iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69006

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Acoustic Bird Detection System Using Deep Learning

Prof. Aarti Bhise[1], Danish Nagaonkar[2], Vrushabh Ghodke[3], Varun Toshniwal[4], Tushar Shinde[5]
Computer Department Smt. Kashibai Navale College of Engineering Pune, India

Abstract: ***Researchers conduct a study that unites audio and image recognition methods to identify bird species through a workflow which combines spectrograms with Convolutional Neural Networks (CNNs). The preservation of species diversity requires effective monitoring of bird populations because biodiversity keeps declining. The methodology employs bird vocal recordings that get transformed into spectrograms through spectral analysis to extract temporal along with frequency signal characteristics. The analysis includes images captured in bird habitats that enhance the audio record evaluation process. The system starts with gathering different data sets which contain audio recordings together with images of multiple bird species. The processing of audio data creates spectrograms through Short-Time Fourier Transform (STFT) and the image data gets preprocessed for uniform presentation. The designed CNN model uses dual-input architecture to process spectrograms alongside images simultaneously. The training combines transfer learning with pre-trained networks for two purposes: improved execution and decreased computational demands.***

Keywords: BirdSpeciesIdentification,SpectrogramAnalysis,ConvolutionalNeuralNetwork.

I. INTRODUCTION

Research on bird species decline contributes critical concerns to both conservationists and ecologists and scientists worldwide. Ecosystems heavily depend on birds because they act as crucial pollinating agents and seed dispersers while simultaneously serving as environmental health indicators. The destruction of natural habitats combined with global warming and human habitat expansion has triggered an extremely concerning drop in global bird populations. Monitoring bird species effectively becomes essential to conservation work and ecological research because scientists need innovative methods to identify birds accurately. The current identification methods based on expert field surveys conduct time-consuming and labor-intensive processes to identify birds. Bird species recognition has received new possibilities through recent technological advancements in audio and image processing systems. The recording of bird vocalizations serves as a valuable evidence source which reveals both their presence and their behavioral patterns while interacting with other species. Acoustic signals undergo transformation to spectrograms for display of temporal frequency distribution which enables researchers to extract complex acoustic features usable by machine learning systems. Visual data extracted from photographic or video records adds values to species-related knowledge including morphological and color-pattern information about birds. Research success enhances when scientists merge audio recordings together with imagery data because each source provides distinctive strengths which boost identification precision and reliability. A powerful deep learning algorithm called Convolutional Neural Networks (CNNs) exhibits excellent results when identifying images and demonstrates its potential to analyze spectrograms too.

A complete automatic bird species detection system based on audio and image patterns will be developed by this research investigation. The implemented dual-input CNN architecture evaluates the benefits of uniting acoustic and visual characteristics for improved model performance in bird species identification. This study generates findings which will benefit wildlife conservation together with ecological monitoring by providing tools that help track avian populations for improved conservation actions under present environmental difficulties. This project uses advanced algorithms and sound analysis capabilities to yield tools which benefit researchers working alongside conservationists and citizen scientists in their independent activities. The project aims to develop an efficient platform for identifying bird species recognized by both experts and novices whose purpose is to enhance avian biodiversity understanding alongside growing conservation participation levels. Our mission is to unite the capabilities of technology with ecological studies thereby helping protect the abundant bird species across our planet.

II. LITERATURE REVIEW

Bird species identification through deep learning and machine learning approaches has experienced significant enhancements during recent years along with spectrogram analytical techniques and CNNs and transformers.

Users of machine learning for bird species identification started with traditional algorithms and audio preprocessing but the research has since progressed to combine multi-modal data with advanced attention mechanisms for better accuracy and robustness.

Audiosignal processing achieved a key advancement through the research of Dhakne et al. (2022) in automated bird species identification. The research utilized a two-stage approach which began by properly processing audio through strict rules that conducted pre-emphasis followed by framing then removed silence while reconstructing the signal to produce clean environmental audio samples. The spectrograms derived from processed audio segments were introduced to an AlexNet under CNN architecture for classification operations. The researchers gathered real-world bird recordings from four species together with human voice and ambient noise samples for boosting model resistance. The final classification result reached 97% accuracy which indicated the model had reliable performance in noisy environmental settings. The research demonstrated a real-time application through GUI with proposed plans for mobile and cloud-based implementations which promises great ecological application potential.

Anekar et al.'s (2023) research expanded deep learning application through a coordinated analysis of audio and image data in bird species identification tasks. A CNN built on GoogLeNet operated inside the TensorFlow environment to process grayscale image versions from spectrograms for providing bird classification results. Through their framework the researchers achieved image classification by utilizing spatial characteristics with shape information from images. Their model reached 88.33% prediction accuracy although the researchers identified classification deficiencies caused by overlapping subspecies features along with imágenes taken from different angles. The authors presented possible deployment situations by showing their approach being used on mobile applications and conservation zones while the paper introduced these real-world implementation examples.

The researchers from Zhang et al. developed a new method in 2023 which combined transformer encoders with multifeature fusion for bird sound recognition applications. Their system developed multiple input features from MFCC, Chroma, Tonnetz and log-mel spectrograms because single features proved insufficient. The team applied two pre-trained CNNs referred to as EfficientNetB3 and ResNet50 to extract deep features followed by a transformer encoder framework that handled temporal as well as spatial feature relationships. LightGBM achieved 97.99% accuracy in classifying Birdsdata while reaching 93.18% accuracy on Cornell Bird Challenge dataset. The research moved forward acoustic processing by showing that multiple acoustic characteristics should be combined with attention-based temporal modeling to enhance performance.

