Advanced Bird Species Classification Using Convolutional Neural Networks and Transfer Learning

Tanuja Kodali[1], CH. Ramadevi[2]

[1]*MTech Student, Department of CSE, Sir C R Reddy College of Engineering(A), Eluru W G Dt Andhra Pradesh*

[2]*Associate Professor, Department of CSE, Sir C R Reddy College of Engineering(A), Eluru W G Dt Andhra Pradesh*

Abstract- Many people find solace in visiting bird sanctuaries, where they may observe various birds and appreciate their unique characteristics and colors. Data about species is a cornerstone of biodiversity protection efforts. Because there are so many similarities between bird species, both within and between classes, it can be quite challenging to determine which bird species you're looking at. Convolutional neural networks (CNNs) and transfer learning models like VGG16, VGG19, ResNet-50, ResNet-101, Inception-V3, DenseNet, MobileNet, and EfficientNet have recently been useful. Quite a few cutting-edge algorithms for picture classification have accomplished outstanding results. Improving a deep learning platform for image-based bird species recognition has been the focus of this article. For both classification and prediction, a convolutional neural network is utilized. A paradigm geared at skip connections in neural networks is being presented to enhance feature extraction. For the training image, the suggested technique attained a classification accuracy of 99.00%. The suggested method makes it easy for inexperienced bird watchers to identify the species of birds in a photograph.

Keywords- ***Transfer Learning, Deep Learning, Pretrained Models, Convolutional neural network (CNN) · Bird species · VGG-19, ResNet-50, Inception-V3, DenseNet, MobileNet, EfficientNet.***

I. INTRODUCTION

BeginningChanges in bird numbers can be a marker of changes in the environment, such as pollution, habitat loss, or climate change. This phenomenon means that bird species can be used to tell if an ecosystem is healthy. Bioacoustics is a useful way to measure how many distinct species there are, identify and classify different species' vocalizations, and keep track of populations over time. Identifying bird species is one of these tasks. It includes looking at different auditory properties, such as frequency, duration, and other factors, to get the right classification [1]. This capacity is very important for both conservation efforts and scientific inquiry. Because data is growing so quickly, traditional bird monitoring methods are becoming boring, time-consuming, and expensive for ecologists because they need highly skilled people to identify bird species. One benefit of using auditory data instead of visual data is that it can be used to find birds when sight is limited or not possible. Many bird species can be distinguished by their sounds (such as songs, screams, and noises) instead of their looks. Also, sound recordings are less impacted by things like the weather or the amount of light, so birds can make calls at any time of day or night. Finally, audio recordings give us information about how birds act, which we need to use learning models to figure out which birds are smart enough to identify.

Identifying bird species is crucial for monitoring and safeguarding biodiversity, particularly in India, home to a diverse range of bird species. Most prior investigations have concentrated on estimating bird species within brief recordings. This is much easier than identifying the bird and when it sings. It usually uses manually segmented bird sounds, which could make the performance seem better than it really is when classifying bird species in continuous recordings. Identifying species aids researchers in refining migration patterns, tracking population dynamics, and assessing environmental effects.

However, traditional methods that rely on visual observation and a professional's ability to tell species apart have many problems, like bad lighting, blocked vistas, and birds with identical feather patterns. Sound-based identification, which uses bird calls, also has problems such as background noise and changes in how birds sound. These constraints underscore the necessity for sophisticated techniques. Although considerable research has been conducted in avian classification, the implementation of multimodal methodologies is still insufficiently investigated, revealing a significant research vacuum in the integration of visual and aural modalities for species identification within India's unique ecological framework.

