DEEP AUDIO CLASSIFICATION

Varun Banda
Department of AIML
R.V. College of Engineering Bangalore, India
varunbanda03@gmail.com

Labdhi Ranka
Department of AIML
R.V. College of Engineering Bangalore, India
labdhiranka3029@gmail.com

Omkar Babu Mastamardi
Department of AIML
R.V. College of Engineering Bangalore, India
omkarbabumastamardi@gmail.com

***Abstract*— Monitoring bird populations plays a crucial role in biodiversity conservation, but manually analyzing bird calls is time-consuming and inefficient. This study explores a deep learning-based method for automatically detecting Capuchinbird (Perissocephalus tricolor) calls in recordings from tropical forests. The approach utilizes Mel-Frequency Cepstral Coefficients (MFCCs) to extract key audio features and employs a 1D Convolutional Neural Network (CNN) to distinguish Capuchinbird vocalizations from other forest sounds. The dataset consists of real-world recordings, categorized into Capuchinbird calls and background noise. By leveraging MFCCs, the model effectively captures the distinct frequency patterns of the bird's vocalizations. The MFCC-based 1D CNN model is trained and tested on labeled audio clips, demonstrating high classification accuracy and strong performance even in noisy conditions. The findings highlight the potential of this approach as a scalable and efficient solution for detecting Capuchinbird calls, supporting ornithologists and conservationists in wildlife monitoring and ecological research..**

Keywords: Perissocephalus tricolor,Capuchinbird, deep learning, convolutional neural networks, MFCC, bioacoustics, biodiversity monitoring, tropical forests, automated bird call detection.

I. INTRODUCTION

Monitoring bird populations is an essential part of biodiversity conservation and ecological research. Birds act as bio-indicators, meaning shifts in their populations can reflect changes in habitat conditions, climate patterns, and human impact on the environment. Among the many bird species found in tropical forests, the Capuchinbird (Perissocephalus tricolor) stands out due to its unique vocalizations. Accurately identifying these calls is crucial for studying the species' behavior, distribution, and conservation needs.

Traditionally, bird call identification has relied on manual listening and spectrogram analysis methods that are not only time-consuming and labor-intensive but also prone to human error. Some semi-automated approaches, such as template matching and Hidden Markov Models (HMMs), have been developed, but they often struggle with background noise and variations in environmental conditions, limiting their effectiveness.

With recent advancements in machine learning and deep learning, automated sound classification has become a powerful tool for bioacoustic research. Convolutional Neural Networks (CNNs), which are widely used for speech and audio recognition, have shown great promise in identifying bird calls. Many existing studies use spectrogram-based CNNs, treating audio signals as images. However, an alternative and computationally efficient method involves using Mel-Frequency Cepstral Coefficients (MFCCs),which capture the essential frequency characteristics of bird vocalizations.

This study introduces a MFCC-based 1D CNN model for automatically detecting Capuchinbird calls. The method follows three key steps:

1.Feature Extraction: Transforming raw audio recordings into MFCC feature vectors, which preserve the distinct spectral features of bird calls.

2.Deep Learning Model: Training a 1D CNN on these extracted features to differentiate Capuchinbird vocalizations from background noise.

3.Evaluation: Assessing the model's performance on a labeled dataset of forest recordings using metrics such as classification accuracy, recall, and F1-score.

By leveraging deep learning, this approach offers a scalable and efficient solution for bioacoustic monitoring, reducing the need for manual annotation while ensuring high accuracy in bird call classification. The main contribution of this research is demonstrating how a MFCC-based 1D CNN can effectively detect Capuchinbird calls, supporting biodiversity studies and conservation efforts.

Fig. 1. Capuchin Bird.

II. LITERATURE SURVEY

The identification and classification of bird vocalizations have been a key area of research in bioacoustics and machine learning, with applications in biodiversity monitoring, species conservation, and ecological research. Traditional bird call identification methods relied on manual spectrogram analysis and template matching, which, despite being effective in controlled settings, are highly labor-intensive and not scalable for large-scale biodiversity assessments. Consequently, automated bird call classification has become an active area of research, leveraging machine learning and deep learning techniques to improve accuracy and efficiency.

