iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.82931

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 14 Issue V May 2026- Available at www.ijraset.com

Bird Species Detection from Audio Signals Using Transfer Learning

Trishika K[1], Divyashree R[2]
Dept. of Data Science, AMC Engineering College, Bangalore, Karnataka, India

Abstract: ***Automatic identification of bird species from audio recordings is an important task in ecological research and biodiversity monitoring. This study proposes a deep learning-based framework that analyzes bird sounds using signal processing and transfer learning techniques. Audio signals are first transformed into frequency-based representations such as Fast Fourier Transform (FFT) and spectrograms. The use of pre-trained networks enhances learning efficiency and improves classification performance. A comparative evaluation between FFT features and spectrogram inputs reveals that spectrogram-based representations capture richer acoustic patterns, leading to better accuracy. The proposed system demonstrates reliable performance and can be effectively used in real-time environmental monitoring applications.***
Keywords: Bird Sound Classification, Deep Learning, Transfer Learning, FFT, Spectrogram, CNN, Bioacoustics.

I. INTRODUCTION

Bird sounds carry important information about species identity, behavior, and environmental conditions. Studying these sounds helps researchers monitor ecosystems and track biodiversity changes. However, manual analysis of bird audio files is time-consuming and requires specialized knowledge.

Recent developments in artificial intelligence have enabled automated analysis of such audio data. They have shown strong performance in recognizing patterns in both images and audio signals. When audio signals are converted into visual formats like spectrograms, they can be effectively processed by CNNs.

This work focuses on building an automated system that identifies bird species from audio recordings. The approach combines signal processing methods with transfer learning to achieve efficient and accurate classification.

The organization of the paper:

1) Section II provides the literature survey.
2) Section III gives a detailed implementation.
3) Section IV shows the result.
4) Section V concludes the paper.

II. SCOPE OF THE PROJECT

The scope of this project focuses on developing an intelligent system that can discriminate between bird species from audio recordings using signal processing and deep learning techniques. The system is intended to analyze the bird vocalizations by transforming raw audio signals into meaningful representations, such as FFT and spectrograms, followed by classification using a transfer learning-based CNN model.

This project covers multiple dimensions:

1) Audio-Based Classification: The system works entirely on sound data, making it useful in situations where visual identification is not possible, such as dense forests or low-light environments.
2) Feature Representation Analysis: It explores different audio feature extraction techniques, particularly FFT and spectrograms, and evaluates their effectiveness in bird species classification.
3) Application of Transfer Learning: The project utilizes pre-trained deep learning models, reducing the need for large datasets and computational resources while improving accuracy.
4) Scalability: The model can be extended to classify a larger number of bird species by training on expanded datasets.
5) Real-World Deployment Potential:
   - Mobile applications for bird identification
   - Smart environmental monitoring systems
   - IoT-based wildlife tracking devices

6) Research and Conservation Use:
It can assist researchers and environmentalists in studying bird populations, migration patterns, and ecosystem health.

III. RELATED WORK

Bird species recognition using audio recordings has become an important research area in machine learning and deep learning. With the growth of bioacoustic studies, different techniques have been developed to improve the accuracy and speed of classification systems.

Earlier methods mainly depended on traditional machine learning algorithms. These systems used manually extracted features such as MFCC, spectral centroid, and zero-crossing rate. Although these features were useful, the process required expert knowledge and often gave lower performance when recordings contained environmental noise or other disturbances.

In recent years, deep learning approaches have gained more attention, especially Convolutional Neural Networks (CNNs). These models can automatically learn useful patterns from the data without the need for manual feature extraction. Many studies use spectrogram images generated from bird sounds as input to CNN models, which has resulted in better classification accuracy.

Transfer learning has also improved the performance of bird sound recognition systems. Popular pre-trained models such as ResNet, VGG, and MobileNet can be adapted to bird audio datasets through fine-tuning. This helps reduce training time and provides better results when only a limited amount of labeled data is available. Research comparing different sound representations has shown that spectrograms are more effective than basic frequency-based methods like Fast Fourier Transform (FFT). Spectrograms provide both time and frequency information, making them suitable for capturing detailed bird call patterns. Because of this advantage, they are widely used in modern bird classification systems. Even though significant progress has been made, some challenges still remain. Background noise, overlapping calls from multiple birds, and imbalance in dataset classes can reduce model performance. Current research focuses on solving these problems using data augmentation, noise filtering methods, and advanced neural network architectures.

IV. METHODOLOGY

Tools and Technologies used:

A. Programming Language: Python

Python is selected for developing this project because it is easy to understand and provides a wide range of libraries for machine learning, deep learning, and audio signal analysis. It also allows quick implementation and testing of models.

B. Development Environment: Google Colab

Google Colab is used as the development environment for this project. It provides a cloud-based platform where Python code can be executed through notebooks without installing software locally. It is useful for testing code, analyzing data, and training models with free GPU and TPU support.

C. Audio Processing Libraries

- Librosa

Librosa is used for working with audio recordings. It helps in loading sound files, cleaning and preprocessing signals, extracting useful features, and generating spectrogram images for analysis..

D. Data Visualization

- Matplotlib

Matplotlib is used to create graphical representations of data. It is helpful for displaying waveforms, spectrograms, accuracy curves, and loss graphs during training.

E. Deep Learning Frameworks

- TensorFlow / Keras

TensorFlow and Keras are utilized for constructing and training deep learning models. These tools provide support for designing CNN architectures and implementing transfer learning methods.

