Towards Lightweight and Efficient Mobile Melanoma Diagnosis

Adam Sun
adsun@stanford.edu

Maggie Sun
mgsun@stanford.edu

June 2023

1 Introduction

Skin cancer is the most prevalent type of cancer, with melanoma accounting for 75% of all skin cancer deaths. The American Cancer Society estimates that 97,610 new melanomas will be diagnosed in 2023, with almost 8,000 people expected to die. Dermatological medical expertise concurrently does not stretch equitably or fully across global populations, meaning that swathes of people lack methods for detection or prevention before a late diagnosis. Automated methods for diagnosis or classification have the potential to dramatically reduce specialized overhead and extend care to a larger amount of patients. It is thus apparent that early detection and diagnosis with AI can be greatly beneficial in reducing these statistics.

Early attempts to automate diagnosis of melanoma yielded limited accuracy or a lack of general capability. [10] Results might have been experimentally similar to dermatology accuracy, but the lack of generality meant that methods optimized for experimental settings might not apply to real-life situations. Recent studies have shown that deep convolutional neural networks can achieve dermatologist-level efficiency in classifying skin cancer images. By using transfer learning with a dataset of 129,450 clinical images on a pretrained CNN, Esteva et al achieved skin cancer classification accuracy comparable to or better than dermatologist opinion. [5] Other approaches with similarly optimistic results exist: Haenssle et al used transfer learning on Google's Inception v4 model, fine-tuning the weights at every layer, and their model surpassed the majority of 58 dermatologists who participated for comparison. [6] Despite these successes, many of these models rely on heavy preexisting architectures or many feature vectors and layers. The accuracy of heavy highly optimized models is not as applicable to practical situations for patients without access to institutionalized care.

While prognoses are better in economically developed nations, a report by the World Health Organization estimates that skin cancer is diagnosed in one out of three people worldwide. Without methods for early detection or awareness, melanoma is far more dire for patients with less access to medical care or specialized skin cancer expertise. Patients in less fortunate areas may not have access to experienced doctors who are able to help them diagnose their lesions. While deep learning models are able to achieve high accuracy on these data, many of them require huge amounts of compute to generate predictions, which is not feasible for patients with access to more basic technology such as smartphones. There thus arises the need for a more lightweight and affordable way to diagnose melanoma from images.

Attempts to make these models faster, more lightweight, or suitable for usage on mobile phones have shown success in the past. Using a pretrained MobileNet CNN with 10,015 dermoscopy images, Chaturvedi et al were able to achieve a high classification accuracy with a lightweight architecture, comparable to the accuracy of dermatologist diagnosis. [4] An app proposed by Abuzagleh et al with the purpose of assisting in melanoma prevention and detection is designed to take dermoscopy images from a mobile phone camera, preprocess and normalize, and feed the image through a classifier for the user to self-diagnose skin lesions. [1]

Applications like this are potentially useful because of of their ability to extend medical expertise or preventative methodologies to patients who otherwise do not have access to specialized care. Accordingly, in the interest of generality, speed, and accessibility, we wanted to train a lightweight model with mobile applications that could classify the prognosis of a skin lesion, allowing for an equitable yet convenient way to observe if a patient requires more care.

In this work, we propose and train a model that can diagnose melanoma from an image. We propose a framework to mitigate the imbalance in our data. In addition, by experimenting with different models of different parameter counts and inference times, we observe that our proposed framework results in a good AUC on melanoma images.

2 Dataset

We use the publicly available SIIM-ISIC Melanoma Classification Dataset, which contains 33,126 unique dermoscopic images of benign and malignant skin lesions from over 2,000 patients. Each image is annotated and labeled with a classfiication of melanoma via a histopathological diagnosis or expert agreement. The dataset was generated by the International Skin Imaging Collaboration with images from a number of global hospitals. [11] Since we plan to utilize models pretrained on ImageNet in our approach, we resize all images to 224x224x3, downsizing the images to a manageable size.

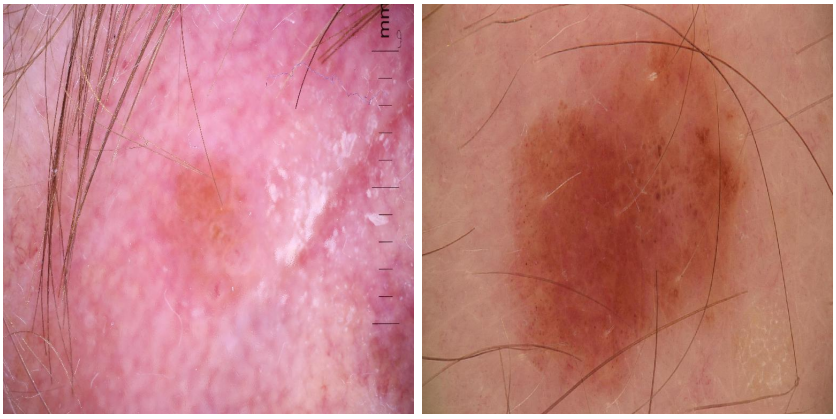

Figure 1: Benign (left) and Malignant (right) examples taken from our dataset.

We note that there is a severe class imbalance in the dataset. Notably, there are only 500 instances of melanoma in the dataset.

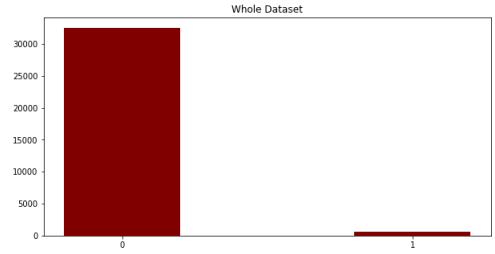

Figure 2: Class distribution in our dataset

This severe class imbalance can cause significant bias in our models. Models trained on imbalanced data frequently tend to predict the majority class, resulting in a model that is essentially useless [3]. In order to prevent this, we utilize data augmentation and focal binary crossentropy loss to train our model. We elaborate more in depth on the methods we use in section 4.2.

3 Methods

3.1 Setup

To accelerate the training process, the model was developed and trained on a Google Cloud instance. Our model was trained on a Tesla T4 GPU. Training took approximately 2 hours.

3.2 Baseline

We use a radiomics approach with computational features as our baseline. We use a sample of the data, using only 5000 examples. We calculate global shape, texture, color descriptors (Hu Moments, Haralick features, Color Histogram). After normalizing all the features with a min-max normalization scheme, we train a support vector

Augmentation	Amount/Probability
Horizontal Flip	0.5
Vertical Flip	0.5
Rotation	[0, 90]
Zoom	[0, 0.2]
Shear	[0, 0.2]
Saturation Shift	[0.4, 0.6]
Brightness Shift	[0.7, 1.3]

Table 1: Data Augmentation Methods. By imposing flips with a probability of 0.5 and rotation up to 90 degrees, we hope to create more spatial invariance in our model, since malignant skin tumors that are upside down are still malignant. In addition, we impose zoom and shear to further emphasize the spatial invariance of our model. To create invariance to color, we utilize hue, saturation, and brightness shift.

machine model on the resulting features, comparing it with the other models. We perform 3-fold cross validation and perform grid-search hyperparameter search across kernel type and C.

3.3 Mitigating Class Imbalance

Common methods to mitigate class imbalance include resampling methods, namely oversampling and undersampling. Oversampling involves sampling additional examples from the minority class – however, this can cause overfitting as the model sees the same example multiple times per epoch. Undersampling involves sampling less from the majority class - however, this is also not desirable since this results in using less data, potentially wasting valuable insights.

According to [3], convolutional neural networks do not significantly overfit when trained on an oversampled dataset, making oversampling with data augmentation a valid method to mitigate this class imbalance. Therefore, we utilize oversample the malignant examples to half the size of the benign examples, causing the malignant examples to make up 33% of our data. In order to prevent overfitting from reusing examples, we utilize data augmentations to create more variance in our data. The data augmentations we use, as well as the ranges of augmentations, can be found in Table 1.

Our data augmentations create spatial and color variance, which allows our model to become more robust to these changes. In addition, by imposing these augmentations, our model is less likely to overfit on our limited dataset, allowing it to potentially generalize well to our test data, as well as new datasets and beyond. To demonstrate the performance of our augmentation approach, we demonstrate numerous augmentations of the same image in 3.

To further mitigate the imbalance in our data, we utilize a class-weighted binary focal crossentropy loss [9] with a gamma of 2. This loss focuses training on hard examples (i.e malignant images), preventing the model from being overwhelmed and biased by the abundance of negative examples in our dataset. We found this loss to be the most effective for preventting overfitting on our extremely imbalanced dataset.

3.4 Model

MobileNetv3 [8] is a model that uses depthwise separable convolutions, linear bottleneck, and inverted residual structure to create a model that is not only accurate, but also efficient. MobileNetv3 is able to achieve a 67.4% accuracy on ImageNet while only taking 14.4 seconds to make an inference on a Google Pixel 3 phone [8]. This efficiency, combined with the reasonable accuracy, make the MobileNetv3 model a perfect candidate to serve as the backbone of our model.

Our model consists of a MobileNetv3 model pretrained on ImageNet. Then, we add fully-connected layers of 128 and 1 node on top of the frozen MobileNet model. By freezing the MobileNet model, we perform fine-tuning, allowing the frozen MobileNet model to serve as a feature extractor [2].