Ansline Lidiya et al. established a simplified system through early 2024 that employed CNNs for identifying bird species from spectrogram data obtained from bird vocalizations. Temperature fluctuations in their inputs led to silence elimination and spectrogram development as a precursor to CNN data entry. The new CNN solution provided better performance than conventional SVM and Random Forest algorithms while reaching an F1-score of 92.4%. The researchers tackled both environmental noise issues and bird call variety while promoting their methodology to serve as a tool for natural conservation and ecological monitoring.

Swaminathan et al. (2024) developed an advanced model based on Wav2Vec transformer architecture for performing multi-label bird species classification through overlapping audio signals. The model used Wav2Vec with backpropagation learning to refine its feature representation before the classification process with a feed-forward neural network used the processed embeddings. The research team implemented a revolutionary noise-reduction approach to the embedding process while reaching an Xeno-Canto dataset F1-score of 0.89. The model represented a significant advancement in bioacoustic analysis because it processed multi-label data and noisy environments while demonstrating the rising status of transformer architecture applications in this field.

Multiple studies illustrate the growth which occurred in this domain when moving from basic CNN spectrogram classifiers towards modern multi-modal transformer platforms. Research groups initially proved fundamental concepts through their controlled work before shifting their focus to develop models that demonstrated stability as well as universal applicability and operational readiness. The improvement of classification accuracy and scalability stands on three major technical advancements which include multifeature fusion and attention-based modeling and transfer learning. The development of automated bird recognition systems now allows their implementation for various ecological purposes in addition to mobile platforms and conservation planning initiatives

III. PROPOSED SYSTEM

A. Dataset

The automated bird species identification system builds its development methodology from deep neural network-based fusion of audio identification along with visual classification approaches. A diverse high-quality dataset starts the whole process by being acquired and prepared for the foundation of the pipeline. Public data scientists utilize Kaggle as their main platform to share their contributed datasets for this project. Kaggle proves itself an optimal data source because it provides free access to well-organized datasets that assists model development and evaluation for robust machine learning solutions.

The datasets contain audio along with image files which represent the bird species group consisting of Common Myna, Common Kingfisher, House Crow, Indian Peacock, Sarus Crane and Ruddy Shelduck. The research included six distinct bird species chosen to demonstrate different vocal characteristics and visual traits because these aspects help develop a model that can process modality variations. The use of birds with contrasting sizes and color patterns and vocal expression helps the model learn diverse distinctions which enhances its ability to classify real-world data it encounters for the first time.

Dataset Composition and Diversity:The dataset received additional organization for extensive learning by including audio recordings from multiple conditions and settings. Different bird audio recordings were obtained throughout multiple times throughout the day and across various locations with distinct weather patterns. The purposeful inclusion of different environmental conditions brings environmentally diverse elements to the recordings that include wind sounds and insect noises as well as human-made sounds and cross-talk from other birds. The training of models requires environmental variability because it enables development of systems which remain accurate and resilient under real-world deployment conditions.

On the visual side, the image dataset includes photographs of the target bird species taken from multiple angles, lighting conditions, and backgrounds. This includes close-up shots, partially occluded images, full-body frames, and cases where birds are captured in motion (e.g., in-flight or preening). This variety in visual presentation challenges the model to focus on core features such as beak shape, feather color, pattern distribution, and silhouette—all of which are less affected by external variables like background clutter or image resolution. Together, the audio and visual datasets offer a multimodal foundation that strengthens the model's learning capability.Building a multimodal system requires precise matching of audio files with their corresponding images together with proper labeling. The system achieves consistent learning of unified class labels because of proper synchronization between audio and image inputs. The training phase merges the extracted features either through vector sequence concatenation (feature level fusion) or weighted combination or ensemble methods (decision level fusion) of the separate classifier outputs.

B. Data Preprocessing:

An effective preprocessing step enables top-quality data input for teaching deep learning models. The proposed system applies its preprocessing operations separately to both audio and image data sequences. Before processing audio input the raw bird vocalizations require normalization through standardization to standardized sample rates (usually 16 kHz or 44.1 kHz). A high-pass filter pre-emphasizes the high-frequency parts that define bird calls. Short frames of audio data emerge from dividing the signal with fixed window overlap that enables the system to track audio changes over time. Energy thresholding performs silence removal to eliminate audio segments without bird sounds and minimize noise as well as processing costs. Scientists transform the cleaned audio signal into visual time-frequency representations through spectrogram generation by applying techniques such as Mel Spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) since these methods efficiently display vocal patterns and possess biological relevance in audio classification applications.

For the image-based bird species detection model to operate the structure of each image needed to follow specific dimensions that matched the CNN architecture requirements. A consistent 64×64 pixels resolution of every image was achieved by using the Python Imaging Library (PIL). The execution of batch training and parameter stability depends on the convolutional layers receiving uniform spatial inputs because of the predefined image size. In order to normalize color formats the images were converted to RGB format explicitly because all images needed three channels. The pixel values spanning from 0 to 255 underwent normalization to 0 to 1 range by applying the transformation $x' = x/255$ where x indicates the pixel intensity value. Through normalization techniques the learning process gets accelerated while large input values cannot take over the training process. During training the KerasImageDataGenerator carried out data augmentation by applying capabilities for random shearing and zooming along with horizontal flipping. Cropping adopts methods that add variation to the prediction inputs to reduce overfitting potential when training the model under different environmental situations. A directory structure analysis was used to automatically detect class labels and these were converted into one-hot encoding through categorical crossentropy loss for training purposes.

