Grenze International Journal of Engineering and Technology, June Issue

Comprehensive Identification of Indian Bird Species using Advanced Deep Learning for Acoustic Recognition

Sushopti Gawade[1], Aatmaja Joshi[2], Aditya Kachare[3] and Kadambari Patel[4]
[1-4]Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India
Email: {sushopti.gawade, aatmaja.joshi, aditya.kachare, kadambari.patel}@vit.edu.in

***Abstract*— The project utilizes various machine learning and deep learning algorithms to analyze birds' voices and provide respective predictions. The dataset is collected from the Indian Bird Song website which contains diverse bird's vocalization. After the data collection the key steps include audio processing, feature extraction (e.g. MFCCs and spectrograms) and model training using different algorithms. The application also includes features that help in continuous improvement and bird's species dataset to enhance user interaction. Hence the project aims to contribute towards bird's conservation and reliable tools for bird species identification through a web application.**

***Index Terms*— birds identification, deep learning, acoustic recognition, spectrograms.**

I. INTRODUCTION

Bioacoustics needs sound recognition systems for birds to support research and conservation work as well as birdwatching activities. Despite traditional methods, earlier attempts to classify received poor results due to their time-intensive processes. The development of deep learning specifically Convolutional Neural Networks (CNNs) achieved higher accuracy through spectrogram analysis.
The research findings by Sulupk et al. (2018) proved that MobileNet works best with Jet colormaps to deliver superior performance than past algorithms. RNNs demonstrate promising achievements as well as other neural networks. Through its BirdCLEF competition and existing datasets, the initiative facilitated both innovation development and visibility growth while improving the reliability of the automated bird sound recognition system.

II. LITERATURE REVIEW

Bird call recognition has become difficult and an important factor for research in recent years. Classifications of species by using their calls by using earlier methods was a tedious task. Applications which include environmental monitoring, species conservation and ecological studies has reduced alot of work done when comparing with traditional methods which was time consuming and prone to error. However, with the advent of deep learning this process is now significantly automated with improved accuracy.
CNN has demonstrated good performance in audio classification. The audio is classified by converting the audios which were collected into the spectrograms first and followed by analysing them as images.

Grenze ID: 01.GIJET.11.2.616

During the research Gavalo and Banu[1] showed the effectiveness of combining CNN and LSTM which achieved a good amount of accuracy. Similarly, swaminatahan et al.[2] employed Wav2vec model which captured the pattern effectively and helped in making the process more understandable.
Advancing the concept, Wang et al[3] proposed a hierarchical taxonomy-aware neural network that combined attention mechanism and hierarchical embedding.Ugarate and arias[4] explored feature selection methods and MFCC reinforcing the importance of feature extraction in improving classification accuracy.
Strengthening CNNs, Indumathi et al[5] developed cnn based model where spectrograms were given as input, while Magumba et al[6] contributed to making a diverse dataset. Focusing on practical applications,Amenyedzi et al[7] used Conv1D based pest bird detection system. MacPhail et al. [8] addressed challenges that were related to data compression and MP3 compression with resulted in better kodel performance.
Gupta et al[12] combined CNNs and RNNs to improve temporal pattern reconstruction which achieved 67% accuracy across 100 species.
These results indicate the upgradation in changing scenarios of bird classification, where deep learning models and feature extraction methods are increasingly contributing to accuracy improvements.

III. PROPOSED SYSTEM

The methodology for this project involves building a bird species recognition system using machine learning and deep learning techniques, combined with real-time data retrieval and dynamic display through Google Sheets. The process was divided into several key steps, ensuring smooth data handling, model training, and frontend integration. The proposed system is described in Fig. 1.

