A Combined ViT-FNN Model for Bird Species Classification from Audio Recordings

Bardia Parmoun
Systems and Computer Engineering
Carleton University
Ottawa, Canada
bardiaparmoun@cmail.carleton.ca

Huda Sheikh
Systems and Computer Engineering
Carleton University
Ottawa, Canada
hudasheikh@cmail.carleton.ca

Nadia Ahmed
Systems and Computer Engineering
Carleton University
Ottawa, Canada
nadianahmed@cmail.carleton.ca

Prianna Rahman
Systems and Computer Engineering
Carleton University
Ottawa, Canada
priannarahman@cmail.carleton.ca

I. Introduction

This document is a detailed summary of the final project implementation for the BIOM/SYSC 5405 course, Pattern Classification and Experiment Design. The goal of this project was to recognize the chirps of 10 different bird species from 1-minute recordings. The following birds were considered:

- AMRO - American Robin
- BHCO - Brown-headed Cowbird
- CHSW - Chimney Swift
- EUST - European Starling
- GRCA - Gray Catbird
- HOSP - House Sparrow
- HOWR - House Wren
- NOCA - Northern Cardinal
- RBGU - Ring-billed Gull
- RWBL - Red-winged Blackbird

The data included recordings from 11 different locations and were gathered in such a way that each recording included zero to multiple different bird species. The team had access to data from three different years, 2021 to 2023 (each year having around 3000 recordings). For this project, the data from the years 2021 and 2022 were used for training, and the data from 2023 were held back as blind test data to measure the model's performance.

II. Methodology

A. Approach

For the final project, the team was assigned to identify the specific birds heard in multiple audio recordings, given data from various locations and years. This was a multi-label problem, as each recording could have multiple bird sounds. In addi-

tion, both the audio recordings and CSV files with the corresponding spectrograms were given, making the problem multi-modal as well. Ultimately, the team decided to use meta-learning and ensemble learning techniques to combine predictions from both a Feed-forward Neural Network (FNN) and Vision Transformer (ViT). Here is a quick overview of the approach that was used:

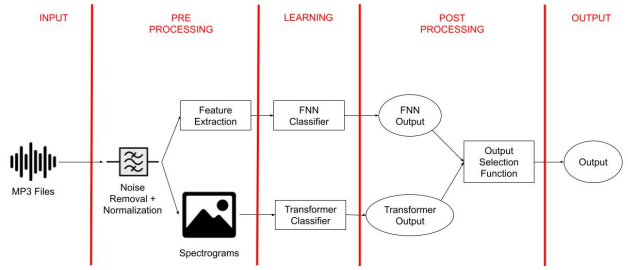


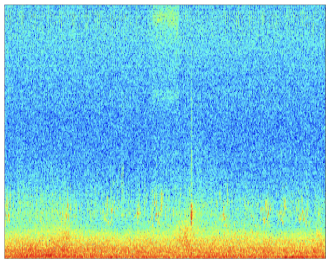
Fig. 1: A diagram of the team's approach.

First, the given MP3 files were processed by applying noise reduction and normalization. Then, proper audio features such as mel-frequency cepstral coefficients (MFCCs), chroma, spectral contrast, and location were extracted for the FNN model. In addition, the processed audio was turned into a spectrogram for the ViT model. The outputs of each model were then combined through hard voting to come up with the final prediction. Each aspect of this method will be discussed further in the following sections.

