E-CNN: Ensemble based Convolutional Neural Network for Animal Species Identification

Kuppusamy P*
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh
drpkscse@gmail.com

Naga Chaitanya C.A
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh
nagachaitanyaca07@gmail.com

***Abstract*—As emerging of technology and increasing population requires the urbanization in the most of the countries. The urbanization leads to destroy the large trees, plants and bushes that provide the shelter, food and living ecosystem. Due to this urbanization changes, various animal species are in need of searching the new shelter. Hence, the forest management should identify the various animal species using unmanned aerial devices, and transfer them into right living shelter. This study proposed the Ensemble Convolutional Neural Network (E-CNN) for detecting and classifying various animal species. The modified light weight MobileNetV2 proposed to classify the various animal species. Four CNN models InceptionResNetV2, DenseNet201, and MobileNetV2 outputs are combined and applied the ensemble method to improve the better performance. The transfer learning is applied in top layers and few convolution layers to improve the training time and accuracy. The proposed model achieved the accuracy of 98.2% in classifying the various 10 animal species. This research considers 10 animal species from the entire dataset that contains 90 species. The ablation study proved that the ensemble learning performs better comparing to the individual CNN model.**

Keywords— Animal species, Biodiversity conservation; CNN, ecosystem, image classification.

I. Introduction

Biodiversity is crucial for maintaining life on Earth, in particular in growing countries facing restrained assets and standard challenging situations. Extreme climate and weather crisis are threatening factors to biodiversity. The biodiversity is a primary requirement for healthy long life of human. Moreover, biodiversity enriches ecosystems, purifies water, and gives natural fortification against to floods and storms. Additionally, the numerous species contribute to weather and climate preservation in terms of seeds spreading and preserves the sensitive ecosystems [1].

In 2017, a number of animal attacks were reported within the United States, affecting millions of people every year. The incidents included various types of animal bites, reports suggesting about 2 million cases each year. The frequency of these attacks varied depending the territory, indicating that certain regions were more susceptible to such animal attacks. In popular parlance, the term "Man-Eater" is used to describe species known to actively prey on humans. In these, elephants are associated with higher mortality rates compared to other animals. Understanding the risks associated with animal attacks and taking precautions is important, especially in areas where potentially dangerous animals and populations coexist [2, 3].

This research is focused on ensemble based animal image classification using four CNN models named Inception ResNetv2, DenseNet201, MobileNetV2 with transfer learning [4]. These models were used to extract relevant information from a dataset consisting of 12000 images representing 6 different animal species [15].

A. Contribution of the research

The objective of this research is to design a majority based ensemble classifier that categorizing the various species from the dataset. This study included a detailed investigation and analysis of various deep learning algorithms such as DenseNet201, InceptionResNetV2, and MobileNetV2 in classifying the species. Also, explore the performance of the algorithms using transfer learning.

II. Related Work

In the domain of image classification, many researchers have extensively used techniques based on deep learning and convolutional neural networks for identifying the object classes, diseases, tumors using image datasets. However, there is a noticeable gap in research in identifying animal species using animal image datasets. In this section, we explored the existing approaches implementation to classify the objects, animals, vehicles, etc., using transfer learning.

The VGG19 is proposed for classifying the animals using muzzle and shape features. It shows commendable accuracy on a balanced data set, but a significant drop in accuracy for an unbalanced dataset. Specifically, the results show an accuracy of 94.1% on the balanced dataset, while achieving a lower accuracy of 54.8% on Unbalanced dataset [4-6]. The customized own dataset JONATHAN is prepared using aerial camera about marine bird. The customized seven layers CNN is proposed to detect the marine bird using automated feature extraction and classification through collected RGB image dataset. It achieved the accuracy of 95% in marine bird detection [7].