F. Machine Learning Techniques

- Convolutional Neural Networks (CNNs)

CNN models are employed to identify hidden patterns from spectrogram images and classify bird species effectively.

- Transfer Learning

Transfer learning is used by adapting pre-trained models to the bird audio dataset. This approach reduces training time and can improve classification performance.

G. Feature Extraction Techniques

- Fast Fourier Transform (FFT)

FFT is used to transform audio signals from the time domain into the frequency domain, which helps in analyzing sound characteristics.

- Spectrogram Analysis

Spectrograms provide a combined time and frequency representation of audio signals. These images are used as input data for the classification model.

H. Dataset

- Bird Audio Dataset (Example: BirdCLEF)

A bird sound dataset containing recordings of multiple species is used for training, validating, and testing the proposed model.

I. Hardware Requirements

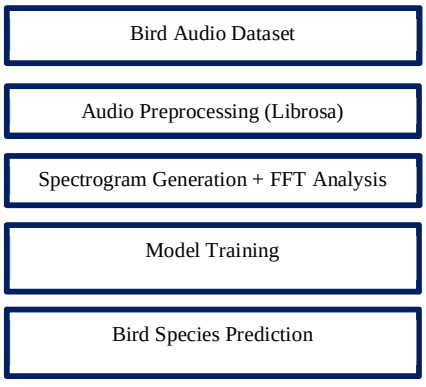
A system with internet access is required to use Google Colab.
GPU support available in Google Colab can be used for faster model training.
Adequate RAM is necessary for storing datasets and processing audio files efficiently.

Methodology Table

Stage	Component	Description
1	Audio Input	Collection of raw bird sound recordings from the dataset
2	Preprocessing	Noise reduction, normalization, and sampling rate standardization
3	Feature Extraction (FFT)	Conversion of a time-domain signal into a frequency-domain representation
4	Feature Extraction (Spectrogram)	Generation of time-frequency visual representation of audio
5	Data Preparation	Conversion of features into model-compatible format (images/arrays)
6	Model Selection	Selection of a pre-trained CNN model for transfer learning
7	Model Training	Fine-tuning the model using the training dataset
8	Validation	Evaluation using a validation dataset to monitor performance
9	Classification	Prediction of bird species based on learned features
10	Output	Final identified bird species label

System Architecture:

Bird Species Detection from Audio Signals Using Transfer Learning



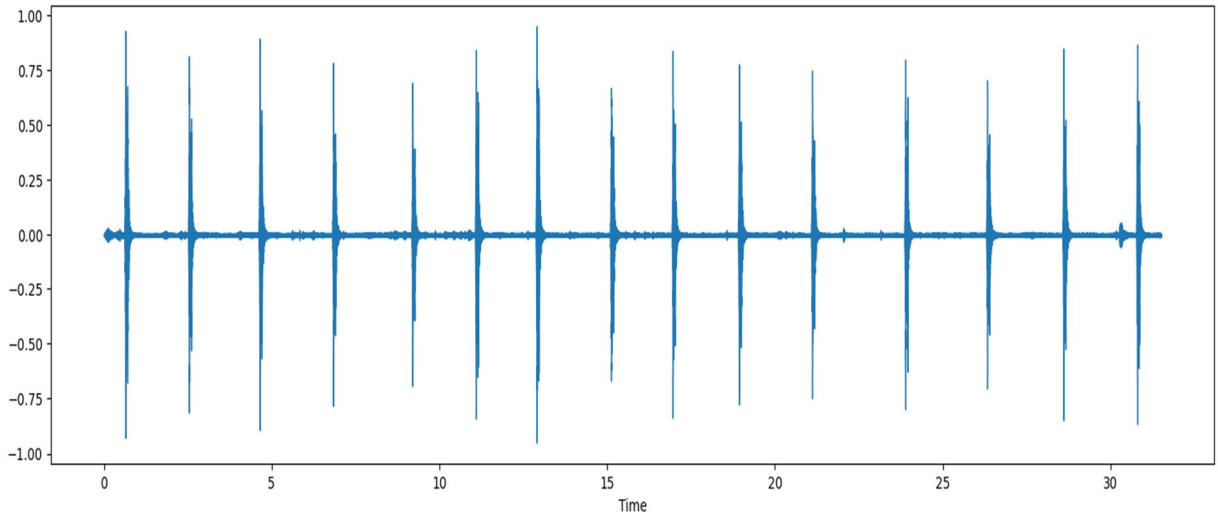


Fig.1: Audio Waveform Representation of Bird Sound Signal

The figure shows the time-domain waveform of a bird audio recording, illustrating amplitude variations over time. It highlights the periodic patterns and intensity peaks corresponding to bird vocalizations.