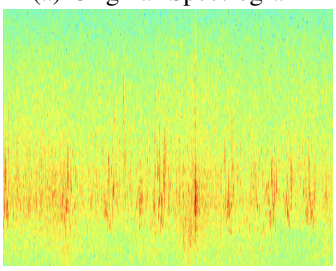
B. Data Pre-Processing

At the beginning of the project, two sources of data were given: MP3 and the corresponding CSV files containing spectrogram data of the recordings. When listening to the audio recordings, it was noticed that bird sounds were largely drowned out by the noise (traffic noise, other environmental noise, etc.). To hopefully reduce the effects of the noise, the audio was pre-processed by performing band-pass filtering. Bird sounds are known to range from roughly 1000 Hz to 8000 Hz [1]. Traffic noise is usually in the range of 500 to 2500 Hz [2]. To be slightly aggressive, all audio recordings were band-passed from 3000 Hz to 8000 Hz. To perform this noise reduction, a Butterworth band-pass filter with an order of 3 in `MATLAB` was used. After processing the audio, the bird sounds were much more noticeable, and the noise was significantly reduced. The team then proposed to amplify the bird sounds further by applying a gain to each of the audio recordings. Although this did work in intensifying the bird sounds, it did cause some undesirable effects such as increased processing time and clipping. Ultimately, it was decided that reducing the noise was sufficient enough. To avoid the effects of clipping from filtering the audio, the recordings were normalized so the amplitude did not exceed the range [-1 1]. Since pre-processing was performed on the audio, the spectrograms were created directly from the recordings, rather than from the provided CSV files. An example of the effects of this pre-processing can be seen in Figure 2. In Fig. 2 (a),

there is a noticeable amount of irrelevant data (blue areas of the spectrogram), and the bird sounds are mainly located at lower frequencies (red areas of the spectrogram). Fig. 2 (b) shows a cleaned version of the spectrogram, after bandpass filtering and data normalization. In this spectrogram, the bird chirp pattern is much more clear (shown in red).



(a) Original Spectrogram



(b) Cleaned Spectrogram

Fig. 2: (a) The original spectrogram and (b) The original spectrogram filtered through a band-pass filter.

As mentioned previously, the team's proposed method was to combine both an FNN model and ViT model. The audio recordings were fed to the FNN model, and the spectrograms were fed to the ViT model. Using both sources of data ultimately led to more accurate predictions.

For feature extraction, the most relevant features from the audio were extracted for classification using the `librosa` library. As noted earlier, these include the MFCCs, chroma, and spectral contrast. MFCCs show the power spectrum of a signal by taking the audio and applying a series of Fourier transforms and transformations to find the proper coefficients. These help define frequency patterns and harmonics. Chroma is another useful feature. For this feature, the audio spectrum is split into 12 distinct bins, which represent the 12 distinct semitones across all octaves [3]. The chroma can help differentiate between different pitches. The spectral contrast is the difference between peak and valley energies within sub-bands of an audio signal [4]. A high spectral contrast indicates a sound closer to a pure tone, and a low spectral contrast is closer to noise. In addition, the location of each recording was added as an additional feature to the feature set by encoding each location into a number from 1 to 11. This ensured that the model was aware of the bird species that were not present in certain locations.

C. Feature Importance

The chart in Figure 3 illustrates Feature Importance Based on Mutual Information, which measures how much each feature contributes to predicting bird species. Higher mutual information scores indicate stronger relevance to the classification task.

The top three features identified are:

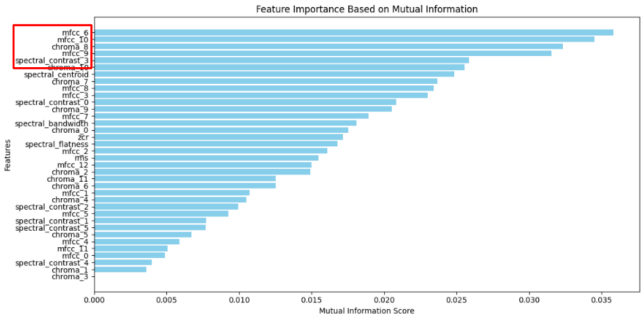


Fig. 3: Feature Importance Based on Mutual Information. Higher scores indicate stronger relevance of features for predicting bird species.

- **MFCC_6**: Captures spectral properties, highlighting tonal and frequency details.
- **MFCC_10**: Reinforces the importance of spectral features in vocal classification.
- **Chroma_8**: Represents pitch class distribution, capturing tonal and harmonic structures.

These features can be categorized into two main types:

- **MFCCs**: Focus on frequency and spectral details to analyze tonal qualities. MFCCs, particularly MFCC_6 and MFCC_10, emphasize the fine-grained spectral properties that are essential for identifying unique vocal patterns in bird calls.
- **Chroma Features**: Highlight pitch patterns and melodic structures. Chroma_8, in particular, emphasizes harmonic and tonal content, making it useful for identifying repetitive tonal patterns in bird vocalizations.

Together, MFCCs and chroma features complement each other. MFCCs provide a detailed spectral analysis, while chroma features emphasize tonal and harmonic patterns. This combination is crucial for distinguishing bird vocalizations, as spectral properties help separate tonal bird calls from environmental noise, and harmonic structures enhance the detection of melodic patterns unique to bird species.

