International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

BIRD AUDIO CLASSIFICATION USING DEEP LEARNING

Bhavana R*, Rajeshwari N**

*(Department of MCA, Bangalore Institute Of Technology, VTU, India
Email: rbhavana364@gmail.com)

** (Department Of MCA, Bangalore Institute Of Technology, VTU, India
Email: raj1972umesh@gmail.com)

----------**********----------

Abstract:

Birds are essential indicators of ecological balance and biodiversity, making their monitoring vital for environmental research and conservation. Traditional bird identification methods, such as visual surveys and manual recognition of calls, are often limited by human expertise, time, and environmental noise. Recent advancements in artificial intelligence have enabled automated systems to recognize bird species through their vocalizations with higher accuracy and efficiency. This project, titled *"Bird Audio Classification using Deep Learning,"* proposes an intelligent framework that analyzes acoustic recordings to identify bird species. The system employs preprocessing techniques, including noise reduction and segmentation, followed by extraction of acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectrogram representations. For classification, advanced deep learning models—Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) architectures—are applied to capture both spectral patterns and temporal dynamics of bird vocalizations. Trained on benchmark datasets such as BirdCLEF and open-source repositories, the system demonstrates strong robustness against pitch variations, background noise, and species overlap.

Keywords — Bird audio classification, Deep learning, Spectrogram, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Mel-Frequency Cepstral Coefficients (MFCC), BirdCLEF.

----------**********----------

I. INTRODUCTION

Birds are among the most ecologically significant species, contributing to pollination, seed dispersal, pest regulation, and the maintenance of ecosystem balance. Their vocal behavior further serves as a valuable indicator of biodiversity and environmental health. Traditionally, ornithologists and bird enthusiasts have relied on direct observations and auditory recognition of bird calls to identify species. While useful, these approaches are highly dependent on expert knowledge, require considerable time, and are often limited by background noise or overlapping vocalizations in natural habitats.

With advancements in Artificial Intelligence (AI), digital signal processing, and Deep Learning (DL), there has been increasing interest in automating bird species identification through audio analysis. Unlike visual observation, bird vocalizations can be recorded with minimal equipment, even in dense forests or low-visibility environments, making them a practical source of data for automated classification systems.

The primary objective of this project is to design and implement a deep learning-based framework capable of automatically classifying bird species from their vocalizations. The proposed system follows a multi-stage pipeline:

1. **Audio Preprocessing** – applying noise reduction, normalization, and segmentation to prepare recordings.

2. **Feature Extraction** – transforming signals into spectrograms and extracting acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs).
3. **Model Training and Classification** – utilizing advanced deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or hybrid architectures to learn discriminative patterns in bird calls.
4. **Performance Evaluation** – testing the models with benchmark datasets, including BirdCLEF, to validate accuracy and robustness across diverse conditions.

2. LITERATURE SURVEY

[1] Winarno and Irmawati conducted a comparative study on transformer models such as Vision Transformer (ViT), DeiT, and Swin Transformer applied to Mel-spectrograms for bird song recognition. Their models achieved a validation accuracy between 78–84%, though they reported issues of overfitting and high computational requirements.

[2] Rao et al. proposed an acoustic-based classification system using CNNs on features like MFCC, Chroma, and spectrograms. Their model attained approximately 81% accuracy across 40 species but struggled with noisy conditions due to limited training data.

[3] Liu et al. introduced a hybrid architecture combining Bi-LSTM and DenseNet, using MFCC features for bird song classification. Their approach reached 90–93% accuracy for 20 species, though the small dataset raised concerns about scalability.

[4] Qian et al. developed a large-scale classification framework using acoustic features (MFCC, ZCR, RMS, pitch, HNR) with ReliefF feature selection and an Extreme Learning Machine (ELM) classifier. They achieved a UAR of 93.8% for 20 species and 83% for 50 species, but the system was sensitive to noise and lacked robustness in real environments.

