Oregon State University

COLLEGE OF ENGINEERING | School of Electrical Engineering and Computer Science

Enhancing Robustness in Bird Species Classification through Adversarial Training

Chelsi Jain, Shubhankar Joshi, Shrirang Patil

Deep Learning Final Project

Winter 2024

Problem

Objective: Develop robust and reliable bird species classification models resilient to adversarial attacks and maintain high accuracy in real-world settings.

Key Points:

- ➔ Vulnerability of models to adversarial attacks.
- ➔ Effectiveness of adversarial training in enhancing model robustness.
- ➔ Comparative analysis of different models (InceptionV3, EfficientNetB3, WideResNet50) and attack scenarios (FGSM, BIM, PGD).

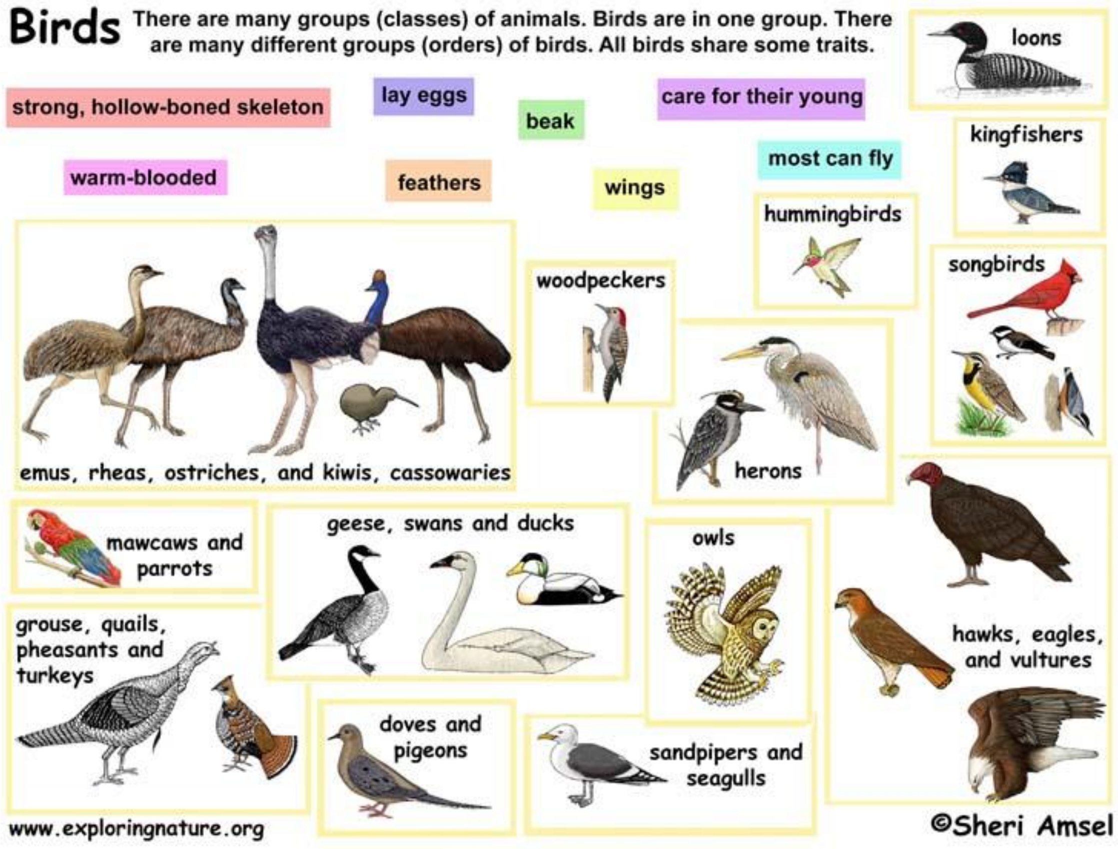

Birds
There are many groups (classes) of animals. Birds are in one group. There are many different groups (orders) of birds. All birds share some traits.
strong, hollow-boned skeleton
lay eggs
beak
care for their young
loons
warm-blooded
feathers
wings
most can fly
kingfishers
hummingbirds
songbirds
woodpeckers
emus, rheas, ostriches, and kiwis, cassowaries
herons
mawcaws and parrots
geese, swans and ducks
owls
grouse, quails, pheasants and turkeys
hawks, eagles, and vultures
doves and pigeons
sandpipers and seagulls
www.exploringnature.org
©Sheri Amsel

Fig 1: Exploring Nature Bird Classification

Dataset

Oregon State University
College of Engineering

525 bird species, with a total of **84,635** training images. Each species have 5 test images and 5 validation images, making up **2,625** images for each category.