D. Meta-Learning

Meta-learning is a machine-learning approach that helps models perform well on new tasks by learning from patterns across different datasets or models [5]. It focuses on combining strengths from multiple models to improve overall accuracy and adaptability, especially when data is limited [5]. In this project, meta-learning was used by combining the predictions of two models: an FNN trained on audio features and a ViT trained on spectrograms. These models focused on different types of data, and their predictions were combined using a hard voting method, which averaged their outputs to make final decisions. To ensure balanced and consistent predictions, the team optimized the decision threshold. This approach allowed the combined model to take advantage of each individual model's strengths, resulting in better overall performance and reliability in identifying bird species from audio recordings and spectrograms.

E. Implementation

To implement this project, the group utilized `Python` and `Google Colab`. Additionally, the team used popular `Python` libraries such as `PyTorch` and `TensorFlow` to implement the ViT and FNN models. Furthermore, as previously mentioned, the team implemented various `MATLAB` scripts for the pre-processing phase. A `GitHub` repository containing the complete implementation of the project is located at:

https://github.com/bardia-p/Bird-Classifier

III. Experimentation

A. Training and Validation

In the beginning stages of the project, the team decided to work with ViT and long short-term memory (LSTM) models. However, after some experimentation with sequential and non-sequential data, the team moved to using FNN instead of LSTM because of its better performance. This is due to the fact that each audio recording has one label, and not a label for each timestamp within the recording.

As previously mentioned, raw audio recordings were pre-processed to extract numerical feature vectors. For the FNN, these extracted features were combined with a location-specific feature (representing the geographical context) to form the final input vector. For the ViT, spectrograms were generated from the audio data, divided into fixed-size patches, and embedded into a 768-dimensional space for input. The training data consisted of labeled audio recordings from multiple locations. Each audio file was processed to generate the appropriate input format for the respective model. Each of the models was trained for a fixed number of epochs, with early stopping mechanisms based on validation performance to avoid overfitting. The team chose to perform training using stochastic gradient descent (`SGD`) with momentum of 0.9, rather than adaptive moment estimation (`ADAM`).

A separate validation set was created, ensuring no overlap with the training data. Validation performance was monitored using metrics like accuracy, precision, recall, and F1-score. Validation loss trends were used to adjust hyperparameters, including learning rates, dropout rates, and regularization strength.

The team also optimized and refined the hyperparameters using iterative experiments. The learning rate, dropout rate, `L2` regularization strength, and transformer-specific settings (e.g., patch size, embedding dimension) were tuned based on validation performance. `ReduceLROnPlateau` was also used to dynamically adjust the learning rate during training, ensuring the models could fine-tune effectively.

Figure 4 shows the final metrics collected during training. Further steps can be taken to improve these

metrics, such as regularization for the ViT.

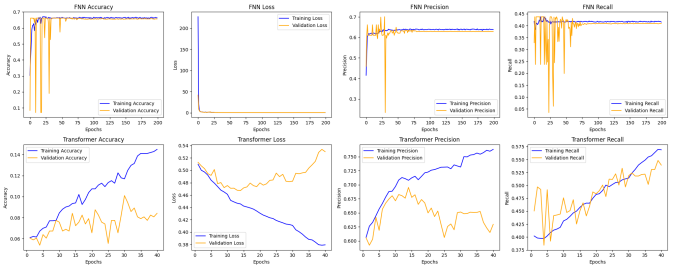


Fig. 4: A summary of the metrics of the transformer and the FNN during training.

B. *Estimation of Model Performance*

After predictions from both models were averaged to create combined probabilities, various thresholds (ranging from 0 to 1) were tested to convert these probabilities into binary predictions, and the F1-score was calculated for each threshold. The F1-score (micro average) is a performance metric that calculates the harmonic mean of precision and recall across all classes, treating every sample equally regardless of its class labels. It uses the true positives, false positives, and false negatives across all classes and then computes the precision and recall to calculate the F1-score.

This metric is useful for addressing class imbalance because it ensures that each sample contributes equally to the overall score, regardless of the class distribution. Unlike the macro average, which gives equal weight to each class, the micro average prevents smaller classes from being dominated by larger classes. Therefore, the micro-averaged F1-score provides a balanced assessment of how well the model handles both majority and minority classes in imbalanced datasets.