To rectify this deficiency, we present a framework that integrates optical and auditory data to identify Indian avian species. We combine visual data that a Deep Convolutional Neural Network (DCNN) has processed with auditory data that Long Short-Term Memory (LSTM) networks have analyzed using early and late fusion methods. Our method improves the reliability and accuracy of identification by using the iBC53 (Indian Bird Call Dataset), which is a complete collection of bird pictures and sounds from India [2]. We employed an LSTM network for acoustic data due to its suitability for sequential data, shown by bird vocalizations. LSTMs adeptly capture the temporal dependencies intrinsic to bird sounds, essential for differentiating across species with analogous tonal qualities yet distinct call sequences. The Mel-frequency cepstral coefficients (MFCCs) inputted into the LSTM encapsulate the spectrum characteristics of bird calls by condensing essential auditory information into a low-dimensional feature set [3]. We integrated the two data types employing early and late fusion methodologies to enhance identification precision and resilience. Early fusion integrates features from both visual and audio modalities simultaneously, enabling the model to comprehend the interactions between the two input sources during training. In late fusion, the two modalities are processed individually, and the outputs of the DCNN and LSTM networks are amalgamated at the decision stage using a weighted average of the probability distributions produced by each model [4].

The primary objective of this research is to develop and evaluate a novel Visual-Acoustic Fusion avian classification model employing Convolutional Neural Networks (CNNs) and transfer learning architectures such as VGG16, VGG19, ResNet-50, ResNet-101, Inception-V3, DenseNet, MobileNet, and EfficientNet.

The rest of this manuscript is set up like way. In Section 2, we give a short overview of the new ways to classify fruit freshness that have been created. In Section 3, we talk about the method we came up with. Section 4 lists the outcomes of the simulations and comparisons, and Section 5 wraps up our findings and suggests what we might do next.

II. RELATED WORKS

Transfer learning strategies have already been the subject of a number of different kinds of research projects with the aim of classification.

Current state-of-the-art CNNs can classify bird poses from various perspectives and bird situations by using pre-trained networks. Accuracy has increased by these techniques to roughly 85.4% [7]. With fixed feature extraction, another study that used transfer learning of pre-trained AlexNet Convolutional NeuralNetworks (CNN) has attained an accuracy of 46% . Their second version is a multi-class SVM that uses computer vision techniques to extract RGB histogram values and Histogram Oriented Gradients to identify the birds, increasing their accuracy by 5% to 9% compared to HOG alone.

The DWT can break down an input signal into a set of approximate coefficients (low-frequency components) and detailed coefficients (high-frequency components) at multiple degrees of decomposition. This graph illustrates how sparse and multi-resolution it is. The first is seen to be a beneficial signal, whereas the second is thought to be noise. Researchers like Xie et al. and Gong et al. have used this great denoising method a lot to recognise birdsong, which has made signal analysis and interpretation more reliable [8]. Chi et al. used the WT, convolutional neural network (CNN), inception module, and long short-term memory (LSTM) network to find out what was wrong with a structure

[9]. For denoising to work well, it is critical to choose the correct threshold and threshold functions. People know how to employ threshold functions in both the rigorous and soft methods that Donoho and Johnstone came up with.

MFCC is the main feature used to sort bird calls, however it has several problems that are commonly solved by combining it with additional features. For instance, Bonet-Solà and Alsina-Pagès utilised k-nearest neighbour (kNN), neural network (NN), and Gaussian mixed model (GMM) classifiers to amalgamate MFCC, gamma-tone cepstral coefficients (GTCCs), and narrowband (NB) data [10]. Andono et al. likewise employed MFCC and GTCC to extract features, and they were able to correctly categorise birdsong 78.11% of the time. When they employed particle swarm optimisation (PSO) on 264 bird vocalisations [11], the success percentage increased up to 78.33%. Xie et al. especially used a CNN to get deep features from the Hilbert-Huang transform (HHT), short-time Fourier transform (STFT), and wavelet transform (WT). They also used MFCC to get great results [12].