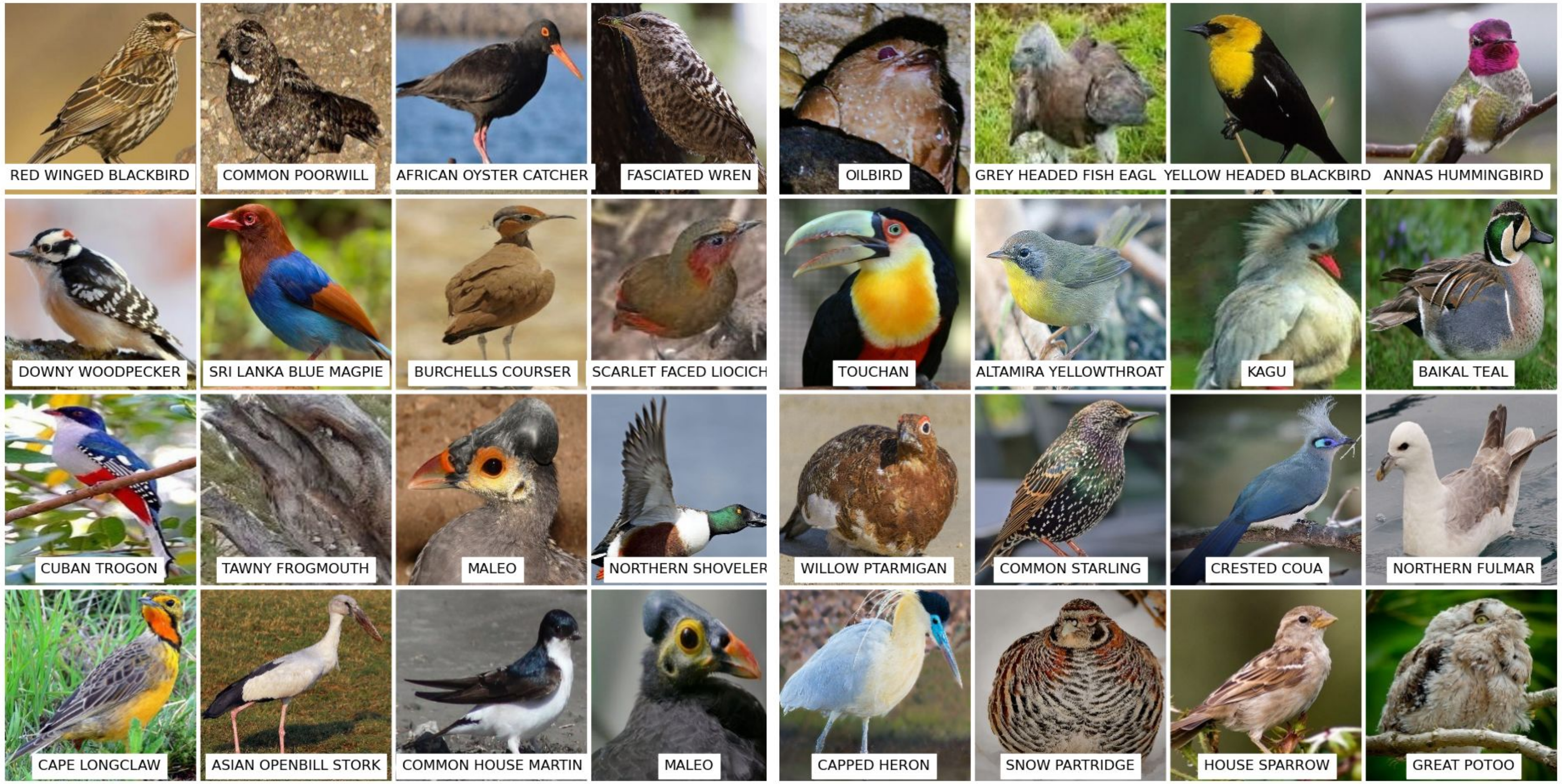
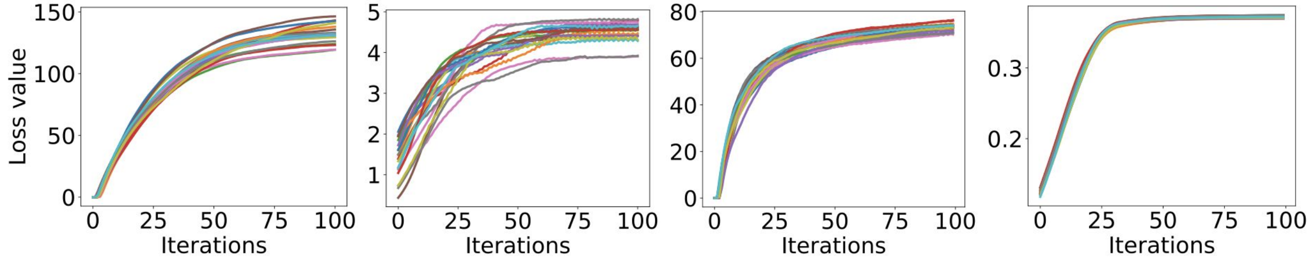


Fig 2: Birds 525 Species - Image Classification

Related Work

- A CNN model classifies 525 bird species with 87% validation and 86.7% test accuracy using transfer learning and data augmentation.[1]
- discusses CNN architectures to enhance image classification performance on various datasets.[2]
- adversarial examples can fool classifiers even after physical transformations like printing and photographing, with simple methods proving more robust.[3]
- a saddle point formulation to train robust deep neural networks against adversarial attacks, which can be efficiently solved using first-order methods.[4]



(a) MNIST
Standard training

(b) MNIST
Adversarial training

(c) CIFAR10
Natural training

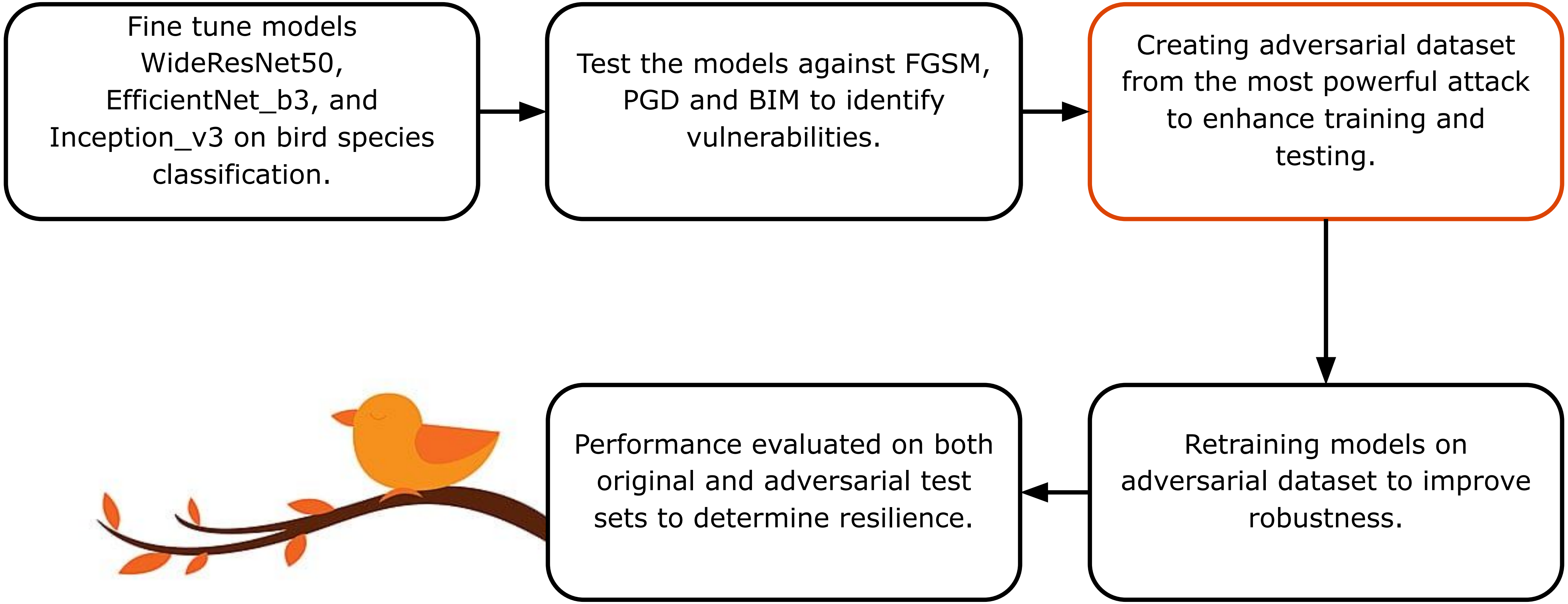
(d) CIFAR10
Adversarial training

Fig 3: Natural and adversarial training on MNIST & CIFAR10 [4]

[1] Cai, R. (2023). Automating bird species classification: A deep learning approach with CNNs. *Journal of Physics: Conference Series, 2664*.
[2] Nguyen, A.H., & Pham, M.T. (2023). Enhancing Convolutional Neural Network Architectures with Long Short-Term Memory for Improved Image Classification. *2023 8th International Scientific Conference on Applying New Technology in Green Buildings (ATiGB)*, 227-232.
[3] Kurakin, A., Goodfellow, I.J., & Bengio, S. (2016). Adversarial examples in the physical world. *ArXiv, abs/1607.02533*.
[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv, abs/1706.06083*.

Proposed Solution

Our previous method:



Fine tune models WideResNet50, EfficientNet_b3, and Inception_v3 on bird species classification.

Test the models against FGSM, PGD and BIM to identify vulnerabilities.

Creating adversarial dataset from the most powerful attack to enhance training and testing.

Retraining models on adversarial dataset to improve robustness.

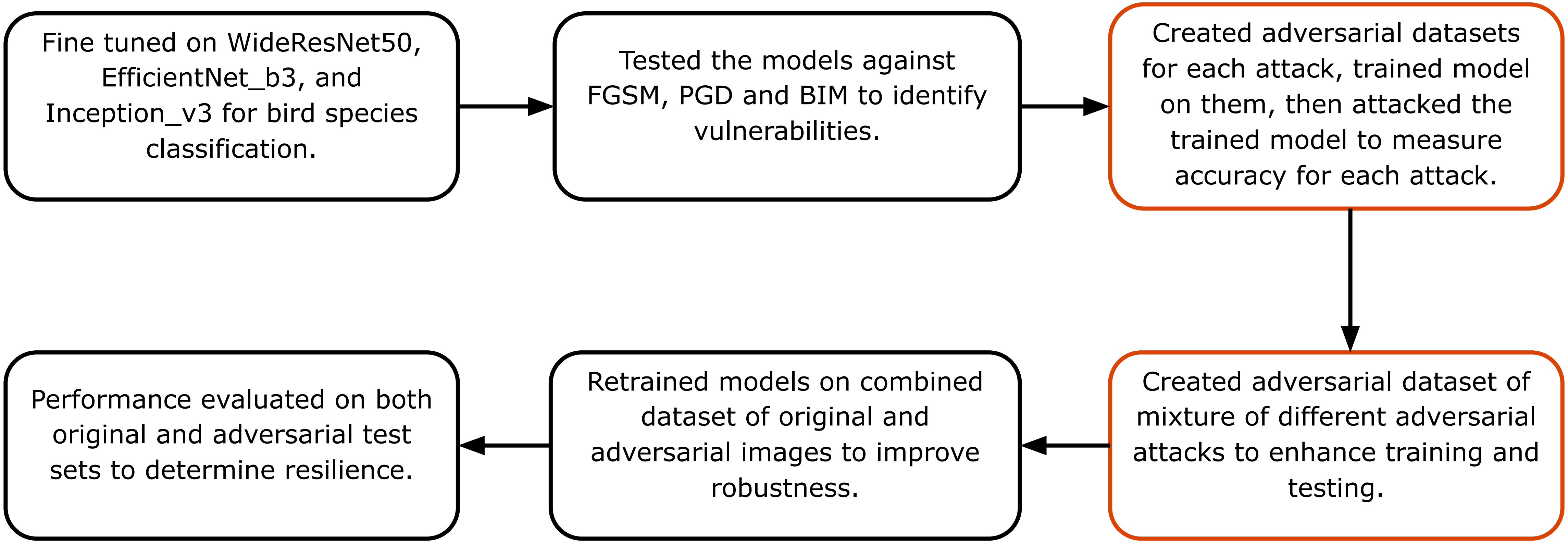
Performance evaluated on both original and adversarial test sets to determine resilience.

Fig 4: Block diagram of previous method

Current Approach

Oregon State University
College of Engineering

Our new method:



Fine tuned on WideResNet50, EfficientNet_b3, and Inception_v3 for bird species classification.

Tested the models against FGSM, PGD and BIM to identify vulnerabilities.

Created adversarial datasets for each attack, trained model on them, then attacked the trained model to measure accuracy for each attack.

Created adversarial dataset of mixture of different adversarial attacks to enhance training and testing.

Retrained models on combined dataset of original and adversarial images to improve robustness.

