Title: Predicting Skin Cancer using Machine Learning

Abstract:

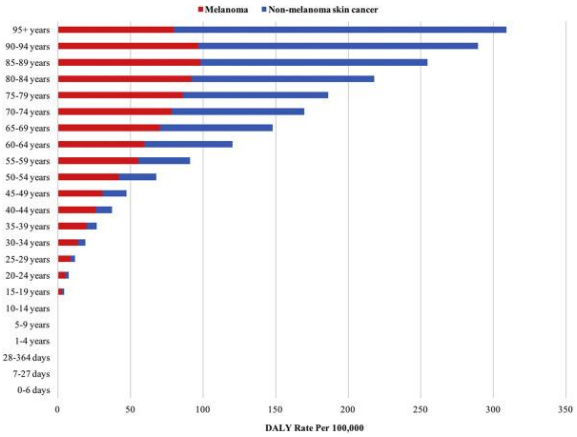
The motivation for this research project was to create an accurate model that could be used to identify the presence of different types of skin cancers. The latter objective is to create an application that could be easily used in rural areas, as they do not have access to modern medical technology. The approach began with loading the dataset into a readable format. I used the load_dataset import from Hugging Face [4] to complete this task. The models that were used for this research project were all from Hugging Face [4]. The highest accuracy was 0.8633 and the lowest accuracy was 0.6647. The models were all run on a colab notebook. The meaning of these results is that on the best model the image was predicted correctly 86% of the time. The conclusion of this research is that more research must be conducted and more models must be used in order to achieve an accuracy that would warrant the model to be used in the real world.

Intro draft:

Melanoma and other skin cancers affect 5.4 million people every year. Basal cell cancer (BCC) is around 80% of all skin cancer cases, Squamous skin carcinoma (SCC) is around 20% of all skin cancer cases, Melanoma accounts for 1% of skin cancer cases but causes the majority of skin cancer deaths. Skin cancer affects an increasing number of people which is the main motivation for my research. The goal is to develop a model that is accurate enough for real world application and that could be distributed to more rural areas which have decreased access to medical technology. Identifying skin cancers early could help save many of the lives lost to melanoma, as melanoma has a high chance of being cured if caught early. The image data used is from the HAM10000 and the ISIC database. The model used is an image classification model from Hugging Face [4] which was trained on data from the HAM10000.

What Causes Skin Cancer?

The purpose of the research conducted entails different types of skin cancer. In order to understand how to effectively use the models, we must understand what causes skin cancer in the first place. The main cause for skin cancer is over exposure to sunlight. Ultraviolet rays from the sun damage the DNA in your skin. This causes abnormal cells to form, which causes rapid cell division and cancer. Skin cancer is much more common than many people think and the risk for skin cancer increases dramatically as one ages. The most effective way to lower the chances of having skin cancer is to not overexpose yourself to the sun, wear clothing that covers your arms and legs, and use sunscreen with an SPF above 15.

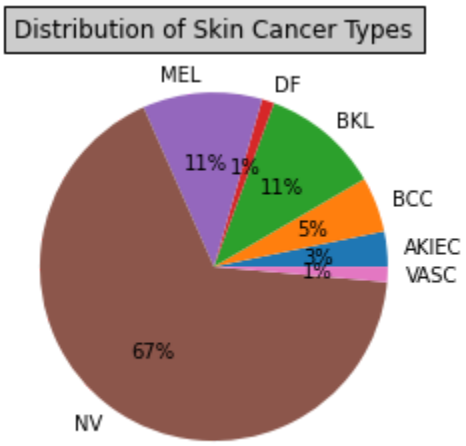


About the Dataset:

The dataset used in my research was the HAM10000 dataset. This dataset contains ten thousand images of skin lesions with seven different types of skin cancers. The types included in the dataset are Actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions. The dataset was pre-processed by first uploading it to a google drive folder, so it could be accessed from the colab notebook, and then input into the load_dataset import from Hugging Face.

Distribution of the Data Categories:

The amount of images for each category of skin cancer was different and does not reflect the real world ratios of skin cancer occurrences. Melanocytic nevi had the greatest percent of images at 67%, which is more than double the amount of melanocytic nevi occurrences at



25-33%. Both melanoma and benign keratosis-like lesions were 11% of images in the dataset. Melanoma only accounts for 1% of skin cancers in the United States, however it causes the most deaths. Benign keratosis-like lesions are much more common however they are not nearly as fatal as melanoma. Around 10% of the remaining images is split between basal cell carcinoma (5%), actinic keratoses (3%), dermatofibroma (1%), and vascular lesions (1%).

