Handling class imbalance by GAN based Data Augmentation in Medical Images

Amitkumar M Maheshwari

Final Report

Master of Science in Machine Learning and Artificial Intelligence

November 2022

Acknowledgement

First, I would like to show my gratitude towards Liverpool John Moores University, for providing me an opportunity to pursue the master degree and this research work. I'd like to express my gratitude towards Dr. Ahmed Kaky, Dr. Sahil Sharma, and Akash Choudhury for being so informative and guiding me towards right direction throughout the research period and helping me with my doubts. I'd also like to thank Mr. Glisten D'Souza for always being available for any unacademic help and support.

Lastly, I'd like to thank my family and loved ones for always being supportive, kept me motivated and understanding my absence for many occasions that allowed me to focus on this research work.

Abstract

Deep learning-based models have proven their strength in medical fields, especially working with medical images. In recent times, many open-source platforms collaborated with medical institutes and experts had attempted to address the fundamental obstacle of the lack of reliable training datasets by making the data available to the community with proper annotation. However, this attempt doesn't solve the other significant problem which is the lack of particular class(es) in the available training dataset. It is generally observed in medical images that some anomaly/abnormality/condition would occur very rarely in comparison with other cases. Such class imbalance impacts the performance of the models by leading the output to be biased towards the dominating class(es). The class imbalance issue isn't hidden from the research community and there has been fair enough research has been done to address the lack of training image by synthetically augmenting. Although in many cases of radiographic image datasets, successful image augmentation has been presented still in the case of camera-based or natural medical images that contain a high degree of variance in visual appearance and colors, the performance of synthetical augmentation is still not satisfactory. Also, traditionally used generative models often require high computational cost and consume too much time to be trained and showcase some degree of instability. This research is aimed to further improve image augmentation for camera-based medical images by using GAN-based image synthesis. The dataset augmented with the generated synthetic images delivers 0.57 sensitivity score, in comparison with the sensitivity score of 0.37 obtained by untreated dataset, it is 72.7% improvement. In addition to the improved classification, to reduce the GAN network complexity, computational cost, and bring robustness an autoencoder is integrated along with the GAN. The proposed method of integrated system of autoencoder and GAN contains 38.7 times less trainable parameters in the architecture and thus requires 78.8% less time to get trained in comparison with conventional GAN network. This research utilizes skin lesion dermoscopic images to train and validate image augmentation carried out using GAN variants where input to GAN is generated with the help of autoencoder. The augmented dataset is independently evaluated as well as the classification models trained on the dataset.

Table of Content

List of Figures

List of Tables

List of Abbreviations

ACGAN	Auxiliary GAN
AE	Autoencoder
BratS	Brain Tumor Segmentation
CBIS	Curated Breast Imaging Subset
CESRGAN	Cascade ensemble super resolution GAN
CNN	Convolutional Neural Network
CT	Computed Tomography
DCGAN	Deep Convolutional GAN
DDSM	Digital Database for Screening Mammography
EDA	
FCGAN	Face conditional GAN
GAN	Generative Adversarial Nets
GFV	Global feature vector
IS	Inception Score
ISIC	International Skin Imaging Collaboration
MA	Moving Average
MRI	Magnetic resonance imaging
PGGAN	Progressive GAN
RL	Reinforcement Learning
SNGAN	Spectrally normalized GAN
SPGGAN	Self-attention progressive GAN
TMPGAN	Texture-constrained Multichannel Progressive GAN
TTUR	Two timescale update rule
VAE	variational autoencoders
VGG NET	Visual Geometry Group Net
YOLO	You Only Look Once

1. Introduction

The scarcity of medical images is always a big challenge for researchers and machine learning professionals as in general, obtaining labelled medical images are extremely time-consuming and expensive in nature.

1.1 Background

Machine learning, especially deep learning based models and AI is continuously making their prominent place in modern-day medical science. From routine checks, to assisting in complex surgical operations AI solutions have been established as digital assistance to doctors and other medical staff. However, for better-performing models, a better training dataset is needed. An ideal training dataset should have sufficient and diverse enough training data. But in the medical domain, there are often cases of unavailability of training data, or even if the data is available, the number of positive cases of rare anomalies is very less in comparison with the number of negative cases which results in either overfitted or extremely biased detection/classification model. Often misclassification of any medical condition can be as bad as fatal, so it is important to develop an unbiased and reliable classification model. Additionally, medical experts are required to get the training data reviewed to label them. This process is manual, time-consuming, and cost inefficient. On top of that, it is highly dependent on the expertise of the medical professional and prone to human error.