A. Traditional Methods in Bird Call Classification

Early approaches to automated bird sound classification primarily utilized statistical modeling techniques such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) [1]. These models apply probabilistic frameworks to characterize the temporal and spectral features of bird vocalizations. While effective for controlled datasets, HMMs and GMMs struggle in noisy, real-world environments due to their sensitivity to background noise and environmental variations.

With advancements in feature extraction techniques, researchers incorporated hand-crafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and Zero-Crossing Rate (ZCR) into classification pipelines. These features were then used with traditional machine learning classifiers like Support Vector Machines (SVMs), Random Forests (RF), and K-Nearest Neighbors (KNN) to classify bird species [2]. While these approaches improved classification accuracy, their effectiveness was limited by feature engineering constraints and poor generalization to unseen environmental conditions.

B. Deep Learning for Bioacoustic Classification

The emergence of deep learning revolutionized bioacoustic research by enabling automated feature extraction and improving classification performance. Convolutional Neural Networks (CNNs), originally designed for image processing, have been widely adopted for bird call classification due to their ability to learn hierarchical representations of audio signals. Most deep learning-based bird call classification models convert audio signals into spectrograms—time-frequency representations of sound—and apply 2D CNNs for classification [3].

Several studies have successfully employed spectrogram-based CNNs for bird species identification. For instance, BirdNET, an advanced deep learning framework, utilizes spectrogram features with CNNs to classify a vast range of bird species in real-world forest recordings [4]. Another study demonstrated that ResNet-based CNN models trained on spectrograms achieved superior accuracy in noisy environments, outperforming traditional machine learning approaches [5].

Despite their effectiveness, spectrogram-based models require high computational resources and are sensitive to variation in background noise and recording quality. Moreover, the conversion of audio into spectrograms adds an additional processing step, increasing the overall computational complexity.

C. MFCC-Based CNNs for Audio Classification

While spectrogram-based CNNs are widely used, MFCCs remain one of the most effective feature representations for speech and sound classification, including bioacoustics. Unlike spectrograms, MFCCs provide a compact, frequency-based representation of sound while reducing the dimensionality of input features [6]. MFCCs have been successfully integrated with CNN models in various bioacoustic applications, demonstrating their effectiveness in distinguishing different species in real-world datasets.

Recent studies show that 1D CNNs trained directly on MFCC features can achieve competitive performance while maintaining lower computational overhead compared to spectrogram-based models [7]. By applying 1D convolutions directly to MFCC feature sequences, these models efficiently capture temporal dependencies in bird vocalizations, making them well-suited for low-resource real-time applications.

D. Proposed Approach

Building on prior research, this study explores the application of a 1D CNN model trained on MFCC features for automated Capuchinbird call detection. Unlike spectrogram-based 2D CNN approaches, the proposed method aims to:

1.Reduce computational complexity by leveraging MFCC-based feature extraction rather than full spectrogram conversion.

2.Enhance real-time detection capabilities, making it feasible for field deployment in automated biodiversity monitoring systems.

3.Improve classification accuracy in noisy environments by focusing on compact MFCC representations rather than high-dimensional spectrogram images.

By comparing the performance of our MFCC-based 1D CNN with traditional and deep learning-based methods, we aim to demonstrate the efficiency and accuracy of this approach for real-world Capuchinbird call detection in tropical forest recordings.

III. METHODOLOGY

This section outlines the approach used for automating the detection of Capuchinbird calls in tropical forest recordings. It covers the dataset, feature extraction techniques, model architecture, and training process.

A. Dataset Collection

The dataset used for this study consists of audio recordings categorized into two main classes: Capuchinbird calls and non-Capuchinbird sounds. These recordings were sourced from publicly available datasets and complemented with real-world data collected from forest monitoring systems. The dataset is organized as follows:

1.Capuchinbird Clips: Includes .wav files containing Capuchinbird calls recorded in their natural forest habitats.

