Pet Breed Identification Case Study Rubric

DS4002 - Lucas Vallarino - Spring 2025
Due: Monday, April 28th, 2025
Submission Format: GitHub link uploaded to Canvas and Hard Copy Turn-In

Individual Assignment

Why am I doing this? This case study allows you to apply machine learning and computer vision to real-world problems. By building a pet breed identification model, you will deepen your skills in data preprocessing, CNN modeling, and model evaluation.

What am I going to do? The link for the GitHub repository is https://github.com/lucasvg133/CS3-DS4002. You will download the Oxford-IIIT Pet Dataset from the link in the GitHub and will work with it. Using all available breeds (dogs and cats), you will preprocess image data, fine-tune a ResNet-50 model through transfer learning, and evaluate model performance. Your goal is to develop a robust CNN achieving at least 90% test accuracy. To visualize results, you will produce a confusion matrix as well as graphs that show the results of the training vs validation data accuracy and the training vs validation data loss.

Final deliverables include:

- Well documented source code: Colab notebook with comments and markdown explanations
- Confusion matrix: visual representation of true vs predicted labels
- Training vs validation accuracy graph: analyzed model generalization over time
- Training vs validation loss graph: checks for overfitting or underfitting
- Final accuracy values (results): reports the test dataset performance
- A GitHub repository: organized repository containing all project materials, code, output figures, and documentation

How will I know I have succeeded? You will meet expectations by following the criteria in the rubric below:

Spec Category	Spec Details

Formatting	• One GitHub repository organized clearly (submitted via link on Canvas)
	• The top-level page of the repository includes:
	  - A README.md File
	  - A LICENSE.md File

	- A Scripts Folder - A Data Folder - An Output Folder - A REFERENCES.md File (if any references were used)
README.md	● Goal: This file serves as an orientation to everyone who comes to your repository, it should enable them to get their bearings. ● This should give a brief introduction to the case study and what you produced: ○ Project overview ○ Dataset used ○ Structure of repo
LICENSE.md	● Goal: Goal: This file explains to a visitor the terms under which they may use and cite your repository. ● Select an appropriate license from the GitHub options list on repository creation. ● Usually, the MIT license is appropriate.
Scripts Folder	● Goal: This contains your well documented Colab notebook file that contains the code used to create your CNN model and results. ● The folder must have a script for analysis containing the following: ○ The CNN model ○ The accuracy results ○ The confusion matrix ● It is optional, but recommended, to include data acquisition scripts for both data organization and pre-processing.
Data Folder	● Goal: This contains the data used for the model, NOT the original dataset. ● A single or multiple csv file of your cleaned and processed image data. ● Include a data appendix to provide an overview of the datasets and transformations applied, from raw input data to analysis data.
Output Folder	● Goal: This folder contains all of the output generated by your project, e.g. figures, tables, etc. ● It must include: ○ Confusion Matrix ○ Training vs Validation Accuracy Results ○ Training vs Validation Loss Results

REFERENCES.md	• Markdown file citing any resources used to build the model