

Fake It Until You Make It

**Exploring if synthetic GAN-generated images can improve the
performance of a deep learning skin lesion classifier**

Nicolas Carmont Zaragoza
s1632536@sms.ed.ac.uk

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh
2020

Abstract

In the U.S alone more people are diagnosed with skin cancer each year than all other types of cancers combined [1]. For those that have melanoma, the average 5-year survival rate is 98.4% in early stages and drops to 22.5% in late stages [2]. Skin lesion classification using machine learning has become especially popular recently because of its ability to match the accuracy of professional dermatologists [3]. Increasing the accuracy of these classifiers is an ongoing challenge to save lives. However, many skin classification research papers assert that there is not enough images in skin lesion datasets to improve the accuracy further [4] [5]. Over the past few years, Generative Adversarial Neural Networks (GANs) have earned a reputation for their ability to generate highly realistic synthetic images from a dataset. This project explores the effectiveness of GANs as a form of data augmentation to generate realistic skin lesion images. Furthermore it explores to what extent these GAN-generated lesions can help improve the performance of a skin lesion classifier. The findings suggest that GAN augmentation does not provide a statistically significant increase in classifier performance. However, the generated synthetic images were realistic enough to lead professional dermatologists to believe they were real 41% of the time during a Visual Turing Test. Areas of further study for skin lesion GANs and potential future applications of the synthetic skin lesion images in educational material were explored.

Acknowledgements

To Robert Fisher for his ongoing support and guidance throughout this project, for our insightful weekly conversations and help in the pronunciation of “Wasserstein”.
To the anonymous professional dermatologists for their participation in the study and the Visual Turing Test.

To Mom, Dad, Blanca and Oli who were always there for me every step of the way a phone call away through out my 4 years at university. Constantly providing love, laughter and reminding me of how pale I look.

In loving memory of Nanny, who always taught me to “Smile, be happy and don’t forget the doughnuts”.

In loving memory of Dottie, an exemplary fighter.

To Poppy, Avia, Avia and my family in Australia and Spain for their unmatched love and constant source of inspiration and joy.

To the Redfort Mafia, Ashish, Justin, Sameer, Jay, Angus, Adityha and Gurz for making this the most enjoyable 4 years.

To Joanna for her constant positivity and enthusiasm even across time zones.

To my friends in Amsterdam and Barcelona for some great memories.

In loving memory of Xana, a beautiful young soul who left us too soon.

Contents

1	Introduction	5
1.1	The Problem	5
1.2	A Potential Solution	6
1.3	Malignant Lesions	7
1.4	Research Question	9
1.5	Roadmap	9
2	Background & Literature	11
2.1	The role of Artificial Neural Networks in skin lesion classification . .	11
2.2	Convolutional Neural Networks	12
2.3	State-of-the-art in skin lesion classification	13
2.4	DERMOFIT Dataset	14
2.5	Work on the DERMOFIT Dataset	15
2.6	What is Data Augmentation?	16
2.7	How GANs create synthetic images	17
2.8	Existing work on skin lesions GANs	18
3	GAN Implementation	20
3.1	Building the GAN	20
3.2	Selecting the dimensions of the generator input and output	21
3.3	Training the GAN	22
3.4	Challenges with mode collapse	23
3.5	The Wasserstein GAN	24
3.6	Many GANs or a single GAN?	25
3.7	When to finish training the GAN	26
4	Testing Methodology	28
4.1	Test Framework	28
4.2	Transfer Learning as Feature Extraction	30
4.3	Choice of Hyper-parameters	31
4.4	Baseline and Comparison	31
4.5	Reproducibility and Reliability	32
4.6	Performance Metrics	32
5	Results	34
5.1	Training Performance of the VGG-16	34

5.2	Performance of the 4 Augmentation Techniques	35
5.2.0.1	No Augmentation	36
5.2.0.2	Affine augmentation	38
5.2.0.3	WGAN Augmentation	40
5.2.0.4	Both Affine and WGAN Augmentation.	42
5.3	Statistical Analysis	43
6	Qualitative Evaluation: Visual Turing Test	46
6.1	Methodology of VTT	46
6.2	Results	47
6.3	Analysis	49
7	Conclusion & Evaluation	50
7.1	Analysis & Conclusion	50
7.2	Comparison to Literature	51
7.3	Method Strengths	52
7.4	Method Weaknesses	52
7.5	Potential Applications	53
7.6	Further Study	53
8	Appendix	59
8.1	Visual Turing Test (VTT) Survey	59

Chapter 1

Introduction

“Reality is merely an illusion, albeit a very persistent one.”

-Albert Einstein [6]

1.1 The Problem

Approximately two in three Australians will be diagnosed with skin cancer by the time they turn 70 [7]. It is the most common type of cancer worldwide, positioning it at 18th worldwide on the rank of global health threats [8]. The largest sufferers are pale-skinned populations in sun-exposed countries such as Australia, the United States and New Zealand [9]. Despite evidence showing individuals having less sun exposure, the total frequency of skin cancer incidence has been rising recently perhaps due to a variety of factors including more intense UV presence and rising rates of longevity [4] [10].

Fortunately, medical tools used for skin cancer detection have greatly improved over the past decades [11]. Vast improvements in skin lesion classification particularly have made it possible for dermatologists to augment their diagnosis capabilities using technology. Today, various medical devices and even apps exist to give a additional information to dermatologists on diagnosis [12].

However, studies have shown that the accuracy of these devices is limited [12][13]. This is concerning as a False Negative misdiagnosis by a skin lesion classifier can lead a doctor to believe that a patient with a malignant skin cancer is fine. Skin lesion classification research papers often blame a lack of data as the primary reason for not achieving higher performance [4]. Most datasets used for machine learning with neural networks often contain 50,000-100,000 images, however most skin lesion datasets only contain about 800-8000 images [14]. The problem with getting more data is that it often takes a lot of time (typically years), much legal approval from patients and extensive collaboration with various medical facilities [14]. Nevertheless, there may exist other forms of acquiring more data without needing to gather new skin lesion images.

1.2 A Potential Solution

In the past few years, Generative Adversarial Neural Networks (GANs) have earned a reputation for their ability to generate highly realistic synthetic images from a dataset [15]. To verify this, the reader is encouraged to classify which of the face images below are real or generated by AI.



Figure 1.1: Mixture of Real and GAN-generated Face images [16]

From the 6 images above, all have been generated by StyleGAN2 [16].

One of the most distinctive parts about a GAN is that it learns implicitly [17]. This means that the component in the GAN responsible for generating synthetic images (the generator) learns only through feedback it is given by another component in the GAN which acts as a supervisor (the discriminator). This means the generator never actually sees real images. Every step of training, the generator creates a sample image and is then told whether it believed this is accurate or not. Much like humans, the generator learns by failure. So it essentially learns to “paint pictures” by itself, correcting itself when it gets bad grades, rather than by looking at the solutions. Because of this, the generator can create synthetic images that do not necessarily look like any single image in the dataset, but instead incorporate a variety of features from each in a realistic way [18]. Chapter 2 explores further how GANs work.

From the standpoint of a skin lesion classifier, generating more data, specifically for rare malignant lesions, is highly desirable. This is as many skin lesions are rare and having more variations of a malignant lesion can help in its detection. Hence if GANs could generate realistic skin lesion images that add new information then they could potentially help increase the the performance of a skin lesion classifier and hence help save lives.

However, in order to help better understand what “realistic” means in the context of skin lesions, we must first look into the what the different types of skin cancer and pre-cancer look like.

1.3 Malignant Lesions

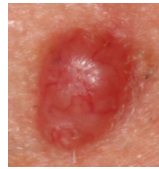

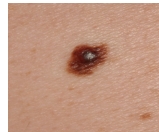
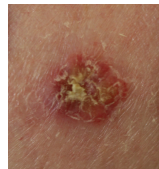

Legal note: The following material is meant as background on skin lesions and should not be taken as medical advice. The author is not a certified doctor.

When talking about malignant skin lesions in this project we specifically refer to the different types of skin cancer and pre-cancer.

Skin cancer is typically divided into two main categories: melanoma skin cancer and non-melanoma skin cancer. Within non-melanoma skin cancer there exists two types: Squamos cell carcinoma and Basal cell carcinoma. Overall, Basal cell carcinoma (non-melanoma) is the most common type of skin cancer, with Squamos cell carcinoma (non-melanoma) following soon after. Meanwhile, Melanoma is the least common skin cancer of the three, but it is also the most deadly with twice the death rate per incidence in the US [1].

There also exists different types of pre-cancer that one should also be wary of as they can develop into cancer. These are: Intra-Epidermal Carcinoma and Actinic Keratosis. If left alone 10% of Actinic Keratosis and 3-5% of Intra-Epidermal Carcinoma are likely to turn into cancerous Squamos cell carcinomas.

The following table shows the distinct appearances of these malignant lesions.

Malignant Lesion	Appearance	Image
<p>1.) Basal cell carcinoma (BCC)</p> <p>Non-melanoma Skin Cancer</p>	<p>Open sore, red patch, pink growth, shiny bumps, scars or growths with slightly rolled edges [19]</p>	<p>Figure 1.1 - BCC Sample from DERMOFIT [20]</p> 
<p>2.) Squamous cell carcinoma (SCC)</p> <p>Non-melanoma Skin Cancer</p>	<p>Scaly red patches, open sores, rough wart-like skin, or raised growths. May look distinct on each person [21].</p>	<p>Figure 1.2 - SCC Sample from DERMOFIT [20]</p> 
<p>3.) Melanoma (MEL)</p> <p>Melanoma Skin Cancer</p>	<p>Presence of a new mole or a difference in existing one. May be bigger than usual and be itchy or bleed [22].</p>	<p>Figure 1.3 - MEL Sample from DERMOFIT [20]</p> 
<p>4.) Intraepidermal Carcinoma (IEC)</p> <p>Pre-cancer</p>	<p>Well-defined pink or red scaly, fairly flat, similar to superficial BCC but often with more scale and dull in color. Difference may be absent [23].</p>	<p>Figure 1.4 - IEC Sample from DERMOFIT [20]</p> 
<p>5.) Actinic keratosis (AK)</p> <p>Pre-cancer</p>	<p>A Patch which is scaly and rough in texture and most likely found on the arms, head and face [24].</p>	<p>Figure 1.5 - AK Sample from DERMOFIT [20]</p> 

It is important to note the above list is not exhaustive and many other malignant skin lesions exist. Less common types of skin cancer include Merkel Cell tumors and Derma Fibrosarcoma Proturans (DFSP).

Apart from malignant lesions, there also are many benign skin lesions which are common for people to have with potentially no harm. These include Moles, Dermatofibroma, Seborrheic keratosis, Pyogenic Granuloma and Vascular lesions. These are more clearly visualised in Chapter 6.

Whilst skin cancer is one of the most visually detectable types of cancer, the main problem with skin lesion classification also tends to be their similarity in appearance such as the similar appearance of a Mole and a Melanoma. Even sophisticated classifiers can have a difficult time in differentiating between these. Hence, any additional

information in a dataset can potentially also be helpful to classifier in distinguishing between lesions. Whether adding GAN-generated skin lesion images can do this is still in question.

1.4 Research Question

The motivation of this thesis project is to answer the following research question:

Can GAN-generated images added to a skin lesion dataset improve the performance of deep learning classifier?

To test this research question, a VGG-16 deep learning classifier was used as a test framework due to its high performance in image recognition tasks. Furthermore, this classifier achieved high accuracy in a past DERMOFIT dataset classification paper making it suitable for the chosen dataset [25] (explored further in Chapter 2). An assumption made by the research question is that the the deep learning model serves as an effective and accurate “information extractor“. To explore this assumption further, the GAN-generated images were also presented to professional dermatologists to verify their realism through a Visual Turing Test (explored in Chapter 6). This served as a form of qualitative evaluation for the results achieved using the test framework.

1.5 Roadmap

The aim of this project was to explore if GAN-generated images could add any new information to skin lesion datasets and hence improve deep network classifiers.

1. The initial step was investigating what work had already been done in the field of skin lesion classification and data augmentation (explored in Chapter 2).
2. Next, several types of GANs were investigated. A Wasserstein GAN was chosen for its ability to help avoid mode collapse (a common problem when training GANs) compared to the original GAN loss function. Various class-specific GANs were chosen over one large conditional GAN as the large class imbalances led to biased output class samples (explored in Chapter 3)
3. A custom test framework was then built using the VGG-16 deep learning classifier to verify whether the generated skin images actually improved accuracy. The hyper-parameters for the VGG-16 were chosen and various metrics were selected for reporting test results (explored in Chapter 4).
4. The generated GAN images were evaluated using the test framework with 5-fold cross-validation to ensure the results are reproducible. This was then compared with the accuracy of the classifier with no data augmentation, affine data

augmentation, and then both GAN and affine data augmentation. Statistical results were found using a paired T-test (explored in Chapter 5).

5. To evaluate the qualitative accuracy of the skin lesion images, a Visual Turing Test (VTT) was conducted with professional dermatologists . They were presented a randomised sample of real and fake GAN-generated images and asked to classify them visually. Results were drawn from the sample (explored in Chapter 6).
6. The methodology used and results were analysed and evaluated. Potential areas of improvements and further exploration were suggested to improve the accuracy of skin lesion classifiers through data augmentation (explored in Chapter 8).

Chapter 2

Background & Literature

The aim of this chapter is to provide background material on recent research into skin lesion classification, research using the DERMOFIT skin lesion dataset [20] and the use of GANs in generating realistic skin lesions.

2.1 The role of Artificial Neural Networks in skin lesion classification

Many of the classifiers that have achieved the highest levels of accuracy in skin lesion classification have leveraged the use of Artificial Neural Networks [4] [26] [27].

An Artificial Neural Network (ANN) is a directed graph which acts as an information or feature extractor [28]. The primary function of a neural network is to give a prediction based on certain input information it receives. The way it computes this prediction is by learning over time the relevance of the input information to the output task [29]. The figure below illustrates a Perceptron, one of the simplest types of Artificial Neural Networks.

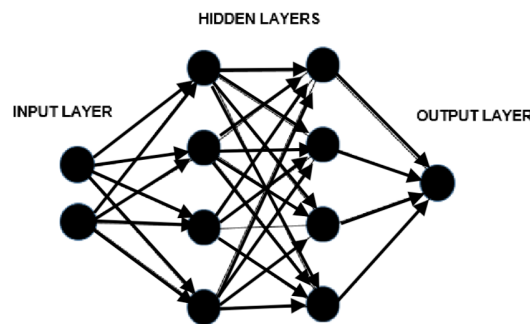


Figure 2.1: Example of an Perceptron, the simplest type of Artificial Neural Network [30]

As seen in the diagram above, the input nodes on the left contain input information fed into the network. Whilst the last node on the right takes in the processed infor-

mation from the middle layers and outputs a final outcome. The middle columns of nodes (or layers) is where the main feature extraction process occurs [30].

The network essentially learns through feedback. When it is training, it takes in an input, creates a prediction and then calculates how far off its predictions were from the real value. The error of this prediction is then used to adjust the weights and biases (or significance) of its middle nodes so that it is less likely to make the same mistake again. This process of the network adjusting its weights and biases according to the error of its predictions is called back-propagation [31].