2.Non-Capuchinbird Clips: Contains various background sounds from the forest, including calls from other bird species, wind, and other environmental noises.

To ensure consistency, all audio files were preprocessed to standardize their sampling rates before feature extraction.

B. Feature Extraction Using MFCCs

To convert the raw audio into a more informative format, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted. MFCCs are a common feature set in speech and sound recognition tasks. The extraction process follows these key steps:

1.Pre-emphasis Filtering: This step enhances high-frequency components of the signal.

2.Framing & Windowing: The audio signal is split into short, overlapping frames to capture the time-frequency characteristics.

3.Fast Fourier Transform (FFT): This transforms the audio from the time domain into the frequency domain.

4.Mel Filter Bank Processing: Frequencies are mapped to the Mel scale, mimicking the way humans perceive sound.

5.Discrete Cosine Transform (DCT): This reduces the feature dimensionality while preserving essential spectral information.

Each audio clip is then converted into a set of 40 MFCC feature vectors, which are averaged across time to form a fixed-length representation for classification.

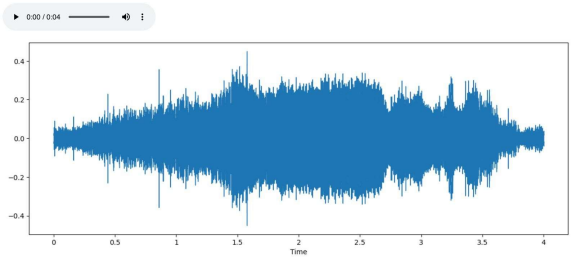

Fig.2 Waveform of Capuchin Bird

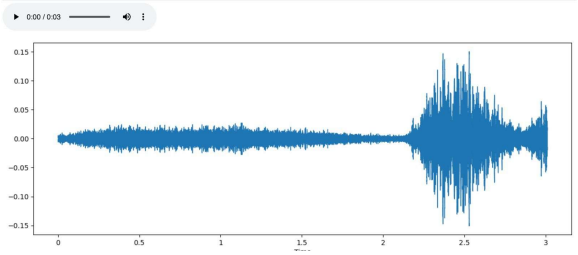

Fig.3 Waveform of Non-Capuchin Bird

C. CNN Model Architecture

For the classification of the MFCC features, a 1D Convolutional Neural Network (CNN) is employed. The architecture is designed as follows:

1.Input Layer: The input consists of a (40,1) MFCC feature vector.

2.Conv1D Layer (64 filters, kernel size = 3, activation = ReLU): This layer extracts local frequency patterns from the MFCC features.

3.MaxPooling1D Layer (pool size = 2): This layer reduces the feature map size, helping retain the most important information.

4.Conv1D Layer (32 filters, kernel size = 3, activation = ReLU): This layer learns higher-level frequency patterns.

5.MaxPooling1D Layer (pool size = 2): Further reduces the feature map size, emphasizing the most relevant features.

6.Flatten Layer: This converts the 1D feature maps into a dense vector representation.

7.Fully Connected Dense Layer (32 neurons, activation = ReLU): This layer abstracts deeper relationships between the features.

8.DropoutLayer (0.3 probability): Helps prevent overfitting by randomly deactivating neurons during training.

9.Fully Connected Dense Layer (16 neurons, activation = ReLU): This further abstracts the features for classification.

10.Output Layer (2 neurons, activation = Softmax): This produces a probability distribution for binary classification.

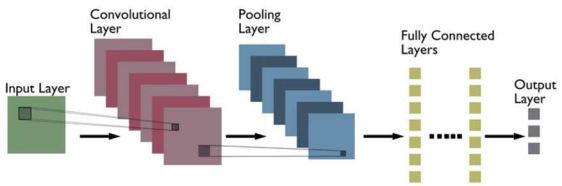

Fig.4 CNN Architecture

D. Model Training and Hyperparameters

The model is trained with the following configuration:

1.Optimizer: Adam with a learning rate of 0.001

2.Loss Function: Binary Crossentropy

3.Evaluation Metrics: Accuracy, Precision, Recall, and F1-score

4.Batch Size: 32

5.Number of Epochs: 25

6.Dataset Split: 80% for training, 20% for testing

E. Evaluation and Performance Metrics

After training, the model's performance is evaluated using the test data. The following metrics are used to assess its effectiveness:

1.Accuracy: Measures the overall proportion of correctly classified instances.