A wildlife recognition model is proposed Lite AlexNet to identify animal species with accuracy at 96.6% for detecting images consists animal, and 90.4% for recognizing almost three common species in the Wildlife Spotter imbalanced dataset of wild animal images [8]. The authors described the significance of machine learning and deep learning algorithms in ecological image processing to implement the segmentation, detection and classification of birds, fish, flowers, and leaves. This work highlighted the utilization of deep learning to extract non-linear features across diverse fields, including ecological, scientific and commercial applications [9].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The well-known deep learning architectures ResNet, and DenseNet proposed for classifying diseases identification from the plant leaves. These architectures proved with the accuracy of 99.54% and 97.34% respectively. The training time of these models is more due to not using the transfer learning in this research [10]. The ResNet has been used as optimal feature extractor to learn the animal behavior in time-series data. The student- teacher learning approach is utilized to share the knowledge of extracted features with Gated Recurrent Unit (GRU) to identify the animal behavior [16]. This study delved into the consequences of land use change particularly affecting forest ecosystems. The changes are affect and decline the bird species. The researchers conducted a comprehensive survey during the breeding season across six distinct land use types. Their findings revealed that all altered land uses experienced significant species loss when contrasted with forests comprised of natural oak trees [11].

The comprehensive assessment report on biodiversity and ecosystem services offered an evidence-based analysis of the current state of global biodiversity. It deeply investigates the intricate factors contributing to biodiversity loss and sheds light on its far-reaching impacts, encompassing effects on food security, health, and livelihoods worldwide. This also offers recommendations on how to increase coverage and accurately identify potential risks and positive scenarios of environmental degradation Further efforts will overcome information gaps, local barriers, addressing implementation challenges and other key global issues, creating opportunities for growth [12]. The study emphasized the importance of automated recognition system developed using the Deep Neural Networks (DNNs) to recognize the species and categorize the animal's behavior with camera-trap images, sound recordings data [14].

III. CNN Methodology for Pharmaceutical Drug Classification

This study aims to investigate the significance of transfer learning in ensemble based DNNs to categorize the animal species. The transfer learning utilizes the pre-trained weights of the ImageNet for the improving the performance of the proposed Densenet201, InceptionResNetV2 and MobilenetV2. The convolutional layers received the input images and transfer learning models captured the distinctive features to understand the different animal species in the dataset. The proposed ensemble based transfer learning DNN architecture is shown in Fig. 1. It provides valuable insights of the framework design and implementation.

A. Data Pre-processing

To achieve maximum performance of proposed DNNs architecture, data preparation is very imperative stage since the data samples contain different size of images $such\ as\ 256 \times 256\ and\ 600 \times 500$. All images in the dataset are *resized* into consistent formats suitable for each *CNN* model. The images are resized to 224×224 pixels for DenseNet201, MobileNetV2 while InceptionResNetV2 require 229×229 pixels.

B. Transfer Learning Models

The transfer learning models selected for this study such as MobileNetV2. InceptionResNetV2 and Densenet201 due to their extensive usage and results proved an excellent performance for diversified tasks. The base layers are frozed by assigning ImageNet values for the weight and bias parameters. The top layers are replaced for learning the texture, color, edges, and pattern of the input images for different CNN models. The transfer learning decreases the training time since frozen layers are not trained, and fully connected layers are trained to learn the patterns. Although DenseNet201 requires longer training time due to denser connections, it proffers better overall accuracy. The MobileeNetV2 is proposed in this study by considering its number of trainable parameters that proved as a compact model with faster processing without significant loss.