As one progresses along the layers of a neural network from left to right, each node in the layer suddenly takes into account more paths from other past nodes and hence more information. This can be seen by tracing back the different paths to a node in a deep layer and seeing how many inputs it is connected to [29].

The size and depth of an Artificial Neural Network is up to the designer, however the more layers a neural network has the “deeper“ it is seen to be [32]. An ANN which is very deep with many layers is called a Deep Neural Network. A rough rule of thumb is that the more layers a ANN has, the more feature extraction capabilities it is deemed to have [33]. This is because the more layers it has, the more combinations of higher and lower level features are taken into account by the network [33] [34]. The types of layers used in an ANN are not limited to just fully connected layers (connecting all possible combination of nodes) like in a perceptron, but instead can be of many different types such as pooling or activation layers. These different types of layers can have different effects on the significance and strength of the nodes and connections [29]. One type of layer popularly used for visual prediction tasks is the convolutional layer. This layer is typically used in a type of ANN called a Convolutional Neural Network, which is what most deep learning skin lesion classifiers are based on.

2.2 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of Artificial Neural Network which is most popularly used for vision tasks such as object detection. The difference between a CNN and a Perceptron (the simple network illustrated previously) is that a CNN uses convolutional layers within its network structure. What a convolutional layer does is that it strides a filter of a certain size across an image and applies the dot product of that filter with the part of the image it overlaps with [35].

Filters are good for detecting different features or shapes within an image [36]. For example a filter containing an L-shape can be useful in attempting to detect L-shaped corners in an image. As the filter is dot producted with the overlapped part of the image, the more the overlapped segment matches with the L-shape, the higher the output of the dot product will be and hence the higher the significance. Below can be seen an example of a CNN being given the image of a woman and applying a filter to a sub-segment of the image.

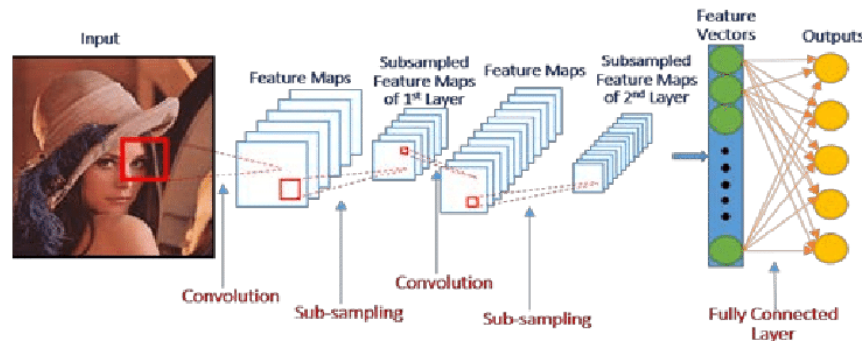


Figure 2.2: Example of Convolutional Neural Network applied to the image of a woman [37]

CNNs can be mostly seen as feature extractors. As we see in the image of the woman above, the convolutional layer strides many different types of filters onto the image that help detect the relevant lower level features (such as wrinkles on a face) and higher level features (such as the shape of a face) within the image [36]. During training, back-propagation is used to adjust the significance (weights and biases) of different combinations of inputs and features in the image and use those to predict the classification of the image. The ability for a CNN to determine important features by itself is one of the reasons it is popular model in skin lesion classification research.

2.3 State-of-the-art in skin lesion classification

There are various studies on skin lesion classification which achieved performance comparable to the state-of-the-art in accuracy. Most of these use Deep Convolutional Neural Networks.

One such study is Mendes et al's 2018 ResNet-152 architecture which achieved 96% accuracy for melanoma classification and 91% accuracy for Basal Cell Carcinoma classification [4]. A ResNet-152 is a Deep Convolutional Neural Network with 152 layers which has the benefit of using residual layers. Residual layers connect previous layers in the network to layers deeper in the network [4]. The reason for this is that it helps feed information forward faster from past layers and helps the gradient stay strong even through the many layers. Mendes et al's ResNet-152 was trained on 3797 skin lesion images from 12 different skin lesions classes and tested on 956 test images [4]. This means the dataset was made of 4753 images, which is a fairly low amount of images compared to most deep neural networks and can often limit total accuracy.

Furthermore, another model which achieved very high accuracy on skin lesion classification was Matsunaga et al's ensemble binary classifiers which won first place at the ISBI Challenge 2017 [26]. Their ensemble binary classifiers used a pre-trained object recognition Deep Neural Network trained with data augmentation as well as using the two optimisers of AdaGrad and RMSProp [26]. An ensemble classifier is essentially a combination of different neural networks which help each other in the prediction of a task. For example the input of a certain neural network may be the output of

the neural network before it. Matsunaga et al's ensemble binary classifier essentially broke up the classification of skin lesions into smaller binary decisions undertaken by Deep Neural Networks and at each stage used the output of one network as an input into a new binary Deep Network classifier.

Finally, Kwasigroch et al also achieved a high accuracy of 84% on all skin lesion classes using a transfer learning method with a VGG-19 (also a type of deep Convolutional Neural Network) and ResNet50 already trained on the 1K ImageNet classes [27]. Their results showed higher accuracy for the VGG-19, but mainly due to the ResNet-50 not having enough training data due to how deep it was [27]. These findings hence suggest that using less deep of a Neural Network such as a VGG-19 may be optimal for skin lesion classification when there is not many training images available.

One clear conclusion from various studies on the use of deep learning is that Deep Neural Networks learn better with more data. This is as more data allows more information to better tune the nodes and layers in the network and recognise common features better. However, apart from the quantity of images, the quality of data is also crucial. In terms of skin lesion datasets, the DERMOFIT dataset is recognised for having one of the highest medical and photographic consistency [20].

2.4 DERMOFIT Dataset

The skin lesion images used in this thesis project come from the Edinburgh DERMOFIT Image Library (or DERMOFIT dataset). All of these skin lesion images have a gold standard diagnosis based on expert opinion (including dermatologists and dermatopathologists) [20].

The DERMOFIT dataset consists of 1300 skin lesion images belonging to 10 different classes. Five of these classes are considered malignant lesions: Squamous cell carcinoma (SCC), Basal cell carcinoma (BCC), Melanoma (MEL), Intra-Epidermal Carcinoma (IEC) and Actinic Keratosis (AK). Whilst the other five classes are considered benign: Moles (ML), Dermatofibroma (DF), Seborrheic keratosis (SK), Pyogenic Granuloma (PYO) and Vascular lesions (VASC). It is one of the highest quality skin lesion datasets in terms of medical consistency as each image was taken at the same distance and with the same camera conditions.



Class	Number
AK	45
BCC	239
ML	331
SCC	88
SK	257
MEL	76
DF	65
VASC	97
PYO	24
IEC	78
Total	1300

Figure 2.3: Breakdown of different skin lesion classes in the DERMOFIT Dataset [25]

2.5 Work on the DERMOFIT Dataset

Various papers have used the DERMOFIT dataset for skin lesion detection and classification and have achieved considerably high average class accuracy.

One of them being Di Leo et al's 2015 paper titled "Hierarchical Classification of Ten Skin Lesion Classes" [38]. Here, a hierarchical classification system based on the k-Nearest Neighbours (kNN) classifier was used to classify the DERMOFIT skin lesions into the 10 different classes. This approach achieved 93% accuracy in distinguishing malignant from benign lesions and 67% overall classification accuracy for the 10 classes [38]. The benefit of using a K-Nearest Neighbours approach is that the most similar looking lesions will be classified the same. However, the downfall of this method is that the overall similarity of two images may not necessarily be a more important factor. Instead, it could be the presence of certain specific features within an image that decide whether its a certain kind of lesion. For example, Melanomas and Moles share much similar features, but tend to be distinguished by the smoothness and roundness of the lesion.

A more recent study in 2018 by Bertrand instead attempted to use a Deep Learning approach to classify the 10 DERMOFIT skin lesions classes. This approach investigated different Deep Neural Networks such as the VGG-16 and ResNet50 to improve the classification accuracy on the 10 classes [25]. This study achieved an accuracy of 78,5% using the VGG-16 network, 78,7% using the ResNet50 and an accuracy of 80,1% using a hierarchical ensemble method which instead divided the classification into binary decisions much like Matsunaga et al [25].

A continuation study of this in 2019 by Fisher et al compared the findings of Bertrand to an optimised Hierarchical K-Nearest Neighbour classifier. Fisher demonstrated that the Hierarchical K-NN and the Deep Neural Network methods were reasonably comparable in terms of accuracy on the DERMOFIT dataset [39]. Fisher's K-NN classifier achieved 78.1% accuracy on the 10 classes, whereas the Deep Neural Network method achieved 78,7% accuracy [39].

Whilst, GAN-generated images could be added to the training dataset of either of the two types of classifiers, it was decided to use only the Deep Learning classifier for

this project. This was because the Deep Learning classifier is designed to focus more on extracting important features of an image rather than focusing on classification through similarity. As the goal of this project was to understand if GAN images could create images that add new and useful information, a Deep Learning classifier was preferred due to its ability to adjust and learn new features.

The addition of GAN-generated images to the dataset can be considered a form of data augmentation, a technique commonly used in deep learning classifiers to improve the variability of the data.

2.6 What is Data Augmentation?

Data augmentation is a standard technique used to expand data sets in image classification tasks with the goal of solving overfitting. Overfitting occurs when a machine learning model has focused too much on learning about the training data and does not perform well on unseen data. Data augmentation procedures typically involves affine image transformations such as rotations, shifts, crops flips, addition of noise and changes in lighting and color settings. An affine transformation is a transformation where parallel lines in the first image remain parallel in the transformed image [40]. Figure 1.1 gives examples of simple augmentations that are shown to be commonly effective in helping generalise the overall training data [41].

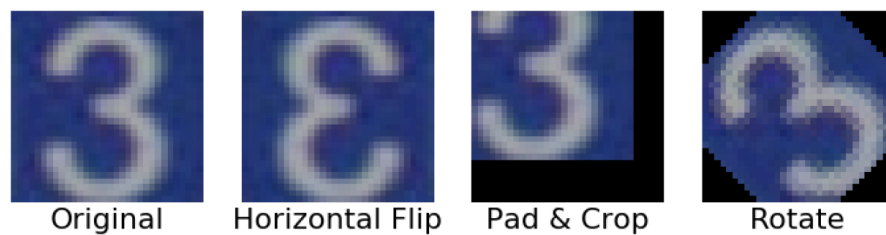


Figure 2.4: - Example of typical affine data augmentation [42]

This typically helps capture the images that “should have been” in the training set. The variation of camera-related factors such as different angles, scales and lighting levels are reduced by data augmentation. This way the model learns to better recognise genuine features of an object and not conditions caused by the camera or environment

Furthermore, applying data augmentation can help combat imbalanced class distributions where one certain type of image may be more common in a dataset than another type. This can often cause problems in weight-based machine learning models such as Deep Neural Networks because large enough image classes can dominate weight updates and lead the classifier having a preference for certain predicted classes for any given input. In Perez et al’s paper “The Effectiveness of Data Augmentation in Image Classification using Deep Learning“, the authors show that standard affine data augmentation can be an effective way to reduce overfitting of vision machine learning models [41]. The standard data augmentation resulted in an accuracy increase of 7% for a mid-accuracy model and a 3.5% accuracy increase for a higher accuracy model

[41]. Hence, the effectiveness of data augmentation may be larger for models that are not achieving top 85%-90% accuracy.

In summary, more data is likely to lead to greater accuracy for deep learning models. One such way of adding more data is to leverage the use of GANs to generate synthetic images.

2.7 How GANs create synthetic images

Since their popularisation in 2014, Generative Adversarial Neural networks (GANs) have become renowned for their distinct ability to produce highly realistic artificial images given a training set [43] [44]. The way GANs work is by having a generator (a neural network which generates synthetic samples) and a discriminator (a neural network classifying whether an image as real or fake) adversarially trained against each other to simultaneously improve [43]. This structure of a GAN is shown in the following figure.

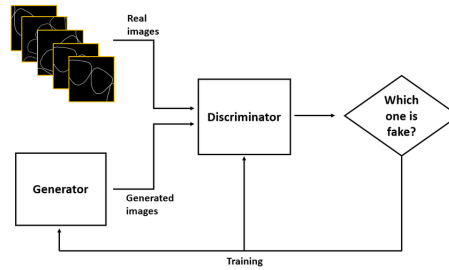


Figure 2.5: Diagram of GAN discriminator and generator network [45]

From the structure above, it can be seen that the generator and discriminator play a game against each other. In this game, the generator attempts to fool the discriminator feeding it synthetic data and the discriminator attempts to correctly classify if a given image comes from the training set or the generator. A common analogy used is that the generator is a con artist, trying to create a fake painting and the discriminator is an art detective trying to verify whether paintings are real or fake [43]. They constantly compete, each getting better at their own task.

The game played between the generator and the discriminator can be described by their loss function [43]:

$$\min_G \max_D V(D, G)$$

$$\text{where } V(D, G) = E_{x \sim p_{data}}[\log(D(x))] + E_{z \sim p_z(z)}[\log(1 - D(G(Z)))]$$

And:

- D = the discriminator,
- G = the generator,
- Z = random input into the generator,
- X = input to the discriminator

Here we see that the generator is incentivised to fool the discriminator (by minimising $1 - D(G(Z))$), whereas the discriminator is incentivised to correctly classify the generator's fake images and correctly classify real images (by maximising both $1 - D(G(Z))$ and $D(X)$). This is called a minimax game as to the discriminator and generator, maximising their score is equivalent to minimising that of the opponent.

One of the most interesting aspects about GANs is that they set up the problem of generating realistic images as a self-supervised problem [46]. This is because the job of the discriminator network is essentially to supervise the images produced by the generator network and tell the generator if it believes the images produced are realistic or not. GANs very cleverly set up the problem of generating realistic images through the use of Game Theory and reinforcement learning to some extent [43]. As the two players (the generator and discriminator) are fighting over the maximum reward as their loss functions above contradict each other in a zero sum game. By fighting over this reward, they essentially both get better at their individual tasks and if successful, end up letting the generator produce highly realistic images.

2.8 Existing work on skin lesions GANs

Limited research has been conducted using GANs to generate realistic skin lesions images and none has been done using the DERMOFIT dataset.

In 2018, Baur et al explored the generation of realistic skin lesions images using a DCGAN and LAPGAN. Their findings show that with the help of progressive growing (PGAN), high quality skin lesion images can be created that are difficult even for professional dermatologists or deep learning experts to distinguish [47]. However, while the images were deemed highly realistic, it is still unclear as to whether these images could be used to enable skin lesion classifiers to improve in accuracy.

In Bissoto et al's 2019 paper on GAN skin lesion generation, they opt for an approach where the GAN generator begins from semantic label maps rather than random noise. In the study they compare the use of a DCGAN, pix2pixHD, PGAN and found that combining both semantic and instance maps led to the most realistic generated images [48]. This allows for the generator to start from a more informed image, which may be beneficial when using a limited dataset, but also requires more labelling which may be detrimental to the scalability of this technique.