The best threshold for maximizing the F1-score was recorded for that fold. After all folds, the mean and standard deviation of the F1-scores was calculated in order to reflect the model's average performance and variability across different splits. This cross-validation approach ensured a thorough evaluation of the model's performance. Additionally, the mean and standard deviation of the selected thresholds were computed to assess consistency. The final selected threshold was 0.148 ± 0.01. While the final threshold value reflects data imbalance, its low variability indicates consistent threshold selection across folds.

In the end, the estimate of the meta-model's performance based on the five-fold cross validation process was 59.2 ± 0.1.

IV. Testing

A. *Testing Process*

As previously mentioned, the team used data provided from 2023 to verify the model. The team utilized the models that were trained during the experimentation phase (200 epochs for the FNN and 40 epochs for the ViT conducted on 80% of the data collected in 2021 and 2022). To prepare the test data, the same pre-processing scripts were run on it. After that, the cleaned audio files and

spectrograms were fed through the same feature extraction process. It is worth mentioning that the team made sure to apply the same encoding logic, that was done during the training process, to the location labels.

After running both models on the test data, the team combined their results by applying the same meta-learning technique of hard voting. Finally, to obtain the final binary predictions, the mean value of the best threshold that was obtained during cross-validation (0.15) to the final weights was applied.

B. Results

After submitting the test results, the team received an F1-score of 59.16 which was very similar to the estimate that was submitted (59.2 ± 0.1). This resulted in a very high precision of 39.862 for the team. These results indicate that the model's performance was as expected and quite consistent when faced with new data.

V. Discussion and Reflections

A. Strengths

As mentioned in the testing section, a big strength of the developed model is consistency and satisfactory performance when faced with blind test data. The team believes that various factors contributed to this:

First, the team utilized a lot of data to train the models. Through the precise feature selection process, both 2021 and 2022 datasets were utilized which greatly reduced the meta-model's risk of overfitting. In addition, it should be noted that the team avoided using too many features since those would also increase the risk of overfitting.

Additionally, as mentioned in the experimentation section, the team tried different threshold values to account for the inconsistent behaviours of the models. In other words, each model was trained separately with a threshold of 0.5 to ensure they learned as much as possible about the data; however, after combining the two, the team decided to reduce the overall threshold to around 0.15 since it was found that the model was underfitting.

Furthermore, the team decided to utilize meta-learning which allowed for the usage of both audio files and the spectrograms. Training two separate models and combining them also reduced the variance of the overall model, since the FNN and ViT behaved quite differently and were given very different inputs. The overall bias of the model is also very low, as both the FNN and ViT are deep learning models that do not make a lot of assumptions about the shape of the distribution. An observation that the team had which can clearly be seen in Figure 4 is that the FNN model was particularly good at precision (shown by its validation precision) and similarly the ViT model was good at recall (shown by its validation recall metric). This meant

by averaging the two models together the team was able to get consistent and satisfactory precision and recall values which together resulted in a consistent F1-score.

To reduce the risk of overfitting even further, the team performed five-fold cross-validation on the data. In each fold, both classifiers were given the same train, validation, and test data to obtain the best F1 score and threshold that can be calculated by both models. The team repeatedly adjusted the various hyper-parameters until these values were sufficiently consistent.

Finally, the team made the conscious decision to tackle the class imbalance problem through providing the additional location feature as opposed to utilizing popular methods such as class weights or under-sampling. This is due to the fact that some bird species might not be available in certain locations meaning using a class weights or under-sampling on all the data will incorrectly punish the model for learning useful information about the non-minority classes. The team also noticed that the distribution of data changed greatly between 2021 and 2022, meaning there is a good chance that the data from 2023 did not have a similar distribution to them. This is another reason why the use of class weights might lead to overfitting. On the other hand, the use of location as a feature ensures that the model only learns about the overpopulated classes when applicable and would be another factor to distinguish the minority classes from each other.

B. Challenges

The team noticed that the training accuracies of both the FNN and the ViT were always around 70% no matter the number of epochs or the hyperparameters used. This indicates a lack of data and shows that there is still room for the model to grow given more data or better features. This is especially true since there are $2^{10} = 1024$ possible label combinations for the data, but the team was only provided with around 3000 samples per year which included a lot of similar label combinations. If the data had more examples and more unique labels, the mvariation would help the model tp fprm better pattern associations. In the future, the team would like to experiment using more features or using external feature extraction libraries such as the `Wavelet` library [6].