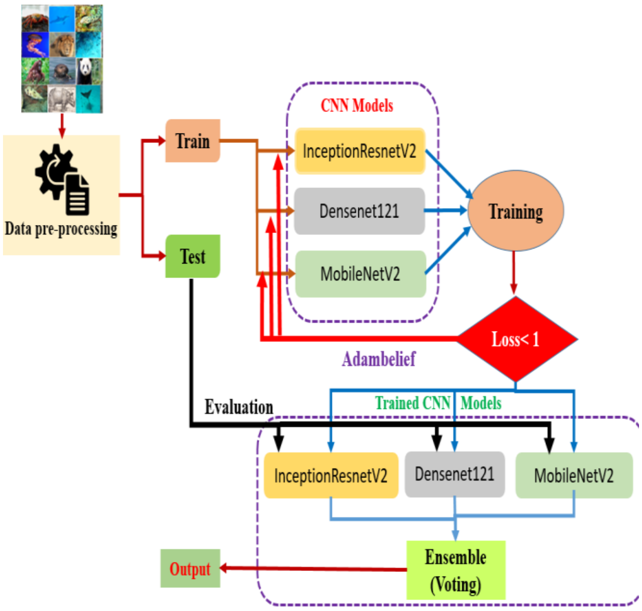

Fig.1 . Ensemble based proposed system architecture

The Densenet201 is proposed due its specialized property reusing of the features during forward propagation. It is designed with 201 parameterized layers that contains 10 dense blocks, 6 transition layers and output layer. The dense blocks processing the previous layers feature maps that reuse the information extracted in previous layers. This reuse property diminishes the loss and maximize the accuracy. The transition layer applies the batch normalization to standardize the feature map values, and reduce the feature maps count using 1×1 convolution block. The dimension of feature maps diminished by average pooling approach. It controls the information flow and feature reuse between levels efficiently. It is computationally economical, processing massive volumes of visual input precisely. Because of its optimal processing power, it learns a wide range of discriminative characteristics, textures in object identification, image categorization, and segmentation [17].

This proposed InceptionResnetV2 is a powerful 164 layers DNN utilizes the special properties of inception module and dense module from Inception, and Resnet architectures respectively. Inception module facilitates the various scale filters to grasp the complex features with different size in parallel. It reduces the training time. Meanwhile the dense blocks facilitate the feature reuse and reduce the vanishing gradient [21]. It consists of symmetric and asymmetrical blocks, which combine a convolution, average aggregation, maximal aggregation, reduction, concatenation, and fully connected layers. These building blocks contribute to the model's capacity to grasp local and global features, improving model robustness and accuracy in image classification tasks. The MobileNetV2 is a highly efficient DNN with the inverted residual connections property. This architecture comprises inverted residual blocks, each consisting of a bottleneck layer, an expansion layer, and linear bottleneck layer. The bottleneck reduces computational cost, while the expansion layer increases the receptive field of the network. MobileNetV2 achieved an impressive performance with depth-wise separable convolutions, squeeze-and-excitation blocks. It contains less trainable parameters than other DNNs. Hence, MobileNetV2 is well-suited for deployment on resource-constrained devices and real-time image recognition applications [20].

The proposed voting ensemble model applies the majority voting method to enhance the performance of the algorithm. It received the classified label from three CNN classifiers as input. It selects majority number of the received classes j from various CNN classifiers i as an predicted output y' as given equation 1.

$$y' = argmax_{j\in classes} \sum_{i=1}^{n} output_{i,j} \qquad (1)$$

For example, consider the two CNN classifiers Densenet121, MobileNetV2 are predicted the given input as cow, but InceptionResnetV2 predicted as bull. The ensemble voting approach selects the majority values cow as a finalized predicted output. This voting ensemble approach increases the performance of the model using majority class labels.

IV. Dataset Description

The dataset comprises 2000 images per class that includes 6 different animal species such as Elephant, Gorilla, Hippo, Monkey, Tiger, and Zebra [15]. This work considered 12000 samples that is divided as training phase (8400), validation phase (1200), and testing phase (2400) with the ratio of 70%, 10%, and 20%. The sample images are RGB colored with size 256×256. The samples are pre-processed to fit for the proposed models as 224×224 for Densnet201, MobileNetV2 and as 299×299 for InceptionResNetV2 algorithm. The output labels are encoded in the range from 0 to 5.

V. Results and Discussion

The experiments have been executed and results are generated with the accuracy delivered by an individual model. The metrics computed based on the correct positive predictions (TP), wrongly predicted as positives (FP), correct negative predictions (TN), and wrongly predicted as negatives as the following equations:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \qquad (2)$$

The total loss is estimated by adding the wrong classification of the samples FP and FN. The experiment has been configured using the various parameters given in Table I.

Table I. Experimental configuration

Description	**Values**
Learning_step_rate	0.01
Maximum_Epochs	40
Batch size	16
Earlystopping	Patience_p=3
Loss	Categorical_cross entropy
Optimizer	Adabelief
Layers_removed for transfer learning	Fully_connected and classification layers
Pre-trained weights	ImageNet_weights
FC_layers units	512
Output_classes	6

The models were evaluated based on various factors, all of which were designed using the same parameters to guarantee an equitable comparison of their performance. Among the models considered, MobileNetV2 showcased exceptional efficiency, requiring less training time and boasting a small model size. However, when it comes to overall performance, DenseNet201 took the lead as the top performer, achieving the highest test accuracy and demonstrating a comparatively lower test loss. InceptionResNetV2 also exhibited solid performance, scoring well in the test accuracy and loss metrics.

The utilization of early stopping as a strategy to prevent overfitting during model training. This entails overseeing the model's performance on the validation dataset and halting training if there is no progress over a predefined number of epochs (patience).This prevents the model from storing noise in the training data and promotes better generalization to unseen data. An early stop feature with a patience period of 3 epochs, the practice will stop if no improvement is seen for 3 consecutive epochs. This saves computational resources and ensures that the model is saved at the optimal validation performance point. Early `stopping improves model stability and performance during deep learning.

The training duration corresponds to the duration required for the alogorithm or model to ingest training data and fine tune its parameters (biases and weights) to optimize the loss function.Time required to train can vary depending on a number of factors, including the intricacy of the model's architecture, the scale of the training data, accessible computational capabilities, and the hyperparameters employed during training.

Fig. 2 showed the loss varying while increasing the ephochs. The training loss is initiated with 1.42, validation loss with 0.76. The loss difference is gradually varying step by step by increasing the training epochs. The model generated optimal results for the configured parameters.

A. *Discussion*

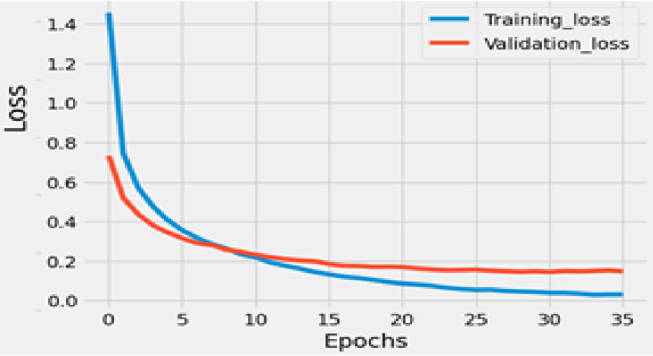

Fig. 2. Loss variation based on epochs

The Fig. 3 presented the accuracy of the proposed DNN that increase while incrementing the epochs till 35. The difference amidst the training accuracy and validation accuracy also less. It proved the reduction of overfitting by propoised model.