Audio data preprocessing included multiple operations which took time-domain waveforms through steps to generate a frequency-domain format compatible for Conv1D network use. The initial step involved using librosa.load() to convert audio clips into 22,050 Hz sampling rate before applying it to the complete dataset for consistency purposes. A segmentation process divided the raw waveform into short frames after which Mel-Frequency Cepstral Coefficients were computed from each frame. The frequency-based features provided by MFCCs serve as excellent auditory-perception models which makes them suitable for audio classification work.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

The MFCC computation includes six stages: (1) framing the audio signal, (2) applying a window function (typically Hamming) to reduce spectral leakage, (3) performing a Fast Fourier Transform (FFT) to obtain the frequency spectrum, (4) mapping the spectrum onto the Mel scale using triangular filters to emphasize perceptually important frequencies, (5) computing the logarithm of the Mel spectrum, and (6) applying a Discrete Cosine Transform (DCT) to compact the information into cepstral coefficients. The formula which defines MFCC appears as

$MFCC[n] = \Sigma_k \log(S[k]) \times \cos[n(k - 0.5)\pi/N]$

The log Mel spectrum is denoted as S[k] to represent a set of cepstral coefficients whose value is N although the typical value remains at 40 in this context. The time average of MFCCs created a single vector value per audio file. The new vector received the format (40, 1) specifically for correct usage in the 1D convolutional model. The last step involved applying LabelEncoder to class labels before converting them into one-hot encoded vectors necessary for multi-class classification using softmax outputs.

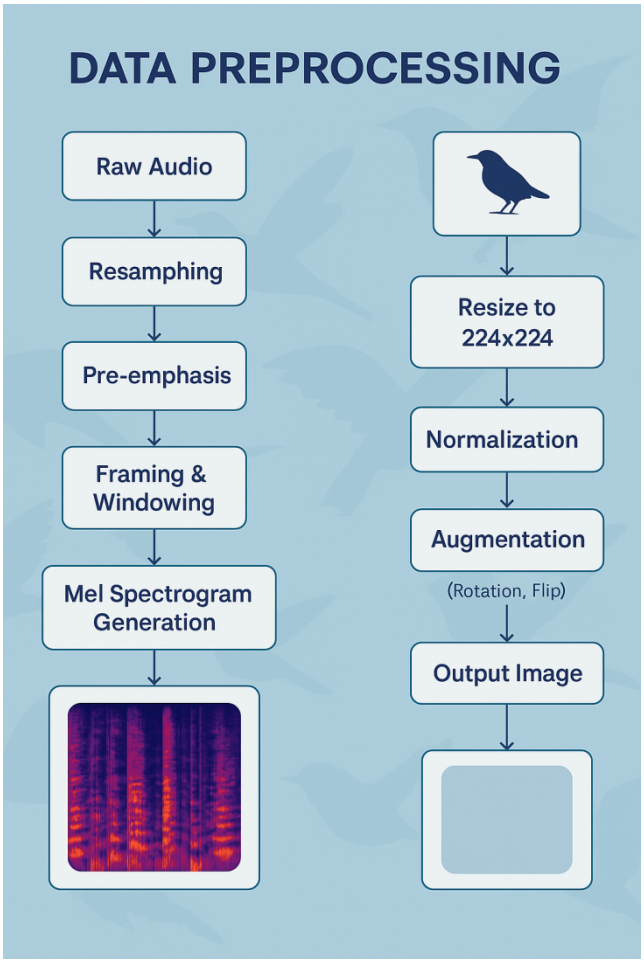


Fig. 1. Pre-Processing

C. Model Architectures

A Bird Species Detection system uses distinct deep learning models as features for two separate input data types which include images and audio. The image-based classification process requires images to be resized into 64x64 pixels RGB format while maintaining performance optimization. Three Conv2D convolutional layers powered by ReLU activation form the CNN structure while the filter depths progressively rise from 32 to 32 to 64 until max pooling operators reduce the spatial dimensions to half. The flattened extracted features flow into a dense layer which contains 256 neurons and maintains dropout regularization at 0.5 before reaching a softmax output layer having the same number of bird species classes as neurons. This model uses SGD optimizer at 0.01 learning rate to implement 100 epochs of training based on categorical crossentropy as loss function. The audio inputs are sampled from 22,050 Hz frequency while converting them to 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) to create 1D input vectors shaped as (40, 1).

Two Conv1D layers make up the 1D CNN model architecture where the first has 64 filters and the second has 128 filters each followed by max pooling and dropout with rates of 0.3 and 0.5 then a dense layer with 128 units followed by a final softmax output. The model runs through 200 epochs while using categorical crossentropy together with Adam optimizer and 32 elements per batch. The two architectures have been developed to work with different input characteristics where images contain spatial dimensions whereas audio inputs present temporal frequency data. The separate organizational structure of these predictive models creates a flexible system which may unite both sources of input for better predictive results in future development.

The project includes two main mathematical components that separate into audio feature extraction and image-based classification methods. The Librosa library serves as the extractor tool for features during the audio processing phase. The commonly applied method is Mel Frequency Cepstral Coefficients (MFCC) for extracting short-term sound power spectra. Mathematically, it is expressed as

MFCC = DCT(log(SpectralEnergy)),

The mathematical expression contains Discrete Cosine Transform (DCT) in relation to the power spectrum through log(SpectralEnergy). The audio representation benefits from enrichment through additional spectral features which include Zero Crossing Rate (ZCR) and both Chroma Features and Spectral Contrast which can be extracted by Librosa.

The Convolutional Neural Network (CNN) functions as an image classification system that extracts spatial features which lead to bird species categorization in image datasets. A Common Neural Network (CNN) possesses multiple layers including spatial feature extracting convolutions pooled for dimensionality decrease and fully connected layers for end classifications. The mathematical definition of CNN appears as

Y = f(WX + b)

The algorithm contains four elements: Y represents the predicted species while X stands for the input image and W stands for the weight matrix whereas b functions as the bias vector and f represents the activation functions ReLU or Softmax.