Performance evaluated on both original and adversarial test sets to determine resilience.

Fig 5: Block diagram of new method

Research Questions

RQ1: Does adversarial training improve the robustness of neural network models against adversarial attacks in automated bird species classification?

RQ2: Does the integration of a combined dataset, encompassing examples from FGSM, PGD, and BIM adversarial attacks, affect the accuracy of the models?

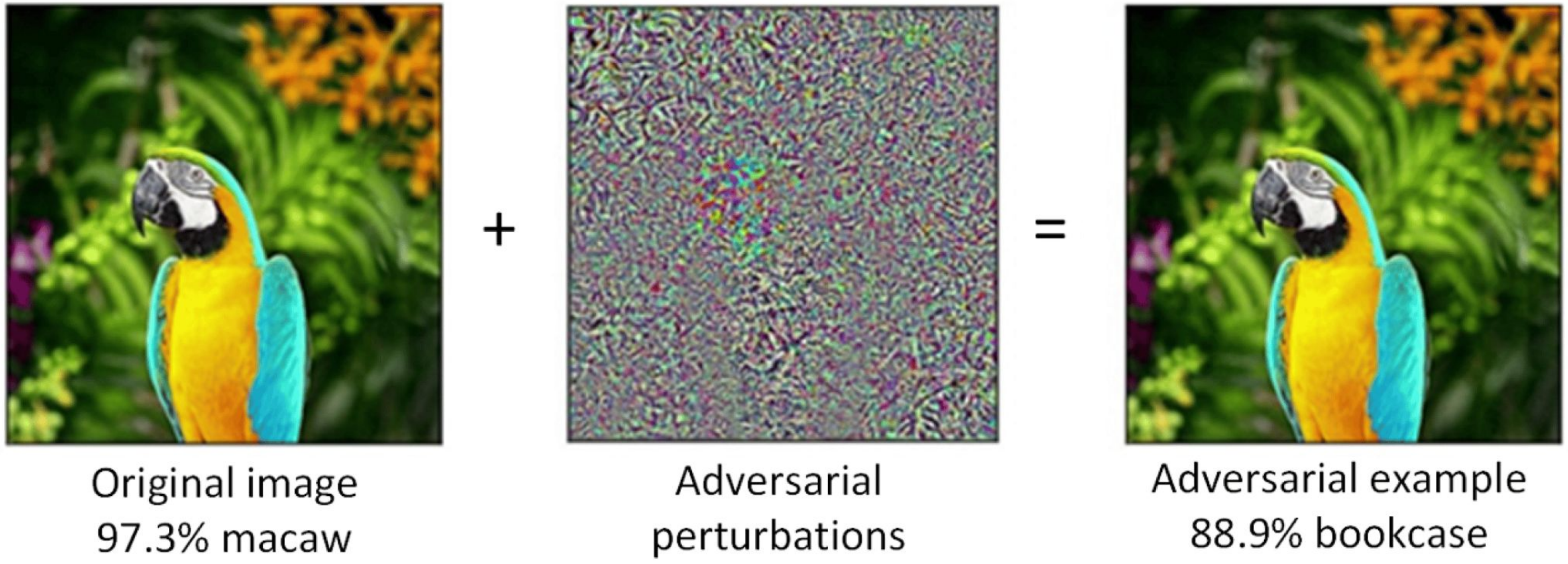


Fig 6: DNN provides the wrong prediction with high confidence by adding imperceptible perturbations to the original image [5]

[5] Shi, Y., Fan, C., Zou, L., Sun, C., & Liu, Y. (2020). Unsupervised Adversarial Defense through Tandem Deep Image Priors. *Electronics*.