Use with Machine Learning:

The image data in the HAM10000 was used with various machine learning models found on Hugging Face. However the data had to be manipulated slightly in order to be used efficiently with the machine learning models. First, The image data was downloaded into Google Drive which made the data accessible in Google Colab. Second, The image data was all lumped into two folders, one with all the images and one with all the masks for the images. This meant that I had to create a folder for each category of skin cancer, which was accomplished through the use of the Ground Truth csv which contained the information for each image. The data was prepared for machine learning models through the datasets import. This import automatically splits the data into test and train with 90% of images going to train and 10% of images going to test.

Science Underlying the Objective:

The main objective for my research is to create a model that could accurately predict skin cancer types. Some concepts that are useful to complete this objective are concepts about image classification and how skin cancer occurs. Image classification models have complicated internal

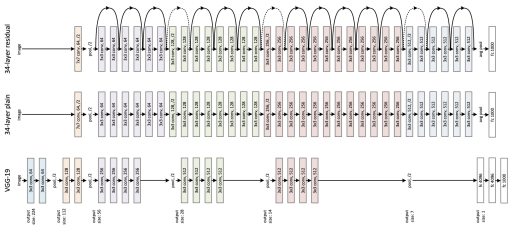
workings that are used to output some sort of value. Especially many advanced image classification models, which are trained on millions of different images. The effectiveness of these models is crucial in achieving the objective of the project. Which is to find a model that is accurate and easy to use so that people can check skin lesions without going to the doctor.

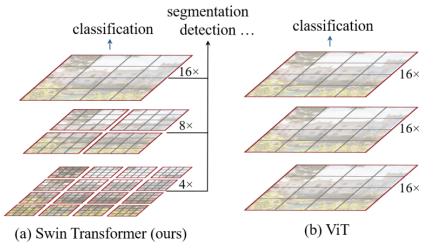
Methodology:

My approach to complete the objective started with organizing my data. I organized the images into their respective categories and put all the categories under one folder so that the dataset could be read easily. The data was read through the dataset import. This import automatically separated the data into train and test splits. The train split was 90% of the images while the test split was the remaining 10% of the total images. Once the dataset was in a readable format, it was used with the models from Hugging Face. The other models used were advanced models found on Hugging Face. There were six different models used and they are "keithanpai/dit-base-finetuned-rvlcdip-finetuned-eurosat"[11], "keithanpai/resnet-50-finetuned-eurosat"[12], "keithanpai/swin-tiny-patch4-window7-224-finetuned-eurosat"[13], "keithanpai/tiny-random-vit- finetuned-eurosat"[14], "keithanpai/vit-base-patch16-224-finetuned-eurosat"[15], "keithanpai/vit-base-patch32-384-finetuned-eurosat"[16]. The first model had an accuracy of 0.7315 and a loss of 0.7997. The second model had an accuracy of 0.6647 and a loss of 1.0488. The third model had an accuracy of 0.6677 and a loss of 1.1981. The fourth model had an accuracy of 0.8423 and a loss of 0.4381. The fifth model had an accuracy of 0.8633 and a loss of 0.3953. The sixth model had an accuracy of 0.8074, and a loss of 0.5495.

How the Models Work:

The models employed were Swin Transformer, Vision Transformer, ResNet, and Document Image Transformer. The Document Image Transformer model was pre-trained on IIT-CDIP, which is a dataset that includes 42 million images and was fine-tuned on RVL-CDIP, which is a dataset that includes 400,000 images in grayscale with 16 different classes. The ResNet model was pre-trained on ImageNet-1k at a resolution of 224x224. It is important to note that the images used in my dataset were set to 180x135 to decrease training time for the models. The Vision Transformer is a transformer encoder model pre-trained on images in a supervised fashion, namely ImageNet-21k. The Swin Transformer is a vision transformer that builds hierarchical feature maps by merging image patches.



