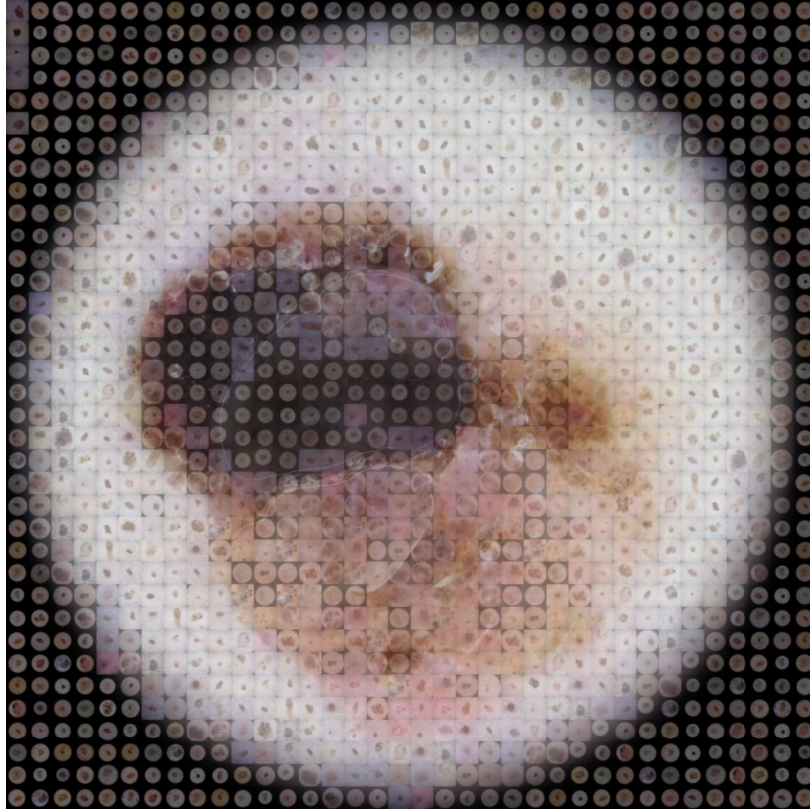




CHALMERS
UNIVERSITY OF TECHNOLOGY



Chalmers University of Technology

Generative Modeling for Melanoma Detection

Master's thesis in Complex Adaptive Systems

ANNA ROSÉN
MOHAMAD KHIR ZOUBI

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022

www.chalmers.se

MASTER'S THESIS 2022

Generative Modeling for Melanoma Detection

Quantitatively and qualitatively evaluating skin cancer images
synthesized using deep learning

ANNA ROSÉN
MOHAMAD KHIR ZOUBI



Department of Physics
Sahlgrenska AI Competence Center
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022

Generative Modeling for Melanoma Detection
Quantitatively and qualitatively evaluating skin cancer images synthesized using
deep learning
ANNA ROSÉN & MOHAMAD KHIR ZOUBI

© ANNA ROSÉN
MOHAMAD KHIR ZOUBI, 2022.

Supervisors: Sandra Carrasco, Sylwia Majchrowska, Juulia Suvilehto, and Lisa
Sjöblom, Sahlgrenska AI Competence Center
Examiner: Mats Granath, Department of Physics

Master's Thesis 2022
Department of Physics
Sahlgrenska AI Competence Center
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A mosaic displaying a malignant melanoma comprised of synthetic images
of lesions.

Typeset in L^AT_EX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Acknowledgments

We would like to acknowledge the people that made the completion and success of this project possible. We want to extend a special thank you to our supervisors Sylwia Majchrowska and Sandra Carrasco who guided us through the technical areas of the project. We also wish to show our appreciation for AI Sweden for supplying us with a space to work and the computational power to carry out our research. We would like to thank our supervisors from Sahlgrenska Dimitra Lappa, Lisa Sjöblom, and Juulia Suvilehto and our examiner Mats Granath at Chalmers who helped us finalize the thesis. Lastly we would like to extend our gratitude towards our families who have supported us through all of our years of school.

Anna Rosén & Mohamad Khir Zoubi
Gothenburg, May 2022

Abstract

Early detection significantly reduces deaths associated with melanoma, a skin cancer. Despite this information, 80% of skin cancer related deaths are attributed to malignant melanoma. Melanomas are difficult to diagnose by a dermatologist (skin doctor) therefore many patients undergo unnecessary surgeries to get a biopsy that can confirm the disease. Minimizing unnecessary surgeries would leave more resources that in turn could lead to a higher frequency of earlier diagnosed melanomas. Machine learning algorithms have shown a great potential in the field of medicine and could be deployed to help doctors diagnose melanomas. To obtain a high performing model it is crucial to have a large and balanced dataset. The scarcity of labeled publicly available medical images makes applying machine learning an obstacle in this field, thus hindering development. A solution to this problem could be to synthesize realistic looking images using deep neural networks. One such network is Generative Adversarial Network (GAN), which has been shown successful in producing images in the field of medicine. This thesis explores the generation of synthetic image data for medical purposes and how such data can be evaluated. We utilize StyleGAN2-ADA to generate synthetic images of melanoma lesions that we evaluate using both qualitative and quantitative measures. A survey was made to establish if experts can identify generated images in a mixed dataset. The expert dermatologists found the images difficult to distinguish from real ones, accordingly proving that we can synthesize realistically looking images of melanomas. Using a classifier trained on synthetic melanoma and non-melanoma images we are also able to reach a high accuracy when validating against real data. Our results show that synthetic images are verifiably realistic looking. From our research we are able to conclude that synthetic data can be the answer to further development of classification algorithms in a clinical setting.

Contents

1	Introduction	1
1.1	Thesis Background	1
1.1.1	Medical Background	2
1.1.2	Technical Background	2
1.2	Organization Background	3
1.2.1	Related Work	4
1.3	Aim	5
1.4	Scope	5
2	Theory	7
2.1	Artificial Neural Networks	7
2.1.1	Feed Forward Neural Networks	8
2.1.2	Convolutional Neural Networks	8
2.2	Generative Modeling	9
2.2.1	Generative Adversarial Networks	10
2.2.1.1	GAN types	11
2.2.2	GAN Extensions	11
2.2.2.1	StyleGAN	11

2.2.2.2	StyleGAN2-ADA	12
2.3	GAN Metrics	13
2.3.1	Fréchet Inception Distance	13
2.3.2	Perceptual Path Length	13
2.4	Classification Metrics	14
2.4.1	Confusion Matrix	14
2.4.2	Other Classification Metrics	15
3	Materials & Methods	17
3.1	Data	17
3.1.1	Datasets	17
3.1.2	Bias	19
3.2	Data Preprocessing	20
3.2.1	Bias Removal	20
3.2.2	Image Resizing	21
3.3	StyleGAN Training and Experimentation	21
3.3.1	Computational Power	22
3.3.2	Hyperparameters	22
3.3.3	Manipulating the Latent Input	22
3.3.3.1	Latent Vector Manipulation Using a Binary Classifier	23
3.3.3.2	Latent Vector Manipulation Using PCA	23
3.3.3.3	Latent Vector Manipulation Using Semantic Factor- ization (SeFa)	24
3.4	Synthetic Data Evaluation	24
3.4.1	Visual Evaluation - Fool the Doctor	24

3.4.2	Classification Model	26
4	Results & Discussion	29
4.1	Data Preprocessing	29
4.2	StyleGAN2-ADA Experimentation	32
4.2.1	StyleGAN2-ADA Training Results	32
4.2.1.1	Training Results on ISIC Dataset	32
4.2.1.2	Training Results on SUH Dataset	34
4.2.2	Manipulating the Generator’s Input for Bias Removal	35
4.2.2.1	Image Editing Using a SVM	35
4.2.2.2	Image Editing Using PCA	36
4.2.2.3	Image Editing Using SeFa	37
4.3	Synthetic Data Evaluation	38
4.3.1	Survey Evaluation	38
4.3.2	Classification Model	43
5	Conclusion	47
6	Appendix	53
6.1	StyleGAN2-ADA Network Architectures	53
6.1.1	Generator	53
6.1.2	Discriminator	55
6.2	Survey	56
6.2.1	Survey Layout	56

List of Figures

2.1	Feed forward neural network. The connection strength between the nodes is given by a weight matrix W for two each connected layers. The calculated error of the difference between the output of the network and the real output is used to update the weights of the network using backpropagation [1].	8
2.2	A schematic of a CNN architecture. Note that convolution layers contains the feature maps that aim to extract the most prominent features in an input image. Pooling layers reduce the dimensionality of the feature maps for faster computations. The fully connected layer learns the latent representation that the repeated convolutional operations produce [2].	9
2.3	General scheme of a generative model. The generative model (yellow) is a function that adjusts the parameters θ as a way to map the input Gaussian vectors (red) to images that matches the true data distribution [3].	10
2.4	StyleGAN generator unlike traditional generators that map a Gaussian distributed latent input to an image, StyleGAN generator learns the input to the generator. The elements from the resulting vectors from the \mathcal{W} -space are each supposed to correspond to a feature in the output image. This makes it easier to edit the image through changing the input \mathbf{w} -vector [4].	12
2.5	A demonstration of the confusion matrix. The abbreviations represent; true positive (TP), true negative (TN), false positive (FP), and false negative (FN).	14
3.1	These graphics represent the general work flow of the project.	17

3.2	(a) Examples of malignant melanoma in the SIIM-ISIC dataset. (b) Images of non-melanoma cases from the SIIM-ISIC dataset. (c) Visual of images with both labels (invasive and in-situ) from the Sahlgrenska dataset.	18
3.3	(a) class distribution in the SIIM-ISIC dataset. The total dataset has 37648 images of various skin lesions where 5106 of them are malignant melanomas. Note that the used dataset contained extra images of melanoma from external datasets as well. (b) distribution of classes in the smaller Sahlgrenska dataset with 632 invasive melanomas and 683 in-situ melanomas.	19
3.4	Example images with the common biases in the datasets.	19
3.5	Distribution of biases in the malignant melanoma cases in the SIIM-ISIC dataset. Clarification of the biases; hair is in regard to any body hair and black frames are from the dermascope lens.	20
3.6	Algorithmic chart showcasing the approach with bias removal using a binary classifier. Note that the used vectors here are ones sampled from the learned \mathcal{W} space.	23
3.7	This image shows projected synthetic images that have been classified by a pretrained classifier that was trained on real images. The cluster on the left are the benign melanomas and the cluster to the right in the figure are malignant melanomas. The high dimensional data was clustered and represented in 3D using PCA.	25
4.1	This graphic illustrates how the images were cropped to remove all the frames from the dermascope.	30
4.2	shows 3 of the different preprocessing techniques we tried to remove the dermoscopic frames. The top row shows the original photos from the dataset. The following rows below show the removal techniques applied on the same images.	31
4.3	Example showing images before and after algorithm described in section 3.2.1 was applied.	31
4.4	Example of synthetic melanoma images generated through unconditional GAN (melGAN). The training for this model was resumed from a benign skin lesion generating model.	33

4.5	Example of synthetic non-melanoma skin lesion images generated through unconditional GAN (benGAN). The training for this model was started from scratch as the training set used was significantly larger than the melanoma skin lesion set.	33
4.6	Example of synthetic skin lesions generated using the conditional GAN model (cGAN). Here, we see the per image quality for both benign skin lesions (top row) as well as for melanoma (bottom row). .	34
4.7	Synthetic skin lesions generated using a conditional GAN model. Note how some hair patterns are repeated among the the images. This was due to the small number of images the model was trained on. The dataset used here was from SUH and had a sample size of 1.3k images.	35
4.8	(a) An illustration of moving an image perpendicularly towards the decision boundary of a trained SVM. Note that while the frame is gradually removed from the image, the melanoma gradually changes its form as well, making it hard to know if the resulting frameless image remains a melanoma. (b) Simplified illustration of how the binary classifier separate images with labeled biases. Here, the circle represents the learned style space \mathcal{W} and each image correspond to a latent input \mathbf{w} -vector.	36
4.9	Results from shifting the latent vector \mathbf{w} along the 3 first principal directions, along with the unwanted frame bias being removed. The center image is the original, the images to the right and left represent opposite and constant directions. Other features in the image, such as the mole's shape and size change as well, and now we begin to question if this image still represents a melanoma.	37
4.10	Examples of image editing using the SeFa framework. The image in the middle in all three rows is the original image. The rows correspond to the \mathbf{w} -vector being moved along different eigenvectors. Left and right of the original image are positive and negative eigenvectors. . .	38
4.11	Ability to distinguish real images from fake images. (a) Metrics visualization of the rates for the individual participants in the survey. The abbreviations here are; true positive rate (TPR), false positive rate (FPR), and accuracy (ACC). (b) Showing the number of correct guesses per synthetic image.	40
4.12	This figure displays all images that were incorrectly classified as false negatives by all survey participants.	40

4.13	The images in this figure were all the images that the participants correctly classified as true positives.	40
4.14	(a) displays 9 out of 14 images that were classified as true negatives by all survey participants. (b) displays all images that participants classified as false positives.	42
4.15	Participants average certainty per each image. The certainty is in regards to how sure they were that they labeled the image correctly (synthetic or not).	42
4.16	(a) is a confusion matrix of the diagnosis answers to the survey where they choose from malignant melanoma and not melanoma. These images are the dermatologist diagnosis of the real patient data. (b) displays a confusion matrix of the diagnosis of the synthetic data for the two dermatologists.	43
4.17	ROC curves comparing the sensitivity (true positive rate) to the false positive rate at varying thresholds for each of the different classifiers.	44

List of Tables

3.1	Training setups for different data processing methods.	27
4.1	lists the separate models trained and their abbreviated name. Training set refers to the total number of real images that was used to train each GAN. The type of GAN that was used to train them is also specified; unconditional (uncond.) and conditional (cond.). . . .	32
4.2	FID and PPL metrics derived from the trained models using SIIM-ISIC dataset.	33
4.3	Results from survey evaluation of synthetic data. Each participant's confusion matrix answers are listed in this table.	39
4.4	Survey related metrics. Listed is the accuracy, sensitivity, specificity, and precision in relation to each participants answers.	39
4.5	Test results from trained classifiers. The first 4 rows are the results from the confusion matrix. ACC is the test accuracy of the classifiers. True positive rate or sensitivity is the TPR and TNR is the specificity. AUC is the area under the ROC curve.	45

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ACC	Accuracy
ADA	Adaptive Instance Normalization
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CNN	Convolutional Neural Network
DCGAN	Deep Convolutional Generative Adversarial Network
DL	Deep Learning
FFNN	Feed Forward Neural Network
FID	Fréchet Inception Distance
FN	False Negative
FP	False Positive
FSR	Frequency Selective Reconstruction
GAN	Generative Adversarial Network
ML	Machine Learning
PPL	Preceptual Path Length
ROC	Receiver Operating Characteristics
SeFa	Semantic Factorization
SUH	Sahlgrenska University Hospital
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate

1

Introduction

Melanoma skin cancer although an infrequent disease makes up about 80% of deaths from skin cancer [5]. Linked to the fair skin of the population and not enough use of skin protection this is a crucial issue in Sweden. Most skin lesions are diagnosed through subjective visual examination, which is time consuming, often require second opinions, and often lead to misdiagnosis. Waiting time at clinics are often long, which in turn can lead to prolonged disease. Survival rates of patients can be improved if detection and diagnosis can be reached earlier. This is an opportunity where computer power could be employed to reduce the time from first appointment to diagnosis along with improved accuracy.

Computer algorithms that use data to get better at classification tasks have seen big improvements in recent years. With the release of powerful computer hardware, computationally heavy tasks in image analysis can now be performed at higher speeds. In this project our main focus is to implement a generative adversarial network to generate images of melanoma. Having a large and balanced dataset is crucial when developing a good classification algorithm.

In the following sections a brief background on the organization and a general description of the thesis work will be discussed.

1.1 Thesis Background

There are two aspects of the project that are important to understand. The medical background will introduce the general markers of the disease and why this project can prove useful. The other is the technical background that is a comprehensive summary of the systems used in this project.

1.1.1 Medical Background

As mentioned previously malignant melanoma is a skin cancer and the prevalence of the disease is steadily increasing. Most commonly the disease appears in an already existing birthmark where the symptoms conventionally are changes in the birthmark's appearance, in particular, color, shape, size, and/or bleeding [6]. The only way to certainly establish if a skin lesion is melanoma is through a biopsy. A dermatologist will investigate the mole using a dermascope, which can be thought of as a magnifying glass with a light that also is capable of taking images. The dermatologist will remove the mole if there is any suspicion that it is malignant melanoma. There are also different types of melanoma where a surface level melanoma is called a *in-situ* melanoma and a melanoma that grows on the depth is an *invasive* melanoma [7]. An invasive melanoma can get contact with lymph-system and veins which sequentially can lead to the cancer metastasizing in other areas of the body.