In this research, ISIC 2020 skin lesion images (International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset., 2020) are used to demonstrate the issue of class imbalance. ISIC 2020 dataset is the dataset consist of dermoscopic skin lesion images. Table 1.1: Shows frequency of images in different types of skin lesions in ICIS 2020 dataset.

Table 1.1: Number of images per class in the ISIC 2020 dataset

Diagnosis	Count of diagnosis
atypical melanocytic proliferation	1
cafe-au-lait macule	1
lentigo NOS	44
lichenoid keratosis	37

melanoma	584
nevus	5193
seborrheic keratosis	135
solar lentigo	7
unknown	27124
Total images	33126

Figure 1.1: show the distribution of different cases of skin lesions. It can be clearly seen that the dataset is being dominated by two categories.

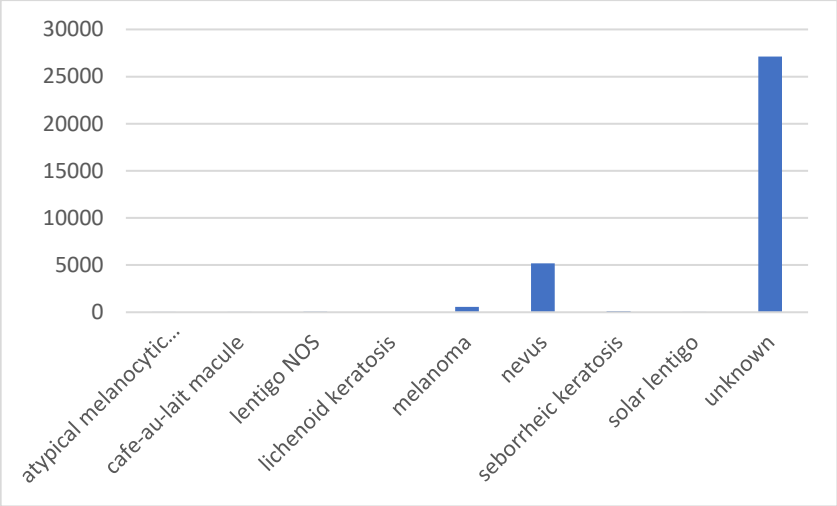


Figure 1.1: Class distribution in ISIC 2020 dataset

Class ‘unknown’ and class ‘nevus’ are highly dominating the entire distribution and it is obvious if this dataset is used to train the skin lesion classification model as is, the resultant model will be biased towards these two classes. The condition becomes too dangerous given the fact that ‘melanoma’ type skin lesion is critical to be detected especially when dermoscopy is the only reliable source of traditional detection as naked eye examination is proven to be less accurate (M E Vestergaard et al., 2008).

Two general approaches are there to handle class imbalance, under sampling and over sampling. Oversampling, the process of increasing the training data using data augmentation techniques (or just duplicating the data) is a more appropriate approach as just like the most cases of medical images, under-sampling of the two dominant classes to balance class distribution can’t

be the possible approach as it is observed in the Table , availability of the images in other classes are extremely less and an attempt to under-sample the dataset will result in underfitted model.
A combination of two independent deep learning based networks, one responsible for image generation and the other for image classification, interacting with each other can build an innovative image generation model (Goodfellow et al., n.d.). In their research, they proposed two deep learning models being trained parallelly, a Generative model G which learns the data distribution to produce the image as output and a Discriminative model D that takes the generated image as input and estimates the probability of the input image is from real training dataset rather than generated by G. Together both model can work as one unit that is capable of generating realistic synthetic images and it is known as generative adversarial nets (GAN). Figure 1.2: basic architecture of GAN shows the basic architecture of GAN.

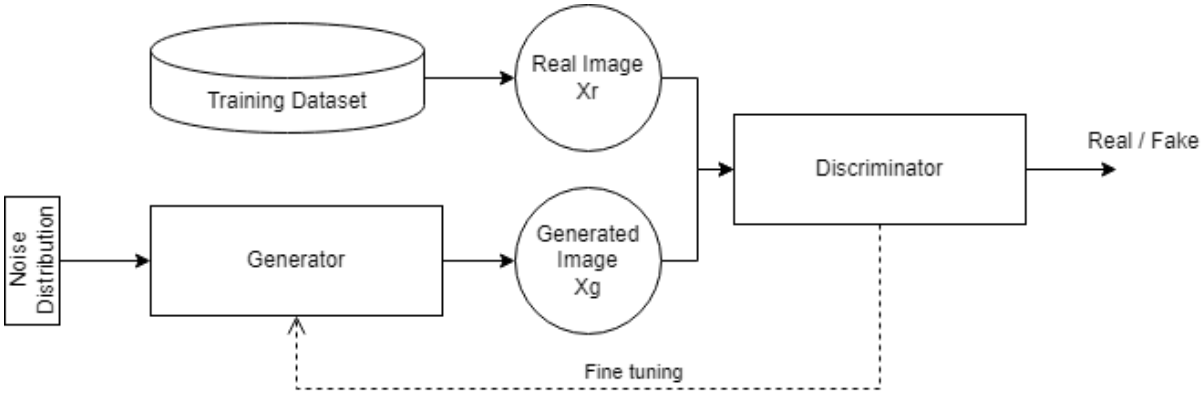


Figure 1.2: Basic architecture of GAN

Applications of GANs have a wide range in the computer vision field, there are many cases such as image augmentation, image registration, medical image generation, image reconstructions, and image-to-image translation where GANs are proven to be useful. Basic/Vanilla GAN has issues when working with high resolution images or more complex features like Mode collapse and gradient vanishing. Also, it performs limited on complex tasks such as image-to-image translations. Many researchers extensively worked on GAN to propose different variants of GAN to overcome the limitations of original GAN architecture like, AC-GAN to introduce the conditional operation, Progressive GAN to be able to progressively enhance the resolution of generated images, pix2pix GANs to be able to perform image to image translations and fusing segment of one image (or entire image) on other images to produce out of the box results.

1.2 Research Questions

On the bases of reviewing the prominent works of literature so far and by understanding the existing gaps and limitations of various techniques of medical image systemization, below mentioned questions are formulated that the current research will ultimately explore.

- Does class imbalance present in the dataset affect the outcome of the classification of skin lesion images?
- Does GAN based data augmentation help in creating a synthetic dataset for camera/dermoscopic skin lesion images that can improve classification performance?
- Does the autoencoder network can produced more suitable and low dimensioned input for generator to achieve early and stable convergence?
- Can reinforcement learning be applied on AE network to pick the best latent representation for GAN input?
- Does the skin lesion dataset generated by GAN based data augmentation outperform the dataset generated by traditional image augmentation techniques?
- For the classification of skin lesion images, does the model train on data augmentation perform better than the model train on data anonymization?

1.3 Aim and Objective

The main aim of this research is to develop a stable GAN model that can generate reliable synthetic medical images. Also, this research is aiming to address existing common issues of GANs in images generation like early convergence and robustness in the architecture and smarter way to generate an input for the generative architecture. The skin lesion dataset is highly imbalanced and biased, the end goal is to be able to generate synthetic images for a specific class(es) to handle the class imbalance present in the dataset that ultimately results in better trained and reliable classification models.

To achieve the aim following objectives are formulated:

- Proper EDA on the dataset, identify and eliminate any error/impurity in the dataset, and perform the image preprocessing to normalize the images and bring them to a uniform size
- Training the autoencoder system that can pick the best input latent representation for GAN network

- Developing and training the GAN networks using different techniques to identify the most suitable GAN with early feedback loop breaker mechanism, based on the nature of the given dataset
- Developing and training a stable classification models being trained on the augmented dataset.
- Evaluating the performance of GAN and classification models using proper metrics and derive the inference.

1.4 Scope of the study

This research work needs to be well defined and well directed towards the mentioned objective. To keep the research focused and feasible to be completed in given time duration, the scope of the research work has been limited as below:

- This research will explore only two approaches to image augmentation, traditional image transformation, and GAN based image synthesis.
- Only noise-based Image generative GANs will be explored and only DC-GAN and Style-GAN variants will be further implemented for image augmentation. Image translation-based GAN techniques are not included in the research and so does the image segmentation.
- This research will explore the possible applications of autoencoders network to support the GAN network.
- Reinforcement learning will be exploited with autoencoders to learn the policy that can pick the best input for generator network of the GAN.
- The classification models are only meant to evaluate the dataset balanced by image augmentation techniques and further improvements of the classification models are not in scope.

1.5 Significance of the study

This research is contributing to the synthetic medical camera image generation by using different variants of GAN models to handle the ‘class imbalance’ problem in dataset and scarcity of training images which leads to poor performance of classification models. Dermoscopic skin lesion images are selected to be used in this research as in this dataset, images

are camera-based images and demonstrate extreme class imbalance. Among all types of skin cancers, ‘melanoma’ is the most lethal one thus it becomes very critical for medical science to have a stable and reliable melanoma detection mechanism as early diagnosis can greatly improve the survival rate of patients.