Finally, a recent paper published in May 2019 by Pollastri et al compared a Deep Convolutional GAN (DCGANs) to a Laplacian GAN (LAPGAN) as a form of data augmentation. They showed that using LAPGAN-generated images were able to increase baseline accuracy by 0.82% [49]. Furthermore segmentation masks were used and generated by the GANs to help further improve the accuracy of the GAN.

Whilst there has been some research conducted on creating realistic skin lesions with GANs, their use as data augmentation for Deep Neural Network classifiers remains still quite an unexplored field. Especially, given that GAN data augmentation has not

been conducted on a dataset similar to that of the DERMOFIT dataset.

The DERMOFIT dataset contains skin lesion images with some of the highest quality medical consistency [20], but this also means due to the higher standard that it has lower image amounts compared to image datasets such as HAM10000 with 10,000+ images. This makes it an ideal candidate for data augmentation, which seem to achieve greatest performance gain on smaller datasets. [41].

Chapter 3

GAN Implementation

“I have not failed. I’ve just found 10,000 ways that won’t work.”

— Thomas A. Edison [50]

The purpose of this chapter is to explain the methodology and experimentation behind the selection and implementation process of the GAN chosen for generating synthetic skin lesion samples.

3.1 Building the GAN

A GAN is made up of two main components - a discriminator and a generator network. The task of the generator is to create synthetic images that are realistic enough to fool the discriminator. Meanwhile, the task of the discriminator is to supervise the generator and correctly label whether the images coming to it are from the training set (real) or the generator (fake) [43].

The discriminator typically consists of a Convolutional Neural Network (CNN), as explored in Chapter 2, which has the simple binary classification task of predicting whether an image given to it is real or synthetic. On the other hand, the typical generator used in a GAN can be thought of as the reverse of a CNN structure [46]. A CNN takes in an image (a matrix of pixels) and down-samples them to output a series of features which it then uses to make a prediction.

The generator does the exact opposite of this - it takes a series of latent vectors (in our case a random input vector) and up-samples it until an image of the required size is produced. The structure of the generator is almost identical to that of the discriminator, except instead of using convolutional layers (which down-sample an image), it uses transposed convolutions which up-sample an image [51]. Transposed convolutions do this by applying a large stride and using weights to insert likely pixels in between the known pixels [51]. This is done in sequence until the required size image is reached. Below can be seen the comparison of a convolution (a - discriminator)

and a transposed convolutions (b -generator) taking place.

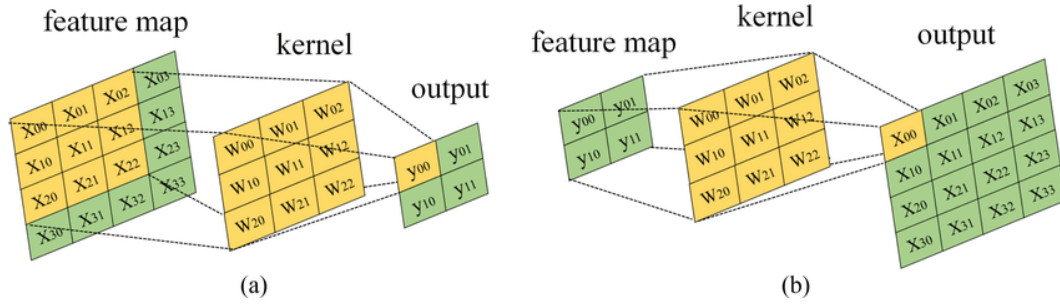


Figure 3.1: - Comparison of convolutions (a - discriminator) and transposed convolutions (b - generator) [52]

As observed, the generator is essentially the same as the discriminator, except for the use of transposed convolutions which up-sample images [52]. Furthermore, the input and output to both networks is almost the opposite. As the generator takes in a random noise vector instead of an image and then outputs a synthetic skin lesion sample instead of a feature vector or prediction.

The reason the input vector to the generator is random is so that the generator is encouraged to produce different variations of synthetic images and not just the same image from a given input [46]. This is as the generated synthetic image is created by passing the input vector through a number of up-sampling layers attempting to make it look more like a realistic skin lesion. Varying the size of this input vector can lead to changes in the quality and smoothness of the synthetic images produced.

3.2 Selecting the dimensions of the generator input and output

The size of the random input vector that is given to the generator essentially defines to some extent the quality and range of the output synthetic image [46]. This is because the output image can only vary according to the degrees of freedom (or dimensions) of the input vector. However, increasing the dimensionality of the input noise vector greatly affects the total training time and computational resources needed.

Furthermore, a higher noise vector dimension does not ensure the output image will necessarily be of greater or more realistic quality. In fact it may make the synthetic image look less smooth as up-sampling tends to create pixels similar to those around it. Therefore there is a trade-off between lower level details and computational complexity and smoothness when selecting the size of the input vector.

In our case, 2 sizes of input vector were tested. Firstly, an input vector of 128 dimensions, secondly an input vector with twice the dimensions at 256 dimensions. The results of both can be seen below.

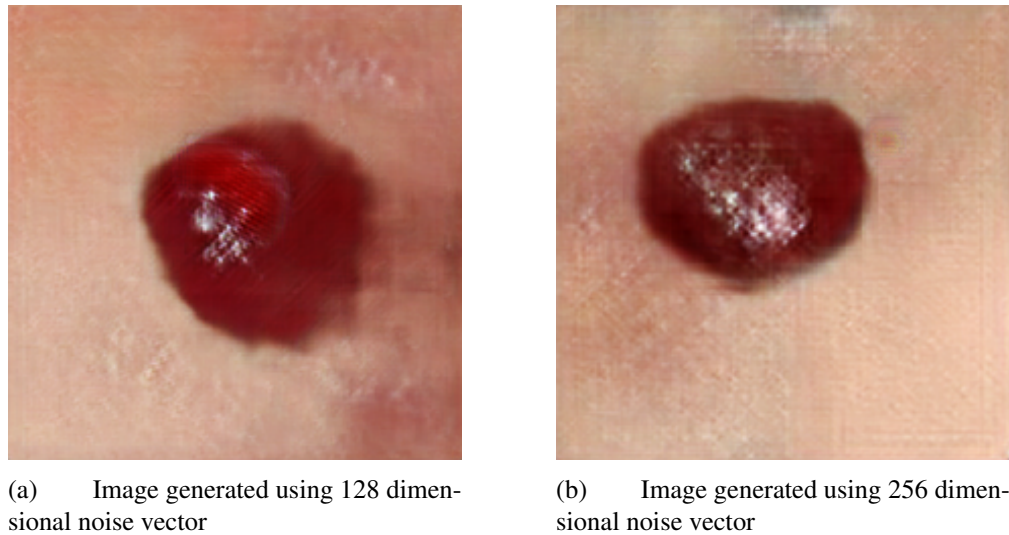


Figure 3.2: Comparatively the 128 dimensional noise vector image is similar yet slightly less detailed than the 256 noise vector.

The 256 dimensional vector was chosen, however this performed at the peak of the hardware resources used. Hence, larger dimensionality was not tested and is left for further exploration. Furthermore, upon inspection the 256 dimensional images did not appear very visually different to the 128 dimensional noise images, yet did show slightly more detail. Hence, the 256 dimensional input was chosen.

Meanwhile, the output dimensions of the generator were chosen to be a $3 \times 224 \times 224$ pixel sized matrix. This is because, following [25], the VGG-16 takes in an image with a length and width of 224 pixels and 3 RGB color channels.

Therefore, these input dimensions (1x256 pixels) and output dimensions ($3 \times 224 \times 224$ pixels) were chosen as network parameters when training the GAN.

3.3 Training the GAN

Whilst exploring the ideal GAN for synthetic skin lesion image generation, a GAN from Goodfellow's original "Generative Adversarial Neural Networks" paper was implemented [43]. This is because it allowed a baseline for comparison to other more modern types of GANs. We adjusted the original GAN network and layers to allow a 1x256 dimension input and $3 \times 224 \times 224$ dimension output to fit our specific task.

Training the GAN involves taking turns between training the generator and the discriminator networks. The most interesting part about training a GAN is that the generator never truly sees the images in the training set [46]. Instead, it updates itself according to how well its synthetic samples manage to fool the discriminator.

In training the generator, it is first fed random noise vectors of a certain dimension. It then up-samples this noise and outputs an image which is fed to the discriminator.

The generator is then updated according to whether the discriminator correctly or incorrectly classified the synthetic images [43].

In training the discriminator, it is first fed equal amounts of images from the generator and the training data to make sure it is exposed to both real and fake images equally. The discriminator is then evaluated and updated using back-propagation with the binary cross entropy of how well it was able to classify whether an image is real or fake [46].

Although the discriminator and generator share the same loss function, they do not update their weights simultaneously and instead each take turns updating individually. For example, the discriminator typically trains for 2 batches, whilst the generator weights are frozen. Then the discriminator weights are frozen and the generator trains for 2 batches. This process is repeated until convergence [46]. This way each network then has time to train and adjust without simultaneous influence from the other's direct actions.

We trained our GAN by letting the discriminator network train for 2 batches, whilst the generator trained for 2 batches. This is typically recommended as then the discriminator and generator improve more equally [46]. However, despite attempts at equal training, a common problem encountered when training GANs is that of mode collapse.

3.4 Challenges with mode collapse

Initially we used a custom GAN using the original GAN loss function (explained in Chapter 2) to produce realistic skin images. The problem encountered with this approach was its proneness to mode collapse.

Mode collapse is the main challenge encountered when training GANs. What occurs is that the generator learns to map the random noise vectors fed into it to the same shapes, colours and patterns [53]. This means essentially very similar images are generated. The figure below shows how the custom GAN experienced mode collapse, mapping random colour vectors to the same patterns of melanoma images.

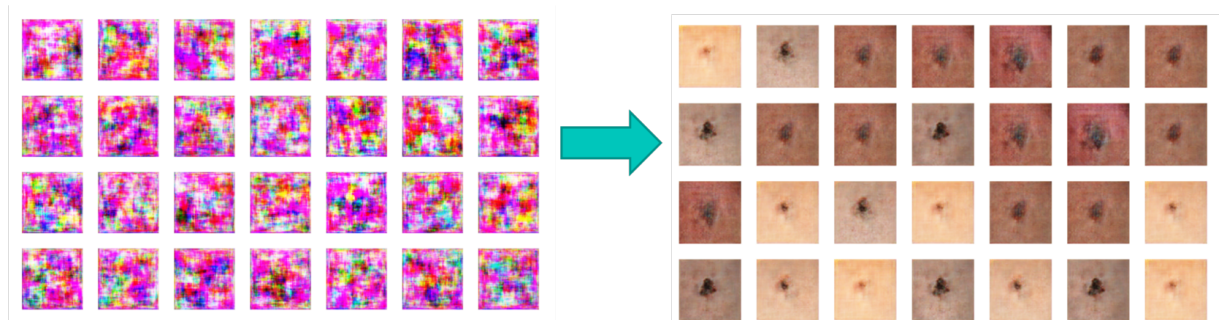


Figure 3.3: - Mode Collapse experienced using original GAN loss equation on Melanoma Images

This ends up partially ruining the purpose of the random input noise vectors that are

meant to help the generator create a variety of types and styles of generated images. This is observed in the figure above, where the initial random noise vectors after one iteration are presented on the left and the generator output after 5000 iterations are shown on the right. We observe the problem as the random noise vectors map to a similar looking melanoma spot. This is because the generator has learnt to fool the discriminator by essentially mapping different random noise vectors to the same melanoma shape and pattern, but in different pigmentations.

Several newer types of GANs such as the Wasserstein GAN have been created in order to help reduce the effects of mode collapse.

3.5 The Wasserstein GAN

A Wasserstein GAN is a GAN that uses the Wasserstein distance in its loss functions instead of the original GAN loss function that uses Jensen-Shannon Divergence as a distance metric. Jensen-Shannon Divergence is a probability distance metric which is based on measuring the vertical distance between two overlapping probability distributions [54].

The Wasserstein or Earth Mover's distance function is instead defined as the minimum amount of energy taken to move all the "dirt" composing a given probability distribution to all the "dirt" forming another probability distribution. Wasserstein distance measures the distance between two probability distributions according to the horizontal difference of the points in the distributions rather than their vertical difference [55].

The main reasons the Wasserstein distance is seen as smoother than the original GAN loss function is that Jensen-Shannon Divergence (JSD) suffers a discrete jump when two probability distributions are not overlapping. This is because JSD uses a vertical distance to compare distributions rather than a horizontal measure such as the Wasserstein Distance meaning non-overlapping distributions are harder to measure [55]. This is shown in the following figure.

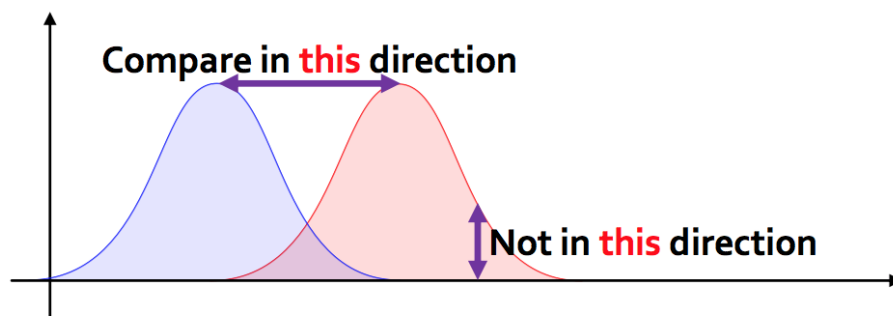


Figure 3.4: - Example of Wasserstein or Earth Movers Distance [56]

The Wasserstein distance gives the advantage of providing a smooth distance metric

across all differences in probability distributions, which makes training of the GAN generator much more stable. This can be visually observed with the synthetic images mapping to various different shapes and patterns under the Wasserstein loss function implementation.

A Wasserstein GAN based on Gulrajani et al's paper, "Improved Training of Wasserstein GANs" was implemented and the various up-sampling and convolutional layers were changed to fit our chosen input dimensions (1x256) and output dimensions (3x224x244) [57]. This was then trained and the following figure illustrates our practical results using the improved WGAN implementation.

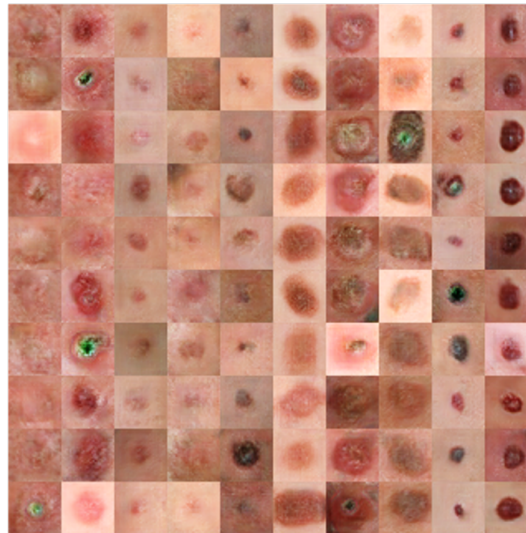


Figure 3.5: - Synthetic lesion images produced by the Wasserstein GAN

The above figure demonstrates how the Wasserstein GAN is able to produce images with a much greater variety of shapes and patterns compared to the GAN with the original GAN loss function in Section 3.6. This is as the Wasserstein GAN is able to better avoid mode collapse through its use of a smoother loss function.

The choice of GAN used in experimentation was hence chosen to be a Wasserstein GAN (WGAN) due to its ability to avoid mode collapse and generate a greater variety of skin lesion samples. Nonetheless, as multiple classes of skin lesions needed to be generated, it was investigated whether one conditional WGAN or multiple WGANs (one for each skin lesion class) was optimal to create realistic synthetic lesion images.

3.6 Many GANs or a single GAN?

The typical objective of a GAN is to train its generator to create realistic synthetic samples of the training data given to it. However, when trying to get a GAN to create multiple types or classes of images, there exist two main solutions.

Firstly, a GAN can simply be created for each class of skin lesion or secondly, a conditional GAN can be created to generate multiple types of classes [58].

The way a conditional GAN works is very similar to that of a normal GAN, however instead of feeding in just a random noise vector as input to the generator, a one-hot encoding label is also appended to it, indicating what class of lesion it should try to create [58]. Next, the loss function of the generator is also adjusted to instead be a categorical cross-entropy loss instead of a binary cross-entropy loss which essentially just allows updates to take into consideration the different categorisation of the classes.

Both of these approaches were tested, with a conditional GAN and a 10 different GANs (one for each skin lesion class). In practice, it was found that the 10 different skin lesion GANs performed much better than the conditional GAN in generating subjectively realistic skin lesions for each class. The reason we believe this was the case was due to the large class imbalances within the DERMOFIT dataset. For example, whilst the conditional GAN may have attempted to generate a Pyogenic Granuloma (PYO) lesion of which there are only 24 in the dataset, the added class label input into the generator may have simply been ignored and over-powered by the weight updates created by larger classes such as the 331 Moles (ML) in the dataset. This is one of the large problems of imbalanced class distributions, as more images in a certain class can tend to unfairly bias a classifier towards a certain class over another. This was experienced in the conditional GAN as the main output tended to appear similar to a Mole (ML), despite the intended input class to be another type of lesion.

Therefore, 10 individual WGANs, one for each class, were selected as the final GAN to be used to generate skin lesion images. This was because the many individual WGANs were able to avoid the problems associated with class imbalances and bias towards a certain type of lesion.

Whilst, the implementation of the multiple class WGAN was selected, it was still in question as to how long to train each of these WGANs to generate the most realistic-looking skin lesions.

3.7 When to finish training the GAN

Currently, there is no practical method for objectively deciding if a GAN has finished training or reached its final outcome. Whilst theoretically, the point of convergence is the Nash Equilibrium or a joint local maximum payoff strategy, this is rarely reached in practice due to the vanishing gradient problem and failures of convergence [59].

Hence a typical stopping criteria is when the discriminator reaches an accuracy close to 50%. This is as the discriminator is practically guessing between whether an image is real or fake and hence the generator has successfully produced images that fool the discriminator. Nonetheless, this is still not an objective stopping criterion, hence the main measure of GAN performance tends to be primarily based on the subjective image quality of the samples produced [46].

The stopping criteria used to train the WGAN on each individual class of skin lesions was subjective as described. After running each until the point of over-training, it

was evident that the WGAN tended to produce the highest quality subjective images after 3500 iterations of training. The effects of training iterations on subjective image quality can be observed in the figure below.

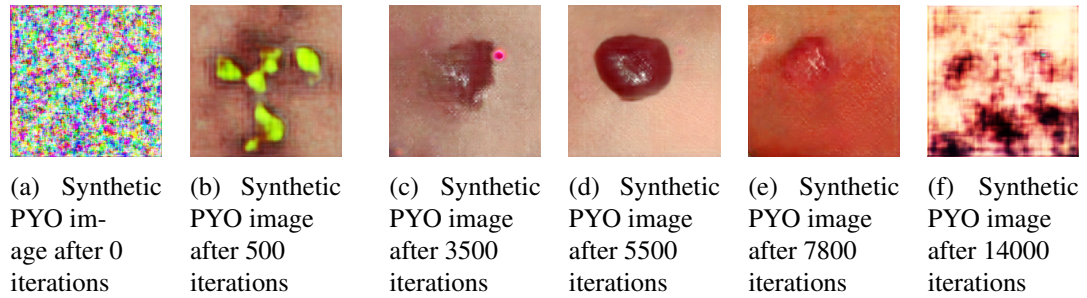


Figure 3.6: Comparing synthetic PYO images generated by GANs across different amount of training iterations

Furthermore, the WGAN started over-training and producing unrealistic images at around 7500 iterations where the GAN became imbalanced and started producing unrealistic image samples with high colour contrast. Hence, this was used at the upper limit.

As explored in this chapter, after much experimentation, all the network parameters were chosen for the final WGAN implementation. After training each WGAN on the given iteration ranges, synthetic images were generated for each class. Chapter 6 shows what some of these WGAN-generated lesions look like due to their use in the Visual Turing Test. However, to test whether these generated skin lesion images were adding any new information, a testing methodology had to be created.

Chapter 4

Testing Methodology

The aim of this chapter is to explore the methodology and metrics used for the test framework in verifying whether the WGAN-augmented images provide any performance increase to the VGG-16 Deep Learning classifier.

4.1 Test Framework

The main question that arises from using WGANs as data augmentation is whether these generated images really create any information gain over the existing training images. A limited amount of research (covered in Chapter 2) attempts to create realistic synthetic skin lesion images using GANs. However, some of these research papers hold the assumption that the more realistic a synthetic images looks, the more information gain it provides. However, this may not necessarily be the case. As the Generative Teaching Network paper by Such et al at Uber AI Labs recently showed, the images that create the most information gain in training datasets may not actually look very realistic, but rather they encode the greatest information and variability of an image [60].

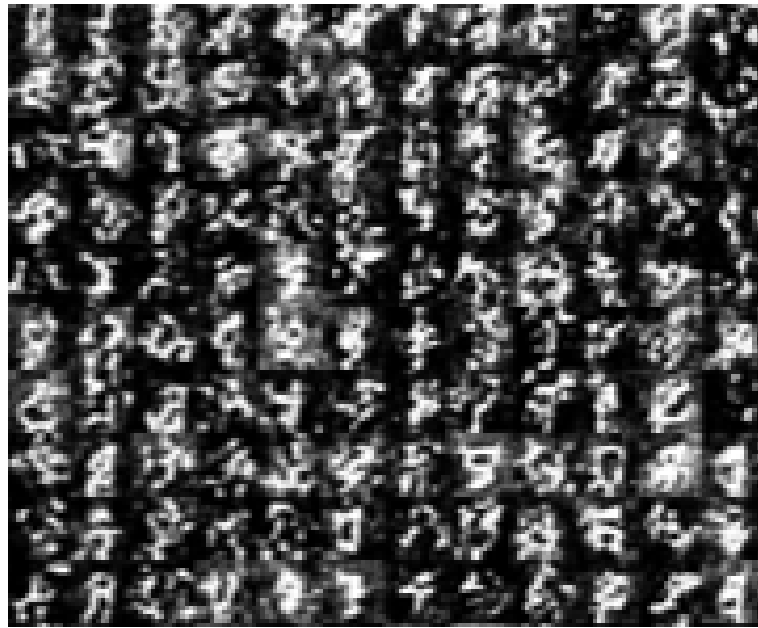


Figure 4.1: - Example of Synthetic MNIST Images produced by GTN [60]

The figure above by Such et al shows curriculum training where progressively the GAN augmented training images become more and more realistic from left to right. However, the images on the right still do not embody very accurate illustrations of the digits. Instead it appears as though the Generative Teaching Network has learnt that slightly distorted images of digits represent a more compact and regularised version of the digit images and hence are better as training images.

These findings hence suggest that there may be more objective ways to measure the amount of useful information in GAN generated image. In fact, the improvement in accuracy of a classifier trained using GAN-augmented images may be a better measure of the information gain provided by synthetic images than judging if the synthetic images look realistic or not. To test the performance of a GAN as a form of data augmentation, a Deep Neural Network from Bertrand's 2018 DERMOFIT skin lesion paper was used [25].

The discriminator used was a VGG-16 known for its high accuracy in object recognition tasks. The baseline accuracy of Bertrand's pre-trained VGG-16 classifier was 78,5% using affine data augmentation techniques and masking [25]. This network was used as the baseline to determine the accuracy improvement when using the WGAN generated images as data augmentation.

However, as our WGAN is not able produce masks for synthetic images, the same VGG-16 was used but without the masking. Furthermore, Bertrand used a 80%, 20% split for training and test sets without the use of a validation set. However, Bertrand later suggests that including a validation would be an improvement as it helps remove the bias of hyper-parameter tuning on the reported accuracy.

With these two modifications, the baseline accuracy of our pre-trained VGG-16 was roughly 63%. This drop in accuracy is expected as the VGG-16 classifier had no

masking and less training images in exchange for being able to use WGAN augmented images in the training data and also reducing the bias of hyper-parameter tuning. Below is the comparison of a normal training image, a WGAN-generated image and the equivalent masked image as used in Bertrand's paper.



Figure 4.2: Comparison of WGAN synthetic image, real skin lesion and the equivalent masked and cropped lesion

As we can see above, the WGAN-generated AK lesion was trained on the un-cropped and un-masked skin lesion AK images. However, the cropping and masking seen in image (c) as conducted by Bertrand would perhaps allow the GAN to focus more on generating the characteristic part of the lesion rather than the surrounding skin. In future work, the WGAN could be modified to create synthetic masks too as in [48], however this is beyond the scope of this project.

Given the limited data, the VGG-16 had less data to learn from. This is why the common technique of transfer learning was used to help pre-train the network.

4.2 Transfer Learning as Feature Extraction

When using limited datasets, transfer learning can help a deep learning model learn how to extract general features from a variety of different images before being trained to specialise on a smaller, more limited dataset [61]. Transfer learning is when one trains a deep learning network on a larger dataset, until its weights are properly tuned before training it on the dataset of choice.

Transfer learning can be thought of as teaching an infant (in this case the discriminator) how to distinguish between many different types of animals before using these transferable skills to distinguish between different breeds of dogs.

As in Bertrand's 2018 paper, a VGG-16 was pre-trained on the ImageNet dataset and then fine-tuned on the DERMOFIT dataset [25]. This is because transfer learning allows the VGG-16 to learn to recognise general object features, before specialising itself on skin lesion classification. Furthermore, Mutsunaga et al and Mendes et al similarly used transfer learning methods in their skin lesion classification papers to improve the feature extraction capabilities of their classifiers [26] [4].

Apart from the use of transfer learning, a variety of different hyper-parameters were tested to try and achieve the highest baseline accuracy for the VGG-16.

4.3 Choice of Hyper-parameters

A variety of hyper-parameters were tested for the VGG-16 using a grid search approach. In this grid search method different hyper-parameters were adjusted one by one, keeping whichever had the highest accuracy. The hyper-parameters tested included the learning rate, optimiser function and batch size. The following hyper-parameters were found to produce the highest accuracy for the VGG-16 of 63.9%:

Hyper-Parameter	Optimal Value found
Batch Size	20
Optimizer Function	RMSprop
Learning Rate	0.0001
Output channels of last FC Layer	1024

A larger batch size seemed to lead to a slightly higher accuracy. The reason for this is thought to be because then the updates to the weights of the VGG-16 are conducted more smoothly after a larger amount of images have been seen. Furthermore, RMSProp was found to be the preferred optimiser. This may be because compared to a normal stochastic gradient descent, RMSProp allows an adjusted and smoother descent due to its implementation of moving average descent using the square of the gradient. Finally, a small, but not overly small learning rate seemed preferred as this way the VGG-16 was able to find a better local minima without overshooting or undershooting too much. Furthermore, the choice of output channels neurons was 1024 for the last fully connected layer as any amount above this seemed to have no influence on the accuracy obtained, perhaps due to a redundancy of neurons.

As the hyper-parameters were selected, the VGG-16 deep learning classifier was implemented. Following this the different scenarios for comparing the performance of the WGAN-augmentation were investigated.

4.4 Baseline and Comparison

To test the performance gain of the WGAN augmentations, the VGG-16 was trained and tested on the following 4 scenarios of data augmentation:

1. Skin lesion dataset without data augmentation
2. Skin Lesion dataset with affine augmentation
3. Skin lesion dataset with WGAN augmentation
4. Skin lesion dataset with both affine augmentation and WGAN augmentation

These 4 scenarios allowed to compare the effect that the WGAN augmentation had over no augmentation, but also how it compared to the typically used affine data

augmentations. In the final scenario both WGAN and affine augmentations were combined to see if WGAN augmentation could perhaps provide any information gain when combined with affine augmentations.

Although the 4 scenarios allowed a better comparison of the performance of WGAN-augmentation, focus had also to be placed on the reproducibility and reliability of the results achieved.

4.5 Reproducibility and Reliability

The above 4 augmentation scenarios were conducted using a stratified train-validation-test split (60%, 20%, 20%). Although the total dataset size was small (1300 images), it was still decided to split this into a training, validation and testing set. This is because as suggested by Bertrand's 2018 paper, using solely a validation set makes it difficult to distinguish whether tuning hyper-parameters leads to a real increase in classifier accuracy [25].

5-fold cross validation was used to verify whether the specific allocation of images in the training, validation and test sets truly impacted the reported accuracy. This meant that the dataset was split randomly into 5 equally-sized stratified sets. These were then swapped around in all 20 possible combinations between the training, validation and test sets to ensure the results accurately reflected the full dataset.

The final reported accuracy was hence an average of the different allocations of the 20 different combinations of the 5 folds between the training, validation and test sets.

Nonetheless, as certain skin classes such as malignant lesions, are more important to detect correctly than others, a variety of performance metrics were tracked for the test framework.

4.6 Performance Metrics

To evaluate the performance of the 4 augmentation scenarios, the following metrics were tracked:

1. Confusion Matrix
2. Malignant vs. Benign Classification Accuracy
3. Precision, Recall and F1-score
4. Total Accuracy and Average Accuracy by Class

The confusion matrix was useful as it allowed a fine-grained evaluation of the performance on each skin class. The malignant vs. benign classification accuracy was crucial in analysing the performance of the data augmentations as the largest risk posed by skin lesions are their potentially harmful effects. The precision, recall and F1-score helped to understand what affects the data augmentation had in terms of generalisability and different classification of class-imbalanced images. Finally, as in

any machine learning model, the overall accuracy and class-specific accuracy were analysed. Statistical Paired T-tests were then conducted to verify whether any of the four augmentation scenarios produced statistically significant results.

As the test framework and desired performance metrics were established, the VGG-16 was tested using the 4 augmentation scenarios with 5-cross-fold validation and the results were recorded.

Chapter 5

Results

This chapter aims to analyse the results of the training, validation and testing for the 4 data augmentation scenarios using 5-cross-fold validation and verify if statistically significant results were achieved through the use of paired T-tests.

5.1 Training Performance of the VGG-16

The four augmentation scenarios were conducted and their results and metrics were analysed. As stated in the previous chapter, the pre-trained VGG-16 was fine tuned on the DERMOFIT dataset. This fine tuning was done for 25 epochs and the model with the highest accuracy on the validation set was recorded and then used for testing. The following figure shows the training and validation accuracy graphs across the 25 epochs for each of the 4 augmentation types.

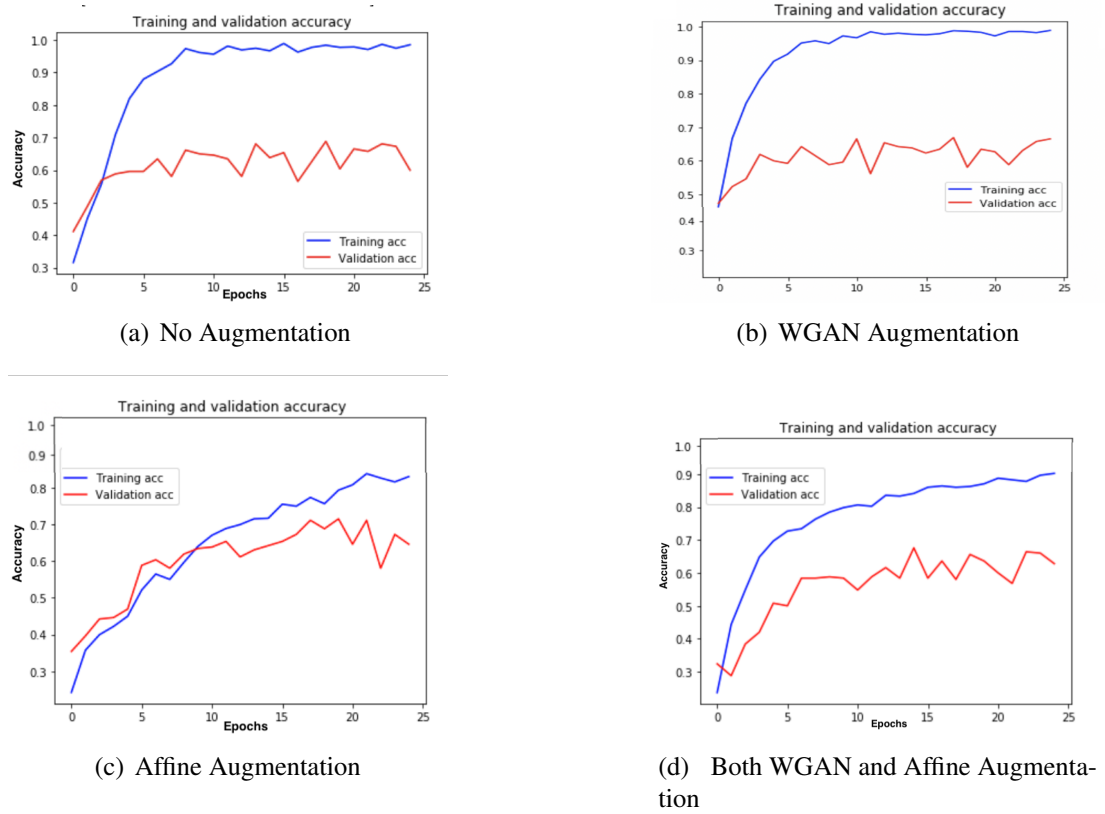


Figure 5.1: Training and Validation Accuracy versus epoch Graphs

As observed above it seemed as though both the No Augmentation and WGAN Augmentation scenarios seem to increase in training accuracy very quickly and plateau in validation accuracy after 10 epochs. However, the Affine Augmentation and Both Affine and WGAN Augmentation scenarios appear to more gradually increase in both training and validation accuracy, starting at a lower value and then reaching a higher accuracy overall. This can perhaps be explained by the larger amount of augmentation provided by affine transformations which the VGG-16 gradually learns over several epochs.

Once training was finished, the distribution of accuracy across the 20 folds were reported for each augmentation scenario.

5.2 Performance of the 4 Augmentation Techniques

To report the performance of the 4 different augmentation techniques, 5-fold cross validation was used. This meant the dataset was split into 5 equal parts and all 20 permutations of the 5 folds were tested. The cumulative results of this are shown below for each augmentation scenario.

5.2.0.1 No Augmentation

The mean and standard deviation for the 20 data permutations for 25 epochs of training of the No Augmentation scenario are shown below.

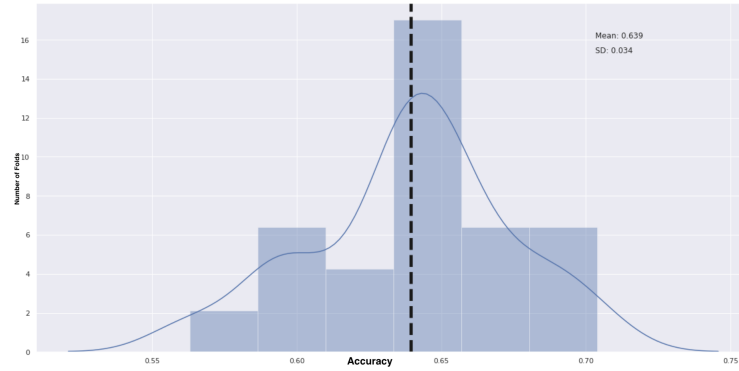


Figure 5.2: - Mean and Standard Deviation of the VGG-16 Average Accuracy with 5-fold cross validation on No Augmentation Scenario

As observed above the mean accuracy of the VGG-16 on the DERMOFIT dataset was about 63.9%. With a standard deviation of 0.034 which we can see has a range of 12.6%. This seems like quite a large range, however the smaller standard deviation means that most fold accuracies lie fairly close to each other. As these are the first training results, they will serve as a baseline for comparison to the other 3 scenarios.

Observing the aggregated confusion matrix for the 20 runs gives a better insight into how the No Augmentation VGG-16 performed on the distinct classes.

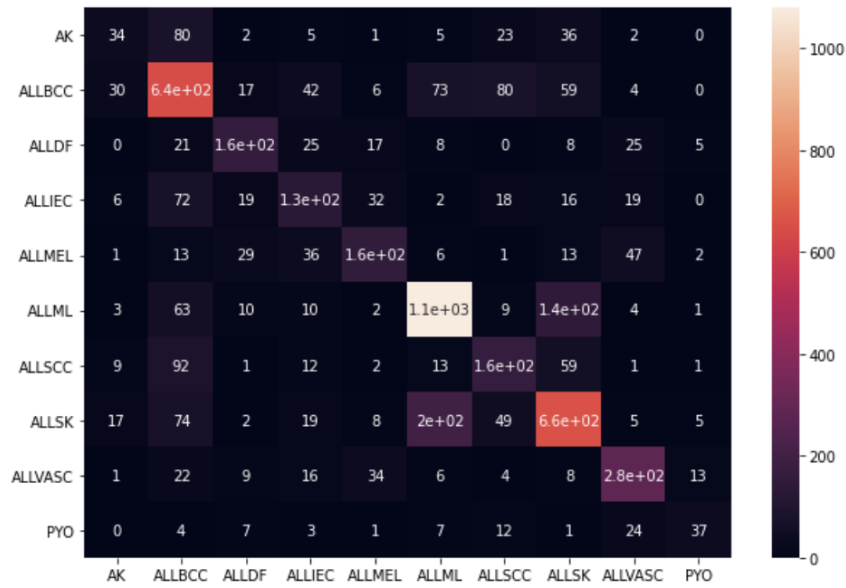


Figure 5.3: - Confusion Matrix for the No Augmentation Scenario

Class	Precision	Recall	F1-score	Support
AK	0.34	0.18	0.24	188
ALLBCC	0.59	0.67	0.63	956
ALLDF	0.62	0.59	0.61	269
ALLIEC	0.43	0.41	0.42	312
ALLMEL	0.60	0.51	0.55	304
ALLML	0.77	0.81	0.79	1327
ALLSCC	0.45	0.46	0.46	352
ALLSK	0.66	0.64	0.65	1037
ALLVASC	0.69	0.72	0.70	398
PYO	0.58	0.39	0.46	96
Accuracy			0.64	5239
Class avg	0.57	0.54	0.55	5239
Weighted avg	0.63	0.64	0.63	5239

The confusion matrix and classification report table indicate that under the No Augmentation scenario the VGG-16 seems to best classify Moles (ML) as it achieves the highest precision, recall and F1-Score of any class. This makes sense as coincidentally moles are also the largest class in the dataset in terms of number of images. We see this is a trend holds for most classes such as SK which has the second most images and the third highest F1-score. However, some classes such as Vascular Lesions which have less than a tenth of the size of the total dataset, still achieved some of the highest Precision, Recall and F1-score. This may simply mean that the classifier has been able to most easily classify this type of lesion accurately because vascular lesions are fairly more distinctive in color and shape compared to most lesions. The precision, recall and hence F1-score appears to be quite similar for most classes meaning that the amount of false positives and false negatives is quite

balanced. This is with exception of AK and PYO which seem to have fairly higher precision than recall. This is due to their False Negative count being larger than their False Positive count. A possible explanation for this may be that both these classes contain the least amount of images in the entire dataset, hence there is a bias of the classifier towards not predicting these.

Comparing the class and weighted results, we see that the class-specific results for Precision, Recall and F1-Score are roughly 6-10% smaller than those for the weighted results. This suggests that the performance on larger-sized classes was moderately greater than that of individual classes. Finally, observing the benign and malignant accuracy, the classifier seemed to perform considerably better at classifying benign lesion correctly (70.8% accuracy) than malignant lesions (53.6%). This may be due to the fact that the types of malignant lesions look similar to each other especially as some of the pre-cancers like IEC and AK are early stage types of SCC and look similar. Meanwhile, some of the benign lesion classes have no other classes that look very similar to them such as VASC or PYO lesions.

5.2.0.2 Affine augmentation

The mean and standard deviation for Affine Augmentation scenario using the 20 data permutations for 25 epochs of training are shown below:

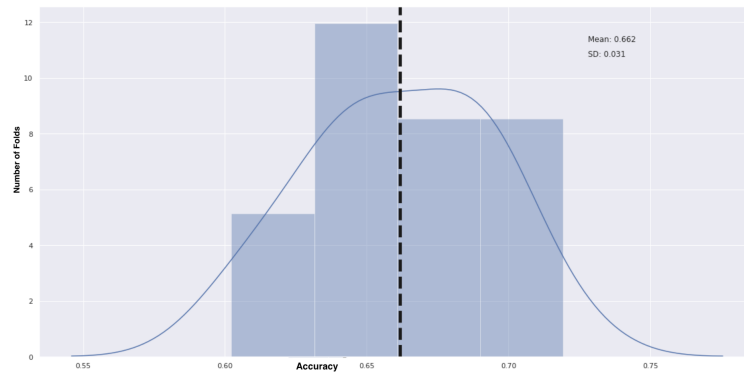


Figure 5.4: - Mean and SD of the VGG-16 Average Accuracy with 5-fold cross validation using Affine Augmentation

As observed above the mean accuracy of the VGG-16 with the Affine Augmentation is about 66.2%. With a standard deviation of 0.031 which we can see has a range of 11.5%. This is a 2.3% increase in accuracy over no augmentation which is not a large margin, but still notable. The standard deviation and range is similar to that of the non-augmented scenario.

Observing the aggregated confusion matrix for the 20 runs gives a better insight into how the Affine Augmentation VGG-16 performs on the distinct classes.

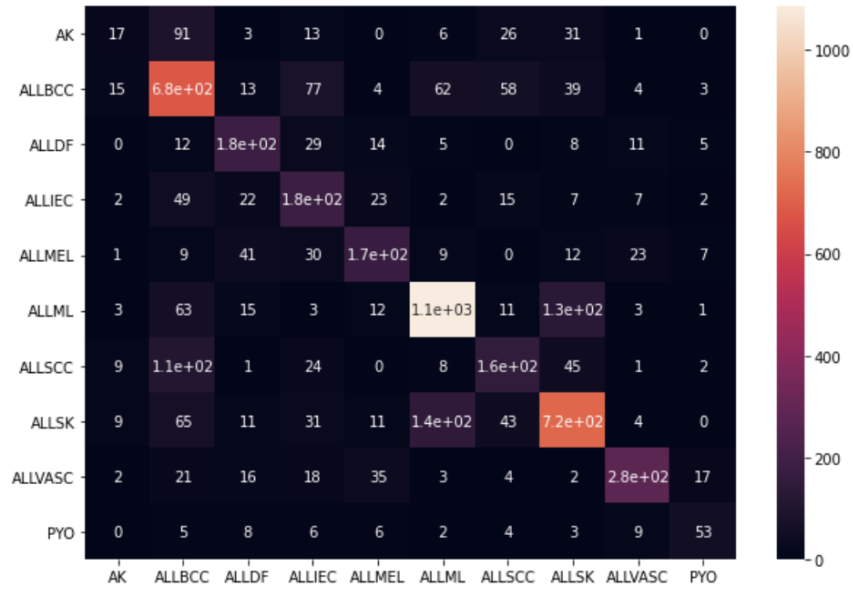


Figure 5.5: - Confusion Matrix for the Affine Augmentation Scenario

Class	Precision	Recall	F1-score	Support
AK	0.29	0.09	0.14	188
ALLBCC	0.62	0.71	0.66	956
ALLDF	0.59	0.69	0.63	269
ALLIEC	0.44	0.59	0.50	312
ALLMEL	0.62	0.57	0.59	304
ALLML	0.82	0.82	0.82	1327
ALLSCC	0.49	0.44	0.47	352
ALLSK	0.72	0.69	0.71	1037
ALLVASC	0.82	0.70	0.76	398
PYO	0.59	0.55	0.57	96
Accuracy			0.67	5239
Class avg	0.60	0.59	0.58	5239
Weighted avg	0.67	0.67	0.67	5239

Comparing the classification results of the Affine Augmentation scenario to that of No Augmentation we can observe a slight increase in accuracy in almost all F1-scores (as well as precision and recall) except for that of AKs. This essentially means that for most classes, Affine Augmentation has helped to slightly improve the accuracy of most classes. The classes most heavily impacted again were those of AK and PYO. The AK accuracy experienced a large drop in both precision and recall, whilst PYO experienced a large gain in recall of 16%. Nonetheless, these results may not be notable as the small amount of images in both these classes may make them more susceptible to more volatile accuracy changes.

Furthermore, there was a similar increase in both class-specific and weighted precision and recall of about +3-4%. This indicates perhaps that the different rotation and flips helped the VGG-16 better learn to identify the lesions under different

conditions and orientations. Finally, in terms of benign and malignant accuracy, the Affine Augmentation scenario showed a 9% increase in the benign accuracy (79.5%) and a 2% increase in the malignant accuracy (55.5%). Indicating that most of the accuracy gained by the Affine Augmentation seemed to be attributed to benign lesions.

5.2.0.3 WGAN Augmentation

The mean and standard deviation for WGAN Augmentation scenario using the 20 data permutations for 25 epochs of training are shown below:

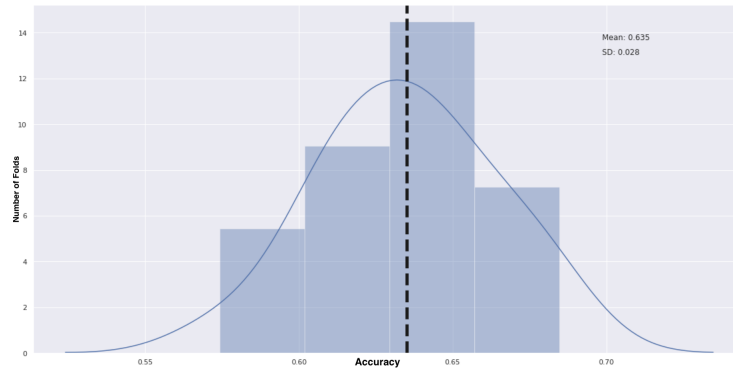


Figure 5.6: - Mean and SD of the VGG-16 Average Accuracy with 5-fold cross validation using WGAN Augmentation

As observed above the mean accuracy of the VGG-16 for the WGAN Augmentation scenario is about 63.5%. With a standard deviation of 0.028 which we can see has a range of 10%. This accuracy is essentially the same as the No Augmentation scenario. This perhaps suggests that little information gain was provided by the WGAN Augmentation. Observing the aggregated confusion matrix for the 20 runs gives a better insight into how the WGAN Augmentation VGG-16 performs on the distinct classes.

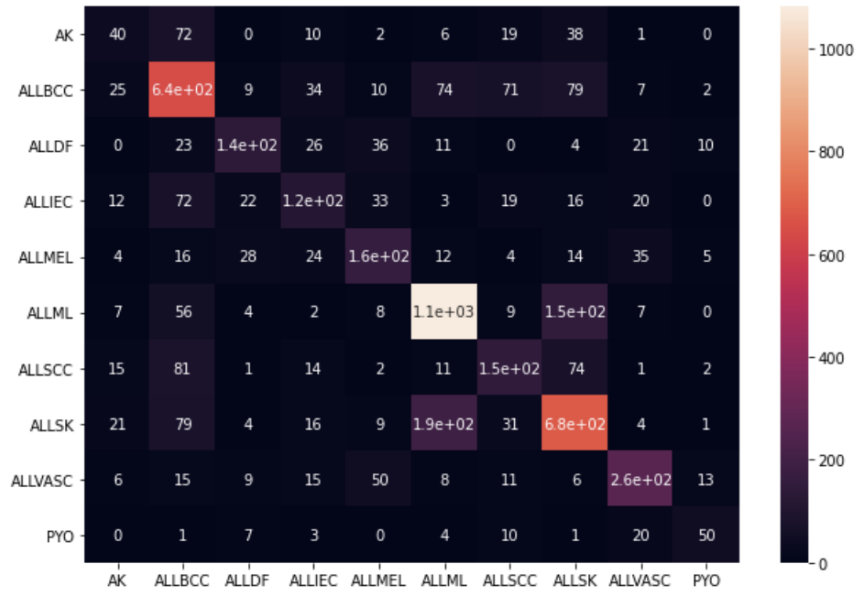


Figure 5.7: - Confusion Matrix for the WGAN Augmentation Scenario

Class	Precision	Recall	F1-score	Support
AK	0.21	0.12	0.15	188
ALLBCC	0.58	0.66	0.62	956
ALLDF	0.62	0.56	0.59	269
ALLIEC	0.42	0.37	0.39	312
ALLMEL	0.50	0.51	0.51	304
ALLML	0.78	0.79	0.78	1327
ALLSCC	0.53	0.39	0.45	352
ALLSK	0.62	0.67	0.64	1037
ALLVASC	0.71	0.70	0.71	398
PYO	0.61	0.56	0.59	96
Accuracy			0.63	5239
Class avg	0.56	0.53	0.54	5239
Weighted avg	0.62	0.63	0.62	5239

Overall, comparing the classification results of the WGAN Augmentation scenario with the No Augmentation scenario we see there is very small difference between both. The precision and recall of some classes are slightly larger in the WGAN Augmentation case, whilst others are slightly lower in the No Augmentation case. Furthermore, this is reflected in the class and weighted averages for Precision, Recall and F1-score which tend to differ by just 1% between both scenarios. This may indicate that the WGAN Augmentation has had little impact on the classifier and hence did not add much useful information to the VGG-16. Observing the benign and malignant accuracy in the WGAN Augmentation scenario we see almost no difference with that of No Augmentation with about a 1% difference in both benign accuracy (69.6%) and malignant accuracy (52.1%).

5.2.0.4 Both Affine and WGAN Augmentation.

The mean and standard deviation for the Both Affine and WGAN Augmentation scenario using 20 data permutations for 25 epochs of training are shown below:

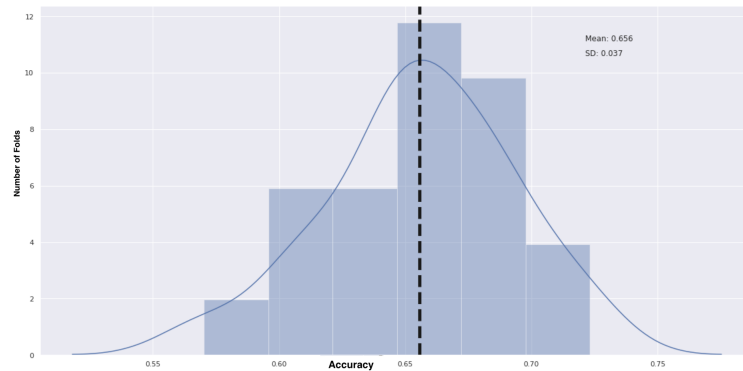


Figure 5.8: - Mean and SD of the VGG-16 Average Accuracy with 5-fold cross validation using Both Affine and WGAN Augmentation

As observed above the mean accuracy of the VGG-16 with Both the Affine and WGAN Augmentation was about 65.6%. With a standard deviation of 0.037 which we can see has a range of 15%. This is comparable to that of the Affine Augmentation alone and we see a 1.6% accuracy increase compared to No Augmentation.

Observing the aggregated confusion matrix for the 20 runs gives a better insight into how Both the Affine and WGAN Augmentation of the VGG-16 performs on the distinct classes.

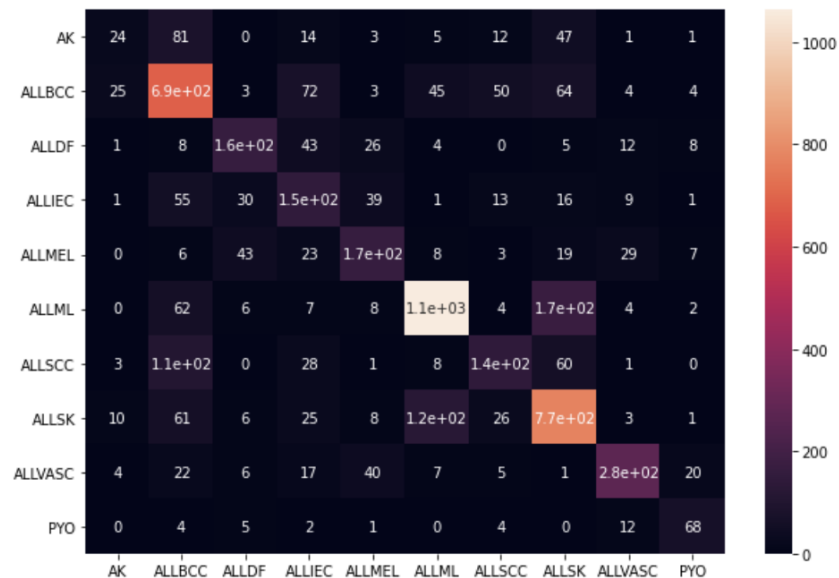


Figure 5.9: - Confusion Matrix for the Both Affine and WGAN Augmentation Scenario

Class	Precision	Recall	F1-score	Support
AK	0.35	0.13	0.19	188
ALLBCC	0.63	0.72	0.67	956
ALLDF	0.62	0.60	0.61	269
ALLIEC	0.39	0.47	0.43	312
ALLMEL	0.56	0.55	0.55	304
ALLML	0.84	0.80	0.82	1327
ALLSCC	0.55	0.41	0.47	352
ALLSK	0.67	0.75	0.71	1037
ALLVASC	0.79	0.69	0.74	398
PYO	0.61	0.71	0.65	96
Accuracy			0.67	5239
Class avg	0.60	0.58	0.58	5239
Weighted avg	0.67	0.67	0.67	5239

Comparing the individual class Precision, Recall and F1-Score between the Both Affine and WGAN Augmentation scenario and the Affine Augmentation scenario we observe little differences between each. Some classes such as AK tends to perform better in all three metrics with Both augmentations, whilst others perform slightly less well in the former. However, overall it can be seen that the majority of the metrics are quite similar. This is reflected in the class and weighted averages as the class precision (60%) and weighted precision (67%) are the same in both scenarios. Whilst the class Recall is just 1% less at 58% and the weighted Recall is the same (67%). This helps support the theory that WGAN Augmentation truly had little impact on the information added in the dataset. This is as even with the Affine Augmentation applied to the WGAN-generated images, there was still not a very noticeable change in results. Analysing the benign and malignant accuracies we see that the malignant lesion accuracy is very similar, but slightly higher at 55.8%, whilst the benign accuracy is slightly lower at 74.5% accuracy.

As little differences had been detected in accuracy between the No Augmentation and WGAN Augmentation scenario. Meanwhile, a slight yet noticeable difference had been found between the No Augmentation and Affine Augmentation scenario. To further investigate whether these results were statistically significant, paired T-test were conducted between the main augmentation scenarios.

5.3 Statistical Analysis

Using a Paired T-test we can verify whether these results are truly statistically significant with a p-value of 0.05. There are 4 important comparisons for conducting the Paired T-tests:

1. The No Augmentation results versus the WGAN Augmentation results
2. The No Augmentation results versus the Affine Augmentation results
3. The Affine Augmentation results versus WGAN Augmentation results

4. The Affine Augmentation results versus Both Affine and WGAN Augmentation results

These 4 Paired T-tests are shown.

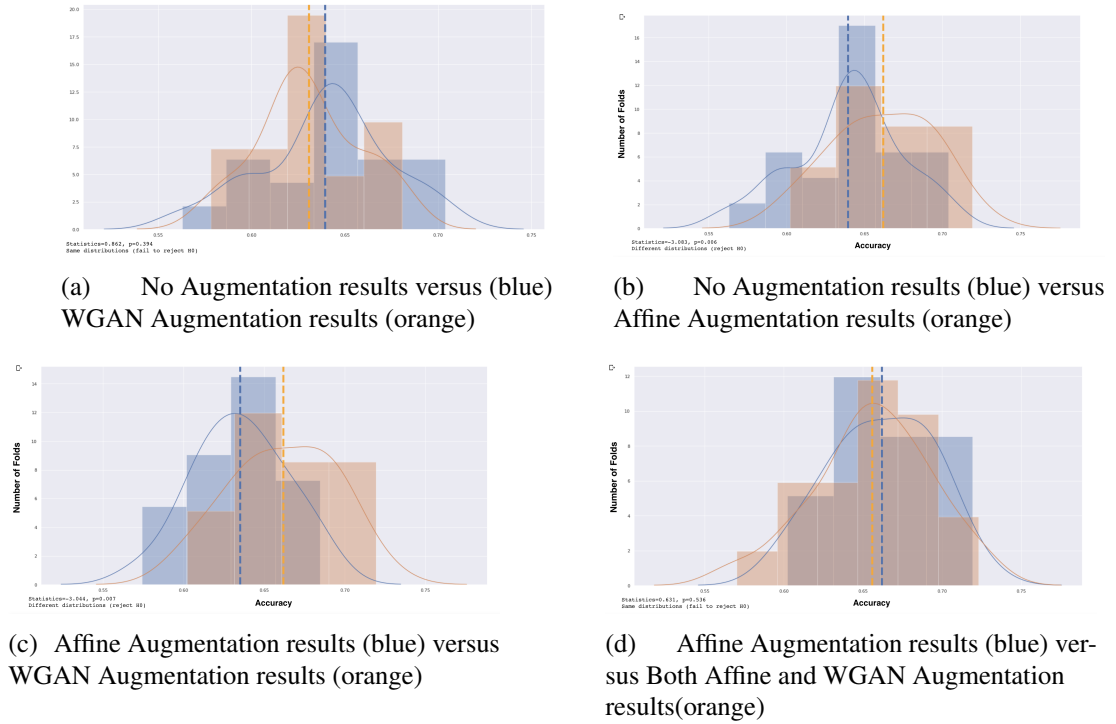


Figure 5.10: Paired T-test results between the main scenarios

As shown above we see that at a p-value of 0.05, there is no statistical difference between the No Augmentation results and the WGAN Augmentation results. The WGAN Augmentation results show a slightly smaller Standard Deviation (0.028) than the No Augmentation results (0.034), however they still show no statistical difference at the given p-value.

In turn, the Affine Augmentation did show a statically significant improvement in accuracy over No Augmentation, with an overall mean accuracy that was 2.3% higher. Whilst this is not a very large margin, it is still notable. Hence, perhaps the total achievable improvement using data augmentation is limited to the total information already within the existing training data.

The Paired T-tests conducted between the WGAN Augmentation results and the Affine Augmentation results also showed a statistical significance which is expected as there was a statistical significance with the similar No Augmentation results. This

may also be due to the relatively smaller amount of augmentation applied using the WGAN Augmentation compared to the Affine Augmentation.

Lastly, no statistical significance was achieved between the Affine Augmentation results and Both the Affine and WGAN Augmentation results suggesting that the WGAN Augmentation did not greatly affect the performance.

As the WGAN Augmentation seemed to provide no statistically significant accuracy gain from the VGG-16, it was further investigated whether it was due to the WGAN-generated lesions simply being not realistic enough. For this a Visual Turing Test with professional dermatologists was conducted to verify whether the synthetic images were flawed.

Chapter 6

Qualitative Evaluation: Visual Turing Test

The aim of this chapter is to explain the methodology and results obtained from the Visual Turing Test (VTT) conducted with professional dermatologists. The VTT was as form of qualitative evaluation for the realism of the GAN-generated images.

6.1 Methodology of VTT

To better understand the qualitative accuracy of the generated images we conducted a Visual Turing Test (VTT) with professional dermatologists. A visual Turing Test is essentially a test where a mix of synthetic and real images are given to a volunteer which is asked to correctly label these as synthetic or real. If the volunteer is unable to correctly classify these with reliable accuracy, then the Visual Turing Test can conclude that the synthetic images are fairly indistinguishable from real ones.

The purpose of the VTT was to see if the WGAN was generating skin samples that had realistic higher-level skin lesion features. The reason higher level features such as shapes and colours were tested instead of fine-grained skin texture is because the main focus of the WGAN was to generate correct higher level characteristics of a lesion. Furthermore, GANs tend to struggle to generate skin texture as this is very fine-grained and unique to each individual lesion.

The VTT involved giving dermatology professionals a survey with a random mix of synthetic WGAN-generated images and real skin images. They were then asked to classify these images as either real or synthetic to the best of their abilities. Both of these sets had a light gaussian blur applied to them to ensure the WGAN was producing the correct higher level feature rather than correct lower level details. An example of a gaussian blur applied to both the synthetic and real skin lesions can be seen below.

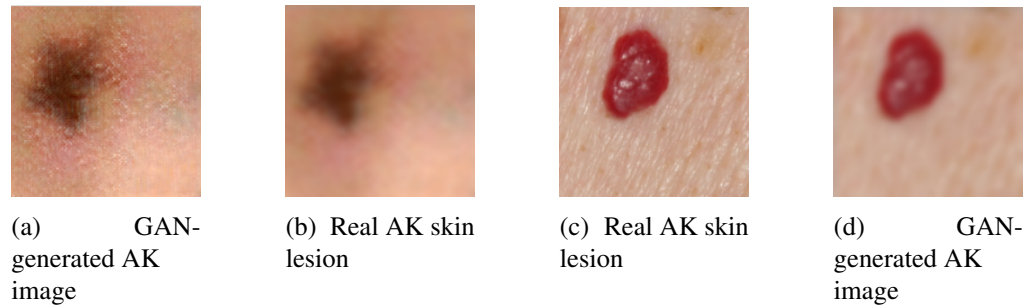





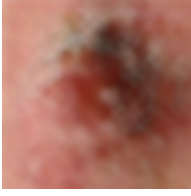








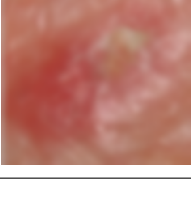

Figure 6.1: Impact of light gaussian blur to create focus on higher level details

The reason for this was that GANs tend to find it difficult to simulate very intricate textures such as that of the skin and instead would generate checkered-like skin patterns. A light gaussian filter was hence applied to simply smoothen the skin texture and reduce the effect of lower level patterns in preference for higher level features.

For the survey, 49 skin lesion images were used from the 7 skin lesions classes that were most underrepresented in the dataset and which were going to be used in augmentation. 22 of these 49 images were WGAN-generated and 27 of these were from the original training set. The split of images within the 7 classes were decided not to be a 50%-50% split but instead were mixed randomly. This is because an even image split would give the dermatologists additional unfair information on what ratio to predict their answers. The following section describes the results observed when conducting the survey. The VTT survey can be found in Appendix 8.1.

6.2 Results

The survey was answered by 3 professional dermatologists and the results shown below were analysed anonymously. An image sample from each class of both real and WGAN-generated lesions were added to the table for visual comparison.

Skin lesion class	Real Image Sample (Gaussian filter)	Synthetic Image Sam- ple (Gaussian filter)	Dermatologist Accuracy (Average)
Pyogenic Gran- ulomas (PYO)			62.50%
Squamous Cell Carci- noma (SCC)			50.00%
Melanoma (MEL)			80.95%
Intra-epidermal Squamous Cell Carcinoma (IEC)			66.67%
Vascular Lesion (VASC)			90.47%
Dermatofibra (DF)			77.78%
Actinic Ker- atosis (AK)			50.00%
Average			68.33%

6.3 Analysis

The average accuracy of the dermatologists in distinguishing between real and fake images for each class was 68.33% and the total accuracy across all images was 67.34% . Whilst this shows that for the most part, the dermatology professionals are able to distinguish correctly, it also means that roughly 1 of out of every 3 images were misclassified.

Furthermore, if we disregard real images and just focus on the accuracy on synthetic images, the average accuracy by class was 63.89% and the total accuracy across all synthetic images was 59.09%. This indicates that the WGAN-generated synthetic images were realistic enough to fool the dermatologists roughly 41% of the time. This lowered accuracy suggests that on this limited sample the WGAN generated reasonably realistic skin lesion images. The results of the 3 dermatologists classifying the 49 images are observed in the following cumulative confusion matrix:

	Real Images	Synthetic Images
Predicted Real	60	24
Predicted Synthetic	21	42

Furthermore, a large distinction to be made is that if a dermatologist deemed a synthetic image was fake, it does not necessarily mean the image does not contain the correct information of its intended skin lesion class. This is observed in the results table in the previous section where the dermatology professionals were able to classify vascular images with 90.47% accuracy, however as shown above, the synthetic vascular images still contain relevant colour tones, shapes and patterns as typical vascular lesions do.

A limitation to these results however was that a light gaussian filter was applied to both real and synthetic images. This may have hindered how realistic both images may have been portrayed. However, the purpose of the VTT was to test whether the GAN was producing the correct higher level features rather than fine-grained details like skin texture. Furthermore, Gaussian filtering is a type of data augmentation that can still be done before inserting GAN-generated images into a training set and could be applied regardless.

Overall, it appears that under a qualitative evaluation using a Virtual Turing Test, the WGAN-generated images appear to have fairly realistic higher-level features according to the results by dermatology professionals.

Chapter 7

Conclusion & Evaluation

The purpose of this chapter is to draw potential justifications for the conclusion reached from the paired T-tests and results found in Chapter 5 as well as the outcomes of the VTT from Chapter 6. This chapter also provides an evaluation of the overall methodology used in this project to answer the research question.

7.1 Analysis & Conclusion

Whilst the results found in the Chapter 5 indicate that not a statistically significant increase in accuracy ($p=0.714$) was witnessed using WGAN augmentation with a p-value of 0.05, this is still a notable result. No major differences in precision, recall or F1-score were witnessed for any individual skin lesion classes between the No Augmentation scenario and the WGAN Augmentation scenario. Meanwhile, in comparison the Affine Augmentation scenario found a small statistical significance ($p=0.006$) in accuracy improvement (+2.3%) over the No Augmentation scenario. This leads to the conclusion that the WGAN augmentation used did not provide a significant performance gain for the VGG-16 classifier. Even under a p-value of 0.20 this conclusion can still be drawn. Three possible reasons can be attributed to justify this conclusion.

Firstly, it is possible that the WGAN-generated synthetic images simply do not provide a large information gain over the existing training data images. This justification can be explained by the WGAN's design purpose being to create synthetic samples using features already present in the training data. This may also be because the Convolutional Neural Network used in the discriminator of the WGAN is quite similar to the VGG-16 in structure. Hence, if the only feedback the generator receives is from the discriminator, then this may mean the WGAN-generated images do not encode any information not already extracted by the VGG-16. Section 7.2 will show how the existing skin lesion GAN-augmentation literature also supports this conclusion.

Secondly, another possible justification is that the WGAN Augmentation used was insufficient. Evidence for this is provided by the comparatively larger amounts of

augmentation done using the affine augmentation which showed a more steady increase in accuracy over various epochs. In comparison, the total augmented images using the WGAN was doubling the size of the already small classes, whereas the affine augmentation increased the number of total images by almost 10 fold. Perhaps testing different amounts of WGAN augmentation could yield larger and more significant performance increases. Furthermore, the VTT test provided evidence that the WGAN generated fairly accurate lesion samples and hence perhaps some information gain is still present in the images.

Thirdly, a last and less likely explanation for the conclusion is that the VGG-16 simply could not extract the correct information gain from the generated synthetic WGAN images. The VTT test conducted indicated that the synthetic generated GAN images appeared to create realistic lesions and perhaps this information was not able to be processed by the VGG-16 feature extractor. This may be evidenced by the fact that the overall accuracy of the VGG-16 was still fairly low at about 63% as a baseline and hence was not extracting the maximum information gain from the training images to begin with.

Overall, the first and second justifications seems the most likely out of the three posed. This is also as Pollastri et al use a LAPGAN and DCGAN as a form of GAN skin lesion augmentation also showing a very small increases (and sometimes decreases) in accuracy for their Deep Neural Network (further discussed in the next section) [49]. Furthermore, there is evidence that the amount of augmentation created by the WGANs was perhaps insufficient compared to the potential of information gain by the WGAN. This is as the chosen images for WGAN augmentation were chosen according to a threshold of iterations that appeared stable rather than being cherry-picked for which were most realistic. Hence the methods for selecting WGAN augmentation images overall could perhaps be improved in future with more intelligent selection methods.

7.2 Comparison to Literature

As observed in Chapter 2, there is limited literature covering the topic of accuracy gain using WGAN Augmentation in skin lesion training sets. The only studies found with results on the impact of GAN-augmented training on skin lesion classification accuracy is by Pollastri et al in 2019. They compared the accuracy changes of Deep Convolutional Neural Networks with different types of either DCGAN or LAPGAN augmentation on the ISIC 2017 skin lesion dataset. Their results indicated that the LAPGAN augmentation achieved an average increase in accuracy of +0.783% for the DCGAN augmentation, +0.817% for the LAPGAN augmentation and +0.367% for a combination of both GAN augmentations [49].

These results are fairly similar to the ones observed in our project using the WGAN Augmentation on the DERMOFIT dataset (explored in Chapter 5). This is as the WGAN Augmentation created a similar accuracy to that of no augmentation which is similar to Pollastri et al's accuracy increases of less than 1% [49].

Furthermore, Pollastri et al did not conduct cross fold validation, hence it is difficult

to determine if these small increases are truly statistically significant. Moreover, they did not compare this to regular affine augmentation hence its comparative efficacy as a data augmentation technique is unclear. However, our project found that both affine and GAN augmentation had similar performance to affine augmentation. This reinforced the notion that GAN augmentation has little effect on the information gained in a dataset or accuracy increase in a classifier using that dataset.

7.3 Method Strengths

One of the most notable strengths of the methodology used in this project was that reproducible and reliable processes were favoured over higher possible reported accuracy. This is shown in the chosen training-validation-test split being 60%,20%,20% which could have been one of the reasons the reported accuracy overall was lower than Bertrand's paper. Furthermore, Bertrand describes in his Deep Neural Network classifier that the sole use of a testing and validation set can lead to a bias in the reported accuracy due to the fine tuning of hyper-parameters. Hence, the use of a validation set was in line with recommendations from past research in attempts to reduce biases in reported results.

Furthermore, in comparison to Pollastri's paper using PAGAN and DCGAN augmentation, this project explored the use of a WGAN augmentation which due to the Wasserstein distance metric has been preferable for its ability to avoid mode collapse [49] [55]. Furthermore, the use of 5-cross-fold validation compared to single runs in Pollastri's paper ensured that the results were perhaps more replicable and we were able to gauge the variability of the accuracy across folds. Lastly, using a paired T-test meant that we were also able to verify whether the gains in accuracy witnessed were statistically significant.

Aside from quantitative evaluation, a qualitative evaluation of the realism of the WGAN-generated images was also assessed. This was conducted through a Visual Turing Test (VTT) where 3 professional dermatologists were challenged to verify whether slightly blurred skin lesion images were real or synthetic. This allowed us to gauge the realism of higher level features of the lesion sample the GAN produced for the range of the different lesion classes.

7.4 Method Weaknesses

An important weakness to be identified in this project is the limitation of the test framework being used. This is as only one type of Deep Neural Network was tested, namely a VGG-16. However, many different Deep Neural Networks exist whose different structures and layers allow them to learn in distinct forms [49]. Hence, perhaps the results reported may not extend to other types of Deep Neural Networks. Therefore, improvements that could have been made in the methodology are firstly, like in Pollastri et al's research, testing out the effect of the GAN augmentation on a variety of different Deep Neural Network classifiers [49].

Moreover, transfer learning was used for the classifier as in Bertrand's deep network paper to allow the Deep Neural Network to learn transferable feature extraction capabilities before specialising on skin lesion classification [25]. The VGG-16 was pre-trained on Imagenet which is one of the most expansive visual datasets there is with many classes. However, the transferability of the feature extraction process of Imagenet to skin lesion classification is not clear. Hence, different datasets could be attempted for use in transfer learning process which are perhaps more closely related to the task of skin lesion classification.

Lastly, DERMOFIT is one of the highest standard skin lesion datasets, however because of its high quality it is also on the lower end of total images compared to for example the HAM10000 dataset with 10,000 skin lesion images [5]. Therefore, there is a trade-off witnessed here as data augmentation tends to have higher impact with less data, but also the realism and range of GAN images improves with more data [41]. Therefore, it would be interesting to further explore the effects of this trade-off by experimenting with GAN augmentation on different sized skin lesion datasets.

7.5 Potential Applications

Although, the results indicated that the GAN augmentation did not provide a statistically significant improvements in VGG-16 accuracy, the Visual Turing Test suggests potential future applications of the realistic synthetic images. Whilst in informal discussion with a medical professor on the outcomes of this project, there seemed a genuine interest for the potential of the WGAN-generated images being used in printed and educational material.

The reason for this being that medical images used for educational purposes typically require large legal consent paperwork to be signed by the original patient that the image belongs to [14]. However, it is often very difficult to track down the original patients if the image is taken from a dataset or secondary source as is typical [20]. Furthermore, as discussed in Chapter 1, attempting to gather one's own skin lesion images is an exhaustive task. Hence perhaps the use of GAN-generated images could allow a way to easily use realistic-looking skin lesion images for education purposes with a reduced worry for privacy concerns. Nonetheless, the legal implications behind using GAN-generated images commercially when they were inspired by real images from a dataset is a topic with little exploration as of yet due to the recency of this technology.

7.6 Further Study

A potential avenue for further study is firstly to create an improved process for creating masks for WGAN-augmented images as in Bissoto et al's skin lesion GAN [48]. This would be an interesting exploration as Bertrand's Deep Neural Network paper showed that masking of skin lesion images can help improve the accuracy of the classifier [25].

Moreover, as it is a difficult process to verify when a GAN is finished training, the synthetic skin lesions used for augmentation were chosen randomly from the generator output from 3500 to 7500 iterations. This was the period when the WGAN was found to be most visually stable. However, in theory images found within this set may be similar to each other due to mode collapse or simply unrealistic. An automated method to select the best images from the GAN generator output, for example by avoiding images with extreme colors (such as yellow or pink) and trying to only select images with visually different SIFT markers could be an interesting for further exploration.

Lastly, it is necessary to address the problem of the VGG-16 test framework being separate from the augmentation network of the GAN. This was a limitation because the loss function of the GAN was not directly influenced by the accuracy of the classifier, which was the main goal of this project. It may be interesting in further exploration to try to combine both the WGAN and test framework loss functions together into a single system. This way the WGAN would be directly designed to produce the images that create the most accuracy gain for the discriminator, and not just the ones that best fool the WGAN critic.

Bibliography


- [1] T. S. C. Foundation. (2020). Skin cancer facts statistics, [Online]. Available: <https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/> (visited on 04/13/2020).
- [2] M. R. Alliance. (). Melanoma survival rates, [Online]. Available: <https://www.curemelanoma.org/about-melanoma/melanoma-staging/melanoma-survival-rates/> (visited on 04/13/2020).
- [3] C.-X. Li, C.-B. Shen, K. Xue, X. Shen, Y. Jing, Z.-Y. Wang, F. Xu, R.-S. Meng, J.-B. Yu, and Y. Cui, “Artificial intelligence in dermatology: Past, present, and future”, *Chinese Medical Journal*, vol. 132, p. 1, Aug. 2019. DOI: 10.1097/CM9.0000000000000372.
- [4] D. B. Mendes and N. C. da Silva, *Skin lesions classification using convolutional neural networks in clinical images*, 2018. arXiv: 1812.02316 [cs.CV].
- [5] P. Tschandl, C. Rosendahl, and H. Kittler, “The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions”, *Scientific Data*, vol. 5, Mar. 2018. DOI: 10.1038/sdata.2018.161.
- [6] J. E. Donlan, *Ordaining Reality Made Easy: A Guide for Creating the Future*. BrownWalker Press, 2009, ISBN: 1599429101.
- [7] C. C. Australia. (). Skin cancer, [Online]. Available: <https://www.cancer.org.au/about-cancer/types-of-cancer/skin-cancer.html> (visited on 04/13/2020).
- [8] L. E. M. Dubas and A. M. Ingraffea, “Non-melanoma skin cancer”, 2013.
- [9] W. C. R. Fund. (). Skin cancer statistics, [Online]. Available: <https://www.wcrf.org/dietandcancer/cancer-trends/skin-cancer-statistics> (visited on 04/15/2020).
- [10] C. E. Lan, “Effects and interactions of increased environmental temperature and uv radiation on photoageing and photocarcinogenesis of the skin”, 2019.
- [11] F. M. Walter, H. C. Morris, E. Humphrys, P. N. Hall, A. T. Prevost, N. Burrows, L. Bradshaw, E. C. F. Wilson, P. Norris, J. Walls, M. Johnson, A. L. Kinmonth, and J. D. Emery, “Effect of adding a diagnostic aid to best practice to manage suspicious pigmented lesions in primary care: Randomised controlled trial”, *BMJ*, vol. 345, 2012. DOI: 10.1136/bmj.e4110. eprint: <https://www.bmj.com/content/345/bmj.e4110.full.pdf>. [Online]. Available: <https://www.bmj.com/content/345/bmj.e4110>.
- [12] C. X. Li, D. DN, and S. LC, “Skin cancer detection technology”, 2019.

- [13] G. Zouridakis, T. Wadhawan, N. Situ, R. Hu, X. Yuan, K. Lancaster, and C. Queen, “Melanoma and other skin lesion detection using smart handheld devices”, *Methods in molecular biology (Clifton, N.J.)*, vol. 1256, pp. 459–96, Jan. 2015. DOI: 10.1007/978-1-4939-2172-0_30.
- [14] T. Shaikhina and N. A. Khovanova, “Handling limited datasets with neural networks in medical applications: A small-data approach”, 2016.
- [15] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, “How generative adversarial networks and their variants work: An overview”, 2017.
- [16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan”, *ArXiv*, vol. abs/1912.04958, 2019.
- [17] S. Mohamed and B. Lakshminarayanan, *Learning in implicit generative models*, 2016. arXiv: 1610.03483 [stat.ML].
- [18] H. Hukkelås, R. Mester, and F. Lindseth, “Deepprivacy: A generative adversarial network for face anonymization”, *Advances in Visual Computing. ISVC 2019*, 2020. DOI: https://doi.org/10.1007/978-3-030-33720-9_44.
- [19] T. S. C. Foundation. (2020). Basal cell carcinoma overview, [Online]. Available: <https://www.skincancer.org/skin-cancer-information/basal-cell-carcinoma/> (visited on 04/13/2020).
- [20] L. Ballerini, R. B. Fisher, R. B. Aldridge, and J. Rees, “A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions”, 2013.
- [21] T. S. C. Foundation. (2020). Squamous cell carcinoma overview, [Online]. Available: <https://www.skincancer.org/skin-cancer-information/squamous-cell-carcinoma/> (visited on 04/13/2020).
- [22] NHS. (2020). Overview-skin cancer (melanoma), [Online]. Available: <https://www.nhs.uk/conditions/melanoma-skin-cancer/> (visited on 04/13/2020).
- [23] D. N. Z. Trust. (2020). Intraepidermal squamous cell carcinoma, [Online]. Available: <https://dermnetnz.org/topics/intraepidermal-squamous-cell-carcinoma/> (visited on 04/13/2020).
- [24] M. Clinic. (2020). Actinic keratosis, [Online]. Available: <https://www.mayoclinic.org/diseases-conditions/actinic-keratosis/symptoms-causes/syc-20354969> (visited on 04/13/2020).
- [25] A. Bertrand and R. B. Fisher, “Classification of skin lesions images using deep nets- intern report”, 2018.
- [26] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, *Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble*, 2017. arXiv: 1703.03108 [cs.CV].
- [27] A. Kwasigroch, A. Mikołajczyk, and M. Grochowski, “Deep neural networks approach to skin lesions classification — a comparative analysis”, in *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*, 2017, pp. 1069–1074.
- [28] N. Malik, “Artificial neural networks and their application”, 2005.
- [29] E. Grossi and M. Buscema, “Introduction to artificial neural networks”, 2008.

- [30] R. Bartzatt, “Determination of dermal permeability coefficient (kp) by utilizing multiple descriptors in artificial neural network analysis and multiple regression analysis”, *Journal of Scientific Research and Reports*, vol. 3, pp. 2884–2899, Jan. 2014. DOI: 10.9734/JSRR/2014/13125.
- [31] M. Buscema, “Back propagation neural networks”, 1998.
- [32] J. Schmidhuber, “Deep learning in neural networks: An overview”, 2014.
- [33] H. W. Lin, M. Tegmark, and D. Rolnick, “Why does deep and cheap learning work so well?”, 2016.
- [34] W. Zou, L. Li, and A. Tang, “Effects of the number of hidden nodes used in a structured-based neural network on the reliability of image classification”, 2009.
- [35] J. Koushik, “Understanding convolutional neural networks”, 2016.
- [36] A. Montserrat, Alvarado-Gonzalez, G. Fuentes-Pineda, and J. Cervantes-Ojeda, “A few filters are enough: Convolutional neural network for p300 detection”, 2019.
- [37] S. Sarwar, P. Panda, and K. Roy, “Gabor filter assisted energy efficient fast learning convolutional neural networks”, May 2017.
- [38] C. Di Leo, V. Bevilacqua, L. Ballerini, R. Fisher, B. Aldridge, and J. Rees, “Hierarchical classification of ten skin lesion classes”, Mar. 2015. DOI: 10.13140/RG.2.1.5178.8323.
- [39] R. Fisher, J. Rees, and A. Bertrand, “Classification of ten skin lesion classes: Hierarchical knn versus deep net”, in. Jan. 2020, pp. 86–98, ISBN: 978-3-030-39342-7. DOI: 10.1007/978-3-030-39343-4_8.
- [40] M. Berger, *Geometry I*. Springer, 1987, ISBN: 3540116583.
- [41] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning”, 2017.
- [42] B. A. Research. (). Why should you care about data augmentation?, [Online]. Available: https://bair.berkeley.edu/blog/2019/06/07/data_aug/ (visited on 04/13/2020).
- [43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].
- [44] Y. Liu, Q. Zhao, and C. Jiang, “Conditional image generation using feature-matching gan”, Oct. 2017, pp. 1–5. DOI: 10.1109/CISP-BMEI.2017.8302049.
- [45] Kim, Cho, and Chang, “Tooth segmentation of 3d scan data using generative adversarial networks”, *Applied Sciences*, vol. 10, p. 490, Jan. 2020. DOI: 10.3390/app10020490.
- [46] I. Goodfellow, *Nips 2016 tutorial: Generative adversarial networks*, 2016. arXiv: 1701.00160 [cs.LG].
- [47] C. Baur, S. Albarqouni, and N. Navab, “Melanogans: High resolution skin lesion synthesis with gans”, *ArXiv*, vol. abs/1804.04338, 2018.
- [48] A. Bissoto, F. Perez, E. Valle, and S. Avila, “Skin lesion synthesis with generative adversarial networks”, *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pp. 294–302, 2018, ISSN: 1611-3349. DOI:

- 10.1007/978-3-030-01201-4_32. [Online]. Available:
http://dx.doi.org/10.1007/978-3-030-01201-4_32.
- [49] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, “Augmenting data with gans to segment melanoma skin lesions”, *Multimedia Tools and Applications*, pp. 1–18, 2019.
 - [50] T. Edison, *The Diary and Sundry Observations of Thomas Alva Edison*. Greenwood Press, 1968, ISBN: 9780837100678.
 - [51] W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, A. Tejani, J. Totz, C. Ledig, and Z. Wang, “Is the deconvolution layer the same as a convolutional layer?”, 2016.
 - [52] J. Feng, X. He, Q. Teng, C. Ren, H. Chen, and Y. Li, “Reconstruction of porous media from extremely limited information using conditional generative adversarial networks”, *Physical Review E*, vol. 100, Sep. 2019. DOI: 10.1103/PhysRevE.100.033308.
 - [53] Y. Hong, “Comparison of generative adversarial networks architectures which reduce mode collapse”, 2019.
 - [54] J. Pardo, L. Pardo, and M. Pardo, “The jensen-shannon divergence”, *Journal of The Franklin Institute-engineering and Applied Mathematics - J FRANKLIN INST-ENG APPL MATH*, vol. 334, pp. 307–318, Mar. 1997. DOI: 10.1016/S0016-0032(96)00063-4.
 - [55] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan”, in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 214–223.
 - [56] J. Kun. (). Earthmover distance, [Online]. Available:
<https://jeremykun.com/tag/wasserstein-metric/> (visited on 04/13/2020).
 - [57] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved training of wasserstein gans*, 2017. arXiv: 1704.00028 [cs.LG].
 - [58] M. Mirza and S. Osindero, “Conditional generative adversarial nets”, 2014.
 - [59] S. Barnett, *Convergence problems with generative adversarial networks (gans)*, Jun. 2018.
 - [60] F. P. Such, A. Rawal, J. Lehman, K. O. Stanley, and J. Clune, “Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data”, 2019.
 - [61] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, *A comprehensive survey on transfer learning*, 2019. arXiv: 1911.02685 [cs.LG].

MELANOMA LESION #2




Is the above MELANOMA lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

MELANOMA LESION #3




Is the above MELANOMA lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

MELANOMA LESION #4




Is the above MELANOMA lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

MELANOMA LESION #5



Is the above DERMATOFIBROMA lesion real or fake (AI-generated)? *

☐ Real


☐ Fake (AI-generated)

Is the above MELANOMA lesion #5 real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

MELANOMA LESION #6




Is the above MELANOMA lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

MELANOMA LESION #7



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

Generate

MELANOMA (MEL)

NOTE: For testing reasons, the number of fake and real skin lesion images are not necessarily the same.

INTRA-EPIDERMAL SQUAMOUS CELL CARCINOMA (IEC)

NOTE: For testing reasons, the number of fake and real skin lesion images are not necessarily the same.

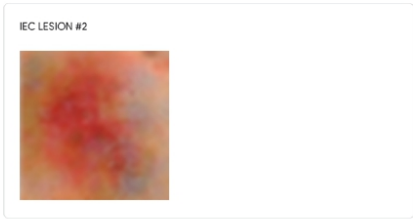
IEC LESION #1



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

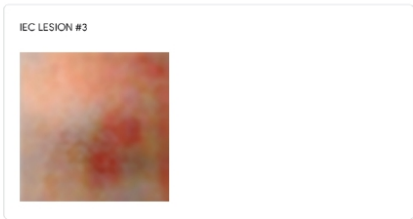
☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

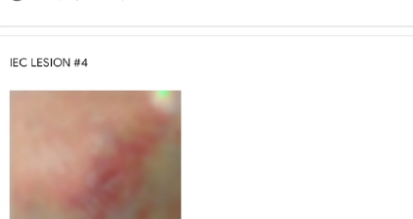
☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

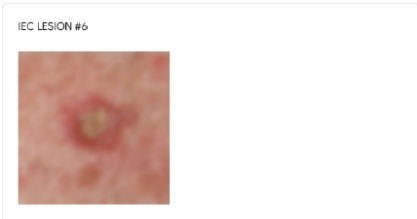
☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

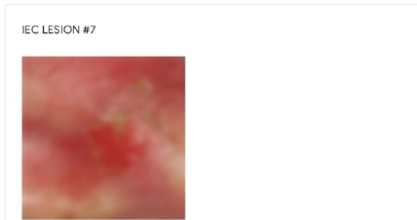
☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

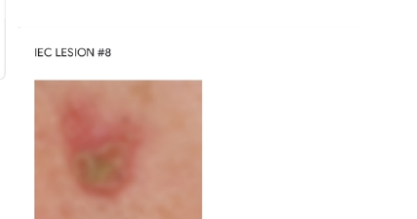
☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)



Is the above IEC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

PYOGENIC GRANULOMAS (PYO)

NOTE: For testing reasons, the number of fake and real skin lesion images are not necessarily the same.

PYO LESION #1



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

PYO LESION #2



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

PYO LESION #3



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

PYO LESION #4



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

PYO LESION #5



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

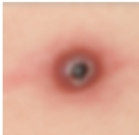
PYO LESION #6



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

PYO LESION #7



Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

PYO LESION #8



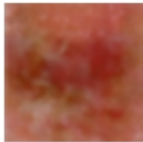
Is the above PYO lesion real or fake (AI-generated)? *

- ☐ Real
- ☒ Fake (AI-generated)

ACTINIC KERATOSIS (AK)

NOTE: For testing reasons, the number of fake and real skin lesion images are not necessarily the same.

AK LESION #1




Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #2

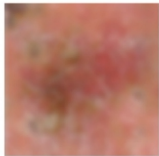


Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #3




Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #4

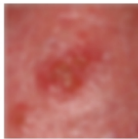


Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #5

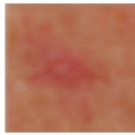


Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #6

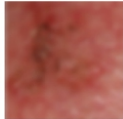


Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #7

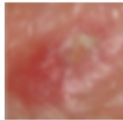


Is the above AK lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

AK LESION #8



Is the above AK lesion real or fake (AI-generated)? *

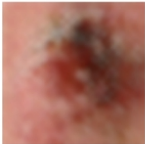
☐ Real

☐ Fake (AI-generated)

SQUAMOUS CELL CARCINOMA (SCC)

NOTE: For testing reasons, the number of fake and real skin lesion images are not necessarily the same.

SCC LESION #1




Is the above SCC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

SCC LESION #2

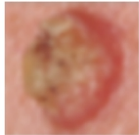


Is the above SCC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

SCC LESION #3

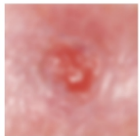


Is the above SCC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

SCC LESION #4



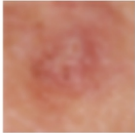
Is the above SCC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

1004 Paul Dupont Borey - Act for 6/8/2019: 4/4 Images real (generated by AI)

SCC LESION #5




Is the above SCC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

SCC LESION #6



Is the above SCC lesion real or fake (AI-generated)? *

☐ Real

☐ Fake (AI-generated)

Any other comments you would like to add about the quality of the above images? Or how easy/difficult it was to classify them?

Your answer