Today melanoma account for 6.5% of cancer diagnosis and the 10 year survival rate is 86% for men and 92.6% for women [6]. However the prognosis looks better when patients are treated early when the tumor is in-situ. Melanoma appears on skin that has been damaged through burns by the sun's UV radiation [5]. Therefore the tendency to burn in the sun is a risk factor, which is why the disease is prevalent in Sweden with its large population of fair skin people. The sun damage intensifies with time as well as the sun's power. The melanocytes, the cells that makes your skin tan is also what gives you melanoma when they turn cancerous. The best way to avoid sun damage is through wearing clothes, sunscreen, or staying away from the sunshine.

Biopsy is the only way to determine if a mole is melanoma, because of this many people undergo unnecessary surgeries since in medicine it is better to be safe than sorry. Because of this there are many resources that are used needlessly in terms of dermatologists' time and physical resources. Consequently the unnecessary use of resources can lead to long waiting times. This leaves an opportunity to introduce *artificial intelligence* (AI) in effort to improve the speed and accuracy of diagnosis, the topic that we will explore in this thesis.

1.1.2 Technical Background

Machine learning (ML) is a part of artificial intelligence. Putting it simply, ML is composed of algorithms whose performance improves with data. A sub branch of ML is *deep learning* (DL), which uses *artificial neural networks* (ANNs) with multiple processing layers to learn a function through training on a given dataset [8]. The learned function can then be used to annotate new unconnected data that are of similar characteristics.

Image classification with machine learning is not hard to achieve with current knowl-

edge and technology. A simple example would be the classification of images of dogs and cats, where a *convolutional neural network* (CNN) is able to classify images into one of the two categories with an accuracy of 97% [9], when trained on images of cats and dogs. In the medical field ML have found its purposes in tasks where a trained neural network can be used to reduce the workload on healthcare workers or as a second opinion for a diagnosis for a doctor [10]. This could also be applied to the images taken with the dermoscope to assist in diagnosing melanomas. One of the challenges of training ML models in the medical field is that one is dealing with patient data and it is difficult to come across large sets of data. It is a problem within the medical image category that many datasets are small. Classifiers trained on a small dataset can lead to increased uncertainty in the model, as the model may be unable to generalize to new data. One proposed approach to solve this problem is to use ML to generate new data of similar distribution to the target data as a way to expand datasets. These types of models that are trained to create new data are called generative models.

In recent years different approaches to tackle generative modeling have been published such as variational autoencoders, Boltzmann machines, etc. Among the proposed methods, *generative adversarial network* (GAN) was the most promising, The GAN framework was introduced by Ian Goodfellow in 2014 [11]. Generative adversarial networks are based on two networks. The first is a *discriminator*, which is trained as a classifier to differentiate real data from fake data. The second is a *generator* that trains to trick the discriminator by generating data that is too hard to classify as fake. After getting enough feedback from the discriminator the generator will be able to generate data that is good enough to trick the discriminator [12]. Hence this synthetic data will be good enough to substitute for real data.

Another issue that presents itself in generators trained on images is that anything present in the dataset will likely be regenerated in the synthetic images. In view of that fact it is reasonable to work with a diverse dataset in order to avoid having *biases* regenerated. Biases are unwanted artifacts (noise) that may exist in a dataset. Such artifacts may be a determining factor for the classifier's ability to distinguish between classes in a dataset.

1.2 Organization Background

Sahlgrenska is located in Gothenburg Sweden and is the largest hospital in Sweden [13]. The hospital functions both as the University Hospital (SUH) as well as an academic side. Sahlgrenska is where most research into life-sciences in the Nordic countries takes place. The AI Competence Center was started at Sahlgrenska in September of 2021, an investment that will make it possible for SU to increase their competencies within AI. Their vision is to fully utilize artificial intelligence in clinical work, research, education, development, and innovation. The belief is that an increased knowledge and research in the AI fields could greatly benefit the

patients.

1.2.1 Related Work

Training a good classification model requires a large and balanced dataset, these requirements do not exist in every dataset. In the healthcare sector, acquiring data is generally difficult. One has to go through long legal procedures to get the permission needed to use patient data in training an algorithm [14]. One way to deal with this issue is to use generative modeling to expand available training datasets. Several studies have been conducted on using generative modeling for such this purpose.

Zihwie Q et al [15] implemented a modified version of StyleGAN architecture to generate synthesized images of melanoma. The changes were made to; *(i)* handle the scarcity of the available dataset and *(ii)* reduce the style variation of the generated images since skin lesions have significantly less features than complex images that StyleGANs were built for. StyleGAN was primarily built to be used for images of faces [4] and similar feature rich motifs. The authors stated that their modified GAN architecture achieved the best quantitative metrics when compared to other GAN frameworks. A classifier was then trained on generated images as a measure of quality. Using synthetic images the trained classifier showed 1.6% improved accuracy compared to one trained on real images. This motivates the use of the network architecture StyleGAN for image generation.

Bauer et al [16] did a comparative study between 3 GAN frameworks where the goal was to generate high quality melanoma images and validate the images qualitatively on expert dermatologists. The images used had a resolution of 256×256 . The study concluded that the synthetic samples generated using progressive GANs were highly realistic looking. The synthetic images however were difficult for dermatologists to distinguish from real ones. In addition, Gonçalves [17] did a comparative study on data augmentation techniques for image classification. One of the conclusions they made was that training a classifier on synthetic data generated using the StyleGAN2-ADA model yield a 2.1% accuracy improvement. This percentage improvement was in comparison to a CNN model that did not incorporate augmented data in the training set.

1.3 Aim

The aim of the thesis is to investigate the problems mentioned in the background:

- Using GANs, can we generate images that are realistic enough to deceive dermatologists into thinking they are real?
- Can we train a CNN on generated data to classify different images into malignant melanoma vs not melanoma?

In this project we intend to work with a conditional GAN model and use it to generate new images with corresponding labels from a dataset in the context of dermatology. The datasets that will be used contain images of melanomas from patient cases. We will tune the hyperparameters, preprocessing, and play with network configuration to optimize the artificial image generation. We strive for the generated images to be good enough to deceive a dermatologist into believing that the images are from real patients. The quality of the generated data will be scrutinized through training a CNN on synthetic data and validating the trained model on real images with two classes (*malignant melanoma* and *not-melanoma*). Achieved performance will be compared to results obtained with model trained only on real cases. Our hope is that we will achieve comparable accuracy of classification into the two classes. We will try to explore characteristics of generated data and look into possible methods for bias removal from generated images as well as real images.

1.4 Scope

The main focus of this thesis is to train a GAN model that can generate realistic looking images of melanoma. Additionally, we evaluate the validity of generated images by showing them to experts in the fields of both dermatology and AI.

Building a model from scratch can be time consuming since there are many parameters to be taken into account to increase the model's performance. StyleGAN2-ADA is the framework the we will utilize for our study. This have shown to generate diverse and high-quality images of melanoma, while being trained on several thousand images [15]. Although this model is capable of generating high resolution images, the project will mainly look into generating images with a 256×256 resolution.

2

Theory

This section of the thesis focuses on the theoretical framework that the project was built on. The chapter is divided as follows: A short recap on what artificial neural networks are, the related theory on how ANN can be applied for image classification. An introduction to generative modeling. A description of how GANs work, as well as to different GAN extensions. In addition, some metrics on how results from GANs can be quantified and how they work will be discussed. The theory related to how the output of trained generative models can be manipulated will be discussed as well.

2.1 Artificial Neural Networks

Using biologically inspired algorithms to solve computer problems have shown to yield near-optimal solutions when applied to large-scale complex problems [18]. One such biological system is the brain, the center of thinking for mammals. The brain consists of billions of neurons that receives input from connected neurons that forward the information throughout the body. This is what takes place when we interact with the surrounding environment. ANNs are designed as a way for computers to mimic the process of how neurons in the brain interact. In ANNs, *nodes* have the function of the neurons in a biological brain. Nodes can be thought of as light bulbs that are represented by an *activation function* σ that takes in an input signal (numerical value) and in turn switches on (activates) the node when its value is higher than some threshold. The value of the threshold depends on how the activation function is defined. ANNs understands the data it is fed by changing the strength of the connections between the nodes in accordance to what gives a correct interpretation of the input data. A simple example of an ANN is a *feed forward neural network* (FFNN).

2.1.1 Feed Forward Neural Networks

Putting it simply FFNNs work as a black box that maps an input vector \vec{x} to its corresponding output vector \vec{y} by learning an arbitrary function f . The connection strength between the nodes are determined by the the weight matrix W and the value that controls when the node activates is determined by the *bias* \vec{b} , as seen in Figure 2.1. In detail this network type consists of layers of nodes that are connected sequentially. Each layer consists of nodes that get their value from the nodes in the previous layer. The nodes \vec{h} in the network obtain their value via a pre-specified activation function σ in the form of $\sigma(W^T \vec{h} + \vec{b})$.

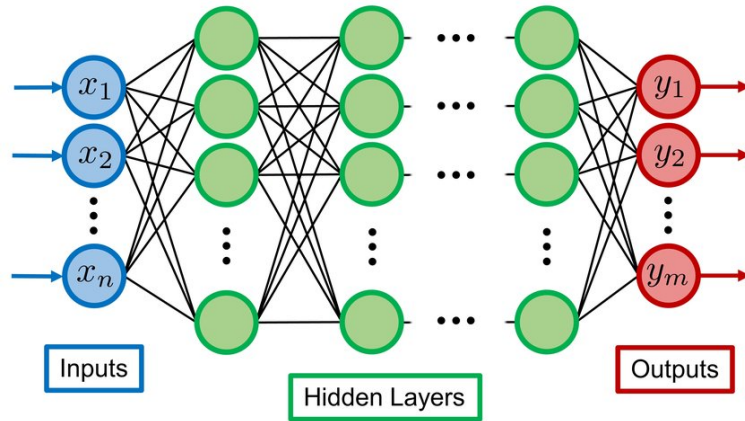


Figure 2.1: Feed forward neural network. The connection strength between the nodes is given by a weight matrix W for two each connected layers. The calculated error of the difference between the output of the network and the real output is used to update the weights of the network using backpropagation [1].

The network learns to map given inputs to desired outputs through adjusting the weights and biases of the network's node connections. This is commonly done using *backpropagation*; an algorithm that adjusts the value of the weights and biases based on the difference between what the network outputs and the real output. This difference can be modeled by using some error function, which has to be minimized in order to obtain a correct mapping function.

FFNNs can be used for different applications such as regression, classification or prediction problems. However, they mainly perform poorly in tasks where the input data are grid-shaped. To this end one make use of CNNs.

2.1.2 Convolutional Neural Networks

CNNs work well with position dependent input like time series, images or graphs. This variant of networks mainly rely on a set of operations with the purpose of extracting features and downsample an input image. These operations are done

repeatedly until we obtain an implicit (*latent*) representation of the input image. The representation obtained is in the form of a vector that is later fed into a FFNN, which learns the function that correctly maps the input image to the correct output, as seen in Figure 2.2. Image processing is done by using convolution operations of a number of filters over an image with the aim to extract the most representative features of that image. A filter is passed over the image obtaining a *feature map*; an image that highlights the parts where a target feature exists in an input image. The convolution process is also useful for upsampling an image in a process called *transposed convolution*. Furthermore, *pooling* layers are a part of the network that works as a feature extraction tool as well as for dimensionality reduction. There are no weights or biases accompanied by using pooling layers.

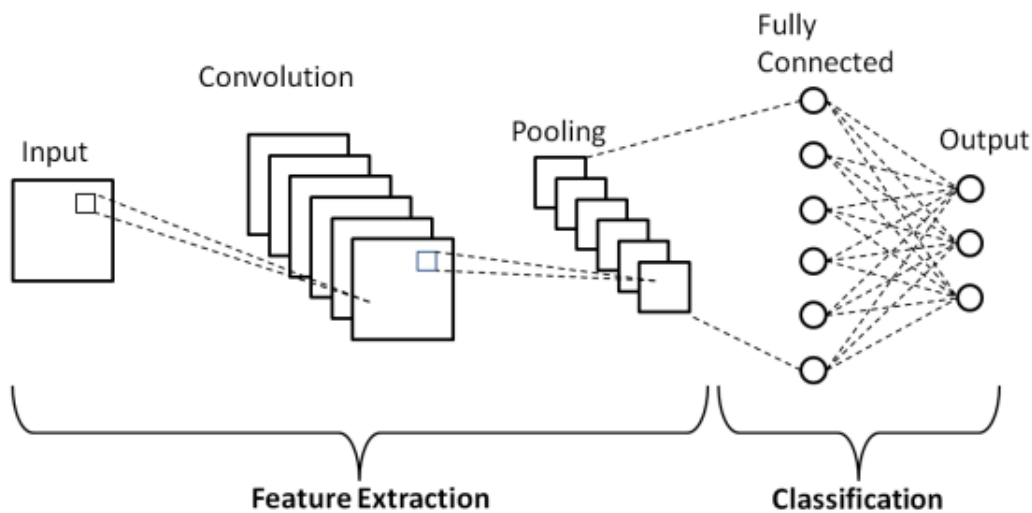


Figure 2.2: A schematic of a CNN architecture. Note that convolution layers contains the feature maps that aim to extract the most prominent features in an input image. Pooling layers reduce the dimensionality of the feature maps for faster computations. The fully connected layer learns the latent representation that the repeated convolutional operations produce [2].

2.2 Generative Modeling

Generative models aim to learn the inherent features of a given dataset by learning how it is distributed. This concept has been rapidly advancing in recent years, and it has found application in different areas such as image generation [19] [20], video generation [21] [22] as well as audio synthesis [23]. These applications, among other, shows that there are potential uses for generative modeling for data manipulation or data generation. One of the most successful frameworks in generating high resolution and detailed images are GANs.

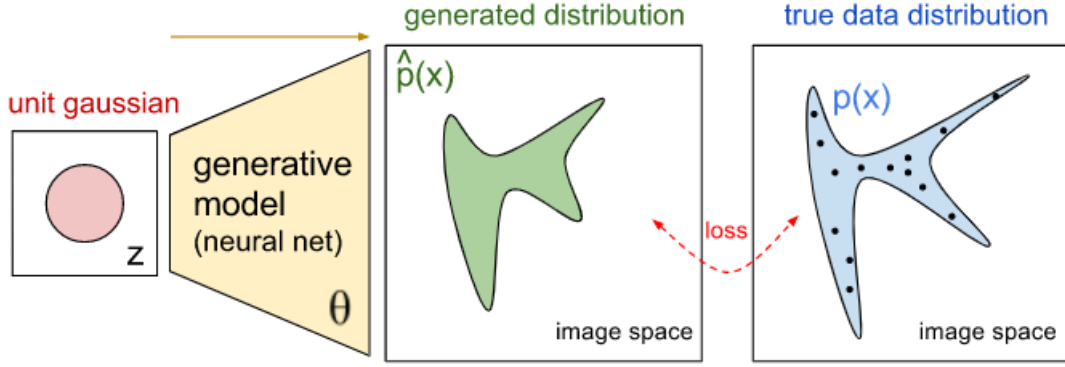


Figure 2.3: General scheme of a generative model. The generative model (yellow) is a function that adjusts the parameters θ as a way to map the input Gaussian vectors (red) to images that matches the true data distribution [3].

2.2.1 Generative Adversarial Networks

In rough terms, GANs map input noise or *latent code* to an output that has the same characteristics as a given training set. This is done by making two networks compete against each other, where the generator network, G aims to create indistinguishable replicas of the input training set. What G generates is validated in with the discriminator network D , which trains to get better at distinguishing real data from the fake one.

In this framework, the generator tries to trick the discriminator by generating realistic looking data, while the discriminator learns to improve at distinguishing fake data while being presented with both real and fake. This in turn make the generator strive to generate data that has the same characteristics as the input data. A converged network is obtained when the discriminator's accuracy reaches 50%, which corresponds to random guessing by the discriminator's side.

GANs are useful when we have a dataset x with the distribution p_{data} and we want to learn a function G that maps input noise z to the data space of x . The resulting output of the function G will have a distribution p_G that is close to that of p_{data} . It is the discriminator's job to distinguish between data coming from p_{data} or p_G by giving a probability of how certain it is that the input is real [11]. G and D competes against each other through the value function V in a way such that

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2.1)$$

where G trains to minimize V by minimizing $\log (1 - D(G(z)))$ and D trains to maximize V by maximizing $\log (D(x))$. Here $0 < D < 1$ and it trains in order to yield $D(x) = 1$ and $D(G(z)) = 0$, while the generator trains to create data that yield $D(G(z)) = 1$.

2.2.1.1 GAN types

There are two setups for training GANs, (i) conditional and (ii) unconditional. In the conditional setup, the dataset that the networks train on contain data with multiple classes. The class labels are concatenated with the latent vector to the input layer of the GAN. With the resulting model, one is able to generate data that has a certain class label. With the unconditional setting input data belongs to one class only [24].

2.2.2 GAN Extensions

GANs can be used to generate different types of data such as images, time-series or tabular data. However, for our application for image generation, we will make use of the StyleGAN framework [4] for its ability to generate high quality images as well as make use of augmentation techniques that are applied on smaller dataset, which suits our case. In the following section, a walkthrough of how StyleGANs work is discussed.

2.2.2.1 StyleGAN

The authors of StyleGAN mainly focused on reconstructing the generator in their framework. These changes were done to obtain a generator with image controlling properties. In this architecture, the generator's input starts from a Gaussian latent vector $\mathbf{z} \in \mathcal{Z}$, and then is mapped to another vector $\mathbf{w} \in \mathcal{W}$. This is done by learning a mapping function f such that $f : \mathcal{Z} \rightarrow \mathcal{W}$. Image controlling properties of the StyleGAN architecture are exploited when each element in the \mathbf{w} -vector controls a feature in the output image. In contrast, one element in the Gaussian \mathbf{z} -vector may change multiple features in the output image at once. In such case, the space from which the latent vectors are sampled is said to be *entangled*. This may also be the case with \mathcal{W} space depending on the training set of the model.

Another feature that the network has is *Adaptive Instance Normalization* (Ada-IN) as a way to incorporate the style of target images into the generated images. Ada-IN for one image (instance) i can be formulated as

$$\text{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i} \quad (2.2)$$

where \mathbf{x}_i is a feature map and $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$ contains the scale and the bias vectors, respectively. The normalization is obtained by inserting the vector \mathbf{w} to the mapping

network **A**. Before Ada-IN operation, a Gaussian noise is inserted to the network's layers at different points and scaled with the parameter **B**, see Figure 2.4.

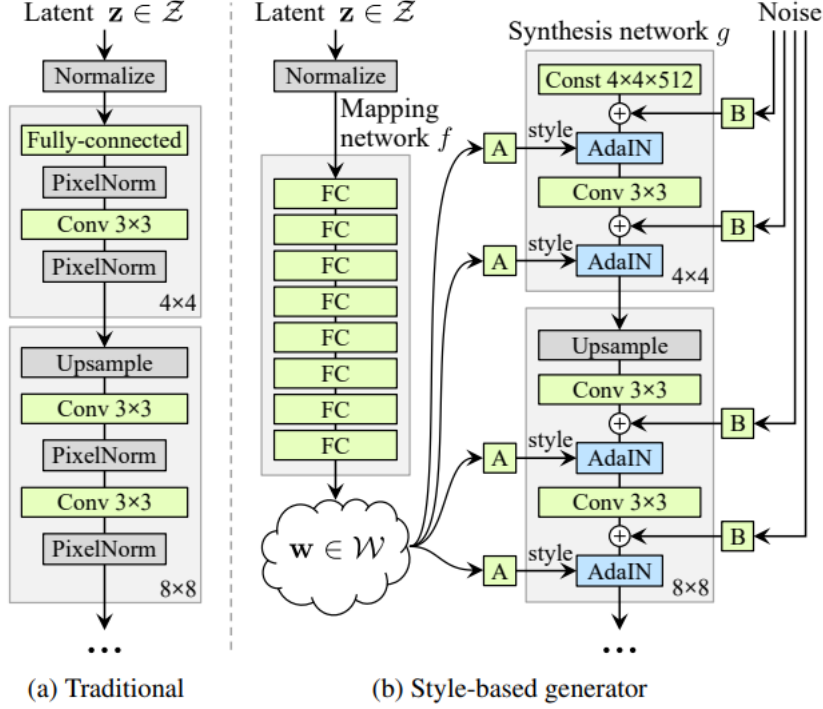


Figure 2.4: StyleGAN generator unlike traditional generators that map a Gaussian distributed latent input to an image, StyleGAN generator learns the input to the generator. The elements from the resulting vectors from the \mathcal{W} -space are each supposed to correspond to a feature in the output image. This makes it easier to edit the image through changing the input \mathbf{w} -vector [4].

As a demonstration of the generative quality of StyleGAN, readers are encouraged to try distinguishing real images from fake by checking out this website; <https://www.whichfaceisreal.com/index.php>.

2.2.2.2 StyleGAN2-ADA

When training GAN models on a small dataset, there is a chance that the discriminator overfit to the training set, which leads the model to diverge in training. A different variant of the StyleGAN framework was built to tackle this issue by applying augmentations to images that are fed to the network effectively increasing the data sample size. The newer version, StyleGAN2 with *adaptive discriminator augmentations* (ADA), provides a framework that is capable of generating more realistic looking images when trained on a dataset that is small in sample size. This is possible due to the use of the image augmentation techniques [25]. These augmentations are applied to all the images that passes thorough the discriminator network

(real or generated) with a probability that dynamically changes in order to avoid overfitting of the model. This adjustment is done by the model where it uses a part of the training set of the model as a validation set as well. When training, the model overfits when it treats the validation set as generated images, this triggers the model to increase the probability of augmenting the images.

2.3 GAN Metrics

To test how a GAN training performs, qualitative evaluation of GAN generated images on a large scale is hard to perform given that the model is capable of generating an infinite number of images, in theory. To this end, one make use of metrics that return values of how good the resulting generated images are. The meaning of these values is put into context when comparing them. To quantitatively evaluate the synthetic data we make use of two metrics, namely *Fréchet Inception Distance* (FID) and *Perceptual Path Length* (PPL).

2.3.1 Fréchet Inception Distance

Fréchet inception distance (FID) score provides a measure of similarity between generated images and real images. This metrics is accomplished using the encoder of InceptionV3 model that is trained on the Imagenet dataset. The encoder takes real images and generated images and encodes these into embeddings. The embeddings are then fitted into multivariate normal distributions and the distance between these distributions are calculated. This means that a lower FID score value implies that the generated images' features are closer to those of the real images [26].

2.3.2 Perceptual Path Length

Since the StyleGAN framework has a learnable latent space that is linearly separable in terms of controlling separate factors of variation in the output image. Measuring this separability will tell us how disentangled the features are. To this end, we make use of *perceptual path length* which is a metric used to measure how separate the features are when interpolating a vector in the latent space. This is done by taking two end points in the latent space, then choose a random point that lies on the line between the end points as well as another point in its neighborhood. The images that correspond to these points are generated and the perceptual distance of the images can be calculated. This is done by acquiring the embeddings of these images through the VGG16 network whose weights are fit to match those of the human perceptual judgment [4]. PPL score can then be obtained by taking the expected

value of repeatedly doing this process multiple times. A lower PPL score means a more linearly separable latent space.

2.4 Classification Metrics

To evaluate the performance of classification algorithms, many different approaches can be used. Relevant metrics that were considered in this project will be addressed in this section.

2.4.1 Confusion Matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Figure 2.5: A demonstration of the confusion matrix. The abbreviations represent; true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The confusion matrix as seen in Figure 2.5 can be used to describe the performance of classification. Answers are compared to what is known to be the truth to visualize the error. In this example there are two labels or classes, however a confusion matrix can be used for multiple classes as well. This is a good metric that can be used for evaluate the performance of an algorithm [27]. The different labels are as follows, true positive: the positive class was correctly labeled as the positive class. True negative is when the negative classes are correctly labeled. False positive is when the the class is incorrectly predicted to be positive, and false negative is the opposite of that.

2.4.2 Other Classification Metrics

Many metrics build on the understanding of the confusion matrix as discussed above. The following four are some of these:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2.4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2.5)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.6)$$

In equation 2.3 accuracy is defined. Accuracy is the rate of correctly predicted classes compared to all classifications. Sensitivity seen in equation 2.4 is sometimes referred to the *true positive rate* (TPR). This compares the correctly predicted positive answers compared to all actual positive labels. Equation 2.5 defines the specificity which is the *true negative rate* (TNR), same as for sensitivity but for negative classes. Lastly equation 2.6 is how many of the positively predicted labels were true positives. These metrics are valuable when evaluating the performance of CNN [15].

3

Materials & Methods

This chapter goes through the process of which the thesis project was executed. We demonstrate exploratory data analysis (EDA), going through the content of the used datasets and their limitations, we then describe the experimentation that was done using the StyleGAN2-ADA framework. A part of our project was doing some exploratory work in image editing using generative modeling, which is also described in this chapter. Figure 3.1 shows the large process steps involved in executing the project.



Figure 3.1: These graphics represent the general work flow of the project.

3.1 Data

This section explores the datasets that were used to train our generative models, we showcase image examples as well as class and bias distribution of the given datasets. When taking images of melanomas doctors use what is called a Dermoscope, a camera which has a circular lens that makes sure that the distance from the subject is where it is supposed to be.

3.1.1 Datasets

Two datasets were utilized in this project, reference examples of the images in the dataset are visualized in figure 3.2. One of the datasets is from a collaboration between Society for Imaging Informatics in Medicine and International Skin Imaging (SIIM-ISIC) Melanoma Classification [28]. The second one is a dataset collected by MD PhD Sam Polesie at Sahlgrenska University Hospital [7]. Both of these datasets contain images of skin lesions.

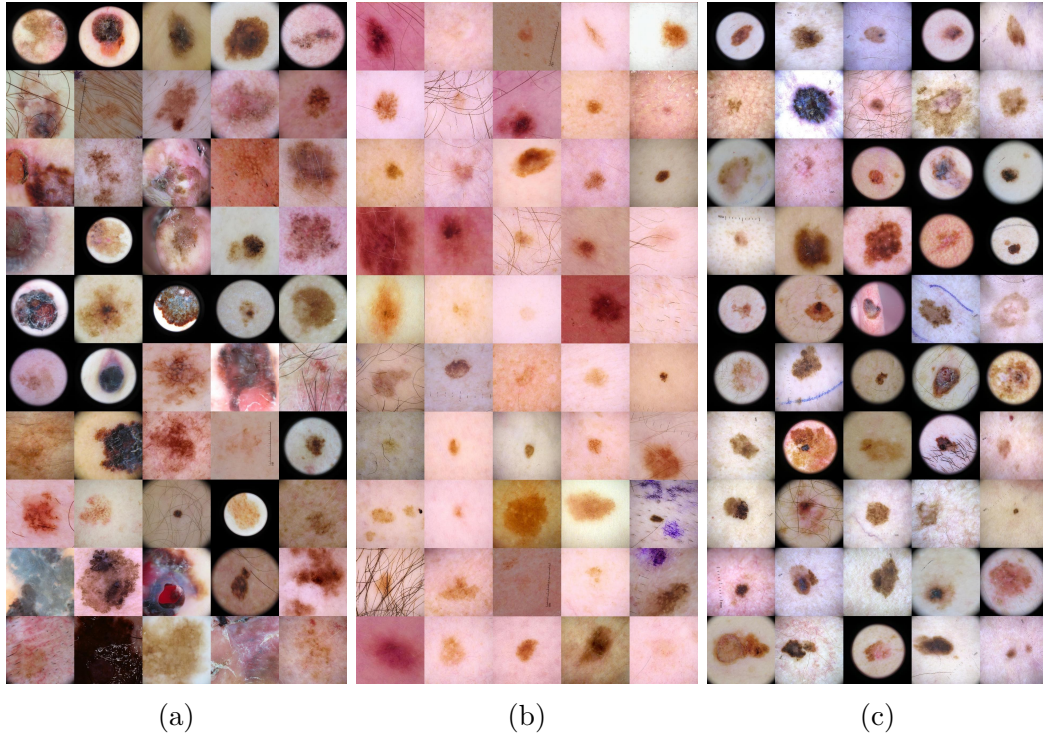


Figure 3.2: (a) Examples of malignant melanoma in the SIIM-ISIC dataset. (b) Images of non-melanoma cases from the SIIM-ISIC dataset. (c) Visual of images with both labels (invasive and in-situ) from the Sahlgrenska dataset.

The SIIM-ISIC dataset is open source and publicly available and is the larger out of the two, contains around 37k images. Along with the images follows a metadata file that identifies features of the individual images. Information for each image includes; labels stating melanomas or benign skin lesions as well as other diagnosis if applicable along with data of gender, age, the lesions location on the body, as well as an anonymous patient identifier. As seen in Figure 3.3a the large majority of images are benign skin lesions, with the imbalance in the dataset, about 13.5% are malignant melanomas. In addition to the main dataset, SIIM-ISIC, there are also extensions that we utilized [29]. This extended set has additional images from 3 datasets representing melanomas and therefore the total dataset is more balanced between the labels.

The dataset collected at Sahlgrenska University hospital is balanced and contains around 1,300 images. The labels in the Sahlgrenska dataset have some more information related to each image in addition to the ones mentioned for SIIM-ISIC. This dataset only contains images of melanomas and therefore list if the moles are in-situ or invasive melanomas. The class distribution of this dataset is a lot more balanced as seen in figure 3.3b.

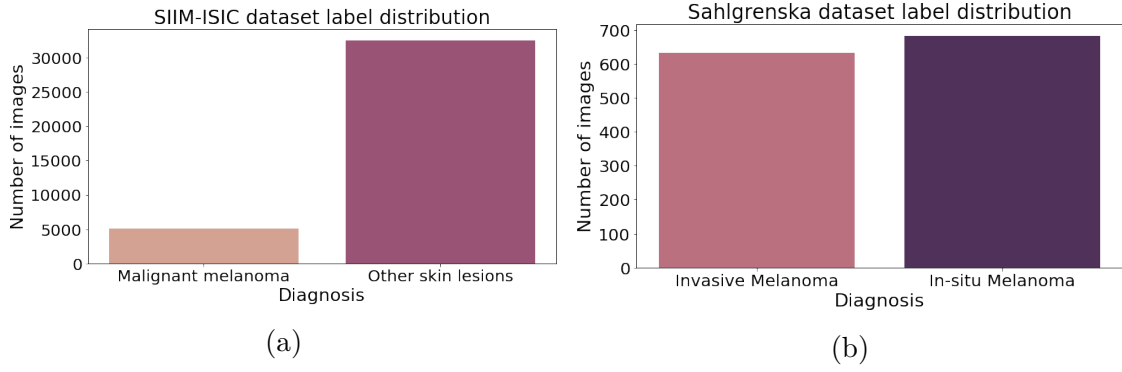


Figure 3.3: (a) class distribution in the SIIM-ISIC dataset. The total dataset has 37648 images of various skin lesions where 5106 of them are malignant melanomas. Note that the used dataset contained extra images of melanoma from external datasets as well. (b) distribution of classes in the smaller Sahlgrenska dataset with 632 invasive melanomas and 683 in-situ melanomas.

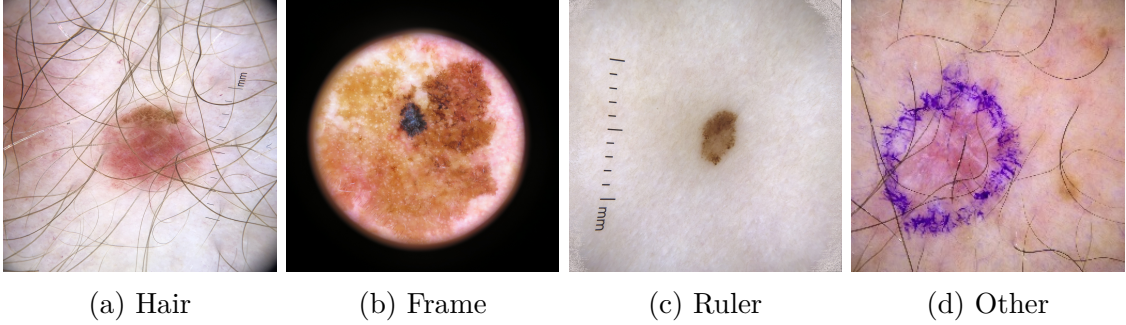


Figure 3.4: Example images with the common biases in the datasets.

3.1.2 Bias

A pretrained Multiclassifier with a measured accuracy of 98% was used to evaluate existing biases in the dataset [30]. Essential biases and their frequency in the images can be seen in figure 3.5. The biases include; frames from the dermoscope lens which is a residue in the image, different types of hair (coarse, fine, short, and long), and rulers that are put on the skin so that the provider can know the size of the lesion. Along with other biases such as; markings made with pens or dust particles. A representation of example images that have biases can be viewed in image (3.5).

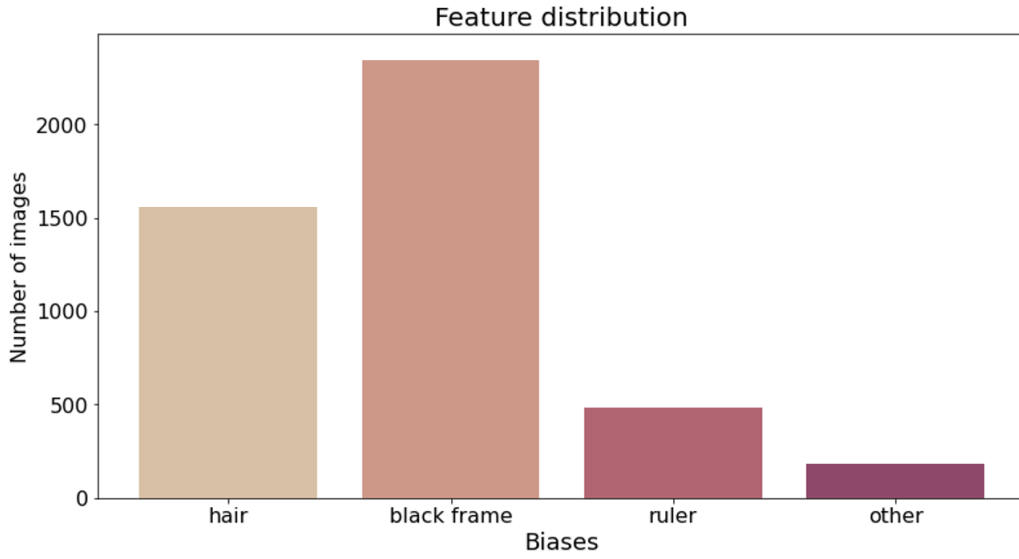


Figure 3.5: Distribution of biases in the malignant melanoma cases in the SIIM-ISIC dataset. Clarification of the biases; hair is in regard to any body hair and black frames are from the dermoscope lens.

3.2 Data Preprocessing

Preprocessing of data is very important in machine learning as the quality of the data highly relates to the quality of the models that can be trained with it. Using the mentioned datasets, we have tried some methods to remove biases such as hair and frames. The reason for this approach was to get rid of biases that would affect classification accuracy per class.

3.2.1 Bias Removal

Training models using biased data can lead to influence performance and result in classification inaccuracies. For this reason we try to remove the biased images from the datasets. We used a classifier which is trained on the SIIM-ISIC dataset to determine if an image has any one of the known large biases in the dataset. Some of these features are; hair, frames, and rulers. The Dermoscope used for capturing and diagnosing the melanomas leaves a frame on images as a black circular surrounding per image, see figure 3.4b. This is a residue from the design of the device. Since this device is only used when documenting melanomas, it creates a significant bias in the dataset where a trained network would classify any image with this suggested frame as a melanoma. In addition to the frame feature many of the images also contain hairs due to the nature of human skin as well as rulers. Rulers are used when taking these images to show the singular size of the lesions.

A couple different approaches were applied to the images with the aim to remove the

frames. The results from the different techniques were compared to see what gave the best result in terms of qualitatively looking at the images. The comparison of the results is discussed in the results section 4.1. Firstly we created an in-painting algorithm that found the RGB values of the skin area in the image while disregarding the melanoma. These colors were then randomly in-painted in the area where the highly contrasted frames are in the original data. Secondly we tried another type of in-painting, Frequency Selective Reconstruction [31] where the frames were used as a mask for where this in-painting should be. The masks served as an edge between the non-biased part of the images and the biased frames. A nearest neighbor approach was then used when in-painting the masked part of the images. Another approach used to remove the frames from the images was to crop the images without removing the content of the image. Lastly, images were also cropped to remove the majority of the frame bias. The performance of these approaches were evaluated visually through output inspection.

In addition, we have looked into image editing methods that were applied post-training. These were done by doing changes in the latent input vector of a trained generator.

Furthermore we also removed other biases such as the hair and rulers from the images. As seen in Figure 3.5 about another 40% of the SIIM-ISIC dataset contains these biases. This process was more straight forward where any highly contrasted thin features in the images was assumed to be these biases and were therefore removed. Inpainting was done using the nearest neighbor technique.

3.2.2 Image Resizing

To work with the desired GAN architectures the dimensions of the images are constrained to $\{64, 128, 256, 512, 1024\}$ pixels with a 1 : 1 ratio. In the datasets there were multiple different sized images. There is a trade-off between the quality of images that needs to be good enough to see details in the melanomas and the speed of training the networks. The decision was to use 256×256 pixels for the images to achieve the mentioned result. To downsize the images, the Python library Open CV was used and method `INTER_LANCZOS4` that interpolates over the nearest 8 by 8 neighbors.

3.3 StyleGAN Training and Experimentation

The StyleGAN2-ADA framework [25] was employed to learn from the datasets. Transfer learning was utilized to improve model performance given that the number of targeted image to generate was limited to learn from scratch, this may affect the color tone of the resulting images.

3.3.1 Computational Power

GANs are computationally hungry models [32] that takes a long time to obtain fully trained generator. To this end we made use of NVidia DGX A100 workstation to train our generative models. These resources, among others were provided to us by AI Sweden’s *Data Factory*.

3.3.2 Hyperparameters

In order to train a GAN there are hyperparameters that need refining in order to get the best performing generator model. One of these are the choice of unconditional and conditional StyleGAN where the conditional GAN uses labels during the training process. This in turn gives the final model the alternative to generate images of each label. On the contrary an unconditional GAN has no relation and takes no consideration to the labels in the dataset. Another hyperparameter that is taken into consideration is the duration of training which is measured in *king*. This is referring to the number of real images that are passed through the discriminator during training. This parameter should be around 20000 king for a converged model, according to StyleGAN2-ADA developers [33].

3.3.3 Manipulating the Latent Input

To be able to adjust the output images, the latent space was altered. Using the same set of seeds data was generated using the generator. The same multiclassification algorithm as previously mentioned was used to find what biases were present in each image. From this we knew which seeds resulted in what biases. Our main objective is to remove the frames from the bias in the datasets. A first approach was to find the regression hyperplane between the data latent vectors that generated frames and the ones that did not. In the first experiments we evaluated each latent vector to see where it was within the space. If the image would be generated with a frame, we avoided that latent vector and continued with the next seed and so forth.

We also tried many experiments when it came to the best way to alter the images we found through the use of a frame-classifier which classifies which images that should have frames. We investigated which direction in the latent space that affected the frames in the images and altered the vector to minimize this effect. In order to find what direction to alter in the latent space we tried different methods. The proficiency of the latent space manipulation was assessed through visual inspection.

While the described methods in above section aimed to make the data more presentable for before training, we have also looked into processing the output data of the generator by training our generator on raw data without any sort of preprocess-

ing. This was done by making changes in the latent point input of the generator which resulted in bias removal while minimally affecting the content of the overall image. One problem with this approach is that it was hard to qualitatively evaluate the overall quality of the image after performing latent shifts on input vectors, so we did this part of the thesis as an exploratory work without focusing much on large scale similarity measures of the images. The proficiency of the latent space manipulation was assessed through visual inspection.

3.3.3.1 Latent Vector Manipulation Using a Binary Classifier

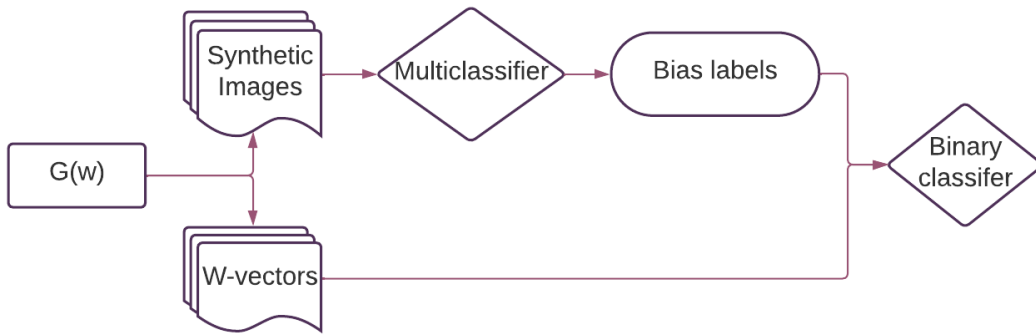


Figure 3.6: Algorithmic chart showcasing the approach with bias removal using a binary classifier. Note that the used vectors here are ones sampled from the learned \mathcal{W} space.

Here, we looked into image editing using a binary classifier as a way to draw a decision boundary between latent inputs that represent images with a specific bias and images without that bias. The goal is achieved by *(i)* generating a large sample size of images with their corresponding latent code, *(ii)* use a multiclassifier to label generated images that contain the unwanted bias and *(iii)* use the acquired labels along with latent code to train a binary classifier that draws a linear boundary between classes, e.g. *support vector machine* (SVM). With the obtained binary classifier, we can shift latent points representing images with biases along the plane normal of SVM to reduce the influence of the unwanted bias.

The downside of this method is that a large sample size of images is needed along with a pre-trained classifier on that specific bias, which is done in a supervised manner. To this end, we have also looked into another method which has an unsupervised approach.

3.3.3.2 Latent Vector Manipulation Using PCA

In this presented method, we projected the sampled latent code on the principal directions where maximum variance can be obtained, using principal component

analysis (PCA). The result is a covariance matrix containing the eigenvectors (principal components) which we experimented with by shifting the latent points along different principal direction and observe different changes in the input image [34].

3.3.3.3 Latent Vector Manipulation Using Semantic Factorization (SeFa)

Contrary to image editing by moving latent points in different principal directions, it is possible to obtain similar results to those using PCA by decomposing the weight matrix of a trained generator. The result is a matrix containing eigenvectors ranked from the ones with highest noticeable change in an image to the lowest [35].

3.4 Synthetic Data Evaluation

Two evaluation methods were used to establish the quality of the generated images. The first method included qualitative input from experts in both fields of dermatology and deep learning. This test was set up to conclude how realistic the images are visually. The second method is a metric version where the images generated are used in various constellations to train a classifier. The classifiers are then assessed on their classification accuracy of real images from the SIIM-ISIC dataset.

3.4.1 Visual Evaluation - Fool the Doctor

A survey was constructed to give the participants that evaluated our images a simple interface to use. The survey was constructed with 200 images where they were divided up in 50% real images from the SIIM-ISIC dataset and 50% generated synthetic images from a model that was trained on SIIM-ISIC images as well. Out of these sets each was made up of 50% malignant melanoma images and 50% benign melanoma images. The real images were picked from the dataset. The reason they were not randomized was because we wanted to remove certain biases, therefore images with rulers were left out of the set. We want the classification to be based on the quality of the melanomas and not given away by the byproducts from the images such as the rulers. For the synthetic set we followed the same procedure as for the real images however we also picked images that were outliers. These outliers were not in the clusters that were visible when projecting classified images using *embedding projector*; the latent representation of input images in the final layer of a classifier before being returning a label for the class. Each class had their own cluster, see figure 3.7. With an embedding projector we can get a good visualization of our high dimensional data points using clustering methods.

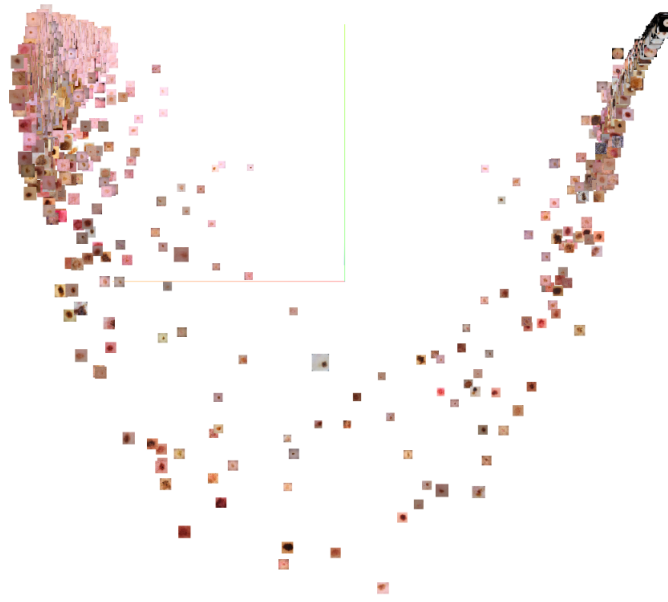


Figure 3.7: This image shows projected synthetic images that have been classified by a pretrained classifier that was trained on real images. The cluster on the left are the benign melanomas and the cluster to the right in the figure are malignant melanomas. The high dimensional data was clustered and represented in 3D using PCA.

Furthermore, the survey included edited images using SeFa to qualitatively assess how good the resulting images were. These images initially contained frames but we manually changed the value of their input to obtain unbiased images.

To setup the survey, Google Forms was chosen for the survey due to its simplicity. The email address of the participants was collected to ensure no two responses were the same. We also asked the persons' background to know their expected proficiency of the task. The answers in question read as follows:

Medical Background:

- ☐ Dermatologist
- ☐ Deep Learning Expert

Each image was showed separately to the professionals, they had to answer the same three questions in regards to every image. These questions were the following:

1. Diagnosis of lesion:
 - ☐ Malignant Melanoma
 - ☐ Not Melanoma
2. Is this image synthetic?
 - ☐ Yes
 - ☐ No
3. Level of certainty:
 - ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Level of certainty is in regards to the answer to question 2, how certain are you that you knew if the image was real or fake. The scale is from 1 to 5 where; ,1=not at all certain, 5=very certain (how certain are you that you knew if the image was fake or real).

The dermatologists that participated in the survey were recruited from Sahlgrenska University Hospital with the assistance of Dr Sam Polesie We reached out to people with experience in generated data to receive expert opinions from them. It is worth mentioning that the data collected from the deep learning experts were not in regards to what type of melanoma they observed as that is not part of their competence. As the participants were taking the survey they did not find out the correct answers. The answers to the survey were compiled and analyzed to find the proficiency of our images.

3.4.2 Classification Model

The second evaluation method used on the generated images was testing using classification accuracy. A number of classifiers were trained using EfficientNetB2 and the same general set of synthetic images to label melanomas and not melanomas. EfficientNet was chosen as it outperform other CNNs with higher accuracy and fewer trainable parameters, as Tan describes [36]. The classifiers were also all validated and tested using real images of the same 256×256 resolution from the same SIIM-ISIC dataset. Five different setups of the training images were tried to see how the

Classifiers		
Classifier	Type	Description
1	Real Baseline	No processing except 256 resolution
2	Synthetic Baseline	No processing except 256 resolution
3	Synthetic No bias	Remove all images that have biases
4	Synthetic No frames	Remove all images with frames
5	Mixed Baseline	40% real and 60% synthetic

Table 3.1: Training setups for different data processing methods.

biases could be handled. The training setups were per table 3.1. The test accuracy was compared to see how well the classifiers comprised of different datasets performed on classification tasks.

Seen in table 3.1 are the specifics of the limitations to the images used when training the different classifiers. 10,000 images of each label, malignant melanoma and not melanoma were used to train the classifiers composed of synthetic data. For the validation 1,500 images of real skin lesions from the SIIM-ISIC dataset were used, these are divided equally between both labels. Another 1883 images were used for the test set, which was unbalanced with 255 images of melanoma. Note that a separate set of images was used to train the generator and hence the classifiers, these were not a part of the validation and test set.

These classifiers listed above (3.1) will also be compared to a classifier trained only on real images to see the accuracy lost when generating images. The real classifier due to the lack of diversity in the SIIM-ISIC dataset will be trained on 4,000 images of each class.

4

Results & Discussion

In this chapter we will discuss and present the results achieved with using the methods mentioned in the previous chapter.

4.1 Data Preprocessing

The focus of preprocessing was to remove the frames resulting from the Dermoscope device as 38% of melanoma images in the SIIM-ISIC dataset have frames, while there are no images containing this artifact in the not melanoma data. We do not want a trained classifier to associate frames with melanoma labels and therefore classify every image with frames as melanomas, this is not a relevant real metric for diagnosis.

In our efforts to remove the frame bias from the images in the dataset, we tried different preprocessing approaches. See Figure 4.2 for the results of the experimentation. As can be viewed in the figure the purpose of this preprocessing step was to remove the black frame that is a residue from using the dermoscopic lens to take the photos. We found that 38% (Figure 3.5) of the malignant melanoma cases have this frame bias which from earlier research has proved to cause a prejudice in the classifier to connect the occurrence of frames to malignant melanoma labels [30].

The first approach was to cut the image so that the frames would disappear. In order to cut all the dark pixels we lost some of the data of the melanomas. As can be viewed in the second and third picture in the second row in the figure 4.2 much of the melanomas have been lost in the cropping of the image. This happens since the information in the image is round while we need a square resolution when working with GANs. When removing all pixels from the dermoscope we also have to remove some information when the melanomas are larger and take up a larger proportion of the picture. See figure 4.1 for a representation of how the cropping takes place. Because of the loss of information we decided to explore other alternatives.

The second alternative that we tried was to find where the dark pixels from the

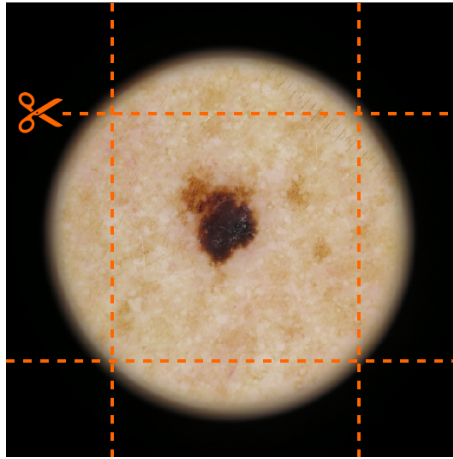


Figure 4.1: This graphic illustrates how the images were cropped to remove all the frames from the dermascope.

dermascope were and then use an algorithm to "paint over them". The rest of the pixels in the image were evaluated for color and a normal distribution was fitted to the data. This was assumed to be the skin colors as the samples in the dataset are from fair-skinned people. These colors were then randomly sampled and replaced the dark dermascope images. This worked well in terms of removing the frames as well as not having the same issue as the cropping technique did since all information in the melanomas remained. However, we found that when we trained GANs on these images it learned the preprocessing method and regenerated the inpainted random pixels. For this reason we kept exploring other alternatives as these fragments were undesirable in the efforts to fool a human into thinking that the images are real. See the third row in figure 4.2 for examples of the resulting images from this technique.

As seen in figure, 4.2 above, we tried an additional third technique as mentioned in the method section (3.2.1). This preprocessing technique was using a *frequency selective reconstruction* (FSR) algorithm. This used a similar idea as the random inpainting above. Seen in the last row in figure 4.2 the result is also very similar to that of the row above.

All of the preprocessing methods tested succeeded at removing the frame bias. However, the desired result is an image that a professional dermatologist would believe to be real. After analyzing the images created using the discussed techniques we did not believe that they looked realistic. Since the images' realism was not convincing to us, we do not believe that they would look convincing to an expert. Due to this discrepancy between preprocessed and original images, we decided to move away from preprocessing. Different options were considered with the prospect of removing the frames. Alternatives for manipulation and variation of the latent space vector was chosen as a viable solution. These experiments involving latent space are discussed in section 4.2.2.

In Figure 4.3 the bottom row shows the images that had hair removed using the

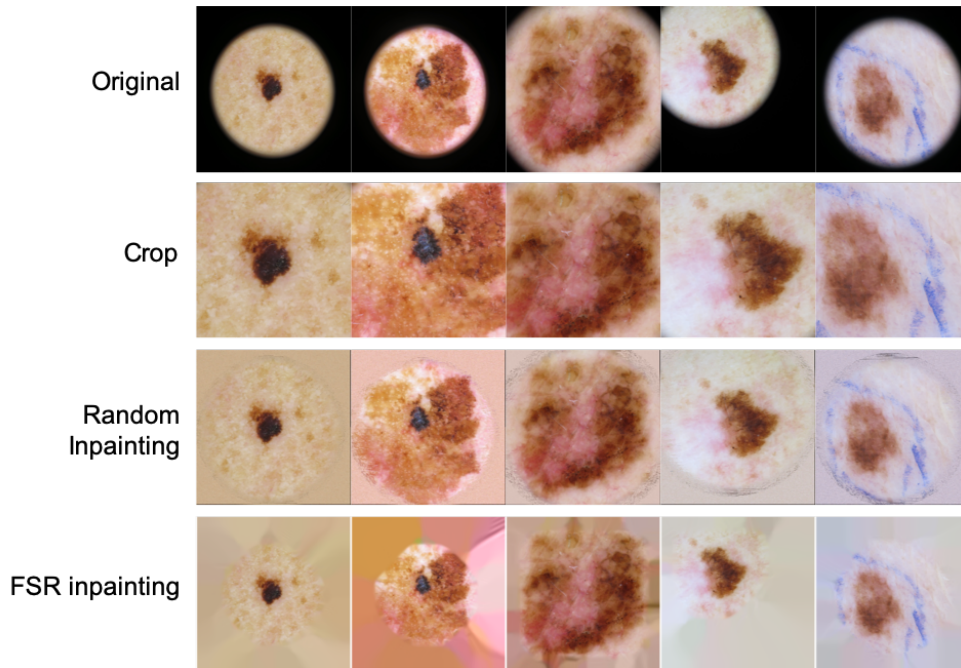


Figure 4.2: shows 3 of the different preprocessing techniques we tried to remove the dermoscopic frames. The top row shows the original photos from the dataset. The following rows below show the removal techniques applied on the same images.

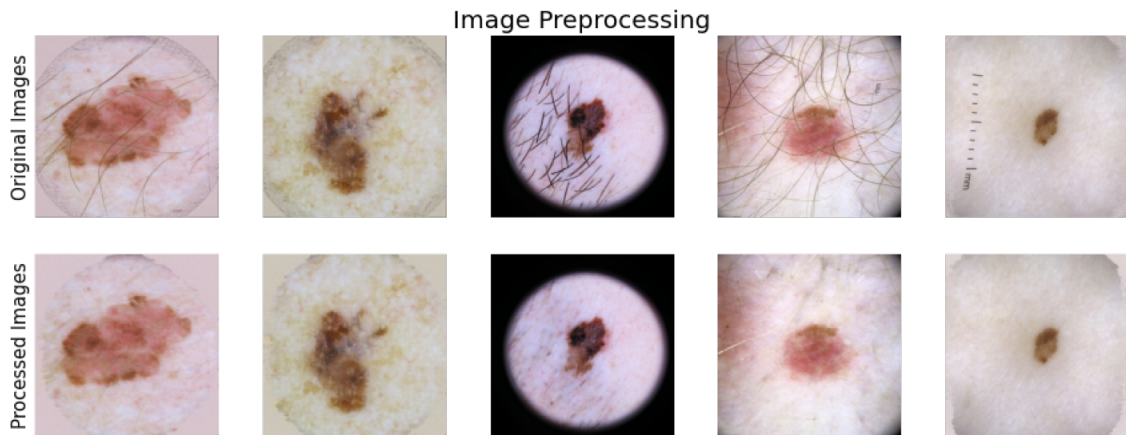


Figure 4.3: Example showing images before and after algorithm described in section 3.2.1 was applied.

technique described in methods (3.2.1). As seen in the images the algorithm worked fairly well. Although, images with more dark features in addition to hair or rulers also lost some of the definition in the features. We decided to keep training with data containing hair as both labels in the dataset had hairs so it was not considered a bias that could skew the behavior of a classifier.

4.2 StyleGAN2-ADA Experimentation

This section exhibits the experiments done to optimize the synthetic images. We will also discuss the methods that were evaluated to edit the features of the output.

4.2.1 StyleGAN2-ADA Training Results

The generators of our StyleGAN models were trained using the ISIC dataset as well as the SUH dataset. In this section we present the results obtained by both models qualitatively and quantitatively.

4.2.1.1 Training Results on ISIC Dataset

Per table 4.1 different GAN models were trained. This was done using either conditional or unconditional type StyleGAN2-ADA, hence the data that the models were trained on also varied as listed in the table.

Model	Description	Training set
benGAN	Uncond. model generating data of benign skin lesions	32k
melGAN	Uncond. model generating data of melanoma images	5k
cGAN	Cond. model generating both classes	37k

Table 4.1: lists the separate models trained and their abbreviated name. Training set refers to the total number of real images that was used to train each GAN. The type of GAN that was used to train them is also specified; unconditional (uncond.) and conditional (cond.).

Generally, conditional cGAN achieved better melanoma image quality than unconditional melGAN, see Figure 4.4 and Figure 4.6. Note how the conditional model is able to generate artifacts like hair and frames with attention to details while the unconditional model had more synthetic looking biases. Through acquiring metrics for both conditional and unconditional GANs shown in table 4.2 we can also see that the unconditional melGAN images were worse in comparison to the cGAN images. Seen in the table, lower PPL score means that the \mathcal{W} -space is more regularized, indicating a more linearly separable space for image editing. The results also indicates that the low FID metric for the conditional cGAN model is prone to generating more realistic looking melanoma images than the unconditional melGAN model. This is a result of having trained the unconditional model on a dataset with low sample size. Overall, the training for both benGAN and cGAN were started from scratch, whereas the training for the unconditional melGAN model was obtained by finetuning benGAN on only melanoma images. This was done to offset the implications of having a small sample size of melanomas overall.

	FID	PPL
benGAN	7.5	59.0
melGAN	14.4	59.3
cGAN	7.2	77.2

Table 4.2: FID and PPL metrics derived from the trained models using SIIM-ISIC dataset.

Note the low FID score for conditional GAN in table 4.2, this suggest a better model, which in turn makes it more suitable for qualitative analysis. The low PPL score for unconditional models indicates a more disentangled feature space allowing for easier feature manipulation. Although the quality of the unconditional melGAN images were worse in comparison to cGAN, image editing through changing the generator’s input was easier applied to unconditional GANs.

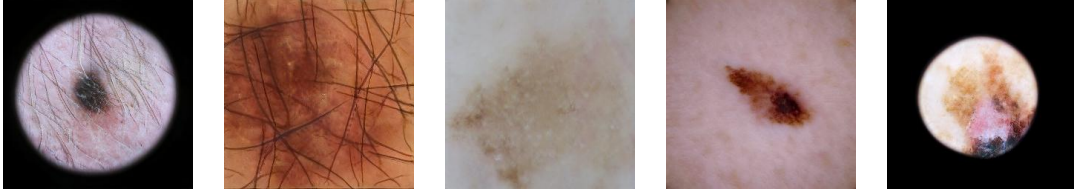


Figure 4.4: Example of synthetic melanoma images generated through unconditional GAN (melGAN). The training for this model was resumed from a benign skin lesion generating model.

The generated images from unconditional GAN succeeded at capturing a variety of features from the dataset. However, finer details such as hair and the dermoscope frame were not entirely captured by the model at the same quality as the original images. This may be related to the fact that even though transfer learning was used as a starting point for the model the number of images that the model resumed training with was not large enough to achieve desired results, as displayed in Figure 4.4.

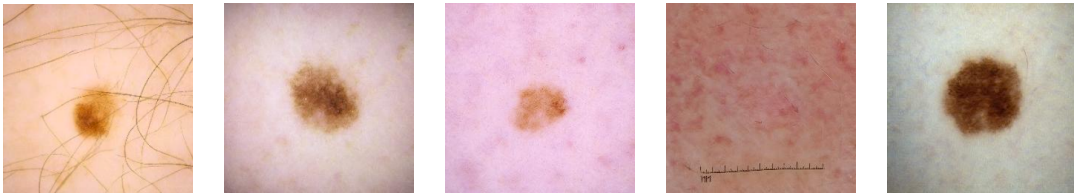


Figure 4.5: Example of synthetic non-melanoma skin lesion images generated through unconditional GAN (benGAN). The training for this model was started from scratch as the training set used was significantly larger than the melanoma skin lesion set.

Compared to images generated using melGAN, the benGAN model yielded images that generally looked better in quality than the ones trained on melanoma skin

lesions. The model was able to generate images with finer more realistic details, like hair strands and dermoscope ruler, see Figure 4.5.

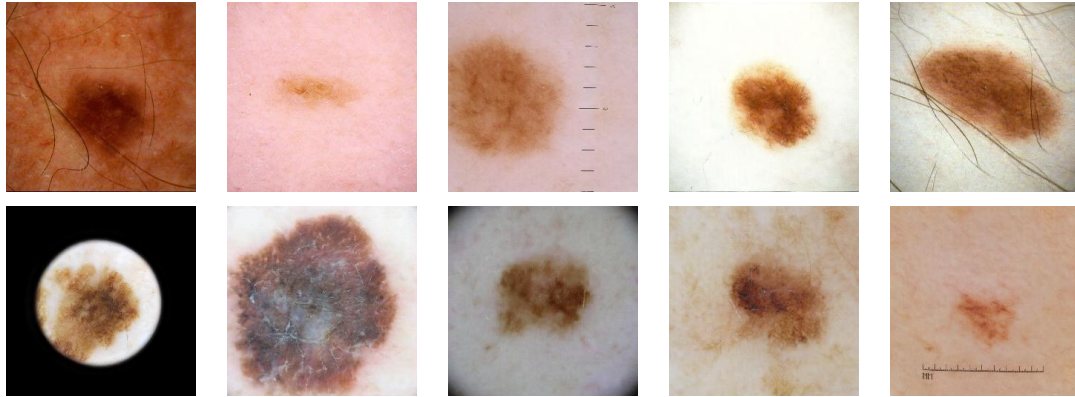


Figure 4.6: Example of synthetic skin lesions generated using the conditional GAN model (cGAN). Here, we see the per image quality for both benign skin lesions (top row) as well as for melanoma (bottom row).

The details in the images are more realistic looking in Figure 4.6. Qualitatively the images are hard to distinguish from the real images of the training set. These images are from the trained conditional model, that had the largest training set (30k images) in comparison with the unconditional models, this yielded more detailed images, see table 4.1.

4.2.1.2 Training Results on SUH Dataset

As a consequence of having a small sample size in SUH dataset (1,300 images), we trained a conditional model that was resumed from the trained conditional cGAN model in an effort to get a more generalized model. The trained model had a FID score of 22.5, which is relatively high when compared to the models trained to generate ISIC dataset images. As seen in Figure 4.7, the images generated through this model weren't able to synthesize artifacts like hair, ruler or melanoma in a qualitatively convincing way, see figure 4.7.

Since the quality of the trained GANs depend on the size of the training dataset, models trained on the SUH dataset to generate images of invasive and in-situ melanoma did not yield realistic looking results. Subsequently we made the decision to focus on the images we generated using only the SIIM-ISIC dataset so that we could focus on acquiring answers to our research questions. Furthermore, we chose to continue the project using the cGAN, because it had the lowest FID score as discussed above as well as when visually evaluated, they looked the best. The cGAN model was used to generate images for the survey as well as the training sets for the classifiers.

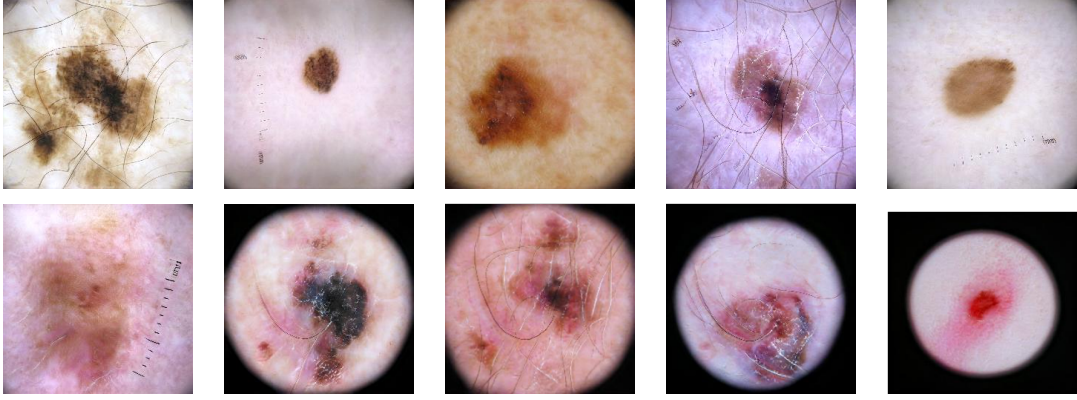


Figure 4.7: Synthetic skin lesions generated using a conditional GAN model. Note how some hair patterns are repeated among the the images. This was due to the small number of images the model was trained on. The dataset used here was from SUH and had a sample size of 1.3k images.

4.2.2 Manipulating the Generator’s Input for Bias Removal

In this section we demonstrate the executed experiments done in order to remove biases through manipulation of the latent input vector for the generator. This was carried out in a systematic way to obtain interpretable outputs. Our goal was to remove unwanted biases from the images without changing the overall information in the image. We present examples of edited images using the different methods.

4.2.2.1 Image Editing Using a SVM

The first technique explored followed the steps discussed by the authors of StyleGAN [4]; separate biased and un-biased images. This was done through training a binary classifier on a sample size of $20k$ latent codes sampled from \mathcal{W} -space. The classifier defines a boundary (decision boundary) between images that contain frames from the Dermascope and images that do not, see figure 4.8b for an illustration. Utilizing the network model Multiclassifier for bias labeling along with the large sample size of latent codes, we obtained a binary classifier with an accuracy of 94.5%. The classifier is able to predict if a generated image is going to have that bias or not from the corresponding latent code of said image.

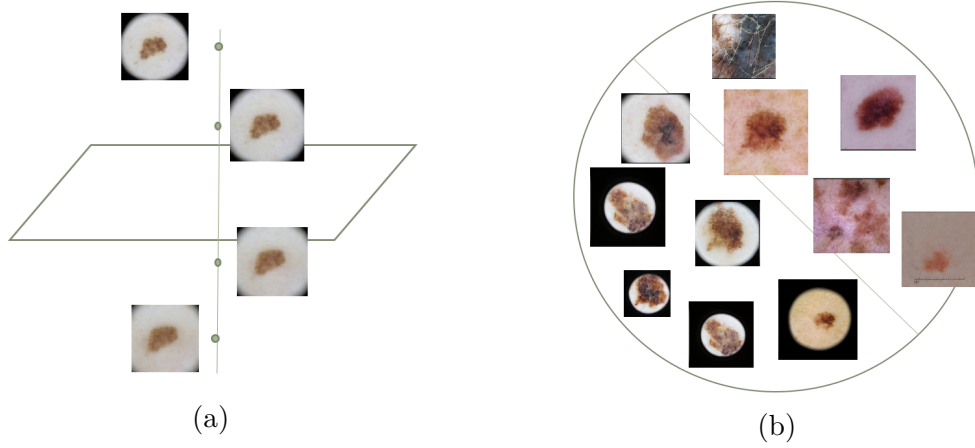


Figure 4.8: (a) An illustration of moving an image perpendicularly towards the decision boundary of a trained SVM. Note that while the frame is gradually removed from the image, the melanoma gradually changes its form as well, making it hard to know if the resulting frameless image remains a melanoma. (b) Simplified illustration of how the binary classifier separates images with labeled biases. Here, the circle represents the learned style space \mathcal{W} and each image correspond to a latent input \mathbf{w} -vector.

As 38% of the images in the dataset have the frame bias, we were concerned about the distribution of the images if we choose to not consider images with frames for generator training purposes. Therefore we tried using SVM to edit the images instead of just removing them from the sample. The image editing was done using the normal plane of the trained SVM classifier. We explored the effects of shifting a \mathbf{w} -vector that corresponds to a biased image along the normal plane. This was attempted as another technique to remove the bias of that image, see figure 4.8b for resulting images. This method of image editing may not be optimal since it depends on how accurate the Multiclassifier is at labeling biases compounded with how accurate the SVM model is.

4.2.2.2 Image Editing Using PCA

Instead of relying on the Multiclassifier to label biases, we explored our trained generator to obtain the principal directions for a large sample size of latent codes sampled from the \mathcal{W} space. This in turn yields the directions where maximum variance in the latent code occurs. We took an input vector \mathbf{w} that corresponded to an image with bias and aimed to remove the bias by shifting the vector along the principal directions. The resulting data are shown in figure 4.9.

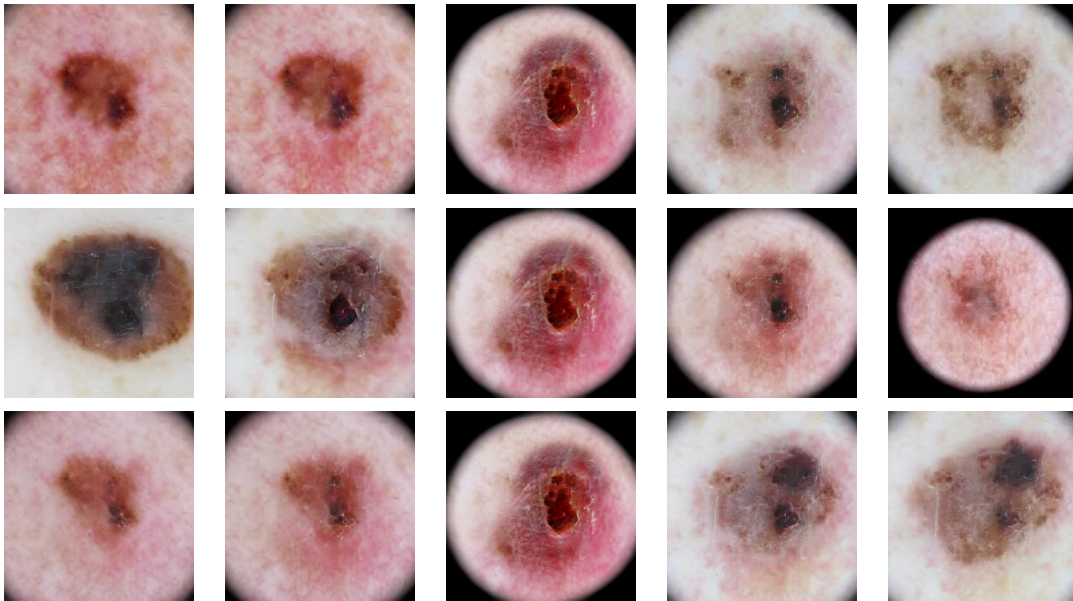


Figure 4.9: Results from shifting the latent vector \mathbf{w} along the 3 first principal directions, along with the unwanted frame bias being removed. The center image is the original, the images to the right and left represent opposite and constant directions. Other features in the image, such as the mole's shape and size change as well, and now we begin to question if this image still represents a melanoma.

For this technique we also see feature entanglement, a result from shifting the vector as visualized in Figure 4.9. This way of image editing is prone to changing other not targeted features in the original image. As a consequence, this affects the certainty of this image still having the features of a melanoma skin cancer after editing.

4.2.2.3 Image Editing Using SeFa

The performance of SVM and PCA based image editing is dependent on a large sample sizes. We decided to explore the effect of a method that does not require a sample size of latent code, namely, semantic factorization. The corresponding latent \mathbf{w} -vector to the image in Figure 4.10 was shifted along the second, fourth and sixth eigenvectors. These were obtained using the SeFa framework [35]. Looking at the figure we see that applying SeFa image editing suggests less entangled features from visual inspection of the images of different directions. The eigenvectors displayed are chosen from a larger qualitative evaluation of multiple images and directions. The second, fourth, and sixth eigenvectors showed the best result in removing the frames while leaving the other features intact. We observe a better result using this method of image editing compared to the other tested methods.

For the purpose of this project many images were evaluated using each method (in addition to the example figures) and SeFa was determined the most reliable to

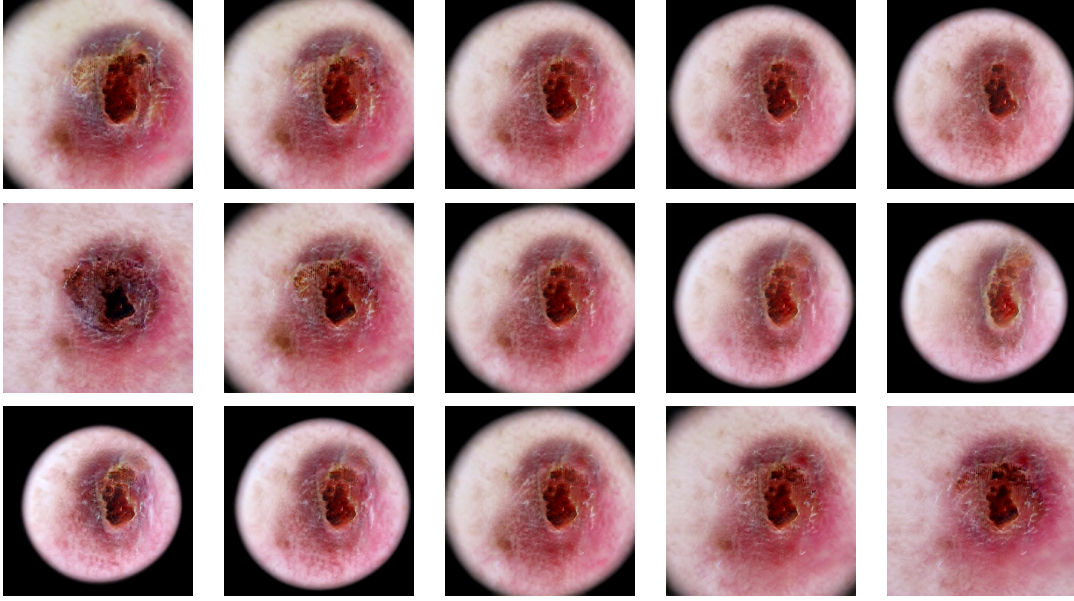


Figure 4.10: Examples of image editing using the SeFa framework. The image in the middle in all three rows is the original image. The rows correspond to the \mathbf{w} -vector being moved along different eigenvectors. Left and right of the original image are positive and negative eigenvectors.

remove the unwanted frames while leaving the other information intact. We therefore moved forward with this method and edited images using the 6th eigenvector. These images were then used in a training set for a classifier discussed in section 4.3.2. We noted that different \mathbf{w} -vectors correspond to different places in the latent space, meaning that the same eigenvector does not affect every image the same way. We may not always obtain an image that still looks like a malignant melanoma after editing.

4.3 Synthetic Data Evaluation

As discussed in Methods (section 3.4), generated data was evaluated using two methods; *(i)* showing it to people that have experience in synthetic data or in melanomas, and *(ii)* training classifiers using synthetic images and evaluating them on real data. This section presents and discusses the results using mentioned methods.

4.3.1 Survey Evaluation

We received 4 answers to the survey, the answers were provided by two doctors with expertise in the field of dermatology (ED) as well as two experts in deep learning (DLE). The results in regards to the visual quality of the synthetic data can be

viewed in table 4.3. Our main objective from this visual test is to find out how many synthetic images (positives) were classified as real images (negatives).

	ED1	ED2	DLE1	DLE2
TP	70	36	38	64
FP	33	31	45	48
FN	35	69	67	41
TN	62	64	50	47

Table 4.3: Results from survey evaluation of synthetic data. Each participant’s confusion matrix answers are listed in this table.

With the small number of participants in the survey we can not draw any statistical conclusions, however, we can analyze the trends and get an overall impression of the quality of the data and responses. Looking at the results found in Baur’s research [16] where they used PGANs (a different older framework) to generate melanoma images we can compare our results. Baur also tested the subjective quality of the images through input from dermatologists and deep learning experts. The average accuracy in Baur’s study was 0.58% compared to our 0.54%. This indicates that the images generated by our StyleGAN2-ADA may be the better option compared to PGAN.

	Accuracy	Sensitivity	Specificity	Precision
ED1	0.66	0.67	0.65	0.68
ED2	0.50	0.34	0.67	0.54
DLE1	0.44	0.36	0.53	0.46
DLE2	0.56	0.61	0.49	0.57

Table 4.4: Survey related metrics. Listed is the accuracy, sensitivity, specificity, and precision in relation to each participants answers.

Analyzing the accuracy (table 4.4) of the survey participants responses there is no obvious trend that lead us to believe that the quality of the synthetic data is unsatisfactory. The synthetic data does pass for real as per the precision of the participants answers. As seen in figure 4.11b there were seven generated images that were classified as reals by the experts but were actually synthetic. These images are displayed in figure 4.12.

An important metric for the project is to know if there are any obvious synthetic images. A qualitative analysis help us understand if there are features of the generated images that the experts picked up on. Looking at figure 4.11b we see that the distribution of how many of the experts that guessed true positive on each image is dispersed. As mentioned there were seven images that none of the experts assigned correctly, nonetheless there were also five images that all experts classified as true positives, see figure 4.13. The top four images in figure 4.12 are images that had been edited to remove the frames. This shows us that even when manipulating the image, the integrity remains and it still looks like a realistic lesion.

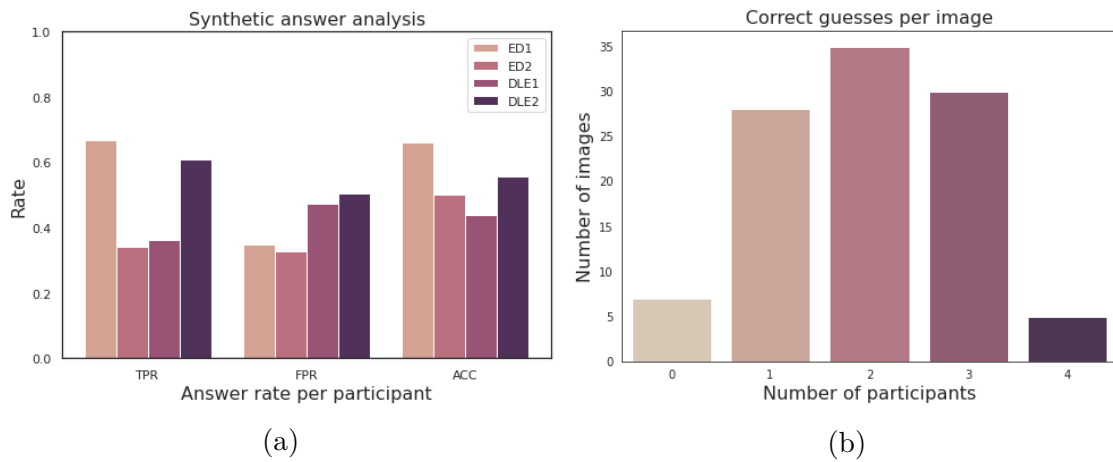


Figure 4.11: Ability to distinguish real images from fake images. (a) Metrics visualization of the rates for the individual participants in the survey. The abbreviations here are; true positive rate (TPR), false positive rate (FPR), and accuracy (ACC). (b) Showing the number of correct guesses per synthetic image.

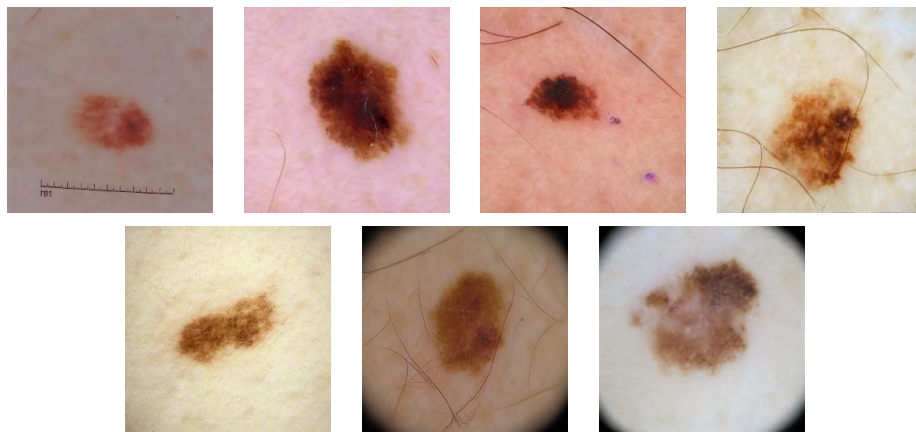


Figure 4.12: This figure displays all images that were incorrectly classified as false negatives by all survey participants.



Figure 4.13: The images in this figure were all the images that the participants correctly classified as true positives.

Looking at the images in figure 4.12 we see that the majority of them have one or more of the biases discussed in section 3.1.2. It could be the case that people are predisposed to put more emphasis on these features when evaluating an image. With this hypothesis we want to say that a ruler in an image could skew the believability of said image as it is perceived to be realistic. Although looking at the images in

figure 4.13 hair does not seem to bias the evaluators as they classified images with visible body hair as both real and synthetic. These two figures (4.12 and 4.13) are representations of what are both the most successful and least successful generated images of our synthetic data set. Qualitatively there is no obvious resemblance of the lesions. Because of this we can not draw a conclusion that there are specific traits of generated images that would give them away to either medical or technical expert. We also received some comments regarding how convincing the synthetic images were. These comments are displayed below:

"I would say they were very convincing. Hair structures can sometimes help in the assessment. I found this task really challenging. "

(Dermatologist 1)

"Some pretty convincing, some clearly fake"

(Dermatologist 2)

"Hard to say when you don't have the correct answers yet! But overall it was hard to tell them apart. "

(Deep Learning Expert 1)

"very convincing! well done! I was mainly focusing on the skin around the mole, hairs and the centering of the mole in the image but was still very unsure how it went"

(Deep Learning Expert 2)

There is a clear correlation between the comments made and the result from the survey. Most of the comments are in agreement that the images are convincing. Recall, the participants did not find out which answers they got right or wrong so these comments are all based on how they perceived they performed. Hairs in the images was something they mention as bias that gave the synthetic images away. The hair removal algorithm as mentioned in section 4.1 could be improved further and used to eliminate this bias. This being said, as mentioned and displayed in figure 4.12 this is not the case for every image. Overall, the participants believed the experiment to be successful as they had a difficult time telling the real and generated images apart, this is consistent with the result from their evaluation (table: 4.4).

In figure 4.14 are real images that all participants agreed on. In these cases there are no specific biases in the images that were correctly identified as real patient data per the figure 4.12. This might suggest that the bias also affects the experts' ability to classify lesions, as all but one false negative image had some sort of bias. The images with a bias could suggest some false notion of realism to the experts.

Another measurement that was collected through the survey was how confident the participants were with their answer regarding if the image is generated or not. This is a subjective measurement and will inherently have more variation between persons. The resulting average certainty for all images and all participants was 3.0.

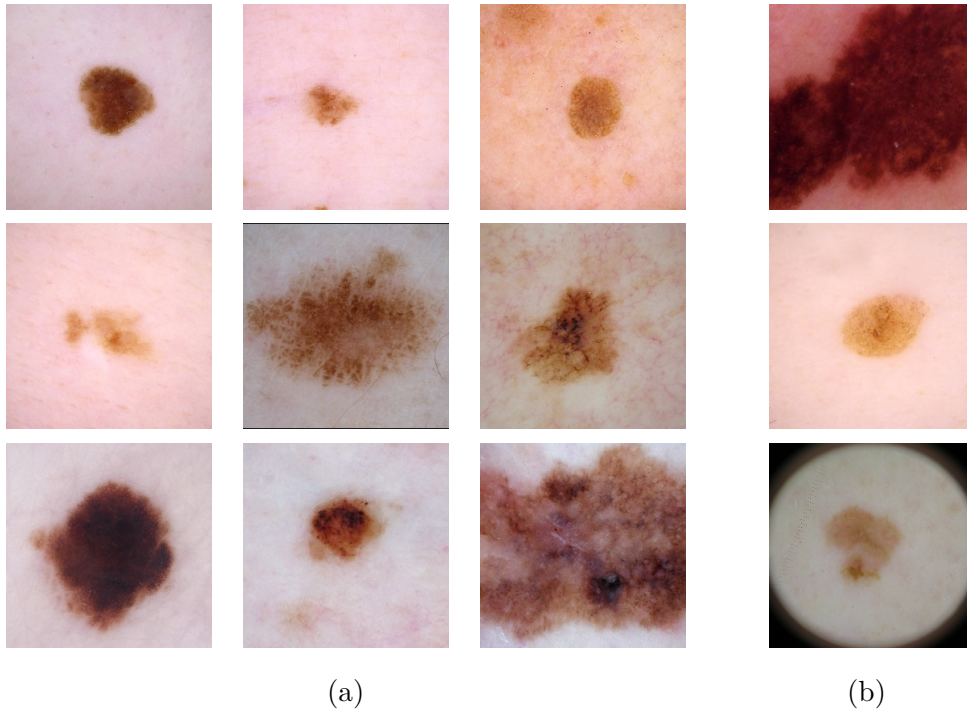


Figure 4.14: (a) displays 9 out of 14 images that were classified as true negatives by all survey participants. (b) displays all images that participants classified as false positives.

Knowing the average certainty and looking at graph 4.15, the certainty distribution is not uniform. From this data, the labels assigned to the images were not obvious choices.

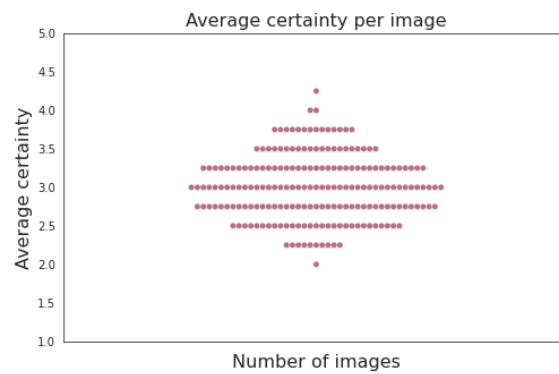


Figure 4.15: Participants average certainty per each image. The certainty is in regards to how sure they were that they labeled the image correctly (synthetic or not).

Another question on the survey was for the experts to diagnose the lesions. The only answers taken into considerations were the ones made by the dermatologists. The result from this is shown in a confusion matrices, see figure 4.16. As seen, the lesions in the survey were not easy to diagnose and the average accuracy between the two doctors was 74%. There is not a large variation between the answers to the

real melanomas and the answers to the synthetic melanomas.

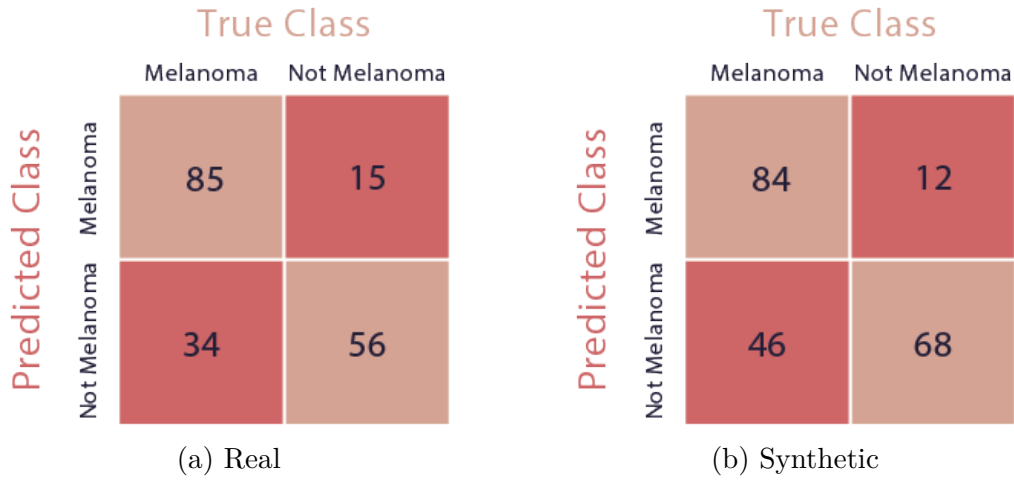


Figure 4.16: (a) is a confusion matrix of the diagnosis answers to the survey where they choose from malignant melanoma and not melanoma. These images are the dermatologist diagnosis of the real patient data. (b) displays a confusion matrix of the diagnosis of the synthetic data for the two dermatologists.

For the experiment we chose synthetic images that were outliers, as discussed in section 3.4.1. A point of concern arose when generating a specific class, would the generated image represent that class with reasonable fidelity. To evaluate this we chose to use images that fell close to images of the other label. We chose images that were difficult to classify. We wanted to assess if the label they were suppose to represent was definite. From the results of the survey it appears that the performance of the classifying doctors was representative for both types of data (real and synthetic). There is some uncertainty as the doctors' accuracy was rather low at 74% and the precision is quite low. Though the distribution of the confusion matrix being very similar between real and synthetic suggests accurate labeling or similar label distribution. This result is nevertheless an indication that the synthesized images actually represent the class that they are supposed to.

4.3.2 Classification Model

The second part of synthetic data evaluation was to use synthetic data to train a classifier to classify either melanoma or not melanoma. To compare the different ways of training classifiers introduced in section 3.4.2 they were evaluated on the same balanced dataset from the SIIM-ISIC set. The result from the evaluation is arranged in the table 4.5.

All datasets but the real image baseline were of 10,000 images from each class. There are only about 5,000 melanoma images in the SIIM-ISIC set and 20% of them were used for test and validation. The small number of test images is a limitation in this research. We had around 4k images remaining to use when training the classifiers.

The last column in the table 3.4.2 is for the performance of a model that was expanded using synthetic data. Expanding the dataset to get a larger and balanced one resulted in a higher accuracy when predicting melanomas.

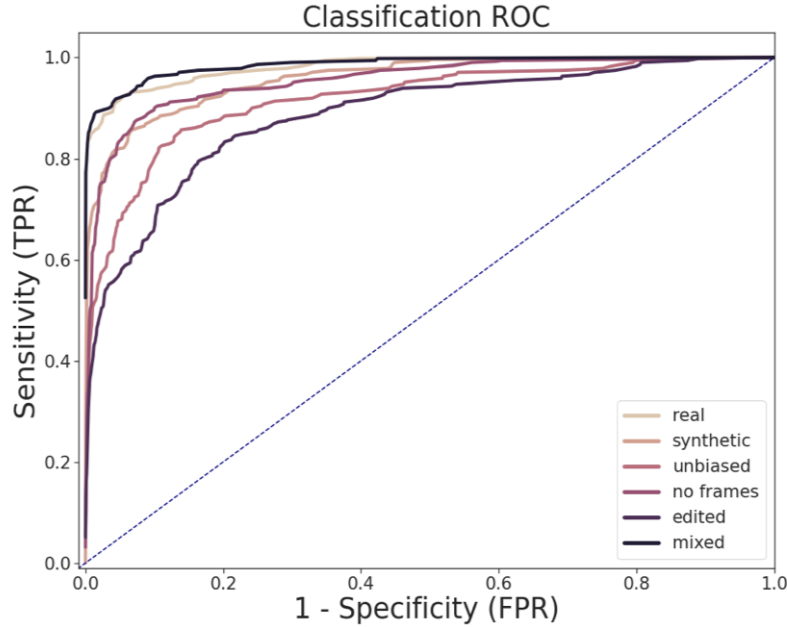


Figure 4.17: ROC curves comparing the sensitivity (true positive rate) to the false positive rate at varying thresholds for each of the different classifiers.

The *Receiver Operating Characteristics* (ROC) curves as visualized in Figure 4.17 helps us understand how well the models can distinguish between the classes. We can get a sense of which classifier is better through the *Area Under the Curve* AUC, listed for each classifier in Table 4.5. In our experiments the classifier that performed the best was the mixed dataset. The mixed dataset contains the images in the real image baseline, however, complemented by synthetic data so that the training set could be expanded.

The results from the adjusted datasets show a lower AUC overall compared to the baselines. One plausible reason for this is the earlier mention of bias in the dataset. The frame bias as discussed is only present in the melanoma images. When removing these images from the training set we assure that when the classifier is presented with an image with frame from the test set, it will not classify it as melanoma due to it being associated with the frame. The result suggests that the baseline classifiers may have a bias to classify any image with frame as a melanoma. Images with frames would either have to be added to the non melanoma class to offset this or as discussed throughout this thesis, the frames would have to be removed.

The classifier trained on the edited images resulted in the lowest AUC. This might be because of the entangled latent space. The StyleGAN framework was developed for feature rich images such as faces, as there are many more details in these types

	Real	Synthetic				Mixed
	Baseline	Baseline	No Bias	No Frames	Edited	Baseline
TP	227	209	186	212	167	226
FP	50	75	13	77	155	21
FN	28	46	69	43	88	29
TN	1578	1553	242	1551	1473	1607
ACC	0.958	0.935	0.911	0.936	0.871	0.973
TPR	0.891	0.820	0.700	0.831	0.655	0.886
TNR	0.970	0.954	0.945	0.952	0.905	0.987
AUC	0.982	0.960	0.920	0.954	0.870	0.986

Table 4.5: Test results from trained classifiers. The first 4 rows are the results from the confusion matrix. ACC is the test accuracy of the classifiers. True positive rate or sensitivity is the TPR and TNR is the specificity. AUC is the area under the ROC curve.

of images that are all related, the latent space will also be entangled. We found no systematic way of editing the images using latent directions that only directly implicated one feature while leaving the other ones intact. This makes sense in faces as a smile does not only stem from the mouth, but the eyes and many muscles in the face are involved in this expression. Therefore an entangled latent space is helpful for manipulating facial expression, although when removing a bias in a dataset it implicates more than just the one feature. Qualitatively when inspecting the edited images we thought they looked less like melanomas. 38% of the melanoma images have frames which could be the reason why the edited images look less like melanomas. Hence, the classifier is trained on fewer severe looking melanoma image samples. This suggests that the severeness of melanoma is entangled to the presence of frames in the images. We may have been training the classifier on non-melanoma looking images that were labeled as melanomas, that could be the reason for the low accuracy.

As it is important in a medical field to not send someone home that has a disease. We might wish to increase the sensitivity at the expense of having more false positive. This would in practice lead to more unnecessary surgeries to establish definite clinical diagnosis of lesions. Even so less people would go home being misdiagnosed as healthy. As discussed in the introduction the prognosis for the patients is better when the disease is discovered early [5]. Therefore it is crucial that we lower the threshold to increase sensitivity. Increasing the number of false positives does also come with a trade-off. We discussed the need for more accessible resources in the introduction, more false positives would lead to more biopsies and with that tied up resources.

The performance in our testing comparing the real and mixed result we see that the AUC is similar. Although looking at the ROC curve (Figure 4.17) we see that the mixed classifier performs with higher sensitivity at lower false positive rate. This result aligns with our initial statement in the introduction of classifiers being most

successful when trained on a large and balanced dataset [37]. Our results suggest that small patient datasets can be expanded with synthetic images generated from GANs trained on said small dataset. This leads to the expanded dataset improving the classifier achievement.

5

Conclusion

The aim of this research was to evaluate if we could generate data of skin lesions that qualitatively looked realistic enough to trick an expert into thinking they are real. As discussed in the previous section, this was accomplished as both dermatologists and deep learning experts were unable to reliably distinguish the synthetic data from that of real patients. The second question we were wanting to answer was if it is possible to train a CNN to classify melanomas. We know that a large and balanced dataset is important when training a CNN. Our results suggest that using synthetic data to expand a smaller dataset will improve the accuracy and sensitivity of the classifier. These discoveries led us to be confident that synthetic data is a good response to a lack of balanced data and data scarcity in a medical setting. We are left with one uncertainty and that is the reliability of the labeling. More experiments need to be performed as the experts' survey verdict did not have a high enough accuracy to conclude that the labels are representative.

StyleGAN2-ADA has proven to be a good framework when reaching for generation of realistic looking skin images for small datasets. What we discovered in our project is that the latent space is entangled. More work would need to be put into exploring and experimenting with the latent space in order to see if it may be possible to alter one feature while leaving the rest of the data in the image intact.

The threshold for the classifiers can be altered so that the algorithm acts with higher sensitivity in a "better safe than sorry manner" so that no melanomas are missed. With this implementation a next step could be to test the algorithm in clinic as a second opinion to the dermatologist's initial diagnosis. Eventually such algorithm could be deployed as a product, although this might come with some regulatory and technical difficulties. An important thing to contemplate is the unconscious bias that the algorithm has, since the dataset mostly consists of people with fair skin. The performance would have to be evaluated on all skin types before being launched as a product.

Bibliography

- [1] S. Kelly, A. Lupini, and B. Epureanu, “Data-driven modeling approach for mistuned cyclic structures,” *AIAA Journal*, vol. 59, pp. 1–13, 04 2021.
- [2] Phung and Rhee, “A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets,” *Applied Sciences*, vol. 9, p. 4500, 10 2019.
- [3] A. Karpathy *et al.* Generative models. [Online]. Available: <https://openai.com/blog/generative-models/>
- [4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” *CoRR*, vol. abs/1812.04948, 2018. [Online]. Available: <https://arxiv.org/abs/1812.04948>
- [5] A. J. Miller and M. C. Mihm, “Melanoma,” *New England Journal of Medicine*, vol. 355, no. 1, pp. 51–65, 2006, pMID: 16822996. [Online]. Available: <https://doi.org/10.1056/NEJMra052166>
- [6] “Cancer i siffror 2018,” *Social Styrelsen*, 2018. [Online]. Available: <https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/statistik/2018-6-10.pdf>
- [7] M. Gillstedt, E. Hedlund, J. Paoli, and S. Polesie, “Discrimination between invasive and in situ melanomas using a convolutional neural network,” *Journal of the American Academy of Dermatology*, 2021.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: <https://doi.org/10.1038/nature14539>
- [9] J. Brownlee, “How to classify photos of dogs and cats (with 97% accuracy).” [Online]. Available: <https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-to-classify-photos-of-dogs-and-cats/>
- [10] S. Debnath, D. P. Barnaby, K. Coppa, A. Makhnevich, E. J. Kim, S. Chatterjee, V. Tóth, T. J. Levy, M. d. Paradis, S. L. Cohen, J. S. Hirsch, T. P. Zanos, L. B. Becker, J. Cookingham, K. W. Davidson, A. J. Dominello, L. Falzon, T. McGinn, J. N. Mogavero, G. A. Osorio, and the Northwell COVID-19

- Research Consortium, "Machine learning to assist clinical decision-making during the COVID-19 pandemic," *Bioelectronic Medicine*, vol. 6, no. 1, p. 14, Jul. 2020. [Online]. Available: <https://doi.org/10.1186/s42234-020-00050-8>
- [11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661>
- [12] B. P. Jason. (2019, 06) A gentle introduction to generative adversarial networks (gans). [Online]. Available: <https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/>
- [13] (2022-03-13) Om sjukhuset. [Online]. Available: <https://www.sahlgrenska.se/om-sjukhuset/>
- [14] E. Piacentino, A. Guarner, and C. Angulo, "Generating synthetic ecgs using gans for anonymizing healthcare data," *Electronics*, vol. 10, no. 4, 2021. [Online]. Available: <https://www.mdpi.com/2079-9292/10/4/389>
- [15] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A gan-based image synthesis method for skin lesion classification," *Computer Methods and Programs in Biomedicine*, vol. 195, p. 105568, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169260720302418>
- [16] C. Baur, S. Albarqouni, and N. Navab, "Generating highly realistic images of skin lesions with gans," *CoRR*, 2018. [Online]. Available: <https://arxiv.org/abs/1809.01410>
- [17] G. M. Gonçalves, "A comparative study of data augmentation techniques for image classification: generative models vs. classical transformations." [Online]. Available: <http://hdl.handle.net/10773/30759>
- [18] "Chapter 4 - bio-inspired algorithms: principles, implementation, and applications to wireless communication," in *Nature-Inspired Computation and Swarm Intelligence*, X.-S. Yang, Ed. Academic Press, 2020, pp. 49–63. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780128197141000130>
- [19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015. [Online]. Available: <https://arxiv.org/abs/1511.06434>
- [20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, vol. abs/1710.10196, 2017. [Online]. Available: <https://arxiv.org/abs/1710.10196>
- [21] A. Liu, R. Tucker, V. Jampani, A. Makadia, N. Snavely, and A. Kanazawa, "Infinite nature: Perpetual view generation of natural scenes from a single image," *CoRR*, vol. abs/2012.09855, 2020. [Online]. Available: <https://arxiv.org/abs/2012.09855>

-
- [22] Synced. (2019) DeepMind DVD-GAN: Impressive Step Toward Realistic Video Synthesis. [Online]. Available: <https://medium.com/syncedreview/deepmind-dvd-gan-impressive-step-toward-realistic-video-synthesis-12027d942e53>
- [23] M. Huzaifah and L. Wyse, “Deep generative models for musical audio synthesis,” *CoRR*, vol. abs/2006.06426, 2020. [Online]. Available: <https://arxiv.org/abs/2006.06426>
- [24] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: <https://arxiv.org/abs/1411.1784>
- [25] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training generative adversarial networks with limited data,” *CoRR*, vol. abs/2006.06676, 2020. [Online]. Available: <https://arxiv.org/abs/2006.06676>
- [26] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, “The role of imagenet classes in fréchet inception distance,” *CoRR*, vol. abs/2203.06026, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2203.06026>
- [27] J. Brownlee. (2020, Aug) What is a confusion matrix in machine learning. [Online]. Available: <https://machinelearningmastery.com/confusion-matrix-machine-learning/>
- [28] “Siim-isic melanoma classification.” [Online]. Available: <https://www.kaggle.com/c/siim-isic-melanoma-classification/data>
- [29] “melanoma external malignant 256.” [Online]. Available: <https://www.kaggle.com/datasets/nroman/melanoma-external-malignant-256>
- [30] S. Carrasco, “Stylegan2-ada for generation of synthetic skin lesions,” <https://github.com/sandracl72/stylegan2-ada-pytorch>, 2022.
- [31] J. Seiler, M. Jonscher, M. Schöberl, and A. Kaup, “Resampling images to a regular grid from a non-regular subset of pixel positions using frequency selective reconstruction,” *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4540–4555, 2015.
- [32] S. M. Sandra Carrasco. (2022) Artificial Intelligence in Healthcare Part II. [Online]. Available: <https://medium.com/mlearning-ai/artificial-intelligence-in-healthcare-part-ii-157301b51c0f>
- [33] T. Karras, “Stylegan2-ada - official pytorch implementation,” <https://github.com/NVlabs/stylegan2-ada-pytorch.git>, 2021.
- [34] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “Ganspace: Discovering interpretable gan controls,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9841–9850. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/6fe43269967adbb64ec6149852b5cc3e-Paper.pdf>

- [35] Y. Shen and B. Zhou, “Closed-form factorization of latent semantics in gans,” vol. abs/2007.06600, 2020. [Online]. Available: <https://arxiv.org/abs/2007.06600>
- [36] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: <http://arxiv.org/abs/1905.11946>
- [37] S. Sharma, N. Yu, M. Fritz, and B. Schiele, “Long-tailed recognition using class-balanced experts,” *CoRR*, vol. abs/2004.03706, 2020. [Online]. Available: <https://arxiv.org/abs/2004.03706>

6

Appendix

6.1 StyleGAN2-ADA Network Architectures

6.1.1 Generator

Generator	Parameters	Buffers	Output shape
---	---	---	---
mapping.fc0	262656	-	[16, 512]
mapping.fc1	262656	-	[16, 512]
mapping	-	512	[16, 14, 512]
synthesis.b4.conv1	2622465	32	[16, 512, 4, 4]
synthesis.b4.torgb	264195	-	[16, 3, 4, 4]
synthesis.b4:0	8192	16	[16, 512, 4, 4]
synthesis.b4:1	-	-	[16, 512, 4, 4]
synthesis.b8.conv0	2622465	80	[16, 512, 8, 8]
synthesis.b8.conv1	2622465	80	[16, 512, 8, 8]
synthesis.b8.torgb	264195	-	[16, 3, 8, 8]
synthesis.b8:0	-	16	[16, 512, 8, 8]
synthesis.b8:1	-	-	[16, 512, 8, 8]
synthesis.b16.conv0	2622465	272	[16, 512, 16, 16]
synthesis.b16.conv1	2622465	272	[16, 512, 16, 16]
synthesis.b16.torgb	264195	-	[16, 3, 16, 16]
synthesis.b16:0	-	16	[16, 512, 16, 16]
synthesis.b16:1	-	-	[16, 512, 16, 16]
synthesis.b32.conv0	2622465	1040	[16, 512, 32, 32]
synthesis.b32.conv1	2622465	1040	[16, 512, 32, 32]
synthesis.b32.torgb	264195	-	[16, 3, 32, 32]
synthesis.b32:0	-	16	[16, 512, 32, 32]
synthesis.b32:1	-	-	[16, 512, 32, 32]
synthesis.b64.conv0	1442561	4112	[16, 256, 64, 64]
synthesis.b64.conv1	721409	4112	[16, 256, 64, 64]
synthesis.b64.torgb	132099	-	[16, 3, 64, 64]
synthesis.b64:0	-	16	[16, 256, 64, 64]

6. Appendix

synthesis.b64:1	-	-	[16, 256, 64, 64]
synthesis.b128.conv0	426369	16400	[16, 128, 128, 128]
synthesis.b128.conv1	213249	16400	[16, 128, 128, 128]
synthesis.b128.torgb	66051	-	[16, 3, 128, 128]
synthesis.b128:0	-	16	[16, 128, 128, 128]
synthesis.b128:1	-	-	[16, 128, 128, 128]
synthesis.b256.conv0	139457	65552	[16, 64, 256, 256]
synthesis.b256.conv1	69761	65552	[16, 64, 256, 256]
synthesis.b256.torgb	33027	-	[16, 3, 256, 256]
synthesis.b256:0	-	16	[16, 64, 256, 256]
synthesis.b256:1	-	-	[16, 64, 256, 256]
---	---	---	---
Total	23191522	175568	-

6.1.2 Discriminator

Discriminator	Parameters	Buffers	Output shape
---	---	---	---
b256.fromrgb	256	16	[16, 64, 256, 256]
b256.skip	8192	16	[16, 128, 128, 128]
b256.conv0	36928	16	[16, 64, 256, 256]
b256.conv1	73856	16	[16, 128, 128, 128]
b256	-	16	[16, 128, 128, 128]
b128.skip	32768	16	[16, 256, 64, 64]
b128.conv0	147584	16	[16, 128, 128, 128]
b128.conv1	295168	16	[16, 256, 64, 64]
b128	-	16	[16, 256, 64, 64]
b64.skip	131072	16	[16, 512, 32, 32]
b64.conv0	590080	16	[16, 256, 64, 64]
b64.conv1	1180160	16	[16, 512, 32, 32]
b64	-	16	[16, 512, 32, 32]
b32.skip	262144	16	[16, 512, 16, 16]
b32.conv0	2359808	16	[16, 512, 32, 32]
b32.conv1	2359808	16	[16, 512, 16, 16]
b32	-	16	[16, 512, 16, 16]
b16.skip	262144	16	[16, 512, 8, 8]
b16.conv0	2359808	16	[16, 512, 16, 16]
b16.conv1	2359808	16	[16, 512, 8, 8]
b16	-	16	[16, 512, 8, 8]
b8.skip	262144	16	[16, 512, 4, 4]
b8.conv0	2359808	16	[16, 512, 8, 8]
b8.conv1	2359808	16	[16, 512, 4, 4]
b8	-	16	[16, 512, 4, 4]
b4.mbstd	-	-	[16, 513, 4, 4]
b4.conv	2364416	16	[16, 512, 4, 4]
b4.fc	4194816	-	[16, 512]
b4.out	513	-	[16, 1]
---	---	---	---
Total	24001089	416	-

6.2 Survey

6.2.1 Survey Layout

5/22/22, 6:25 PM

Synthetic Data Evaluation

Synthetic Data Evaluation

The purpose of this survey is to evaluate the performance of a synthetic data generator for melanomas.

In the following survey there are 200 images. These images depict benign moles and melanomas. Some of the images have been generated by a computer program, i.e. they are synthetic, and some of them are real images taken from real patients. All images are shown with 256x256 resolution. The objective of the survey is to evaluate how realistic the computer-generated images are. The anonymized ratings may be shared with other researchers but any identifying information (e.g. your email address) will only be handled by the researchers on this project.


You will see the images one by one. Your job is to, for each image, indicate whether you think the mole in the image is benign or malignant and whether you think the image is a real photo or a computer-generated synthetic image. You will even rate how confident you are in your evaluation of the realness or fakeness of the image. You will do this on a scale from 1 to 5 where; ,1=not at all certain, 5=very certain in relation to your answer in question 2 (how certain are you that you knew if the image was fake or real). You can see some examples of the images similar to those you will be rating below.

Each image has three questions attached to it that all are in regards to the image:

1. Is this a melanoma?
2. Is the image synthetic or real?
3. How certain are you in your answer?

If you have any questions before you begin, you can reach us at annro@student.chalmers.se


anna_frosen@hotmail.com [Switch account](#)




* Required

Email *

XXX





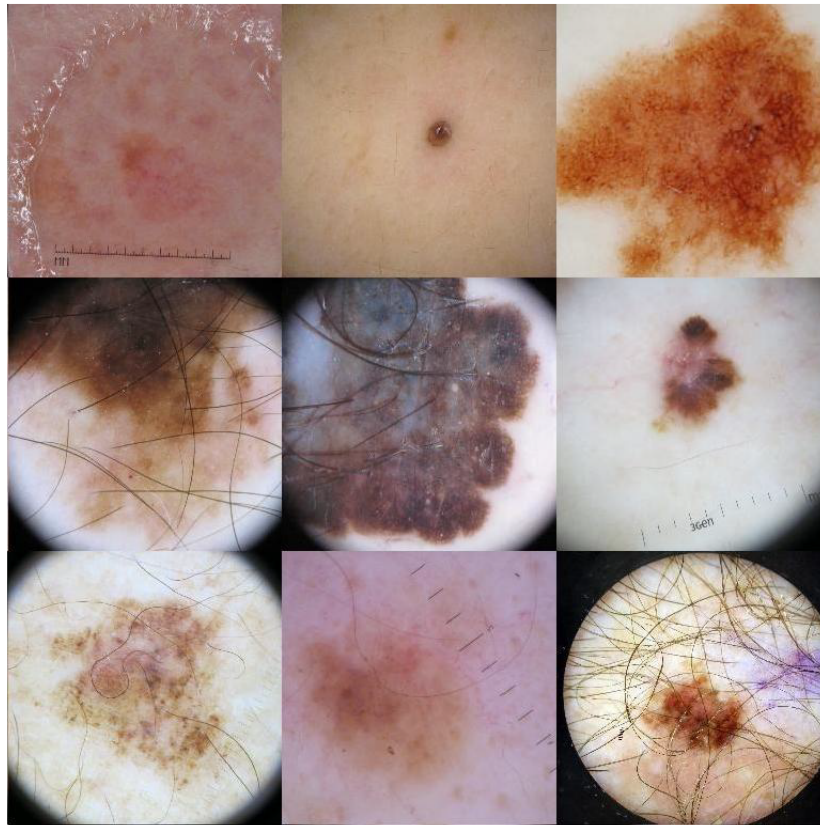
https://docs.google.com/forms/d/e/1FAIpQLSeOZgh55Y1OrK8upovWwBGEQ4Mv_FGuD8FsQnbK4Qp8Oz9Djw/view form

1/3

Figure 6.1

5/22/22, 6:25 PM

Synthetic Data Evaluation

Real samples of a melanoma skin lesionshttps://docs.google.com/forms/d/e/1FAIpQLSeOZgh55Y1OrK8upovWwBGEQ4Mv_FGuD8FsQnbK4Qp8Oz9Djw/viewform


2/3

Figure 6.2

5/22/22, 6:25 PM Synthetic Data Evaluation

Synthetic Data Evaluation

XXX [Switch account](#)



Medical background

- ☐ Dermatologist
- ☐ Non-dermatologist doctor
- ☐ Medical student
- ☐ Other medical personnel
- ☒ Lay person (i.e. non-medical)
- ☐ Deep learning expert

Clear selection

Page 2 of 203

Back



Next

Clear form

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. [Report Abuse](#) - [Terms of Service](#) - [Privacy Policy](#).

Google Forms




https://docs.google.com/forms/d/e/1FAIpQLSeOZgh55Y1OrK8UpovWwBGEQ4Mv_FGuD8FsQnbK4Qp8Oz9Djw/form/Response 1/1

Figure 6.3

5/22/22, 6:25 PM Synthetic Data Evaluation

Real samples of benign skin lesions



Page 1 of 203

Next **Clear form**

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. [Report Abuse](#) - [Terms of Service](#) - [Privacy Policy](#)

Google Forms

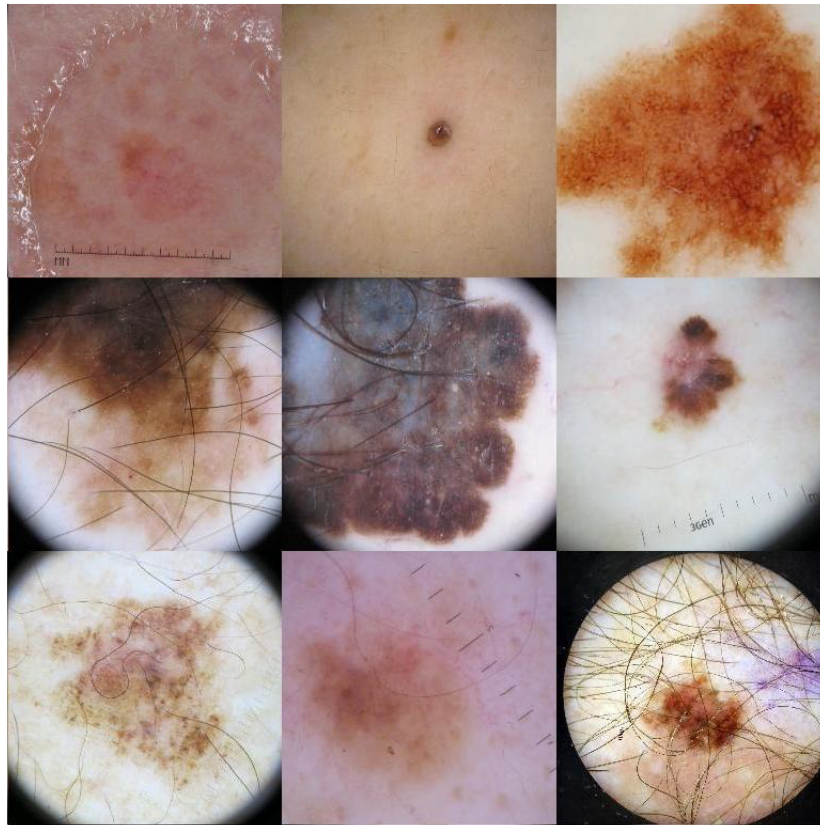
https://docs.google.com/forms/d/e/1FAIpQLSeOZgh55Y1OrK8upovWwBGEQ4Mv_FGuD8FsQnbK4Qp8Oz9Djw/viewform 3/3

Figure 6.4

5/22/22, 6:25 PM

Synthetic Data Evaluation

Real samples of a melanoma skin lesions



https://docs.google.com/forms/d/e/1FAIpQLSeOZgh55Y1OrK8upovWwBGEQ4Mv_FGuD8FsQnbK4Qp8Oz9Djw/viewform


2/3

Figure 6.5

5/22/22, 6:26 PM Synthetic Data Evaluation


Synthetic Data Evaluation

XXX Switch account



* Required

Mole 1



Diagnosis of lesion *


☒ malignant melanoma


☐ not melanoma

Is this image synthetic? *

☐ Yes

☒ No





https://docs.google.com/forms/d/e/1FAIpQLSeOZgh55Y10nK8upovWwBCEQ4Mv_FGuD8FsQnbK4Qp8Oz9Djw/fomResponse

1/2

Figure 6.6

5/22/22, 6:26 PM

Synthetic Data Evaluation

Level of certainty:

1

2

3

4

5

☐

☐

☐

☐

☐

Page 3 of 203

Back

Next

Clear form

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. [Report Abuse](#) - [Terms of Service](#) - [Privacy Policy](#).

Google Forms

Figure 6.7