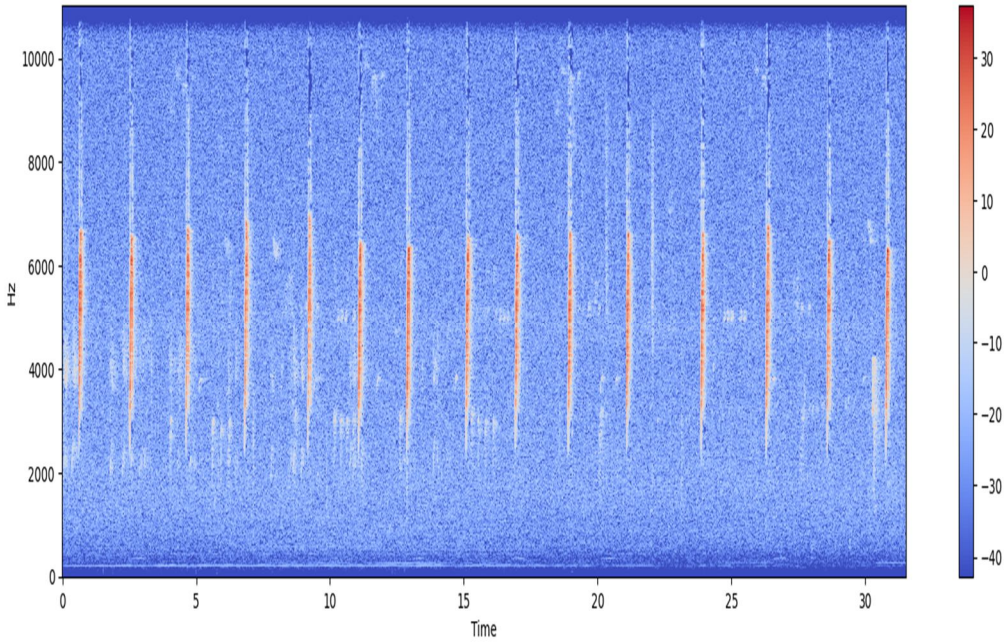


Fig.2: Spectrogram Analysis of Rhythmic Acoustic Signals

The figure displays the spectrogram, which displays a series of high-intensity, periodic sound pulses characterized by vertical bands spanning frequencies from approximately $2,000\text{ Hz}$ to $7,000\text{ Hz}$. The consistent temporal spacing and narrowband frequency peaks, indicated by the red heat map, suggest a repetitive biological or mechanical signal recorded over a 30-second duration.

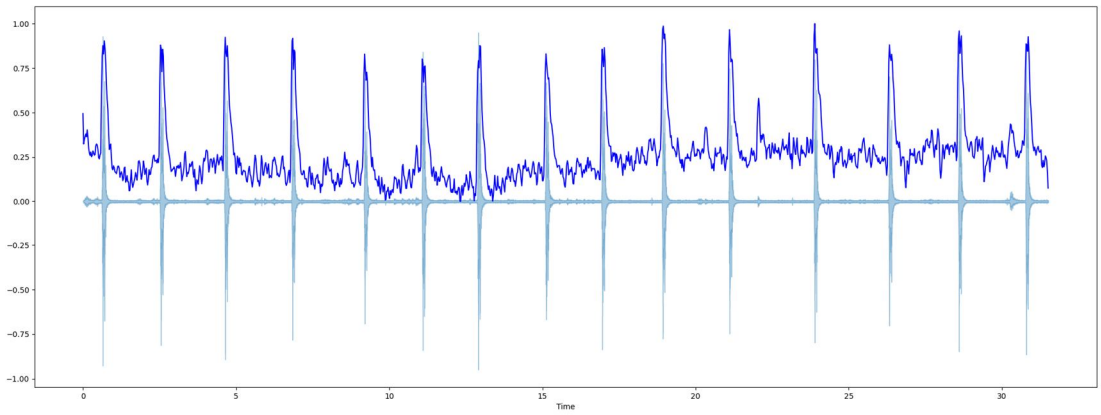


Fig.3: Spectral Centroid - The spectral centroid is a measure used in digital signal processing to characterise a spectrum. It indicates where the center of mass of the spectrum is located.

This visualization illustrates the raw acoustic pressure waves (light blue) aligned with a corresponding amplitude envelope (dark blue) over a 30-second interval. The plot highlights the precise temporal synchronization between the peak signal energy and the rhythmic pulses, emphasizing the impulsive nature of the sound source against the ambient noise floor.

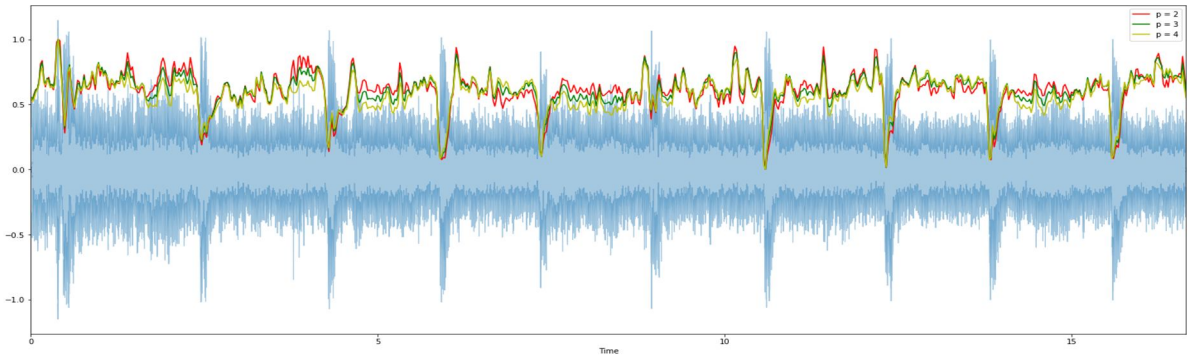


Fig.4: Spectral bandwidth.

This figure displays a time-domain acoustic waveform (light blue) overlaid with three spectral flux curves calculated using different p-norm values ($p=2, 3, 4$). The sharp, downward-pointing peaks in the spectral flux lines correlate precisely with the impulsive transients in the audio signal, demonstrating how varying the parameter p affects the sensitivity and smoothing of onset detection.

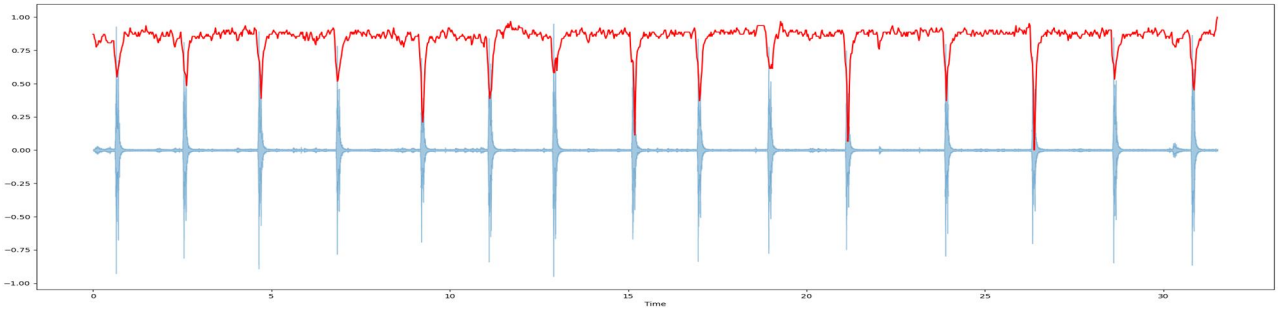


Fig.5: Spectral Rolloff

This plot, which superimposes a spectral flatness measure (red line) onto the time-domain waveform (light blue) characterizes the tonality of the signal. The sharp dips in the red line coincide perfectly with the acoustic pulses, indicating a transition from broadband noise to a more structured, tonal signal during each event.

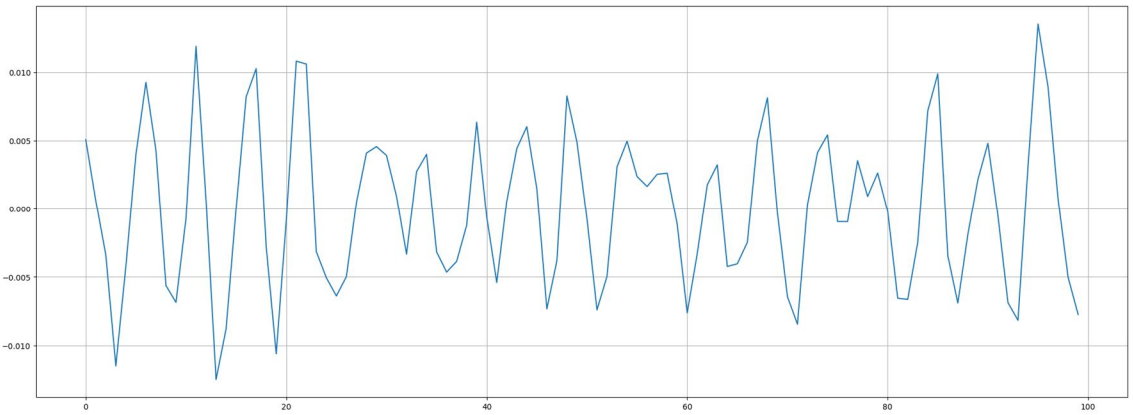


Fig.6: Zero-Crossing Rate - The zero-crossing rate (ZCR) is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive.

This line chart represents the instantaneous amplitude of a signal segment over 100 discrete time samples, revealing a complex, oscillating waveform with a maximum magnitude of approximately ± 0.0125. The jagged peaks and inconsistent periodicity suggest a transient or non-stationary signal, possibly representing a specific grain or micro-texture within a larger recording.

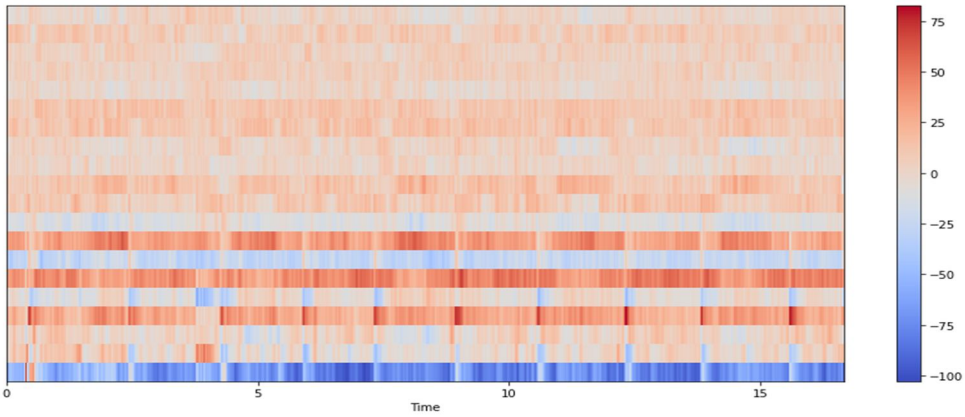


Fig.7: Mel-Frequency Cepstral Coefficients (MFCCs)

This visualization represents the spectral characteristics of an audio signal by plotting MFCCs over a 15-second duration, with the vertical axis corresponding to various coefficient bins. The color gradient—ranging from deep blue to dark red—indicates the magnitude of each coefficient, revealing horizontal bands of high energy that characterize the underlying timbral texture of the sound.

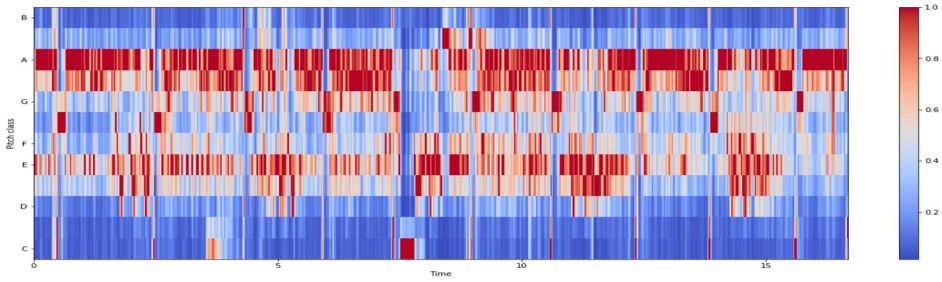


Fig.8: Chrome features.

This heatmap represents the intensity of the twelve musical pitch classes over a 17-second period, with the vertical axis mapping notes from C to B. The distribution of red regions highlights dominant harmonic content—particularly around the pitch classes A, E, and G—providing a clear visual profile of the melodic or chordal structure within the recording.

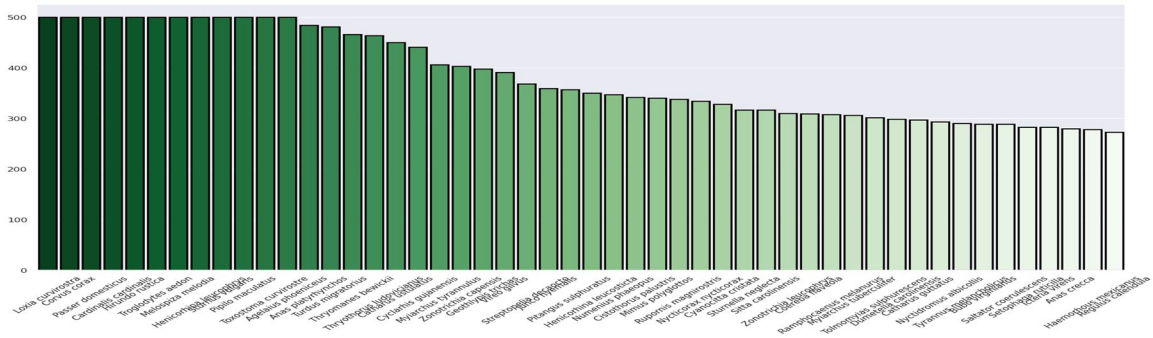


Fig.9: Frequency Distribution of Avian Species Observations

This bar chart illustrates the relative abundance of various bird species, with the most frequent species—including *Loxia curvirostra* and *Corvus corax*—reaching a capped frequency of 500 occurrences. The data follows a gradual long-tail distribution, visualized through a green-to-white color gradient that tracks the decline in observations across dozens of species.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 24, 128)	512128
batch_normalization (BatchNo	(None, 24, 128)	512
max_pooling1d (MaxPooling1D)	(None, 6, 128)	0
conv1d_1 (Conv1D)	(None, 4, 64)	24640
batch_normalization_1 (Batch	(None, 4, 64)	256
flatten (Flatten)	(None, 256)	0
dense (Dense)	(None, 256)	65792
dense_1 (Dense)	(None, 49)	12593

Total params: 615,921
Trainable params: 615,537
Non-trainable params: 384

Epoch 1/2
101/101 [==============================] - 1513s 15s/step - loss: 0.1887 - binary_accuracy: 0.9220 - val_loss: 0.0527 - val_binary_accuracy: 0.9846
Epoch 2/2
101/101 [==============================] - 1501s 15s/step - loss: 0.0402 - binary_accuracy: 0.9882 - val_loss: 0.0453 - val_binary_accuracy: 0.9848

Fig.10: Model 1 - FFT-based 1D CNN Neural Network

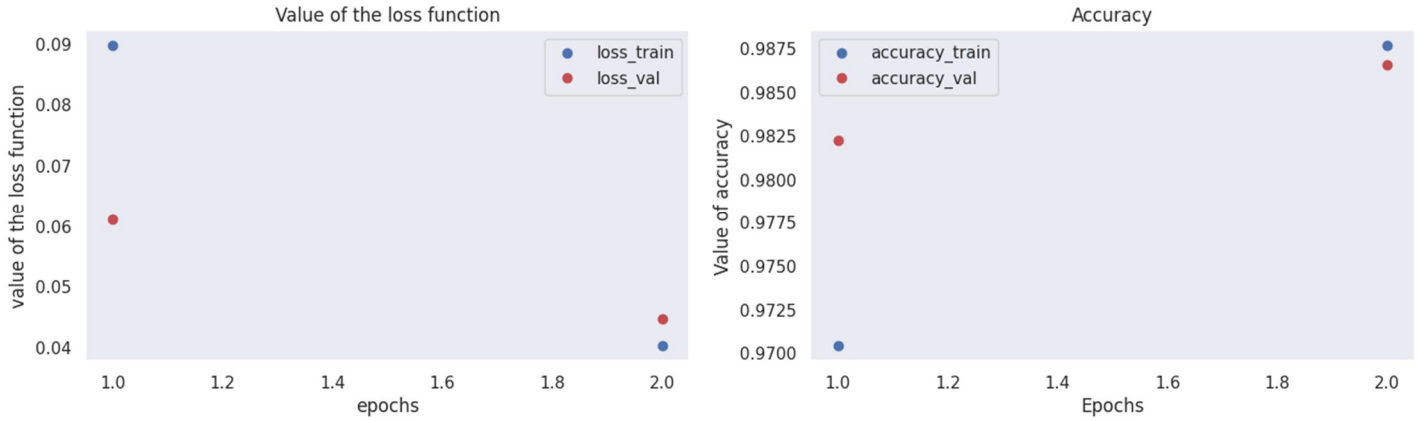


Fig.11: Training and Validation Performance Over Initial Epochs

This dual-pane plot tracks the convergence of a machine learning model, showing a sharp decrease in loss and a corresponding increase in accuracy between the first and second epochs. The close proximity of the training (blue) and validation (red) metrics suggests stable learning and minimal overfitting during the early stages of the training process.

```
STEP 1) CREATING A SUBSET OF DATASET:

[DATASET]: (62874, 14) : LABELS 397

RATING LIMITER APPLIED:
[SUBSET]: (38226, 14) : LABELS 397

 200+ RECORDINGS ONLY BIRDS LIMITED:
[SUBSET]: (8548, 14) : LABELS 27

 BIRD LABELS AVAILABLE AFTER FILTER:
amerob    comrav    mallar3   rucspa1
barswa    comyel    norcar    sonspa
bewwre    eursta    normoc    spotow
blujay    gbwwre1   redcro    wbwwre1
bncfly    grekis    rewbla    wesmea
carwre    houspa    roahaw    yeofly1
compau    houwre    rubpep1

LIMITING AUDIO FILES ...
[SUBSET]: (1500, 14) : LABELS 27
```

```
SUCCESSFULLY EXTRACTED 4157 SPECTROGRAMS
```

```
Found 3335 images belonging to 27 classes.
Found 822 images belonging to 27 classes.
```

Fig.12: Model 2 - Spectrums based 2D CNN

```
Downloading data from https://storage.googleapis.com/tensorflow/keras-
applications/resnet/resnet101_weights_tf_dim_ordering_tf_kernels_notop.h5
171450368/171446536 [==============================] - 1s 0us/step
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
resnet101 (Functional)       (None, 2048)              42658176
_________________________________________________________________
flatten_1 (Flatten)          (None, 2048)              0
_________________________________________________________________
dense_5 (Dense)              (None, 1024)              2098176
_________________________________________________________________
dropout_2 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_6 (Dense)              (None, 27)                27675
=================================================================
Total params: 44,784,027
Trainable params: 2,125,851
Non-trainable params: 42,658,176
_________________________________________________________________
```

```
Epoch 00031: early stopping
```

Fig.13: Transfer Learning models

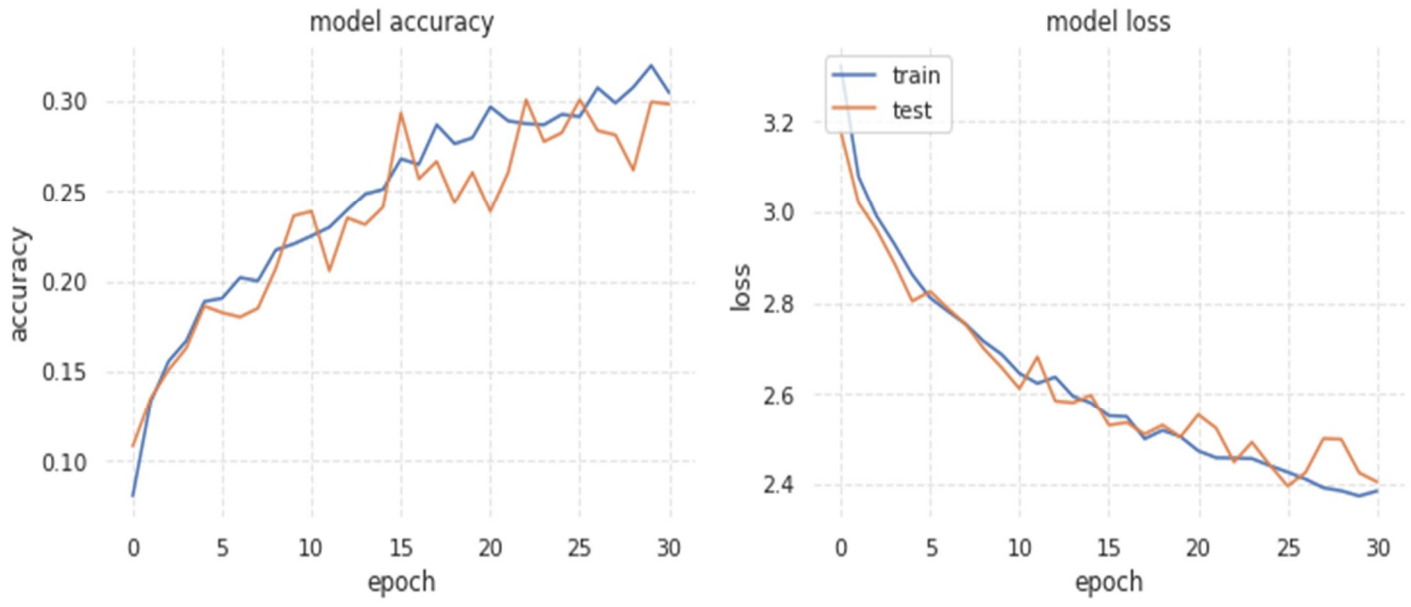


Fig.14: Training and Test Performance Metrics over 30 Epochs

This dual-pane visualization tracks the evolution of model accuracy and loss across a 30-epoch training cycle, showing a steady improvement in predictive performance. The close alignment between the training (blue) and test (orange) curves indicates that the model is generalizing well to unseen data with minimal evidence of significant overfitting.