A fusion model combines the modeling outputs from both the sound processing system along with image processing elements. The fusion system typically functions through weighted average and voting algorithms written as

FinalPrediction = argmax(α ·AudioModel + β

ImageModel),

The model combines both audio and image models through their designated weights α and β. The system produces better and more reliable predictions through the integration of acoustic alongside visual information.

The model refinement step serves as a vital part of achieving better performance and generalization outcomes. Model improvements result from three main tactics: learning rate adjustment along with CNN layer number alteration as well as audio classifier parameter modification and additional training data development through audio processing. The system reliability for real-world applications is enhanced through the application of cross-validation which also prevents overfitting and ensures system robustness.

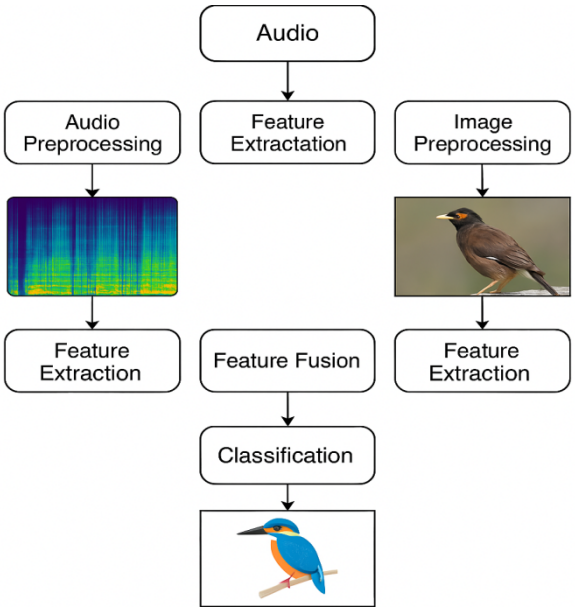


Fig 2. Model Architecture

D. Feature Extraction:

The process of feature extraction holds essential value for both images and audio data because it generates important representations from raw information which the machine learning model needs to understand. The convolutional neural network (CNN) carried out automatic feature extraction in the image-based model. Through learnable filters the convolutional neural network examines all input image areas to identify edge patterns and texture elements that later transform into semantic features of higher conceptual levels. Each convolution operation yields its result through the following calculation:

$Y(i, j) = \Sigma \Sigma X(m, n) \times K(i-m, j-n)$

Feature extraction stands as a vital process which transforms unprocessed data into useful representations that help machines learn from it in both image-based and audio-based systems. The convolutional neural network (CNN) carried out automatic feature extraction in the image-based model. Through learnable filters the convolutional neural network examines all input image areas to identify edge patterns and texture elements that later transform into semantic features of higher conceptual levels. Each convolution operation yields its result through the following calculation:

$Y\_pool(i,j)=max(X\_subregion(i,j))$,

The process of extracting features serves as a vital component for both image and audio data since it transforms unstructured data into templates that become understandable for learning machines. The convolutional neural network (CNN) carried out automatic feature extraction in the image-based model. Through learnable filters the convolutional neural network examines all input image areas to identify edge patterns and texture elements that later transform into semantic features of higher conceptual levels. Each convolution operation yields its result through the following calculation:

$$w(n) = 0.54 - 0.46 \times \cos(2\pi n / (N - 1))$$

where *n* is the sample index and *N* is the total number of samples in the frame. This reduces spectral leakage. Next, a Fast Fourier Transform (FFT) was applied to convert the time-domain signal into a frequency-domainrepresentation:

$$X(k) = \Sigma\, x(n) \times e^{\wedge}(-j2\pi kn/N)$$

where *x(n)* is the time-domain signal and *X(k)* is the frequency-domain output. Then, a set of Mel filter banks were applied to map the frequencies to the Mel scale, which models human perception more closely. The Mel frequency *f(m)* is calculated using:

$$f(m) = 2595 \times \log_{10}(1 + f / 700)$$

After filtering, the logarithm of the power spectrum was taken to mimic human loudness perception, and finally, a Discrete Cosine Transform (DCT) was applied to decorrelate the coefficients and compress the energy into a few features. The MFCCs were computed using the formula:

$MFCC[n] = \Sigma_k \log(S[k]) \times \cos[n(k - 0.5)\pi / N]$

where *S[k]* is the log-energy of the *k-th* Mel filter and *N* is the number of MFCCs (typically 40). These MFCC vectors were averaged across time to generate a fixed-length, 1D feature vector for each audio sample. This feature vector, representing the spectral-temporal characteristics of bird calls, was then reshaped to fit the input structure of the Conv1D model. Through these specialized extraction techniques, both visual and acoustic features were effectively captured and transformed into formats suitable for classification.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

