CS231N Project Final Report
Improving Melanoma Classification With Semantic Segmentation

Shelly Goel
Stanford University
shelly23@stanford.edu

Harry Ko
Stanford University
harryxko@stanford.edu

Renasha Mishra
Stanford University
rmbk201@stanford.edu

Abstract

Melanoma is a deadly form of skin cancer but can be cured when detected early. Most of the previous work involving melanoma detection focuses either solely on detection or on finding the area of melanoma in images via segmentation. In our project we present a two-stage melanoma classifier which includes semantic segmentation in the first stage and subsequent binary classification in the second stage. The output of the first stage is thereby used as input for the second stage minimizing the area of visible skin around the lesions. We use the HAM10000 dataset to train our segmentation model and the ISIC2020 dataset to train the classification model. Both datasets are publicly available. We experiment with different architectures and approaches dealing with class imbalance. Our best model is trained on segments and achieves a balanced accuracy of 77.1% while surpassing the performance of a model trained on original data. In addition, we also explored the problem with racial biases in melanoma detection and how we can address this issue.

1. Introduction

We are investigating dermoscopic image classification for melanoma diagnosis using the International Skin Imaging Collaboration (ISIC) datasets from 2020. [19] Melanoma incidence has risen in the last 50 years. It accounts for the majority of skin cancer deaths and costs. Stage III and IV disease have 60.3% and 16.2% 5-year survival rates, respectively, whereas stage I/II melanoma has a 5-year survival rate of 97.6%. [1] The magnitude of difference in mortality between early and late stage diagnosis is quite large despite recent advancements in immunotherapy with checkpoint inhibitors. Survival with immunotherapy ranges from 8 to 12 months. [21] Advancements in late stage melanoma treatment have only increased costs with modest survival benefits. Melanoma accounts for $3.3 billion of the $8.1 billion direct annual skin cancer costs in the US. [10] Dermatologist appointments are often difficult to come by, often requiring scheduling 6 months or more in advance. Moreover, specialty care such as dermatology is difficult to access physically and financially due to lack of dermatology presence in more rural and under-resourced communities. If melanoma is caught early, treatment consists of a biopsy and regular skin exams, simple and cost effective. Early diagnosis in stages I/II will improve survival rates and reduce the cost burden on patients and the healthcare system. The dermoscopic image classification project we propose has the potential to improve survival rates, reduce annual melanoma costs, and democratize access to dermatologic care.

We approach the skin cancer detection problem as a two stage problem. In the first stage we propose a novel approach by adding a segmentation step before classification. In this step we take as input images from the ISIC dataset and use our segmentation model trained on the HAM10000 dataset to produce predicted masks of the skin lesion in the input images. We then use these masks to segment the skin lesion in the original input images. In the second stage we use these segmented images of the skin lesion and feed these as training data into an existing neural network architecture (DenseNet121[11]) with weights pre-trained on ImageNet. We replace the final fully connected layer of 1000 outputs with a fully connected layer of 2 outputs for our binary classification task. The binary classifier is then used to predict whether a given input image has a skin lesion that is malignant or benign. We choose the best model by evaluating using balanced accuracy as our metric on the validation set. We then use the best model to report the final results of the performance on the test set. The rationale for this approach is that we hypothesize that only the region of the mole is relevant for the skin cancer classification and weight sharing will hinder the classifier to learn relevant features of the moles by sharing weights with irrelevant regions of the image including the background and surrounding skin. Therefore, we expect that minimizing the surrounding background will help improve the performance of our classifier.

2. Related Work

We have included many helpful readings in our References section but there are ten main papers we focused on: [15] The authors tried segmenting the lesion or leaving it unprocessed before the classification training step. They tested on two datasets (HAM, ISIC) and used a U-net architecture for segmentation and ResNet50 for classification. They found that segmentation impact is unclear as it can decrease common lesion-adjacent confounding factors but classification accuracy is greatly dependent on segmentation quality which was dependent on data source. [6] The authors tried expanding the segmentation border for the target lesion which produced better results. Their model used CNN classifiers (VGG-Net and InceptionV3) with transfer learning to perform a binary classification (melanoma vs. benign). [13] The authors use the AlexNet architecture but replace the classification layer of AlexNet with a softmax layer. Additionally, three classes are used, namely Melanoma, Seborrheic keratosis (noncancerous, benign skin growths), and Nevus (birthmark or mole). On the ISIC dataset, they were able to achieve a 95.91% accuracy, 88.47% sensitivity, and 93% specificity. [17] The authors evaluated 12 different data augmentation scenarios across three networks (Inception-v4, ResNet, and DenseNet) and found that the basic set scenario which combines geometric and color transformations resulted in the best AUC values for all three networks. [3] The authors use a full resolution convolutional network segmentator (FrCN) to segment skin lesions and use the results as input into a neural network classifier. They find that the Inception-ResNet-v2 architecture performs best with an overall 82.59 F1-score. Translating the authors' reported confusion matrix into comparable metrics as our intended metrics, their best result translates into a 85.2% benign accuracy, 68% malignant accuracy and 76.6% balanced accuracy. [22] The authors use a pretrained ResNet34 as the encoder in a U-Net style architecture to segment skin lesions. Their best result uses a LinkNet34 pretrained on ImageNet and achieves a dice coefficient of 0.89 using multiple datasets combined as training data. [12] The authors propose a framework by incorporating transfer learning for segmenting lesions using an underlying UNet architecture with a novel multi-scale convolutional (MSC) block. This block is built by using four convolutional blocks at different scales. Their trained model outperforms state-of-the-art approaches for ISIC2016 and ISIC2017 segmentation task. [4] The authors propose addressing data imbalance by using data augmentation techniques. They use rotations, flips, and random transformations. They show that data augmentation increases accuracy by approximately 5%. [14] While lighter-skinned people are at the highest risk of getting skin cancer the mortality rate of darker-skinned people is much higher (African Americans have a 73% first-year survival rate compared to white Americans who have a 90% first-year survival rate) since skin cancer is often diagnosed at a late stage for people of color. To address this, AI models have been created for earlier detection of melanoma; however, people of color are still largely at risk due to the fact that the datasets used to train the model are largely composed of data of "fair-skinned populations in the United States, Australia, and Europe". [9] Two challenges mentioned with developing AI models for skin cancer detection that we decided to address was having an unbalanced dataset where there are many more benign skin lesions compared to malignant skin lesions and how most current skin lesion datasets mainly included images from fair-skinned individuals. The paper also presented how there is a lack of transferability of the model's learned features for one population to another.

3. Methods

3.1. Segmentation

We train a semantic segmentation model using a U-Net [18] encoder-decoder architecture and use images of skin cancer as training data and the corresponding black and white masks as the ground truth labels. We prefer semantic segmentation to instance segmentation since in our training set there is only one skin lesion in each image. The U-Net architecture has been shown to train fast and to work well in a medical imaging context in previous work. The U-Net predicts a class label for every pixel of the image. It follows a U shaped encoder-decoder architecture and starts with a contracting downward path to capture context of the image and corresponds to a typical convolutional neural network architecture. In the following symmetric expanding upwards path, the network learns precise localization. We use upconvolutions instead of upsampling. As an additional feature, at every layer there is a direct skip connection from the contracting to the expanding path to avoid losing information from the downsampling in the downwards contracting convolution and the subsequent upwards expansion back to the original size of the image. A restriction with regards to the dimensions of the input image is that it needs to be possible for height and width to be evenly divided by 2 times the number of layers, 2 to the power of 5 for the original U-Net architecture. This is to prevent rounding during the max-pooling operation as then in the subsequent upwards path the dimensions will be different. We make two changes to the architecture proposed by the original authors. We add batch-normalization, which at the time of the original publication was relatively new, and use padding at every layer to achieve output masks of the same size as the input images. The padding will result in some loss of information at the edges which is not of high relevance for our case. In addition to the original U-Net architecture, we also use a second method where we replace the 5 layer en-

coder in the U-Net architecture with a DenseNet121 backbone with weights pretrained on ImageNet, the same network we use for our classifier. The encoder backbone still has skip connections on 5 layers but is deeper than the original U-Net encoder. We expect that the pretrained weights will make it faster and easier for the segmentation model to recognize borders of the moles as compared to the original U-Net which has to learn everything from scratch.

We train both segmentation models for 10 epochs with an Adam optimizer and a learning rate of 0.001. The batch size had to be restricted to 4 images due to limitations in GPU memory. For our first approach the width and height are 592*448 pixels and for our second approach the width and height are 576 * 448 pixels. We intended to stay as close as possible to the original 600 * 450 pixel image size, which is the smallest image size of the original dataset. We use dice loss as the loss function to minimize during backpropagation and include pixel-wise accuracy as an additional accuracy metric. Both metrics compare the pixels of the ground truth masks and the predicted masks. Pixel-wise accuracy just calculates the percentage of all pixels predicted correctly. The dice coefficient measures the area of overlap between the segment and the prediction (2*Intersection) divided by the total number of pixels in both the mask and the prediction (Union + Intersection). We calculate the dice loss as 1 minus the dice coefficient. It carries a big advantage over pixel-wise accuracy in that it deals better with class imbalance where the background is much larger that the actual mole. In this case of class imbalance, pixel-wise accuracy could be very high when the model predicts the whole image as background. However, as described earlier, the dice loss is only concerned with the overlap of the area of interest (the mole) and therefore doesn't reward correctly classified background at the expense of incorrectly classified moles.

In our approach using semantic segmentation with subsequent classification, the training data set is segmented with the segmentation model and the moles extracted to minimize background and surrounding skin. The classifier is then trained on the segmented training crops. The results are compared to the baseline which uses the original images. We use the same segmentation pipeline for the training set as well as for the validation and test data sets.

3.2. Classification

We use an existing convolutional neural network architecture with weights pretrained on ImageNet for our melanoma classifier. We investigated the performance of EfficientNetB0 in our milestone report but decided to use DenseNet121 for the final report based on better performance when fully finetuned. The DenseNet architecture is a deeper than EfficientNet's architecture due to the skip connections between each layer and all subsequent layers.

	weights	data	augm.	lr	l2	class w.
1	base froz.	orig.	no	1e-3	0	no
2	base froz.	seg.	no	1e-3	0	no
3	finetun.	orig.	no	1e-3	0	no
4	finetun.	seg.	no	1e-3	0	no
5	fc layers	seg.	no	1e-3	0	no
6	dense-bl.	seg.	no	1e-3	0	no
7	finetun.	seg.	yes	1e-3	0	no
8	finetun.	seg.	no	1e-2	0	no
9	base froz.	seg.	no	1e-2	0	no
10	finetun.	seg.	no	1e-3	1e-3	no
11	finetun.	seg.	no	1e-3	0	2-1

Table 1: We perform an ablation study by changing the training data, the methods for freezing layers, augmentations, learning rates, regularization and class weights.

This alleviates the vanishing gradient problem and allows for a deeper network architecture. It also reduces the number of total parameters. The authors of DenseNet show that it achieves a lower error rates on ImageNet compared to ResNet. We will investigate the performance of a DenseNet classifier trained on original ISIC2020 data and compare it to the performance of a classifier trained on segmented crops from the same dataset. We will experiment with various methods to finetune the pre-trained weights by freezing all, parts or none of the pre-trained baselayers. To address potential overfitting issues we also selectively include data augmentation and l2-regularization. We will also tune the learning rate. The different setups for our ablation study are shown in table 1.

All models use the Adam as the learning rate optimizer and are trained for 15 epochs. All models receive a 224 * 224 pixel RGB image as input and output a class label as prediction. We use the binary crossentropy loss as the loss function to minimize during backpropagation. For all methods we replace the final fully connected classification layer which has 1024 input units and 1000 output units with a layer with 2 output units for binary classification. We train the model freezing the baselayer except for the fully connected layer, freezing the baselayer except for the last denseblock, finetuning all weights and finetuning all weights and adding another fully connected layer with 1024 hidden units on top of the existing one.

The mathematical formula for binary crossentropy loss is shown below. In the base case, both classes receive equal weights of 1. Since the loss is averaged for each batch, the result would be a simple average. Including unequal weights in the loss function can be beneficial for highly imbalanced datasets or to place higher importance on a certain class such as in our case malignant moles. The resulting loss would be the weighted average of the batch. We will mainly

use equal weights since our class imbalance problem has already been solved with random upsampling. However, we will include one model with additional class weighting to place higher importance on the malignant class. We had also explored not using random upsampling and instead solely using class weighting to balance the benign and malignant class but with this method while the malignant accuracy reached near 100% accuracy our benign accuracy was extremely low.

$$
\begin{aligned}
loss(x, class) &= w[class] \cdot -log\frac{\exp(x[class])}{\Sigma_j \exp(x_j)} \\
&= w[class] \cdot (-x[class] + log(\Sigma_j \exp(x_j))) \\
loss &= \frac{\Sigma_{i=1}^{N} loss(i, class[i])}{\Sigma_{i=1}^{N} w(i, class[i])}
\end{aligned}
$$

3.3. Data imbalance

The ISIC2020 dataset is highly imbalanced consisting of only around 1.8% malignant examples. This can lead problems during training because the model can achieve a high accuracy by just predicting every example as benign. This can be mitigated with random upsampling where the malignant examples are repeatedly oversampled to match the number of benign examples. Therefore an artificial even split between the classes is created though the existing examples are presented to the model more than once per epoch. This oversampling is only done to the training set. The oversampling increases the risk of overfitting as the model will now see identical examples of malignant skin lesions more frequently but we will investigate this problem with data augmentation.

3.4. False negatives in medical imaging

One aspect to keep in mind when operating in the medical imaging domain is the importance of avoiding false negatives. In a real world application, the detection model will most likely be used as a first "filter" to alleviate the workload for a dermatologist who will then only look at the skin lesions classified as malignant. Therefore, it is of higher importance for the classifier to learn to classify skin lesions as malignant instead of benign and avoiding false negatives since a benign classification won't lead to a human doctor to take a second look or the patient becoming unconcerned with the skin lesion and not receiving treatment. On the other hand, if a skin lesion is classified as malignant but after closer examination turns out to be benign then no harm is done to the patient. We will investigate the use of class-weights to penalize the classifier when it gets malignant examples wrong.

4. Dataset

We use two publicly available datasets as our primary sources for the segmentation and classification tasks. The International Skin Imaging Collaboration (ISIC) 2020 challenge dataset contains 33,126 dermoscopic training images in JPG format with labels of unique benign and malignant skin lesions from over 2,000 patients.[8] It served as the primary dataset for the "2020 Melanoma Classification Challenge" hosted on Kaggle during the summer of 2020. All malignant diagnoses have been confirmed via histopathology, and benign diagnoses have been confirmed using either expert agreement, longitudinal follow-up, or histopathology. The images are from the following sources: Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School. We use the ISIC 2020 dataset as input data for our classifier.

For the segmentation model we will use the Human Against Machine (HAM) 10000 dataset which consists of 10015 dermatoscopic JPG images.[10] Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions. More than 50% of lesions are confirmed through histopathology, the ground truth for the rest of the cases is either follow-up examination, expert consensus, or confirmation by in-vivo confocal microscopy. The images are collected from different populations, acquired and stored by different modalities. The HAM10000 dataset also served as the training set for the ISIC 2018 challenge. In addition to class labels, the HAM10000 dataset also contains binary segmentation masks for all 10015 images.[21] Masks were initialized with a segmentation network and subsequently manually verified, corrected or replaced by a dermatologist. At the moment we have no intention to use the ground truth labels of different modalities for the classifier and will focus on using the training data and binary masks to train our segmentation model.

4.1. Preprocessing

For our baseline model we preprocessed the ISIC 2020 dataset as follows. First, we resized the images to a resolution of 640*480 pixels. The images in the dataset are of varying sizes ranging from 6000*4000 to 640*480 so we resized all images to the smallest image size. Second, we cropped the sides to achieve a 480*480 image size. Third, we resized all images to 224*224 which is the input size for our DenseNet classifier. In addition, the dataset consisted of 425 duplicates due to clerical error which we removed reducing the dataset to 32,701 images. For the segmentation task, we resized the images of the HAM10000 dataset to a size required for the U-Net encoders, which are 592 * 448 pixels for the original U-Net encoder and 576 * 448 pix-

els for the U-Net with DenseNet backbone. The restrictions are due to the pooling operations in the downwards path to make the shapes evenly divisible to prevent rounding which will change the image size during the subsequent upconvolution. All original images and masks in the HAM10000 dataset initally have a size of 600 * 450 pixels.

4.2. Data augmentation

Reviewing existing literature we decided the best augmentation techniques would be to use a combination from the set {rotations upto 90 degrees, horizontal flips with 0.5 probability, vertical flips with 0.5 probability, scaling by [0.8, 1.2], shearing by upto 20 degrees}. [17] We experimented with various combinations of this set of augmentations on our baseline model by running it for 4 epochs on a small subset of our data. A combination of rotation, shearing and scaling followed by horizontal flips and vertical flips performed the best. However, this combination did not improve our model when it was used on the full data set and run for a larger number of epochs. Hence we did not use any form of augmentation on our dataset in our best performing model.

4.3. Validation split

We split both datasets into 70% training data, 15% validation data and 15% test data. We train our model on the 70% training data and validate per epoch on the 15% validation data. Once the model finishes all epochs we select the best model in terms of validation accuracy for classification and balanced accuracy for segmentation. We evaluate the best model once on the 15% test data and report the results.

5. Experiments

5.1. Segmentation

We trained both segmentation models on the HAM10000 dataset for which ground truth segmentation masks are publicly available. The resulting masks are shown in figure 1. Overall, the U-Net trained with the pretrained DenseNet encoder reaches a lower dice loss of 0.06 compared to 0.09 for the original architecture. Pixelwise accuracy is also better at 94% compared to 92%. Upon visual inspection of the outcomes, the masks produced by the segmentation model with backbone also look better. The edges of the segments are smoother, the regions are filled more cleanly and there are minimal artifacts outside the mole areas. We attribute this to the capacity of the pretrained DenseNet to recognize edges and colors which had to be learned by the original U-Net encoder with a small dataset and during only 10 epochs. We used the trained DenseNet segmentation model to produce the masks for the ISIC2020 dataset. We run the whole dataset including training, validation and test data through the segmentation model and save the resulting masks. Since the resulting masks are black and white with mole regions white and background black, we normalize the masks to a range of 0 to 1 and multiply the mask with the original images. This results in images with black background (multiplied by 0) and the mole areas remaining in original colors (multiplied by 1). We then utilize OpenCV to create contours and bounding upright rectangles around the moles. We then cut out the bounding rectangles resulting in rectangles of different shapes with minimimal background areas in black color. We view the black background color as equivalent to zero padding around the segmented moles. We then resize all images to 224 * 224 to feed into our DenseNet classifier. The images are also redistributed into the same training, validation and test datasets.

5.2. Classification

The experiments we've done with our classifier are explained in detail in the next subsections. A comparison of all results can be seen in Table 2.

The models are evaluated on the classification accuracy for malignant moles. However, due to the class imbalance problem we cannot use overall classification accuracy as the main evaluation metric as this will incentivize the model to predict all examples as benign. On the other hand, we also want to avoid the classifier predicting all examples as malignant, so we cannot solely use malignant accuracy which is also called true positive rate (recall) as the primary evaluation metric. There are two metrics which take into account prediction accuracies for both classes, f1-score and balanced accuracy. While we report both metrics, we select the best models based on the balanced accuracy. To explain the reasoning behind our choice, we consider it of highest importance to minimize false negatives, malignant melanoma predicted as harmless and benign even at the expense of a higher rate of wrongly classified benign moles. In a real life scenario, a doctor would likely take a look at all moles classified as malignant whereas the moles classified as benign would not get a second look. Therefore, wrongly classified benign moles would be less harmful to the individual than wrongly classified malignant moles. The f1-score is calculated as 2*precision * recall / (precision + recall). Precision is calculated as the true positives / all predicted positives while recall is calculated as the true positives / all positives. The f1-score is only concerned with the positives, in our case malignant moles. The weakness of the f1-score for our case is that it rewards a high precision, a high number of true positives and a low number of overall predicted positives. This can offset a high recall, which is a high number of true positives as a percentage of all positives. Therefore an increasing f1-score can lead to a lower absolute number of true positives. Balanced accuracy on the other hand is calculated as the average of malignant accuracy (recall) and benign accuracy. This metric is not

	loss	accuracy	benign acc.	malignant acc.	balanced acc.	precision	f1-score
1	0.51	71%	70.8%	81.6%	76.2%	4.8%	9.1%
2	0.66	64.2%	63.9%	78.2%	71%	3.8%	7.3%
3	0.49	77.5%	77.6%	75.9%	76.7%	5.8%	10.7%
4	0.49	74.9%	74.8%	79.3%	**77.1%**	5.5%	10.2%
5	0.18	92.5%	93.7%	26.4%	60%	7.1%	**11.2%**
6	0.59	60.2%	59.6%	92%	75.8%	4%	7.7%
7	0.45	78.2%	78.3%	70.1%	74.2%	5.6%	10.4%
8	0.67	67.1%	66.8%	80.5%	73.7%	4.3%	8.1%
9	0.57	71.9%	72%	70.1%	71%	4.4%	8.3%
10	0.77	53.5%	52.8%	93.1%	73%	3.5%	6.7%
11	0.6	69.7%	69.5%	78.2%	73.8%	4.5%	8.5%

Table 2: The best performing model measured by balanced accuracy is number 4, the fully finetuned model trained on segmented crops. The gap to the second best number 3 model, the same model trained on original data is 0.04%. Note that the best model according to the f1-score has a very low recall (malignant accuracy).

concerned about the relationship between true positives and predicted positives, which is also the metric we decided to neglect. We just want the percentage of correctly predicted malignant moles to be as high as possible while keeping the percentage of correctly predicted benign moles reasonable. If we can train our model to recognize malignant moles then both accuracy metrics should go up in sync.

5.2.1 Segmented images

To analyze the performance of our two-stage melanoma detection model, we first replicate the experiments from our milestone on the original unsegmented images, a model with frozen baselayer and a model with all pre-trained weights finetuned, each ran for 15 epochs. The results show that the finetuned model outperforms the frozen baselayer model in terms of balanced accuracy. We then run both models on segmented images and compare the results in terms of frozen baselayer with the finetuned weigths and segmented with unsegmented images. We find that for segmented images the finetuned model also outperforms the frozen baselayer model in terms of balanced accuracy and the difference between the two models is larger than for unsegmented images. If we compare the performance of both models in terms of segmented and unsegmented images we find that for the frozen baselayer model, the model trained on unsegmented images outperforms the model trained on segmented images. However, when all pre-trained weights are fine-tuned, the model trained on segmented images outperforms the model on unsegmented images.

5.2.2 Finetuned weights

We experimented with two more approaches to improve the model with all weights finetuned. First, we added more trainable linear layers on top of the model with all weights unfrozen. Second, we freeze the baselayer but unfreeze the last denseblock and make it trainable. We keep all other hyperparameters constant and also train both models for 15 epochs. The best model is selected based on balanced accuracy. The results show that neither model outperforms the model with the original DenseNet architecture with all weights trainable.

5.2.3 Augmentations

Due to the relatively small size of the dataset in particular the number of malignant examples, we also included data augmentations to reduce the risk of overfitting and compared the performance of our best model with and without augmentations. We find that performance does not increase after 15 epochs compared to our best model. However, while performance for other models reached a plateau or started to decrease after 15 epochs, for the model with augmentations the loss seemed to decrease linearly even after 15 epochs and balanced accuracy seemed to have an increasing trend as can be seen in figure 12. To investigate further, we trained the whole model for 50 epochs. However, we found that performance also did not increase after 50 epochs with loss continuing to decrease linearly and balanced accuracy starting to decline after 15 epochs.

5.2.4 Hyperparameters

We also changed the learning rate, added weight decay (as L2 regularization) and class weights. We increased the learning rate from 0.001 to 0.01 since we noticed that the loss for some models plateaued on a high level while loss converged to zero for many others, which might indicate a local optimum. We also added weight decay since validation loss looked much more volatile than training loss. Finally, we also added 2-1 class weights to the crossentropy

loss where we attribute a twice as high importance to malignant example than to benign examples penalizing the model higher when it gets malignant examples wrong. However, none of these changes improved the performance of our initial parameters.

6. Results

6.1. Segmentation

The quantitative results of the segmentation can be seen in table 3 and the qualitative results in figure 1. The segmentation model trained with the DenseNet encoder backbone achieves a lower dice loss of 6% on the test set compared to the fully trained U-Net encoder model which reaches a dice loss of 9%. The best model by the authors of [22] who created the masks for the training dataset reaches a dice loss of 14.7% when trained only on the HAM10000 dataset. Comparing the results of our two segmentation models qualitatively, we can see that the model with the pretrained encoder predicts better masks with smoother boundaries, less artifacts around the mole and filled out surfaces especially with lighter moles. We attribute this qualitative difference to the pretrained weights which enhance the model's capabilities to detect edges, colors and shapes and to be able to distinguish noise as compared to the U-Net model trained from scratch. Upon visual inspection of the predictions for the test set, the masks generated by the pretrained encoder backbone look virtually identical if not better than the ground truth masks.

6.2. Classification

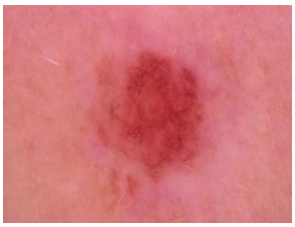
The results show that the classification model trained on segmented data performs better than the model trained on original data. Our initial hypothesis was that when the classifier is trained on original data, it is possible that the model detects benign and malignant moles based on patterns in the surrounding healthy skin instead of the mole itself. We can partly validate this hypothesis qualitatively by looking at saliency maps which show the final convolutional layer activations overlayed on the original images. One example can be seen in figure 2. The model trained on original images shows activations in a wide area of the image, including the mole and surrounding skin. The same mole segmented and cropped shows a different activation map with the model trained on all segmented crops. The activations are much more concentrated on different regions within the mole and the borders of the mole.

We can also see that the overall best model in terms of balanced accuracy is trained on segmented crops and outperforms the same model trained on original data. When we compare our results to two previous papers where the authors use a similar pipeline of segmentation and classification, we achieve a better balanced accuracy of 77.1% compared to 76.6% by the authors of [3]. One thing to note is that the authors of the paper select their best model according to the f1-score instead of balanced accuracy. Comparing our results to [15] we achieve the same balanced accuracy of 77.1% when measured against a holdout test set of the same dataset. However, the authors find that the performance of the model trained on the unsegmented dataset outperforms the model trained on the segmented dataset for all models.

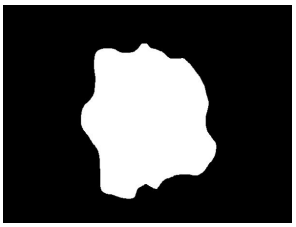
However, the training process is made difficult due to the imbalanced data and the importance of false negatives. We can see in figure 11 that the training loss for the best performing model trends towards 0 and overall accuracy trends towards 1. However, this is due to the data imbalance and the fact that malignant moles make up only 1.8% of the overall data. Even though we use random upsampling to deal with this problem, it doesn't completely resolve the issue. As such, over time, balanced accuracy and malignant accuracy trend downwards while benign accuracy trends upwards. The best model in terms of balanced accuracy is generally achieved in one of the early epochs of the training runs. We investigated this issue with class weights, regularization, augmentation and learning rates. While slight improvements can be observed, the overall issue remained that while the loss decreases, the malignant accuracy trends downwards while benign accuracy trends upwards and the classifier learns to predict more and more examples as benign.

The models trained with frozen baselayers tended to converge to a higher loss plateau (around 0.5) than the models with fully finetuned baselayers but increasing learning rates did not change the plateau. Adding class weights by using a weighted average in the loss function setting the malignant as twice as important as benign examples resulted in the model initially predicting most examples as malignant. However, it then gradually started to predict more and more examples as benign and after 15 epochs also predicted most examples as benign. With data augmentation and L2 regularization we observed an interesting phenomenon that loss did not converge towards 0 after 15 epochs but was decreasing linearly and was still relatively high. In addition, benign and malignant accuracies were on a similar high level. We trained the model including augmentations for 50 epochs but the results show that instead of starting to converge after 5 epochs, the model started to converge after around 20 epochs and showed the same behavior as the other models.

While we have shown that segmentation outperforms training on original data, the inherent class balance problem in our dataset makes it difficult to thoroughly investigate the problem. We would suggest two areas of further research. First, gathering more malignant training examples by combining several datasets or acquiring more samples directly from patients. Second, instead of using binary



(a) Original image

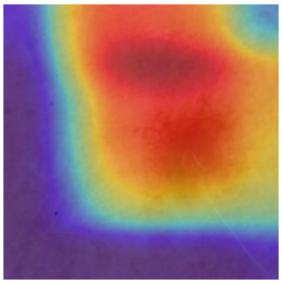


(b) Ground truth mask

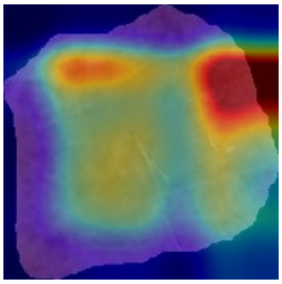
(c) U-Net mask

(d) Backbone mask

Figure 1: The DenseNet based encoder mask has no artifacts and a smoother overall shape compared to the U-Net encoder mask. The pretrained weights enable the DenseNet model to recognize boundaries better while the U-Net encoder needs to be trained from scratch.



(a) Original



(b) Crops

Figure 2: The final layer activations cover a wider area for the model trained on original images. The activations for the model trained on segmented crops show activations of specific areas within the mole and the edges. The pictures refer to activations of the same image for the best model (b) and the same model trained on original data (a).

classification, a multi-class classification based on different skin conditions could yield better results as malignant melanoma and benign moles might group several categories with distinct features into one class exacerbating the class imbalance problem.

For our final model, the metrics we care the most about is minimizing the False Negative rate. Because ROC Curves and the ROC AUC can be more optimistic on imbalanced datasets(we have 49 times as many benign images compared to malignant classifications), we decided to plot the Precision vs. Recall graph as well since both precision and recall focus on the malignant class. The metric graphs are displayed in Figure 1 and the metric calculations are shown in Table 1.

	dice loss	dice coeff.	pixelw. acc.
U-Net	0.09	0.91	92.3%
DenseNet	**0.06**	**0.94**	**93.8%**

	mal.	ben.
mal.	**69**	18
ben.	1,193	**3,541**

	mal.	ben.
mal.	**66**	21
ben.	1,081	**3,737**

Table 3: The DenseNet based U-Net architecture achieves a lower dice loss on the test set (top). Confusion matrices for the best model trained on segmented crops (bottom left) and the second best model trained on original data (bottom right) with horizontal predicted labels and vertical ground truth labels.