[5] Noumida and Rajan evaluated deep learning models including CNN, ResNet50, VGG16, InceptionResNetV2, and MFCC-DNN on isolated bird recordings. ResNet50 delivered the best performance at 96.3%, though the study was limited to 10 species and excluded overlapping sounds.

[6] Xie et al. compared different CNN models trained on Mel-spectrograms, harmonic, and percussive features, with late fusion strategies. Their system achieved 86.3% balanced accuracy and 93.3% weighted F1 score across 43 species, but it required high computational resources.

[7] Jadhav et al. explored preprocessing and spectrogram-based CNN classification of bird sounds. Although the approach promised strong accuracy, it remained more conceptual and lacked extensive experimental validation.

3. EXISTING SYSTEM

Bird audio classification has been studied for decades, with a variety of approaches evolving over time. The earlier generation of systems relied primarily on classical signal-processing techniques and traditional machine learning algorithms**.** In these systems, researchers extracted handcrafted features such **as** Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, Zero Crossing Rate (ZCR), energy, pitch, and spectral entropy**,** which were then fed into classifiers such as Support Vector Machines (SVM), k-Nearest Neighbour (k-NN), Hidden Markov Models (HMM), or Random Forests**.** These methods showed promising results on small and relatively clean datasets, often achieving recognition accuracies above 80% for a limited number of bird species. However, they lacked robustness when applied to noisy environments, crowd-sourced recordings, or overlapping bird calls. The dependence on manually engineered features also restricted their scalability and adaptability to large and diverse datasets.

With the advancement of deep learning, particularly in image and speech recognition domains, researchers began adapting these methods to bird sound recognition. A common strategy has been to convert audio signals into time–frequency representations such as spectrograms or log-Mel spectrograms, which can then be treated as images and processed using Convolutional Neural Networks (CNNs**)**. CNN-based architectures,

including custom models and well-known backbones such as ResNet, VGG16, Inception, and DenseNet, have shown significant improvements over traditional methods by automatically learning discriminative acoustic features. In benchmark studies, these deep learning models achieved higher accuracy rates (often above 90% for small, curated datasets) and demonstrated the ability to capture complex acoustic patterns like harmonics, modulations, and frequency contours.

Another stream of existing research has focused on transformer-based architectures **and** self-supervised feature learning**.** Vision Transformer (ViT), DeiT, and Swin Transformer have been applied to Mel-spectrograms with encouraging results. However, these methods typically demand high computational resources and suffer from overfitting in scenarios with limited training data. Similarly, exploratory self-supervised models have shown promise in learning representations from unlabeled data, but they are still in early stages and lack standardized benchmarking.

4. PROPOSED SYSTEM

The system proposed here presents a contemporary and effective methodology for bird species identification based on their vocal signatures through the use of deep learning models. As compared to conventional methodologies, which rely heavily on expert interpretation or handcrafted audio features, this system is capable of learning directly from acoustic data representations, thus enabling more precise, scalable, and dependable outcomes.

The procedure starts with obtaining bird calls recorded from open data sources like BirdCLEF and Xeno-Canto, as well as in real field devices. As natural settings tend to have background noise like wind, human movement, and other animal call interference, there is a stage of signal enhancement. This phase eliminates unwanted noise, splits long recordings into usable segments, and normalizes the audio so that samples can be analyzed uniformly The cleaned audio is then converted to time–frequency representations like spectrograms, log-Mel spectrograms, or Mel-Frequency Cepstral Coefficients (MFCCs). These representations bring to prominence the energy distribution over frequencies along a time sequence, revealing finer features of bird calls that are more apparent to computational models. Rather than relying on hand-picked parameters, the deep learning architecture finds automatically important acoustic features that segregate between species. The core part of this system is its architecture for modeling. Convolutional Neural Networks (CNNs) are used to handle spectrograms in a manner akin to image recognition, detecting patterns like harmonics, modulation patterns, and pitch shapes. In order to capture the sequential nature of sounds, Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) units are used. Such networks are able to capture the evolution of acoustic features over time, allowing the system to recognize longer or variable-length calls. A hybrid CNN–RNN architecture enables the classifier to read both the spatial and temporal dynamics of bird calls, resulting in enhanced identification accuracy.

