CS3 Rubric – Classifying Skin Tumors

DS 4002 - Fall 2024 - Grace Kang
Due: TBD
Submission format: Submit a GitHub link to Canvas.

Individual Assignment

General Description: Submit to canvas assignments a link to GitHub repository which includes a single page pdf reflecting on the results of the study.

Why am I doing this? This is your opportunity to gain a better understanding of the data science research process by dipping your toes in running through a completed data science project. After following this case study, you will understand the motivations behind crafting a research question, gain more technical experience in terms of handling data and training models, and use critical thinking to analyze results and provide improvements for the study.

What am I going to do? The GitHub repository for this case study can be found at https://github.com/gckang/CS3-DS4002/. You will start out by reading 'CS3 Hook.pdf' as well as the articles in the 'MATERIALS' folder to learn about the project and its motivations. Then you will retrieve the data by following the links in the 'CS3 Image Data.pdf' document in the 'DATA' folder and develop an understanding of the source data. After comprehending the research problem, follow the Python code in the 'SCRIPTS' folder and perform exploratory data analysis on the data through 'EDA.ipynb' where you will explore the demographics of gender, skin region, and age across various groups. Then, using the code in 'finetuning.ipynb', you will fine-tune two EfficientNet models using a different number of classes each time: six classes vs. two classes. Lastly, you will evaluate your results and consider options for future steps to improve the results.

Your final deliverables should include:

- GitHub repository
- A single page document analyzing the results and feedback for improvement

All of this will be submitted electronically via a link to a github repository.

Tips for success:

- Understand what each block is doing in the scripts and challenge the method. Can you think of a different approach for the same output?
- It is ok if results go in an unexpected direction! That is part of the research process and it's a chance to rethink and explore possible reasons of where these results stem from.

How will I know I have Succeeded? You will meet expectations on this case study when you follow

the criteria in the rubric below.

Spec Category	Spec Details

Formatting	• One GitHub repository (submitted via link on Canvas)
	• The repository should contain:
	  ○ Data & Code
	  ■ Submit any code you write for all portions in the repository
	  ■ Include any additional data used into the repository
	  ○ Reflection Document
	  ■ Submit the document as a single page pdf file.

Data & Code	• Goal: All code developed and data used will be into the repository. • The original repository will already contain all scripts needed, titled 'EDA.ipynb' and 'finetune.ipynb'. It will also already contain all data needed including 'metadata.ipynb' which maps each image file to the diagnosis and other subject data, as well as links to download the images. • If any code was added or changed, provide brief comments on each line you edited to explain the functionality and reason for changing. • If you created an additional script file, add that to the SCRIPTS folder with comments describing what you did. • If you used additional data, add that to the DATA folder.

Reflection Document	• Goal: Analyze results from the project and present feedback.
	• Analyze the results from the exploratory data analysis notebook file and the results from the two fine-tuned EfficientNet models in comparison to the baseline, pre-trained EfficientNet model results and in comparison to each other.
	  ○ Two paragraphs long
	• Provide explanations for what the results imply and give feedback on some actions to take to improve the results of the fine-tuned model
	  ○ One paragraph long
	• One page maximum
	• PDF format

Acknowledgements: Special thanks to Professor Alex Gates for the rubric template. This structure is pulled from Streifer & Palmer (2020).