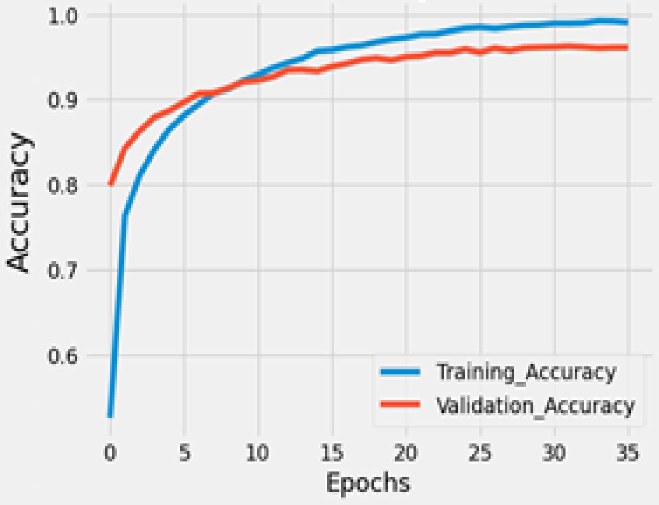

Fig. 3. Accuracy increases based on epochs

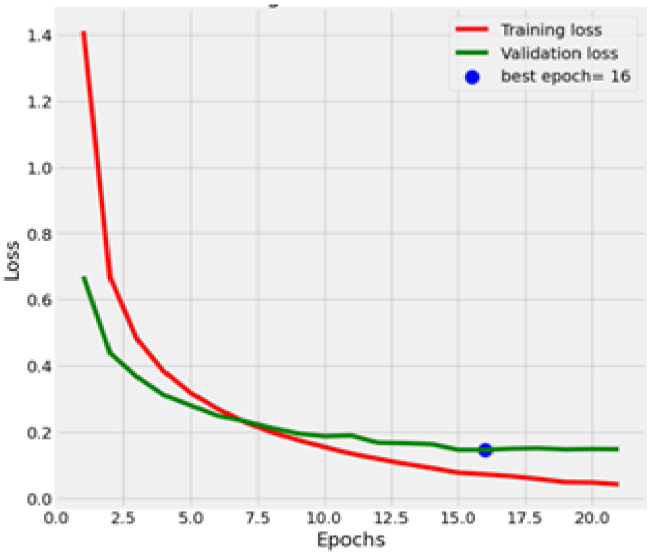

Fig. 4. Loss variation based on epochs

Fig. 4 presented the training and validation loss decerased while increased the ephochs. The training loss is initiated with 1.4, and validation loss with 0.68. The loss difference is gradually varying step by step by increasing the training epochs. The model generated optimal results and the training is halted at the epoch 20 before reaching 40. The Fig. 5 illustrated the accuracy of the proposed DNN that is improved while incrementing the epochs. The difference amidst the training accuracy and validation accuracy also less. It proved the reduction of overfitting by propoised model and selected 21st epoch as best epoch.

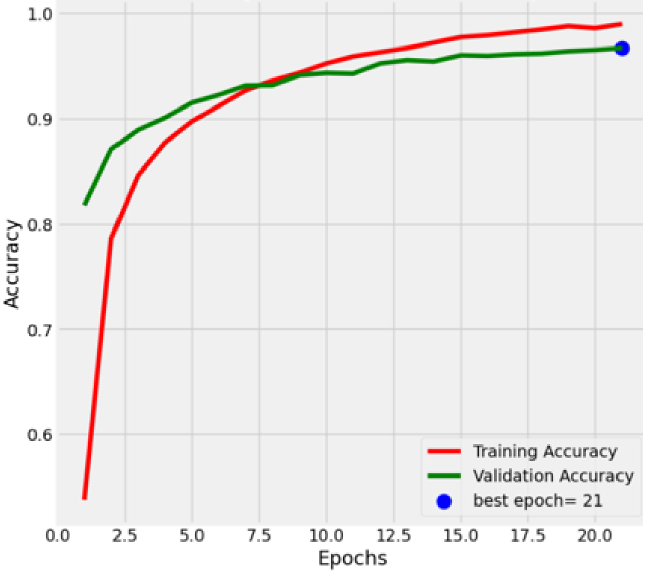

Fig. 5. Accuracy increases based on epochs

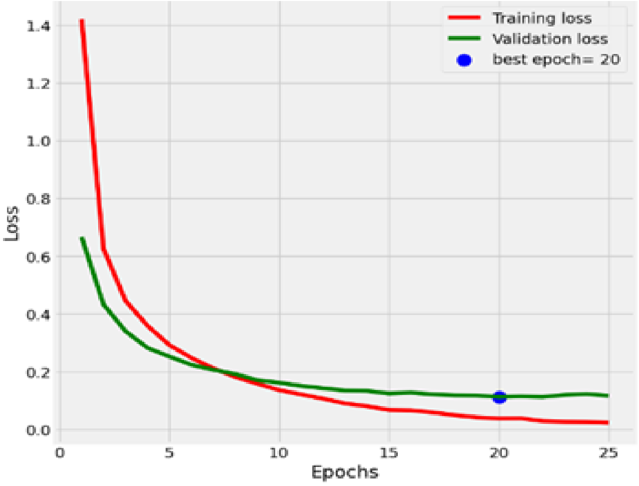

Fig. 6. Loss variation based on epochs

Fig. 6 illustrated the loss of propoised DNN that is decreaseing by varying the ephochs incrementally. The training loss is originated with 1.42, and validation loss with 0.67. The loss difference is gradually varying and model generated optimal results for the configured parameters at best epoch 20. The Fig. 7 presented the accuracy of the proposed DNN that increase while incrementing the epochs till 35. The difference amidst the training accuracy and validation accuracy also less. The best epoch is selected as 20 due it provides optimal accuracy.

