EQUITABLE LIGHTWEIGHT MELANOMA PRESCREENING WITH DEEP NEURAL NETWORKS

By

Caleb Matthew Grenko

A THESIS

Submitted to
Davidson College
in partial fulfillment of the distinction of

Thesis for the Center for Interdisciplinary Studies – Bachelor of Science

2021

APPROVAL

The thesis of Caleb M. Grenko is approved:

Deborah Thurtle-Schmidt Date
Research Advisor

Michelle Kuchera Date
Research Advisor

Raghu Ramanujan Date
Research Advisor

ABSTRACT

EQUITABLE LIGHTWEIGHT MELANOMA PRESCREENING WITH DEEP NEURAL NETWORKS

By

Caleb Matthew Grenko

Melanomas are the most deadly form of skin cancer and early detection of malignant melanoma lesions is critical to improving patient outcome. Self-screening is universally recommended, yet it is still difficult for the layperson to correctly identify suspicious moles. Recent efforts in medical computer vision have attempted to remedy this problem by creating algorithms that are able to automatically classify a skin lesion into benign (noncancerous) or malignant (cancerous). However, these algorithms tend to be largely focused on achieving research-oriented goals such as obtaining the best performance at a competition or finding novel approaches to image classification. This paper will evaluate the feasibility of a small model deemed a *Lambda EfficientNet* capable of high performance on while still being lightweight showing potential utility in mobile prescreening applications. Additionally, automatic clustering of the data reveals patterns of performance disparities on varying groupings of images, as showing the feasibility of hand-crafted data augmentations in mitigating performance disparities cross these clusters. Two clustering techniques (*feature-extraction clustering* and *RGB-mean clustering*) revealed distinctive clusters of images. Model performance on each cluster showed that there is a strong positive correlation between the number of training samples within the cluster and the performance of models on that cluster. Additionally, augmenting the images to show varied skin tone not only did not improve model performance–instead it negatively impacted the ability of models to detect melanomas. While raw performance is most commonly evaluated when discussing machine learning models, it is important to bring a more nuanced discussion of equitable performance when dealing with data that implicitly involves skin color. Additionally, while some augmentations are able to help models achieve better and more generalized performance, the inclusion of racially diverse and representative data cannot be replaced.

ACKNOWLEDGMENTS

This thesis is dedicated to all those who taken the care to shape who I am today. To my parents, thank you for raising me to be passionate and curious. To my mentors, thank you for your patience and pushing. To my friends, thank you for being my everyday role models. And to you, the reader, thank you for taking the time to contemplate this work. Time is very valuable, and I appreciate all those that have invested in me thus far.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

The goal of this research is to begin to evaluate deep learning as a tool for user-friendly prescreening of melanoma as well as expand on usual techniques of evaluating model performance through the assessment of subgroups of images clustered through visual characteristics. This thesis will emphasize the need to evaluate the impact of skin tone on models rather than just the overall performance, thus leading to tools which can more equitably serve marginalized users.

1.2 Background

This thesis consists of elements from the fields of Biology, Medicine, Machine Learning, Statistics, and of course Ethics. This section will provide the pertinent background across this range of disciplines that a reader needs to appreciate this thesis.

1.2.1 Bioinformatics

Bioinformatics lies at the intersection of biology and computer science. More specifically, bioinformatics focuses on the development of tools that enable the analysis of biological data. Bioinformatics is commonly used to analyze genomic, transcroptomic, and proteomic data in order to discover genes associated with diseases and potentially identify targets for treatment [1]. Since the completion of the Human Genome project [2] bioinformatics has been an invaluable tool into understanding molecular interactions within cells, but the scale of biological data is not limited to molecular mechanics. While many equate bioinformatics with genomics, biological data comes in a wide variety of forms such as micromolecular data [3], radiographs [4], whole slide histopathology images [5], and miscellaneous data such as health records [6] or gross images of lesions [7].

1.2.2 Machine Learning

Machine learning is a field which focuses on building algorithms that can learn from data and improve their performance through time and experience. This field tackles tasks such as regression, classification, segmentation, and many others applications [8] and has become a popular tool for biomedical data analysis.

1.2.3 Deep Learning

Deep learning is a discipline within machine learning which aims to create algorithms which can learn from data via deep neural networks. The history of artificial neural networks dates back to 1957 when a psychologist named Frank Rosenblatt [9] created the single-layer neural network dubbed the Perceptron- a machine capable of learning from data and outputting a binary classification which served as the foundation for learning algorithms. Soon Alexey Grigoryevich Ivakhnenko and Valentin Grigor'evich Lapa [10] built upon Rosenblatt's work by transforming the Perceptron into a hierarchical multi-layer network (later known as a "deep" network– referring to the presence of hidden layers within the network). Over the years this work has been expanded greatly and deep learning has been applied with great success to tasks such as computer vision, natural language processing, reinforcement learning, and many other tasks. The two deep learning advances most relevant to this thesis are the fully connected network (*FCN*) [11] and the convolutional neural network (*CNN*) [12].

1.2.4 Fully Connected Networks

Deep neural networks work by transforming an input into an output via successive transformations called *layers*. Layers can come in a variety of forms with the simplest being the fully connected layer. A fully connected layer can be abstracted as a linear function $\Phi(\boldsymbol{x})$ that takes an input $\boldsymbol{x}$ and maps it to an output y through the equation:

$$\Phi(\boldsymbol{W}, \boldsymbol{x}, \boldsymbol{b}) = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

where $\boldsymbol{x}$ is an input vector, $\boldsymbol{W}$ and $\boldsymbol{b}$ are parameters of the layer called the *weights* and *biases*, and σ is a function that scales the output of $\boldsymbol{Wx} + \boldsymbol{b}$ to $[0, 1]$.

A fully connected *network* is a series of these layers which take an input $\boldsymbol{x}$ and map it to an output $\boldsymbol{y}$. Each layer has multiple neurons, each with their own weight and bias, and has each neuron in a layer connected to each neuron in the next layer, making $M \times N$ connections for two layers of size M and N.

Additionally, a fully connected network with L layers can be represented mathematically via successive transformations with fully connected layers:

$$f(\boldsymbol{x}) \ = \Phi_{L-1}(\Phi_{L-2}(...\Phi(\boldsymbol{X}, \boldsymbol{\theta_0}...)\theta_{L-2})\theta_{L-1})$$

Where Φ_N is the *Nth* layer's linear function and θ_N is the parameters (weights and biases) of the Nth layer.

Fully connected networks are generally considered "universal approximators," meaning that they can learn and generally approximate any function mapping a fixed-size input to a fixed-size output given an arbitrarily sufficient width and depth. The term "fully connected" refers to the fact that every neuron in a layer is connected to every neuron in the next successive layer.

Consider an instance where a user passes a $28_{px} \times 28_{px}$ square image of a handwritten digit into a fully connected network for classification. Since fully connected networks require a vector of activations as an input, then the *FCN* would first *linearize* (or flatten) the square image input into a $784x1$ single-column vector before propagating the values through the network. While *FCN*s have the advantage of being incredibly versatile, it comes at the cost of ignoring valuable information about the structuring of the data, and thus are said to be *data agnostic*.

1.2.5 Convolutional Neural Networks

While the spatial layout of values within an input are ignored by a fully connected network, convolutional networks are specifically designed to incorporate spatial information into their processing.

The main component of the convolutional neural network is the *convolutional layer*. However, rather than using the matrix multiplication of weights ($\boldsymbol{W}$) by the layer input ($\boldsymbol{x}$), it uses a convolutional operator and can be written as:

$$\Phi(\boldsymbol{W}, \boldsymbol{x}, \boldsymbol{b}) = \sigma(\boldsymbol{W} * \boldsymbol{x} + \boldsymbol{b})$$

where $*$ denotes a convolutional operation of input $\boldsymbol{x}$ by weights $\boldsymbol{W}$. This convolutional operation incorporates spatial weighting of values and can be done over an arbitrary number of dimensions within a tensor (a uniformly typed multidimensional array) making it incredibly useful for computer vision tasks.

Additionally, convolutional neural networks can be viewed as a combination of two distinct components: *feature extraction* and *classification*. Feature extraction typically consists of a series of convolutional layers and other operations such as batch normalization and max pooling. As the name implies, feature extraction takes the initial input and outputs a set of *latent features* that describe the input through learned numerical representations. From here, the classification portion of the network (typically a series of fully connected layers) weighs the output features and uses them for classification.

Many variations upon the typical convolutional neural network exist for many different tasks (segmentation, regression, generation, etc.), but these techniques are beyond the scope of this thesis and we will instead focus on network as a means of classification.

1.2.6 Data Augmentations

The term *data augmentation* refers to altering data during training to create models with more accurate or generalizable results. For example, if one were training a model to classify species of birds, it would be helpful to not just train the model on the original images, but also flip and rotate the images, insert noise, and use other data alterations such as cropping and ablations to force the model to learn more robust representations of the birds [13].