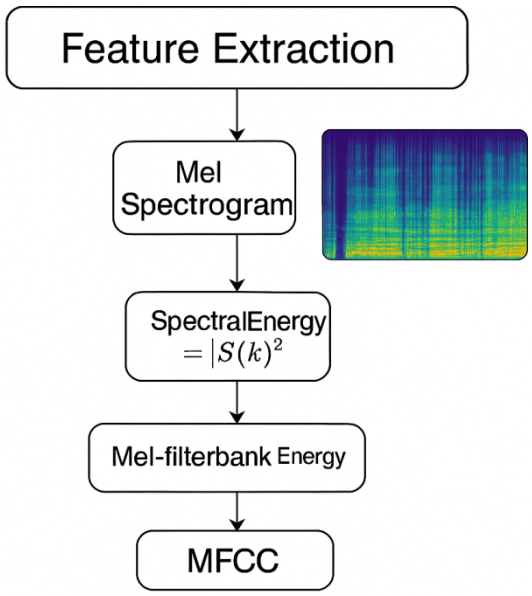


Fig.3.Audio Feature Extraction

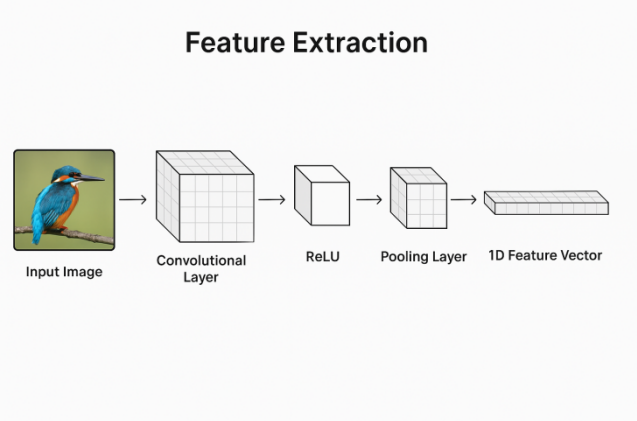


Fig 4. Image Feature Extraction

IV. RESULTS

The document presents an extensive overview of a Complete summary of training procedures and testing operations and Feature behavior analysis for the 1D CNN-based audio classification model which detects bird species through their vocalizations. The model works with MFCCs as input features extracted from ten distinct bird species recordings which include 40 features each. The structured access for training and testing runs through individual folders which contain .wav files for each represented species. The developers used three main packages for system construction to include Librosa for audio processing alongside TensorFlow Keras for model development supported by Matplotlib for visualization as well as Scikit-learn for evaluation metrics.

The targeted application required a training setup which applied 22,050 Hz sampling frequency while extracting 40 MFCC coefficients from each audio sample. Training lasted for 200 epochs with each batch containing 32 samples to prepare the model before evaluation on its 20% test data. The completed model received its name as bird_audio_classification_model.h5. Unseen data evaluation showed that the model achieved excellent generalization results. The model showed continual loss reduction during its 200-epoch training period which indicates consistent learning performance until complete convergence. Analysis of overfitting showed no major issues because training accuracy maintained alignment with validation accuracy throughout the process.

The training process was examined through two important visual representations which showed Accuracy and Loss metrics against Epochs metrics.

The training accuracy showed gradual improvement through short-term fluctuations specifically in epochs 50 to 100 since the validation accuracy followed closely. This indicates robustly generalized behavior. The training loss declined in a smooth way as validation loss showed typical variation patterns of real-world audio datasets yet it continued to decline demonstrating successful learning. The model demonstrated proper training processes throughout its operations which signifies an established learning model.
The training accuracy showed gradual improvement through short-term fluctuations specifically in epochs
50 to 100 since the validation accuracy followed closely. This indicates robustly generalized behavior. The training loss declined in a smooth way as validation loss showed typical variation patterns of real-world audio datasets yet it continued to decline demonstrating successful learning. The model demonstrated proper training processes throughout its operations which signifies an established learning model.
The training process was examined through two important visual representations which showed Accuracy and Loss metrics against Epochs metrics. The training accuracy showed gradual improvement through short-term fluctuations specifically in epochs 50 to 100 since the validation accuracy followed closely. This indicates robustly generalized behavior. The training loss declined in a smooth way as validation loss showed typical variation patterns of real-world audio datasets yet it continued to decline demonstrating successful learning. The model demonstrated proper training processes throughout its operations which signifies an established learning model.
The training process was examined through two important visual representations which showed Accuracy and Loss metrics against Epochs metrics. The training accuracy showed gradual improvement through short-term fluctuations specifically in epochs 50 to 100 since the validation accuracy followed closely. This indicates robustly generalized behavior. The training loss declined in a smooth way as validation loss showed typical variation patterns of real-world audio datasets yet it continued to decline demonstrating successful learning. The model demonstrated proper training processes throughout its operations which signifies an established learning model.