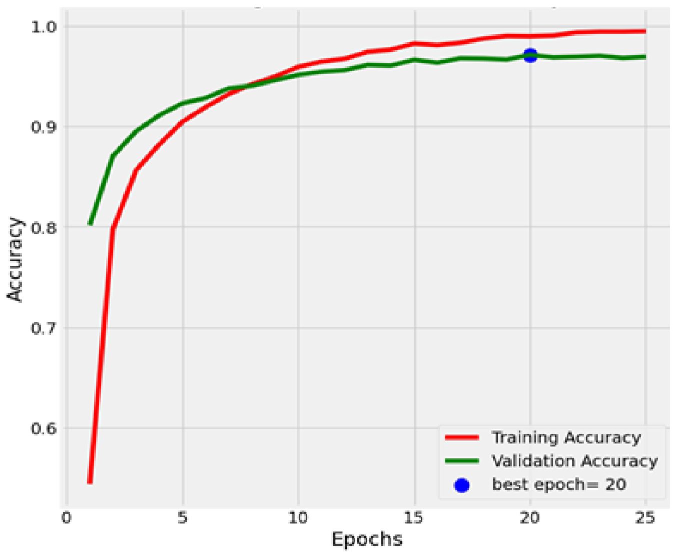

Fig. 7. Accuracy increases based on epochs

Fig. 8 shows the predicted output of the animal species by proposed ensemble based architecture. The animals cow and elephant are predicted with confidence score 1.0.

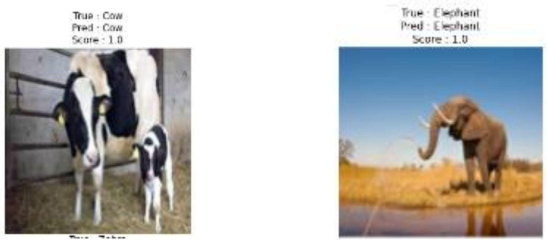

Fig. 8 Prediction of animal species

Table 2 displays the results of various deep learning architectures for image classification tasks on an animal dataset, considering different input sizes and pre-trained freezing layers. Inception-ResNetV2 achieved a promising training accuracy of 98.7% while DenseNet201 achieved the lowest test accuracy of 98.01%. MobileNetV2, with only 20 unfreezed layers, obtained a training accuracy of 98.5% and demonstrated with low accuracy. Overall, Inception-ResNetV2 exhibited the best performance among the architectures, showing higher training and validation accuracy and lower test loss. Further fine-tuning and optimization of the models could be explored to improve the values, resulting in more robust and accurate image classification models.

Table II. Performance analysis of DNNs

Model	Image	Train Accuracy (%)	Validation Accuracy (%)
Inception-ResNetV2	229 × 229 × 3	98.7	97.8
MobileNetV2	224 × 224 × 3	98.5	95.7
DenseNet201	224 × 224 × 3	98.01	95
Ensemble Method	-	**99.1**	**98.2**

Inception-ResNetV2's validation accuracy is steadily increasing, indicating that this model is continuously improving its ability to accurately classify images over time. This steady growth shows that the model is not overfit to the training data. On the other hand, the validation accuracy of other models is also increasing, but they do not show consistent progress like Inception-ResNetV2. This may indicate that these models are not well trained or tend to overfit the training data. For example, MobileNetV2 shows a faster growth rate than Inception-ResNetV2, but it shows signs of more trainig time, implying that it is a more complex model capable of achieving higher accuracy and is at risk over-matched.

Inception-ResNetV2 exhibits less variability in its validation accuracy than other models, suggesting that it is less sensitive to changes in training data. This robustness suggests that Inception-ResNetV2 is less likely to overfit the training data.

The average training time of the 3 architectures shows that MobileNetV2 is the light weight model, with an average training time of 603 seconds per epoch. On the other hand, DenseNet201 requires an average training time of 363 seconds per epoch. This indicates that MobileNetV2 requires more time for training compared to DenseNet201. The shorter training time of DenseNet201 makes it computationally more efficient and faster to train compared to MobileNetV2. However, it is essential to consider this difference in training times along with other factors, such as validation accuracy and potential overfitting, when choosing the appropriate model for image classification tasks, especially in resource-constrained environments.

All 3 models performed similarly on the training data, achieving high accuracy. However, they exhibit different patterns of oscillation upon confirmation. InceptionResNetV2 has the most stable commit accuracy, while MobileNetV2 has the most volatile commit accuracy. This suggests that InceptionResNetV2 is a more robust model with less overfit. InceptionResNetV2 can be considered a more suitable choice than DenseNet201 for image classification tasks. Although DenseNet201 can provide optimal accuracy, training is computationally more expensive and more complex. On the other hand, MobileNetV2's training efficiency, too low matching probability and high accuracy make it a viable choice for an image classification needs.

The ensemble learner voting method is proposed to improve the performance that received the output of all three models and takes the majority of these three classes. It provides a significant trival improvement in the results to enahace the performance.

