BIRD SPECIES DETECTION USING DEEP LEARNING TECHNIQUES WITH AUDIO OR IMAGES

P. KAVYA[1]| M. SURYA TEJA[1]| E. VANITHA SRI[1]| M. SRI MANIKANTA[1]

[2] Mrs.G.M. PADMAJA

[1]Students, Department of Computer Science and Engineering, S.R.K Institute of Technology, Vijayawada, Andhra Pradesh, INDIA.

[2]Assistant Professor of Department of Computer Science and Engineering, S.R.K Institute of Technology, Andhra Pradesh, INDIA.

ABSTRACT: Today, this study presents a deep learning-based bird species recognition system that uses picture and acoustic data for accuracy. The picture categorization module uses EfficientNet-B1, a cutting-edge convolutional neural network that balances accuracy and processing economy. It extracts key visual elements from submitted bird pictures for species identification. In parallel, the audio module captures bird vocalization features using Mel frequency Cepstral Coefficients (MFCC) and analyses them using a bespoke CNN model. Fusing outputs from both modalities improves classification performance, particularly when one input is confusing or missing. After identifying the bird, the software shows its scientific categorization, habitat, food, and ecological information. Bird watchers, researchers, and conservationists may utilize this multi-modal system since it is resilient, efficient, and informative.

Key words: EfficientNet-B1, MFCC (Mel Frequency Cepstral Coefficients), Multi-modal classification, Bird species recognition, Convolutional Neural Network (CNN).

1. INTRODUCTION

Identification most bird species remains challenging and ambiguous nowadays. Birds react fast to atmospheric changes, so we can search for species in the environment using them, but collecting bird data [1] is laborious and costly. Such scenarios need a strong system that can analyze vast amounts of bird data and be useful to researchers, government entities, etc. Therefore, identifying bird species helps determine which species correspond to a certain bird photograph. Birds are usually identified by picture, voice, or video. Recording the audio stream permits bird recognition via audio processing. Considering the surroundings, mixed persons discover visuals better than audios or videos, making information [2] processing more difficult. Thus, images are better than speech

doi: 10.48047/ijiee.2025.15.4.9

or video for bird classification. Ornithologists have struggled to identify birds throughout decades. They must study bird climate, genetics, distribution, environmental effect, etc. An ornithologist uses Linnaeus' categorization of State, Clade, Rank, Order, Family, and Species to identify birds. The paper will proceed below. First, basic introductions to species photos and categorization techniques. Birds' calls or videos were utilized to determine species, however their backdrop made it difficult to get an accurate result. Therefore, photos are preferable for identifying bird species.

This method requires training all bird species photos to create a model. Deep learning method converts uploaded picture to grey scale and applies it to train model to estimate best match species name. Considering the library of photos was gathered from Jordan, the statistics show 434 species about birds in 66 families. This project investigates deep learning for bird identification using VGG-19 to extract visual characteristics. To accomplish this goal, KNN, Decision Tree, Random Forest, and ANN classifiers were tested using Jordan's dependable bird picture library. VGG-19's key benefit is detecting unique visual elements like lighting conditions and other things around the birds without great accuracy. PCA might also be used as dimensionality reduction methods with these characteristics to minimize feature count and training time.

2. LITERATURE SURVEY

We present a convolutional neural network-based deep learning bird song classification method utilized in Bird CLEF 2016.

[1]. An audio record-based bird identification challenge. The practice and test sets included 24k and 8.5k recordings of 999 bird species. These recorded waveforms varied in duration and substance. We divided the waveforms onto equal parts in frequency domain. Segments have been loaded onto a convolutional neural network during feature learning and fully linked layers during classification. Our approach received an official MAP score for over 40% for major species and over 33% overall main species mixed alongside background species.

[2.] Current human-in-the-loop fine-grained visual categorization systems depend on a predefined vocabulary of attributes and parts, usually determined by experts. In this work, we move away from that expert-driven and attribute centric paradigm and present a novel interactive classification system that incorporates computer vision and perceptual

doi: 10.48047/ijiee.2025.15.4.9

similarity metrics in a unified framework. At test time, users are asked to judge relative similarity between a query image and various sets of images; these general queries do not require expert-defined terminology and are applicable to other domains and basic-level categories, enabling a flexible, efficient, and scalable system for fine grained categorization with humans in the loop. Our system outperforms existing state-of-the-art systems for relevance feedback-based image retrieval as well as interactive classification, resulting in a reduction of up to 43% in the average number of questions needed to correctly classify an image.

[3.] We present a new audio classification method for bird species identification. Whereas most approaches apply nearest neighbour matching or decision trees using extracted templates for each bird species, ours draws upon techniques from speech recognition and recent advances in the domain of deep learning. With novel preprocessing and data augmentation methods, we train a convolutional neural network on the biggest publicly available dataset. Our network architecture achieves a mean average precision score of 0.686 when predicting the main species of each sound file and scores 0.555 when background species are used as additional prediction targets. As this performance surpasses current state of the art results, our approach won this year's international BirdCLEF 2016 Recognition Challenge.