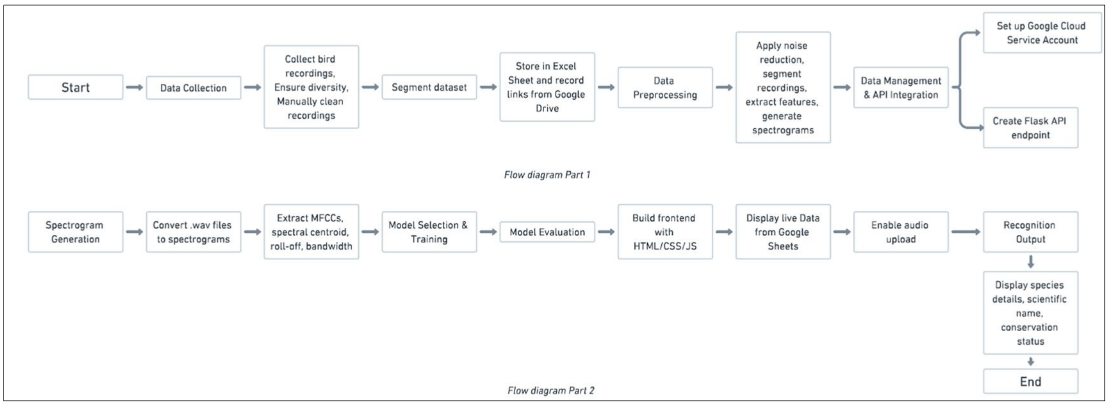

Fig. 1. Proposed System

A. Data Collection

The project began by gathering bird audio recordings from the Indian Bird Song website that provided various metadata about species names along with their locations and audio files. The metadata played a key role in getting the contextual information of the bird recordings which was used to display it in front-end.
The dataset achieved diversity using audio files obtained from various locations, recorded by different contributors. The data collection process considered bird call variations attributed to environmental factors including geographic location and climate along with environmental noise levels. A human-based verification procedure evaluated the recordings to eliminate those with excessive noise levels or overlapping callouts and poor quality to maintain only high-quality data.
We divided the dataset alphabetically to assign individual letters as processing tasks for team members. Each team member received individual letters to process data to manage distribution of work equally and stop accidental repetition. The recordings were transferred to mp3cut.net to be trimmed to appropriate durations before transformation into audio files with .wav format using high-quality encoding methods. After conversion the processed files received storage at Google Drive and their relevant URLs were added to an Excel database for quick accessibility.

B. Data Preprocessing

The gathered audio data needed preprocessing work to refine its quality before it reached an optimal state for feature identification. The identification of bird calls suffered from the environmental noise created by wind conditions and water sounds together with human disturbances in the unprocessed audio recordings. The research used spectral subtraction together with Wiener filtering techniques for reducing noise in the recordings. The spectral subtraction method created background noise estimates from quiet parts of the signal which then subtracted from the recorded sound while Wiener filtering automatically cleansed background noise while improving sound clarity without modifying bird calls.

The meaningful bird calls and all other audios were trimmed into short segments of manageable length during the processing stage. Relevant data obtained through this process was used to keep the model straightforward and free from unnecessary complexity.

The trimmed recordings were saved by separating meaningful bird calls from quiet parts and overlapping sounds. This was done to only include relevant data to process, keeping the model free from unnecessary complexity.

Extracting the key acoustic features next isolates unique aspects of each bird's call. The Mel-Frequency Cepstral Coefficients (MFCCs) were calculated, converting the sound into a representation highlighting the perceptual aspects of sound. Spectrograms were created to offer visual time-frequency representations, which simplify the detection of patterns in bird calls. The extracted features comprising pitch as well as tempo along with tone qualities assisted the model in species identification.

C. Data Management and API Integration

Google Sheets was used to store data because it was adaptable and allowed real-time updation, facilitating effortless changes without altering the backend. The dataset consisted of names of species, scientific names, taxonomy, distribution, and conservation status, in a structured, easily accessible format.

The data was uploaded on Google Drive, transferred to Google Sheets, and merged utilizing the Google Sheets API. A Service Account was made within Google Cloud Console for secure authentication, and permission was provided to access the sheet. The ID of the sheet (SHEET_ID) was extracted for API requests.

A Python script from gspread and oauth2client authenticated and loaded data successfully to verify the configuration. A Flask API with the /birds endpoint was created to deliver the data in JSON form, with Pandas guaranteeing a neat hierarchy. Error checking handled missing information and connectivity loss, with security being maintained by authentication.

The system was being put to the test to validate real-time Google Sheet updates that appeared immediately in the API, providing dynamic, reliable data access.