Along with MFCC, FBank features have also been used a lot in recognition. Sui et al. utilised the transformer model in conjunction with FBank to identify ten bird calls, achieving an accuracy rate of 85.77%. Peng et al. integrated FBank with SincNet-based Sinc spectrograms, utilising 2D Gabor filters for Sinc spectrograms and secondary feature extraction to enhance feature representations [13]. García-Ordás et al. also employed a complete CNN with FBank to categorise 17 birdsongs with an accuracy of 85.71%. Wang et al. employed a distinct methodology by integrating FBank, delta, and delta-delta FBank (3D-FBank) with a Phylogenetic Perspective Neural Network (PPNN), attaining an accuracy of 89.95% in the classification of songs from 500 avian species. In unrelated studies, Liao et al. employed FBank within a deep neural network (DNN) to analyse giant panda vocalisations, while Lin utilised FBank and Linear FBank to discern human voiceprints. Wu et al. also used ResNet18 and 3D-FBank to find targets underwater [14].

Stefan Kah et al. [33] employed their proprietary CNN model for the avian recognition task. The authors employed mel-spectrograms of processed audio data, executed data augmentation and normalisation techniques, and partitioned the recordings into fixed-length segments. In the BirdCLEF 2017 dataset, they received a mean average precision score of 0.605. Hybrid models learn from data that varies over time or in a sequence by using the best parts of both convolutional and recurrent neural networks (RNNs). There exist studies on CNN plus RNN hybrid models.

T. Gupta et al. [35] examined the newest photo classification algorithms that include transfer learning (ResNet18, ResNet50, and VGG16) and found that VGG16 was the most accurate, with an accuracy of 61.9%. They used the embedded hybrid CNN and RNN models to make their model work better, and they got the best accuracy at 67%. The dataset they used was the 2020 Cornell Bird Challenge. This collection has 264 distinct kinds of birds. But they only used 100 different types of birds to fix the problem of class imbalance.

III. THE PROPOSED MODEL

The objective of this research effort is to devise an efficient technique for identifying Deep Learning-Based Bird Species Classification Using CNN and Transfer Learning. The entire process is illustrated in Fig. 1.

Given that image quality significantly affects deep learning results, preprocessing approaches are utilized to eliminate artifacts and improve brightness and contrast. Various pre-trained transfer learning models are assessed based on accuracy, precision, recall, and F1-score. The entire process is illustrated in Fig. 1.