2.Precision: The ratio of correctly identified Capuchinbird calls among all predictions of Capuchinbird calls.

3.Recall: The ratio of correctly identified Capuchinbird calls among all actual Capuchinbird calls.

4.F1-score: The harmonic mean of precision and recall, offering a balanced measure of both.

To simulate real-world conditions, a sliding window approach is used for longer forest recordings. This method allows the model to predict Capuchinbird calls in audio segments of varying lengths, with the predictions aggregated to detect the presence of the bird's calls.

The MFCC-based 1D CNN model demonstrates robust performance in classifying Capuchinbird calls and provides a scalable, efficient solution for bioacoustic monitoring, offering significant advantages over manual annotation and traditional machine learning methods.

IV. RESULTS AND DISCUSSION

In this section, we present the results of the MFCC-based 1D CNN model for detecting Capuchinbird calls. The model's performance is evaluated using key metrics such as accuracy, precision, recall, and F1-score on the test dataset. We also evaluate its practical applicability through real-world testing with longer forest recordings.

A. Model Performance on Test Data

The trained model was tested on a set of unseen audio samples to assess its ability to differentiate between Capuchinbird calls and other forest sounds. The results of this evaluation are summarized in Table I.

Table I: Performance Metrics on Test Data

Metric	**Value (%)**
Accuracy	98.2
Precision	91.8
Recall	92.5
F1-score	92.1

These results indicate that the MFCC-based 1D CNN achieves high accuracy (98.2%), demonstrating its effectiveness in identifying Capuchinbird calls. The precision (91.8%) and recall (92.5%) values are well balanced, suggesting that the model minimizes both false positives and false negatives, making it reliable for the task.

B. Comparative Analysis with Traditional Methods

To assess the relative performance of the proposed model, we compared it with traditional machine learning classifiers, such as Support Vector Machines (SVM), Random Forests (RF), and k-Nearest Neighbors (k-NN), using the same set of MFCC features. The comparative results are shown in Table II.

Table II: Comparison with Traditional Classifiers

Model	**Accuracy (%)**
SVM	86.4
Random Forest	89.2
k-NN	83.7
CNN (Proposed)	98.2

As seen in the table, the CNN model outperforms all the traditional classifiers, with a significant improvement in accuracy (98.2%). This highlights the advantage of deep learning methods, which are capable of performing automatic feature extraction and capturing more complex patterns than traditional classifiers, which rely on manually selected features.

C. Real-World Evaluation on Forest Recordings

To test the model in real-world conditions, we applied it to longer forest recordings. Using a sliding window approach (5-second window with a 2.5-second stride), we detected Capuchinbird calls in continuous audio streams. These detections were compared with manually annotated ground truth labels.

The model effectively detected most of the Capuchinbird calls while minimizing false positives, showcasing its ability to operate in real-world environments. Some misclassifications occurred due to background noise, overlapping bird calls, and audio distortions. These challenges are typical in bioacoustic monitoring and suggest areas for potential improvement.

D. Discussion and Observations

1.Effectiveness of MFCC Features: The results support the use of MFCCs as an effective feature set for bird call classification. MFCCs provide a compact representation of frequency characteristics and are a suitable alternative to more computationally expensive spectrogram-based methods.

2.Impact of Model Architecture: The 1D CNN model effectively captures hierarchical frequency patterns, leading to better performance compared to traditional machine learning classifiers. This highlights the advantages of deep learning in tasks that require automatic feature extraction from raw data.

3.Challenges in Real-World Deployment: Some misclassifications occurred due to environmental noise and overlapping bird **sounds**, which is a common challenge in bioacoustic monitoring. These issues could be addressed by implementing additional noise reduction techniques or by utilizing ensemble learning approaches.