Adversarial images

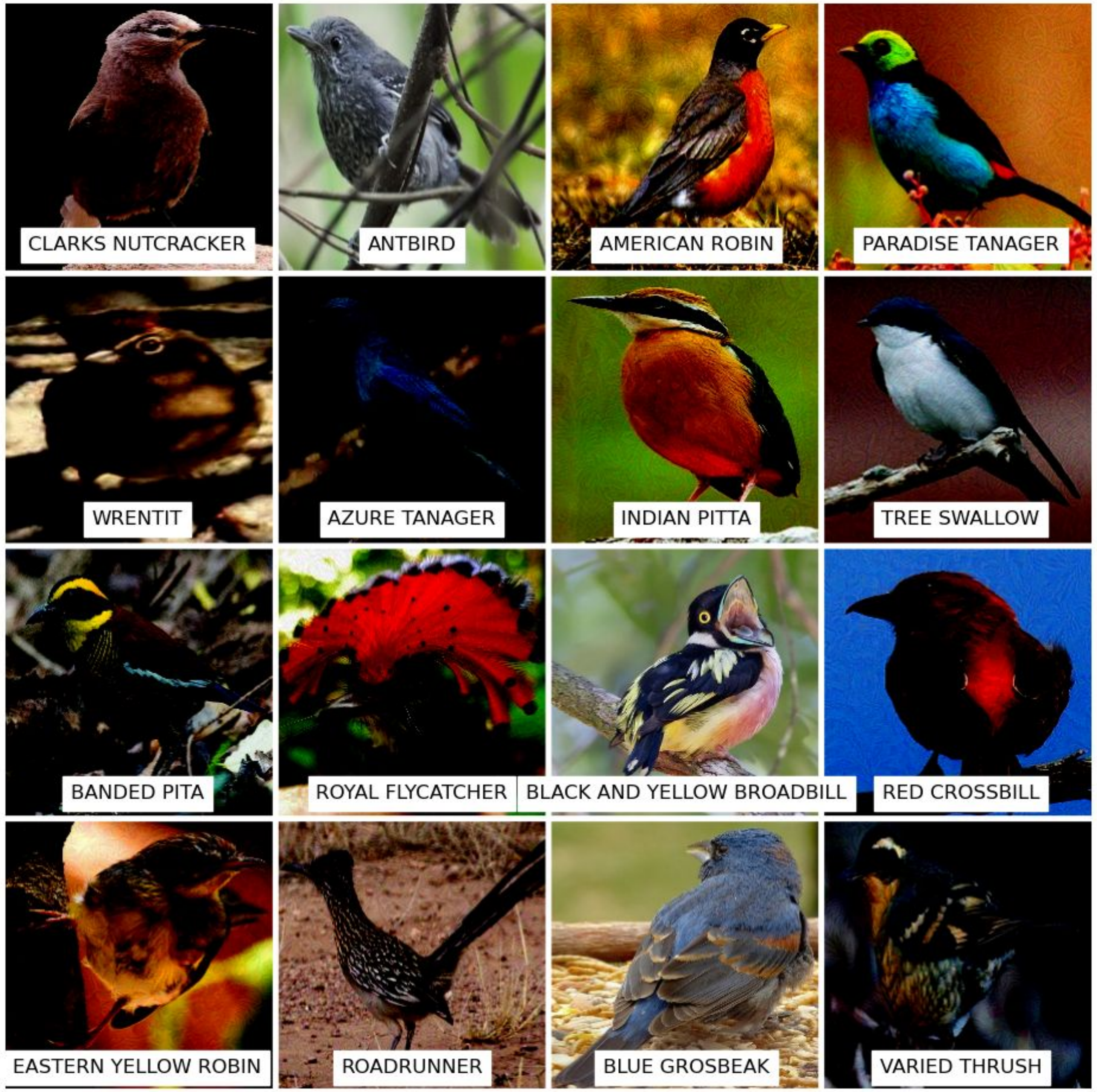
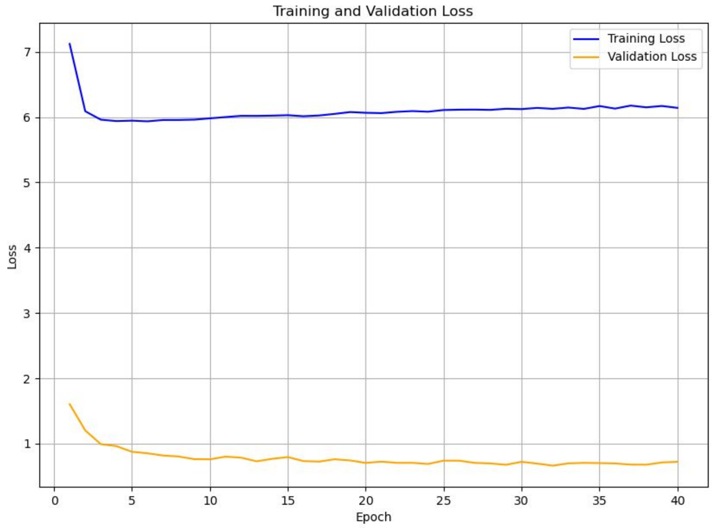