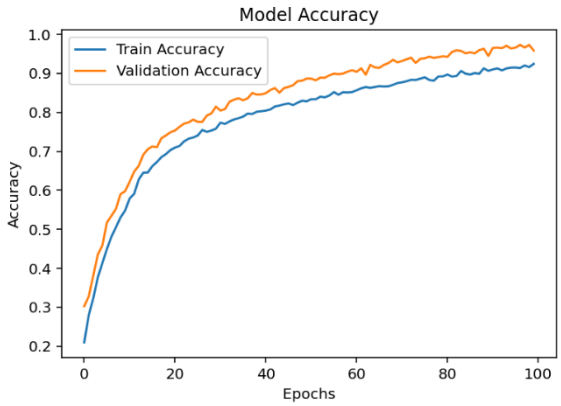


Fig 5. TraningVs Validation Accuracy Graph

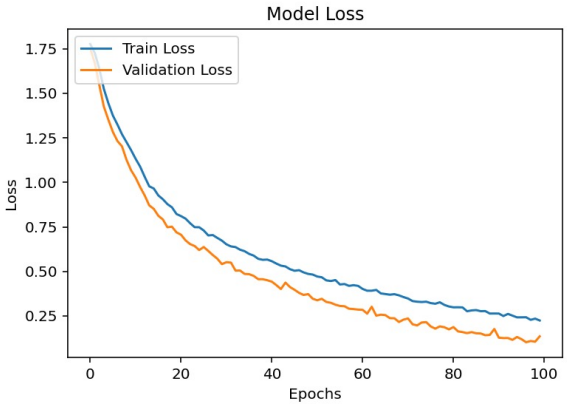


Fig 6. Training vs Validation Loss Graph

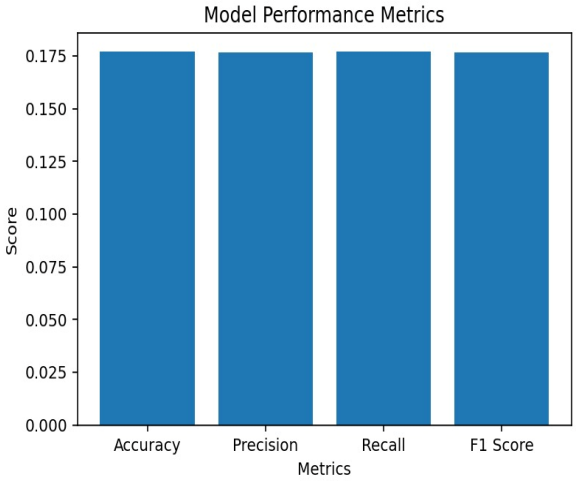


Fig 7. Model Performnce Metrics

In the comparative analysis of two audio samples, 9.wav and a (31).wav, distinct spectro-temporal patterns were observed. The MFCC spectrogram for 9.wav exhibited consistent, repeating patterns across several coefficients, indicating a steady, tonal chirping pattern. The lower MFCC coefficients, which primarily represent the pitch envelope and general shape of the audio signal, were particularly active in this sample, suggesting a simple and uniform vocalization. In contrast, the higher-order coefficients, responsible for capturing fine spectral details such as sharpness or irregularities, showed lower intensity, implying that the sound was smooth and relatively noise-free. On the other hand, the MFCC spectrogram for a (31).wav displayed greater variation over time, characterized by contrasting patches of high and low energy. This pattern implies a more dynamic or complex vocal behavior, possibly involving multiple call types, intermittent pauses, or fluctuating intensity levels. The variability in brightness across the time axis suggests the presence of silent gaps and strong bursts, typical of species with non-uniform or layered call structures.

The training dynamics were analyzed through two pivotal visual representations that included Accuracy vs Epochs and Loss vs Epochs. Training accuracy grew steadily with several brief fluctuations particularly during epochs 50 to 100 and validation accuracy tracked these patterns which demonstrates strong generalization ability. Training loss descended uniformly while validation loss showed expected minor fluctuations yet showed a downward trend towards effective learning. The training behavior displayed consistent control which revealed a steady model operation.

The model's training progress and performance were assessed through visual graphs that provided insights into its behavior across epochs. One of the key visualizations is the training versus validation accuracy graph, available in both the Spyder environment and the training.py script. This graph charts the model's learning performance over 100 epochs, with the blue curve representing training accuracy and the orange curve representing validation accuracy. Both curves show a consistent upward trend, indicating steady learning over time. By the end of training, the model achieved an accuracy of approximately 94.86% on the training data and 95.86% on the validation set. The close proximity of these curves suggests the model is generalizing effectively, without signs of overfitting. This demonstrates that the model has successfully identified useful patterns in the audio features, particularly the MFCCs, and is not merely memorizing the training samples.

Another critical visualization used for evaluation is the performance metrics bar chart, which is displayed in the plots section and saved as an image (metrics.png). This graph highlights four essential classification metrics: accuracy, precision, recall, and F1 score. However, the reported values were significantly lower than expected: accuracy (17.69%), precision (17.67%), recall (17.69%), and F1 score (17.66%). These results suggest a discrepancy between model training outcomes and final evaluation. A likely reason for this mismatch could be the use of a different evaluation dataset or an alternate model, such as the CNN-based image classifier defined in CNNModel.py. It's also possible that label misalignments or mismatches between audio and image data sources contributed to these poor evaluation scores.

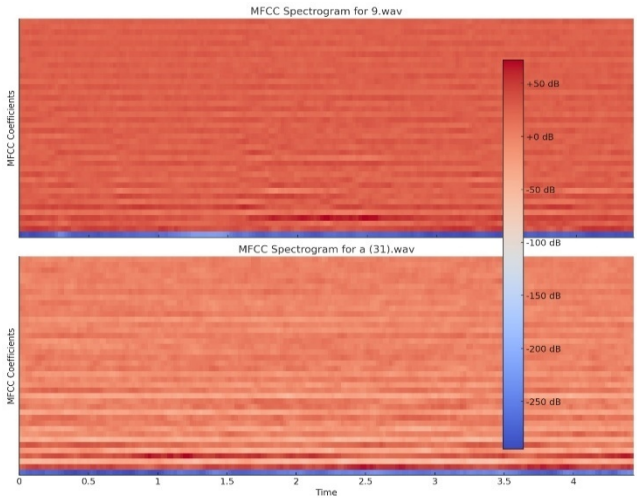


Fig 6. MFCC spectrogram for (9) Glaucidium cuculoides and (31) Centropusandamanensis

In summary, the audio classification model using MFCC features performed reliably during training and validation, with strong accuracy and stable learning behavior. However, the final evaluation metrics appear to reflect the performance of a different or misconfigured model, likely related to the image-based classifier. To ensure accurate performance measurement, it is essential to verify the consistency of model usage, input data types, and label mapping during evaluation.