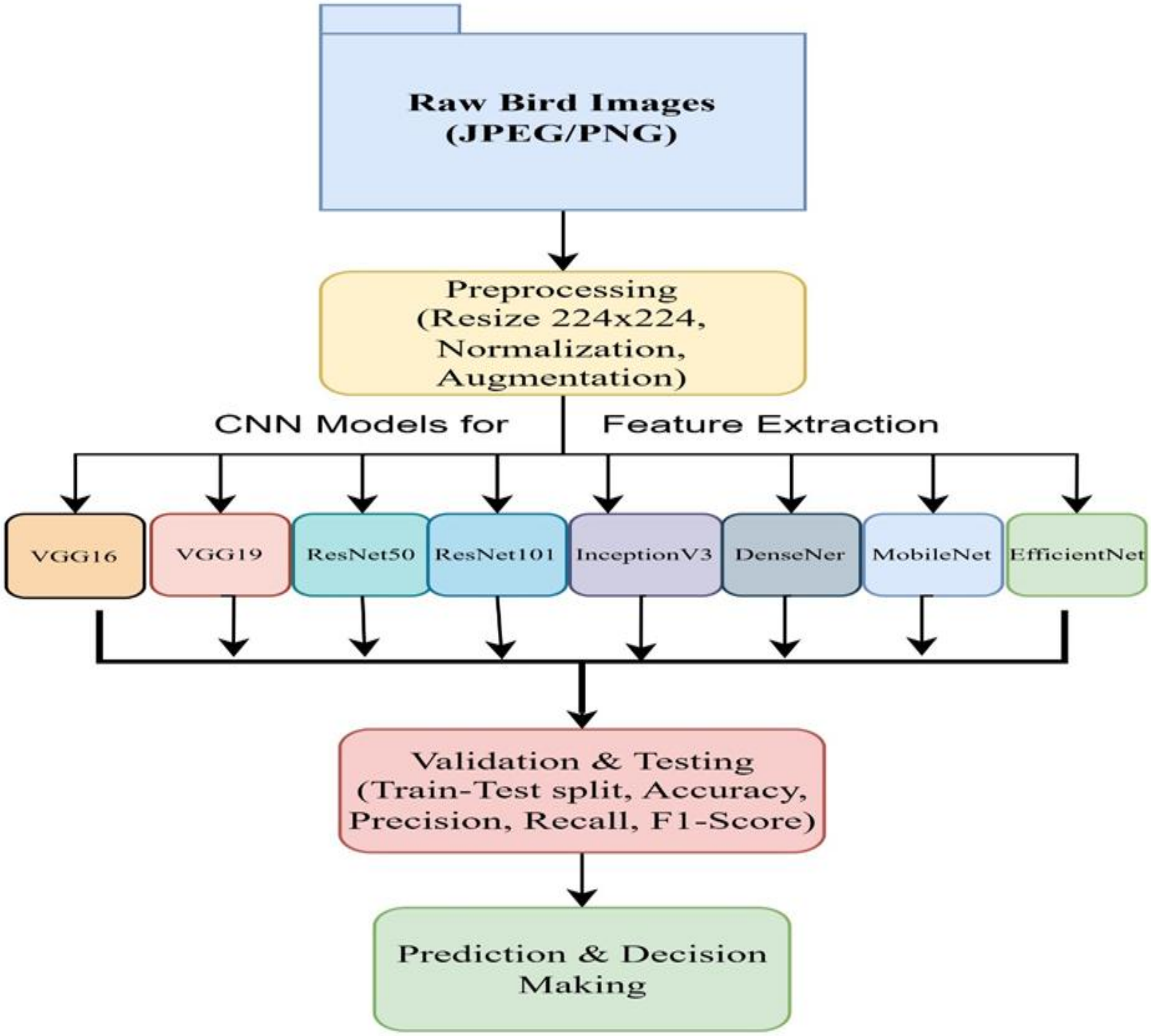

Fig 1. Overall architecture of the system

A. Data set description

A dataset of bird photos from Kaggle [17] is utilized for both training and testing to finish the implementation of the comparison of these transfer learning models. There are 58388 training images of 500 species of birds, and 2000 testing and validation images of 500 species of birds.

B. Data pre-processing

This is mostly about getting the data ready so that we can get the most out of the model when we train it with this data. For the model to be easier to train, all the photos in the dataset must be in the same format. We changed the size of the photographs to 256x256px to make them all the same size before processing them. Then we established groups of data to use for training the model. A group of 32 photos was made. Batch size is the number of samples that are processed before the model is changed. This is necessary because after training using batches, the expected results may be compared to the actual outcomes to find the error. This error is then utilized to change the algorithm to make the model better. All of the pictures are RGB pictures. So, the size of the photos is (256,256,3).

The photos were downsized to 256 × 256 pixels and normalized by dividing the values of each pixel by 255 to meet the input needs of the deep learning models. Deep learning methods usually need more training data than regular machine learning methods. To make the training set bigger and more varied, data augmentation was used. This method uses different image-processing techniques, like zooming, stretching, and changing the contrast, to make numerous versions of the same image. Too much or wrong changing of important aspects for picture recognition could make learning harder and lead to bad results.

This paper chooses six image processing settings (brightness, contrast, gamma, hue, saturation, and central-crop) to keep information safe and stop the algorithm's learning from being distorted. The model will use the new images created by the augmentation process throughout training. This step is an online augmentation that doesn't require storing more data on the computer's hard drive. We looked at a number of different image-processing settings and found the ones that worked best for improving efficiency during augmentation:

Brightness (0.3): changing the brightness; Contrast (0.7): changing the difference in tone; Gamma (gamma = 4, gain = 3): controls the overall brightness and colour balance of blue, green, and red; Hue (0.9): changing the colour spectrum, which includes red, yellow, green, and blue; Saturation (0.3): changing the colour and white light merging;
Central-crop (0.96): dividing the focal region into parts.