Fig 7: Adversarial images from Bird 525 Image Classification

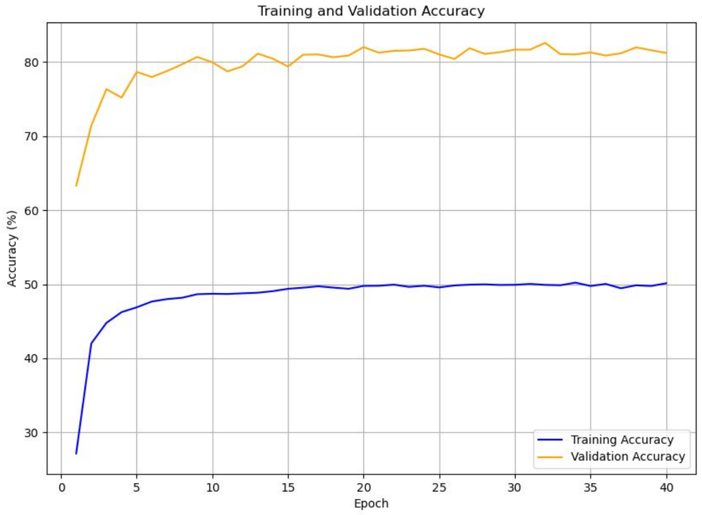
Results & Evaluation

Oregon State University College of Engineering

- InceptionV3 (Before adding adversarial images)



Plot 1



Plot 2

Precision	82.28%
Recall	81.83 %
F1 Score	82.07 %

Table 1

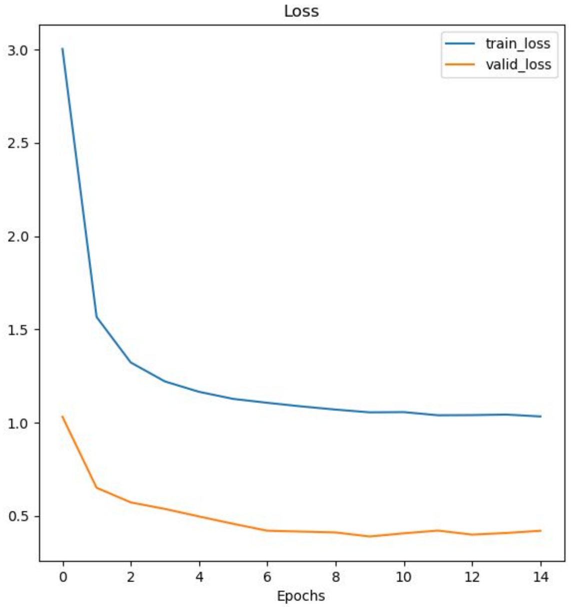
Attack	**FGSM**	**PGD**	**BIM**
Accuracy	11.04%	0.0%	0.0%

Table 2

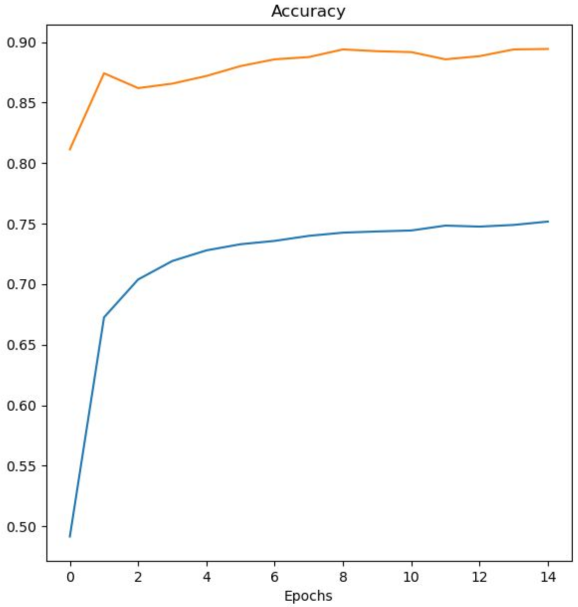
Results & Evaluation

Oregon State University
College of Engineering

- EfficientNetB3 (Before adding adversarial images)



Plot 3



Plot 4

Precision	94.79 %
Recall	93.83 %
F1 Score	93.55 %

Table 3

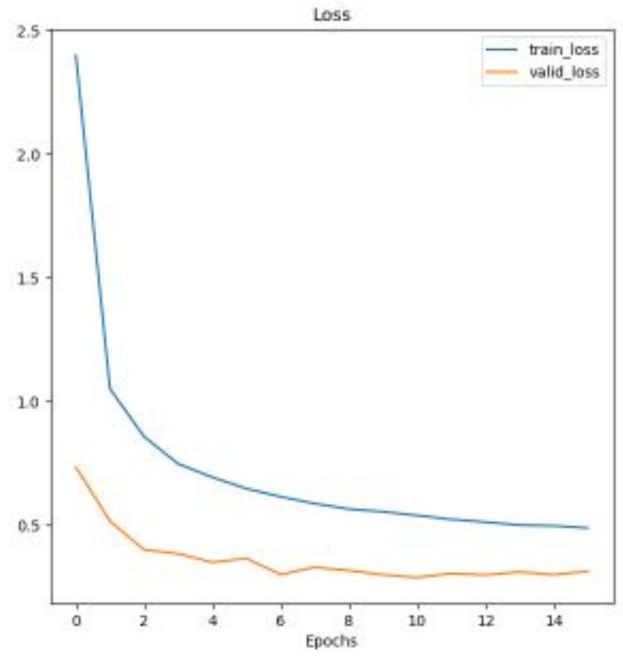
Attack	**FGSM**	**PGD**	**BIM**
Accuracy	30.4%	0.0%	0.419%

Table 4

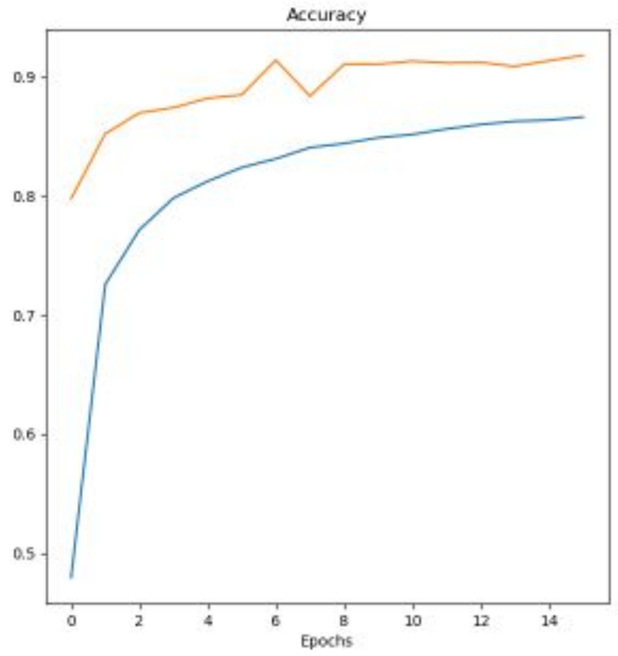
Results & Evaluation

Oregon State University
College of Engineering

- WideResNet50 (Before adding adversarial images)



Plot 5



Plot 6

Precision	97.21 %
Recall	96.87 %
F1 Score	96.81 %

Table 5

Attack	**FGSM**	**PGD**	**BIM**
Accuracy	33.45%	0.0%	10.12%

Table 6

Results & Evaluation

Oregon State University
College of Engineering

		InceptionV3 (FGSM)	EfficientNetB3 (BIM)	WideResNet (PGD)
Robust Accuracy	Normal	87.66 %	98.93 %	98.70 %
	FGSM	56.55 %	72.0 %	66.36 %
	PGD	0.0 %	1.21 %	13.88 %
	BIM	0.0 %	21.25 %	71.88 %
Precision		90.83 %	99.15 %	98.95 %
Recall		87.66 %	98.93 %	98.70 %
F1 Score		87.40 %	98.91 %	98.66 %

Table 7

		WideResNet
Robust Accuracy	Normal	98.17 %
	FGSM	69.96 %
	PGD	17.54 %
	BIM	78.78 %
Precision		98.53 %
Recall		98.17 %
F1 Score		98.14 %

Table 8

Discussion/Conclusion

- Automated bird species classification is crucial for ecology and conservation, but vulnerable to adversarial attacks. Adversarial training enhances model robustness.

- Through rigorous testing against FGSM, PGD, and BIM attacks, and by training on a combined dataset of original and adversarial images, we have enhanced the accuracy and resilience of models like WideResNet50, EfficientNet_b3, and Inception_v3.

- Adversarial dataset training significantly enhances model performance and resilience, with combined datasets further improving generalization capabilities.

- Our research highlights the critical need to address adversarial threats in deep learning development, fostering robust model enhancement across various fields.

Enhancing Robustness in Bird Species Classification through Adversarial Training

Thank You

Chelsi Jain, Shubhankar Joshi, Shrirang Patil

Deep Learning Final Project

Winter 2024