D. Spectrogram Generation

Following data gathering, preprocessing, and API configuration, the second major step was transforming the preprocessed .wav files into spectrograms — graphical representations of sound displaying frequency patterns over time. In contrast to raw waveforms, spectrograms expose important frequency-based information, allowing the model to differentiate bird species based on their distinct vocal signatures.

The.wav format was used due to its uncompressed, high-fidelity nature, which would not lose any frequency data in the conversion. Spectrograms were created with the Short-Time Fourier Transform (STFT), which divides the audio into short pieces to measure frequency changes over time. The Librosa library performed this transformation, using an optimal window size and Hann window function to minimize noise and provide clean frequency resolution. The spectrograms were then transformed to Mel spectrograms, scaling frequencies to the Mel scale — a more perceptually representative account of how birds make and hear sound. Min-Max normalization and contrast enhancement were used to maintain steady intensity and enhance visibility of subtle frequency patterns.

Lastly, the spectrograms were stored as PNG images, with labels by species as illustrated in Fig. 2 at the end of this page. These labeled images formed the primary dataset used to train the deep learning model.

E. Data Augmentation

To enhance the model's strength and resilience to real-world variations in bird calls, data augmentation was carried out on the audio files as well as spectrograms. The approach implemented in this stage created broader diversity within the dataset so that the model gained better ability to perform across different registration conditions and environments.

Audio processing techniques implemented time stretching which modified audio speed only and maintained pitch while simulating how natural bird vocalizations may change in tempo. The technique of pitch shifting adjusted sonic frequencies to correct variations between different bird species or locations and different ages of the birds. The addition of noises including rain and wind helped the model prepare for processing authentic and noisy ecological data. The audio signal's volume received adjustments through scaling tactics to represent changes in sensitivity levels and distance variations and time masking eliminated segments of the signal similarly to noise disturbance in recordings. Spectrograms received visual augmentation procedures to boost their data point variability. The model gained the ability to identify bird calls through frequency masking which blocked random spectrum bands regardless of whether the audio signals were damaged or altered. Time masking digitally damaged parts of the spectrographic image to simulate damaged recordings. The model was prevented from relying on specific patterns through spectrogram image deviations that resulted from rotation and warping procedures. The different spectrogram visualization parameters were manipulated to remove their effect on intensity and brightness variations. The expanded dataset improved the model's understanding of diverse acoustic patterns, enabling accurate bird species identification even in noisy recordings.

F. Feature Extraction

The model gained stronger performance capability through applying data augmentation procedures to both audio files and spectrograms to handle natural variations in bird sound patterns. Data enhancement techniques expanded dataset variety, so the model gained better ability to work across multiple environmental conditions and recording equipment. Time-stretching capabilities shifted the audio playback tempo but left the pitch untouched to duplicate the natural tone variations present in bird vocalizations. The frequency shifting technique adjusted sounds to match different bird ages and locations or personal traits. The addition of noise elements like rain noise with background wind sounds prepared the model for processing actual data that contains background noise. The volume scaling parameter adjusted audio amplitude levels to replicate microphone responsiveness changes or to approximate changing bird positions while time masking sections introduced auditory gaps for imitating noise-related recording data loss. The second augmentation technique manipulated spectrograms to enhance visual dataset diversity. Frequency masking blocked random frequency bands from the model which enabled identification of bird calls even when audio spectrum areas became inaccessible or damaged. Time masking randomly cut segments from spectrogram images to simulate situations when recordings become incomplete. The model was prevented from focusing on specific spectrogram patterns through small rotational and warping modifications of its visual images. Evaluation tests included contrast modifications for eliminating visual differences which stemmed from various spectrogram display parameters. With added transformations, the enlarged dataset improved generalization, enabling accurate bird species detection even in noisy conditions.

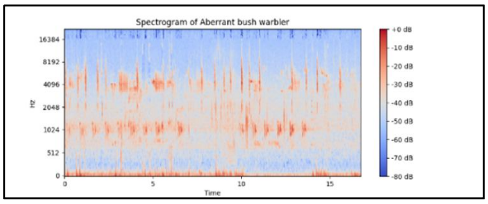

Fig. 2. Spectrogram of an Audio File

G. Model Selection and Training

A variety of deep learning techniques function to establish the best possible framework when using spectrogram images for bird sound classification. The selection of these models occurred because they displayed significant effectiveness and feature analysis strength alongside efficient data processing capabilities in earlier research work. Among the training models chosen by the analysis, lay ConvNeXt-Tiny, ResNet101, ResNet18, RegNet_Y_800MF, Swin-Tiny, DenseNet121, and CoatNet_0_RW_224. A standard testing protocol was used to check model fairness through constant parameters leading to recording the most accurate results.