Transfer Learning Models:
After completing the preprocessing of the input features, the processed data is deployed for training using various traditional transfer learning methods [14] as follows:

VGG16:
VGG16-CNN We used data augmentation methods to look at the data and learn more about it. It works like it should and helps reduce problems with overfitting while training machine learning algorithms. In the field of data augmentation, the VGGNet model is called VGG16. It is a 16-layer CNN model that has learned from more than a million pictures in the dataset. The proposed VGG16-CNN features twelve fire modules. The convolution layer uses these modules to figure out how many parameters to put in the fire module.By increasing the number of input channels by the number of filters, it creates a CNN model with few parameters. The training data will go through a convolution layer, a pooled layer, a flattening layer, and a dense layer during the VGG16-CNN stage. CNN models might help visualisation systems learn new things more easily. VGG16 employs features from raw remote sensing data, and the classification stage uses CNN output feature maps to figure out what the category is. Transferring the well-learned parameters of VGG16, which is trained on the widely annotated dataset, would make image recognition tasks a lot easier. Getting a lot of samples and labelling them is hard. This means that the VGG16 architecture contains millions of variables that need to be trained and a lot of data that has been labelled.

VGG19:
The Visual Geometry Group (VGG) at the University of Oxford produced VGG19, a deep neural network architecture, in 2014. The model has 19 layers: 16 convolutional layers and 3 fully connected layers. This is why it is called VGG19. All of the convolutional layers use a 3 × 3 kernel with a stride of 1 and padding to keep the spatial resolution, while the max-pooling layers use a 2 × 2 kernel with a stride of 2 to make the spatial dimensions smaller. When using transfer learning, it is common to freeze the first few convolutional layers of VGG19 because they contain basic low-level characteristics like edges, textures, and gradients. The deeper layers, which learn more patterns that are specific to the task, are best for the target domain. This study utilises the VGG19 architecture for effective domain adaptation in predicting cardiovascular diseases (CVD) from electrocardiogram (ECG) images. The pretrained layers keep their strong ability to extract features from the ImageNet domain, while the deeper trainable layers learn patterns that are important for medicine, such as waveform shape, segment changes, and rhythm anomalies. This method helps the model find a good balance between transfer learning and acquiring specialised features. It keeps about 5 million trainable parameters that are specific to predicting CVD accurately.

Resnet 50
The graphic displays the ResNet50 design, which is a well-known type of Residual Network (ResNet) that is used to classify images. The design begins with a Zero Padding stage and then moves on to a simple convolutional layer (Stage 1) that includes

Convolution, Batch Normalisation, ReLU activation, and Max Pooling. The model is then made up of five primary pieces, which are named Stage 2 to Stage 5. There are convolutional blocks (Conv Blocks) and identity blocks in each component. These blocks have shortcut connections that are supposed to fix the problem of the gradient disappearing in very deep networks. After the stages for extracting features, the output proceeds via Average Pooling, Flattening, and a Fully Connected (FC) layer to get the final classification result. ResNet50 is great for recognising photos on a large scale since this architecture lets you train deep networks without losing accuracy.

Resnet 101:

ResNet101 is a popular deep learning design that uses the notion of residual learning to fix the problem of deterioration in very deep neural networks. ResNet101 has been successfully employed across multiple fields, including as medical imaging, pattern recognition, and computer vision, owing to its strong feature-extraction capabilities. There are a lot of convolutional layers in the architecture, and they are grouped into residual blocks. The output of one layer can skip over the layers in between and go straight to the output of a deeper layer. There are 104 convolutional layers in the setup, which are grouped into 33 residual blocks. Nine of these blocks use direct skip connections to send the output of the previous layer unmodified. Additionally, four blocks use a 1×1 convolution with a stride of 1 to make sure that the dimensions are the same before sending features to the next normalisation layers.

