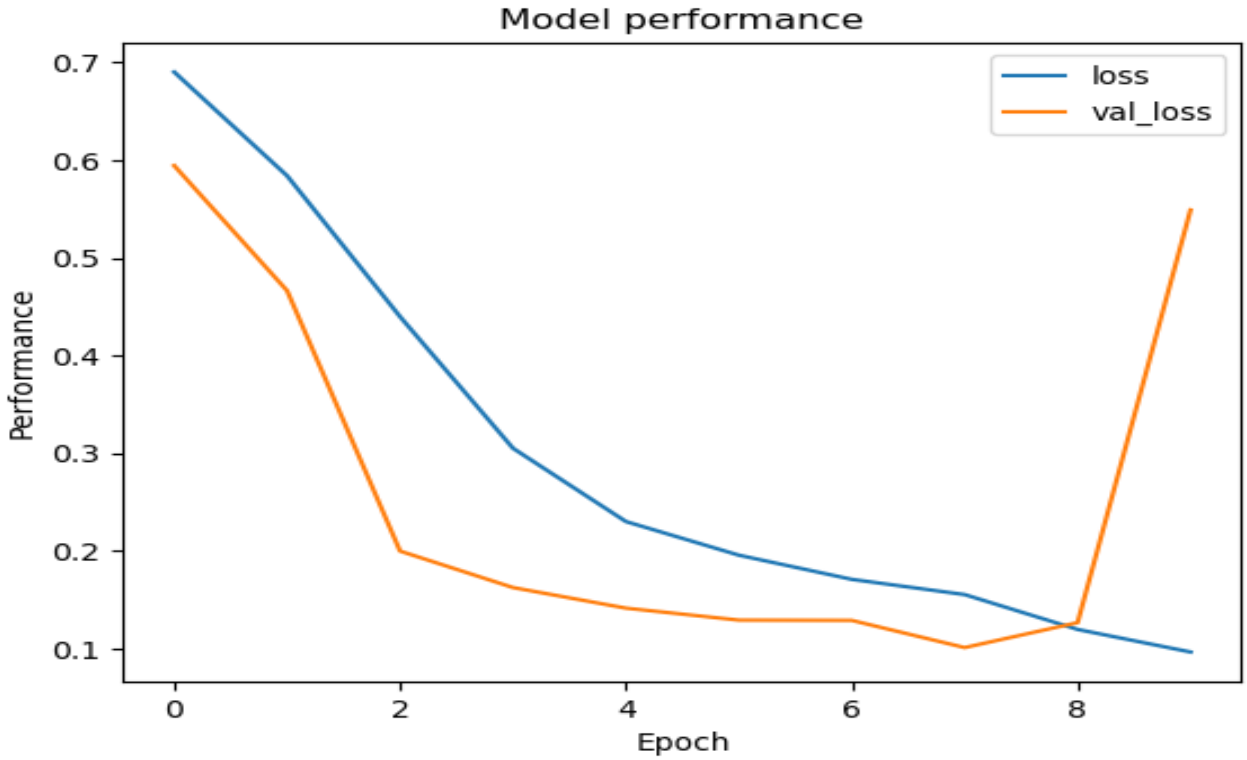
Model 1 (Batch Size: 32):