‘melanoma’ is one of the classes of skin lesions in the dataset which is being shadowed by the dominating class ‘melanocytic nevus (nv)’ the classification models benign trained on such biased datasets mostly perform poorly in melanoma detection. This research is aimed to overcome this issue by oversampling the minority class (here ‘melanoma’) with synthetic images of the melanoma class generated by using GAN.

In addition, a generic GAN model will not only help in balancing the skin lesion images but can also be utilized in generating other camera based medical images like surgical images of rare conditions or endoscopic images of anomalies found. This research will also open gates for further extended research to develop GANs that can be used domain agnostically.

1.6 Structure of the study

In this section, a basic outline of the current thesis report with a brief information regarding the context and content is presented.

Chapter 1:

Chapter 1 talks about the background and motive of the current research work. An Introduction to skin lesions and a well-known dataset for skin lesions ISIC is provided. This chapter provides a sense of gravity of the problem that this research is attempting to address. Further the aims and objective of the current research and research questions that current research is going to explore are given and means to achieve the objectives are discussed. Overall scope is defined.

Chapter 2:

Various research works are discussed in this chapter. Research works are picked and discussed based on their relevance to the domain of the problem statement, methodology used, or both. An attempt to provide a sense of evaluation of the research work doing in the area of medical images synthesis and various GAN variants. Further this chapter discuss around the research work that has tried to apply of autoencoder and reinforcement learning with GAN.

In second half of the chapter, a systematic summary and comparison between relevant and prominent research work is presented. Further the challenges and gaps present in the research work are discussed. Also, advantages and disadvantages of the referred research with respect to current research has been formulated.

Chapter 3:

This chapter talks about the methodologies that are going to be applied in the research work and experiments. A detailed analysis of dataset and pre-processing steps are discussed. A brief discussion on overall process flow is given. Further in the chapter all steps of flow chart are discussed in detail. All the steps are discussed separately and at the end a holistic view is present where all are demonstrated as an integrated system. At last means of evaluations are discussed.

Chapter 4:

This chapter extends the chapter 3 and consists various experiments around the defined methods and data. Starts from data understanding and pre-processing steps, further it talks about various image transformation techniques. Detailed understanding the experiments regarding the generative models like autoencoders and GANs are then discussed, and outcome of the experiments are analyzed.

Chapter 5:

Finalized methods from chapter 5 are further analyzed in line with research objectives and the questions this research has defined and looking to be answered. This chapter talks about the results of the finalized methods and discuss the derived inferences. Later the classification results are also analyzed based the augmentation performed on the dataset. This chapter prepares data that is required to conclude the research.

Chapter 6:

Based on the understanding, observations and analysis done during previous chapters, this chapter tries to answer the research questions and address the defined objectives. This chapter concludes the research with resultant numbers and intuitions behind the numbers. Also, this chapter talks about the findings gathered during the research work that adds up in the knowledge base in the domain and provides possible direction to extend the work.

2. Literature Review

There has been significant research work done in the area to understand skin lesions and understanding different types of skin lesions. However, this study specifically talks about different types of skin lesion images and how to synthetically generate such images to help the dataset be balanced. Further is discussed research works carried out in this field.

2.1 Introduction

Skin lesion images, just like other rare anomaly datasets are extremely imbalanced and biased toward one or more classes over other classes. This becomes a major issue in the classification of different skin lesions. This study is focusing on addressing this issue by studying different means of generating synthetic skin lesion images for less occurring classes to balance the dataset.

Many research works in the area of skin lesion, different means of synthetic image augmentations, and their impact on classification task has been studied in this research work. Different techniques based on deep-learning models are proven to be a significant help and are discussed in the following sections.

2.2 Related work done

While there are many approaches proposed to handle data imbalance like down sampling, and data augmentation using traditional ways, Generative models, especially Generative Adversarial Nets has shown the most promising results in terms of generating realistic images. This section of the study explores some of the significant and state of art research done in the field of synthetic image generation using generative models.

2.2.1 GAN origin and variants

After Goodfellow and his team introduced the concept of Generative Adversarial Nets (GAN) (Goodfellow et al., n.d.) it had opened a new door in the field of synthetically image generation, and soon it become an area of interest for many researchers working in the domain of computer vision, and deep learning and a lot of work has been done in this field so far. Although it was introduced in 2014 a solid trend of using GAN variants to generate synthetic images to be used in other deep learning networks as input can be seen in recent years.