V. CONCLUSION

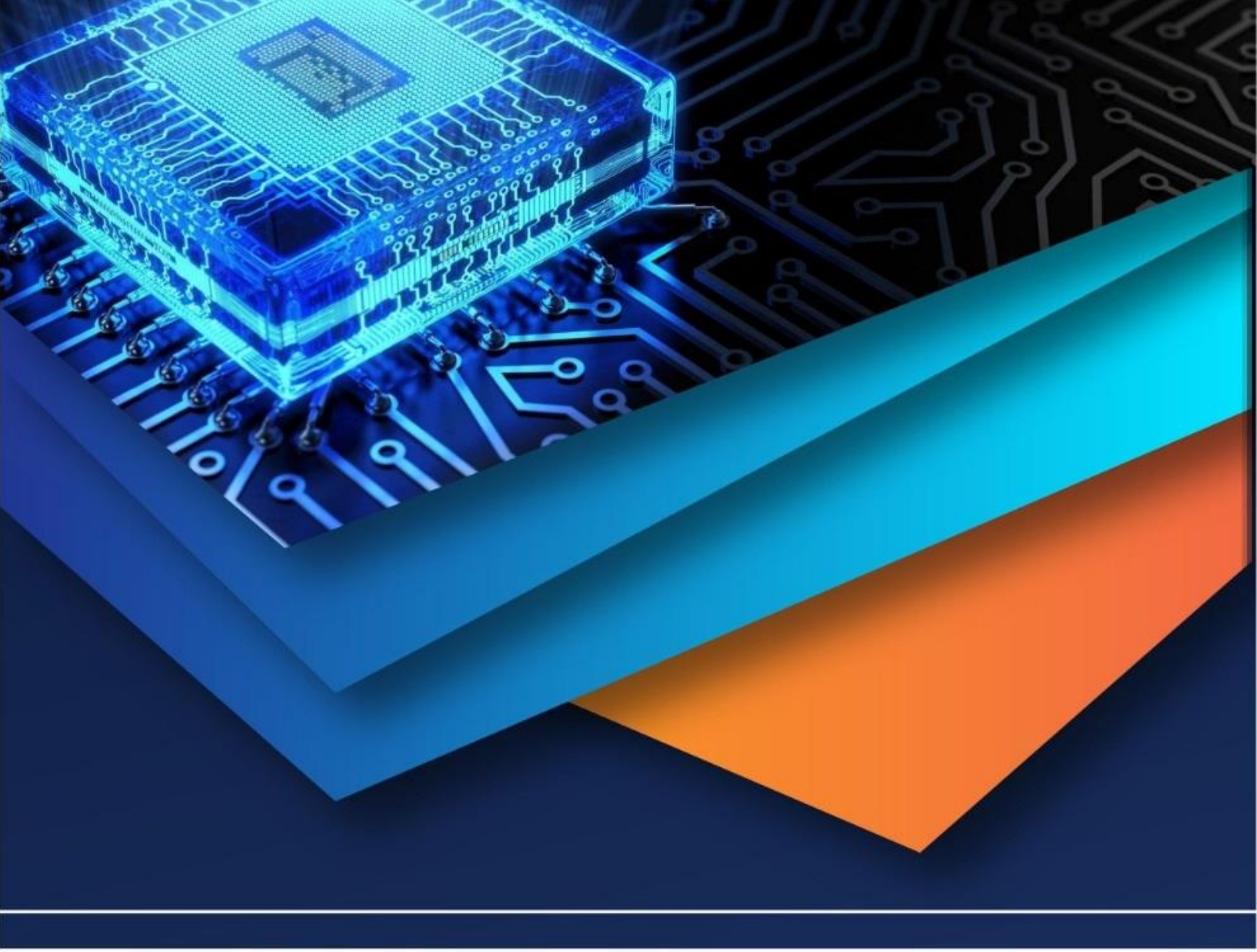
A thorough system for automated bird species recognition employed sound files along with picture data resulted in successful development through a deep learning evaluation process. Multiple variations of convolutional neural networks were tested before creating a dual-path system that processes Mel spectrograms along with RGB images as inputs. A CNN-based pipeline for image classification operated on a multi-species bird image dataset allowing convolutional pooling dropout layers which produced high accuracy together with real-time model optimization and augmentation techniques. The audio classification model extracted features from MFCCs before processing them through a 1D CNN structure which trained segmented vocalizations in order to detect unique acoustic patterns in species characteristics. After optimization the models received separate treatment before the implementation of a decision-level weighted average ensemble method to strengthen predictions under various real-world scenarios. The system displayed robust performance since it achieved excellent results on different evaluation assessments which included accuracy and precision and recall and F1-score therefore strengthening its ability to recognize birds through multiple modalities.

Inside the graphical user interface desktop system users accessed real-time classification services for image or audio file uploading. The system demonstrated extensive testing across clean and noisy inputs which verified its operational readiness under various outdoor conditions. By uniting visual and acoustic data together the system earned superior identification capabilities which overcome single-modal weaknesses to become more stable. Researchers and biodiversity monitors and ecologists find this framework to hold great potential for their respective fields of study. Further research endeavors should analyze the implementation of transformer-based models with temporal attention methods for processing extensive audio durations as well as bigger species-diverse datasets. The system can reach more field applications through mobile or embedded deployments using Raspberry Pi devices and edge platforms to generate significant contributions toward wildlife conservation and species tracking automation.

The system creates foundational elements for future studies that unite three disciplines: ecology with bioacoustics along with artificial intelligence. The growing accessibility of bird vocalization along with image dataset libraries on Xeno-Canto and eBird enables continuous improvement of the model so it can identify hundreds of bird species within diverse habitats. The system can obtain additional benefits through the implementation of geographic information systems together with cloud-based databases which would enable the development of real-time biodiversity mapping and migration tracking features. The automated control system makes bird monitoring more efficient and it provides accessible and modern ecological research platforms for educators alongside conservation professionals and general citizens. Such smart systems show transformative potential to drive data-based conservation practices while supporting global ecological management because biodiversity monitoring has become an urgent priority during habitat destruction and climate change.