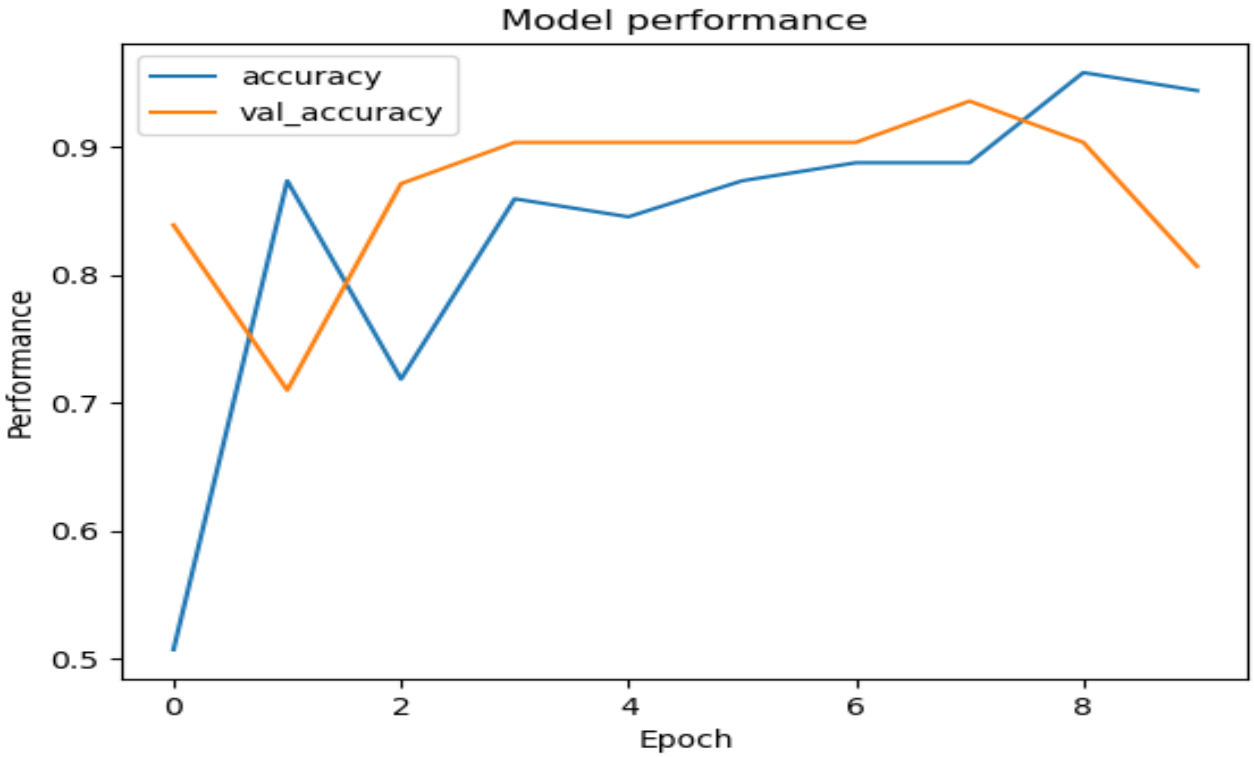




For the Neural Network Model 1, we have the following data:
Neural Network Model Accuracy: 90.32%
Wall time is 1min 20s.

Model 2 (Batch Size=64):