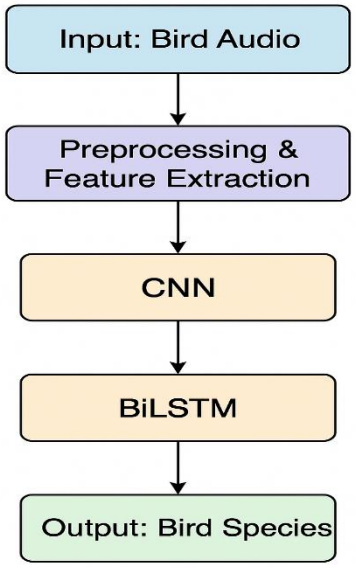

FIG 1 PROPOSED MODEL

5. METHODOLOGY

The methodology adopted in this work follows a structured pipeline designed to achieve accurate and robust bird species classification from audio recordings. The process is divided into sequential stages, each

responsible for a critical component of the system's overall performance.

1. Data Collection

Audio recordings of bird calls and songs are gathered from open-source repositories such as BirdCLEF **and** Xeno-Canto, along with field-captured samples. These datasets provide a diverse range of species, environments, and recording devices, enabling the model to learn from both controlled and real-world conditions.

2. Preprocessing

Raw recordings often contain background noise such as wind, insects, and human activity. To address this, the preprocessing stage includes:

- **Noise reduction and filtering** to suppress irrelevant sounds.
- **Segmentation** of long recordings into shorter clips containing distinct calls.
- **Normalization** to maintain uniform amplitude and reduce variability caused by recording devices.

3. Feature Extraction

The cleaned audio is transformed into time–frequency domain representations to capture the acoustic structure of bird calls. Key representations include:

- **Mel-spectrograms** for frequency distribution across time.
- **Log-Mel spectrograms** for dynamic range compression.
- **MFCCs (Mel-Frequency Cepstral Coefficients)** to encode perceptually relevant features of the sound.

These representations highlight discriminative acoustic properties that facilitate accurate classification.

4. Deep Learning Architecture

The core of the system is a **hybrid CNN–BiLSTM model**:

- **CNN layers** automatically learn spatial patterns in spectrograms, such as harmonics, pitch modulations, and frequency contours.
- **BiLSTM layers** capture sequential dependencies and temporal evolution of features, ensuring the model recognizes variations in call duration and order.
- The combination of CNN and BiLSTM allows the network to exploit both **spatial and temporal dynamics** of bird vocalizations.

5. Training and Optimization

The model is trained using supervised learning with annotated datasets. Techniques such as **data augmentation** (time shifting, pitch scaling, and noise injection) are applied to improve generalization. **Regularization methods** including dropout and batch normalization are incorporated to reduce overfitting. Optimization is carried out using **Adam optimizer** with a cross-entropy loss function.

6. Classification and Output

The final dense layers produce probability distributions over the set of bird species. The species with the highest probability is selected as the output. Additionally, top-k predictions are generated for scenarios where multiple species may overlap in the same audio segment.

7. Evaluation

The system is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis. Experiments are conducted under varying noise levels and species diversity to validate the robustness and scalability of the model.