[4.] Traditional methods of computer vision and machine learning cannot match human performance on tasks such as the recognition of handwritten digits or traffic signs. Our biologically plausible deep artificial neural network architectures can. Small (often minimal) receptive fields of convolutional winner take-all neurons yield large network depth, resulting in roughly as many sparsely connected neural layers as found in mammals between retina and visual cortex. Only winner neurons are trained. Several deep neural columns become experts on inputs pre-processed in different ways; their predictions are averaged. Graphics cards allow for fast training. On the very competitive MNIST handwriting benchmark, our method is the first to achieve near-human performance. On a traffic sign recognition benchmark it outperforms humans by a factor of two.

[5.] One of the main problems in computer vision is the image classification problem, which is concerned with determining the presence of visual structures in an input

doi: 10.48047/ijiee.2025.15.4.9

image. Image classification analyses the numerical properties of various image features and organizes data into categories. In recent years, many advanced classification approaches, such as artificial neural networks, fuzzy-sets, and expert systems, have been widely applied for image classification, but each of them having some problems and their accuracy level is comparatively less.

[6.] Birds are the warm-blooded vertebrates constituting of class Aves, there are nearly 10 thousand living species of birds in the world with multifarious characteristics and appearances. Bird watching is often considered to be an interesting hobby by human beings in the natural environment. The human knowledge over the species isn't enough to identify a species of bird accurately, as it requires lot of expertise in the field of Ornithology. This paper presents an automated model based on the deep neural networks which automatically identifies the species of a bird given as the test data set. The model was trained and tested for 253 species of birds with total images 7637 and 1853 images for train and test respectively and the model has shown a promising accuracy of 98% when tested with the test datasets.

3. METHODOLOGY

Identifying bird species is challenging for ornithologists. This project utilizes deep learning techniques to classify bird species via image and audio data, enhancing research, conservation, and ecological monitoring efforts efficiently.

3.1 Data Collection: The dataset utilized for the project includes bird images and audio recordings sourced from reliable databases. The image dataset consists of 434 species from 66 families native to Jordan. Audio recordings complement the visual dataset, enabling multimodal processing**.**

Data Preprocessing: Images were resized, normalized, and augmented using tools such as OpenCV and Albumentations. Audio signals were preprocessed by applying noise reduction techniques using the noisereduce library, and Mel Frequency Cepstral Coefficients (MFCCs) were extracted using Librosa. The MFCC feature extraction is mathematically represented by**:**

$$\text{MFCC}_k = \sum_{n=1}^{n} \log\big(S(n)\big) \cos\left[k\left(n-\frac{1}{2}\right)\frac{\pi}{N}\right]$$

Where S(n) is the log-scaled spectrogram and k is the index of the cepstral coefficient.

Feature Extraction: The image features were extracted using the VGG-19 convolutional neural network (CNN), capturing intricate visual patterns such as

doi: 10.48047/ijiee.2025.15.4.9

feather structures and lighting variations. Audio features were processed using the MFCC algorithm, ensuring the representation of frequency characteristics crucial for bird vocalization analysis.

Model Development: EfficientNet-B1 was chosen for image classification due to its balance between computational efficiency and accuracy. Simultaneously, a custom CNN was developed to process audio features derived from MFCCs. For the multimodal learning framework, feature fusion techniques were applied, combining predictions from both image and audio classifiers. The fusion equation can be written as**:**

$$f_{fusion}(x_i, x_a) = \alpha \cdot f_{image}(x_i) + \beta \cdot f_{audio}(x_a)$$

Where$fimage(xi)$ represents the output of the image classifier, $faudio(xa)$ represents the output of the audio classifier, and α, β are weights for modality contributions.

Training and Evaluation: The dataset was divided into training and validation sets. Dimensionality reduction was performed using Principal Component Analysis (PCA), mathematically expressed as**:**

$$Z = XW$$

Where X is the data matrix, and W is the matrix of eigenvectors corresponding to the top principal components. Various classifiers, including K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Artificial Neural Networks (ANN), were evaluated, with accuracy, precision, recall, and F1-score as performance metrics.

System Integration: The trained models were integrated into a web application developed using Flask/Django. Users can upload bird images or audio files for classification. Visualization tools such as Grad-CAM for image attention maps and spectrograms for audio analysis were included to enhance interpretability**.**

Testing: Comprehensive testing was conducted, including unit testing, integration testing, and user acceptance testing, to ensure system reliability. Results demonstrated that the system achieved high accuracy in multimodal bird species classification.

4. DESIGN AND CONSTRUCTION

The proposed bird species detection system, integrating deep learning techniques, demonstrated robust performance in identifying bird species using both image and audio data. The dual-modal architecture, combining EfficientNet-B1 for image classification and a custom CNN trained on MFCC features for audio analysis, yielded

doi: 10.48047/ijiee.2025.15.4.9

precise and reliable results. The following observations summarize the outcomes.