VI. Conclusion

In summary, InceptionResNetV2 appears as the most favorable model for classifying animal images with different classes, counts and resolutions. It exhibits consistent growth in validation accuracy without overfitting, exhibits robustness with lower sensitivity to changes in training data, and requires less time. more training time per epoch than other models. Although DenseNet201 offers higher accuracy, it comes with increased computational complexity and cost during training. For future work, in addition to the points mentioned earlier, it is possible to explore improvements in accuracy by refining the decision thresholds of the model and incorporating techniques such as class balancing. By prioritizing accuracy, this study aims to ensure that the model's predictions on the training data are more precise and trustworthy, ultimately reducing overfitting and improving its performance on unseen data. These efforts will contribute to the development of a more robust and general model for image classification tasks. In addition, applying the ensemle learner improves the accualcy significantly showed the impact of ensembling appraoch. This work would be extended by adding the optimization appraoches and vision transformers to improve the performance in real-time applcations.

References

[1] https://news.un.org/en/story/2022/11/1130677

[2] Binta Islam S, Valles D, Hibbitts TJ, Ryberg WA, Walkup DK, Forstner MRJ. Animal Species Recognition with Deep Convolutional Neural Networks from Ecological Camera Trap Images. *Animals*. 2023; 13(9):1526. https://doi.org/10.3390/ani13091526

[3] Li, Hao-Xuan, et al. "An automatic identification method of common species based on ensemble learning." *Ecological Informatics* (2025): 103046.

[4] Nanni, L., Costa, Y.M.G., Aguiar, R.L. *et al.* Ensemble of convolutional neural networks to improve animal audio classification. *J AUDIO SPEECH MUSIC PROC.* **2020**, 8 (2020). https://doi.org/10.1186/s13636-020-00175-3

[5] Ahumada, Jorge A., Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G. O'Brien, Jonathan Palmer et al. "Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet." Environmental Conservation 47, no. 1 (2020): 1-6.

[6] Favorskaya, M. and Pakhirka, A., 2019. Animal species recognition in the wildlife based on muzzle and shape features using joint CNN. Procedia Computer Science, 159, pp.933-942.

[7] Lynda Ben Boudaoud, Fr´ed´eric Maussang, Ren´e Garello, and Alexis Chevallier. Marine bird detection based on deep learning using highresolution aerial images. In OCEANS 2019-Marseille, pages 1–7. IEEE, 2019.

[8] Nguyen, Hung, Sarah J. Maclagan, Tu Dinh Nguyen, Thin Nguyen, Paul Flemons, Kylie Andrews, Euan G. Ritchie, and Dinh Phung. "Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring." In 2017 IEEE international conference on data science and advanced Analytics (DSAA), pp. 40-49. IEEE, 2017.

[9] Guo, Q.H.; Jin, S.C.; Li, M.; Yang, Q.L.; Xu, K.X.; Ju, Y.Z.; Zhang, J.; Xuan, J.; Liu, J.; Su, Y.J.; et al. Application of deep learning in ecological resource research: Theories, methods, and challenges. Sci. China Earth Sci. 2020, 63, 1457–1474

[10] Sadhu, Vamsi Suhas, et al. "Revitalizing Plant Disease Detection using Optimized DNNs for Enhanced Leaf Images." *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*. IEEE, 2023.

[11] Shahabuddin G, Goswami R, Krishnadas M, Menon T. Decline in forest bird species and guilds due to land use change in the Western Himalaya. Global Ecology and Conservation. 2021 Jan 1;25:e01447.

[12] Díaz, S., Settele, J., Brondízio, E., Ngo, H., Gueze, M., Agard, J., Arneth, A., Balvanera, P., Brauman, K., Butchart, S., 2020. Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES).

[13] Sylvain Christin, Eric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. ´ Methods in Ecology and Evolution, 10(10):1632–1644, 2019.

[14] Nguyen, Hung, et al. "Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring." 2017 IEEE international conference on data science and advanced Analytics(DSAA). IEEE, 2017.

[15] https://www.kaggle.com/datasets/utkarshsaxenadn/animal-image-classification-dataset

[16] Arablouei, Reza, et al. "In-situ animal behavior classification using knowledge distillation and fixed-point quantization." Smart Agricultural Technology, 4, 100159, 2023.