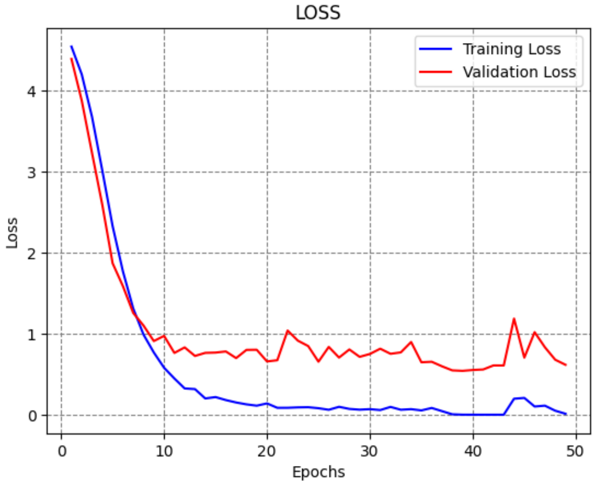
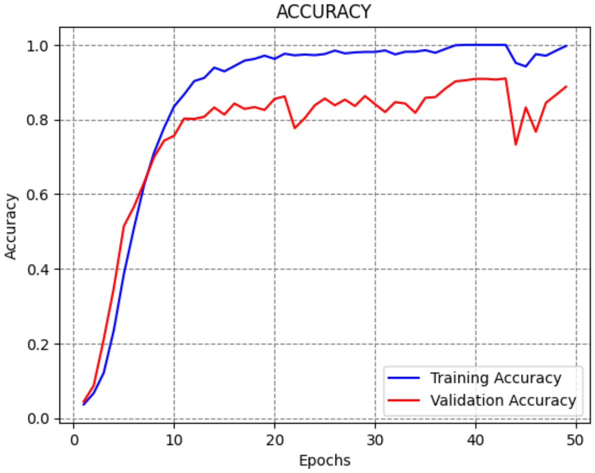
Task-1

a) Classification accuracy Table-

Model	Train set	Validation Set	Test set
Vgg16	0.982	0.834	0.873

b) The training and validation loss and accuracy curves-



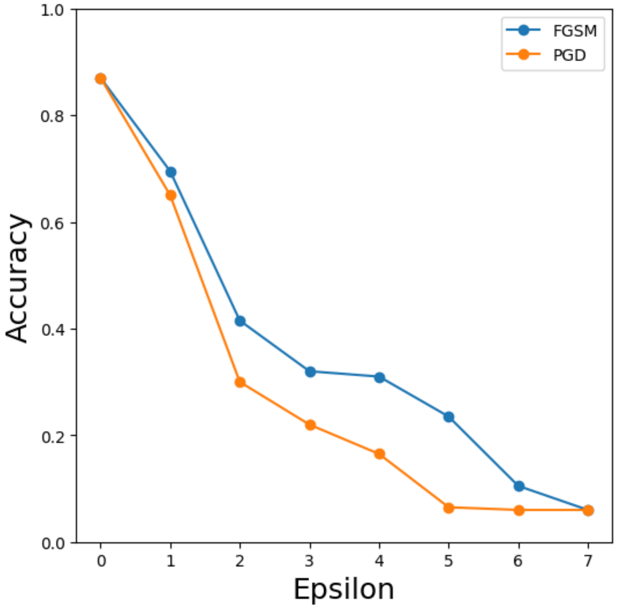


Task-2

a) Classification accuracy on clean and adversarial images Table-

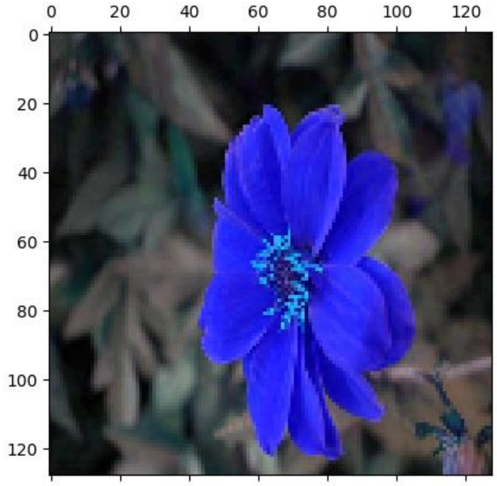
model	Clean images	Adversarial images €=1/255	Adversarial images €=5/255	Adversarial images €=8/255
FGSM attack	87%	69.5%	32%	31%
PGD attack	87%	65%	22%	16.5%