Standardized training procedures ensured equal conditions between multiple models throughout the training process. The collection of spectrogram images occurred before their organization into specific class directories. The images underwent resizing operations which transformed them to 224 × 224 pixels dimensions to match requirements of convolutional neural networks. The model obtained better generalization power by implementing data augmentation methods including random horizontal flips and random rotations in combination with color adjustment procedures. The training and validation data split happened while preserving a stratified distribution of classes because this maintained normal model performance during the process.

The system implemented a specialized dataset class because this approach enhanced image loading and transformation operations. The training pipeline depended on PyTorch alongside torchvision library components to adjust pre-trained network layers into classifiers suitable for the bird sound classification assignment. A training period of 30 epochs together with an AdamW optimizer ran at 1e^-4 learning rate while weight decay

was set to 1e^-5 and batch size was 16. The chosen loss function utilized cross-entropy since it proves effective for multi-class classification challenges. Training of the models took place using GPU hardware for increased computing speed and shortened overall training duration.
The model performance measurements including training loss and validation loss together with training accuracy and validation accuracy were recorded after every epoch during model training sessions. The ConvNeXt-Tiny model delivered the most successful results since it effectively processed spectrogram images which produced the highest classification precision. Also, competitive results were obtained by other models ResNet101 and ResNet18 which established their capability for spectrogram-based audio classification tasks. The outcomes from this research demonstrate that convolutional neural networks excel in bioacoustic assessments by conducting automatic bird species recognition through auditory signals.

H. Model Evaluation

Each trained model received an independent evaluation through unseen data to check its effectiveness in bird species identification. The prediction accuracy acted as the main metric to measure correctness, yet precision provided details about bird misidentification errors. Each species identification performance was evaluated based on recall with combined precision and recall through F1-score measurement to achieve a complete performance evaluation. Analysis of species misidentifications created a confusion matrix that pointed out which species the model repeatedly confused with others. An ROC curve together with AUC evaluation methods allowed us to determine how accurately the model differentiated between various species of birds. The model stability across data splits was verified using cross-validation while detailed inspections on misclassified samples indicated similar bird calls or background noise contributed to most errors.
Performance enhancement through the hyperparameter tuning process enabled modifications of learning rates together with batch sizes and optimization approaches. The model's robustness received essential improvement through data augmentation where transformations incorporated with background noise removal techniques were used. After achieving stable results during model evaluation, the system became ready for deployment purposes, becoming able to classify bird species in real-time.

I. Deployment

The last development period involved working to merge intuitive user interfaces with real-time bird species identification together with improvements between front-end and back-end frameworks. The team began developing a frontend interface following accuracy checks with Streamlit before coding both the frontend and backend components of the application using a mixture of HTML CSS and JavaScript in combination with the Python-based backend. The pre-installed user interface designed its model to achieve maximum performance. The system used a Flask API to connect with Google Sheets for displaying bird data to users who could execute querying and filtering functions. The audio upload function produced spectrograms with species predictions and included species classifications and their conservation status results. During system testing the API received stability enhancements and processing performance improvements that streamlined the efficient and precise bird classification features for users.