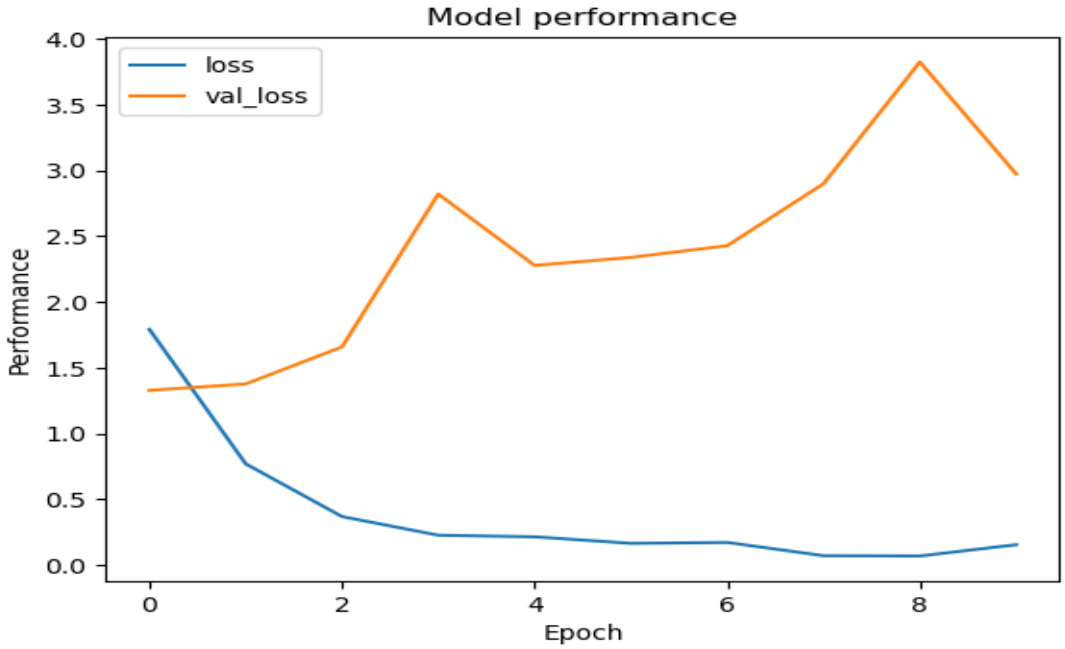


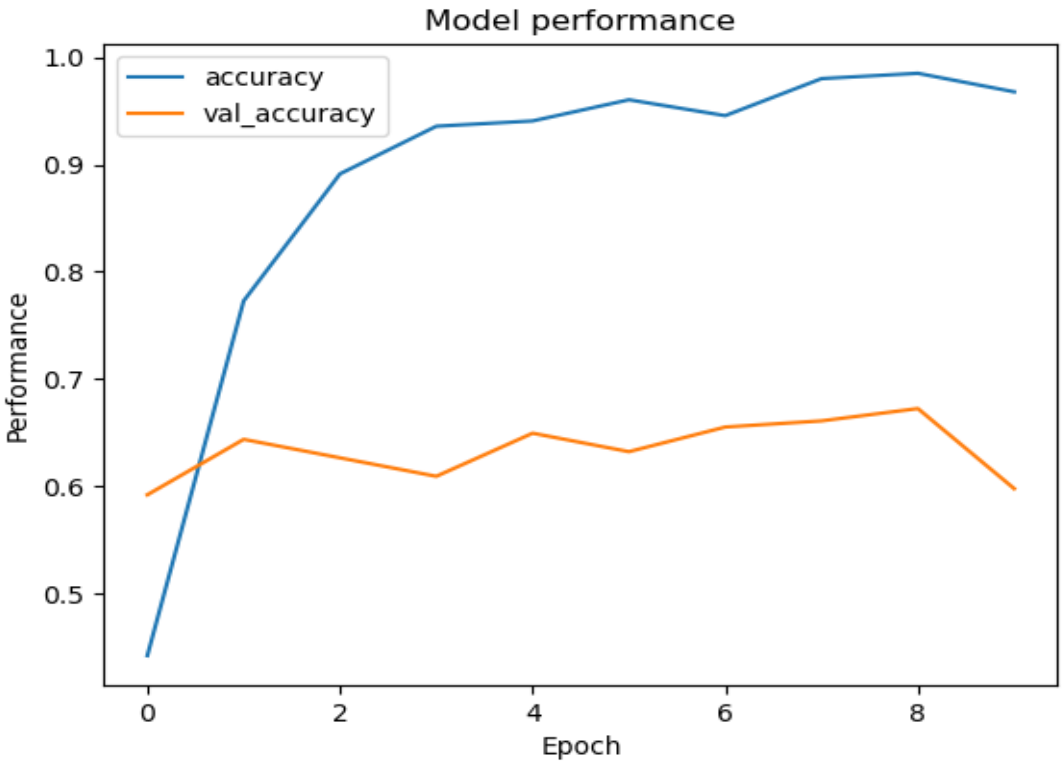


For the Neural Network Model 2, we have the following data:
Neural Network Model Accuracy: 80.64%
Wall time is 1min 20s.

Multiclass Model:

Model 1 (Batch Size: 32)

