Project Proposal:
Dermoscopy-Based Melanoma Classification Using CNNs

CS 221 Project, Fall 2021

Yifu Chen
yifuchen@stanford.edu

Mariana Irene Frangos
mfrangos@stanford.edu

Jacob Alan Smith
jacobas@stanford.edu

Ahsan Raisul
dahsan@stanford.edu

***Abstract*—Melanoma is a type of serious skin cancer that can develop anywhere on the body. Approximately 7,000 deaths are attributed to melanoma each ear in the US. The diagnosis of melanoma often requires both a physical exam and a biopsy (removing a skin sample for testing). The cost of biopsies can range from $100 - 300, which is often not covered by insurance and therefore poses a barrier towards early melanoma diagnosis. We plan to develop an image-based melanoma classifier using deep learning, so that people receive quick diagnosis of their suspicious skin lesions. We found a suitable dataset on Kaggle, and plan to train several state-of-the-art Convolutional Neural Network (CNN) models to perform this task. We will explore three topics: 1) the benefit of transfer learning from pre-tarined image models, 2) the optimal hyperparameters for fine-tuning these models, and 3) whether CNN models may outperform human dermatologists.**

***Index Terms*—Convolutional Neural Networks, Deep Learning, Computer Vision, Melanoma, Cancer**

I. Problem Statement and Task Definition

The diagnosis of melanoma from just an image is hard. Without additional information (e.g., from biopsies), human must rely on visual features such as the color, size, and shape of the lesion. Since these features are able to be encoded by the CNN architecture, we will develop and compare various CNN computer vision models for classifying images of the skin lesion into melanoma or non-melanoma.

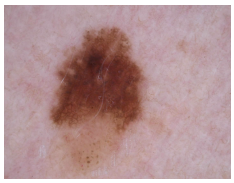
The model will help people who are concerned about skin cancer to receive a quick screening before they decide whether to visit a dermatologist in person. To note, our model is not intended to give the final diagnosis, since physical biopsies provide much more information about the skin cell pathologies than a single image.

II. Input/Output Behavior

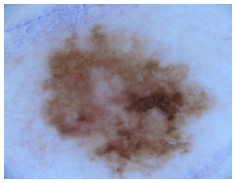
The input is an RGB image obtained from a dermoscopy. Each image contains a skin lesion that is potentially a case of melanoma. Two examples are shown in Figure 1.

The output is a binary label that indicates the presence of melanoma. In the original dataset, each image belongs to one of the following nine categories:

1) Melanoma
2) Melanocytic nevus
3) Basal cell carcinoma
4) Actinic keratosis



(a) Negative for melanoma



(b) Positive for melanoma

Fig. 1: Two examples from the image dataset.

5) Benign keratosis
6) Dermatofibroma
7) Vascular lesion
8) Squamous cell carcinoma
9) None of the above

In order to optimize the classification accuracy of melanoma (since it's the most serious), we group 2-9 as a new "non-melanoma" category. The dataset we will be using contains 25,331 labeled images. We will be using a subset of this dataset to train our model, and the rest will be using for validation and testing.

III. Evaluation Metric

Melanoma is a life-threatening disease. For this binary classification task, we are hoping to detect every melanoma case, even at the cost of potentially higher false positives. In the context of our model, a false negative is a image of melanoma that is incorrectly classified as a picture that does not show melanoma. Since our goal is to correctly identify melanoma in patients, it is imperative that we minimize the number of false negatives so as not to cause a delay in patients getting the treatment they need for this very serious disease, assuming this model is applied to real-world cases. Therefore, we choose the main evaluation metrics to be Accuracy and Recall (Sensitivity). Given a confusion matrix:

		True Diagnosis	
		Positive	Negative
Classification	Positive	TP	FP
	Negative	FN	TN

We compute the following evaluation metrics:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}, \text{Recall} = \frac{TP}{TP+FN}$$

IV. Related Works

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6231861/
https://www.sciencedirect.com/science/article/pii/S0959804919302217
https://ieeexplore.ieee.org/abstract/document/8945133

V. Baseline and Oracle

The **Baseline** performance may be established in two ways. First, we will train and evaluate a Multilayer Perceptron that is a small fraction of the size of the candidate CNNs models. In addition, we will train a CNN from scratch (instead of using transfer learning from pre-trained CNN models), and compare the performance with fine-tuned pre-trained models.
The human dermatologists establish the **Oracle** performance. Since we lack the medical knowledge to hand-label images, we will fuse the human performance from previous similar research papers and give an estimated dermatologist accuracy for our task.

VI. Methodology

Our plan is to apply various pre-trained deep learning models to perform binary classification of images as being melanoma. The first model we want to try is Google's EfficientNet. We will begin by running on default parameters, then tune the model based on knowledge of the data and performance feedback on the dataset.

VII. Challenges

It is challenging for the untrained human eye to perform tasks such as distinguishing between a case of melanoma and a mole. In fact, previous studies have shown that even dermatologists have high error rates (without additional information, e.g., biopsy results). In previous works, researchers found that the dermatologists are 65%-80% accurate in melanoma diagnosis.

Acknowledgment

We are thankful for the support and mentorship from the CS 221 CA team, especially to our team mentor, Yuchen Wang.