1.2.7 Medical Informatics

Recently computer vision (*CV*) techniques have been applied to medical images for bioinformatic analysis. Advances in pathology and radiology have led to greater volumes of medical imaging data at higher resolutions than ever before. This influx of high-quality data has rapidly expanded the field of medical image analysis and led to applications of both traditional computer vision methods as well as the development of more advanced deep learning computer vision methods.

1.2.8 Melanomas

Skin cancer is the most common form of cancer in the United States [14]. Skin cancer is a class of diseases that can be further stratified into basal cell carcinoma (*BCC*), squamous cell carcinoma (*SCC*), and melanomas. Currently melanoma is estimated as the fifth most common cancer in men and the sixth most common cancer in women in the United States [14]. When detected early, melanomas are highly treatable and only require minor surgery. However, melanomas can rapidly metastasize (spread to other organs) to lymph nodes, the lungs, the liver, or other critical organs such as the brain. This means that early detection and intervention is essential to improving the prognosis of melanomas.

The most common sites of melanoma incidence are dependent on sex. For males the majority of cases are located on the back, whereas melanomas are most likely found on arms and legs for women [14, 15]. Additionally individuals from white populations are ten times more likely to be diagnosed with a cutaneous (on the skin) melanomas than black, Asian, or Hispanic individuals [14]. More specific figures estimate 27.5 new cases of melanoma per 100,000 in white populations as opposed to 1.1 per 100,000 in black populations [16]. A notable exception are plantar (on the bottom of the foot) melanomas, which occur at identical rates in white and black populations [14], and non-white populations are at a higher risk for mucosal melanomas [14].

1.2.9 The Data

In May 2020 the Society for Imaging Informatics in Medicine (*SIIM*) and International Skin Imaging Collaboration (*ISIC*) collaborated to publish a large data set (*"SIIM-ISIC"*) on the competition platform *Kaggle* which consisted of 44, 108 images of melanomas and moles along with associated information such as patient age, sex, and the anatomical location of the image [17]. This data set was the result of the multi-institutional collaboration between six hospitals from across the world. This data set came as an expansion upon the smaller ISIC-2019 data set [17]. Additionally, this data set was novel in that it is the first to include multiple moles for individual patients in an attempt to make the data a more similar representation of real dermatological evaluation where multiple moles are evaluated for one individual. All images and diagnoses were validated via retrospective review which included reviewing the pathology report. Plausible diagnoses were referred to expert dermatopathologists for consultation. Each data entry (image, approximate age, sex, and general anatomical site) was labeled as either benign or malignant. Malignant diagnoses included both melanoma *in situ* as well as invasive melanomas. All other diagnoses were designated as benign including severely dysplastic nevi [17].

1.2.10 Model Bias in Machine Learning

Model bias (not to be confused with *weights and biases*) refers the general idea of how well a model is fit to a set of data. A model that does not adequately capture the trends of the training data is said to be *highly biased*. Conversely, a model that is highly suitable for predictions on its training data is said to have *low bias*.

1.2.11 Models Inherit Biases

Machine learning algorithms can also inherit the biases (in *intolerance* sense) of the actual data they were trained on. A common example is that of a model which was built to predict who a company would want to hire [18]. This model was trained on past resumes of both successful company hires and rejected applicants. This historical data inherited the biases of former recruiters, particularly in that *female applicants be almost automatically rejected.* Any phrases found containing "women's" would penalize the application, even if it was "women's chess club captain," and thus in 2018 the project was cancelled. The machine learning algorithm successfully learned from the data and inherited the implicit bias of the past [18].

In a more medically oriented example, one study showed that there are sex, ethnic, and racial disparities in the care and outcomes of patients hospitalized for coronary artery disease ("*CAD*") [19]. The analysis showed that compared to men, women are less likely to receive optimal care for CAD and ultimately had a significantly mortality rate. Additionally, black patients were significantly more likely to die than white patients. If a machine learning algorithm aimed at predicting the severity of CAD was trained on this data, it would pick up on these sex-based and race-based patterns and predict more severe disease for black and female patients than it would for white male patients [19].

If algorithms are able to learn from the aforementioned records and mistakenly learn bias in scenarios such as CAD, then algorithms that directly make inferences from the skin should be thoroughly vetted and understood for how they are influenced by skin-tone.

1.3 Scope and Significance

The synthesis of all of this background is to use machine learning, and more specifically deep learning, to create a tool that allows users to predict whether or not a mole is potentially cancerous. This thesis aims to achieve three notable contributions:

1. To successfully design and train a user-oriented network that uses both patient information and an image as input, and outputs a classification which predicts whether the image contains a cancerous mole or not.

2. Validate the equity of the model across various skin tones to investigate the importance of skin tone in the network's prediction.

3. Evaluate if targeted data augmentations can improve the performance of models across varying skin tones.

While this thesis strives to achieve all of the contributions listed above, all of the goals lead back to a central theme: how do we fairly and equitably use machine learning to create a tool that could have immediate impacts on the prognosis of melanomas? This question is tied up with technical, statistical, and ethical questions that will all be addressed in the coming chapters.

CHAPTER 2

LITERATURE REVIEW

2.1 Machine Learning

Machine learning has an extensive history, but this literature review will mostly focus on the fundamental and relevant contributions in regards to this project. While extensive contributions have built up to the machine learning techniques used in this paper, many fall beyond the scope of this literature review. For a more extensive review on deep learning, see Goodfellow *et al.* [11].

Three broad categories of deep learning can be constructed: supervised, semi-supervised, and unsupervised. Supervised learning utilizes fully labeled data in order to take an input $\boldsymbol{X}$ and map it to an output, y. Supervised learning can further be divided into *classification* and *regression*. Regression seeks to predict an numerical value as an output whereas classification tasks, such as that done in this thesis, involve predicting a categorical output.

The first and perhaps most famous instance of supervised classification was done by Frank Rosenblatt in 1958 [9] which served as the foundational example for machine learning. This network worked by mathematically modeling neurons using backpropagation to fit the parameters to the input data. The perceptron served as a foundational example for machine learning and modern fully connected networks (FCNs).

Later Fukushima and Miyake addressed the challenge of incorporating spatial information into deep neural networks through the Neocognitron– a precursor to the modern convolutional neural network (CNN) [**?**]. Many years and innumerable advances later, CNNs are the predominant algorithm for image classification.

Typically CNN models are developed on restricted resources and later scaled through altering the network depth, width, or resolution. This haphazard scaling led to inaccurate results and wasted resources, so in 2019 Tan and Le proposed EfficientNets [20] which delicately balanced parameters of model scaling to create networks that were significantly smaller, faster, and more accurate than other state of the art methods at the time.

While EffifcientNets changed how researchers view model scaling, others were taking a more divergent approach to machine learning research. In 2017 researchers at Google Brain developed the Transformer [21] which could model long-range interactions within sequences of data. Initially this was applied towards language processing tasks (*e.g. translation*), but in 2018 Yuan and Wang applied transformers to image segmentation [22] to incorporate context into the process. Soon others followed suit [23, 24, 25] and others developed attention blocks [26, 27] for integration into larger network architectures. While attention mechanisms were paradigm shift in image analysis and were breaking records for start of the art accuracy, they also posed a steep memory requirement which hindered the application of attention to multidimensional data such images [28].

In ICLR 2021 a large step was taken to improve memory efficiency while still incorporating the internal data structure through the relative distance between pixels. These networks, called *Lambda Networks* [28], serve to model long-range interactions while decreasing the number of training parameters, increasing the data throughput thus improving on training time, and doing all this while still maintaining state of the art accuracy.

2.2 Melanomas and Diagnosis

Malignant melanomas (also referred to simply as *melanomas*) arise from melanocytes in the skin. Melanocytes are the cells found most commonly in the epidermis and are responsible for producing melanin granules which pigment the skin and protect the underlying cells from ultraviolet radiation [29]. While there are many contributing risk factors for melanomas such as age, race, sex, and family history, the most important risk factor for melanoma is exposure to ultraviolet light due to it's importance in the development of melanomas [29] as well as the ability of the individual to modify their behavior to directly control their risk [14].

Melanomas typically originate in the epidermis, but they can also arise from non-pigmented melanocytes in the brain, mouth, or eye [30]. Additionally, melanomas quickly metastasize to critical organs such as the lymph nodes, lungs, and brain [30] which significantly contribute to their lethal nature. One of the most important factors to improving the prognosis of melanoma patients is catching lesions as soon as they appear thus allowing for early intervention [31]. As such, a long-standing recommendation is for individuals to regularly conduct a self-examination and see a physician about any suspicious moles. While conducting a self examination, a helpful mnemonic ***ABCDE***; **a**symmetry, **b**order (irregular), **c**olor (nonuniform), **d**iameter (large), and **e**volving.