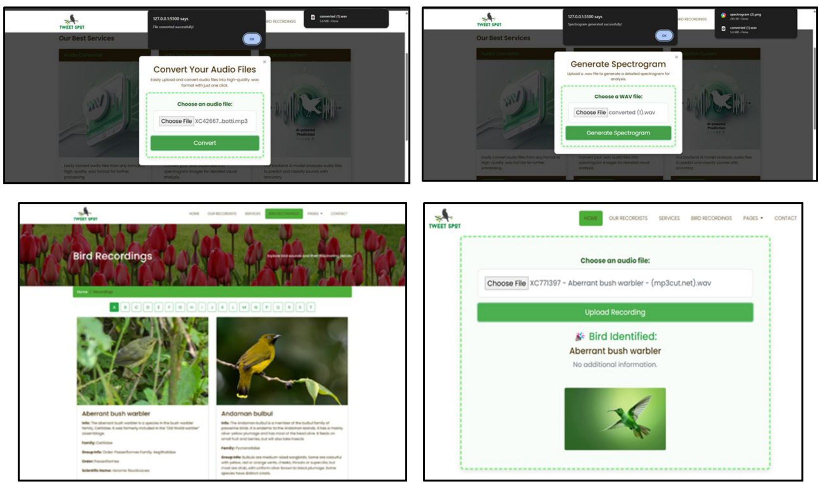

Fig. 3. Overview of the system interface, showcasing high-quality .wav file conversion, spectrogram generation, bird data retrieval from the API, and bird species prediction.

IV. RESULTS AND DISCUSSION

The ConvNeXt-Tiny showed inadequate accuracy levels at the beginning of training which amounted to 0.75% training precision and 1.99% validation precision. The model achieved data extraction competency through epoch three where training precision reached 25.65% and validation precision reached 30.85%. AdamW served as the optimizer in the training process at a learning rate of 1e^-4 together with a weight decay of 1e^-5. The training data received three augmentations involving flipping and rotation operations and color transformation to avoid model overfitting.

During the tenth epoch of ConvNeXt-Tiny's training period the model achieved 98.88% accuracy before validating at 56.22% that proved almost identical to ResNet101's 55.22% validation value. The maximum validation accuracy occurred at Epoch 20 thus training became stalled because further improvement was not possible. The training terminated automatically because the system included an integrated mechanism to prevent overfitting. Our testing demonstrated that ConvNeXt-Tiny led the accuracy charts with 62.69% while ResNet101 achieved 59.70% and ResNet18 ended with 58.21%. The validation accuracy for RegNet_Y_800MF was 57.71% and Swin-Tiny generated an accuracy of 56.72% and DenseNet121 produced 55.72% and CoatNet_0_RW_224 generated the lowest with 53.23%. The model achieved better generalization capability through data augmentation which allowed it to interpret numerous bird calls and environmental noise patterns. The human-assisted noise reduction techniques improved the clarity of spectrograms. The /birds API processed Google Sheets data through an execution time that remained lower than 200 milliseconds. Error handling-maintained system stability. A validation accuracy of 62.69% was achieved by Mini ConvNeXt-Tiny model when tested with a loss value of 1.7518 and execution time of 1–2 seconds for each classification.

V. CONCLUSION

The objective of our research was to apply Deep Neural Networks for bird sound classification while creating an Indian bird species dataset comprising more than 1,000 recorded sounds. Background noise went through manual removal to make the input quality better. An MFCC extraction process produced spectrograms for the classification procedure. The usage of data augmentation techniques enabled better generalization capabilities of the model.

The effectiveness of spectrogram-based bird sound classification was determined through evaluation of deep learning frameworks consisting of ConvNeXt-Tiny, ResNet101, ResNet18, RegNet_Y_800MF, Swin-Tiny, DenseNet121 and CoatNet_0_RW_224. The research shows that larger intricate architectural structures produce better results by extracting useful features from spectrogram images which lead to higher accuracy rates.

Better detection and classification of vocalizations combined with address of delay problems will advance ecological observation capabilities and wildlife protection efforts and bird habitat interaction research.

TABLE I. PERFORMANCE MATRIX

Model	Accuracy	Loss
ConvNeXt-Tiny	62.69%	1.7518
ResNet101	59.70%	1.9175
ResNet18	58.21%	1.8459
RegNet_Y_800MF	57.71%	1.8918
Swin-Tiny	56.72%	2.2162
DenseNet121	55.72%	1.9189
CoatNet_0_RW_224	53.23%	2.5034