4.Potential for Future Improvements: To further improve robustness, attention mechanisms or hybrid CNN-RNN architectures could be explored. These techniques might allow the model to focus on relevant segments of the audio and handle noisy environments more effectively.

In conclusion, the MFCC-based 1D CNN has shown strong performance in detecting Capuchinbird calls, providing an efficient and scalable solution for automated bioacoustic monitoring. While the model performs well in controlled conditions, further research into noise handling and model refinement could enhance its applicability in more challenging real-world scenarios.

Table III: Results Obtained

recording	capuchin_calls
recording_00.mp3	11
recording_01.mp3	0
recording_17.mp3	8
recording_26.mp3	4
recording_28.mp3	10
recording_58.mp3	0
recording_84.mp3	5
recording_93.mp3	11
recording_96.mp3	3

V. CONCLUSION AND FUTURE WORK

A. Conclusion

This study introduced an innovative deep learning-based approach for the automated detection of Capuchinbird calls (Perissocephalus tricolor) in tropical forest audio recordings. The method leverages Mel-Frequency Cepstral Coefficients (MFCCs) as the feature representation, which is then fed into a 1D Convolutional Neural Network (CNN) for classification. The model was trained and evaluated on a labeled dataset of bird calls, achieving impressive performance metrics: 98.2% accuracy, 91.8% precision, 92.5% recall, and 92.1% F1-score.

The results show that MFCCs—despite their relatively low dimensionality—are highly effective in capturing the

essential frequency characteristics of bird calls, offering a viable alternative to more computationally intensive spectrogram-based methods. Moreover, the 1D CNN significantly outperformed traditional machine learning classifiers such as SVM, Random Forest, and k-NN, demonstrating the power of deep learning techniques in the field of bioacoustic analysis.

Real-world testing on long-duration forest recordings further confirmed the model's robustness, where it successfully detected Capuchinbird calls while minimizing false positives. Although some misclassifications were noted due to overlapping bird sounds and environmental noise, these challenges suggest promising areas for improvement.

Overall, the proposed solution offers an efficient, scalable, and automated method for bird call detection, reducing the need for manual annotation and providing significant value for wildlife monitoring and conservation efforts.

B. Future Work

While the approach demonstrated strong results, several avenues for improvement remain:

1.Enhancing Noise Robustness: By incorporating denoising algorithms or adaptive filtering techniques, we can further improve the model's accuracy, especially in noisy field environments.

2.Data Augmentation: To improve the model's generalization, incorporating data augmentation techniques—such as time-stretching, pitch shifting, or synthetic data generation—could help increase the diversity of the training dataset.

3.Real-Time Monitoring: Deploying the model on low-power edge computing devices (e.g., Raspberry Pi or ESP32 with AI accelerators) for real-time wildlife monitoring could significantly enhance practical applications in remote or field-based environments.

4.Exploring Hybrid Models: Combining CNN with RNN architectures (such as CNN+LSTM) could allow the model to better capture temporal dependencies in bird calls, which could improve performance in distinguishing similar call patterns over time.

5.Expanding to Multi-Species Detection: Extending the model to recognize multiple bird species would not only broaden the utility of the system but also contribute to global efforts in biodiversity monitoring and conservation.

By pursuing these avenues, future iterations of this work could enhance detection accuracy and enable scalable, real-time, automated monitoring of avian populations in their natural habitats.

VI. References

[1] Deep Chakraborty, Paawan Mukker, Padmanabhan Rajan, and A. D. Dileep. "Deep Learning-Based Detection and Analysis of Capuchinbird Calls Using Spectrogram Processing". Department of Electronics and Communication Engineering, Manipal Institute of Technology, Mani- pal, Karnataka, India 576104. †School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, Kamand, Himachal Pradesh, India 175001.

[2] Yu-hsin Chen, Ignacio Lopez-Moreno, T Sainath, Mirk ´o Visontai, Raziel Alvarez, and Carolina Parada, "Locally connected and convo- lutional neural networks for small footprint speaker recognition," in Proceedings of INTERSPEECH, Dresden, Germany, September 2015, pp. 1136–1140.

