CS230

Enhancing Melanoma Classifiers with Style Transfer Data Augmentation and Self Supervised Learning

Ben A. Randoing
Stanford University
bar39@stanford.edu
sunID: 06270480

Aastha Jhunjhunwala
NVIDIA
aasthaj@stanford.edu
sunID: 06733973

Anirudh Rao
Samsung Research America
anirao26@stanford.edu
SUID: 06512867

Abstract

In the recent past, deep learning methods have gained popularity in classifying melanoma images, however, most of these classifiers are restrained by the use of lighter skin tone images. In this work, the research team explores Neural Style Transfer as a means to generate darker skin tone images to overcome the bias in preexisting datasets, enhance inclusivity and seek performance improvement on diverse skin tone images. Additionally, the use of self-supervised contrastive learning is explored as a pretraining technique to further improve the performance of the learning algorithm. Recall is used as the primary metric to evaluate the model performance. Neural style transfer data augmentation proved effective in improving the baseline recall from 0.623 to 0.787 in melanoma classification on dark-skin tone images using the baseline model. Self-supervised learning in combination with the NST augmentation further improved the recall to 0.813.

All code is publicly available at: https://github.com/benrandoing20/230_FinalProject

1 Introduction and Related Work

The American Cancer Society (ACS) estimates 100,000 people were diagnosed with melanoma in 2022 in the United States alone [13]. Early detection is essential to minimize the thousands of melanoma related deaths annually. In an effort to detect melanoma as early as possible, health innovators have designed an abundance of melanoma classifiers that feature prediction accuracies nearing those of expert dermatologists. Convolution neural network (CNN) based architectures such as Xception, AlexNet, VGGNet and ResNet have been explored extensively in past implementations providing accuracies as high as 0.92 and similar precision and recall values[18]. Melanoma classifiers often use RGB image inputs with an optimized CNN to classify a skin lesion as malignant or benign. However, the majority of data sets used to train melanoma classifiers are predominantly of light skin tones from the ISIC datasets[17],[18],[19],[20]. The ISIC dataset is primarily curated with skin images from United States, Europe and Australia [21] where light skin colors are more prevalent. The drastic absence of dark skin tone images limits the application of melanoma classifiers and perpetuates institutional discrimination within healthcare. Motivated by this skewness, the team proposes Neural Style Transfer as a technique to artificially generate dark-skin tone images to enhance the applicability and inclusivity of existing melanoma classifiers. As such, the objective of this research work is: 1. To implement data augmentation to develop diverse data sets to improve the performance of melanoma classifiers when applied to dark skin tones. 2. To explore the use of self-supervised pretext tasks to overcome challenges related to the limited availability of annotated data. A recent publication attempted to generate dark-skin tone lesion images for melanoma classifiers using style transfer and deep blending[14]. In this case, generated images were evaluated on the basis of a realness score.

,

2 Datasets

The research team considered it of paramount importance that the development and test sets include real images of varying skill tones. Therefore, the development and test sets are aggregated using data from the Stanford Center for Artificial Intelligence in Medicine Imaging (AIMI) [2]. This data provides 656 images from 570 unique patients. The images were well represented across all Fitzpatrick Skin Type (FST) Scores from I to VI. A score of VI represents darker, more-pigmented skin tones. There were a total of 208 images of FST I-II (159 benign, 49 malignant), 241 images of FST III-IV (167 benign, 74 malignant), and 207 images of FST V-VI (159 benign and 48 malignant). The AIMI data provides real, diverse data (benign and malignant) to evaluate our model. Prior to creating the test and dev sets, 40 benign and 20 malignant data points from the AIMI data were incorporated into the train set to mitigate potential data mismatch conflicts. Subsequently, the test and dev sets were separated and equally represented across each category of FST score and medical diagnosis via random selection in each subgroup. The resulting data sets were randomly separated yet featured equivalent representations of each FST skin tone and distributions of benign and malignant images.

The melanoma classifier was trained with real images and images created through neural style transfer. The real images are predominantly light-skin tone mages from the ISIC Melanoma Classification Challenge in 2020 [1]. The non-light skin tone data is the 60 data points (40 benign, and 20 malignant) chosen from the AIMI data set mentioned above. The data set contains 33,126 (584 malignant) training images. To account for the significant data offset, malignant melanoma cases were duplicated eight times in an attempt to avoid creating a simple model that always predicts benign for every image. Artificially generated dark skin tone data through neural style transfer supplement the training set. Each image depicted in Figure 1 contained 3 channels and was resized to be 299x299x3 in size.

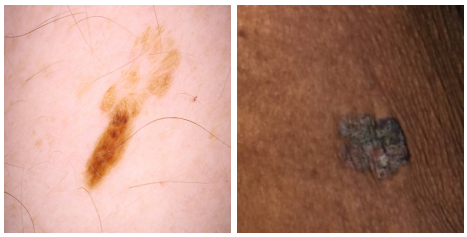


Figure 1: Example of Malignant Melanoma from Light Skin [1] and Dark Skin [2]

3 Modeling Methodology

Our goal is to improve the recall of melanoma classification on dark-skin tone images by supplementing existing light-skin tone data sets with neural style transfer [7] created dark-skin tone images. A promising method for developing the computer vision model is to apply a Convolutional neural network (CNN) [3] and classify RGB melanoma images as malignant or benign comparing it to the annotations by expert dermatologists. Contrastive Learning [9] was explored as a pretext task to enhance the recall of the melanoma classifier.

3.1 Baseline

3.1.1 Convolutional Neural Network Architecture

The baseline was developed from `https://www.thepythoncode.com/article/skin-cancer-detection-using-tensorflow-in-python`. The complete ISIC 2020 training data set was incorporated into the model from the repository. A binary cross-entropy loss function and RMSProp optimizer are applied in tensorflow to create an InceptionV3 model [15]. The 299x299x3 images are fed into a series of convolution layers and a pooling layer before 3 separate inception layers. The final layer implements a sigmoid activation. Transfer learning was implemented starting with the parameters for the InceptionV3 CNN available on ImageNet. As the model was developed, a model checkpoint early-stopping method was implemented. Following each training epoch, the model was evaluated on the dev set. Model parameters were saved only if there

was improved performance on the dev set. Therefore, despite using 100 training epochs as detailed below, the model could converge to optimal parameters at any point during the training process.

3.1.2 Convolutional Neural Network Hyperparameters

CNN training hyperparameters were applied consistently to all models. The adjustable parameters included: batch size (how many samples to update the network parameters with each), shuffle buffer size (the quantity of data points from which a sample to be included in a batch is chosen at random during training), and the number of training epochs (iterations through the complete data set). Based on our preliminary runs on small data samples of the larger population, we decided on the following CNN parameters to use:

- Batch size: 64
- 100 training epochs with > 500 training episodes per epoch
- Shuffle buffer size: 35,000 (intended to exceed population size for a uniform shuffle)

3.2 Neural Style Transfer

3.2.1 Neural Style Transfer Architecture

Neural style transfer (NST) was implemented using the pre-trained VGG19[16] network to generate darker skin tone melanoma images. The content image was used as the starting input image and improved upon iteratively by minimizing the loss values. A dark skin tone color palette was artificially curated using Microsoft paint, 40 images of size 512x512 were generated. For this task, a light skin malignant melanoma image from the ISIC dataset was used as the content image and a darker skin tone image from the curated color palette was used as the style image[12], the images are normalized prior to being sent to the network. The style loss involves the computation of the gram matrix, the product of the given matrix and its transpose, followed by normalization which is done by dividing each matrix element with the total number of elements in the matrix. The standard MSE loss between the gram matrix and the generated output image is used to calculate the style loss. The content loss is computed as the MSE between the content image and the output image.

Figure 2: Sample images from the generated color palette

3.2.2 Neural Style Transfer Hyperparameters

The hyperparameters for NST were determined based on visual inspection of the output image and optimal train time. The content image weight was chosen to be lower than the style image as the input image was chosen to be the content image for faster convergence. The following are the hyperparameter values used for the NST image creation:

- Gradient Descent Optimizer: L-BFGS algorithm
- Content Image Weight: 15
- Style Image Weight: 30,000
- Training Iterations: 2000

It takes about 50s to generate 1 stylized image on one NVIDIA T4 GPU on an AWS EC2 instance.

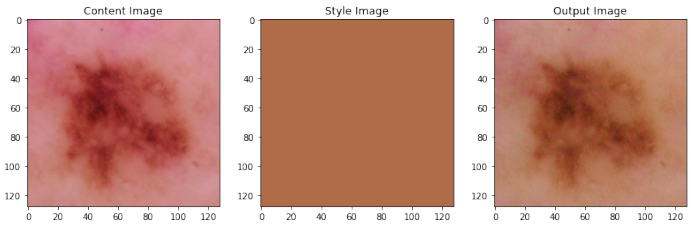


Figure 3: Example of an image generated using NST on a malignant melanoma image

3.3 Self Supervised Learning

3.3.1 Model Architecture

The research team explored self-supervised learning to improve the performance of the baseline classifier. The SimCLR [9] model was adapted from `https://www.kaggle.com/code/aritrag/simclr/notebook`. A pretrained InceptionV3 model was used to derive encodings for a consistent comparison with the baseline models detailed above. First, two random augmentations of each image were created randomly choosing from image flipping, image cropping, grayscale color jitter and Gaussian blur. Each augmentation was then passed through the InceptionV3 model to derive 2048 dimensional encodings after removing the fully connected layer of the pretrained model. Each vector in the latent space was further passed through two dense layers to obtain 256 dimensional projections of the views. The loss function (NTxent) was designed to maximize agreement between augmented views from the same image by combining cosine similarity and softmax. After the pretext training, the projection head was discarded and the learned weights from the network were used to create a binary classifier. For the downstream task, the projection head was replaced with a feed forward neural network with 2 dense layers. Having more than 2 dense layers in the model resulted in overfitting to the training set. A combination of dropout and batch normalization was used to prevent overfitting. Early stopping was used to optimize the training process. Finally, probability scores were obtained using a sigmoid activation. A binary cross-entropy loss was used with each image class weighted inversely to the frequency of occurrence to assign higher weights to the minority class.

3.3.2 Hyperparameter tuning

The hyperparameter tuning was done separately for the pretext task and downstream tasks. The most optimal performance was obtained using the values below:

Hyperparameter Name	Pretext task	Downstream task
Epochs	25	100
Learning rate	0.001	0.0001
Batch size	64	128
Optimizer	Adam	Adam

The final hyperparameter values are based on the optimal validation dataset performance of the model with the NST augmented images. For ease of comparison, the same hyperparameter values were used to train the model without the NST data.

4 Experiments

Since the data set used for training the model is imbalanced, it is imperative to make sure that we measure the performance of the classifier not just using the traditional accuracy metric. Hence, we assign extra emphasis on the recall performance to make sure that our model prioritizes the detection of skin melanoma, when present. A confusion matrix is also implemented to evaluate the model performance on the test data.

Extensive experiments were performed to arrive at the optimal classifier. Each model architecture that was explored involved developing two separate models. As detailed in the data set section above, the malignant data set used to train the second model was augmented with artificially generated dark

skin tone images. The intention was to investigate the model performance of the baseline model architecture with neural style transfer driven data augmentation. Additional experiments involved undersampling the majority benign images to handle class imbalance. Error analysis was performed on the dev set at the end of each training run to identify potential improvements.

5 Results and Discussion

Prior to NST data augmentation with dark-skin tone images, the baseline model performance featured an accuracy of 0.61 and a recall of 0.62. After the incorporation of the NST images, the accuracy of the model was 0.60, and the recall was 0.79. For the self supervised model, the accuracy of the model improved from 0.59 to 0.62 and the recall improved from 0.60 to 0.81. The final precision and F1 score were 0.57 and 0.67 respectively. The presence of neural style transfer augmented images demonstrated elevated recall performance in the classifiers.

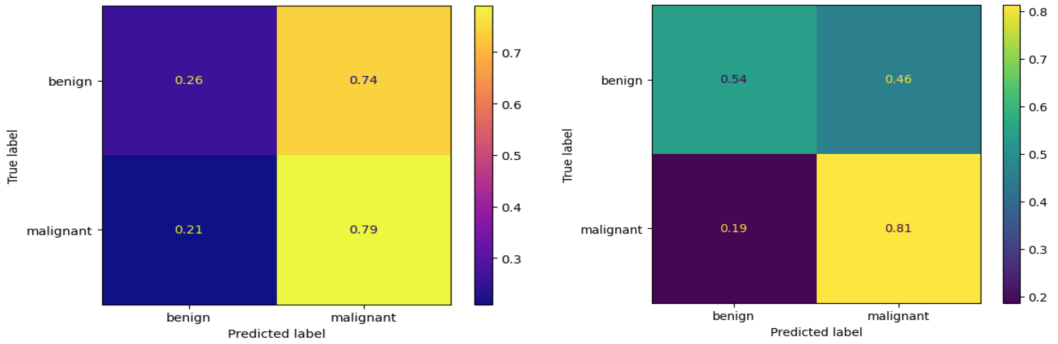


Figure 4: Final Confusion matrices for baseline and SimCLR models, normalized by true label counts

Even though the recall performance of the model is reasonable, the precision and overall accuracy of the model did not achieve a similar improvement. To understand this phenomena, it is important to understand the quality of the dev and test set images. While the Stanford AIMI image database is the only well documented skin lesion database with dark-skin-toned images, the images are not well-cropped around the skin lesions. For example, a white ruler, fingers, and extraneous skin without lesions are common in images. In contrast, the majority of the train set from the ISIC database is cropped closely around a skin lesion to avoid any of the aforementioned artifacts. We understand the reason for the greater quantity of predicted malignant cases in the test set to be the result of the trained model being influenced by the image artifacts and imprecise cropping around skin lesions.

6 Conclusion

Melanoma classification in recent years has become a popular computer vision task for researchers and engineers. While the accuracy of existing classification algorithms improves, the model is likely to be implemented to triage patients with skin lesions in various healthcare environments. However, many literature reviews indicate melanoma classifiers have been developed with predominantly light-skin tone images. The research conducted in this investigation is, to the best of our knowledge, the first to demonstrate neural styles transfer as a means to improve the recall of melanoma classifiers on dark-skin tone lesions.

The team anticipates refining the evaluation techniques on the various models from different data sets in the future. The current neural style transfer created images as depicted in Figure 3 demonstrate a realistic appearance to the untrained observer. As detailed in the results and discussion section above, the current dev and test sets feature images that are not intentionally cropped around a skin lesion. We hope to implement a bounding box lesion detection algorithm to crop the current dev and test set lesion images. We also hope to gather more dev and test set data of dark-skin tone images as the current quantity of images in the dev and test sets is significantly smaller than the number of images in the train set.

7 Appendix

Below is supplemental information referenced in the text above.

7.1 Appendix A: Baseline Model Architecture

type	patch size/stride or remarks	input size
conv	3×3/2	299×299×3
conv	3×3/1	149×149×32
conv padded	3×3/1	147×147×32
pool	3×3/2	147×147×64
conv	3×3/1	73×73×64
conv	3×3/2	71×71×80
conv	3×3/1	35×35×192
3×Inception	As in figure 5	35×35×288
5×Inception	As in figure 6	17×17×768
2×Inception	As in figure 7	8×8×1280
pool	8 × 8	8 × 8 × 2048
linear	logits	1 × 1 × 2048
softmax	classifier	1 × 1 × 1000

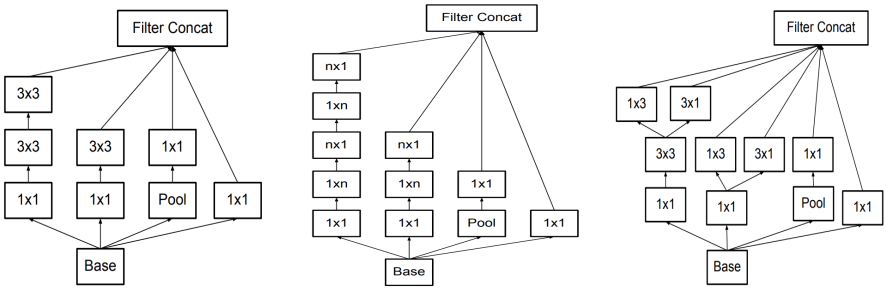


Figure 5: The detailed InceptionV3 model architecture. Figures 6, 7, and 8 are detailed in [15] and are the three small figures above from left to right

7.2 Appendix B: NST Model Architecture

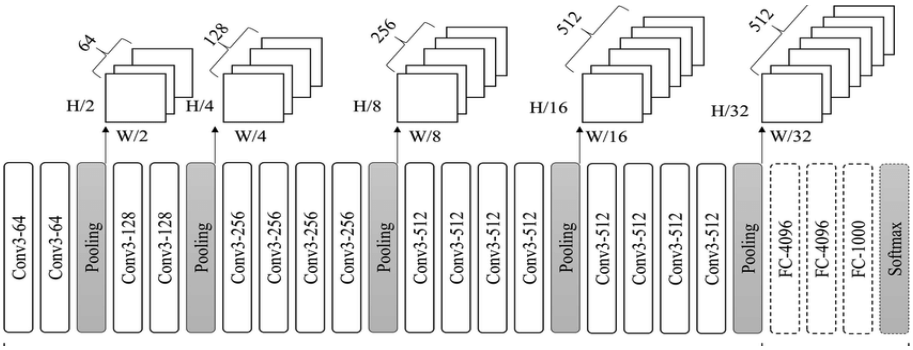


Figure 9: VGG-19 Model Architecture

7.3 Appendix C: NST Examples

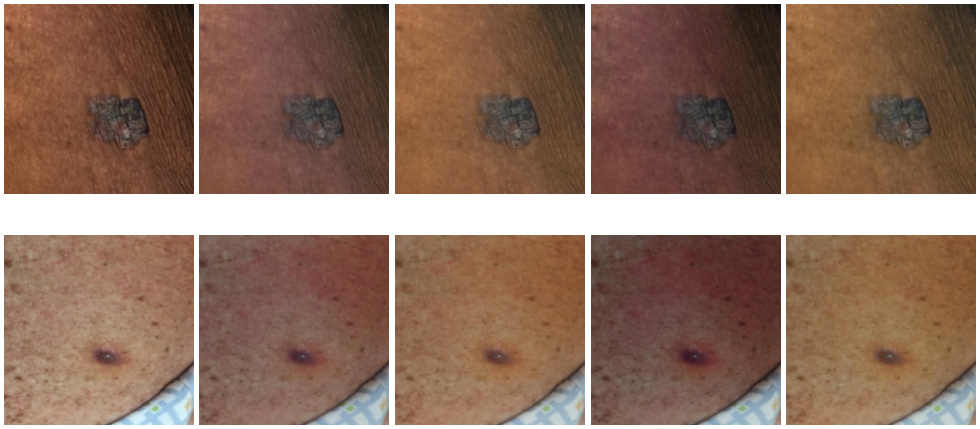
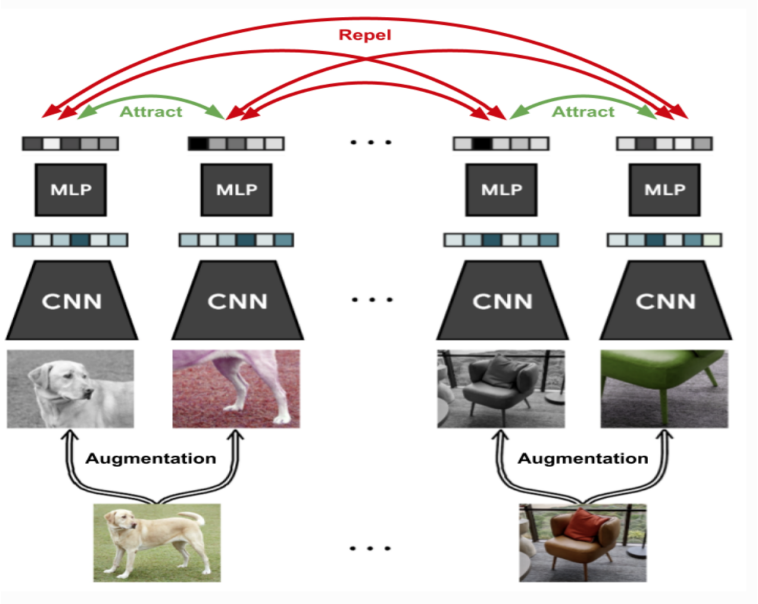
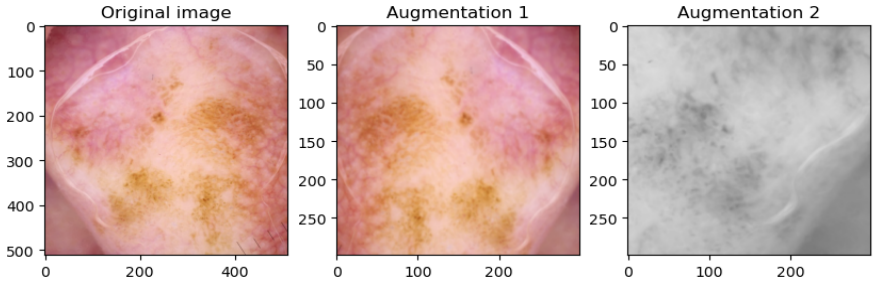


Figure 10: (Left) Original Content Image (Right 4) Generated Stylized Images

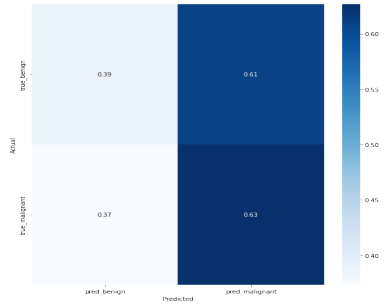
7.4 Appendix D: SimCLR Model Intuition



7.5 Appendix E: Example image augmentations (Horizontal flip and grayscale jitter) applied for SimCLR model

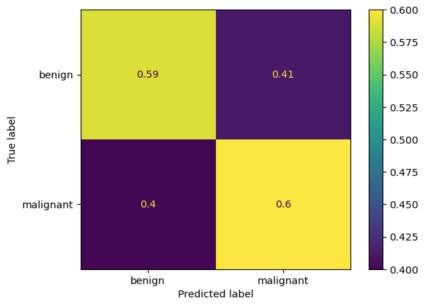


7.6 Appendix F: Baseline pre-NST Confusion Matrix



Confusion matrix for baseline model performance without NST data augmentation

7.7 Appendix G: Pre-NST Confusion Matrix for SimCLR model



Confusion matrix for SimCLR model performance without NST data augmentation

References

[1] International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset. International Skin Imaging Collaboration https://doi.org/10.34970/2020-ds01 (2020).

[2] Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set. Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, James Zou, Albert Chiou. Science Advances (2022).

[3] Esteva, A., Kuprel, B., Novoa, R. et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 (2017). https://doi.org/10.1038/nature21056.

[4] A. Naeem, M. S. Farooq, A. Khelifi and A. Abid, "Malignant Melanoma Classification Using Deep Learning: Datasets, Performance Measurements, Challenges and Opportunities," in IEEE Access, vol. 8, pp. 110575-110597, 2020, doi: 10.1109/ACCESS.2020.3001507.

[5] Kumar A and Vatsa A (2022) Untangling Classification Methods for Melanoma Skin Cancer. Front. Big Data 5:848614. doi: 10.3389/fdata.2022.848614

[6] Xinrong Lu, Y. A. Firoozeh Abolhasani Zadeh, "Deep Learning-Based Classification for Melanoma Detection Using XceptionNet", Journal of Healthcare Engineering, vol. 2022, Article ID 2196096, 10 pages, 2022. https://doi.org/10.1155/2022/2196096

[7] Zheng, X., Chalasani, T., Ghosal, K., Lutz, S., Smolic, A. (2019). Stada: Style transfer as data augmentation. arXiv preprint arXiv:1909.01056.

[8] Krishnan, R., Rajpurkar, P. Topol, E.J. Self-supervised learning in medicine and healthcare. Nat. Biomed. Eng (2022). https://doi.org/10.1038/s41551-022-00914-1

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, "A Simple Framework for Contrastive Learning of Visual Representations". https://arxiv.org/pdf/2002.05709.pdf

[10] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9726-9735, doi: 10.1109/CVPR42600.2020.00975

[11] Misra, Ishan and Laurens van der Maaten. “Self-Supervised Learning of Pretext-Invariant Representations.” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020): 6706-6716.

[12]Rezk E, Eltorki M, El-Dakhakhni W Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach JMIR Dermatol 2022;5(3):e39143 URL: https://derma.jmir.org/2022/3/e39143 DOI: 10.2196/39143

[13]Melanoma skin cancer statistics. American Cancer Society. (n.d.). Retrieved November 27, 2022, from https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

[14]Rezk E, Eltorki M, El-Dakhakhni W Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach JMIR Dermatol 2022;5(3):e39143 URL: https://derma.jmir.org/2022/3/e39143 DOI: 10.2196/39143

[15]Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., amp; Wojna, Z. (2015, December 11). Rethinking the inception architecture for computer vision. arXiv.org. Retrieved December 3, 2022, from https://arxiv.org/abs/1512.00567

[16]Simonyan, K. Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, .

[17] Kaur R, GholamHosseini H, Sinha R, Lindén M. Melanoma Classification Using a Novel Deep Convolutional Neural Network with Dermoscopic Images. Sensors (Basel). 2022 Feb 2;22(3):1134. doi: 10.3390/s22031134. PMID: 35161878; PMCID: PMC8838143.

[18]Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, A comparative study of deep learning architectures on melanoma detection, Tissue and Cell, Volume 58, 2019, Pages 76-83, ISSN 0040-8166, https://doi.org/10.1016/j.tice.2019.04.009.

[19]Devansh Bisla, Anna Choromanska, Russell S. Berman, Jennifer A. Stein, David Polsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 0-0

[20]Marriam Nawaz,Zahid Mehmood,Tahira Nazir,Rizwan Ali Naqvi,Amjad Rehman,Munwar Iqbal,Tanzila Saba, Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering, https://doi.org/10.1002/jemt.23908

[21]Schlessinger DI, Chhor G, Gevaert O, Swetter SM, Ko J, Novoa RA. Artificial intelligence and dermatology: opportunities, challenges, and future directions. Semin Cutan Med Surg 2019 Mar 01;38(1):E31-E37.