Two major types are seen when talking about GAN, image generative, and image-to-image translation. The main difference between these two is where normal image-generative GANs learn data distribution of the actual dataset and use random input to generate realistic images, in image-to-image translation, GANs paste a particular segment of one image into another image to make the target image containing specific features. Numerous different variants of GANs are already introduced ever since the original concept was proposed in 2014.

However, talking about some of the State of the Art or significant studies using different types of GANs. Research works (Waheed et al., 2020; Srivastav et al., 2021) demonstrate two most basic GAN variants AC-GAN and DC-GAN respective, applied in radiological image synthesis. While the F-CGAN, a two-staged conditional GAN proposed in (Fu et al., 2020) works on image-to-image translation style instead of noised based image generation. F-CGAN showcased a significant improvement in generating fine-grained images when compared with previously acclaimed AC-GAN, and SNGAN and the classification models trained on the dataset generated by FCGAN showed better accuracy than the standard model and SNGAN model.

GANs are prone to have mode collapse issue and thus be very unstable while working with higher resolution images and this issue had very much limited its applications in several areas. Progressive GAN, one of the import variants of GAN has addressed this issue and made it possible to generating high resolution synthetic images. An extension of progressive GAN (PGAN), (Guan et al., 2022) have proposed a method of GAN based image augmentation "texture-constrained multichannel progressive GAN (TMPGAN)". The objective was not to handle class imbalance but to generate synthetic images to overcome the issue of less training images available. TMP-GAN applies a progressive generation mechanism that improves image synthesis steadily. Foreground-Generation method is being used in it, which means the model will generate the synthetic lesions in selected areas of normal/actual images to produce positive case images.

In other study (Dumagpi and Jeong, 2021) researchers have used DC GAN for image generation and Cycle-GAN for image translation in addition to traditional image transformation (shown in figure 2.1: Traditional(left) and Generative(right) techniques of Images Augmentation).

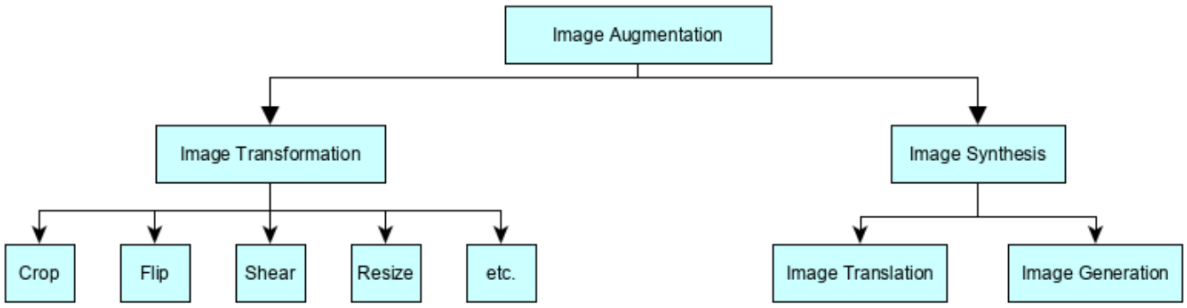


Figure 2.1: Traditional(left) and Generative(right) techniques of Images Augmentation used in (Dumagpi and Jeong, 2021)

In Cycle-based GAN combined with YOLO (you only look once) architecture (Hammami et al., 2020), instead of one set of generators and discriminator, Cycle GAN is made of two sets and works as bidirectional image translation. The output of the Cycle GAN is then fed into YOLO for detection. YOLO style classification makes it works faster than normal variants of GANs.

In another innovative study (Qasim et al., 2020) researchers talk about the class imbalance issue. To achieve the image segmentation task, unlike the traditional GAN where two components, Generator and Discriminator would compete, they introduced a SPADE based GAN with third component called "Segmentor" (Figure 2.2: Basic representation of Red-GAN.) which is fixed and pretrained on the same dataset to obtain the synthetic image segments on the fly.