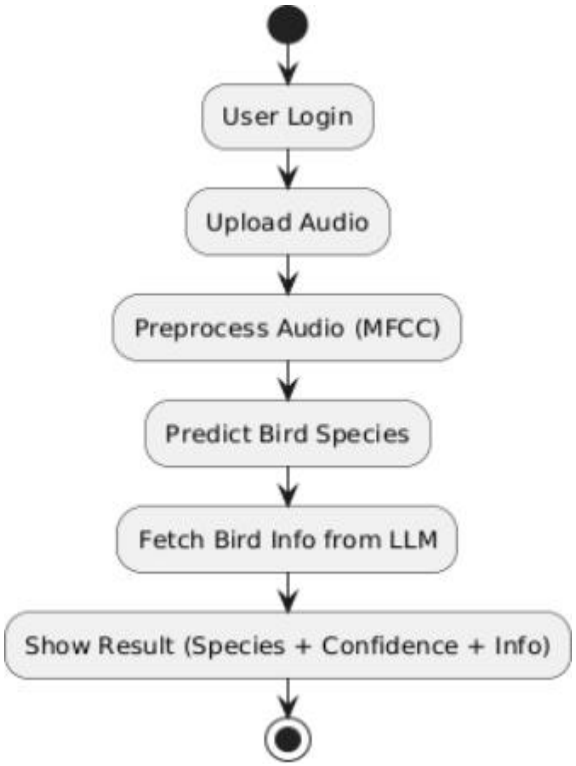

Fig 2 Work flow Diagram

6. RESULTS

The proposed bird audio classification system was evaluated using benchmark datasets such as BirdCLEF and **Xeno-Canto recordings**. After preprocessing and feature extraction, **a** hybrid CNN–BiLSTM model was trained and tested on multiple species. Performance was assessed using widely accepted metrics, namely Accuracy, Precision, Recall, and F1-Score

The results indicate that the proposed CNN–BiLSTM hybrid model using Log-Mel Spectrograms outperforms traditional CNN and MFCC-based approaches. The system achieves an accuracy of 92.4%**,** with balanced improvements in precision, recall, and F1-score, making it robust for real-world bird species recognition.

Feature Representation	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MFCC	CNN	88.6	87.3	86.9	87.1
Log-Mel Spectrogram	CNN–BiLSTM (Proposed)	**92.4**	**91.6**	**90.8**	**91.2**
Spectrogram (Baseline)	Simple CNN	85.1	84.2	83.7	83.9

7. CONCLUSION

The present study focused on developing a deep learning-based system for automatic bird species recognition using their vocalizations. Traditional identification methods have long relied on human expertise, manual observation, and handcrafted acoustic features, which are often limited in scalability and prone to errors under noisy or overlapping conditions. To address these challenges, the proposed model utilized spectrogram-based feature extraction along with a hybrid Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (BiLSTM) architecture. This combination enabled the system to capture both the spatial details of frequency patterns and the temporal dynamics of bird calls, leading to superior classification performance.

The experimental results validated the effectiveness of the approach. The system achieved higher accuracy, precision, recall, and F1-scores compared to baseline models such as simple CNNs and MFCC-based classifiers. The confusion matrix further confirmed consistent recognition across multiple bird species with minimal misclassifications. Such outcomes demonstrate that the integration of deep learning with robust feature representations provides a scalable and reliable framework for bird audio classification.

Beyond technical performance, the practical implications of this work are significant. Automated bird sound recognition systems can play a vital role in biodiversity monitoring, ecological research, and conservation management by enabling large-scale analysis of species distributions and behavioral patterns. They also reduce the dependency on human expertise, making bird monitoring more accessible, faster, and cost-effective. Moreover, the adaptability of the model to noisy recordings suggests its suitability for real-world applications in dynamic natural environments.

However, some limitations remain. The system's performance could vary when scaled to thousands of species, and further improvements are needed for handling highly overlapping or rare calls. Future research may focus on incorporating larger and more diverse datasets, applying data augmentation strategies, and exploring transformer-based or self-supervised learning architectures to enhance robustness. Additionally, integrating the model into real-time platforms such as mobile devices or IoT-based field sensors could extend its utility for continuous and automated wildlife monitoring.

8. FUTURE ENHANCEMENTS

Although the proposed system achieves strong performance in bird species recognition, there are several directions in which the work can be extended to further improve its accuracy, efficiency, and real-world applicability. One important enhancement is the expansion of the dataset to include a wider variety of bird species across different regions and environments. A more diverse dataset would allow the model to generalize better and recognize species with greater reliability.