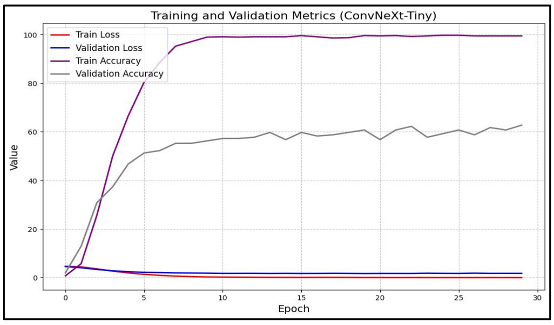

Fig. 4(a). ConvNeXt-Tiny

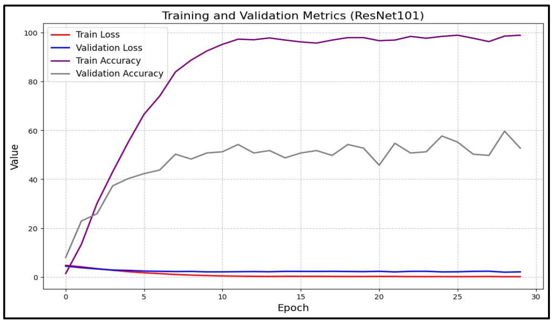

Fig. 4(b). ResNet101

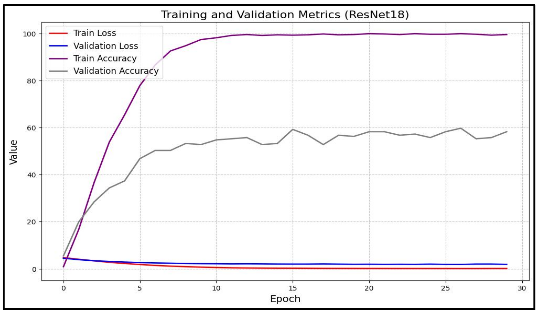

Fig. 4(c). ResNet18

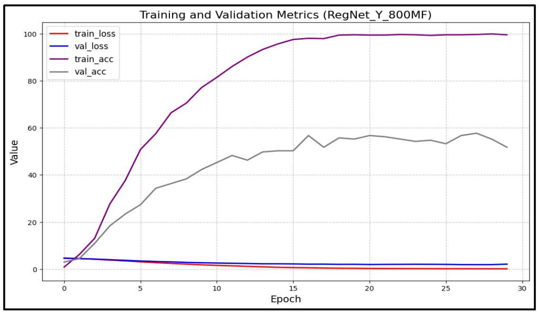

Fig. 4(d). RegNet_Y_800MF

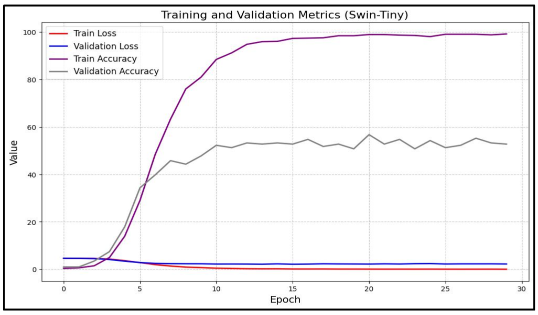

Fig. 4(e). Swin-Tiny

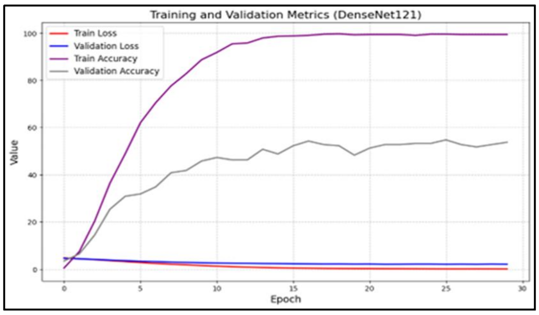

Fig. 4(f). DenseNet121

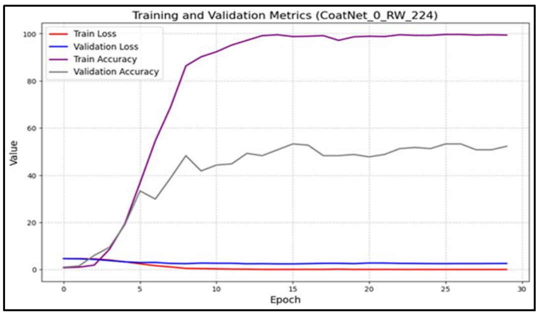

Fig. 4(g). CoatNet_0_RW_224

Fig. 4. Comparison of different deep learning models used for bird sound recognition

REFERENCES

[1] Gavali, P., & Banu, J.S. (2025). A novel approach to Indian bird species identification: employing visual-acoustic fusion techniques for improved classification accuracy. Frontiers in Artificial Intelligence, 8, 1527299.