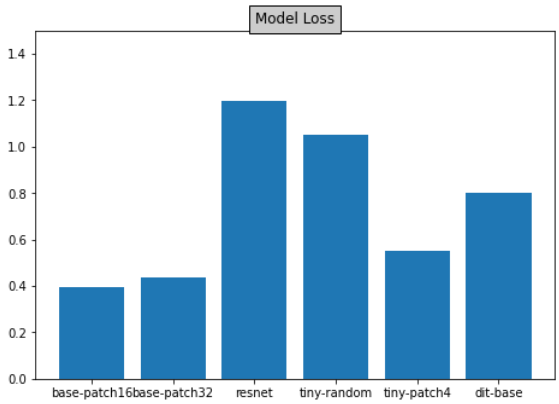


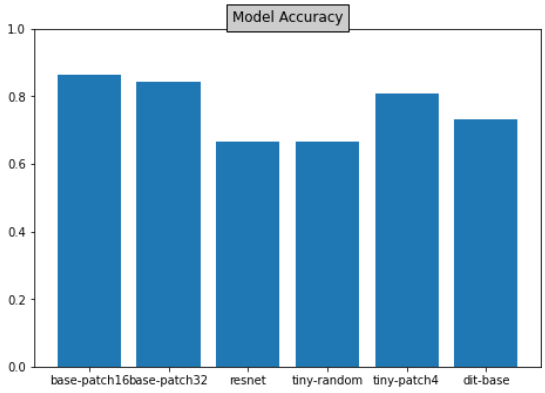
(a) Swin Transformer (ours) (b) ViT

Results and Discussion:

The main objective of this research was to produce an accurate reading of images of skin lesions, preferably in a short time frame. The stated question was whether this goal was even possible. The motivation for this project was brought by my experiences in rural communities. In many rural communities in India, they are not able to access modern medical technology. In order to solve this problem I wanted to create an AI that could detect skin cancer that could work in rural areas. So far, the model is complete but could be improved. The highest accuracy so far was 0.8633. This is relatively good, but I feel that it still might not be reliable enough to use in the real world. There is another experiment done on Hugging Face, with the same dataset but it only had an accuracy of 0.7275. The potential source of errors might arise from incorrectly labeled images. Additionally, when loading the dataset there might have been some uncaught

errors, as the loading of the dataset was automated. Some of the limitations of my approach is that in order to find a high accuracy model I will have to search through many different image classification models, which will take a significant amount of time. Looking at other peoples image classification models for skin cancer might help, however there is only one other model on Hugging Face.





In order to have my AI research reach and positively affect the largest amount of people a product must be made that is cheap and useful. The product will look like a smaller Iphone. It will have a screen with a camera on the other side of the device. The camera will be used to take pictures and the screen will be used to show what the camera is seeing in addition to the results. There are some drawbacks to this approach, such as bias within the model. The images used in the HAM10000 are mostly people with pale skin, which would create inaccuracies when darker skin tones are input. Additionally, the device must also be mass produced ethically. In the ideal sense this device would be distributed to every home within a rural community, or enough devices to which every person in the community would be able to use it. The model must also be very reliable which would be around (90+)% accuracy. Any accuracy below 90 would not be effective enough to have a practical use.

Conclusion:

The main question that was meant to be solved was could a highly accurate model be created and produce outputs quickly. The models ideally should output results within a few seconds, or at least a minute. However accuracy is the greatest concern, while time for results is secondary. Many models were used in my notebook and the highest accuracy was 0.86. The broader implications of the results are the impacts of the model in rural areas. As stated before, rural areas do not have access to modern medical technology. A way to help assist this issue would be to create some sort of application that could predict skin cancer types. A website could be developed or even a mobile app, which would give access to more people. However, many people in rural areas also do not have cell phones or access to the internet. To reach people without access to the internet a handheld device that could take pictures and use the model to predict skin cancers could be possible. The device should also have on-device machine learning for privacy and for ease of use, especially in rural regions with limited access to wifi. Some possible extensions of my work could be to add more data when training the models to create a higher accuracy. Additionally, it could be expanded to more types of skin cancers in the future.

Acknowledgments:

This research project was completed by Eric Bradford, and myself. Eric helped guide me throughout the writing of the paper, creation of the colab notebook, and was very helpful for fixing issues with code. I implemented the models, loaded the dataset, and wrote the paper.

References:

[1] Mayo Clinic. 2022. *Actinic keratosis - Symptoms and causes*. [online] Available at: <https://www.mayoclinic.org/diseases-conditions/actinic-keratosis/symptoms-causes/syc-20354969> [Accessed 15 August 2022].