7. Addressing Racial Biases in Melanoma-Detection AI Models

When exploring the training set from the original ISIC2020 dataset we observed the lack of diversity in the dataset which was mainly composed of lighter skin tone images.

First, we tested the effect of training our model with the original ISIC2020 training dataset on the ISIC2020 test dataset by calculating the results separately across 18 different skin tones (see Figure 3).

Figure 3: Range of 18 Skin Tones, [2]

To do this, we ran the ISIC2020 test dataset through the model and saved all the false negative images. We then wrote a program to capture the skin tone color per false negative test image and for all the benign test images in order to calculate the Sensitivity and the False Negative Rates across 18 different skin tones found in the test set (see Figure 4).

As seen in the results, the False Negative rate for people with darker skin tones (labeled A-C) was much higher than those with lighter skin tones overall (minus the outlier at class P). This suggests that the lack of diverse skin tones in the ISIC2020 training set can result in a much higher rate of not detecting malignant skin lesions for people with darker skin tones.

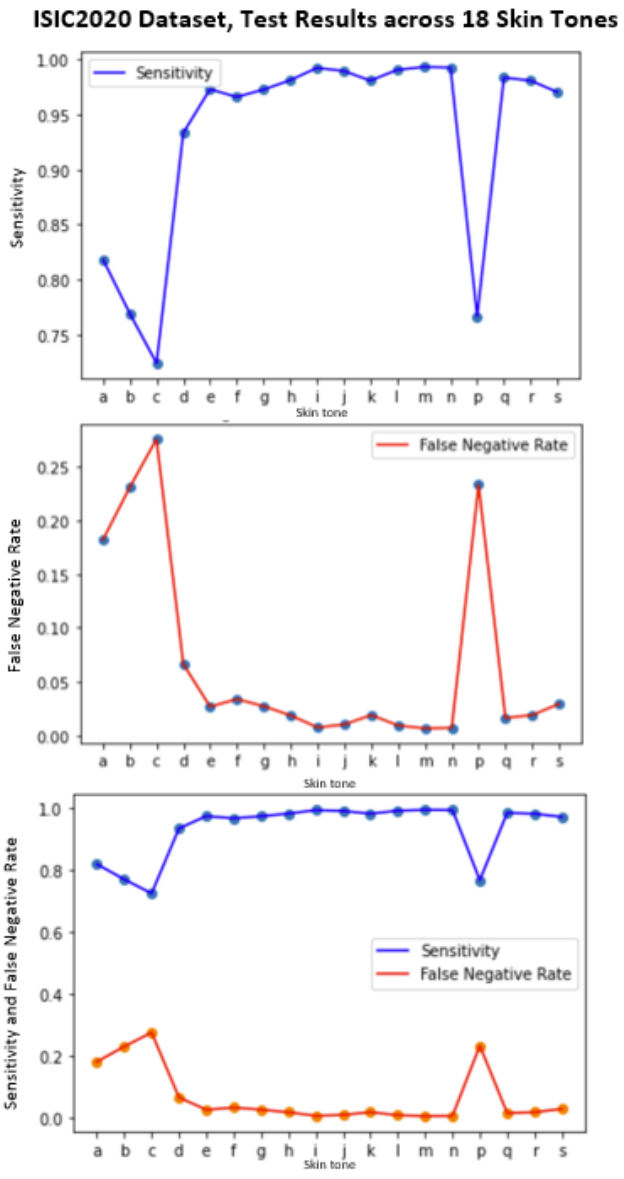


Figure 4: Original ISIC2020 Dataset Sensitivity and False Negative Rate across 18 Skin Tones

We hypothesized that a more balanced training and validation dataset would result in a more balanced Sensitivity and False Negative rate distribution across all skin tones when testing and wanted to research this further.

We took our segmented mole images of the ISIC2020 dataset where the border around the mole was black and wrote a program that randomly samples a skin tone from a range of 18 tones (see figure 3 for skin tone range) and sets that skin tone as the mole background as to replicate the surrounding skin of the mole (see Figure 5).

Using this program, we generated a skin tone train, validation, and test dataset (referred to as balanced skin tone dataset) and then trained our new model feeding in the generated train and validation datasets and used the same model architecture as our best model (number 4).

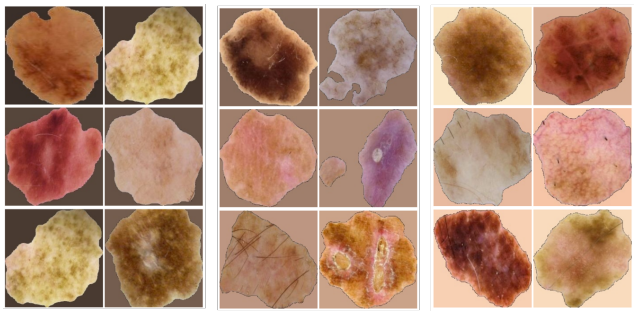


Figure 5: Generation of Balanced Skin tone Dataset across all Skin Tones

We then used our skin tone detection program to calculate the Sensitivity and the False Negative Rates across 18 different skin tones found in the balanced skin tone test set (Figure 6).

As seen in the results, the False Negative rate and Sensitivity were extremely similar across all 18 skin tones and furthermore we had a low False Negative rate overall across all skin tones compared to the ISIC2020 results discussed earlier where the false negative rate was much higher for darker skin tones (see Figure 4).

Our results highlight that skin tone is a feature the model uses to identify malignant images and therefore when training our models it is crucial to ensure that our training set is diverse across all skin tones.

8. Conclusion

In this project we showed that semantic segmentation improves the performance of melanoma classifiers. The best performing model was trained on segmented moles and all pretrained weights were finetuned. The model achieves a balanced accuracy of 77.1%. Data imbalance proves a difficult problem to overcome since the model learns over time to achieve a high accuracy by predicting all examples as benign. We investigated several methods such as data augmentation, class weights and different methods to train the weights but these methods didn't improve performance. Doing a qualitative analysis of saliency maps, we find that the model trained on segmented crops has more fine grained activations inside the mole areas while the model trained on original images activates over larger areas. To improve the results further we propose to sample more malignant examples or combine several datasets together. Another option would be to use multi-class classification and predict different skin conditions instead of grouping them into benign and malignant classes. We also explored the effect of the

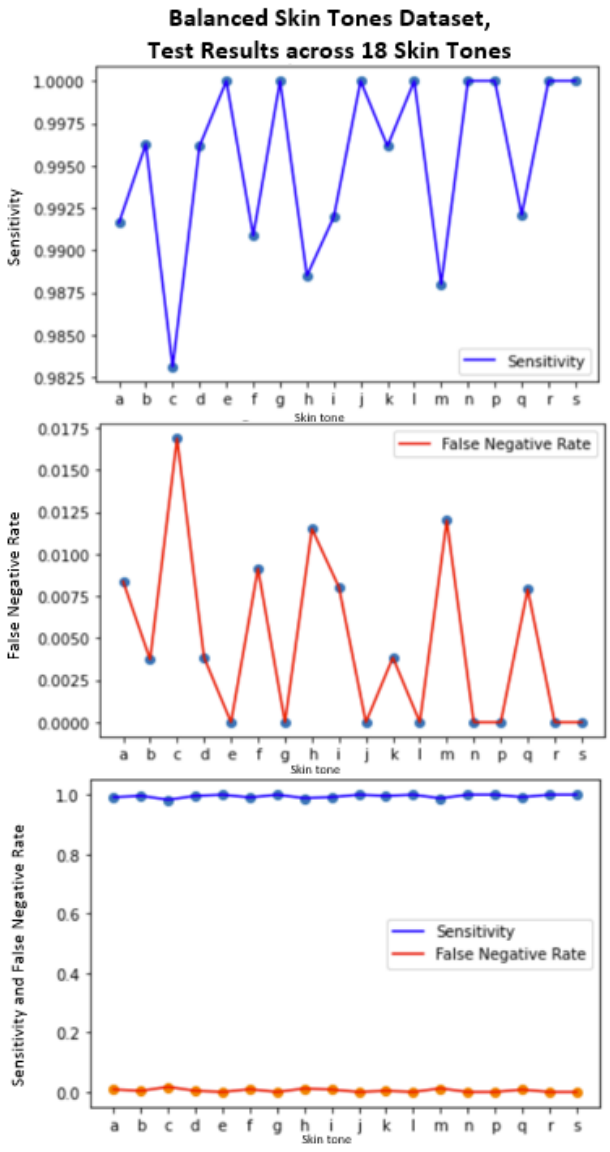


Figure 6: Balanced Skin Tone Dataset Sensitivity and False Negative Rate across 18 Skin Tones

imbalanced (as in lack of diversity) skin tones distribution of the ISIC2020 dataset and how that leads to a much higher false negative rate for individuals with darker skin tones. By generating a balanced and diverse skin tone dataset and retraining our best model with this dataset we were able to achieve high sensitivity and low false negative rates across all 18 different skin tones. This has significant potential to increase early stage melanoma detection for people of color and drastically reduce mortality rates due to late or undiagnosed skin cancer.

9. Contributions

All team members contributed equally to the project. Renasha contributed the problem statement, data augmentations and set up the Google Cloud Platform. Harry coded the baseline segmentation and classification models and ran initial tests. Shelly contributed to the previous work section, and investigated how our method could address the racial biases in melanoma detection models (coded data generation and programs used to test the models' results).

References

[1] Five-year survival rates. http://https://training.seer.cancer.gov/melanoma/intro/survival.html. Accessed: 2021-05-10. 1

[2] A. Akash and A. Mollah, A.and MAH. Improvement of Haar Feature Based Face Detection in OpenCV Incorporating Human Skin Color Characteristic. *Symbiosis*, 2016. 8

[3] M. Al-masni, D. Kim, and T. Kim. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Computer Methods and Programs in Biomedicine*, 190, 2020. 2, 7

[4] Enes Ayan and Halil Murat Ünver. Data augmentation importance for classification of skin lesions via deep learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4, 2018. 2

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] J. Burdick, O. Marques, J. Weinthal, and B. Furht. Rethinking skin lesion segmentation in a convolutional classifier. *Journal of Digital Imaging*, 31(4):435–440, 2018. 2

[7] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[8] A. Filippi, P. Fava, S. Badellino, C. Astrua, U. Ricardi, and P. Quaglino. Radiotherapy and immune checkpoints inhibitors for advanced melanoma. *Radiother Oncol*, 120(1):1–12, 2016. 4

[9] M. Goyal, S. Knackstedt, T.and Yan, and S. Hassanpour. Artificial Intelligence-Based Image Classification for Diagnosis of Skin Cancer: Challenges and Opportunities. *ResearchGate*, 2019. 2

[10] G. Guy, S. Machlin, D. Ekwueme, and K. Yabroff. Prevalence and costs of skin cancer treatment in the u.s., 2002-2006 and 2007-2011. *Am J Prev Med.*, 48(2):183–187, 2015. 1, 4

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[12] Mostafa Jahanifar, Neda Zamani Tajeddin, Navid Alemi Koohbanani, Ali Gooya, and Nasir Rajpoot. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations, 2019. 2

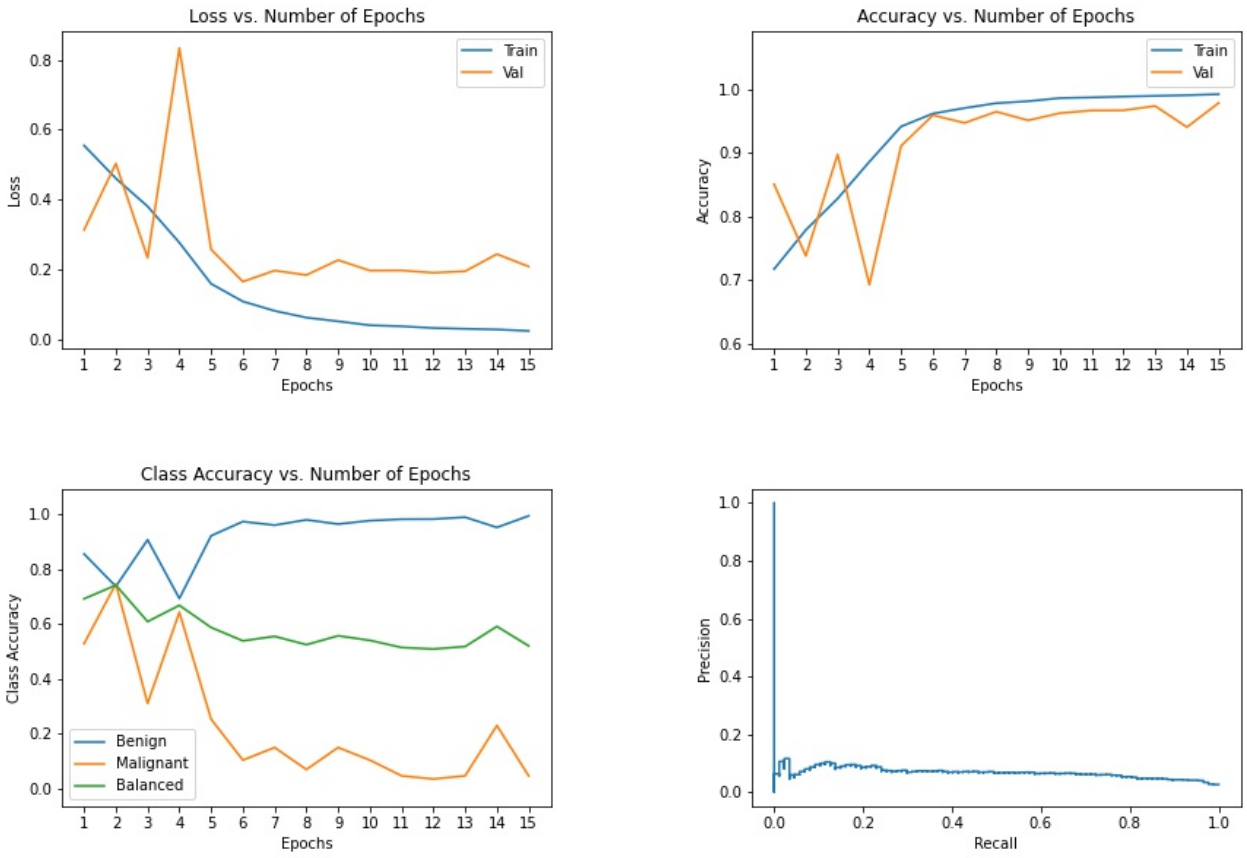


Figure 7: The graphs show that after 15 epochs the loss for the best model converges to zero while accuracy approaches 100%. However, balanced accuracy is on a continued downward trend due to malignant accuracy decreasing and benign accuracy increasing. A similar pattern can be observed with most of our models.

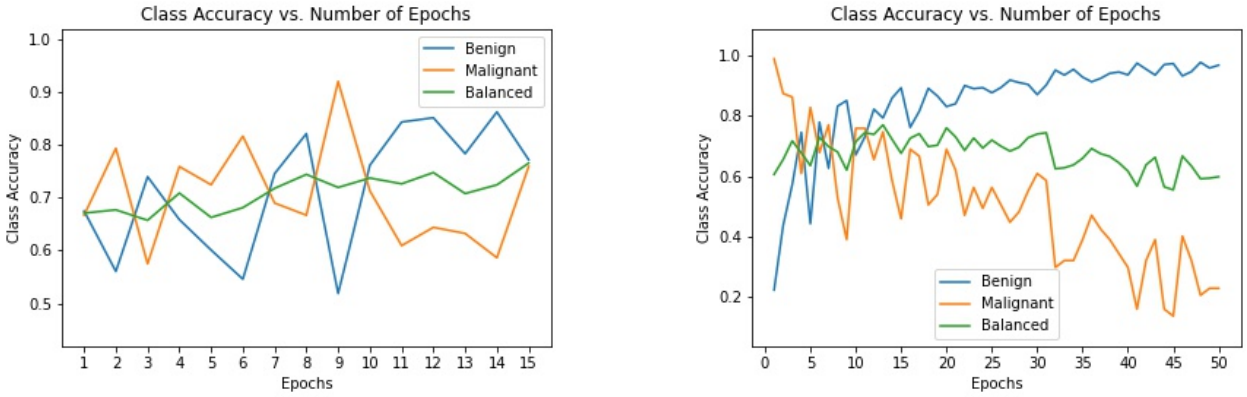


Figure 8: Including augmentations balanced accuracy seemed on a slight upwards trend over 15 epochs with malignant and benign accuracy at a similar level (left plot). To investigate this further, we trained the model including augmentations for 50 epochs (right plot). Unfortunately, after around 15 epochs balanced accuracy went on a downward trend and malignant and benign accuracy started to diverge.

[13] M. Kassem and M. Foaud. Classification of skin lesions using transfer learning and augmentation with alex-net. *PLoS ONE*, 2019. 2

[14] Angela Lashbrook. Ai-driven dermatology could leave dark-skinned patients behind. *The Atlantic*, 2018. 2

[15] R. Maron, A. Hekler, E. Krieghoff-Henning, M. Schmitt, J. Schlager, J. Utikal, and T. Brinker. Reducing the impact of confounding factors on skin cancer classification via image segmentation: Technical model study. *Journal of Medical Internet Research*, 23(3), 2021. 2, 7

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. pages 1–8, 2018. 2, 5

[18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. 2

[19] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and P. Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data*, 8(34), 2021. 1

[20] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*, 5, 2018.

[21] P. Tschandl, C. Sinz, and H. Kittler. Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation. *Computers in Biology and Medicine*, 104:111–116, 2019. 1, 4

[22] P. Tschandl, C. Sinz, and H. Kittler. Domain-specific classification – pretrained fully convolutional network encoders for skin lesion classification. *Computers in Biology and Medicine*, 104:111–116, 2019. 2, 7

[23] Pavel Yakubovskiy. Segmentation models. https://github.com/qubvel/segmentation_models, 2019.

[24] H. Younis, M. Bhatti, and M. Azeem. Classification of skin cancer dermoscopy images using transfer learning. *15th International Conference on Emerging Technologies (ICET)*, 2019.