b) The accuracy versus perturbation ϵ for FGSM and PGD adversarial attacks-

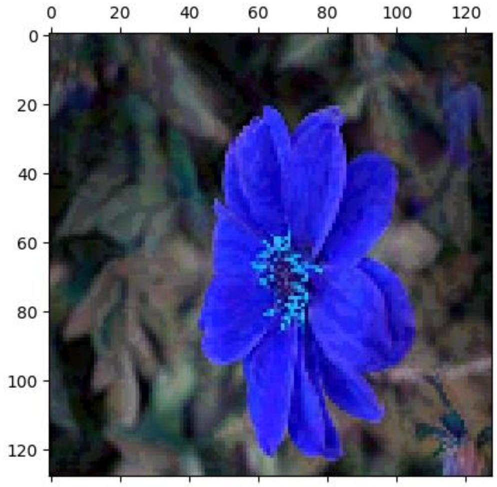


c) Figures with added adversarial perturbation and the labels for ϵ= [3/255, 8/255, 20/255, 50/255, 80/255]

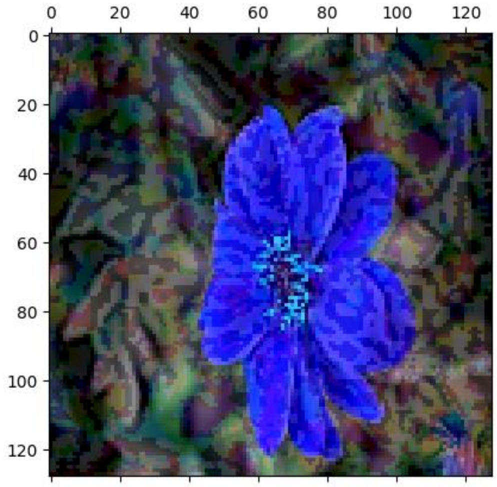
Perturbation maginutude: 0.0118
Predicted label: bishop of Llandaff



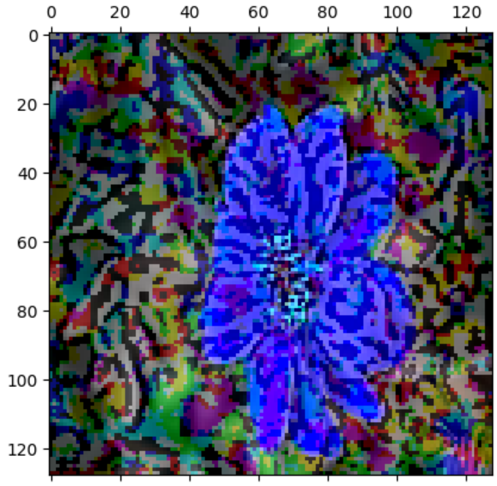
Perturbation maginutude: 0.0314
Predicted label: bishop of Llandaff



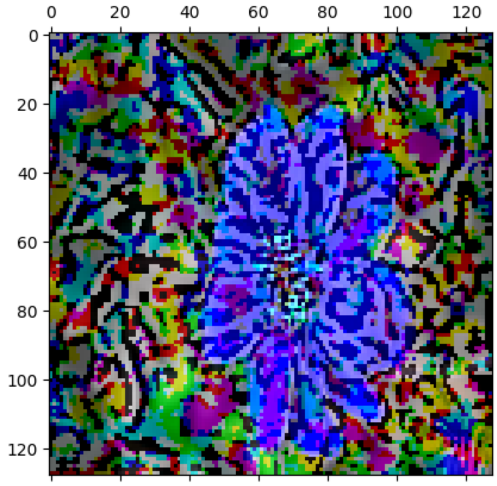
Perturbation maginutude: 0.0784
Predicted label: bishop of Llandaff



Perturbation maginutude: 0.1961
Predicted label: wallflower



Perturbation maginutude: 0.3137
Predicted label: wallflower



d) The model's predictions show robustness to small and moderate perturbations, consistently predicting "bishop of llandaff" up to a perturbation magnitude of 0.0784. However, at a perturbation magnitude of 0.1961, the prediction changes to "wallflower," indicating a threshold where significant input alterations impact the output. Beyond this threshold, the model consistently predicts "wallflower," suggesting that while the model is stable with minor changes, larger perturbations lead to different predictions.

Notebook link- https://github.com/sabidarrow/uidaho