[2] The Skin Cancer Foundation. 2022. *Basal Cell Carcinoma*. [online] Available at: <https://www.skincancer.org/skin-cancer-information/basal-cell-carcinoma/> [Accessed 15 August 2022].

[3] Conditions, G., 2022. *Giant congenital melanocytic nevus: MedlinePlus Genetics*. [online] Medlineplus.gov. Available at: <https://medlineplus.gov/genetics/condition/giant-congenital-melanocytic-nevus/#:~:text=Giant%20congenital%20melanocytic%20nevus%20is,is%20noticeable%20soon%20after%20birth.> [Accessed 15 August 2022].

[4] Huggingface.co. 2022. *Hugging Face – The AI community building the future.*. [online] Available at: <https://huggingface.co/> [Accessed 15 August 2022].

[5] The Skin Cancer Foundation. 2022. *Melanoma*. [online] Available at: <https://www.skincancer.org/skin-cancer-information/melanoma/> [Accessed 15 August 2022].

[6] Mayo Clinic. 2022. *Melanoma - Symptoms and causes*. [online] Available at: <https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884#:~:text=Melanoma%2C%20the%20most%20serious%20type,in%20your%20nose%20or%20throat.> [Accessed 15 August 2022].

[7] Cancer.org. 2022. *Melanoma Skin Cancer | Understanding Melanoma*. [online] Available at: <https://www.cancer.org/cancer/melanoma-skin-cancer.html> [Accessed 15 August 2022].

[8] Mayo Clinic. 2022. *Seborrheic keratosis - Symptoms and causes*. [online] Available at: <https://www.mayoclinic.org/diseases-conditions/seborrheic-keratosis/symptoms-causes/syc-20353878#:~:text=A%20seborrheic%20keratosis%20(seb%2Do,or%20scaly%20and%20slightly%20raised.> [Accessed 15 August 2022].

[9] skinsight. 2022. *skinsight - Dermatofibroma*. [online] Available at: <https://www.skinsight.com/skin-conditions/adult/dermatofibroma#:~:text=Dermatofibromas%2C%20or%20histiocytomas%2C%20are%20common,depressed%20scars%20after%20several%20years.> [Accessed 15 August 2022].

[10] Ssmhealth.com. 2022. *Vascular Lesions*. [online] Available at: <https://www.ssmhealth.com/cardinal-glennon/pediatric-plastic-reconstructive-surgery/hemangiomas#:~:text=Vascular%20lesions%20are%20relatively%20common,Vascular%20Malformations%2C%20and%20Pyogenic%20Granulomas.> [Accessed 15 August 2022].

[11] Pai, K., 2022. *keithanpai/dit-base-finetuned-rvlcdip-finetuned-eurosat · Hugging Face*.

[online] Huggingface.co. Available at: <https://huggingface.co/keithanpai/dit-base-finetuned-rvlcdip-finetuned-eurosat> [Accessed 5 September 2022].

[12] Pai, K., 2022. *keithanpai/resnet-50-finetuned-eurosat · Hugging Face*. [online] Huggingface.co. Available at: <https://huggingface.co/keithanpai/resnet-50-finetuned-eurosat> [Accessed 5 September 2022].

[13] Pai, K., 2022. *keithanpai/swin-tiny-patch4-window7-224-finetuned-eurosat · Hugging Face*. [online] Huggingface.co. Available at: <https://huggingface.co/keithanpai/swin-tiny-patch4-window7-224-finetuned-eurosat> [Accessed 5 September 2022].

[14] Pai, K., 2022. *keithanpai/tiny-random-vit-finetuned-eurosat · Hugging Face*. [online] Huggingface.co. Available at: <https://huggingface.co/keithanpai/tiny-random-vit-finetuned-eurosat> [Accessed 5 September 2022].

[15] Pai, K., 2022. *keithanpai/vit-base-patch16-224-finetuned-eurosat · Hugging Face*. [online] Huggingface.co. Available at: <https://huggingface.co/keithanpai/vit-base-patch16-224-finetuned-eurosat> [Accessed 5 September 2022].

[16] Pai, K., 2022. *keithanpai/vit-base-patch32-384-finetuned-eurosat · Hugging Face*. [online] Huggingface.co. Available at: <https://huggingface.co/keithanpai/vit-base-patch32-384-finetuned-eurosat> [Accessed 5 September 2022].