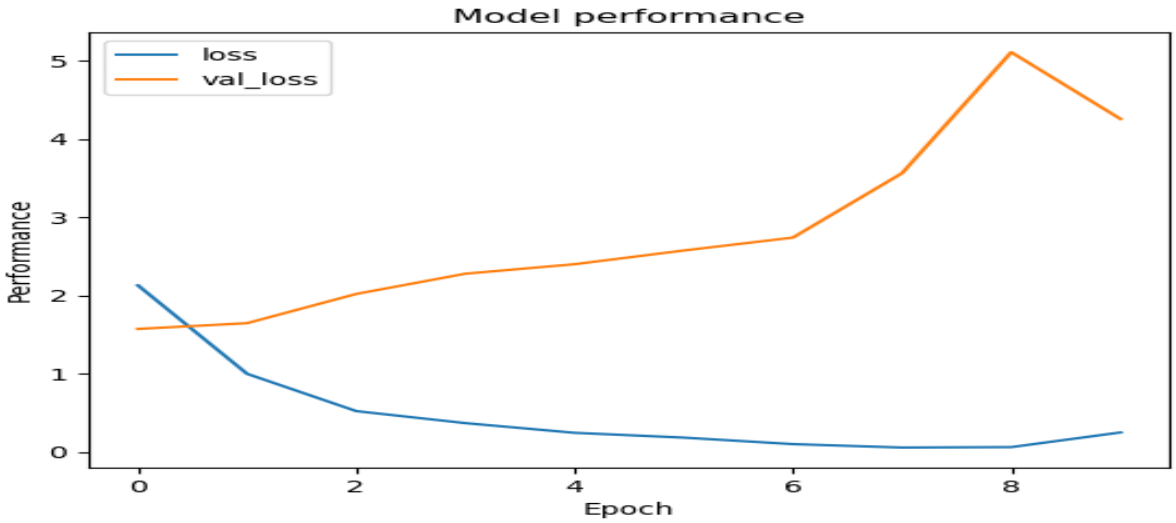


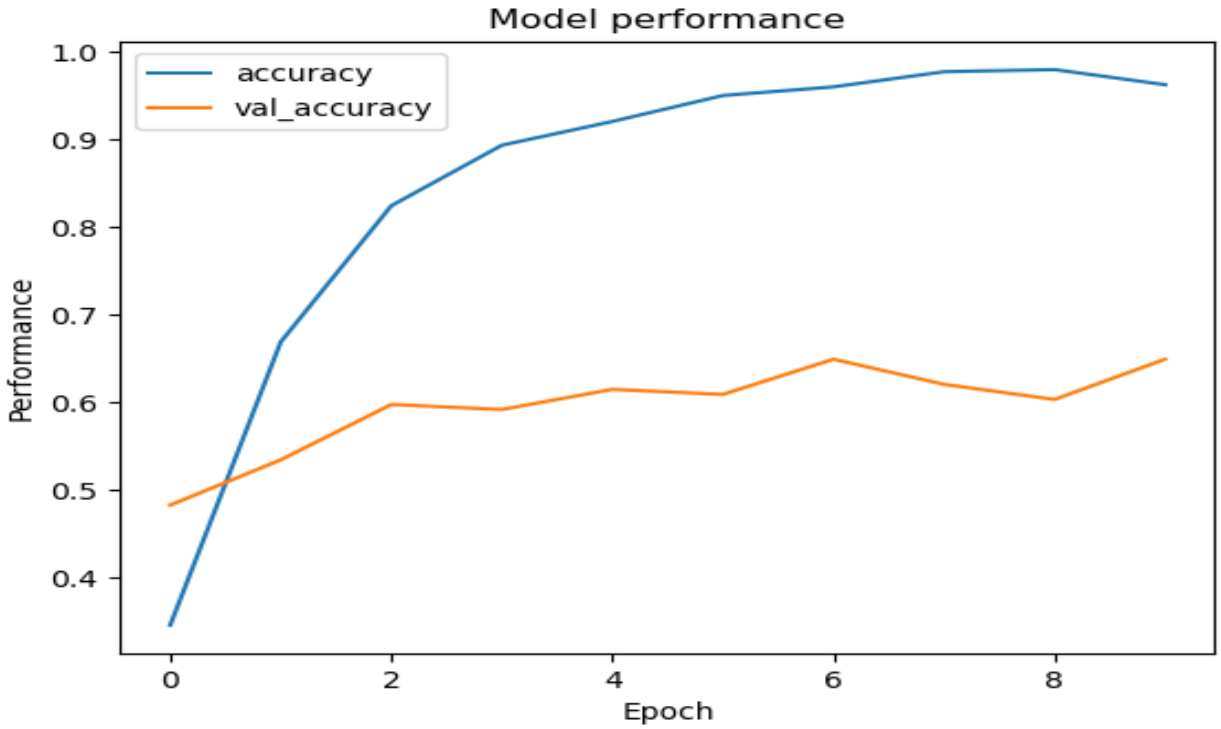


Results obtained from the Multiclass Model 1.

The Neural Network Model Accuracy is 59.77%
Wall Time: 7mins 47seconds.

Model 2 (Batch Size: 64)





Results obtained from the Multiclass Model 2.

The Neural Network Model Accuracy is 64.94%
Wall Time: 7mins 48seconds.

Final model with External Test Data:

Data	Predictions	Probability Percentage
Test 1	American Crow (amecro)	51.67%
	Western Meadowlark (wesmea)	44.45%
	Red-Winged Blackbird (rewbla)	2.18%
	White-crowned Sparrow (whcspa)	1.62%
	Northern Flicker (Norfli)	0.063%
Test 2	Western Meadowlark (wesmea)	86.05%
	American Crow (amecro)	13.51%
	Red-Winged Blackbird (rewbla)	0.33%
	White-crowned Sparrow (whcspa)	0.07%
	Northern Flicker (Norfli)	0.03%
Test 3	Western Meadowlark (wesmea)	49.77%
	American Crow (amecro)	47.55%
	Red-Winged Blackbird (rewbla)	2.15%
	White-crowned Sparrow (whcspa)	0.489%
	Northern Flicker (Norfli)	0.025%

Discussion:

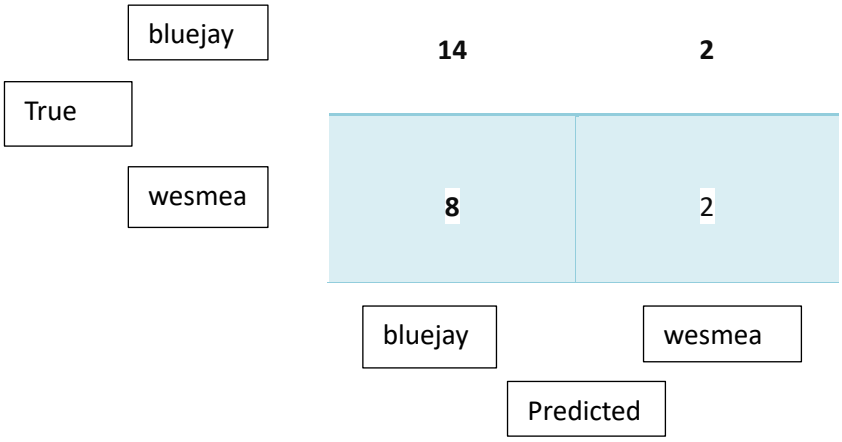
In the binary classification task, both developed models displayed high accuracy; specifically, Model 1 achieved 90.32% accuracy, while Model 2 achieved 80.62% accuracy. The training time for both the model is the same, 1 minute and 20 seconds. These high accuracy values mean that the given dataset was clean, and the chosen features indeed turned out to be very valuable for the task at hand. The models' speed and efficiency in learning and execution highlights the performance of the selected parameters and construction. But the model has some signs of overfitting even here, mainly because the model is very confident in its predictions, despite the high accuracy. This may be since the model over-fits the training data and thus does not generalize well or at all to the unseen data. The architecture included convolutional layers with decreasing filter sizes of 256, 128, 64, and 32, with corresponding max-pooling layers to reduce spatial dimensions. The model then flattened the output and passed it through two fully connected layers with 128 neurons each, before producing a binary output. This model had a total of 1,479,649 parameters, all of which were trainable.

For the multiclass classification, Model 1 achieved an accuracy of 59.77% with a wall time of 7 minutes and 47 seconds, while Model 2 achieved 64.94% accuracy with a wall time of 7 minutes and 48 seconds. The lower accuracy in multiclass classification compared to binary classification indicates the increased complexity and difficulty in differentiating between 12 classes. This suggests that while the neural network can handle complex patterns, there might be a need for more sophisticated techniques or additional data to improve performance. The models struggled to capture the nuances between similar classes, which could be due to insufficient training data for certain species or overlapping features in the spectrograms. The multiclass classification model employs convolutional layers with filter sizes of 256, 128, 64, and 32, followed by max-pooling layers for feature extraction and down-sampling. With an input shape of (256, 343, 1), it processes spectrogram representations of bird audio samples. The architecture includes flattening and dense layers with ReLU activation functions, and a final softmax layer for classification. This design enables the model to learn intricate patterns in bird calls. Training times vary but typically range from several minutes to converge to accurate results, making it computationally feasible for moderate-scale applications. In the multiclass classification models, convolutional layers followed by max-pooling layers were employed to learn features from spectrogram representations of bird audio samples. These models, with a total of 1,481,068 trainable parameters, aimed to classify bird species into twelve distinct categories.

In the analysis of the external test data, the neural network model provided predictions for multiple bird species, with each test file showing varying probabilities across different species. In Test 1, the predicted the American Crow (amecro) with 51.68%, followed by the Western Meadowlark (wesmea) at 44.45%. Additionally, there were smaller probabilities assigned to the Red-Winged Blackbird (rewbla) at 2.18%, the White-crowned Sparrow (whcspa) at 1.62%, and the Northern Flicker (norfli) at 0.063%. In Test 2, the model provided a prediction with Western Meadowlark (wesmea) at 86.05%, with a smaller probability for the American Crow (amecro) at 13.51%. Similarly, Test 3 predicted between the Western Meadowlark and the American Crow, with probabilities of 49.77% and 47.51%, respectively. The model also assigned minor probabilities to the Red-Winged Blackbird, White-crowned Sparrow, and Northern Flicker across all test instances. These predictions highlight the model's ability to differentiate between multiple bird species based on their audio characteristics.