The team faced difficulties with the quality of the data. Unfortunately, the original recordings included a lot of noise. In addition, the recordings seemed to include other background sounds in them. Some of these challenges were mitigated with cleanup scripts; however, there is still room to improve the quality of these audio files.

Additionally, the data could have been formatted better. Audio recordings are ideal for LSTMs and other models that support sequential data; however,

the data only had one label per recording indicating that there is a bird chirp somewhere in the recording. This means that the data is essentially no longer sequential, since we need to consider the recording as a whole. It would have been beneficial if specific timestamps were given for each recording indicating where the different bird chirps occur. This would make the data much better for models such as the LSTM.

Finally, the team was provided with limited time and resources to train these models. This factor became especially prevalent when it came to training the ViT. Transformers usually need a lot of training time, even on T4 GPUs. As such, the team did not have sufficient time or resources to be able to properly tune this model and regularize its output. If given another opportunity, the team would like to experiment with a simpler model such as a convolutonal neural network (CNN) to train the spectrogram data.

C. Conclusions

In conclusion, the team believes that the approach discussed in this report can serve as a starting point for training a model to solve this problem. In this project, it has been shown that it is possible to develop a meta-learning model that behaves consistently when faced with new data. In addition, the team's approach showed how the use of meta-learning and proper data pre-processing can improve the performance of a model. Finally, it was shown that using both spectrograms and audio files will be greatly beneficial in improving and stabilizing the performance of machine learning models. This suggests that the next iterations of the project should continue looking into methods that can utilize both input formats. An example of that could be using an LSTM or RNN model on the audio files and using a CNN on the spectrograms and combining the results.

VI. Contributions

A. Report Contributions

Table I outlines the contributions made toward the preparation and writing of this report:

TABLE I: Report Contributions

Section	**Contributors**
Introduction	Bardia
Methodology	Prianna, Huda
Experimentation	Nadia, Huda, Prianna
Testing	Bardia
Discussion and Reflections	Bardia, Nadia
Contributions	Nadia
Figures and Visualizations	Bardia, Huda, Nadia, Prianna
References and Citation Management	Bardia, Huda, Nadia, Prianna
Proofreading and Editing	Bardia, Huda, Nadia, Prianna

B. Project Contributions

Table II outlines the contributions made to the execution of the project:

TABLE II: Project Contributions

Task	Contributors
Data Preparation and Preprocessing	Prianna, Huda, Nadia
Feature Exploration	Nadia, Huda
Feature Extraction	Huda, Prianna
FNN Training and Validation	Bardia, Huda
ViT Training and Validation	Nadia, Huda, Bardia
Blind Test Evaluation	Bardia

ACKNOWLEDGMENT

This project was completed to fulfill the project requirements for BIOM/SYSC 5405, Pattern Classification and Experiment Design. This project could not have been complete without support from Professor James Green, and the dataset provided by Chris Dennison and Dr. Rachel Buxton from the Department of Biology at Carleton University.

REFERENCES

[1] A. A. Birds, "Do bird songs have frequencies higher than humans can hear?." Available: https://www.allaboutbirds.org/news/do-bird-songs-have-frequencies-higher-than-humans-can-hear/. Accessed: 2024-12-13.

[2] H. Wang, H. Gao, and M. Cai, "Simulation of traffic noise both indoors and outdoors based on an integrated geometric acoustics method," *International Journal of Environmental Research and Public Health*, vol. 16, no. 14, p. 2491, 2019. Accessed: 2024-12-13.

[3] C. University, "Chroma feature analysis and synthesis." Available: https://www.ee.columbia.edu/~dpwe/LabROSA/matlab/chroma-ansyn/. Accessed: 2024-12-13.

[4] librosa, "librosa.feature.spectral_contrast." Available: https://librosa.org/doc/main/generated/librosa.feature.spectral_contrast.html. Accessed: 2024-12-13.

[5] A. Al Masud, S. Hossain, M. Rifa, F. Akter, A. Zaman, and D. M. Farid, "Meta-learning in supervised machine learning," in *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pp. 222–227, 2022.

[6] pywavelets, "Wavelet transforms in python." Available: https://pywavelets.readthedocs.io/en/latest/index.html. Accessed: 2024-12-13.