There are multiple steps leading from the initial self-examination of a melanoma to the diagnosis. When initially seeing a physician about a suspicious mole, the physician will typically first visually examine the mole using a *dermatoscope*. Dermatoscopy (also known as *dermoscopy*) is non-invasive technique which allows for the visual assessment of the surface of skin and provides a fine enough resolution to show morphological features that are invisible to the naked-eye. This allows physicians to diagnose melanoma based on pigment networking, irregular dots and globules, irregular pigmentation, regressive structure, vascularization, or other microscopic features [14]. The other method of melanoma diagnosis relies on excision and histopathological analysis [32], but these approaches are out of this paper's scope.

2.3 Melanoma Classification

There have been a variety of attempts to create networks capable of classifying melanomas and other skin lesions, but most have been within the last decade due to the rapid advancement of computer vision techniques. Since dermatologists use both dermatoscopy and histopathology to diagnose melanoma, two disciplines have evolved around melanoma classification as well. While histopathology analysis is an ever-growing field, these techniques are out of the the scope of this thesis.

Earliest attempts at gross melanoma classification relied on non-standard color spaces and hand-crafted texture features inspired by dermatologist evaluation criteria. These techniques utilized the macroscopic characteristics of melanoma such as irregular streaking [33], regression structures [33], blotching [34], inconsistent color [35, 36, 37], pigment networking [37], and granularities [38]. These techniques would typically mine features related to shape (aspect ratio, diameter, circularity, symmetry, etc.), color features (mean and standard deviation, etc.), and textural features (gray level co-occurrence matrices) [39].

These features would be then passed to algorithms such as K-nearest-neighbors [**?**], Support Vector Machines [39], or basic artificial neural networks [40] to classify the images [41].

Later techniques became more sophisticated with the popularization of deep learning in computer vision. The most popular approaches utilized deep convolutional neural networks [42]. These models Then move in to the more specific discussions relied on large networks or even multiple large networks (referred to as an *ensemble*) combined into one system [43]. While these networks produce accurate results, they also become increasingly difficult to validate due to their complexity and black-box nature. One of the most recent state-of-the-art approaches to melanoma classification used an ensemble of EfficientNets creating a large model capable of accurately predicting cases of melanoma [44].

2.4 Melanoma Data

There are three main data sets for the gross diagnosis of skin lesions. While dermatoscopy has been accessible for imaging due to it's simple and non-invasive nature, there have not been many large and accessible data sets until recently due to technological limitations. One of the first notable dermatoscopy data sets, the Interactive Atlas of Dermoscopy, was published in 2004 [45]. However, due to low-bandwidth internet and insufficient storage technology, the data was only able to be distributed a set of CD-ROMs that could be purchased for $200 (with an accompanying booklet). The accompanying information within the booklet and CD-ROMs described how to diagnose skin lesions while avoiding false-positives, lookalikes, and other pitfalls. While this information was through yet brief, it meant that the Interactive Atlas of Dermoscopy was intended for training physicians– not algorithms.

One of the later and more widely used dermatoscopic skin lesion data sets was PH^2 [46] which was published in 2015 and made freely available for educational and research purposes. This data set contained detailed information on the formal diagnosis, assessment of several dermatoscopic features, and manual segmentation. Due to the time-consuming nature of creating these annotations, PH^2 only contained 200 images which is considered sparse compared to traditional machine learning data sets. However, PH^2 still remains a valuable database for model validation as well as medical training [46].

One year after the publication of PH^2, the International Skin Imaging Collaboration (*ISIC*) released their first data set (ISIC 2016) intended for melanoma detection at the International Symposium at Biomedical Imaging (*ISBI*) in 2016 [7]. This initial data set contained 900 dermatoscopy training images and an additional 379 testing images. In subsequent years, the ISIC expanded the data set to approximately 2,000 in 2017, 12,500 in 2018 , 25,000 in 2019, and eventually 44,000 images in 2020 [**?**] after collaborating with the Society for Imaging Informatics in Medicine (*SIIM*). As of the writing of this paper, [17] is the largest and most robust data sets to date and includes multiple moles for each patient which is more representative of true dermatological evaluation. Additionally, has the advantage of containing images with a variety of perspectives and lighting helping develop robust algorithms [17].

2.5 Algorithmic Bias

As artificial intelligence becomes more entwined in healthcare it will potentially improve patient outcomes– but maybe not for all. There are many inequalities presented in healthcare in America in the current era [47]. While some view artificial intelligence as the solution to these inequalities [48], others view these complex algorithms as ways of perpetuating the historical bias [**?**].

Black-box models present the challenge of being not easily interpretable. Due to this, medical applications of neural networks need to be thoroughly vetted before deployment. There are a number of instances in which machine learning learned to predict medical outcomes through patient data recorded in electronic health records, but rather than providing insightful answers they instead inherited the bias of those who wrote the records.

One hypothetical example presented by Parikh *et. al* [49] relied on a real statistical analysis of the National Library of Medicine's MEDLINE database from 1970-2005. This analysis showed that records of older women presenting with acute coronary syndrome (“ACS") tended to present with atypical symptoms and may by frequently misdiagnosed [**?**]. If a machine learning model were fit to this data, it may recommend not testing for ACS in older women due to their frequent lack of diagnoses [49]. While this is a very specific example, electronic health records are heavily prone to biased data due to various processes within the healthcare system itself [19].

Gianfrancesco *et. al* [50] categorized medical data bias into three classes: *sample size*, *missing data*,and *misclassification or measurement error*. Thus it is important not to only look at what is in the data, but also what is missing from the data. *Missing data* results from vulnerable populations having disparate health outcomes due to poor medical literacy, poor access to health care, or other difficulties. This may be reflected in data for vulnerable populations which only represent the worst cases and may skew the expected risk and result in erroneous inferences for certain groups. A small *sample size* results in some populations of patients not being represented in sufficient numbers in medical data. Inferences on members of these groups cannot be made with statistical confidence due to the small sample size potentially not being representative of the whole. Finally, *misclassification* or *measurement error* stems from vulnerable patients being more likely to only have access to teaching clinics or poorly staffed hospitals, for example. This results in less accurate diagnoses and algorithms trained on this data will result in incorrect predictions as a result of incorrect diagnoses.

2.6 Smartphone App and Need for Validation

Since melanomas are typically first discovered by the actual patient and the early detection of melanoma is critical to improving prognosis, making melanoma screening as accessible as possible has been a long-standing priority [**?**]. This can be seen in the development of early smartphone applications ("*apps*") aimed at helping users self-screen.

Early apps were simple and focused on educating users on risk factors and self-examination strategies (such as the *ABCDE* mnemonic). Many apps also allowed users to save and store images of skin lesions for review either by themselves of a dermatologist and reminded users to monitor their skin lesions. A few even offered expert review of images. However, very few of these apps allowed the users to actually asses the risk of individual skin lesions, and when they did it was simply though a series of questions and the diagnostic accuracy had not been validated [51].

Since the integration of the smartphone into personal life, the technology in phones has advanced exponentially. Today smartphones are equipped with megapixel cameras, gigahertz processors, and gigabytes of ram. As phones have advances, so have analysis techniques. Notable approaches have used convolutional neural networks to provide fine-grained classification of skin lesions [**?**], while others remain apps focused on education and monitoring [52]. Most recently, a handful of popular skin lesion detection apps (TeleSkin, skinScan, and SkinVision) were evaluated and shown to have poor diagnostic accuracy when identifying premalignant or malignant lesions [53].

Today, an estimated 6.3 billion camera-equipped smartphones are in the pockets of individuals around the world [54] proving that a smartphone app could serve as widely-available low-cost screening. However, the accuracy of these apps must be thoroughly vetted for bias against skin tone to ensure the equity of algorithms among users.

CHAPTER 3

MATERIALS AND METHODS

3.1 Data Acquisition

Data was downloaded from the online SIIM-ISIC 2020 competition page posted on online competition platform *Kaggle.com* using Kaggle's API.

3.1.1 Hardware

Training was done on a high-performance computer at the University of Pennsylvania running a modified distribution of RedHat Enterprise with a Nvidia V100 and 16GB of VRAM allocated. An additional 32GB of RAM was reserved. Additional testing was done on a workstation equipped with an Intel i7-6700k, 32GB of DDR4 RAM, and an Nvidia GTX 1070 with 8GB RAM running Debian.

3.1.2 Preprocessing

The initial training data consisted of 37,649 RGB image entries encoded in standard as standard jpegs. Each image had associated data pertaining to a pseudoanonymized patient ID, approximate age, biological sex, and approximate anatomical location of the lesion. Each image was initially loaded and center cropped to provide maximum width in one dimension with a square aspect ratio. After cropping, the selection was scaled to a 256x256 RGB image to standardize data format with minimal loss of context.

The .csv files for training and testing data were then loaded and parsed. This involved mapping $male$ and $female$ to 1 and 0 respectively. Additionally the approximate age of patents was scaled to a range of [0, 1] by normalizing via the highest age in both dataframes (90 years old). The approximate anatomical site was converted from a categorical variable (with 10 possible values: ‘anterior torso’, ‘head/neck’, ‘lateral torso’, ‘lower extremity’, ‘oral/genital’, ‘palms/soles’, ‘posterior torso’, ‘torso’, ‘upper extremity’, and ‘nan’) to a dummy/indicator variable.

3.1.3 Network Design

The final iterations of the network consisted of two branches: an image processing branch and a metadata branch.

The image processing branch consisted of an EfficientNet modified with the last convolutional layer replaced with a lambda convolution to incorporate global context. Each block of the EfficientNet consisted of a mobile inverted residual bottleneck which contained an expansion, depthwise convolution, pointwise convolution, and a batch normalization. The final convolutional bottleneck was replaced with a lambda convolution to incorporate long-range interactions [28].

The metadata processing branch consisted of a standard fully connected network with three hidden layers. Additionally a batch normalization, rectified linear unit function (*ReLU*), and dropout function ($p = 0.2$) was placed between each fully connected layer.

Near the end of both branches, the extracted latent feature vectors were concatenated and propagated through two more layers of a fully connected network before a final classification layer.

3.1.4 Training

3.1.4.1 Training Protocol

To begin training the data was split into an 80/20 training testing split. While a designated testing set is available from the competition organizers on Kaggle.com, the labels for this data remains private due to it’s usage in calculating competition leader boards. From there the training data was further divided into 6 folds for training and validation. Stratified K-Folds was used in order to preserve the percentage of samples from each class. This was done to ensure that each batch contained both cases of melanomas and benign lesions for the calculation of evaluation metrics which otherwise would be difficult due to the large class imbalance.

Each fold was trained for a maximum of 50 epochs, but early stopping was implemented so that if the model's validation ROC-AUC did not improve for 7 sequential epochs then training terminated for the fold.

During training the adaptive moment estimation (*"Adam"*) optimizer was used with an implementation of a binary cross entropy with sigmoidal activation (BCE-logits) for the loss function. Additionally a scheduler was used so that if there was no improvement to the model's validation ROC-AUC after 3 epochs, it would decrease the learning rate by a factor of 0.2 to avoid plateaus.

3.1.4.2 Augmentations

After each image was cropped and resized to 256x256, images were ran through a pipeline for randomized data augmentations. There were two types of augmentations: standard augmentations and color augmentations. Standard augmentations consisted of the following:

- Small ablations (16x16 black boxes)
- Simulated hair (black lines simulating random hairs)
- Simulated microscopy ring (black ring surrounding the image)
- Random horizontal flipping
- Random vertical flipping
- Random Gaussian noise

These augmentations changed the orientation or basic properties of the image and intended to modify in ways which would not impact the process an actual dermatologist would use to asses the lesion (i.e. the *ABCDE* mnemonic).

In addition to these standard augmentations, a second augmentation pipeline was used which included the following additional augmentations:

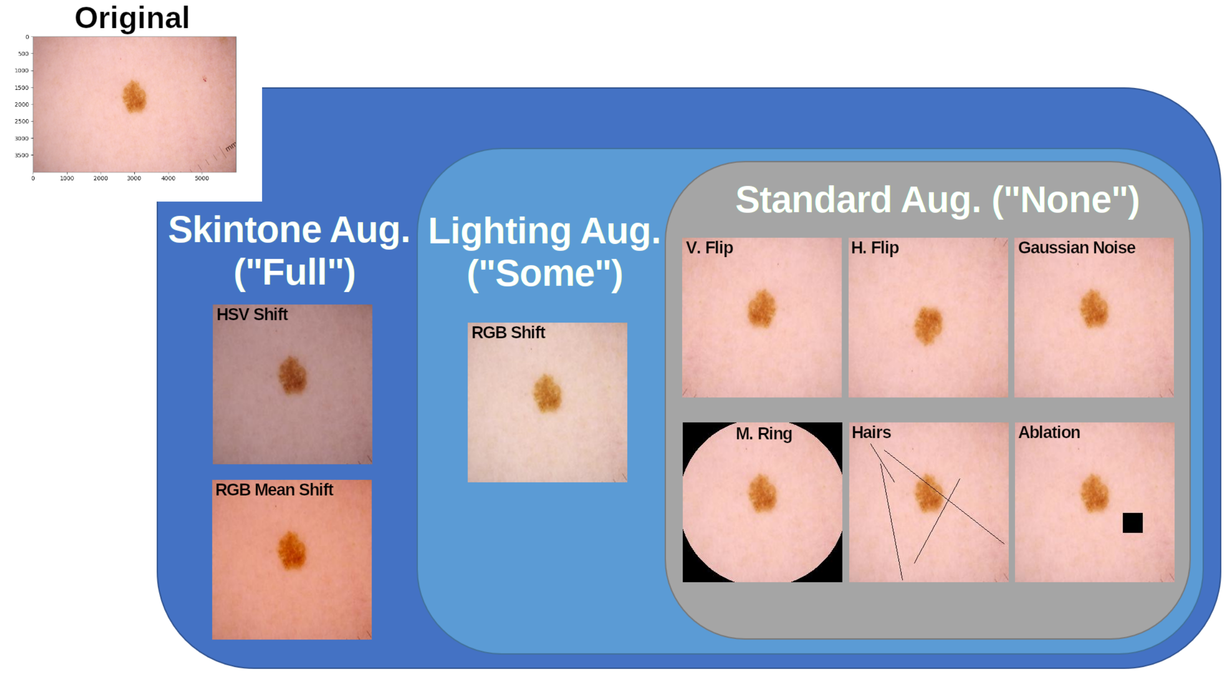


Figure 3.1: Augmentations by type.

- Light shift: simulate a change in the lighting (temperature and brightness) of the lesion in the RGB colorspace.
- Mean Shift: take the weighted of the existing RGB image with a randomly generated skin tone.
- HSV Shift: a limited perturbation of the the hue, saturation, and value of the image to provide a slightly different color profile.

Each augmentation had a 10% of being called with each accession of an image during training. Additionally, two data formatting steps were done after the augmentations to reshape the data into a suitable tensor for each batch and the data was normalized to mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$. The effects of these augmentations on an example image are seen in figure 3.1. Parameters for all color augmentations were randomly determined at call-time, meaning that the lighting augmentation, HSV Shift, and RGB Mean shift augmentations could all produce numerous appearances when called repeatedly on the same image.

3.2 Network Evaluation

3.2.1 Performance

Several metrics were used to evaluate the performance of the networks. Most technically the gross performance was evaluated using the area under the receiver operating characteristic curve (*"AUC-ROC"*). Intuitively this metric shows the variation of sensitivity and specificity at every threshold within the data. This translates into a rough understanding at how the model performs at distinguishing each class.

Additionally the positive predictive value (*"PPV"*) and negative predictive value (*"NPV"*) were used as more intuitive measures for the performance. Intuitively, positive predictive value is the probability of a positive prediction being correct. Similarly, negative predictive value is defined as the probability a negative prediction is true. Each is defined as:

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Where TP, FP, TN, and FN are the true positives, false positives, true negatives, and false negatives respectively.

3.2.2 Skin Tone Analysis

To analyze the impact of skin tone on model performance, the training and testing were subjected to automatic clustering. Inference was run on each cluster within each cluster and results across clusters and data sets were compared.

To automatically cluster, two types of clustering were employed. In (*feature-extraction clustering*) the images were run through a pretrained VGG-16 network and the latent representation (a 1000-element vector) were clustered with k-means ($k = 7$) and visualized with t-distributed stochastic neighbor embedding (*"t-SNE"*).

For the second types of clustering (*mean-rgb clustering*), RGB images were converted to the *CIE-LAB* color space. The lightness channel (often luminance) was thresholded using Otsu's method, and the resulting mask was used to remove RGB pixels darker than the determined threshold. This was done to mask out hair, microscopy rings, and the actual mole. The mean of the red, green, and blue channels were then taken to find the mean color of the remaining image (the skin) given a 3-element vector for each image. Similarly to the feature-extraction clustering, the image vectors were then clustered by k-means ($k = 7$) and visualized with t-SNE.

These clusters correlated to basic visual properties– most notably skin tone and hair. Next the entire data set (training and testing) was partitioned into the given clusters, and inference was run on each cluster. Metrics were then evaluated to determine how the clustering impacted AUC-ROC, PPV, NPV, sensitivity, and specificity.

3.2.3 Cluster Interpretation

Additionally, the number of images in each cluster were plotted against the AUCs of models evaluated on that cluster, and Spearman's correlation coefficient was found for the two distributions. The 95% confidence interval was also plotted as a translucent region surrounding the regression line for each model.

CHAPTER 4

RESULTS

The aim of this research was three fold: create a well-performing network in terms of raw performance, identify sub-cohorts of data withing the training set, and evaluate model performances across the sub-cohorts to explore the impact of targeted data augmentation on model performance.