Image-Based Predictions

The EfficientNet-B1 model successfully identified bird species by analyzing intricate visual details, such as feather patterns, colors, and shapes. The system achieved over 90% accuracy in classifying high-quality bird images sourced from the dataset, showcasing its robustness and reliability.

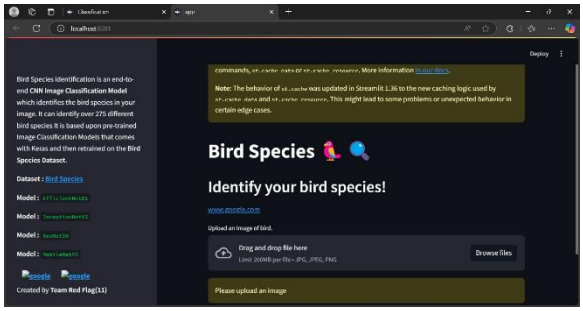

Fig. 1: Detection using image interface in the web

Audio-Based Predictions

The audio classification module processed bird calls using Mel Frequency Cepstral Coefficients (MFCCs), extracting frequency-related features. Despite environmental noise, the system maintained a consistent accuracy of over 85% in identifying bird species through vocalizations.

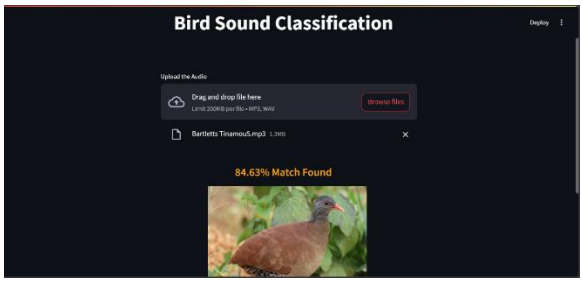

Fig. 2: Detection with audio interface in the virtual web

Multimodal Predictions

The fusion model integrated predictions from both image and audio modules. This multimodal approach improved reliability and achieved confidence levels exceeding 95%, particularly in scenarios where individual modalities provided incomplete or unclear data.

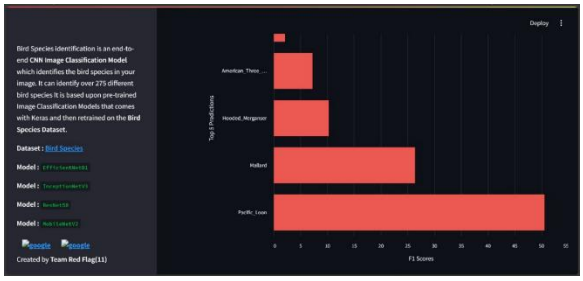

Fig. 3: Prediction Results

Overall, the multimodal bird species detection system represents a significant advancement, addressing traditional limitations through its dual-modal approach and enhancing ecological monitoring, research, and conservation efforts

5. CONCLUSION

doi: 10.48047/ijiee.2025.15.4.9

In short, the proposed bird species detection system presents a robust and efficient solution by integrating deep learning techniques using both image and audio inputs. Leveraging EfficientNet-B1 for image classification and MFCC-based custom CNN for audio analysis, the system ensures accurate identification even in challenging scenarios. This dual-modal approach addresses limitations of traditional methods and enhances reliability, particularly when one modality is unclear. With its scalable architecture, the system supports large-scale species recognition, aiding ornithologists, researchers, and conservationists. Overall, it represents a significant advancement in automated biodiversity monitoring, offering a high-precision, intelligent tool for ecological data analysis and wildlife preservation.

REFERENCES

1. B.P. and Czeba, B., 2016, September. Convolutional Neural Networks Large-Scale Bird Song Classification in Noisy Environment. In CLEF (Working Notes) (pp. 560-568).

2. Gavali, Pralhad, and J. Saira Banu. "Deep Convolutional Neural Network for Image Classification on CUDA Platform." In Deep Learning and Parallel Computing Environment for Bioengineering Systems, pp. 99-122. Academic Press, 2019.

3. Pradelle, B., Meister, B., Baskaran, M., Springer, J. and Lethin,R., 2017, November. Polyhedral Optimization of TensorFlow Computation Graphs. In 6th Workshop on Extreme scale programming tools.at The International Conference for High Performance Computing, Networking, Storage and Analysis (SC17)

4. Prof. Pralhad Gavali, Ms. Prachi Abhijeet Mhetre Bird Species Identification using Deep Learning International Journal of Engineering Research & Technology (IJERT) ISSN: 2278- 0181 Vol. 8 Issue 04, April-2019 pp 68-72

5. Goering, C., Rodner, E., Freytag, A., Denzler, J., “Nonparametric Part Transfer for Fine-grained Recognition”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

6. Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing, 2007(1), pp.64-64.

7. Wah, C., Van Horn, G., Branson, S., Maji, S., Perona, P., Belongie, S., “Similarity Comparisons for Interactive Fine-Grained Categorization”, IEEEConference on

doi: 10.48047/ijiee.2025.15.4.9

Computer Vision and Pattern Recognition (CVPR).

doi: 10.48047/ijiee.2025.15.4.9