[2] Swaminathan, B., Jagadeesh, M., & Vairavasundaram, S. (2024). Multi-label Classification for Acoustic Bird Species Detection Using Transfer Learning Approach. Ecological Informatics, 80, 102471.

[3] Wang, Q., Song, Y., Du, Y., Yang, Z., Cui, P., & Luo, B. (2024). Hierarchical-taxonomy-aware and attentional convolutional neural networks for acoustic identification of bird species: A phylogenetic perspective. Ecological Informatics, 80, 102538.

[4] Ugarte, J.P., & Arias-Arias, J. (2024). Unveiling relevant acoustic features for bird species automatic classification. Expert Systems With Applications, 257, 125046.

[5] Indumathi, C.P., Diviyalakshmi, K.R., & Mahalakshmi, R. (2024). Bird Sound Identification System using Deep Learning. Procedia Computer Science, 233, 597–603.

[6] Magumba, M.A., Denton, T., & Bashir, M. (2024). A bird vocalisation dataset of birds in Uganda for automated bio-acoustic monitoring and analysis. Data in Brief, 54, 110433.

[7] Amenyedzi, D.K., et al. (2024). System Design for a Prototype Acoustic Network to Deter Avian Pests in Agriculture Fields. Agriculture, 15, 10.

[8] MacPhail, A.G., et al. (2024). Audio data compression affects acoustic indices and reduces detections of birds by human listening and automated recognisers. Bioacoustics, 33(1), 74-90.

[9] Kshirasagar, S., Guntoro, A., & Mayr, C. (2024). Impact of Sliding Window Variation and Neuronal Time Constants on Acoustic Anomaly Detection Using Recurrent Spiking Neural Networks in Automotive Environment. Algorithms, 17(440).
[10] Michaud, T., Harvieux, C., & Charbonneau, N. (2023). Deep Learning for Bird Sound Classification in Noisy Environments. Journal of Applied Acoustics, 200, 103943.
[11] Ramashini, M., et al. (2022). Bird Sound Classification Using Support Vector Machine with LPCC, MFCC, and GTCC Features. Journal of Audio Engineering Society, 70(4), 245-257.
[12] Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., & Ferres, J.L. (2021). Comparing recurrent convolutional neural networks for large scale bird species classification. Scientific Reports, 11, 17085.
[13] Gunawan, K.W., Hidayat, A.A., Cenggoro, T.W., & Pardamean, B. (2021). A Transfer Learning Strategy for Owl Sound Classification Using Image Classification Models with Audio Spectrogram. International Journal of Electrical Engineering & Informatics, 13, 546-553.
[14] Tuncer, T., Dogan, S., & Subasi, A. (2021). Multilevel Ternary Patterns and Iterative ReliefF for Bird Sound Recognition. Journal of Computational Science, 53, 101393.
[15] Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. Methods in Ecology and Evolution, 10(3), 368–380.
[16] Stastny, J., Munk, M., & Juranek, L. (2018). Automatic bird species recognition based on bird vocalization. EURASIP Journal on Audio, Speech, and Music Processing.
[17] Incze, Á., Jancsó, H.B., Szilágyi, Z., Farkas, A., & Sulyok, C. (2018). Bird Sound Recognition Using a Convolutional Neural Network. SISY 2018 - IEEE 16th International Symposium on Intelligent Systems and Informatics, 295-300.
[18] Sathyapriya, R., & Muthuswamy, S. (2023). Bird species recognition using deep learning models and ensemble techniques. Neural Computing and Applications, 35(4), 1224–1239.
[19] Kim, J., & Lee, S. (2022). Bird sound classification using capsule networks and attention mechanisms. IEEE Transactions on Audio, Speech, and Language Processing, 30, 1742–1755.
[20] Li, Y., Zhang, H., & Feng, J. (2023). Acoustic event classification using hybrid CNN-RNN architectures. Pattern Recognition Letters, 160, 150–156.