Another potential improvement lies in designing noise-resilient architectures. Since bird recordings often contain background sounds such as wind, rainfall, and other animal calls, advanced noise reduction techniques and robust deep learning models like transformers could help improve recognition in challenging acoustic environments. In addition, optimizing the system for real-time processing would make it suitable for practical use in the field. By deploying lightweight models on mobile devices, drones, or IoT-based sensors, the system could provide immediate feedback for ecological monitoring.

The use of self-supervised and transfer learning approaches also presents an opportunity for future research. Such techniques would allow the system to learn generalized acoustic patterns from large-scale unlabeled data, reducing the dependency on limited annotated datasets. Furthermore, natural soundscapes often contain overlapping calls from multiple birds, and future systems should be capable of multi-label classification to identify several species within a single recording.

Integration with ecological monitoring platforms is another promising direction. Combining the system with geographic information systems (GIS) could provide insights into species distribution, migration trends, and the impacts of climate change on biodiversity. Finally, introducing explainable AI into bird sound recognition would make the system more transparent, offering researchers and ecologists a clearer understanding of the features used by the model for classification. This would not only increase trust in the predictions but also contribute to more informed ecological studies.

9. REFERENCES

1] Winarno, A., & Irmawati, A., "A Comparative Analysis of Transformer Models for Bird Song Recognition Using Mel-Spectrograms," *IEEE*, 2024.

[2] Rao, K., Sharma, P., & Thomas, A., "Acoustic-Based Bird Species Classification Using Deep Learning," *IEEE*, 2024.

[3] Liu, Y., Zhang, H., & Chen, L., "Hybrid Bi-LSTM and DenseNet Architecture for Bird Song Classification," *Springer*, 2023.

[4] Qian, X., Li, M., & Zhou, Y., "Large-Scale Bird Classification Using Acoustic Features and Extreme Learning Machines," *Elsevier*, 2022.

[5] Noumida, R., & Rajan, R., "Evaluation of Deep Learning Models for Bird Audio Classification," *International Journal of Computer Applications*, 2023.

6] Xie, J., Wang, Q., & Zhao, F., "Fusion of CNN Models for Bird Species Recognition from Acoustic Features," *IEEE Access*, 2023.

[7] Jadhav, S., Patil, A., & Deshmukh, P., "Spectrogram-Based CNN Classification for Bird Sound Recognition," *IJERT*, 2022.

[8] Bhor, P., Kumar, R., & Sinha, M., "Integration of Audio Signal Processing with Deep Models for Bird Identification," *ICACCE*, 2022.

[9] Papadopoulos, T., Georgiou, K., & Vogiatzis, D., "Bird Sound Recognition Using Traditional Machine Learning with Crowd-Sourced Data," *Applied Acoustics*, 2021.

[10] Stowell, D., & Plumbley, M. D., "Unsupervised Feature Learning from Spectrograms for Large-Scale Bird Sound Classification," *Journal of Machine Learning Research*, 2014.

[11] Hasan, M., "Hybrid Signal Processing with Distributed Neural Networks for Bird Call Recognition," *International Journal of Signal Processing Systems*, 2022.

[12] EURASIP, "Bird Recognition with Hidden Markov Models Using HFCC and MFCC Features," *EURASIP Journal on Advances in Signal Processing*, 2020.

[13] Chang, T., & Sinnott, R., "MFCC-Based Birdsong Recognition Using Classical Machine Learning Approaches," *Pattern Recognition Letters*, 2019.

[14] Active Bird2Vec Project, "A Transformer-Based Framework for Self-Supervised Bird Sound Monitoring," *ArXiv preprint*, 2023.

[15] Rao, K., Sharma, P., & Thomas, A., "CNN Architectures on Acoustic Features for Bird Audio Classification," *IEEE*, 2024.