4.1 General Network Performance Evaluation

The networks were trained with 6-fold cross validation, and the best performing networks on the validation data were selected and ran on the testing data that was set aside in the initial data partitioning. The area under the receiver operating characteristic curve metric (otherwise known as AUC-ROC or just "AUC") shows how well a model is able to distinguish between positive and negative classes within the data. The ROC curve is created by taking all predictions (within the continuous interval $[0, 1]$) and matching them up with their ground truth (0 and 1 as integers). It then takes all possible thresholds within the continuous interval $[0, 1]$ and uses it to classify the model's predictions into 0 or 1, then compares the thresholded predictions against the ground truth. Figure 4.1 shows the receiver operating characteristic in full detail. The high curve above the diagonal represents that models are able to very clearly distinguish between positive cases (melanomas) and negative cases (benign lesions). The ROC curves were summarized by calculating the are under the curve, giving AUCs close to 1 for nearly perfect model performance, AUCs close to 0.5 for random chance, and models close to 0 for worse than random model performance.

Models trained with each data augmentation condition showed a reasonably high curve representing good model performance. Table 4.1 shows the best performing models across each of the three training conditions. Training AUCs ranged from 0.935 to 0.987, and testing AUCs ranged from 0.961 to 0.983. The light color augmentation showed the best training AUC (followed by the no color augmentation and then the skin tone augmentation), and the no color augmentation and light color augmentation tied in testing, with the skin tone augmentation being the worst of the

Condition	Best Val. AUC	Testing AUC
No Color Aug.	0.976	0.983
Light Color Aug.	0.987	0.983
Skin Tone Aug.	0.935	0.961

Table 4.1: The ROC-AUC curves of the model selected for each training condition.

three.

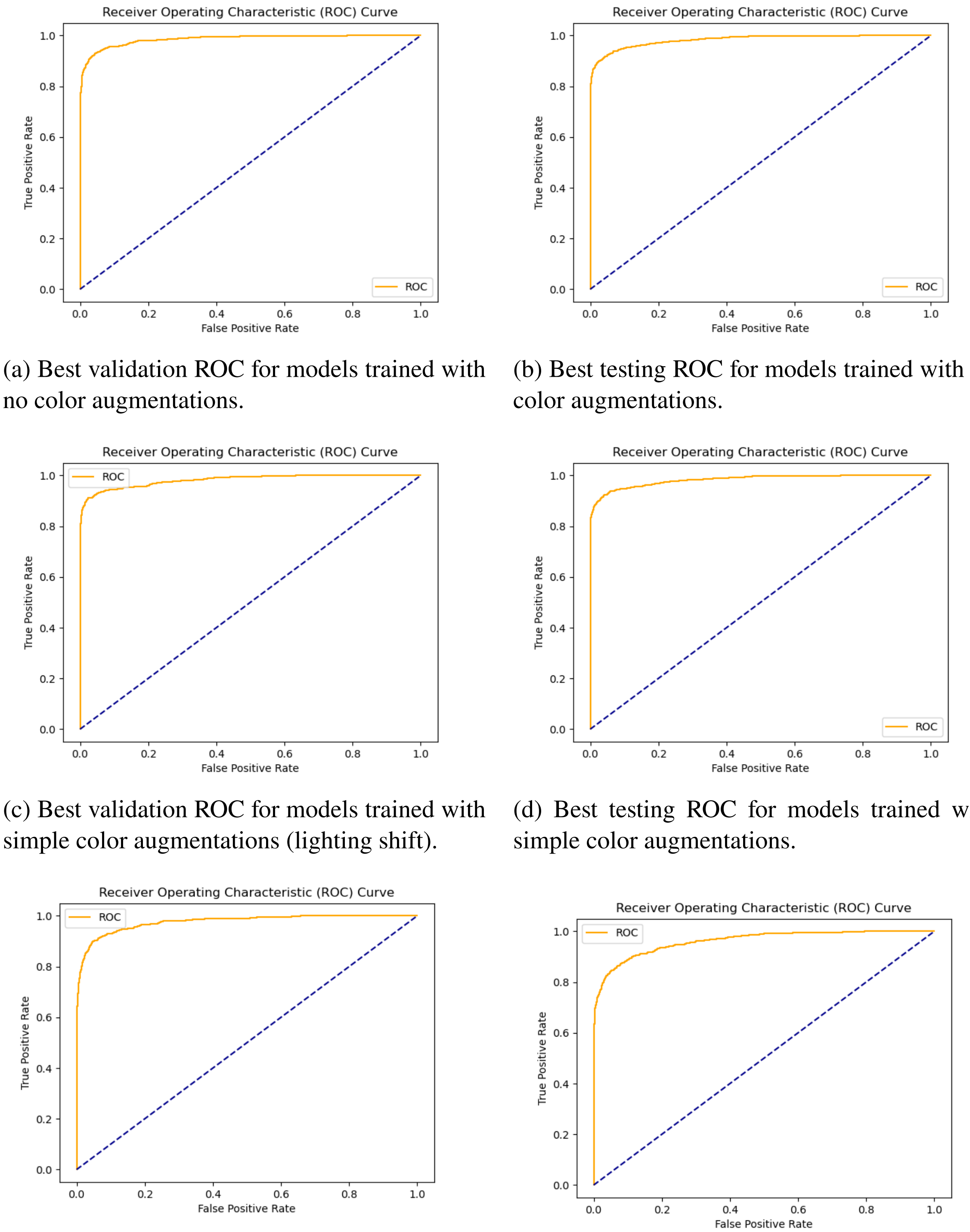
4.2 Image Clustering Results

Both methods of clustering showed groupings roughly based on skin tone and other general image characteristics. Figure 4.2 shows the t-SNE visualization of the clustering for both the VGG-16 feature extraction method as well as the RGB-mean clustering. Additionally, 9 images from each cluster were randomly selected and visualized. The number of images in each cluster as well as number of positive cases in each cluster were also found.

Feature-extraction clustering took images with the dimensions of 256x256x3 and distilled them to 1000 element latent vectors representing numerous features utilized in typical image classification.

The t-SNE in Figure 4.2 shows broad clusters with ill-defined boundaries, and visual inspection of samples from these clusters revealed both wide intra-cluster variability as well as visually similar images being grouped across multiple clusters.

Contrarily the RGB-mean clustering compressed the 256x256x3 images into 3-element vectors representing the upper end of the RGB spectrum in the image, thus roughly approximating skin color. The RGB-mean t-SNE in Figure 4.2 shows more distinct clusters with clearly defined borders. Visual inspection of images within the clusters further shows a clear tendency to cluster by color within images. While this does not always represent skin tone (as seen by the hairy image in figure 2), it presents a more easily interpretable clustering of images. Naturally dark skintones tended to be grouped in cluster 1, whereas lighter skin tones were more evenly distributed across clusters 0, 3, and 4 with minor variations between them.



(a) Best validation ROC for models trained with no color augmentations.

(b) Best testing ROC for models trained with no color augmentations.

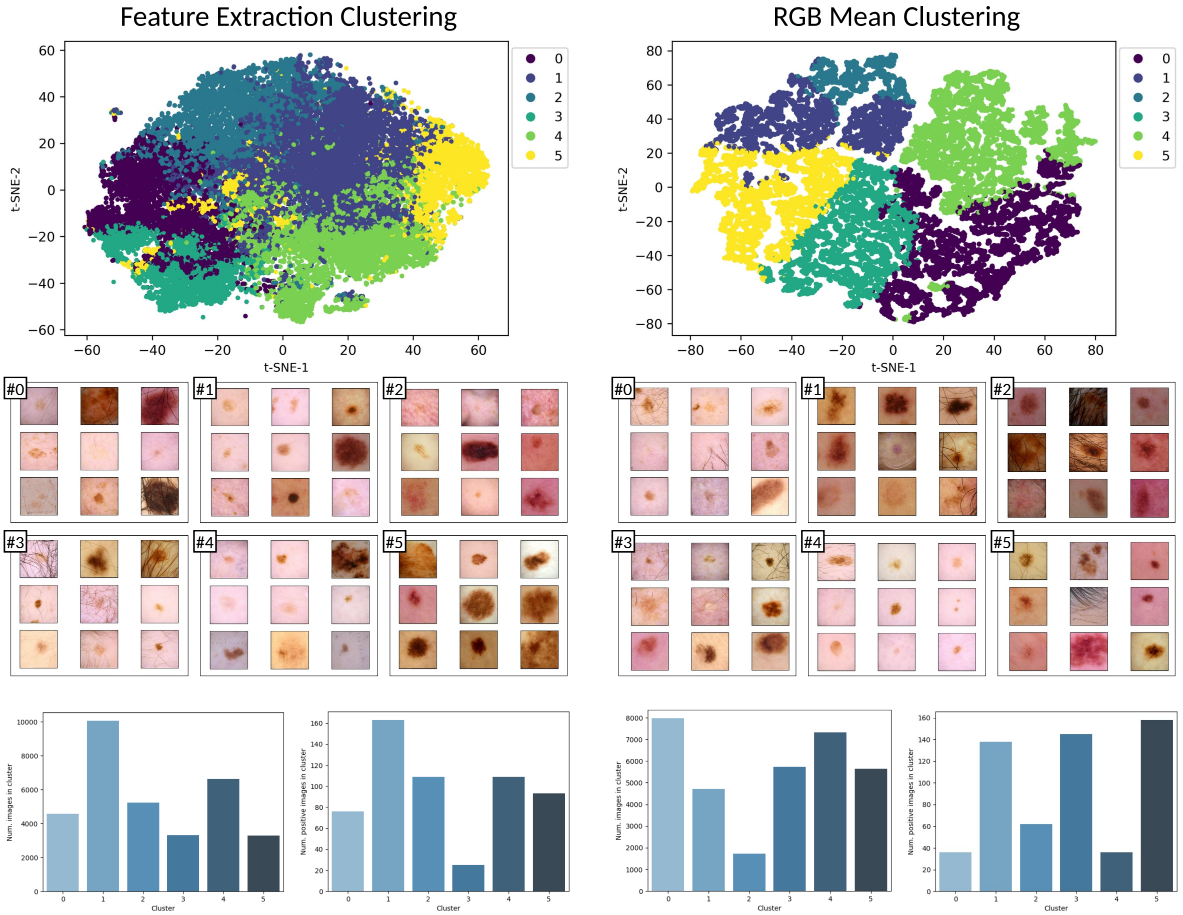
(c) Best validation ROC for models trained with simple color augmentations (lighting shift).

(d) Best testing ROC for models trained with simple color augmentations.

(e) Best validation ROC for models trained with strong color augmentations (lighting shift, HSV shift, mean RGB shift).

(f) Best testing ROC for models trained with strong color augmentations.

Figure 4.1: Validation ROC curves for the best models within each training condition.



(a) Clustering results using VGG-16 feature extraction followed by k-means ($k = 6$).

(b) Clustering results using the mean of RGB content via k-means ($k = 6$).

Figure 4.2: Results of each clustering method summarized for both VGG-16 feature extraction *(left)* and CIELAB clipping followed by taking the mean of the R, G, and B color channels *(right)*. *Top*: t-SNE visualization of the 6-class clustering. *Middle*: nine images from each cluster were randomly selected to show a representative appearance of the cluster. *Bottom*: the number of images within each cluster (left), and the number of positive cases within each cluster (right).

4.3 Evaluation on Individual Clusters

All models across all folds for each condition (no color augmentation, light color augmentation, and skin tone augmentation) were used to evaluate each cluster within the data set. The ROC-AUC, PPV, NPV, Sensitivity, and Specificity were measured. Figure 4.3 shows the results of each metric.

The two clusterings revealed varying numbers of training samples between clusters. The number of images within feature-extraction clusters varied from 3,500 to 10,000, with the mean number of images per cluster equal to 5521 images and a standard deviation of 2332 images. The RGB-mean clusters had the same mean (due to the fixed size of the combined training and testing set), and a slightly smaller standard deviation of 2016 images.

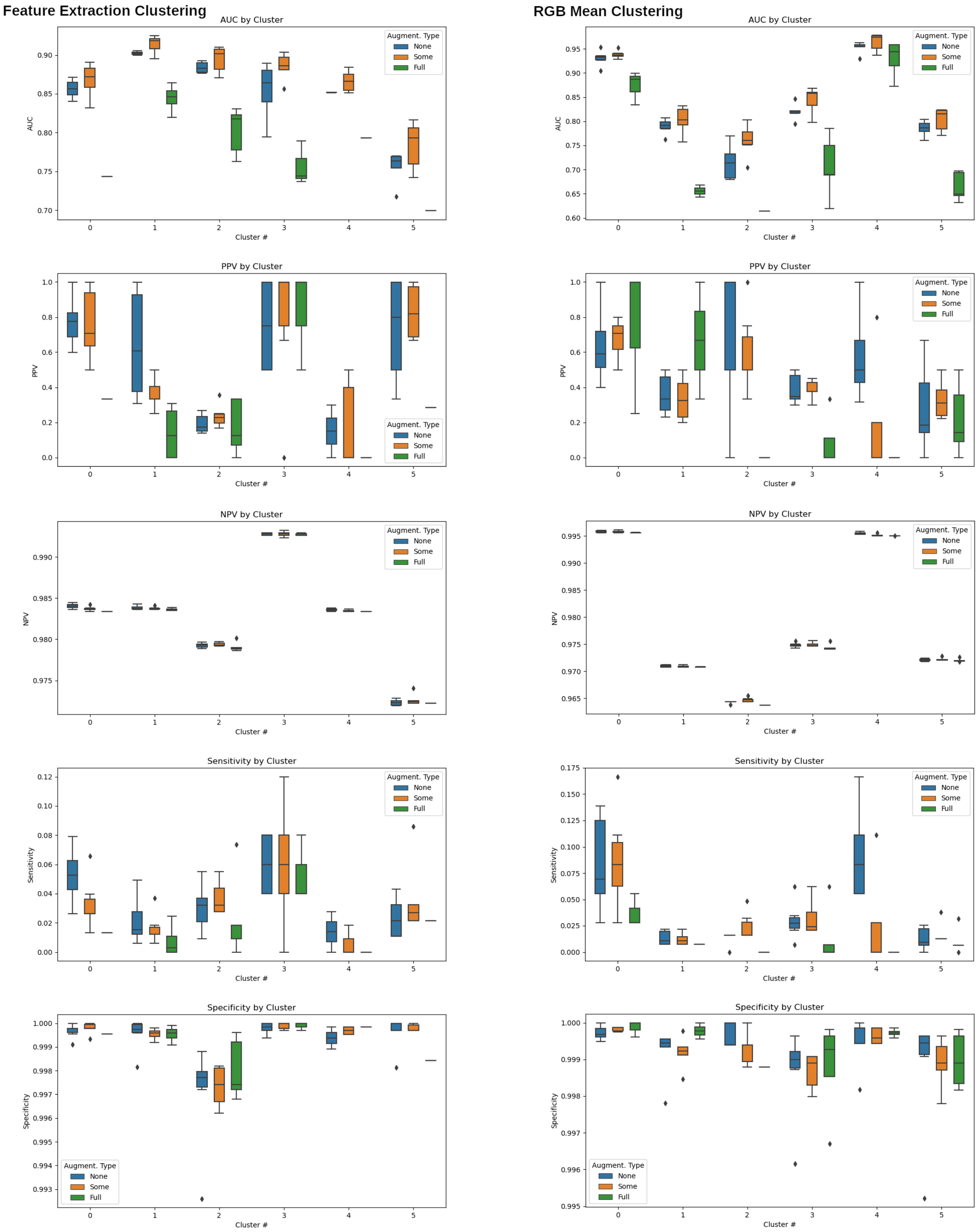
Correlations for both the feature-extraction clustering and the RGB-mean clustering showed that the number of images in a cluster strongly correlate with the performance on that cluster. Clusters with relatively few training samples perform poorly, and models tend to perform the best on clusters with the most training samples.

4.4 Cluster Correlations

A strong positive correlation was found between number of training samples in the cluster and model performance on that cluster for all models regardless of augmentation type 4.4. Both clustering methods showed similar results: the no color augmentation group (‘None’) and light color augmentation group (‘Some’) showed comparable performance, while the skin tone augmentation (‘Full’) showed the worst performance across clusters.

4.5 Lightweight Networks are Feasible for Melanoma Screening

The strong performance we see in both Table 4.1 and Figure 4.1 shows high ROC-AUC values for well-trained models. Each model contained 7,084,717 parameters for a final compressed model size of 27.5 megabytes. Interpretation of the AUC-ROC curves shows a clear ability of the models to distinguish between negative and positive samples in both the training and testing sets.



(a) Model performance evaluated across feature-extraction clusters

(b) Model performance evaluated across rgb-mean clusters

Figure 4.3: Model performance evaluated across clusters with feature-extraction clustering (*left*) and rgb-mean clustering (*right*). Metrics from top to bottom: ROC-AUC, PPV, NPV, Sensitivity, and Specificity.

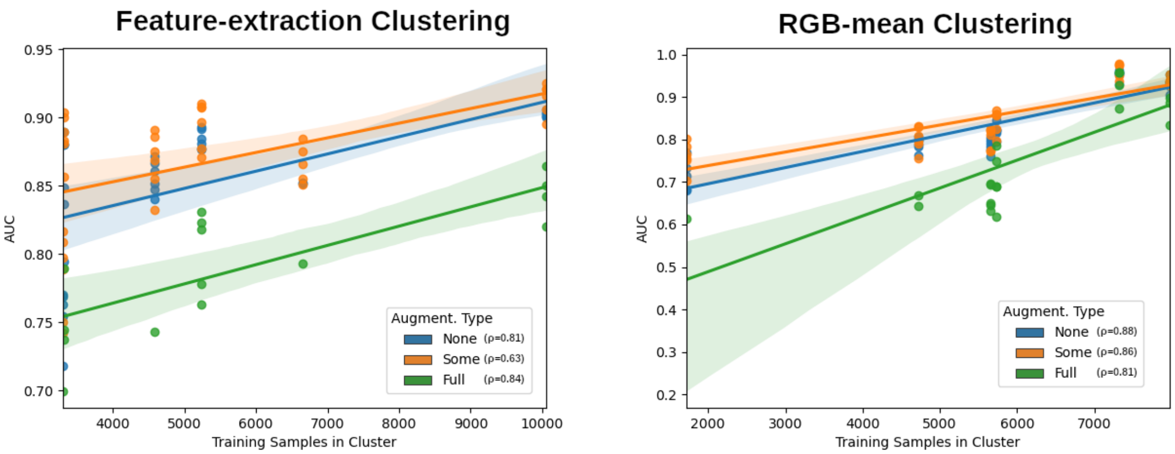


Figure 4.4: The correlation between number of training samples in a data set and the performance of models on that cluster for each augmentation type. *Left*: correlation using feature-extraction clustering. *Right*: correlation using RGB-mean clustering.

CHAPTER 5

DISCUSSION

The models performed quite well when looking at their ROC-AUCs, yet the implications of the classification needs to be weighted more carefully than through raw performance metrics. Automatic clustering and subsequent visualization showed that the data had distinctive clusters based on relatively interpretable visual properties. Among these clusters there was an uneven distribution of both training samples as well as performance on the specified clusters revealing a correlation between the size of a cluster's contribution to the training data and the performance of models trained on that the data when evaluating that cluster.

5.1 Clustering Results

The issue with feature-extraction is that it takes an image that is easily interpretable to humans and converts it to a *latent representation* which is full of vivid detail to a machine, but not decipherable to a human. Techniques like activation maps have historically been utilized to provide insight into the relative importance of image regions, but this is outside the scope of this paper. Instead, a sample of nine images was shown from each cluster to represent the visual properties of that cluster.

The feature-extraction clustering resulted in more poorly defined clusters than the RGB-mean clustering. This is seen both in the t-SNE and visual inspection of figure 4.2. This is due to the two-stage feature extraction clustering producing 1000 element vectors which were then classified, leading to uninterpretable features used for other image classification. While there seemed to be some trend, more experimentation is needed to confirm the visual properties through techniques.

Alternatively, the RGB-Mean clustering showed promise for clustering based on skin color. The handcrafted feature of filtered pixels is not perfect and shows some in-group heterogeneity, but visual trends are easily identifiable and the RGB centroid of each cluster directly corresponds to the mean color within that cluster. This shows potential for RGB-mean clustering to be used in the automatic identification of subcohorts in a more targeted manner rather than relying on feature extraction. Additionally, the condensed RGB vector representation image shows promise for other methods of data clustering such as Gaussian Mixture Models which would more accurately represent skin tone as a continuum rather than discrete classes.

5.2 Augmentations Cannot Fill in For Inclusion

Performing light augmentations (Figure 3.1) to the temperature and intensity of the illumination resulted in what appeared to be slightly better AUCs (Figure 4.1 and Table 4.1), though more experimentation is needed to statistically confirm these findings. Standard augmentations (rotations, cropping, sheering, and ablation– noted by "none") as well as other non-color shifting (simulated hairs, microscopy rings) augmentations produced strong results.

Most notably, targeted data augmentations for skin-tone failed to produce better results even in clusters with darker skin appearances (*feature-extraction clustering* cluster #5, *rgb-mean clustering* clusters #1 and #2). These results show in fact the *opposite* of the expected result, yet they are consistent with how dermatologists visually evaluate lesions. Recall "*C*" of the *ABCDE* mnemonic for melanoma self-screening– "color." Additionally, numerous mentions of vascularity and a "blue-whit veil" [14, 15, 55]. Even whole studies have assessed the importance of the blue color channel in the assessment of melanomas [56]. By dynamically altering the skin-tone artificially, the image may lose important visual cues such as the blue veil, or though increasing saturation it may over-emphasize blood vessels and introduce unfounded indicators of melanoma.

5.3 Implications of Classification

One consideration for the creation of a classifier for consumer use is the thresholding of results. As seen in the ROC curves (Figure 4.1), varying thresholds for flagging a prediction as positive result in different rates of false positives and false negatives. In a perfect world the two classes would be perfectly separable, but medical experts must consider the implications of what a false positive and false negative could mean for a patient.

On one hand, the classifier leaning towards giving false positives could potentially be beneficial. It could encourage individuals with suspicious moles to go to their healthcare provider and get the mole professionally checked and cleared. While there, the physician could address other health concerns or other moles in the process, potentially leading to the detection of actual precancerous or cancerous lesions. Yet on the other side of this, false positives would lead to excessive worry, potentially unnecessary medical bills, and may even increase wait times for medical care thus preventing those who actually need to see a dermatologist from getting care. An alternative is having the model lean towards false negatives. While this would prevent some from unduly worry, it would clearly be detrimental to others who may leave precancerous moles alone until it is too late. Whatever the decision is, it is not one to be made lightly.

CHAPTER 6

CONCLUSION

6.1 Augmentations Fail to Represent the Underrepresented

One of the greatest lessons learned comes from the following revelations: 1) The number of training samples in clusters had wide variation showing an uneven representation, 2) The relative number of training samples in the cluster strongly correlate with the performance of models trained on data containing that cluster, 3) Skintone-specific augmentations failed to improve the generalizability or performance of the models. These three facts lead to the most important conclusions of this paper: there is no replacement for inclusion.

6.2 Evaluation is not Black and White

While competitions and papers often chase state of the art performance through maximizing their ROC-AUC or other evaluative metrics, much more consideration needs to be placed into model evaluation before these algorithms are ready for consumers. For one, the stratification of data shows that there is a distinctive need to ensuring that models work equitably within subcohorts of the data set. If a patient population is underrepresented during training, then the predictions on their population during deployment will similarly lack as seen in Figure 4.4.

6.3 The Need for Racially Inclusive Data in Imaging Data

While several workarounds were done in this paper to attempt to identify patterns among the cohort both through feature-extraction and hand-crafted RGB features, it is nearly impossible to validate this data without a ground truth. This underscores the need for data sets to include information about a patient's complexion (such as melanin index) to be included in medical imaging in order to insure equitable evaluation across subcohorts. Simply recording these measurements would remove the guesswork and provide for more substantial vetting of models.

6.4 Future Directions

In the future, more extensive analysis should be done on prior winning solutions to the SIIM-ISIC challenge to evaluate how ensemble models fair with automatic cluster analysis. Additionally, expert dermatologists should collaborate with image analysts to create acceptable augmentation which enhance the generalizability of melanoma data without negatively impacting equitable performance.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Paulien Hogeweg. The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*, 7(3), March 2011.

[2] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V.

Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, Michael J. Morgan, International Human Genome Sequencing Consortium, Center for Genome Research: Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, Institute of Molecular Biotechnology: Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, The Institute for Systems Biology: Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, Lita Annenberg Hazen Genome Center: Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, also includes individuals listed under other headings): *Genome Analysis Group (listed in alphabetical order, US National Institutes of Health: Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, Keio University School of Medicine: Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, US Department of Energy: Office of Science, and The Wellcome Trust:. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. Number: 6822 Publisher: Nature Publishing Group.

[3] Sarah K. Wooller, Graeme Benstead-Hume, Xiangrong Chen, Yusuf Ali, and Frances M. G. Pearl. Bioinformatics in translational drug discovery. *Bioscience Reports*, 37(4), August 2017.

[4] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. Artificial intelligence in radiology. *Nature Reviews. Cancer*, 18(8):500–510, August 2018.

[5] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, October 2016.

[6] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1:18, 2018.

[7] David Gutman, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin Lesion Analysis toward Melanoma Detection: A Challenge at

the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). May 2016.

[8] What is Machine Learning?, March 2021.

[9] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[10] Alekseĭ Grigorevich. Ivakhnenko, Valentin Grigorévich Lapa, and United States. Joint Publications Research Service. *Cybernetic predicting devices*. CCM Information Corp., New York, 1965. Section: d, 256 pages illustrations 17 cm.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[12] Kunihiko Fukushima and Sei Miyake. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In Shun-ichi Amari and Michael A. Arbib, editors, *Competition and Cooperation in Neural Nets*, Lecture Notes in Biomathematics, pages 267–285, Berlin, Heidelberg, 1982. Springer.

[13] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv:1712.04621 [cs]*, December 2017. arXiv: 1712.04621.

[14] Marco Rastrelli, Saveria Tropea, Carlo Riccardo Rossi, and Mauro Alaibac. Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification. *In Vivo (Athens, Greece)*, 28(6):1005–1011, December 2014.

[15] Gabriella Brancaccio, Teresa Russo, Aimilios Lallas, Elvira Moscarella, Marina Agozzino, and Giuseppe Argenziano. Melanoma: clinical and dermoscopic diagnosis. *Giornale Italiano Di Dermatologia E Venereologia: Organo Ufficiale, Societa Italiana Di Dermatologia E Sifilografia*, 152(3):213–223, June 2017.

[16] Darrell S. Rigel. Epidemiology of Melanoma. *Seminars in Cutaneous Medicine and Surgery*, 29(4):204–209, December 2010.

[17] 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):34, January 2021. Number: 1 Publisher: Nature Publishing Group.

[18] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018.

[19] Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, April 2018. Publisher: British Medical Journal Publishing Group Section: Research.

[20] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:1905.11946 [cs, stat]*, September 2020. arXiv: 1905.11946 version: 5.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. page 11.

[22] Yuhui Yuan and Jingdong Wang. OCNet: Object Context Network for Scene Parsing. *arXiv:1809.00916 [cs]*, September 2018. arXiv: 1809.00916 version: 1.

[23] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, Salt Lake City, UT, USA, June 2018. IEEE.

[24] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-Based Global Reasoning Networks. *arXiv:1811.12814 [cs]*, November 2018. arXiv: 1811.12814.

[25] Songyang Zhang, Shipeng Yan, and Xuming He. LatentGNN: Learning Efficient Non-local Relations for Visual Recognition. *arXiv:1905.11634 [cs]*, May 2019. arXiv: 1905.11634.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]*, October 2020. arXiv: 2010.11929.

[27] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big Transfer (BiT): General Visual Representation Learning. *arXiv:1912.11370 [cs]*, May 2020. arXiv: 1912.11370.

[28] Irwan Bello. LambdaNetworks: Modeling long-range Interactions without Attention. September 2020.

[29] A. Hunter Shain and Boris C. Bastian. From melanocytes to melanomas. *Nature Reviews. Cancer*, 16(6):345–358, June 2016.

[30] William E. Damsky, Lara E. Rosenbaum, and Marcus Bosenberg. Decoding Melanoma Metastasis. *Cancers*, 3(1):126–163, December 2010.

[31] R. J. Friedman, D. S. Rigel, and A. W. Kopf. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA: a cancer journal for clinicians*, 35(3):130–151, June 1985.

[32] Mirna Situm, Marija Buljan, Maja Kolić, and Majda Vučić. Melanoma–clinical, dermatoscopical, and histopathological morphological characteristics. *Acta dermatovenerologica Croatica: ADC*, 22(1):1–12, 2014.

[33] Giuseppe DI Leo, Gabriella Fabbrocini, Consolatina Liguori, Antonio Pietrosanto, and M Sclavenzi. ELM image processing for melanocytic skin lesion diagnosis based on 7-point checklist: a preliminary discussion. *Proc. 13th Int. Symp. measurements for research and industry applications, IMEKO, Budapest, Hungary*, pages 474–479, January 2004.

[34] William Stoecker, Ronald Stanley, Randy Moss, and Bijaya Shrestha. Detection of asymmetric blotches in dermoscopy images of malignant melanoma using relative color. *Skin research and technology : official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)*, 11:179–84, September 2005.

[35] Giuseppe DI Leo, Consolatina Liguori, Alfredo Paolillo, and P. Sommella. An improved procedure for the automatic detection of dermoscopic structures in digital ELM images of skin lesions. pages 190–194, August 2008.

[36] Maryam Sadeghi, Majid Razmara, Tim Lee, and Margaret Atkins. A novel method for detection of pigment network in dermoscopic images using graphs. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 35:137–43, March 2011.

[37] Catarina Barata, Jorge Marques, and Jorge Rozeira. Detecting the pigment network in dermoscopy images: A directional approach. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2011:5120–3, August 2011.

[38] William V. Stoecker, Mark Wronkiewiecz, Raeed Chowdhury, R. Joe Stanley, Jin Xu, Austin Bangert, Bijaya Shrestha, David A. Calcara, Harold S. Rabinovitz, Margaret Oliviero, Fatimah Ahmed, Lindall A. Perry, and Rhett Drugge. Detection of granularity in dermoscopy images of malignant melanoma using color and texture features. *Computerized Medical Imaging and Graphics*, 35(2):144–147, March 2011.

[39] M. Emre Celebi, Hassan Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y Aslandogan, William Stoecker, and Randy Moss. A Methodological approach to the classification of dermoscopy images. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 31:362–73, October 2007.

[40] Pietro Rubegni, Gabriele Cevenini, Marco Burroni, Roberto Perotti, Giordana Dell'Eva, Paolo Sbano, Clelia Miracco, Pietro Luzi, Piero Tosi, Paolo Barbini, and Lucio Andreassi. Automated diagnosis of pigmented skin lesions. *International journal of cancer. Journal international du cancer*, 101:576–80, October 2002.

[41] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques. Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features. *IEEE Systems Journal*, 8(3):965–979, September 2014. Conference Name: IEEE Systems Journal.

[42] Yuexiang Li and Linlin Shen. Skin Lesion Analysis Towards Melanoma Detection Using Deep Learning Network. March 2017.

[43] N. C. F. Codella, Q. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4/5):5:1–5:15, July 2017. Conference Name: IBM Journal of Research and Development.

[44] Qishen Ha, Bo Liu, and Fuxu Liu. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. *arXiv:2010.05351 [cs]*, October 2020. arXiv: 2010.05351.

[45] Peter A. Lio and Paul Nghiem. Interactive Atlas of Dermoscopy: Giuseppe Argenziano, MD, H. Peter Soyer, MD, Vincenzo De Giorgio, MD, Domenico Piccolo, MD, Paolo Carli, MD, Mario Delfino, MD, Angela Ferrari, MD, Rainer Hofmann-Wellenhof, MD, Daniela Massi, MD, Giampiero Mazzocchetti, MD, Massimiliano Scalvenzi, MD, and Ingrid H. Wolf, MD, Milan, Italy, 2000, Edra Medical Publishing and New Media. 208 pages. $290.00. ISBN 88-86457-30-8.CD-ROM requirements (minimum): Pentium 133 MHz, 32-Mb RAM, 24X CD-ROM drive, 800 × 600 resolution, and 16-bit color graphics capability. Test system: Pentium III 700 MHz processor running Microsoft Windows 98. Macintosh compatible only if running Windows emulation software. *Journal of the American Academy of Dermatology*, 50(5):807–808, May 2004. Publisher: Elsevier.

[46] Teresa Mendonça, Pedro Ferreira, André Marçal, Catarina Barata, Jorge Marques, Joana Rocha, and Jorge Rozeira. PH2: A Public Database for the Analysis of Dermoscopic Images. pages 419–439. September 2015.

[47] Mariana C. Arcaya, Alyssa L. Arcaya, and S. V. Subramanian. Inequalities in health: definitions, concepts, and theories. *Global Health Action*, 8:27106, 2015.

[48] Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA journal of ethics*, 21(2):E167–179, February 2019.

[49] Ravi B. Parikh, Stephanie Teeple, and Amol S. Navathe. Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, 322(24):2377–2378, December 2019.

[50] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544–1547, November 2018.

[51] A. P. Kassianos, J. D. Emery, P. Murchie, and F. M. Walter. Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. *The British Journal of Dermatology*, 172(6):1507–1518, June 2015.

[52] Elizabeth Chao, Chelsea K. Meenan, and Laura K. Ferris. Smartphone-Based Applications for Skin Monitoring and Melanoma Detection. *Dermatologic Clinics*, 35(4):551–557, October 2017.

[53] Karoline Freeman, Jacqueline Dinnes, Naomi Chuchu, Yemisi Takwoingi, Sue E. Bayliss, Rubeta N. Matin, Abhilash Jain, Fiona M. Walter, Hywel C. Williams, and Jonathan J. Deeks.

Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ (Clinical research ed.)*, 368:m127, February 2020.

[54] Stephen Carson, Anette Lundvall, Richard Möller, Susanna Bävertoft, Anna Jacobsson, Git Sellin, Michael Björn, Vishnu Singh, Stephen Carson, Reiner Ludwig, Lasse Wieweg, Jonas Edstam, Per Lindberg, and Kati Öhman. Ericsson Mobility Report June 2016. page 32.

[55] T. Schindewolf, W. Stolz, R. Albert, W. Abmayr, and H. Harms. Classification of melanocytic lesions with color and texture analysis using digital image processing. *Analytical and Quantitative Cytology and Histology*, 15(1):1–11, February 1993.

[56] D. Massi, V. De Giorgi, P. Carli, and M. Santucci. Diagnostic significance of the blue hue in dermoscopy of melanocytic lesions: a dermoscopic-pathologic study. *The American Journal of Dermatopathology*, 23(5):463–469, October 2001.