Capstone Project 1 Proposal

Problem:

Every day more than 9,500 people are diagnosed with skin cancer, making it the most common cancer in America. As of 2011, the annual cost of treating skin cancer is $8.1 billion, and the Center for Disease Control (CDC) called it "a growing public health issue".

Skin cancer has a high survival rate when caught early on; however, if left untreated, invasive cancerous cells can migrate to other parts of the body via the blood or lymph supply and become lethal. Thus, early detection and treatment is crucial for optimal recovery. The problem is, many benign conditions of the skin appear similar to skin cancer, and for an untrained individual it may be difficult to differentiate between a such harmless condition and something more malign.

Target audience:

In this capstone project, I would like to create an application of sorts for a non-health professional individual who suspects that he/she may have skin cancer. The individual can upload a picture of their skin condition, and the system will return the most likely diagnosis among those included in the dataset. Using this information, the individual can make a more informed decision on whether or not to pursue further treatment.

Data acquisition and description:

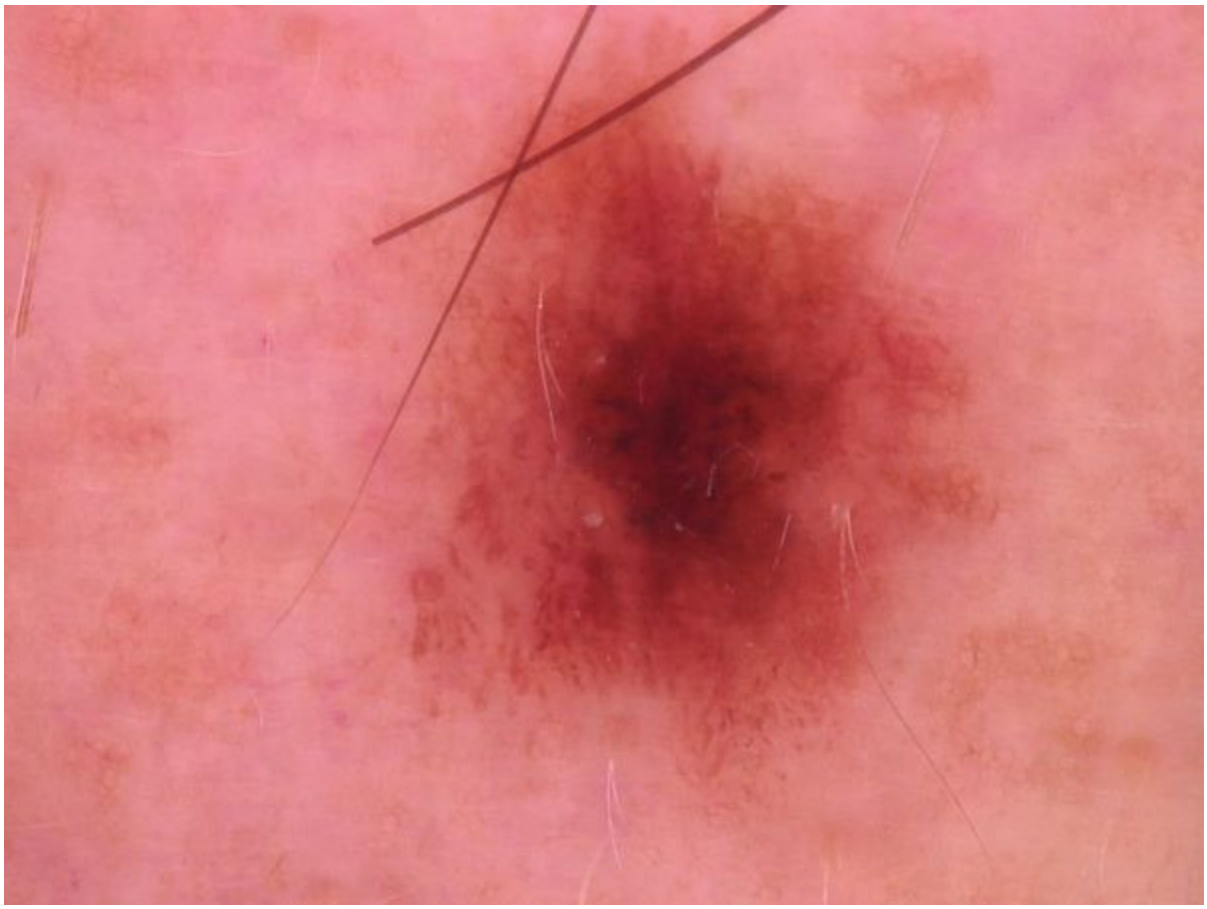
My dataset is the HAM10000 dataset from the Medical University of Vienna (available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T). The dataset contains 10015 photos of cancer-like skin conditions and a .csv file with the lesion id, the image id, the diagnoses, the age of the patient, the sex of the patient, and the location of the condition. Refer to Table 1 for a summary of the skin conditions contained in the dataset, Figure 1 for an example image to be analyzed, and Figure 2 for the first ten rows of the tabular data.

Table 1: A description of the diagnoses included in the HAM1000 dataset .

Abbreviation	Diagnosis	Non/Cancerous	Description	Counts
akiec	Actinic keratoses and intraepithelial carcinoma	Noncancerous	A pre-cancerous area of dry skin caused by UV damage.	327
bcc	Basal cell carcinoma	Cancerous	Manifests as a bump and is	514

			usually caused by prolonged exposure to UV rays	
bkl	Benign keratosis-like lesions	Noncancerous	A benign skin condition on the upper extremity that can be patchy or dry in appearance.	1099
df	Dermatofibroma	Noncancerous	Small modules that form in the upper extremities.	115
mel	Melanoma	Cancerous	Cancerous proliferation of melanocytes. The most common type of skin cancer and manifests as a darkened patch.	1113
nv	Melanocytic nevi	Noncancerous	Noncancerous proliferation of melanocytes. Commonly known as a “mole”.	6705
vasc	Vascular lesions	Noncancerous	Occurs after injury to the skin. Commonly known as a “bruise”.	142

Figure 1. Example of an image to be used for the model.



(For those curious, the diagnosis of this particular picture is Melanocytic nevi, or a mole.)

Figure 2. The first ten rows of the tabular metadata.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear
5	HAM_0001466	ISIC_0027850	bkl	histo	75.0	male	ear
6	HAM_0002761	ISIC_0029176	bkl	histo	60.0	male	face

7	HAM_0002761	ISIC_0029068	bkl	histo	60.0	male	face
8	HAM_0005132	ISIC_0025837	bkl	histo	70.0	female	back
9	HAM_0005132	ISIC_0025209	bkl	histo	70.0	female	back

Method:

I intend to utilize all of the data, and I plan to extract some features from the image and use them to build a multi-class classification model for predicting the diagnosis of a given picture as belonging to one of the 7 categories listed in Table 1.

As the data is imbalanced with the majority diagnosis (67%) being “nv”, accuracy is not an effective metric; a possible alternative metric could be the F1-score, precision, or sensitivity. In addition, an under-sampling technique to balance the distribution of the diagnoses might be necessary.

After performing feature extraction on the images and converting the information to tabular data, I intend to fit the data to several different models and compare the results. Based on the results of previous models in published manuscripts, sensitivity could be around 0.8 and specificity could be >0.95.

Deliverables:

I will have powerpoint slides, a report in PDF, and a jupyter notebook along with interim reports and projects on GitHub.