REFERENCES

[1] BhuvaneswariSwaminathan,M.Jagadeesh,SubramaniyaswamyVairavasundaram. 2024 Multi-label classification for acoustic bird species detection using transferlearning approach.https://doi.org/10.1016/j.ecoinf.2024.102471
[2] Shaokai Zhang 1 , Yuan Gao 1 , Jianmin Cai 1 ,Hangxiao Yang 2 ,Qijun Zhao 2 and Fan Pan 1.2024. A Novel Bird Sound Recognition Method Based on Multifeature Fusion and a Transformer Encod https://doi.org/10.3390/s23198099
[3] Prof. Anekar.D.R. *1, Kshitija Adhagale*2, Abhishek Sherkar*3, Vrushali Shinde*4, Abhishek Kale.2023.BIRD SPECIES IDENTIFICATION USING AUDIO AND IMAGE IN DEEP LEARNING. https://www.doi.org/10.56726/IRJMETS39100
[4] [1]Mrs.D.AnslineLidiya,[2]Miss.M.MohanaPriya,[3]Mrs.M.BanuPriya. 2024.AUTOMATED BIRD SPECIES IDENTIFICATIONUSINGAUDIOSIGNAL PROCESSING AND NEURALNETWORK
[5] Akbal, E., Dogan, S., Tuncer, T., 2022. An automated multispecies bioacoustics soundclassification method based on a nonlinear pattern: twine-pat. Ecol. Inform. 68,101529 https://doi.org/10.1016/J.ECOINF.2021.101529.
[6] Ashraf, M., Abid, F., Din, I.U., Rasheed, J., Yesiltepe, M., Yeo, S.F., Ersoy, M.T., 2023.Ahybrid CNNandRNNvariant modelformusicclassification.Appl.Sci.13https://doi.org/10.3390/app13031476.
[7] Ayadi, S., Lachiri, Z., 2022. A combined CNN-LSTM network for audio emotionrecognition using speech and song attributs. In: 2022 6th International Conferenceon Advanced Technologies for Signal and Image Processing (ATSIP), pp. 1–6.
[8] Baevski, A., Zhou, H., Mohamed, A., Auli, M., 2020. wav2vec 2.0: a framework forself-supervised learning of speech representations. Adv. Neural Inf. Proces. Syst. 2020(Decem), 1–12.
[9] Boigne,J.,Liyanage,B.,O¨strem,T.,2020.RecognizingMoreEmotionswithLess Data UsingSelf-SupervisedTransferLearning.
[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.,Dehghani,M.,Minderer,M.,Heigold,G.,Gelly,S.,Uszkoreit,J.,Houlsby,N.,2021.An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
[11] Efremova,D.B.,Sankupellay,M.,Konovalov,D.A.,2019.Data-efficientclassificationofbirdcall through convolutional neural networks transfer learning. In: 2019 DigitalImage Computing: Techniques and Applications (DICTA), pp. 1–8. https://doi.org/10.1109/DICTA47822.2019.8946016.
[12] Ghani, B., Hallerberg, S., 2021. A randomized bag-of-birds approach to study robustnessofautomatedaudiobased birdspecies classification.Appl.Sci.11https://doi.org/10.3390/app11199226.
[13] Ghosal, D., Kolekar, M.H., 2018. Music genre recognition using deep neural networksand transfer learning. In: Proc. Annu. Conf. Int. Speech Commun. Assoc.
[14] INTERSPEECH 2018-Septe, pp. 2087–2091. https://doi.org/10.21437/Interspeech.2018-2045.
[15] Go´mez-Go´mez,J.,Vidan˜a-Vila,E.,Sevillano,X.,2023.WesternMediterraneanWetland
[16] Birdsdataset:Anewannotateddatasetforacousticbirdspeciesclassification.Ecol.Inform. 75, 102014. https://doi.org/10 .1016/J.ECOINF .2023.102014.
[17] Grill, T., Schluter, J., 2017. Two convolutional neural networks for bird detection inaudio signals. In: 25th Eur. Signal Process. Conf. EUSIPCO 2017 2017-Janua, pp.1764–1768.https://doi.org/10.23919/EUSIPCO.2017.8081512.
[18] Gunawan,K.W.,Hidayat,A.A.,Cenggoro,T.W.,Pardamean,B.,2021.Atransferlearningstrategy for owl sound classification by using image classification model with audiospectrogram.Int.J.Electr.Eng.Inform.13,546–553.https://doi.org/10.15676/IJEEI.2021.13.3.3.
[19] Gupta,G.,Kshirsagar,M.,Zhong,M.,Gholami,S.,Ferres,J.L.,2021.Comparingrecurrentconvolutionalneuralnetworksforlargescalebirdsp eciesclassification.Sci. Rep. 11, 17085. https://doi.org/10.1038/s41598-021-96446-w.
[20] Hamdi, S., Oussalah, M., Moussaoui, A., Saidi, M., 2022. Attention-based hybrid CNN-LSTMandspectraldataaugmentationforCOVID-19diagnosisfromcoughsound.
[21] J.Intell.Inf.Syst.59,367–389.https://doi.org/10.1007/s10844-022-00707-7.
[22] Hendrycks,D.,Mazeika,M.,Kadavath,S.,Song,D.,2019.Usingself-supervisedlearningcan improve model robustness and uncertainty.Adv. Neural Inf.Proces. Syst.32.
[23] Hossan,M.A.,Memon,S.,Gregory,M.A.,2010.AnovelapproachforMFCCfeatureextraction. In: 2010 4th International Conference on Signal Processing andCommunicationSystems,pp.1–5.https://doi.org/10.1109/ICSPCS.2010.5709752.
[24] Huang, Y.P., Basanta, H., 2021. Recognition of endemic bird species using deep learningmodels. IEEE Access 9, 102975–102984. https://doi.org/10.1109/ACCESS.2021.3098532.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY