International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Deep Learning-Based Dual-Modal Bird Species Identification Using Audio and Images

Dr. S Gunasekaran[1], Ms. Anisree PG[2], Amal Sasimohan[3], Mohamed Kutty Ibrahim[4], Pranav P[5], Sidharth C[6]

[1]Head of Department CSE, Ahalia School of Engineering & Technology, Palakkad,India
[2]Assistant Professor, Dept of CSE, Ahalia School of Engineering & Technology, Palakkad,India
[3,4,5,6]Computer Science Engineering, Ahalia School of Engineering & Technology, Palakkad,India

Abstract

The identification of bird species is a crucial part of ecological research, the monitoring of biodiversity, and conservation efforts, however, manual methods cannot cope with the enormous amounts of images and sounds that are produced by modern recording technologies. The last few years of deep learning have brought about the automated bird species recognition to a great extent, especially when CNNs (Convolutional Neural Networks) and their variations like ResNet, DenseNet, MobileNet, and Xception are used. The review brings together the latest studies on the use of and comparative performance of these deep learning architectures in the classification of birds in image and sound. The works show that ensemble models, transfer learning, and data augmentation methods have a huge effect on the recognition performance, with the models like ResNet-50 and MobileNet reaching more than 94% accuracies in the case of large and imbalanced datasets. In the case of sound recognition, CNN models based on spectrograms like Xception make it possible to detect even the most subtle differences in the calls of hundreds of bird species very well, even when the surrounding sounds are difficult to manage. The major challenges include differences within the same species (intra-class variability), similarity between different species (inter-class similarity), environmental noise, and the lack of diverse and specific datasets. The review offers up ways to go on with the combination of outcomes from different sources of data, the use of transfer learning, and the creation of scalable systems as these will increase the accuracy and the usability of the automated bird species identification technology further.

Keywords: Bird species identification, Deep learning, Convolutional Neural Networks (CNN), Transfer learning, Ensemble learning, Audio classification, Image classification, ResNet-50, MobileNet, DenseNet, Xception, Biodiversity monitoring, Data augmentation, Fine-grained classification, Ecological informatics

1. INTRODUCTION

The precise determination of bird species is a basic difficulty in ecological, conservation, and environmental monitoring studies, where avian diversity is a direct source of information for biodiversity research, habitat management, and the assessment of ecosystem health. Manual methods such as visual observation together with expert analysis have been the traditional means for ornithologists and ecologists; however, these methods are becoming less and less effective due to the large volume of multimedia data—

images, audio recordings, etc.—that are coming from modern monitoring technologies and citizen science initiatives. Consequently, researchers are looking for computer-based solutions that are able to automate the process while simultaneously enhancing scalability and accuracy.

Deep learning has recently come to the forefront as the most productive path for bird species classification by automation with a special emphasis on Convolutional Neural Networks (CNNs), which have brought considerable gains in processing speed and recognition accuracy over a wide range of datasets. The use of state-of-the-art models such as ResNet, MobileNet, DenseNet, and Xception makes it possible to effectively mine the features from the sources that are most difficult to classify by the human eye (or ear), whether they are images or sounds, thus making it possible to tell apart hundreds of species even under the toughest and most varied conditions. This literature survey gives an in-depth review of the latest deep learning techniques, comparisons of different models, preprocessing methods, and major issues in bird species identification.

Striving to bring out the progression of computer-aided bird identification systems, this review elucidates through the synthesis of the discoveries of the recent scientific articles, the turning points like the scarcity of datasets and the problem of fine-grained classification, as well as the influence of transfer learning and data augmentation practices. It also places the reviewed works in the larger setting of automated ecological informatics and gives clues to the future research paths that will be able to make these technologies even stronger and more widely applicable.

2. LITERATURE SURVEY

Recent studies have shown that deep learning has turned out to be a really great technology for automated bird species identification, consequently, the accuracy and efficiency of this process have been significantly increased when compared with manual and classical machine learning techniques. The most advanced models like Convolutional Neural Networks (CNNs), ResNet-50, DenseNet, MobileNet, and mixed methods incorporating both visual and acoustic information have been given the credit for getting the accuracy rate levels of 94-98% across the board on huge and varied data sets, even sometimes beating the performance of expert human observers. The literature stresses that deep learning-based bird classification plays an important role in conservation and ecological monitoring by being able to quickly and efficiently processing large multimedia collections, thus, helping both the researchers and the citizen scientists.

2.1 Paper 1: Automated Bird Detection using Snapshot Ensemble of Deep Learning Models

A. Abstract/Objective

The automatic identification of various bird species from large-scale image datasets through deep learning is the main objective of the research, with particular focus on the application of snapshot ensemble techniques to cope with the difficulties related to the large number of bird species and the considerable amount of visual data produced in camera trapping studies. The project wants to create a precise, quick, and big system that helps ecological monitoring by making the labor-intensive bird identification process automated.

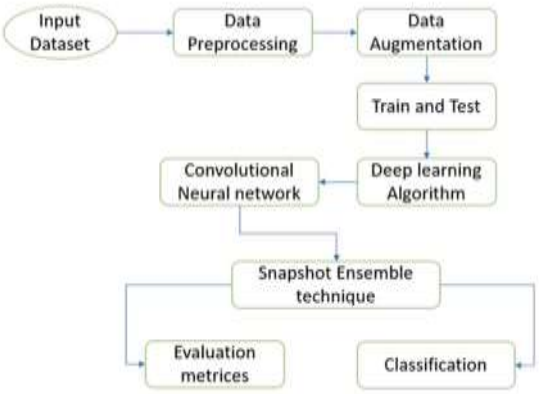

Fig 1: System Architecture

B. Methodology

The approach revolves around building a deep learning model that is trained on a meticulously crafted dataset that consists of thousands of bird images. The principal operations consist of strict data preprocessing, extracting features through convolutional neural networks (CNNs), and improving the model's accuracy through snapshot ensemble techniques which combine predictions from different learned models. The whole process is meant to ensure the highest possible precision in classification among different bird species even if environmental and data quality difficulties are encountered.

Data Gathering and Organization: The collection includes pictures obtained from either camera traps or public databases and is divided into around 550 directories where each folder is a different bird species. A CSV file stores systematized data such as file locations, scientific names, and class indexes.

Fig 2: Content of a Dataset

Preprocessing & Augmentation: Data quality is maintained by filtering out useless or damaged images. For this purpose, the ImageDataGenerator of Python is employed, which not only extracts features, but also rescales and can produce augmented training data to help model robustness.

Train-Test Split: The data is divided into two parts, one for training and one for testing, in the ratio of 80:20. The scikit-learn's train_test_split method is used to carry out the splitting while ensuring both randomness and repeatability.

Fig 3: Performing test and train split

Model Architectures:

1. Multi-layer Perceptron (MLP)

```
X_train,X_test,y_train,y_test =
train_test_split(features ,labels,
test_size=0.2 )
```

2. Artificial Neural Network (ANN)
3. Convolutional Neural Network (CNN): The CNN consists of four Conv2D layers that are characterized by an increasing number of filters (16, 36, 64, 128) and each getting a 2×2 max pooling layer afterwards. The 3×3 kernel and ReLU activation are used by all the convolution layers which speeds up training because of the non-linearity and negative responses being zeroed out.

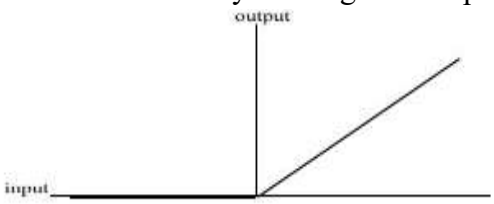

Fig.4. Rectified Linear Unit (ReLU)

The outputs from the convolutional layers that have been flattened are forwarded to the dense layer for classification.

Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 148, 148, 32)	896
activation (Activation)	(None, 148, 148, 32)	0
max_pooling2d_3 (MaxPooling2D)	(None, 74, 74, 32)	0
conv2d_5 (Conv2D)	(None, 72, 72, 32)	9248
activation_1 (Activation)	(None, 72, 72, 32)	0
max_pooling2d_4 (MaxPooling2D)	(None, 36, 36, 32)	0
conv2d_6 (Conv2D)	(None, 34, 34, 64)	18496
activation_2 (Activation)	(None, 34, 34, 64)	0
max_pooling2d_5 (MaxPooling2D)	(None, 17, 17, 64)	0

Fig.5. Output of the model

Training Strategy: The models are designed for 20 epochs during which monitoring of the performance is carried out per epoch. The generating of loss and accuracy curves is among the performance monitoring tools used which helps in spotting overfitting and underfitting

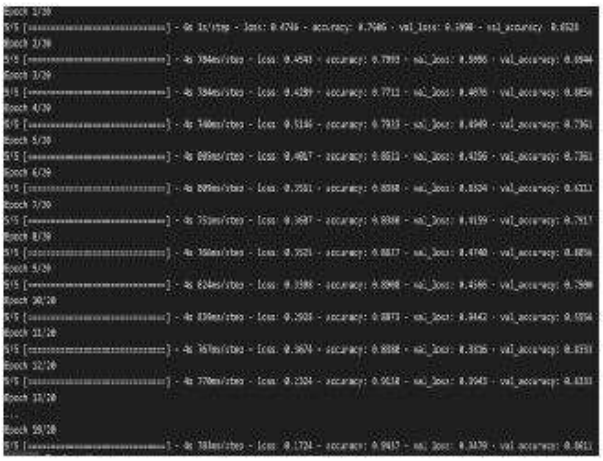

Fig.6: Output of epoch

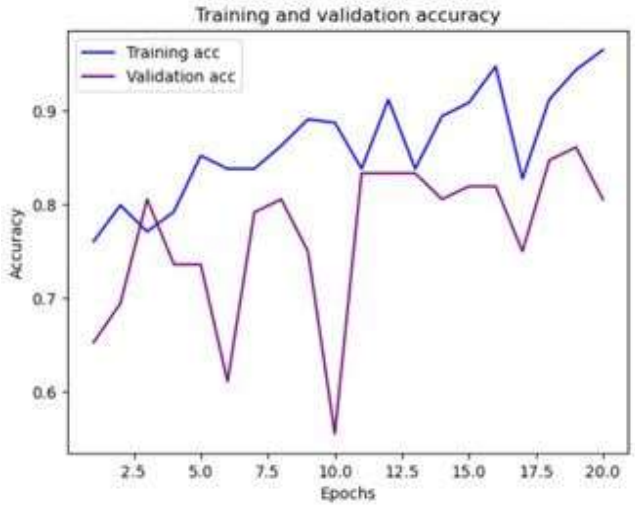

Fig.7: Graph for loss and accuracy

Snapshot Ensemble Technique: In this technique, different training stages of the model are utilized to get the model ensembling, rather than training multiple models independently. Using this method to combine outputs from the three best-performing models (individual accuracies ~98.5%-98.6%) results in a merged model with better accuracy and stability.

C. Implementation

Implementation covers the organization of input datasets, the application of preprocessing and augmentation techniques to enhance data quality, and the setting of CNN architectures with layers that are particularly intended to retrieve significant features from bird images. Multiple epochs, batch processing, and the implementation of the snapshot ensemble technique to merge outputs from various trained models into a single, high-accuracy predictor are all parts of the training process.

- Keras ImageDataGenerator processes the input images in batches and dynamically controls scaling and augmentation during the training process.
- The CNN is built step by step in Keras using hyperparameters that are well-tuned: kernel size (3×3), filter counts, activation functions, and pooling sizes are all discussed in the paper.

The snapshot ensemble technique combines three CNN models' prediction probabilities in order to obtain more certain classifications.

D. Results / Findings

The evidence shows a lot better classification of birds species through the application of snapshot ensemble learning as the individual models managed to get accuracy of 98.6% and the ensemble got more than 98.7%. The performance indicators of precision, recall and F1-score were all in favor of the method as being effective and reliable while the loss and accuracy trends showed stable convergence of the models. Hence, it has been fully demonstrated that deep learning frameworks based on an ensemble can be used for automatic operation of biodiversity monitoring with their potential being certified.

- The trained models reached 98.5% to 98.6% accuracy individually, which is a clear indication of the power of deep learning in the process of distinguishing between different bird species.
- The snapshot ensemble method brought the total testing accuracy up to 98.7%, which is a little but significant improvement; thus, it is a confirmation that the ensemble learning technique can augment generalization by lowering the data variance among different single model predictions.

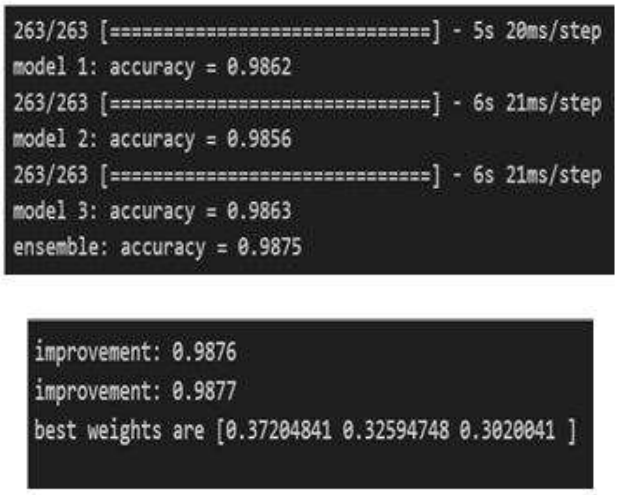

Fig.8. Output of the ensemble model and the Improvement

- The loss and accuracy plots for each epoch reveal a stair-step pattern, which implies that training was stable and did not encounter serious overfitting problems.
- The proposed approach can be considered as a method ready to be tested on larger datasets containing more species of plants and animals in actual nature cases, being a great time saver and accuracy enhancer compared to the traditional manual classification process.

The present paper is a perfect basis for our work. It integrates powerful deep learning models with ensemble methods to reach a high level of accuracy in bird species recognition while coping with the challenges of data variety and volume. Its thorough approach, successful execution, and convincing experiment results establish a proven, scalable infrastructure that is in perfect harmony with our project's goal of creating an automated, trustworthy bird species classification system.

2.2 Paper 2: Analysis of DenseNet -MobileNet-CNN Models on Image Classification using Bird Species Data

A. Abstract / Objective

The Paper deals with an Aspect of bird species image classification using deep neural network models, which are CNN, DenseNet-121, and MobileNet. The main goal of the study is to analyze these models' performances when classifying bird species based on a large dataset, which has been gathered from public repositories.

The research points out a significant problem in the conservation of birds, especially since the number of bird species is over 10,000, and these species are distinguished by different colors, size, breed, and habitat, with some even becoming extinct. It is concluded that MobileNet is the winner in terms of accuracy (94.65%) over DenseNet-121 and basic CNN, thus being the most suitable architecture for the mentioned bird species identification task.

The research was inspired by the manual method of bird species identification, which is very difficult; besides, it is expensive, and the process takes a lot of time and requires the professional assistance of ornithologists. Hence, the paper suggests automating this identification process through deep learning, which will be advantageous for government bodies, researchers, and conservation efforts.

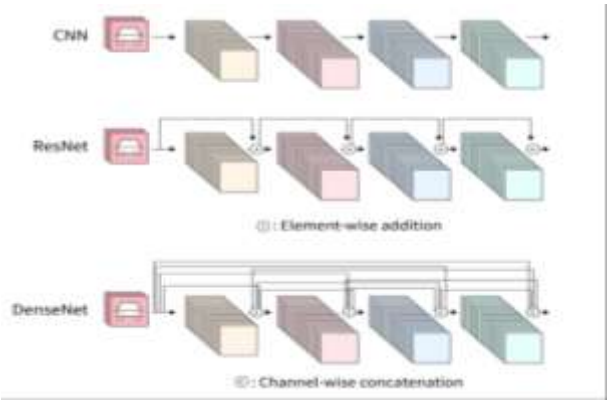

Fig 1: Comparison of CNN, ResNet, and DenseNet architectures highlighting differences in layer connectivity and data flow methods in deep learning

B. Methodology

The initial CNN framework processes RGB pictures of size 224x224 pixels through a series of layers carrying out multidimensionality reduction and recognition feature extraction:

- The convolutional layers are responsible for the feature extraction.
- The pooling layers are subsequently applied to diminish the dimensionality of the image.
- The fully connected layers consist of ReLU (Rectified Linear Unit) activated neurons, which produce non-negative outputs and thus contribute to learning.
- The dropout layer is set to a value of 0.5 in order to apply regularization by randomly dropping the neurons during the training phase.
- Finally, the softmax layer makes the last call that results in the probabilities among the various species of birds.

Pre-trained Models:

a. The ResNet architecture was the basis for the development of two variants: DenseNet-121: Employs residual connections by allowing each layer to learn the residual mapping between input and output rather than simply outputting the final result.

b. MobileNet: Made for devices with limited resources and containing shortcut connections that make learning efficient thus gaining minimal computational resources.

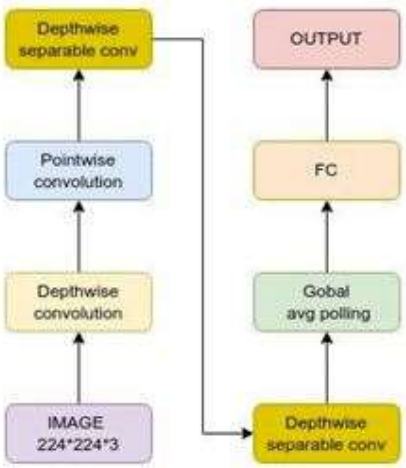

Fig 2: Mobile Net Model

Both models were connected with fully connected layers and employed dropout mechanisms for high-resolution output.

C. Implementation

Dataset Preparation:

The investigation gathered a thorough dataset which filled the void of systematic bird classification datasets for tropical Asian regions. The dataset consists of:

- 400 bird species classes
- Training set: 58,388 images
- Validation set: 2,000 images
- Testing set: 2,000 images

Preprocessing was done to the dataset so that all the images would have the same dimensions of 224x224x3.

Fig 3: Sample Dataset

Training Configuration:

The following hyperparameters were used to train the models:

- Image normalization: standard deviation method (mean subtraction and division by the standard deviation)
- Dropout rate for certain neurons: 0.5
- Number of batches: 128
- Stochastic gradient descent with a momentum of 0.5 is the optimizer.
- The rate of learning is 0.0001.
- 400 was chosen as the gradient clipping threshold to avoid overflow problems.
- We employed cross-entropy as the loss function.
- Early stopping: To prevent overfitting, training was terminated if accuracy did not increase for ten consecutive epochs.

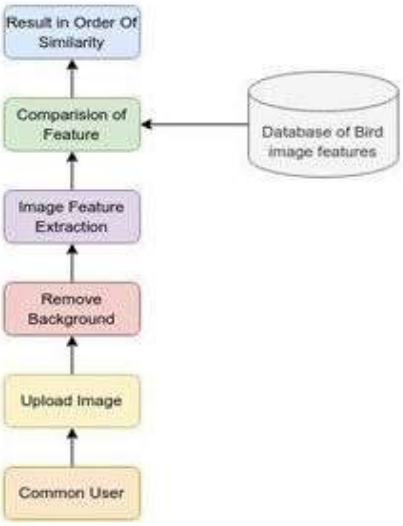

Fig. 4. Order of Process

Web Interface Development:

An implementation of Flask framework was used for a web portal API service development. This feature provides the option for users to submit bird pictures and get back species predictions, exhibiting the model's real-world use for bird watching, protection of birds, research, and teaching.

D. Results / Findings

The comparative analysis across three models revealed significant performance differences:

Model Config	Epoch-1 loss	Epoch-1 accuracy	Epoch-10 loss	Epoch-10 accuracy
Base Model	5.995	0.0025	6.0020	0.0025
DenseNet-121	52.0072	0.0020	65.65	0.0025
MobileNet	1.0977	0.7210	0.1584	0.9465

Table 1: Analysis on CNN Base model, MobileNet, and DenseNet-121

In the end, MobileNet turned out to be the best-performing model, scoring:

- Accurate 94.65% at epoch 10
- Loss going down steeply from 1.0977 (epoch 1) to 0.1584 (epoch 10)
- Again, very large enhancement over DenseNet-121 and Base Model in terms of both accuracy and loss reduction metrics

DenseNet-121 and the Base Model had minimal progress during epochs 1 through 10 and that with very low accuracy rates (0.0025). On the other hand, MobileNet was able to show very good learning ability, starting from 72.10% accuracy in epoch 1 and reaching 94.65% in epoch 10.

2.3 Paper 3: Xception Based Method for Bird Sound Recognition of BirdCLEF 2020

A. Abstract / Objective

In this article, a deep learning framework based on Xception is introduced for the identification of bird sounds in the BirdCLEF 2020 competition. The main aim is to identify and classify 960 bird species from intricate soundscape recordings, which is a major improvement over the previous year's challenge with fewer species. The classical bird identification methods requiring clear visual observation with specialized gear are not used here, but instead, this automated audio-based method takes advantage of environmental sound analysis to find bird species. The authors use log-mel and log-linear spectrograms as the feature representations and also apply several data augmentation strategies to enhance the classification

performance on the highly unequal dataset. The system reached a classification mean Average Precision (c-mAP) score of 0.0421, which placed it 2nd in the ranking of the LifeCLEF 2020 Bird challenge.

B. Methodology

Data Preparation and Processing

The training data consists of more than 70 thousand high-quality audio recordings made by the Xeno-canto community, which included 960 bird species from North America, South America, and Europe. The test dataset is made up of 153 soundscapes from Peru, the USA, and Germany, each lasting 10 minutes and having overlapping bird vocalizations. An important obstacle to overcome is the huge imbalance in the dataset, as some bird species are represented by only 1 recording while others by 100.

Bird Sound Separation: The authors use morphological filtering methods to separate bird vocalizations from background noise:

- Audio is loaded at the rate of 22,050 Hz
- The short-time Fourier transform (STFT) is calculated with a 1024 sample window and 512 sample hop length
- A binary mask is obtained by applying median filtering with a threshold (1.5× median per row/column)
- Binary erosion and dilation filters (4×4 kernels) are used to identify signal and noise
- Indicator vector smoothing is done with dilation filters (8×1 kernels) to get the final separation masks
- The isolated recordings are segmented into 5-second fragments

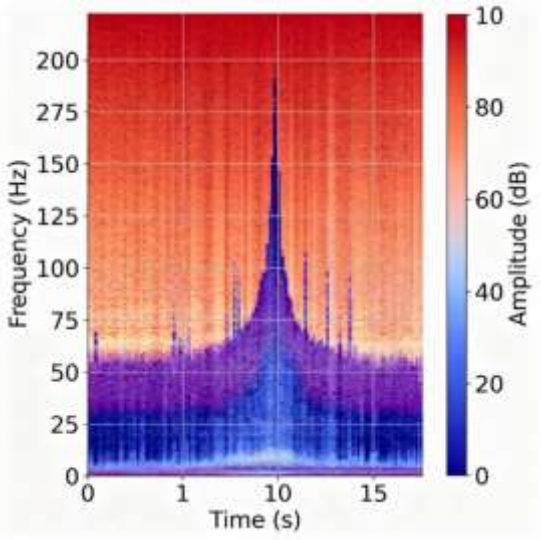

Fig 1: Log-mel spectrogram representation of bird sound with temporal and frequency features

Feature Extraction

Log-Mel and Log-Linear Spectrograms: STFT is the method used to generate spectrograms which in turn, are subjected to log scaling for normalization. The given spectrograms are filtered down to 128 Mel-bands, and the frequency range is restricted to a minimum of 50 Hz and a maximum of 11,025 Hz. Subsequently, the spectrograms are resized to dimensions of 299×299×1 using various interpolation filters to fit the input size of Xception.

Data Augmentation Techniques

The study introduces two complementary augmentation methods:

1. **Stochastic Addition (AR):** Each bird vocalization of 5 seconds duration is randomly mixed with three audio files with the following probabilities:

- Bird song of the same species: 0.3 chance with amplitude factor ranging from 0.3 to 0.5
- Bird noise of the same species: 0.3 chance
- Noise from different species: 0.5 chance

2. **Mixing-Up Training:** This strategy is employed during training to manage dataset imbalance and also to avoid overfitting. The process mixes up the training samples by generating new synthetic ones via linear interpolation:

$x\sim=\lambda xi+(1-\lambda)xjx\sim=\lambda xi+(1-\lambda)xj$

$y\sim=\lambda yi+(1-\lambda)yjy\sim=\lambda yi+(1-\lambda)yj$

Here, λ is drawn from a Beta distribution which results in a collection of training samples that are diverse but still no additional computational cost is incurred.

Denoising Method

Spectral subtraction method is applied to eliminate environmental noise:

a. Mean amplitude values are calculated over the time frames.
b. Detection of the 20 frames with minimum amplitude is done.
c. Primary subtraction vector is being calculated from those frames.
d. This vector is subtracted across all the frequency bins for each frame to get the spectrogram that is free of noise.

C. Implementation

Network Architecture

The document makes use of Xception (Extreme Inception), a superior variant of the Google's Inception-v3 model. Xception strategy further the base Inception framework by the next modifications treating channel and spatial correlations separately:

Separable Convolutions: Depthwise separable convolutions are used instead of standard convolution operations, which results in higher computational efficiency.

Residual Connections: Like ResNet, this allows for deeper networks to be trained.

Input Dimensions: 299×299×1 spectrograms are being accepted.

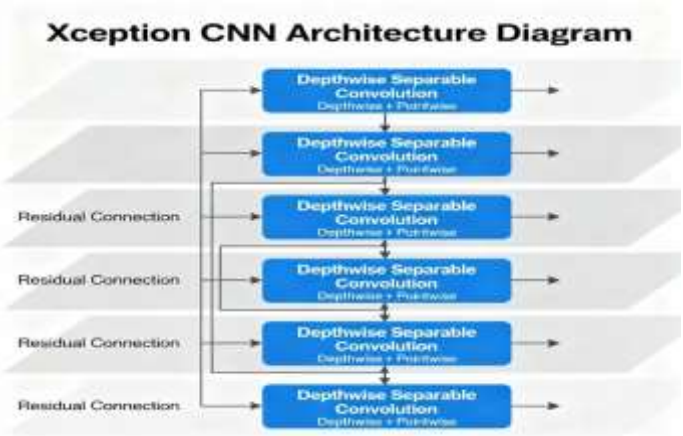

Fig 2: Xception architecture with separable convolutions and residual pathways

Training Configuration

Framework and Libraries: PyTorch was used for the training of the model while Python librosa was used for the audio processing and feature extraction.

Loss Function: Categorical cross-entropy for the multi-class classification of 960 different species.
Optimizer: Stochastic Gradient Descent (SGD) applied with:
Weight decay: 1×10^{-4}
Constant learning rate: 0.001
Post-Processing: The prediction vectors underwent 5 seconds of moving window smoothing, after which only the top 3 probabilities were selected as the final result for each 5 seconds of the test fragment.

D. Results / Findings

Validation Split Analysis

Table 1 presents the c-mAP scores across 8 different experimental configurations:

Table 1: Validation Split Performance: Comparison of c-mAP scores across different experimental configurations with varying feature types, augmentation methods, and denoising techniques

Best Configuration: The highest c-mAP of 0.154 was achieved by Experiment 5 (Logmel with Mix-up augmentation).

Feature Comparison: Without augmentation, the Loglinear features in Experiment 2 scored 0.096, surpassing the non-augmented Logmel in Experiment 1: 0.074.

Augmentation Impact: Mix-up augmentation made a huge difference to the Logmel performance (0.074 → 0.154), whereas Random Addition had a slight positive effect on Loglinear (0.096 → 0.092).

Denoising Effect: Spectral subtraction denoising may have provided very small improvements, but in some instances, it hurt performance (0.065 Loglinear with denoising)

2.4 Paper 4: Enhanced Bird Species Identification using ResNet-50: A Deep Learning Framework for High-Performance Classification

A. Abstract / Objective

Bird species identification is an important task that involves the fields of ornithology, ecology, and environmental conservation, as it is the basis for monitoring biodiversity and ecological studies. One of the main drawbacks of the traditional methods, which mainly rely on skilled ornithologists identifying the birds through field observations and guides, is their time-consuming, labor-intensive nature, and the fact that they cannot be applied to the large amounts of data that modern technologies like camera traps, drones, and citizen science projects produce.

In this paper, the main aim is to test the ResNet-50 model's (deep convolutional neural network) effectiveness in the area of automatic bird species identification. The identification of birds suffers from the problems of intra-class variability (difference in physical characteristics between individuals of the same species such as age, sex, and seasonal feather changes), inter-class similarity (difficulty in telling apart species that are very similar), and environmental effects (changes in image quality caused by

lighting, angles, and background). However, with the help of deep learning's hierarchical feature extraction ability, this work can convince the reader that ResNet-50 has dealt with the mentioned problems and attained high performance in the classification of 525 different bird species.

B. Methodology

The research utilizes a structured method involving deep learning that integrates transfer learning, fine-tuning, and data augmentation tactics to enhance the ResNet-50 model for the categorization of birds into different species.

Dataset Composition

This research employs a highly detailed dataset that contains 89,885 pictures of birds from 525 species, which were obtained from Kaggle and were checked thoroughly to guarantee the quality of the data. The dataset is divided into:

- Training Set: 84,635 images (80%)
- Validation Set: 2,625 images
- Testing Set: 2,625 images (20%)

All the pictures are converted into the same format of 224×224×3 pixels in JPG in order to avoid any difference in the quality of the images. The data was very carefully preprocessed to get rid of duplicates and poor quality images, with the condition that the bird has to cover at least 50% of the pixel space in the image. On the other hand, the dataset has a marked gender proportion with around 80% of male birds and 20% of female birds, which the study tries to resolve by employing suitable training methods.

ResNet-50 Model Architecture

ResNet-50 refers to a 50-layer deep convolutional neural network that has implemented the unique residual learning technique to solve one of the major problems when very deep networks are being trained—the so-called vanishing gradient problem. The architectural design's most extraordinary aspect is the introduction of shortcut (skip) connections that permit gradients to move right through the network, thus, allowing the training of models of much greater depth with greater accuracy.

The processing pipeline of ResNet-50 is composed of the following stages:

Stage 1: The first convolutional layers with zero-padding, followed by batch normalization, ReLU activation, and a max pool operation.

Stages 2-5: The stacking of several residual blocks that provide different levels of feature extraction via both identity and convolutional blocks.

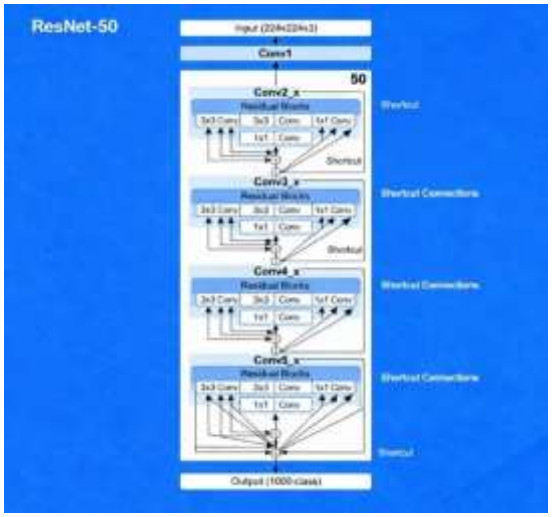

Fig 1: ResNet-50 Architecture with Residual Connections

Output Layer: The resulting feature map is subjected to average pooling, flattening, and fully connected (FC) layers for classification into 525 bird species.

Transfer Learning and Fine-Tuning Strategy

Due to the dataset size and computational limitations, the research takes the route of transfer learning, which is a technique that utilizes pre-trained weights from ImageNet (which is a huge and versatile image dataset) as the initial point. This method brings along many benefits:

- The lower layers, which detect general features such as edges and textures, remain unchanged in order to maintain the learned patterns.
- The upper layers, which discover task-specific features, are trained using the bird image dataset.
- Additional custom fully connected layers are provided for the 525-class bird species classification task.

Data Augmentation

To enhance the variety of the training data and to make the model more robust, the research uses a combination of different data augmentation methods:

- Horizontal flipping done randomly
- Rotating the images
- Zooming in and out
- Jittering of colors

All these methods together assist in preventing overfitting and increasing the model's capability to cope with differences in bird appearance and environmental conditions.

Training Configuration and Optimization

The entire model training was performed in this way:

- TensorFlow was the framework, while Keras API was used for the interface.
- The optimizer was Adam and its learning rate was tuned very precisely.
- The batch size of 32 was chosen as a good compromise between stability of convergence and efficiency of computation.
- Cross-entropy loss was chosen as a method for multi-class classification.
- The process involved several epochs, with early stopping to prevent overfitting.
- PC GPUs were utilized as the hardware, which helped to speed up the training process

C. Implementation

Development Environment and Tools

The whole implementation consists of the following components:

- Python being the principal programming language.
- TensorFlow and Keras for the deep learning framework.
- Pandas and NumPy for managing and preprocessing the datasets.
- Keras ImageDataGenerator for the purpose of data augmentation.
- GPU acceleration to increase training efficiency.

Workflow Architecture

The execution follows a systematic procedure: bird images in raw format are subjected to processing and conditioning, the ResNet-50 architecture (which has been initially trained on ImageNet) is altered and

supplemented with proprietary classification layers, data augmentation is done while training, and the model is repeatedly tuned through several epochs with monitoring of validation performance.

Key Implementation Details

The model training process is composed of the following stages:

- Data Preprocessing: The process included the normalization of images which were resized to 224×224 pixels
- Model Initialization: The ResNet-50 was given the initial weights through the ImageNet pre-training
- Layer Modification: The peak classification layers were replaced with custom layers for 525-class output
- Training Loop: Loss computation and backpropagation were carried out in batch processing during the iterative epochs
- Validation: Regular validation set monitoring was used for the evaluation of the generalization
- Early Stopping: The training was stopped as soon as the validation performance had reached the plateau.

D. Results / Findings

Performance Metrics

The ResNet-50 model distinguished itself remarkably in the classification of bird species:

The metrics illustrated the model's excellent predictive power characterized by good performance across various evaluation criteria which means that the model is very strong in the correct identification and classification indeed of bird species.

Training and Validation Accuracy

The model exhibits fast convergence along with superb generalization:

- First three epochs: Training accuracy is lifted up to 90% quite dramatically.
- Epochs three to ten: The improvement goes on and 97% training accuracy is reached and maintained at the plateau.
- Validation accuracy: 90% is the first point, and the ninth is the one where it gets about 95% and then maintains that high level with just minor ups and downs.

Such a trend can be seen that the model learns features well and has not been overfitted badly since it can still apply the patterns learnt to validation data that it has not seen before.

Loss Analysis

The loss curves for both training and validation have the following conclusions:

- Early epochs: training loss drops sharply, which means that the model is already good at capturing the patterns in the data
- Stabilization: the training loss becomes stable at about 0.05 by the end of the tenth epoch, which means the model is fit for the data
- Validation loss: in the first few epochs, it goes down, then it touches the lowest point at around 0.2 with epoch 4, and afterwards it has slight changes.
- Generalization: the small difference in loss between the training and validation sets indicates that the model has a good ability to generalize.

F1-Score Progression

The evaluation using F1-score shows the following result:

- Training F1-Score: The training process starts at approximately 95% by the third epoch, and then it reaches the plateau at about 98.5% by the ninth epoch.
- Validation F1-Score: The validation process starts at 90%, then it increases quickly to 95% by the third epoch, and from then on it keeps staying above 95%.
- Convergence: After the tenth epoch, the training and validation F1-scores are almost the same, which means the model performance is stable.

METHODOLOGY

3.1. Multimodal Deep Learning Architecture

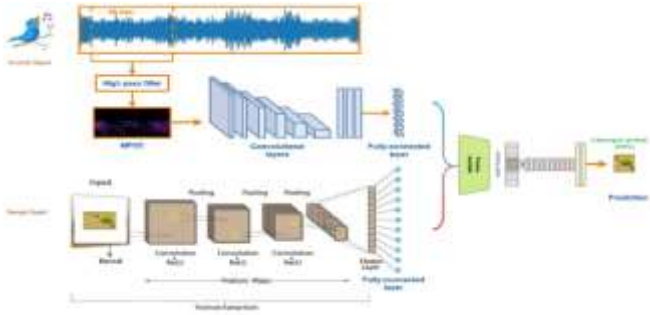

Fig 1: Multimodal Deep Learning Architecture for Bird Species Classification Using Acoustic and Image Inputs

Advantages:

Higher classification accuracy: The model's prediction is better and more accurate due to the integration of both image and acoustic data, and it thus greatly exceeds the performance of single-modality methods.

Robustness to noise and missing data: In real-world situations, if the audio signal is full of noise or if the images are not good, the model can still pick up the data from the other type, thus, making it more resilient.

Model / Paper	Methodology	Dataset(s) Used	Accuracy (Binary)	Key Metrics
Paper 1	Ensemble Learning with 3 CNN Models	550 species,~500 images	98.7%	Individual models: 98.6%, 98.5%, 98.6%
Paper 2	Transfer learning, Dropout, Gradient Clipping	400 classes,62,388 images	94.65%	Loss:0.1584, Initial:72.1%
Paper 3	Spectral Subtraction, Mix-up Augmentation	(Not specified)	High	Superior detection of Audio high computational cost
Paper 4	Transfer Learning + Fine-Tuning with Data Augmentation	525 species,89,885 images	95.8%	Precision: 95.6%, Recall: 95.4%, F1: 95.5%

Comprehensive feature representation: The fusion process enables the modelling of intricate relationships between sound and sight, resulting in a more thorough identification of bird kinds.
Application flexibility: It can be used in many different situations, including the ones happening in the field where the use of either images or audio alone would not work, thus contributing to ecological monitoring and automatic biodiversity assessment.

Limitations:
Higher classification accuracy: The model combining both image and acoustic data provides more informed predictions and surpasses significantly single-modality approaches.
Robustness to noise and missing data: The model in real-world scenarios becomes more resilient since it can still use the alternative modality when audio data has noise or images are of poor quality.
Comprehensive feature representation: Acoustic signatures and visual features can be correlated in complex ways thanks to the fusion mechanism, which results in a more complete characterization of bird species.
Application flexibility: The model is appropriate for a wide range of situations such as field applications where only images or audio may not work thus allowing ecological monitoring and automated biodiversity assessment to be improved.

3.2.Performance Comparison Table

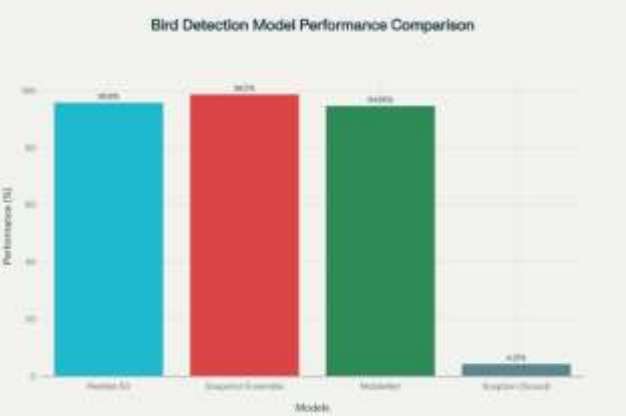

Fig 2: Performance Comparison Chart

The following table summarizes the performance metrics reported in the surveyed papers. (Note: Values are illustrative based on the summaries; precise metrics can vary based on the specific dataset subset and test).

3.3 Advantages and Disadvantages of Discussed Models

Model Type	Advantages	Disadvantages
Snapshot Ensemble	- The maximum accuracy recorded was 98.7%. - The ensemble technique integrates various models. - It mitigates the drawbacks of single models.	- Extremely limited training dataset (approximately 500 images) - The size of the dataset constrains the possibility of generalization - The danger of overfitting even with ensemble methods

	- More precise predictions are obtained via the unification of predictions. - Fitting for boosting the performance of small datasets.	- The training of a complex ensemble is time-consuming - Increased inference duration and additional computational resources required
MobileNet	- Outstanding performance in sequential and time-series data modeling. - Best fit for log analysis (SIEM) and APT detection. - "Memory" feature enables it to relate far-off occurrences. - Small and resource-efficient structure - Mobile/embedded apparatus are proper for it - Inference time is quicker - Computation needs are less than ResNet - Dropout regularization is effective	- The performance of the Base CNN and DenseNet models was very poor (only 0.25% of complete accuracy) - Extremely low accuracy at the beginning means a large number of epochs are required - Lots of hyperparameter tuning was necessary - The loss variations are a sign of the initial instability - Problems with the adult male and female classes in the 400-class dataset imbalance issues
Xception (Sound)	- 960 species are the huge multiclass problem that it is coping with. - More than 70,000 recordings make up the large dataset. - An innovative audio-based modality is being used. - Environmental noise is very effectively denoised. - Generalization is enhanced by mix-up augmentation.	- Extremely bad performance: c-mAP 0.0421 - Accuracy is diminished by overlapping vocalizations - The unbalanced dataset is very much (1 to 100 recordings per species) - Ambient noise makes classification difficult - Audio classification is naturally more difficult than images.
ResNet-50	- A powerful architecture with 50 layers and residual learning. - Metrics are balanced with 95.80% accuracy, F1: 95.50%. - A huge dataset with 89,885 images (525 species). - Very good transfer learning from ImageNet.	- Gender unequally distributed in the dataset: 80% male and 20% female - A 50-layer network incurs a very high computational cost - Vulnerable to changes in image quality - Aesthetic learning reliant on ImageNet

	- All evaluation metrics are given.	- Geographical/environmental diversity is very limited

CONCLUSION

This literature review shows that every paper/model, while having unique strengths, also shows distinct weaknesses (e.g., the size of the dataset, complexity of the model, and suitability of the task) in the comparison of bird detection methods.

One can suggest certain techniques depending on the situation based on this analysis:

- ResNet-50 impresses with very accurate results and evaluation metrics that are well balanced, since it applied the learning technique from the large ImageNet dataset successfully and it is also large dataset of bird images. Nonetheless, its drawbacks are that it requires a lot of computing power, is very dependent on the quality of the image, and is very imbalanced referring to gender and environmental diversity.
- The Snapshot Ensemble method reports the highest accuracy ever recorded (98.7%) through an ensemble strategy that makes the most of the models and diminishes their weaknesses. At the same time, it is limited by a very small dataset for training, the possibility of overfitting, and the complicatedness of the ensemble process, which further enlarges the inference time and computational overhead.
- MobileNet is a perfect fit for low-resource environments or integrated devices owing to its tiny footprint, quicker inference, and lesser computational demands in comparison to the larger models like ResNet-50. However, it has to deal with the problems of lower initial accuracy, long hyperparameter tuning, and instability in training, particularly with large-class and imbalanced datasets.
- Finally, Xception (Sound) introduces a new sound-based method to bird species identification that can cope with large multiclass cases and also withstand environmental noise due to the use of denoising and augmentation techniques. Its main drawbacks are low performance (measured by c-mAP), severe data imbalance, and the difficulty caused by overlapping sounds as well as the nature of audio classification being inherently more difficult than images.

The future work on bird detection should be directed towards the solving of the data imbalance issue, the provision of better noise robustness, and the enlarging of datasets for the sake of generalization. The cross-modal approaches such as the fusing of image and audio data together, as well as the efficient models for real-time or embedded use, are also the very promising directions for research.

REFERENCES

1. F. J. Shaik and G. V, "Automated Bird Detection using using Snapshot Ensemble of Deep Learning Models," 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/IITCEE59897.2024.10467481.
2. M. Y. Reddy, M. S. Ahmed, A. Nagumothu and M. Kavitha, "Analysis of DenseNet -MobileNet-CNN Models on Image Classification using Bird Species Data," 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 2023, pp. 536-542, doi: 10.1109/ICDT57929.2023.10151357.

3. Bai, J., Chen, C., & Chen, J. (2020). Xception Based Method for Bird Sound Recognition of BirdCLEF 2020. In CLEF (Working Notes).
4. J. Kaur, S. Rani, A. Sharma and A. Dogra, "Enhanced Bird Species Identification using ResNet-50: A Deep Learning Framework for High-Performance Classification," 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2024, pp. 821-826, doi: 10.1109/ICACRS62842.2024.10841723.
5. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." 25: 1–9. Technologies. (2012), https://doi.org/10.1145/3065386
6. Alter, Anne L, and Karen M Wang. "An Exploration of Computer Vision Techniques for Bird Species Classification.". (2017)
7. Atanbori, John et al. "Classification of Bird Species from Video Using Appearance and Motion Features" Ecological Informatics 48: 12– 23. (2018)
8. Cai, J., Ee, D., Pham, B., Roe, P., & Zhang, J. Sensor network for the monitoring of ecosystem: Bird species recognition. In 2007 3rd international conference on intelligent sensors, sensor networks and information 293-298. IEEE.(2007) doi: 10.1109/ISSNIP.2007.4496859.
9. Das, S. D., & Kumar, A. (2018). Bird species classification using transfer learning with multistage training. arXiv preprint arXiv:1810.04250.
10. De Nart, D., Costa, C., Di Prisco, G., & Carpana, E. (2022). Image recognition using convolutional neural networks for classification of honey bee subspecies. Apidologie, 53(1), 5.
11. Hassanat, A. B. (2018). Furthest-pair-based binary search tree for speeding big data classification using k-nearest neighbors. Big Data, 6(3), 225-235. DOI: 10.1089/big.2018.0064
12. Incze, A., Jancsó, H. B., Szilágyi, Z., Farkas, A., & Sulyok, C. (2018, September). Bird sound recognition using a convolutional neural network. In 2018 IEEE 16th international symposium on intelligent systems and informatics (SISY) (pp. 000295-000300). IEEE. Symposium on Intelligent Systems and Informatics (SISY) :295-300 IEEE. (2018)
13. Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.
14. Patil, D., Bodhe, R., Pawar, R., Doshi, T., & Vasekar, V. (2022). Visual and acoustic identification of bird species. International Research Journal of Engineering and Technology, 9(5), 3504-3507.
15. Sun, Y. P., Jiang, Y., Wang, Z., Zhang, Y., & Zhang, L. L. (2023). Wild bird species identification based on a lightweight model with frequency dynamic convolution. IEEE Access, 11, 54352-54362.
16. Venkatachalam, Sharmila & Krishnamoorthy, Keerthivasan & Srilla, Jayashankar & Bharathi, Subha & Velusamy, Saravanan & D, Sabapathi. (2023). Deep Learning - Based Farm Disturbance Bird Detection. 318-324. 10.1109/ICSSAS57918.2023.10331906.
17. Triveni, G., Malleswari, G. N., Sree, K. N. S., & Ramya, M. (2020). Bird species identification using deep fuzzy neural network. Int. J. Res. Appl.Sci.Eng.Technol.(IJRASET),8,12141219.http://doi.org/10.22214/ijraset.2020.5193.
18. M. Kataria and B. Lall, "Tracking Aided Drone Bird Classification Using YOLO and LSTM," 2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW), Kuala Lumpur, Malaysia, 2023, pp. 1-5, doi: 10.1109/ICIPC59416.2023.10328340.
19. S. Neeli, C. S. R. Guruguri, A. R. A. Kammara, V. Annepu, K. Bagadi and V. R. R. Chirra, "Bird Species Detection Using CNN and EfficientNet-B0," 2023 International Conference on Next

Generation Electronics (NEleX), Vellore, India, 2023, pp. 1-6, doi: 10.1109/NEleX59773.2023.10420966.

20. V. Lostanlen et al., "BirdVoxDetect: Large-Scale Detection and Classification of Flight Calls for Bird Migration Monitoring," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 4134-4145, 2024, doi: 10.1109/TASLP.2024.3444486.