DenseNet-121:

DenseNet-121 is a deep convolutional neural network that was built to improve feature propagation and reuse by using dense connectivity. DenseNet is different from regular CNNs since each layer only gets input from the layer before it. In a dense block, every layer is connected to all the layers that come after it. This architecture makes sure that feature maps from earlier layers are sent straight to later layers. This cuts down on redundancy and speeds up gradient flow during training. DenseNet-121 has four dense blocks with a total of 121 layers. These layers are convolutional layers, batch normalisation layers, and transition layers. The transition layers, which are between dense blocks, use 1×1 convolutions and 2×2 average pooling to keep the network from being too complicated and to stop feature maps from spreading out.

Inception-V3:

InceptionV3 is a more advanced version of the original Inception architecture. It adds factorised convolutions, better ways to reduce grid size, and better ways to normalise data. The main idea behind the Inception design is to let the network operate with data at several levels at the same time. Each Inception module has many convolutional filters of varying sizes (1×1, 3×3, 5×5) that work at the same time. Inception-V3 takes this idea a step further by breaking down bigger convolutions into smaller, more efficient operations (for example, replacing a 5×5 convolution with two 3×3 convolutions). These improvements greatly lower the cost of calculation while raising the power of representation. The model also includes batch normalisation, extra classifiers, and optimised grid reduction algorithms, which let deep networks be trained without the vanishing gradient problem.

Xception:

Xception is an improved version of the Inception architecture that replaces the usual Inception modules with depthwise separable convolutions. The main idea of Xception "Extreme Inception" is to completely separate spatial correlations and cross-channel correlations in feature maps. Xception uses depthwise convolutions to map spatial relationships for each channel separately. This is different from typical convolutions, which learn spatial and channel-wise filters at the same time, and then pointwise (1×1) convolutions to record channel interactions. This factorisation makes the model easier to use and more powerful while also making it easier to understand. The architecture comprises 36 convolutional layers that are organised into entrance, middle, and exit flows. Each flow has depthwise separable convolution blocks that are repeated and residual connections that let the gradient run smoothly.

MobileNetV3:

MobileNetV3 is a very efficient deep neural network that is built for lightweight, real-time applications. It combines depthwise separable convolutions, which were first seen in earlier versions of MobileNet, with other improvements to the architecture, such as squeeze-and-excitation (SE) modules, inverted residual blocks, and the use of the hard-swish activation function. These new ideas greatly increase representational capacity without adding much to the cost of computing. There are two versions of MobileNetV3: Small and Large. Each is designed for a different balance between performance and efficiency. The approach uses inverted bottleneck layers, which means that feature maps grow before depthwise convolution and then shrink again. This lets the model capture complex patterns without adding more parameters.

IV. RESULTS & DISCUSSIONS

Quantitative and visual studies were used to test how well the suggested model worked. For the quantitative assessment, various widely used performance indicators were employed through the categorisation report, including accuracy, recall, and F1-score. Accuracy is the percentage of true predictions made by the model compared to the total number of forecasts. Recall measures how well the model can correctly identify real positive cases and how often it misses real positive samples. The F1-score, which is the harmonic mean of precision and recall, is a good way to judge how well a model works, especially when the datasets are not evenly distributed. These metrics together give a full picture of how well the proposed model can categorise things.

To ensure reliable assessment of the machine learning model, several performance evaluation metrics were employed. The following measures were used to quantify the effectiveness of the proposed model:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} - - - -(Eq - 1)$$

$$\text{Precision} = \frac{TP}{TP + FP} - - - - - (Eq - 2)$$

$$\text{Recall} = \frac{TP}{TP + FN} - - - -(Eq - 3)$$

$$\text{F1score} = 2 * \frac{Precision + Recall}{precision * Recall} - (Eq - 4)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Table 1: Bird image classification – Performance Summary

TL Classifier	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
VGG16	98.21	98.34	98.27	98.30
VGG19	98.45	98.52	98.48	98.50
ResNet-50	98.67	98.72	98.69	98.70
ResNet-101	98.81	98.88	98.84	98.85
DenseNet121	98.90	98.94	98.92	98.93
InceptionV1	98.54	98.60	98.57	98.58
Xception	98.96	99.01	98.98	99.00
MobileNetV3	98.72	98.79	98.75	98.77
EfficientNet	99.02	99.04	99.03	99.04

The study of different transfer learning (TL) classifiers for classifying bird images shows that all of the studied architectures have good accuracy. Table 1 shows that all of the models had performance metrics between 98% and 99.04%. This shows that deep convolutional neural networks are quite good at finding unique characteristics in bird photos. EfficientNet had the best overall accuracy (99.04%) of all the models tested, as well as the best precision (99.02%), recall (99.04%), and F1-score (99.03%). Xception and DenseNet121 both did very well, which shows that their powerful feature reuse methods and deep hierarchical representations improve classification performance.

The comparison investigation shows that deeper residual networks like ResNet-101 did better than shallower ones like ResNet-50. This shows that having a deeper network can help capture complicated visual patterns. Traditional architectures such as VGG16 and VGG19 also achieved competitive performance, exceeding 98% accuracy, although slightly lower than more advanced architectures. Lightweight models like MobileNetV3 were good at classifying things and were also good at using less computing power, making them good for applications that need to work in real time or that don't have a lot of resources. The experimental results demonstrate that transfer learning substantially improves bird species categorisation performance, with EfficientNet identified as the most successful model in this study.

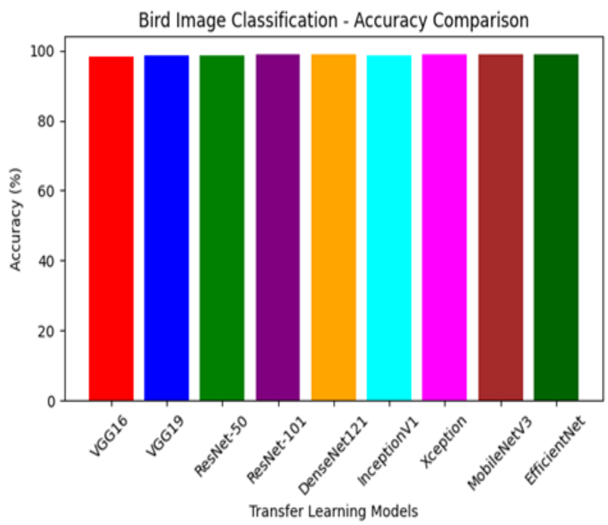

Fig 2: Accuracy Across Transfer Learning Models

V. CONCLUSION

This study provided a thorough assessment of transfer learning-based deep convolutional neural network (CNN) models for the classification of bird images. We looked at several pretrained architectures, such as VGG16, VGG19, ResNet-50, ResNet-101, DenseNet121, InceptionV1, Xception, MobileNetV3, and EfficientNet, to see how well they worked for feature extraction and classification. The experiments showed that all of the models worked well, with accuracy values between 98% and 99.04%. This study shows that deep learning approaches are strong enough to handle fine-grained picture classification tasks. EfficientNet had the best overall performance of all the architectures tested, with an accuracy of 99.04% and better precision, recall, and F1-score values. Advanced designs like Xception and DenseNet121 also did well since they could reuse features efficiently and represent them in a deep hierarchical way. Lightweight models like MobileNetV3 were nonetheless quite accurate and used less processing power, which made them good for use in real-time and resource-limited settings. The results show that transfer learning greatly improves the accuracy of bird species classification and is a good way to automate wildlife monitoring and biodiversity analysis.

REFERENCES

[1] Sprengel, E., Jaggi, M., Kilcher, Y., & Hofmann, T. (2016). Audio Based Bird Species Identification using Deep Learning Techniques. *Infoscience (Ecole Polytechnique Fédérale De Lausanne)*, 547–559. http://infoscience.epfl.ch/record/229232

[2] Swaminathan, B., Jagadeesh, M., & Vairavasundaram, S. (2024). Multi-label classification for acoustic bird species detection using transfer learning approach. *Ecological Informatics*, *80*, 102471. https://doi.org/10.1016/j.ecoinf.2024.102471

[3] Xie, J., & Zhu, M. (2023). Acoustic classification of bird species using an early fusion of deep features. *Birds*, *4*(1), 138–147. https://doi.org/10.3390/birds4010011

[4] Xie, J., & Zhu, M. (2023). Acoustic Classification of Bird Species Using an Early Fusion of Deep Features. *Birds*, *4*(1), 138-147. https://doi.org/10.3390/birds4010011

[5] Turker Tuncer, Erhan Akbal, Sengul Dogan,Multileveled ternary pattern and iterative ReliefF based bird sound classification,Applied Acoustics,Volume 176,2021,107866,ISSN 0003-682X,https://doi.org/10.1016/j

[6] Liu, J., Zhang, Y., Lv, D. *et al.* Birdsong classification based on ensemble multi-scale convolutional neural network. *Sci Rep* 12, 8636 (2022). https://doi.org/10.1038/s41598-022-12121-8

[7] Huang, K., Li, S., Kang, X., & Fang, L. (2015). Spectral–Spatial hyperspectral image classification based on KNN. *Sensing and Imaging*, *17*(1). https://doi.org/10.1007/s11220-015-0126-z

[8] L. Gong, S. Xie, Y. Zhang, Y. Xiong, X. Wang and J. Li, "A Robust Feature Extraction Method for Sound Signals Based on Gabor and MFCC," *2022 6th International Conference on*

Communication and Information Systems (ICCIS), Chongqing, China, 2022, pp. 49-55, doi: 10.1109/ICCIS56375.2022.9998146.

[9] Chi, Y., Cai, C., Ren, J., Xue, Y., & Zhang, N. (2023). Damage location diagnosis of frame structure based on wavelet denoising and convolution neural network implanted with Inception module and LSTM. *Structural Health Monitoring*, *23*(1), 57–76. https://doi.org/10.1177/14759217231163777

[10] Bonet-Solà, D., & Alsina-Pagès, R. M. (2021). A comparative survey of feature extraction and machine learning methods in diverse acoustic environments. *Sensors*, *21*(4), 1274. https://doi.org/10.3390/s21041274

[11] Andono, P.N.; Shidik, G.F.; Prabowo, D.P.; Yanuarsari, D.H.; Sari, Y.; Pramunendar, R.A. Feature Selection on Gammatone Cepstral Coefficients for Bird Voice Classification Using Particle Swarm Optimization. *IJIES* 2023, *16*, 254–264.

[12] Xie, S., Lu, J., Liu, J., Zhang, Y., Lv, D., Chen, X., & Zhao, Y. (2022). Multi-view features fusion for birdsong classification. *Ecological Informatics*, *72*, 101893. https://doi.org/10.1016/j.ecoinf.2022.101893

[13] Peng, C.; Zhang, Y.; Lv, D.; Lv, D.; Huang, Q.; Xiong, Y. FGS-Net: A Mutual Feature Fusion Residual Network Based on FBank and Gabor-Sinc Spectrogram for Birdsong Classification. In Proceedings of the 2024 10th International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 639–644

[14] Wu, J., Li, P., Wang, Y., Lan, Q., Xiao, W., & Wang, Z. (2023). VFR: The Underwater Acoustic Target Recognition Using Cross-Domain Pre-Training with FBank Fusion Features. *Journal of Marine Science and Engineering*, *11*(2), 263. https://doi.org/10.3390/jmse11020263

[15] Kahl S, Wilhelm-Stein T, Hussein H, Klinck H, Kowerko D, Ritter M, et al. Large-Scale Bird Sound Clas-

[16] Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., & Ferres, J. L. (2021). Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific Reports*, *11*(1), 17085. https://doi.org/10.1038/s41598-021-96446-w

[17] https://www.kaggle.com/datasets/gpiosenka/100-bird-species.

[18] V. Subrahmanyam, P. Sudhakar, B. M. Rao, V. J. Ramakrishna, M. K. Vemula, "An Innovative Approach for Detecting the Freshness of Fruits and Vegetables using Transfer Learning Techniques," *2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Tirupur, India, 2025, pp. 999-1005, doi: 10.1109/ICIMIA67127.2025.11200855.