MobileNet's small inference latency that has been optimized on mobile devices allows it to quickly output a feature vector that serves as input to our fully-connected layers. To serve as baselines, we also experiment with other state-of-the-art CNN models, namely EfficientNetv2, ResNet [7] and VGG-16 [12]. These models, while obtaining a larger accuracy on ImageNet, are larger and thus cannot run optimally on mobile devices. Nevertheless, we compare their performance to serve as baselines, taking into account their number of parameters and inference time to serve as a fair comparison between these models.

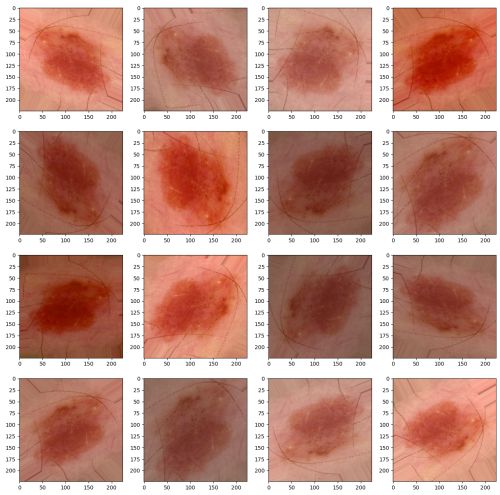

Figure 3: Data augmentation results on a single malignant image. Notice that the images are rotated, and the brightness and saturation varies across images. No two images are exactly the same. Despite the augmentations, all images are realistic instances that could occur in the dataset, thus justifying our approach.

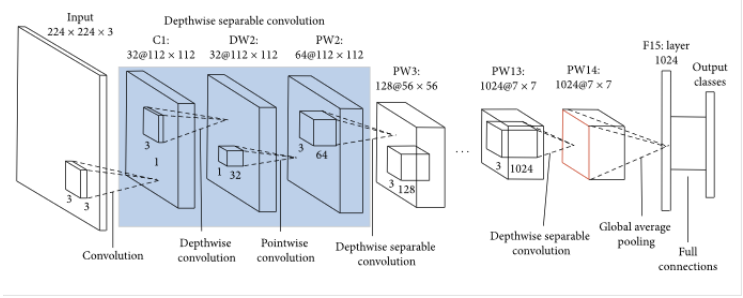

Figure 4: MobileNet Architecture. Notice that the model uses depthwise separable convolutions instead of regular convolutions.

3.5 Training

We split our dataset into 60% train, 20% val, and 20% test. Since we are dealing with imbalanced data, we utilize stratified sampling, ensuring that the ratio of classes is even across all subsets of our data. We trained our model for 8 epochs with a batch size of 32. We use the SGD optimizer with a learning rate of $3e-4$ and a momentum of 0.9. The epoch with the highest validation AUC is kept.

4 Results

We run our test data through our models. The metrics we use include accuracy, precision, recall/sensitivity, specificity, and AUC score. Due to the imbalanced nature of our data, we focus on AUC score as our main metric. To measure the efficiency and lightweightedness of each model, we also factor the number of parameters and the inference time in our comparison of each model. See results in Table 2 and AUC in Figure 5.

5 Discussions

From Table 2 and Figure 5, we can see that MobileNet is the model that has the lowest number of parameters and has the lowest inference time. Bigger models like VGG16, ResNet50, and EfficientNetv2B0 do not significantly

Model	# of Parameters	Inference Time (s)	Precision	Recall	Specificity	AUC Score
SVM baseline	-	-	0.2593	0.3874	0.3942	0.3993
MobileNetv3Small	1,602,929	69.737	0.0841	0.4615	0.8936	0.8284
EfficientNetv2B0	7,394,129	105.477	0.0973	0.2735	0.9375	0.8157
ResNet50	25,947,265	94.390	0.0926	0.5385	0.8892	0.8304
VGG16	15,304,769	71.710	0.0467	0.5299	0.7913	0.7507

Table 2: Comparison between model based on different feature extractors. Inference times are measured on a Tesla T4 GPU and reflect the time it takes to make inferences on the set of 6626 examples.

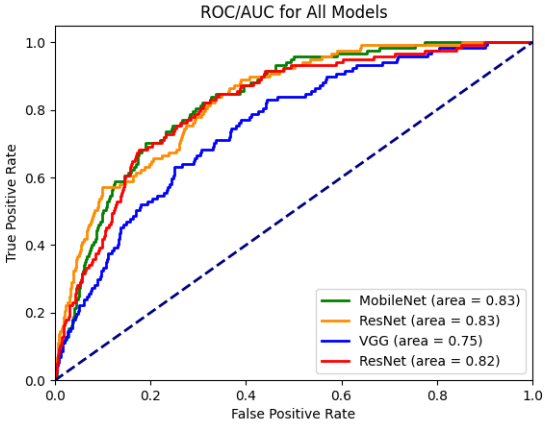

Figure 5: Comparison of ROC/AUC across each different model. We notice that MobileNet performs among the best for AUC, despite its small size and low inference time.

outperform MobileNet in any metrics except recall. As such, we can see that MobileNet allows for maximal performance while being extremely efficient, demonstrating that it is the best feature extractor to use in a mobile application.

Interestingly, VGG16 performs significantly worse than all other models despite its large number of parameters. We attribute this poor performance to the nature of feature extraction – while bigger models like VGG16 may have higher performance on ImageNet, this better performance does not necessarily reflect in our new task, which seems to actually benefit from less precise, more generalizable features.

Our SVM baseline does not perform well relative to the deep learning methods, demonstrating that deep learning approaches using semantic features are superior to computational features.

When dealing with our extremely imbalanced dataset, our results were not promising at first. Our model tended to always predict the majority class, resulting in a very high accuracy but a very low precision, recall, and AUC. We found that our efforts in applying focal loss and in utilizing oversampling with image data augmentation techniques allowed our models to reduce their bias, resulting in a better predictor as a whole.

The high specificity of our model means that our model will have a high number of true negatives and a low number of false positives. This is important, because people who have a positive diagnosis for skin cancer are subject to more expensive testing and treatment.

However, our results also leave room for improvement. In particular, our model's recall is only 0.4615, which does not perfectly align to our problem statement. Ideally, our model should be able to successfully diagnose most positive cases, since a negative diagnosis will instill a false sense of security in patients. The cost of a patient being granted a false sense of security is massive, as they may delay a doctor's visit, resulting in the cancer being worse.

6 Conclusions and Future Work

In conclusion, in this work we train and evaluate a lightweight and efficient model that can diagnose malignant melanoma from images. By taking into account the number of parameters and inference time of each different model, we decide that MobileNetv3Small is the best model that minimizes the inference time and number of parameters

while maximizing performance. The lightweight nature of our model allows it to be able to run on smartphones. Being able to run locally allows this model to be used in areas without access to internet or expensive computing resources, helping make AI more equitable.

In the future, we would like to work on making our model have a higher recall. We largely attribute the low recall of our model on the imbalanced nature of our dataset. So, in order to create a dataset that achieves a higher recall, we would need to collect more examples of malignant melanoma to resolve the imbalance in our data through real reliable examples.

Another future step is to deploy our model in the real world. In order to do so, we would like to convert and deploy our model for use with mobile devices using Tensorflow Lite. Ideally, patients would be able to install an app on their mobile phone, take a picture of a concerning lesion, and get a prognosis of potential severity from our classifier. In order to make this concept a reality, we would need to educate the user on how to take well-lighted images that can be properly interpreted by our model. Ideally, a user with just a smartphone should be able to take similar images to the ones that our model has been trained on.

Finally, generality and applicability to patients is a concern in the work that we've done, especially with the desire to extend this work to those lacking specialized medical care. In particular, the dataset we worked with consists of only images of lighter skin, and thus the model is optimized and trained for use with lighter-skinned individuals. This is a concern that should be addressed both data- and methodology-wise. In the future, we would prefer to train our models on a dataset of diverse medical data across all skin types and colors.

References

[1] Omar Abuzaghleh, Miad Faezipour, and Buket D.Barkana. SkinCure: An innovative smart phone based application to assist in melanoma early detection and prevention. *Signal &amp Image Processing : An International Journal*, 5(6):01–15, dec 2014.

[2] Aidan Boyd, Adam Czajka, and Kevin W. Bowyer. Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch? *CoRR*, abs/2002.08916, 2020.

[3] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, oct 2018.

[4] Saket S. Chaturvedi, Kajol Gupta, and Prakash S. Prasad. Skin lesion analyser: An efficient seven-way multiclass skin cancer classification using mobilenet. *Advanced Machine Learning Technologies and Applications*, 2020.

[5] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 2017.

[6] H.A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018. Immune-related pathologic response criteria.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3, 2019.

[9] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[10] Barbara Rosado, Scott Menzies, and Alexandra Harbauer. Accuracy of computer diagnosis of melanoma. *Arch Dermatol*, 139, 2003.

[11] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, Allan Halpern, Brian Helba, Harald Kittler, Kivanc Kose, Steve Langer, Konstantinos Lioprys, Josep Malvehy, Shenara Musthaq, Jabpani Nanda, Ofer Reiter, George Shih, Alexander Stratigos, Philipp Tschandl, Jochen Weber, and H. Peter Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8(1), January 2021.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.