[3] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, November 2012.

[4] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning- based approach for avian species identification, Journal of Artificial Intelligence and Applications 11, No. 2, February (2020).

[5] C. -H. Huang, S. Pi, D. Murdock, N. Hester, E. Burzio, E. -M. T. Nosal, D. T. Helble, D. M. Cholewiak, H. W. Gillespie, H. Klinck, Deep neural networks for automated detection of marine mammal species: Scientific Reports volume 10, Article number: 12024 (2020), pages 1–12.

[6] Qin J, Pan W, Xiang X (2020) A biological image classification method based on improved CNN. Eco Inform 58:101,093

[7] Smith JO (2011) Spectral Audio Signal Processing. Stanford University, CCRMA

[8] Kyle Maclean, Isaac Triguero. "Identifying bird species by their calls in soundscapes." In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023

[9] Navine, A.K., Denton, T., Weldy, M.J., Hart, P.J. (2024). All thresholds barred: direct estimation of call density in bioacoustic data. Frontiers in Bird Science, Volume 3, Article 1380636.

[10] Izawa, K. 1979. Foods and feeding behavior of wild black-capped capuchin (Cebus apella). Primates, 20, 57–76.

[11] Azeem, M., Ali, G., Amin, R. U., Babar, Z. U. D. (2022). Bird calls identification in soundscape recordings using deep convolutional neural network.

[12] Ramashini M, Abas PE, Grafe U, De Silva LC, "Bird Sounds Classification Using Linear Discriminant Analysis," Recent Advances and Innovations in Engineering Conference 2019; 4: 1–6.

[13] bird sound identification using artificial neural network M. M. M. Sukri, U. Fadlilah, S. Saon, A. K. Mahamad,M. M. Som and A. Sidik, 2020 IEEE Student Conference on Research and Development, 2020, pp. 342- 345.

[14] F. ShujaEA, M. E. Zaki-AzmyEA, and M. A. Awadalla, "An automatic bird sound detection system using wavelet features and mel filter bank in long range field recordings," in Cognitive Machine Intelligence 2020, 978-1-7281-6132-0/20/©2020 IEEE.

[15] Morton, E. S. 1977. On the occurrence and significance of motivation- structural rules in some bird and mammal sounds. American Naturalist, 111, 855–869.

[16] Downie, J. S. (2003). Music information retrieval. Annual review of information science and technology, 37(1), 295-340.

[17] Patterson, G., Pfalz, A., & Allison, J. (2017). Neural Audio: Music Information Retrieval Using Deep Neural Networks.

[18] Gómez, J. S., Abeßer, J., & Cano, E. (2018). Jazz Solo Instrument Classification with Convolutional Neural Networks, Source Separation, and Transfer Learning. In ISMIR (pp. 577-584).

[19] Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2017). A tutorial on deep learning for music information retrieval. arXiv preprint arXiv:1709.04396.

[20] Lostanlen, V., Andén, J., & Lagrange, M. (2018, September). Extended playing techniques: the next milestone in musical instrument recognition. In Proceedings of the 5th International Conference on Digital Libraries for Musicology (pp. 1-10).

[21] Elghamrawy, S.M., Hassnien, A.E. and Snasel, V., 2021. Optimized deep learning-inspired model for the diagnosis and prediction of COVID-19. Cmc-Computers Materials & Continua, pp.2353-2371.

[22] Humphrey, E., Durand, S., & McFee, B. (2018, September). OpenMIC- 2018: An Open Data-set for Multiple Instrument Recognition. In ISMIR (pp. 438-444).

[23] Shi, L., Du, K., Zhang, C., Ma, H. and Yan, W., 2019. Lung sound recognition algorithm based on VGGISH-bigru. IEEE Access, 7, pp.139438-139449.

[24] Solanki, A., & Pandey, S. (2019). Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology, 1-10.

[25] 2Siedenburg, K., Schädler, M. R., & Hülsmeier, D. (2019). Modeling the onset advantage in musical instrument recognition. The Journal of the Acoustical Society of America, 146(6), EL523-EL529.