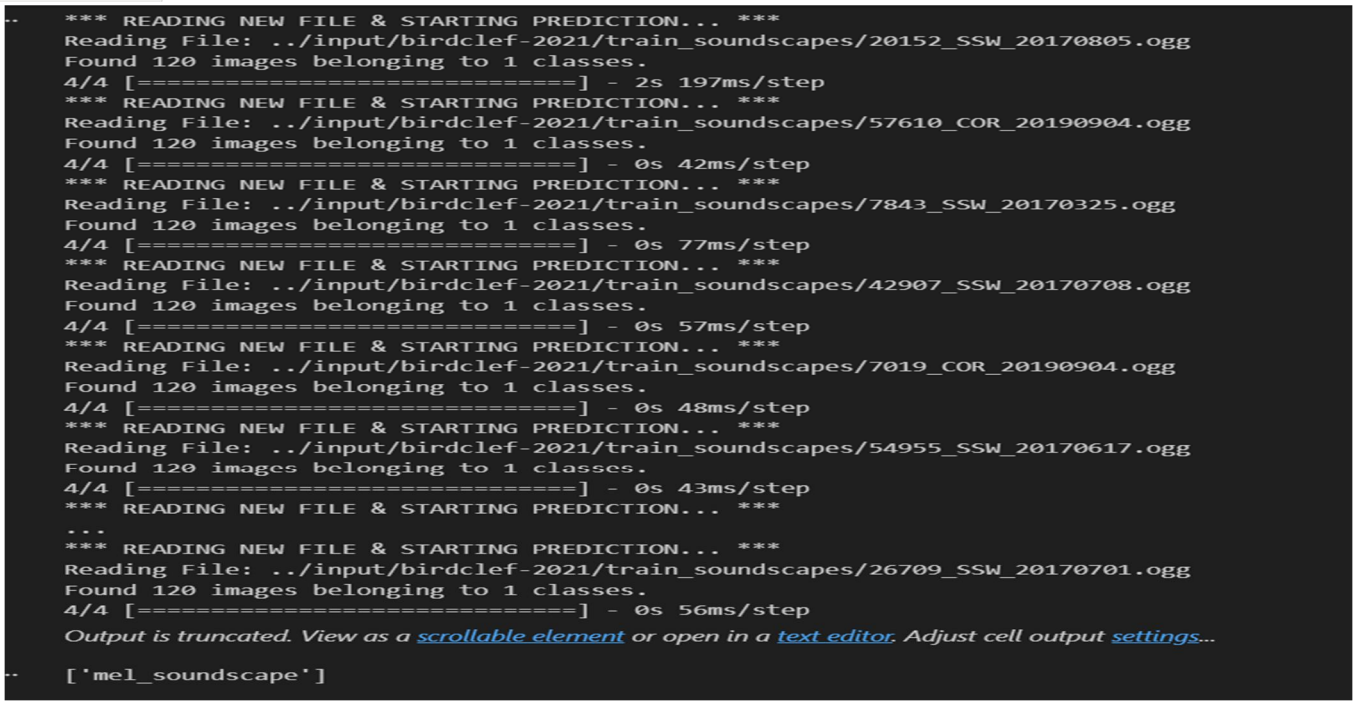


Fig.15: Model Inference Logs for BirdCLEF-2021 Soundscape Processing

This console output captures the sequential execution of a prediction pipeline, showing the automated loading and processing of multiple `.ogg` audio files from the BirdCLEF-2021 dataset. Each log entry details the conversion of audio into image-based representations (spectrograms), with consistent processing times of roughly **40–80 ms per step** across the 120-image batches.

V. RESULT

The proposed bird species detection system produced effective classification results by applying deep learning together with transfer learning techniques. Training was carried out using bird audio recordings that were converted into useful representations through FFT and spectrogram-based feature extraction methods. The results obtained from experimentation showed that spectrogram inputs performed better than FFT features in terms of classification accuracy. This is because spectrograms contain both time-related and frequency-related information, allowing the model to learn sound patterns more clearly. During the training process, the model showed stable improvement, where accuracy increased gradually and loss values reduced over multiple epochs.

The use of transfer learning significantly improved the efficiency of the system. Since the model was initialized with pre-trained weights, fewer training cycles were required to achieve good performance. It also helped in extracting meaningful features and enhanced the prediction capability of the network.

Testing on validation data and unknown audio samples indicated that the model had strong generalization ability. In most cases, it was able to correctly recognize different bird species with high reliability. Overall, the developed system delivered accurate and consistent performance, making it useful for real-world bird sound identification applications.

VI. CONCLUSION

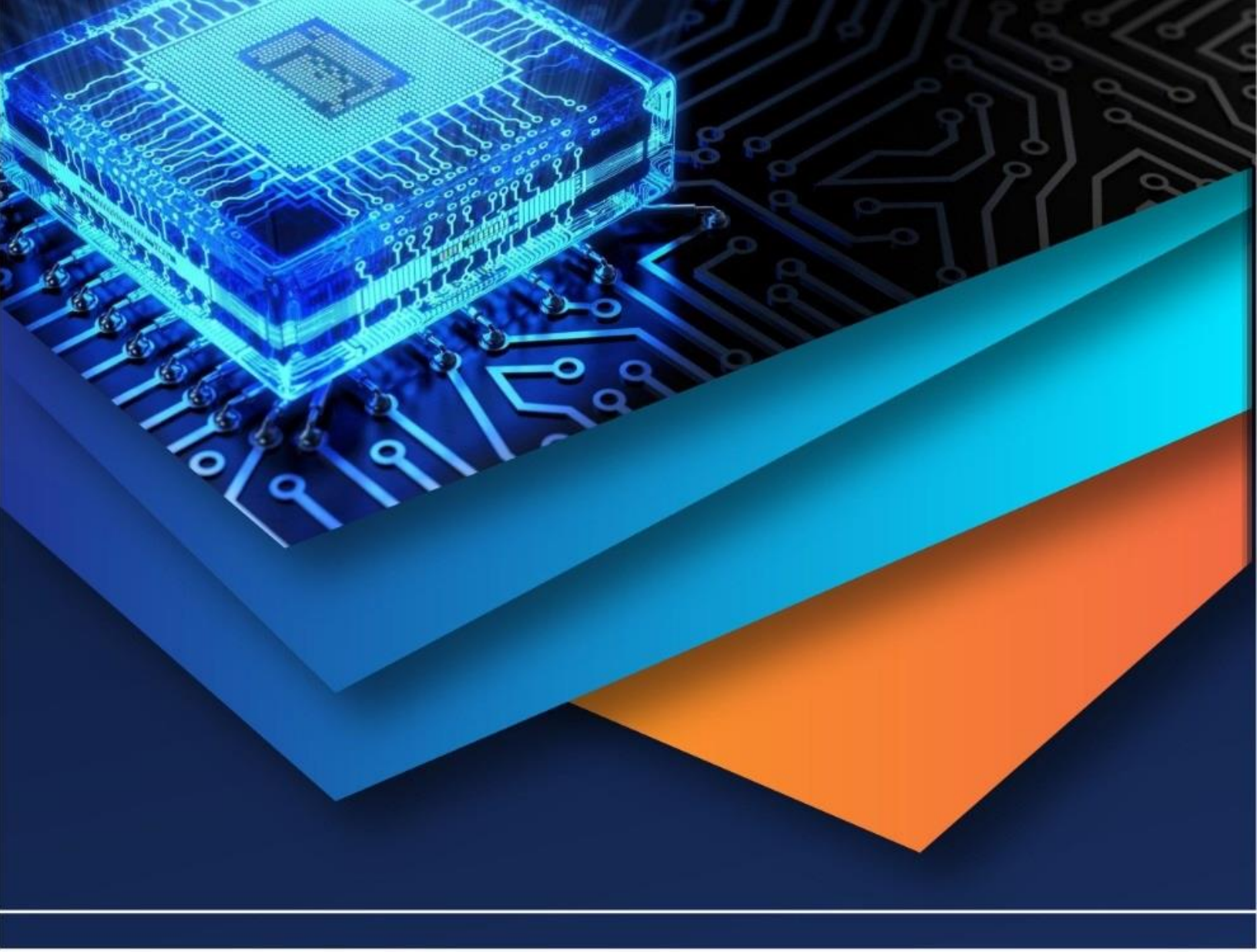
This study introduces an effective system for recognizing bird species through their audio recordings by combining signal processing methods with deep learning techniques. The proposed model uses Fast Fourier Transform (FFT) and spectrogram-based feature extraction methods along with a transfer learning Convolutional Neural Network (CNN) to improve classification accuracy. The experimental results show that spectrogram representations provide better performance than FFT features alone, as they capture both time and frequency information present in bird vocalizations. The use of transfer learning also helps in achieving higher accuracy while reducing the overall training time and computational effort.

Another important outcome of this work is the model's ability to generalize well to new and unseen audio samples. This indicates that the developed system can be applied in practical areas such as environmental observation, wildlife conservation, and biodiversity monitoring.

In conclusion, the proposed approach offers a reliable, scalable, and efficient solution for bird sound classification. Future enhancements may include the use of advanced deep learning architectures, larger datasets, and real-time implementation for field-based applications.

REFERENCES

[1] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning," *IEEE Transactions on Signal Processing*, 2019.
[2] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE Int. Workshop Machine Learning for Signal Processing (MLSP)*, 2015.
[3] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, 2017.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.
[6] BirdCLEF Dataset, LifeCLEF Initiative. [Online]. Available: https://www.imageclef.org/lifeclef/bird
[7] B. McFee *et al.*, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf.*, 2015.
[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[9] "Bird sound identification system using deep learning," *Procedia Computer Science*, vol. 233, pp. 597–603, 2024.
[10] R. Qin and J. Huang, "Towards accurate bird sound recognition through multi-scale texture-aware modeling," *npj Acoustics*, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY