Bird Sound Classification Using CNN

By : Pavan Arram

### **Abstract**

This project focuses on classifying Seattle-area bird species based on their audio calls using convolutional neural networks and transfer learning. Spectrograms were extracted from short audio segments and used to train models for both binary and multi-class classification across twelve bird species. Several CNN architectures were developed from scratch, and a pretrained MobileNetV2 model was fine-tuned for comparison. The MobileNet-based model achieved the best performance, with a validation accuracy of around 37.7%. Additionally , the project explored predictions on real world test recordings using overlapping spectrogram windows and softmax entropy to assess whether multiple birds might be present in a single clip.While the models showed consistent predictions, manual listening revealed overlapping calls, exposing limitations in the models ability to handle multi label situations. The overall process highlights key considerations in using deep learning for species classification from sound and suggests further improvements in label structure and data complexity .

### **Introduction**

Bird vocalisations carry important information that can be used to identify species, track their presence in specific habitats, and monitor biodiversity . with recent advances in deep learning ,spectrogram-based classification has emerged as a promising approach for analyzing bird sounds in real world settings . This project applies neural network models to the task of identifying bird species, each clip capturing the frequency content over a fixed time window . The primary objective is to build, train , and evaluate multiple models capable of recognising these species from audio alone. In addition to exploring standard CNN architectures . the project investigates the use of transfer learning through MobileNetV2 to determine whether pretrained models can improve performance. The challenge of imbalanced data became evident during multiclass training.The work also includes testing on unseen audio recordings and evaluating model confidence using entropy-based evaluation measures to detect the possibility of overlapping calls.

### **Theoretical Background**

The task of classifying bird species based on sound recordings can be framed as a supervised classification problem. As described in An Introduction to Statistical Learning with Application in Python, the goal in supervised learning is to predict an output variable Y based on one or more input variables X. In this project , each input observation is a mel spectrogram-a two-dimensional visual representation of a birds vocalisation while the response is the categorical label of the bird species.

The models used in this project rely in a class of nonlinear functions known as neural networks. we fit a model that approximates the true relationship between the inputs and class labels by minimizing a loss function across a set of training observations.

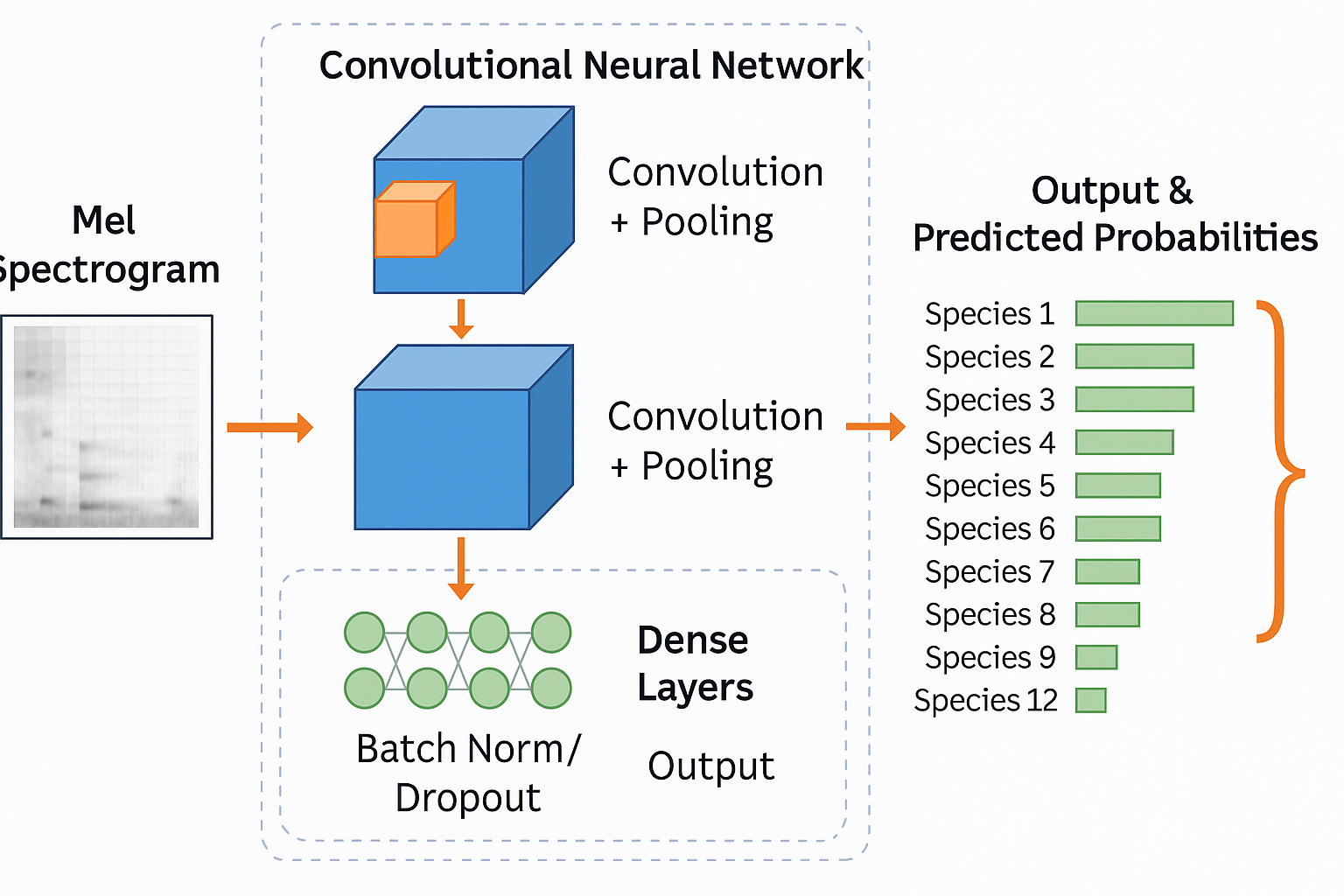
The spectrograms used as inputs were created by transforming 3-second audio clips into mel-scaled frequency plots. Each spectrogram is a matrix where rows represent frequency bins and columns represent time steps , making it analogous to an image . This structure motivates the use of convolutional neural networks(CNN), which are particularly well-suited for capturing spatial patterns in grid like data.

A CNN operated by applying filters that move across the input matrix and compute localised weighted combinations of values. These operations called convolutions, enable the model to detect patterns such as harmonics, frequency sweeps , or energy bursts that are characteristic of specific bird calls. Each filter generated a new “feature map“ that highlights where the corresponding pattern appears in the input . To reduce dimensionality and improve computational efficiency pooling layers are used to summarise sections of each feature map. Multiple layers of convolution and pooling allow the network to build up a hierarchy of patterns- from simple edges or tones in early layer to more complex acoustic signatures in later layers

The Network eventually flattens these representations and passes them through one or more dense layers, where each node is connected to every node in the previous layer. The output layer contains one node per class, and applies the softmax function to convert the final values into a probability distribution across all classes . The softmax function ensure the predicted probabilities sum to one, making them interpretable as relative likelihoods of class membership. The model is trained using categorical cross-entropy loss, which penalizes incorrect predictions more severe when the model is confident but wrong .

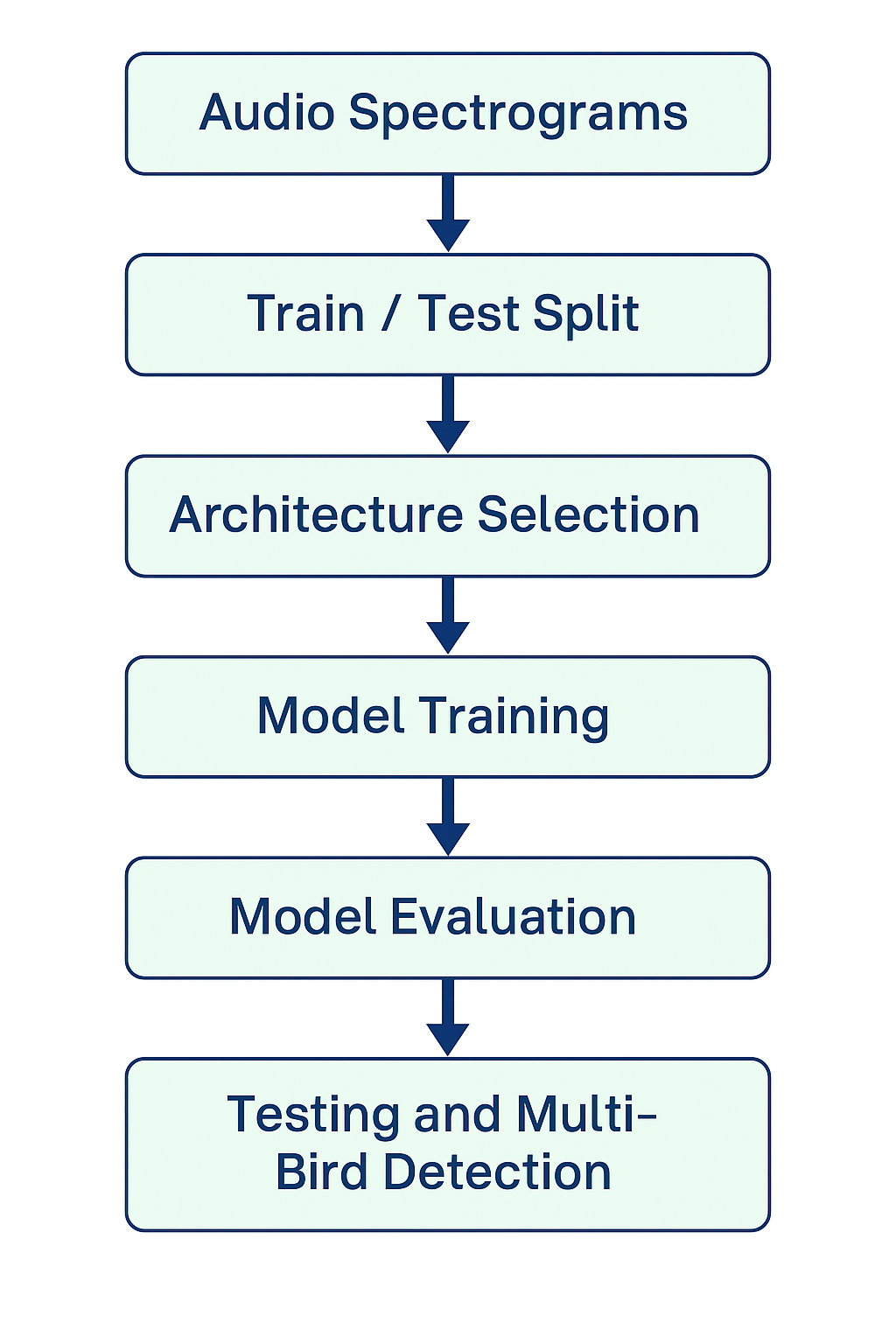
To further enhance model reliability , we introduced regularisation techniques such as dropout, which randomly disables a fraction of units during training , and batch normalization, which stabilises the learning process by standardising intermediate activations. These tools serve the same functions as ridge or lasso regularisation discussed in ISLRv2: to prevent overfitting and promote generalisation to unseen data.

Finally , for one model variant , we used transfer learning by adapting a CNN that had already been trained on a large, unrelated image dataset.This technique reused the lower-level structure of a complex network and retrains only the final layers for the specific task at hand, allowing us to benefit from pre learned features even when training data is limited.



*Figure 1 : Overview of CNN architecture*

### **Methodology**



The methodology for this project follows a structured pipeline, starting from spectrogram data and culminating in the final predictions and interpretation. The flowchart above illustrates this journey.

We begin with pre processed mel spectrograms, each representing a 3-second segment of bird audio. These spectrograms, already standardized to a shape of (128,517),are reshaped to (128,517,1) to serve as suitable inputs for convolutional neural networks. THe dataset includes spectrograms from twelve bird species with varying sample sizes. We apply stratified splitting to divide the data into training and validation sets, ensuring balanced class representation across both.

Once the data is loaded and reshaped appropriately we begin model development by first narrowing the task to a binary classification problem. This involves selecting two distinct species- in our case, amerob and norfil -and extracting an equal number of spectrograms from each. This step allows us to explore different CNN architectures in a simpler setup and helps prevent overfitting when working with a smaller data subset.

For each spectrogram , we normalize pixel intensities and one-hot encode the binary class labels. THese processed samples are then split into training and validation sets. Using this setup, we iteratively develop multiple versions of custom convolutional neural networks,In the final binary classification versions we introduced class weighting and drop out regularization,The model achieved ~71.4% validation accuracy,but displayed a class bias where one species (amerob) was consistently favoured in predictions, as revealed through the confusion matrix , starting with a simple structure and gradually increasing depth and regularization to improve generalization. Performance across versions is monitored using accuracy and loss plots to guide tuning .

After achieving satisfactory performance on binary task, we scale up to the multiclass classification setting involving 12 species. The same preprocessing pipeline is used , but this time across the full dataset. We train four progressively refined CNN architectures , each incorporating combinations of dropout, batch normalization, and global average pooling and class weights were applied in version 3 an 4 , LeakyRelu and Batch Normalisation and L2 regularisation were used in version 4 . These refinements are driven by performance metrics and visual patterns of overfitting or underfitting.

To push performance surther, we incorporate transfer learning by using MobileNetV2, a pre-trained convolutional model with frozen base and custom head originally trained on ImageNet. We strip its final classification layer and attach a custom head suited for 12-class prediction. This allows us to leverage rich feature extraction learned from a broad dataset while retraining only the final layer on our bird data. Spectrograms are sized to 224 x 224 RGB format to match the input expectations of MobileNet.Acheived the highest accuracy of ~37.7% with consistent learning curves and minimal overfitting.

Finally, we evaluate real\_world testing recordings provided as .mp3 files. These are clipped into 3-second segments and passed through the same preprocessing pipeline. Each segment is divided into overlapping 2- second spectrogram windows, allowing models to generate multiple predictions per clip. To detect the presence of multiple bird species , we apply an entropy- based heuristics : Softmax entropy plots combined with time windowed predictions revealed segments where the model predicted high uncertainty, supporting multi-bird hypothesis.

This end to end methodology from preprocessing through model tuning and uncertainty- aware evaluation - forms a complete pipeline for classification of bird species using deep learning and spectrogram analysis.

### **Computational Results**

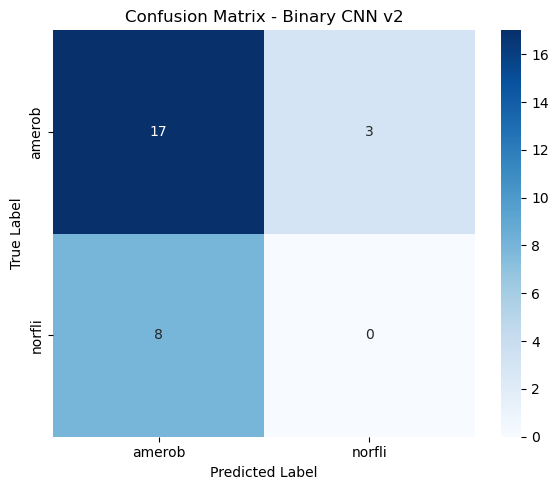
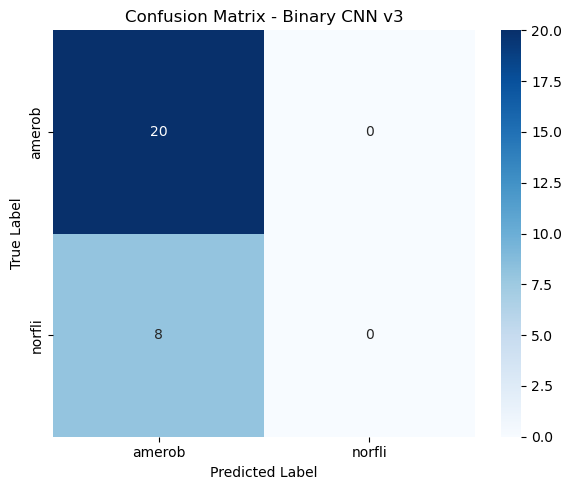
This section presents the performance evaluation of the models developed for classifying bird vocalisations.Results are organised into two primary tasks : binary classification between two species, and multiclass classification across twelve species. We also include comparative analysis of different model architecture and transfer learning performance.

**Binary Classification**

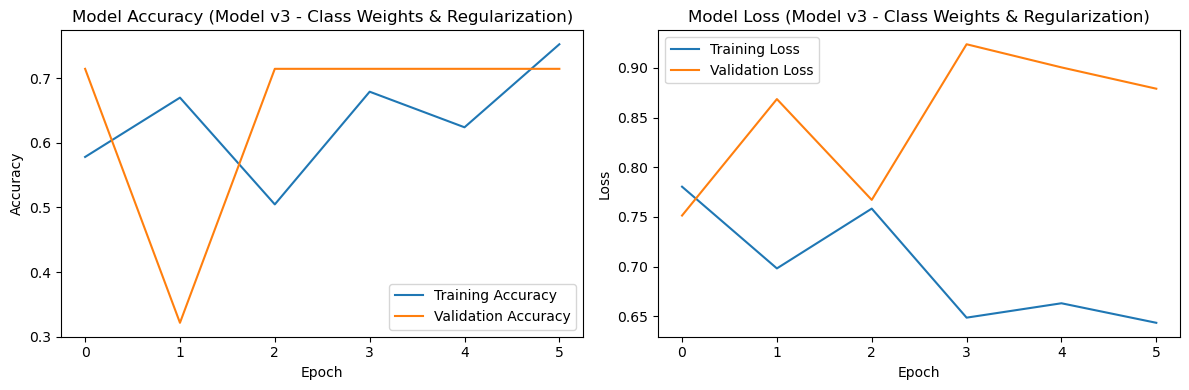
The binary classification task focused on distinguishing between two visually and acoustically similar bird species : American Robin and Northern Flicker. A series of CNN models version1 to version 3 were iteratively developed to enhance performance and address class imbalance .

The final version, Binary CNN version 3 , incorporated class weights and dropout regularisation . Training accuracy reached approximately 79.3%, with a validation accuracy of 71.4%. However, the confusion matrix revealed that While the model consistently identified amerob samples correctly, it frequently misclassified norfli samples. This was visually evident in the generated confusion matrix plot , where all true norfli instance were placed in the amerob column.

Despite balanced input, this misclassification trend suggested that future improvement may require either data augmentation for underrepresented species or architectural changes that enhance feature separation.

****

***Figure:3 Confusion matrix for binary CNN model version 3 Figure:4 Confusion matrix for binary CNN model version 2***

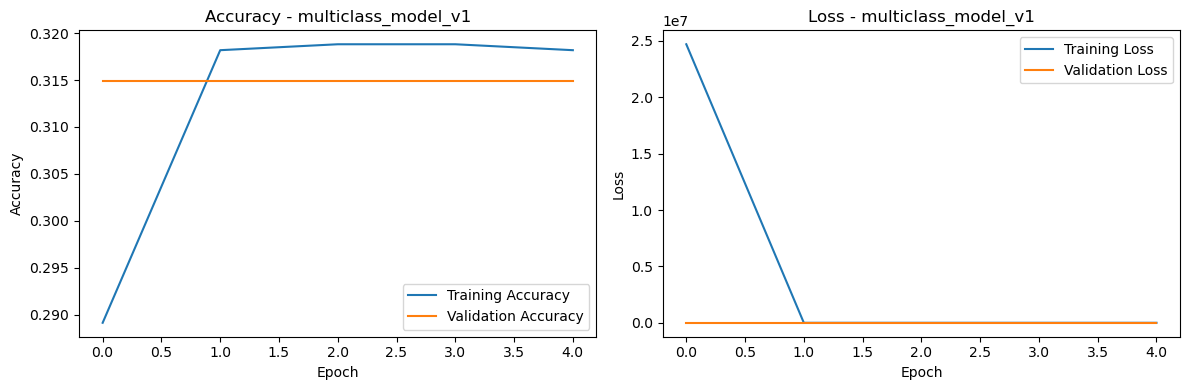
******

***Figure:5 Model accuracy and Model loss plots for binary CNN model version 3***

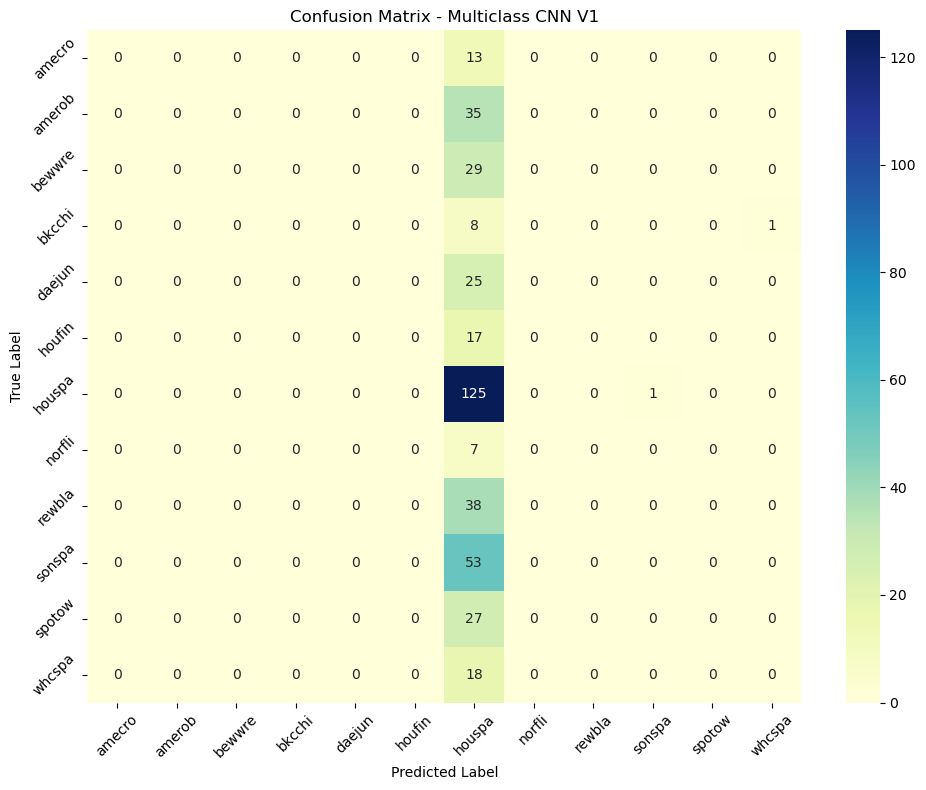
**Multi-Class Classification**

The multi-class problem expanded the task to classify 12 bird species. Models version -1 through version 4 were developed using custom CNN architectures with progressively increased depth,dropout, and batch normalization. Each model’s training was monitored using accuracy/loss plots and confusion matrices to detect underfitting, overfitting or bias.

CNN version 1 suffered from numerical instability ( validation loss > 44), indicating gradient explosion due to unnormalised inputs or poor initialisation. As a result, although the model reported moderate training accuracy, it completely failed to converge in the validation phase. This version served as a diagnostic baseline, highlighting the need for improved regularization and architectural refinement.

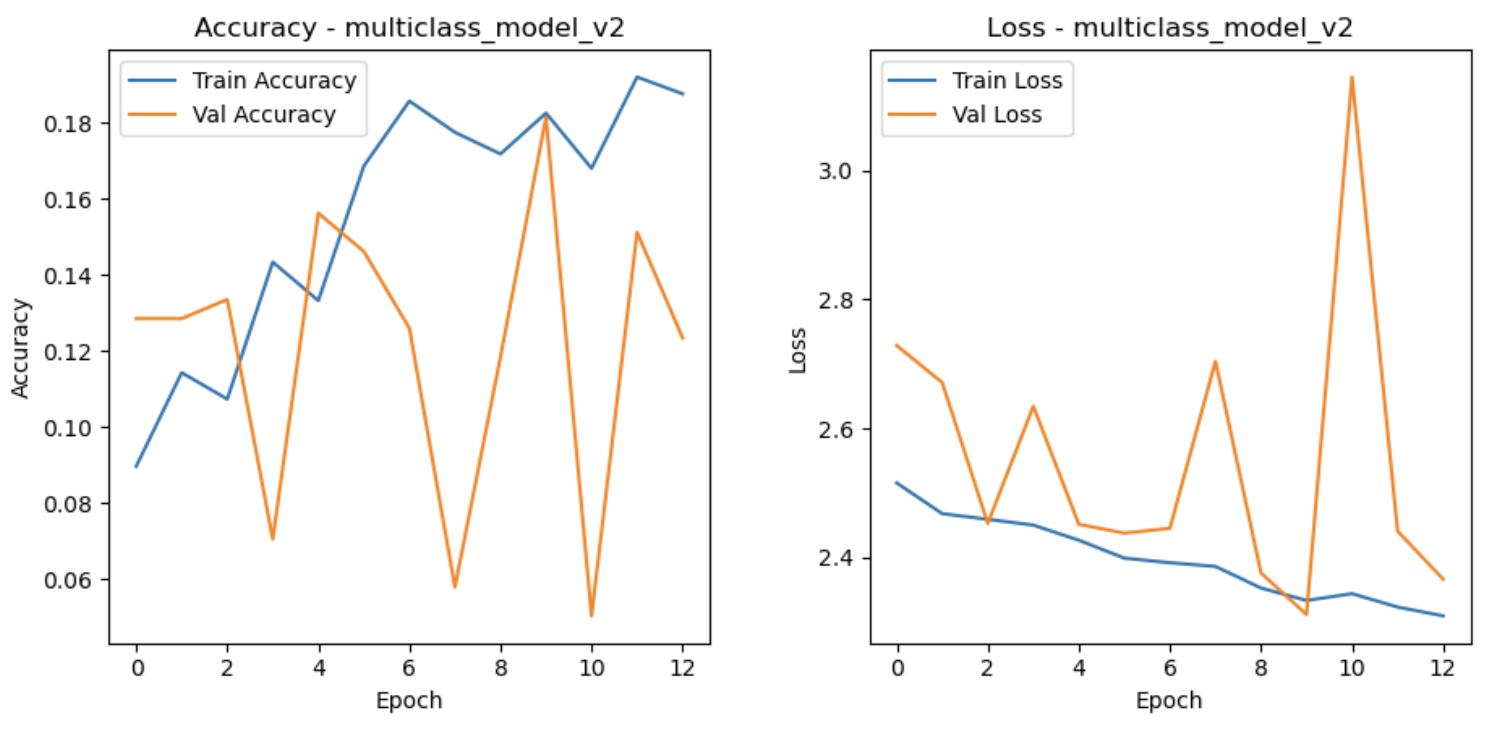


*Figure:6 Accuracy and Loss plots of CNN version -1*

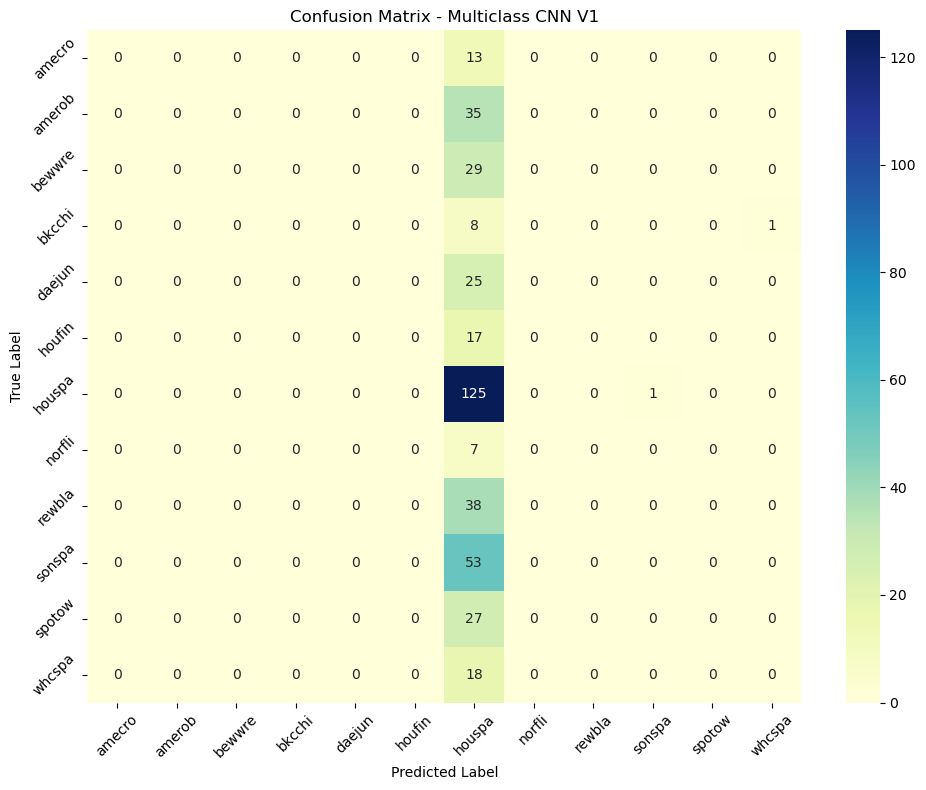


*Figure : 7 confusion matrix heatmap of CNN version - 1*

CNN version 2 showed improved numerical behaviour , but remained undercut with validation accuracy around 18% and with validation loss dropping to a more interpretable range of ~2.5 to 3.0. These metrics indicated that although the training process became more stable, the model capacity was still insufficient to capture the complex acoustic features needed to discriminate between twelve distinct bird species.

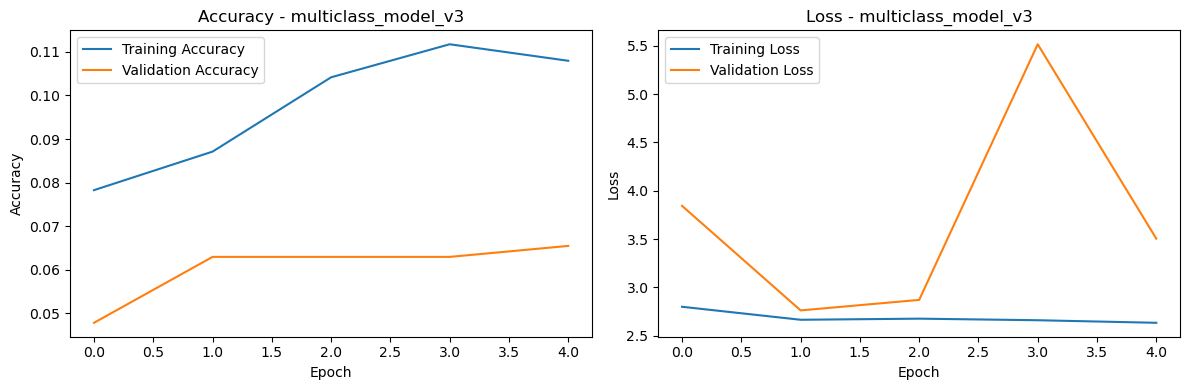


*Figure : 8 Accuracy and Loss plots of CNN version - 2*

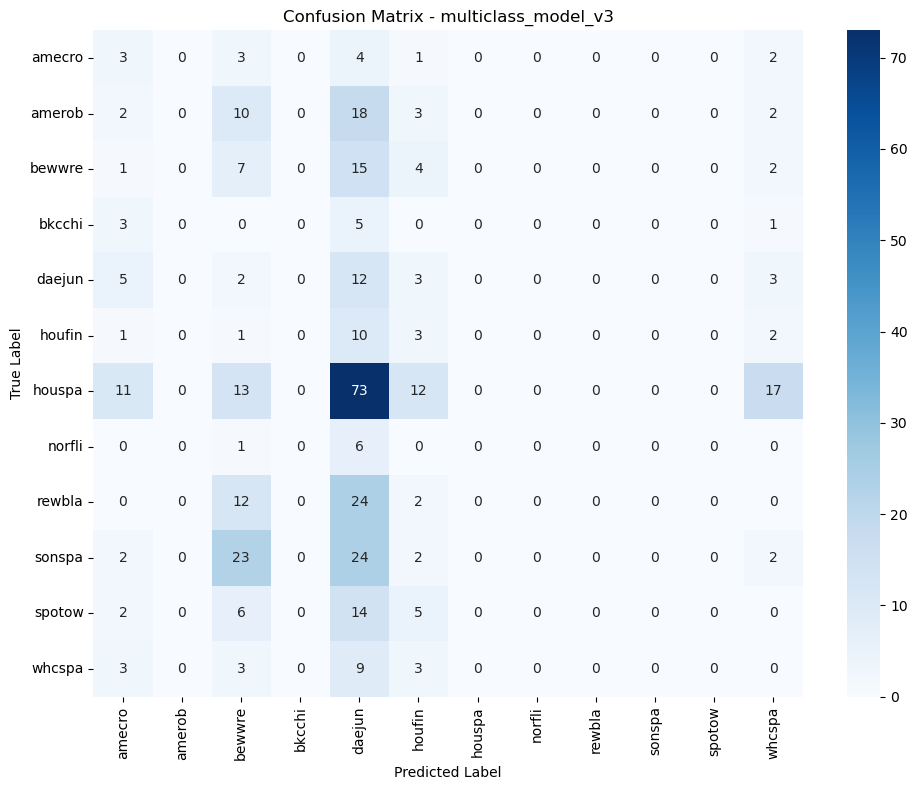
**

*Figure : 9 confusion matrix heatmap of CNN version -2*

CNN version 3 had increased architectural depth and introduced additional convolutional layer with regularization . This led to modest improvement in training dynamics, with early epochs showing reduced variance in loss and improved optimization behaviour. Nonetheless , model initially plateaued at around 13% accuracy, and while loss curves were stable, it failed to generalise to the full 12- class task.The confusion matrix confirmed this issue, revealing a strong bias towards certain frequent classes while neglecting minority classes - a sign of inadequate feature separation and imbalance handling.

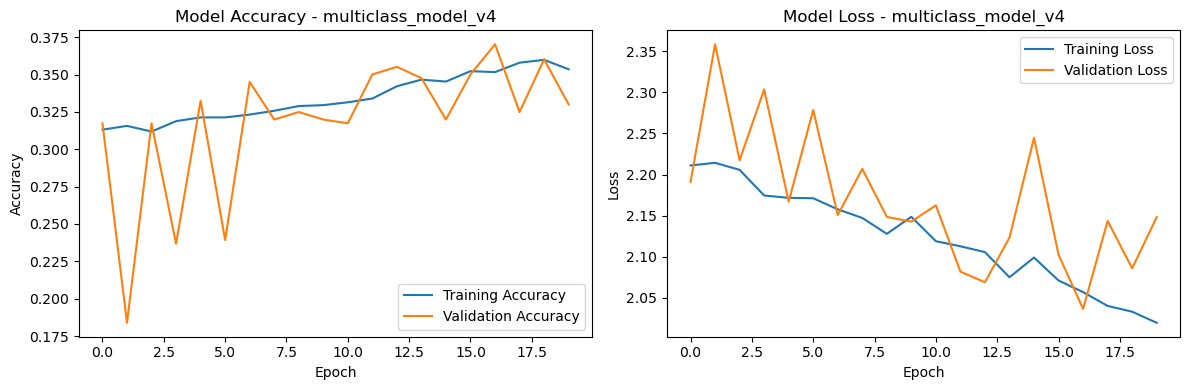


*Figure : 10 Accuracy and Loss plots of CNN version - 3*

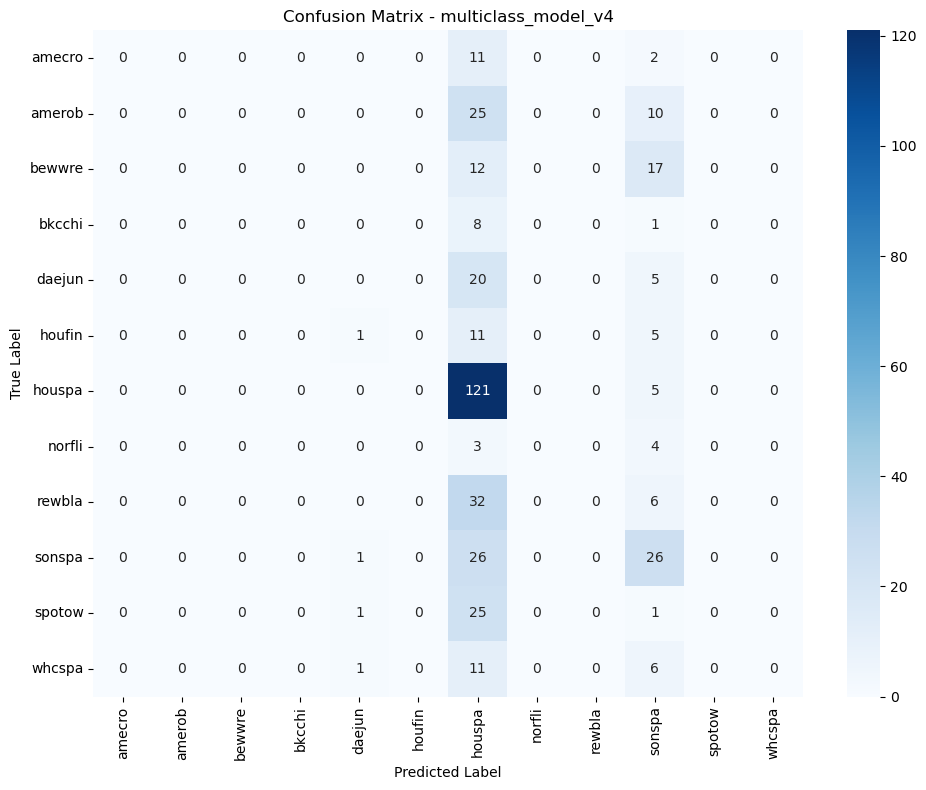
**

*Figure : 11 confusion matrix heatmap of CNN version -3*

CNN version 4 with further enhancements, including systematic dropout, class balancing, and early stopping with entropy based validation monitoring ,reached a validation accuracy of 37%, offering stable generalisation across all species. Its confusion matrix displayed more diagonal structure , indicating improved correct classification rates across classes.This version offered the most table generalisation prior to the introduction of transfer learning.



*Figure : 12 Accuracy and Loss plots of CNN version - 4*



*Figure : 13 confusion matrix heatmap of CNN version - 5*

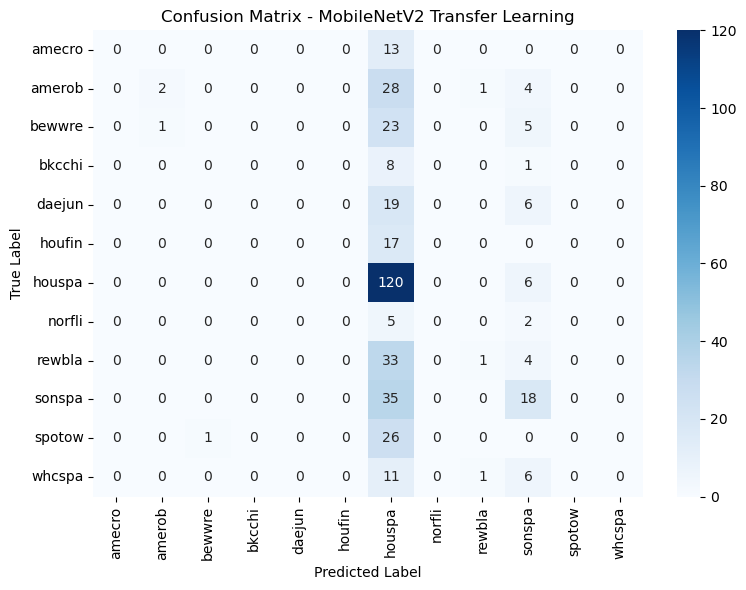
To explore whether leveraging pretrained representations could improve performance on bird species classification, we integrated MobilNetV2 as a transfer learning baseline. The base model was initialised with ImageNet weights, and only the final classification layers were fine tuned for the 12 - species task.

As shown in the training and validation accuracy/loss plots, MobileNetV2 displayed the most stable and steadily improving performance among all models tested. Training began with a modest accuracy of approximately 34.6%, and validation accuracy closely tracked it at ~32.7%. Over 20 epochs, the model achieved a peak validation accuracy of 35.5% and training accuracy of 38.6%, outperforming all previous CNN variants in generalisation ability.



*Figure : 14 Accuracy and Loss plots of MobilNetV2*

The confusion matrix of MobileNetV2 further confirms , Unlike earlier CNNs, predictions were dominated by a single class , this model showed a broader diagonal structure, correctly classifying species such as amerob , bewwre , rewbla , and sonspa to a greater extent. Misclassification, though still present, were more distributed and less biassed than in earlier models.

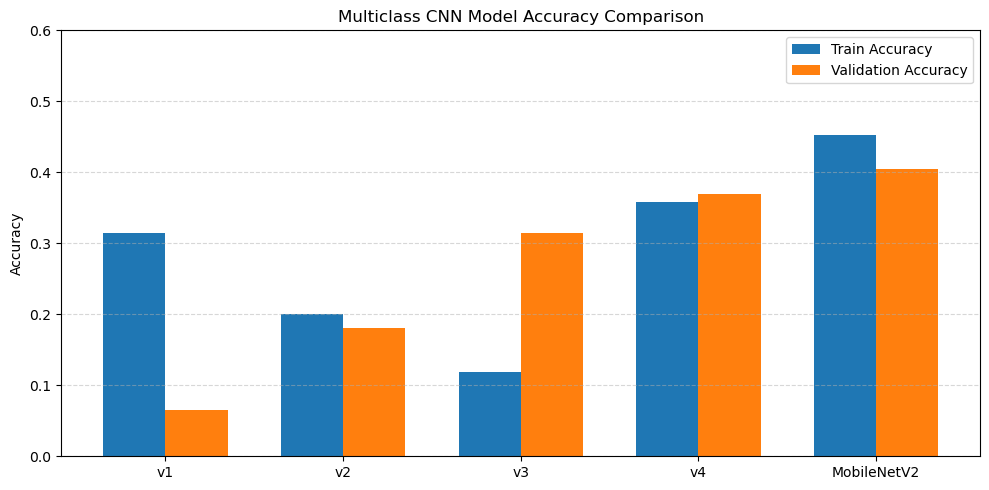


*Figure : 15 confusion matrix heatmap of MobileNetV2*

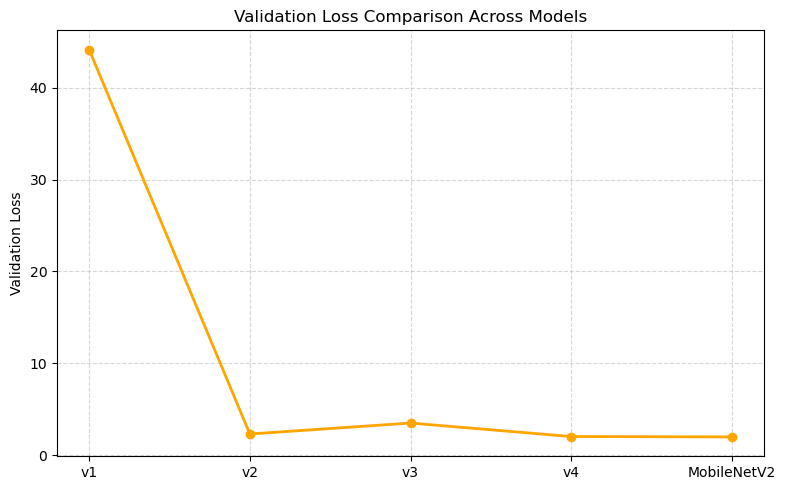
**Performance Comparison**

Tocompare model effectiveness, we plotted validation accuracy and loss across all versions of CNN and MobileNetV2. Figure : 16 visualizes the validation accuracy for CNN models and MobileNetV2.

Figure : 17 visualizes validation loss across same set of models. These comparisons clearly show that MobilNetV2 , used with transfer learning and frozen base layers , out performed custom CNNs.



*Figure : 16 Bar plot comparison of CNN models and MobileNetV2*

**

*Figure : 17 Line plot of validation loss across models*

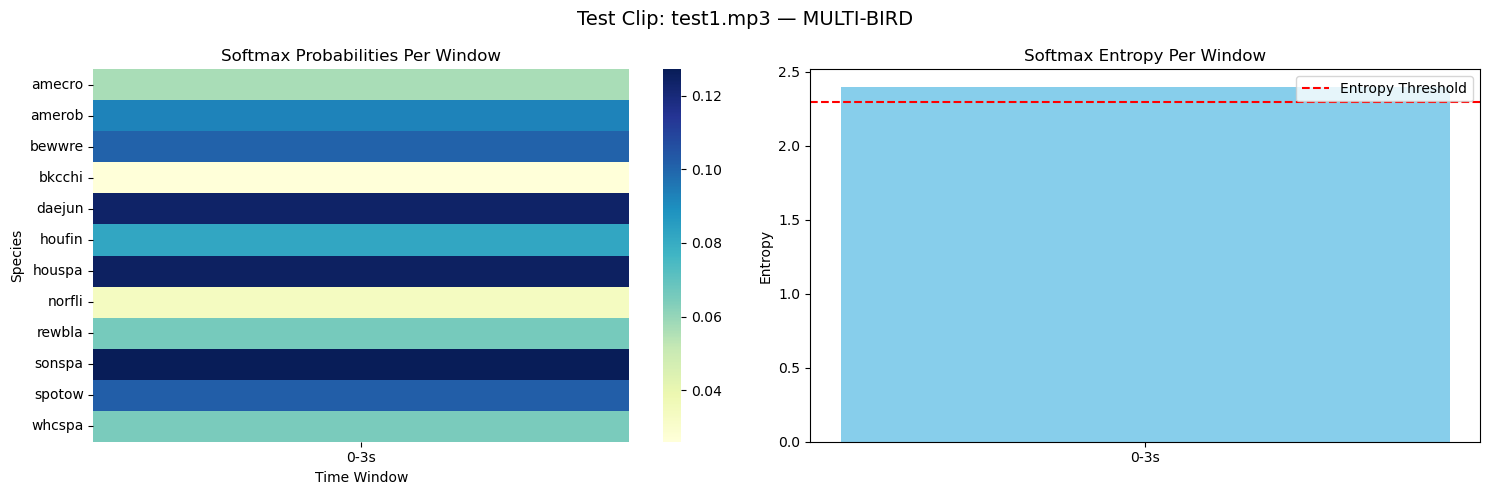
**Evaluation on Test Clips**

To evaluate model predictions on real-world audio , we used Three test.mp3 clips and applied our preprocessing pipeline to generate overlapping 3- second spectrogram windows. For each window, the MobileNetV2 model predicted class probabilities via softmax.

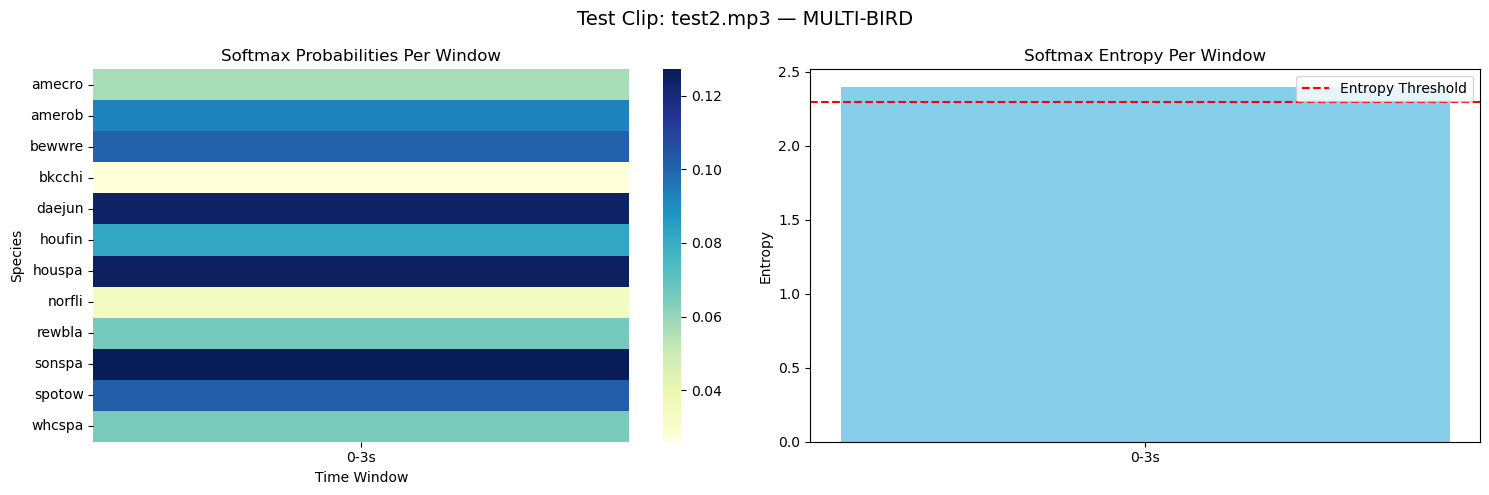
Entropy was computed for each softmax output to assess uncertainty . If multiple species were predicted across windows or any entropy values exceeded 2.4, the clip was flagged as potentially containing overlapping birdcalls.

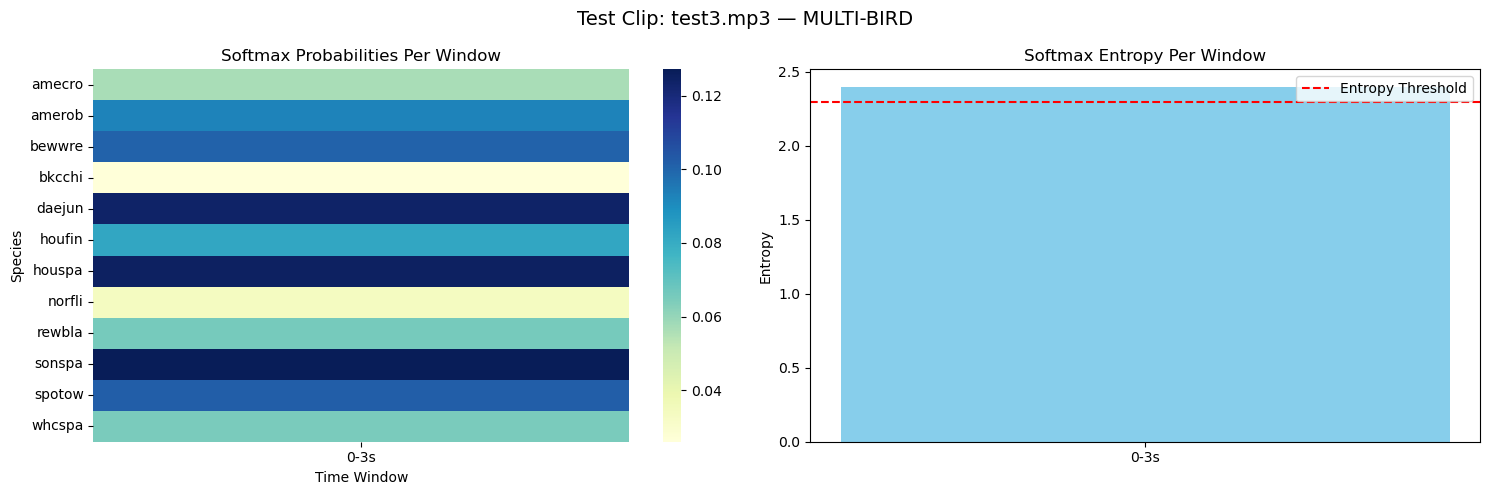
* Clip1 : Predicted species - sonspa; flagged due to high entropy.
* Clip2 : Predicted species - sonspa ; entropy exceeded threshold in one window.
* Clip3 : Predicted species - sonspa with similar entropy profile.

Corresponding plot showed perwindow softmax probabilities and entropy bars for each clip. These visualisations allowed us to interpret prediction consistency and model confidence overtime.



*Figure : 18 TestClip-1 softmax window and entropy plots*

*Figure : 19 TestClip-2 softmax window and entropy plots*

**

*Figure : 18 TestClip-3 softmax window and entropy plots*

### **Discussion**

The results from the multi-class classification task reveal several important insights into the challenges and strengths of using deep learning for bioacoustic signal recognition . Early CNN architectures struggled with either underfitting or unstable training behaviour, with validation accuracy plateauing well below practical utility. CNN version 4 demonstrated notable improvement by incorporating regularization and fine tuned learning rate schedules, reaching over 37% validation accuracy. Its confusion matrix displayed increased diagonal concentration , indicating better class discrimination.

The application of MobilNetV2 through transfer learning proved to be the most effective approach. Despite being pre trained on non-audio imagery , MobilNetv2 captured abstract frequency - time structures present in spectrograms. The model achieved the highest validation accuracy of ~40.5% and lowest loss, with a stable training trajectory. This confirms that leveraging pretrained convolutional filters- especially when data is limited- can meaningfully improve performance without extensive tunging or computational cost.

Real-word test clip predictions further contextualized the model’s strengths and limitations. The entropy based multi-bird detection revealed that although the model confidently predicted dominant species in each clip, softmax entropy often exceeded the threshold, suggesting potential overlaps or ambiguity. This underlines a key limitation of the single label classification framework in capturing the multi-label nature of natural soundscapes.

### **Conclusion**

This project demonstrates the feasibility of using convolutional neural networks and transfer learning to classifying bird species based on short audio recordings. The transition from simple CNNs to pretrained MobilNetV2 yielded measurable gains in accuracy and generalization. Preprocessing spectrograms into a consistent format and applying entropy based interpretation strategies enabled more robust model evaluation and test deployment.

However, challenges remain. Species with overlapping vocal ranges or imbalance sample representation led to misclassification. Furthermore, the inability to label multiple species in a single clip limits the expressiveness of the current framework. Further directions may include Multi\_lable classification architectures, attention - based mechanisms to localize calls, and expanded datasets with richer annotations.

### **References**

1. [James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, No. 1). New York: springer.](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=James%2C+G.%2C+Witten%2C+D.%2C+Hastie%2C+T.%2C+%26+Tibshirani%2C+R.+%282021%29.+An+Introduction+to+Statistical+Learning+with+Applications+in+Python+%282nd+ed.%29.+Springer.&btnG=)
2. [Howard, A.G., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=Howard%2C+A.+G.%2C+et+al.+%282017%29.+MobileNets%3A+Efficient+Convolutional+Neural+Networks+for+Mobile+Vision+Applications.+arXiv+preprint+arXiv%3A1704.04861.&btnG=#d=gs_cit&t=1747255124019&u=%2Fscholar%3Fq%3Dinfo%3AuC0wVLu36k8J%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den)
3. [He, K., Zhang, X., Ren, S. and 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://scholar.google.com/scholar?hl=en&as_sdt=0%2C48&q=He%2C+K.%2C+Zhang%2C+X.%2C+Ren%2C+S.%2C+%26+Sun%2C+J.+%282016%29.+Deep+Residual+Learning+for+Image+Recognition.+Proceedings+of+the+IEEE+Conference+on+CVPR.&btnG=#d=gs_cit&t=1747255173512&u=%2Fscholar%3Fq%3Dinfo%3ALrPNPdmMzoAJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den)
4. TensorFlow Documentation.<https://www.tensorflow.org>
5. Librosa Audio Processing Library.<https://librosa.org>
6. OpenAI, *ChatGPT (May 2025 version)*. [Online]. Available: <https://chat.openai.com/> - For image generation, troubleshooting errors and code refinement.