In this analysis, some limitations were found, primarily related to model training time and computational resources required. The neural network models, especially those handling multiclass classification, demanded significant processing time, with training durations ranging from several minutes to over half an hour (depending on the system power). Additionally, while the models achieved high accuracy rates, they showed signs of overfitting. The multiclass classification models have timing issue as it took a lot of time to run and along with that the model has performed satisfactorily. The test data has more noise like voices other than the bird voice which makes it difficult to predict the correct the species. So overall this model's performance is being affected by some factors where it's not easy to understand which may be providing a different selection. This is also one of the issues in this model where the size and shape of the data becomes more complex, and it does not fit the multiclass model for we must flatten the data and even after doing so, we are not getting some definitive result. In the analysis, it has been analyzed that Blue Jay (blujay) and the Western Meadowlark (wesmea) were the most challenging species to predict and were frequently confused. The spectrograms of their calls showed overlapping frequency ranges making them difficult to distinguish. The Blue Jay (blujay) and the Western Meadowlark's voice shared almost same frequencies, complicating accurate differentiation. The confusion matrix generated also proves to be correct as we can observe that there are 2 misclassifications on Blue Jay (blujay) and 8 misclassifications on Western Meadowlark (wesmea) part. On the other hand, alternative models such as decision trees, SVMs which can be used for classification, logistic regression, gradient boosting machines, and k-NN can also be applied

for binary and multiclass classification tasks; however, neural networks have the advantages of complex data patterns, scalability, flexibility, and high performance. These characteristics are the reasons why neural networks are more suitable for these kinds of solutions, particularly when it comes to big, high-dimensional datasets and applications requiring the recognition of the patterns.



Conclusion:

The performed analysis revealed the effectiveness of the neural network models that can be used to classify bird species from the data containing audio signals. The binary classification models with batch size 32 and 64 have yielded an accuracy of 90.32% and 80.64% respectively. while the multiclass models got an accuracy of 59.77% and 64.94% for 2 batch sizes stated earlier. The results highlighted the relevance of neural networks in providing an accurate analysis of complex patterns in spectrogram representations of bird calls. The analysis had some challenges like over fitting and the computational requirements computational time) when running the models. Neural networks have not been completely replaced because of their capacity to adapt to high dimensional vector data and complex nonlinear identification. While in Part 3 the model's prediction has been very different as the model had predicted more than 2 birds voices in each test set. The model for Test 1 predicted the voices as 'American Crow' with 51.67% confidence along with 'Western Meadowlark' at 44.45%. For the second test, the class 'Western Meadowlark' was predicted with a probability of 86.05% while the at the second position the model had predicted the voice as 'American Crow' with 13.51%. The third test analyzed the class 'Western Meadowlark' at 49.77% while 'American Crow' took the second place at 47.55%. This analysis highlighted the effectiveness of the model in the classification of the bird species across different patterns. The test sets had a lot of noises which makes the model performance low to track the birds' voices. The analysis can be used as a foundation to automate bird species identification from audio data, and it is valuable for effective monitoring of biodiversity and conservation purpose.

References:

1. Xeno-Canto Bird Recordings Extended (A-M) Dataset
   https://www.kaggle.com/datasets/rohanrao/xeno-canto-bird-recordings-extended-a-m
   https://xeno-canto.org/

2. What is Deep Learning? Theoretical Background
   https://www.ibm.com/topics/deep-learning#:~:text=Deep%20learning%20is%20a%20subset,AI)%20in%20our%20lives%20today

3. What is Deep Learning? Theoretical Background
   https://aws.amazon.com/what-is/deep-learning/#:~:text=Deep%20learning%20is%20a%20method,produce%20accurate%20insights%20and%20predictions.

4. Introduction to LIBROSA by Technocrat Theoretical Background and Code
   https://medium.com/coderhack-com/introduction-to-librosa-912c2c109f41

5. Hey, You! What's That Bird? Theoretical Background and Code
   By https://medium.com/@aarroonn?source=post_page-----c1c1e262200a----------
   https://medium.com/digital-futures-publications/hey-you-whats-that-bird-c1c1e262200a

6. Neural Network Models Theoritcial Background
   https://otexts.com/fpp2/nnetar.html