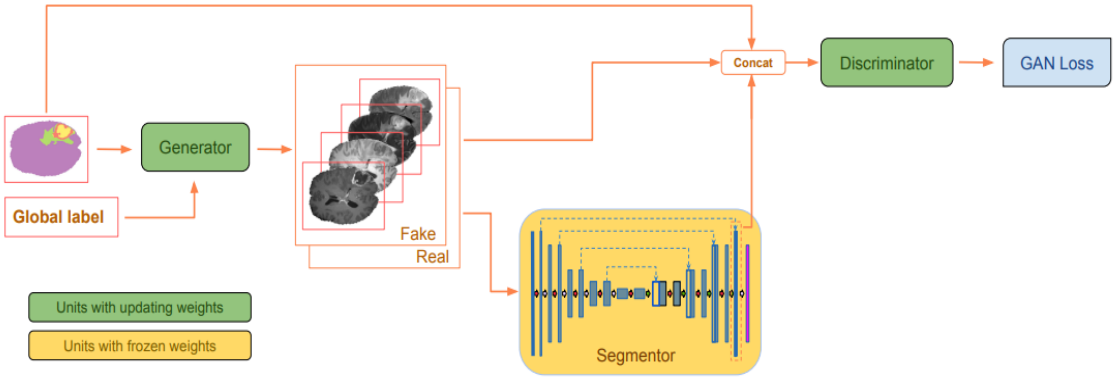


Figure 2.2: Basic representation of Red-GAN. Here we can observe the third component pre-trained "Segmentor" being introduced (Qasim et al., 2020).

(Abdelhalim et al., 2021) proposed self-attention progressive growing GAN (SPG-GAN) combined with two-timescale update rule (TTUR) which shows better stability in comparison with Big-GAN (Brock et al., 2018) while still achieving higher resolutions. TTUR in the architecture decupled the learning rate between generator and discriminator to avoid unhealthy competition between these two components making them independently moving towards optimum loss. When a group of researchers working on to reconstruct super resolution images from low resolution images felt that the single set of GAN might not be sufficient and instead of exploring cycle-based GAN, a study (Shahsavari et al., 2021) was proposed sequential GAN, CESR-GAN – Cascade Ensemble Super Resolution GAN. Figure 2.3: Shows the proposed cascaded GAN architecture. Gates provide much needed flexibility as it provides the decision making if flow is needed to go in further GAN or current result is good enough.

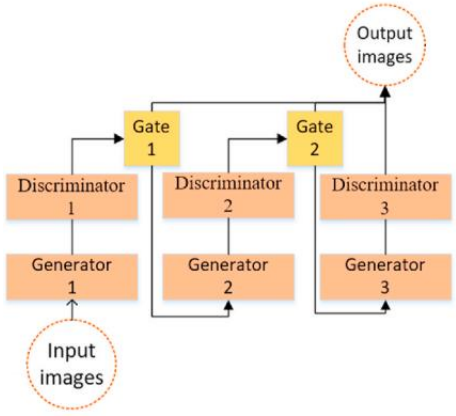


Figure 2.3: Cascading architecture of GANs as proposed in (Shahsavari et al., 2021)

Equation for the gates in CESR-GAN,

$$Q(x) = 1 \; if \; D(x) > t \; else \; 0 \tag{1}$$

where “D(x)” is discriminator’s value and “t” is a threshold.

GANs in studies (Dumagpi et al., 2020; Dumagpi and Jeong, 2021), have been put to generate synthetic images of positive subway security threat X-ray images to balance an extremely unbalanced dataset. While evaluating they noticed that combining all three types of synthesized images can make the classification model generalized enough to bring significant improvement in average precision.

This shows that GANs have applications from normal object classification to as critical as subway and airport security by improving the performance of the classification model. The other field where GAN has been proven to play an important role is Bio-Medical images generation. In further section, the applications of these GAN variants in medical fields are discussed.

2.2.2 GAN for Medical Images

Talking about medical images, most research has been done on radiographic images like X-rays, CT, MRI, etc. while on natural or camera images we can see there was comparatively less focus.

There is a fundamental domain difference in medical images in comparison with other images be it camera images or radiographic images. Deep learning based models like classification model or segmentation model, would, in general, look for certain types of anomalies and in many cases, such anomalies would display very delicate texture or color differences thus Image synthesis for medical images must be sensitive enough to learn such delicate distribution and produce images that contain due features properly.

A study (Frid-Adar et al., 2018) explored two very basic variants of GANs and those were DCGAN and ACGAN. Unlike DCGAN, ACGAN is a conditional GAN and as external conditional information, ACGAN provides class information in the GAN network. Trained on liver CT images for lesion segmentation, their study not only demonstrates the performance improvement but also compares the performance difference of classification when the model is trained on traditional image augmentation and GAN Based augmentation.

While most research related to data augmentation using GAN variants were focused to overcome the scarcity of the data itself, the main challenge in medical images is imbalance dataset. There were some researches focused on the challenge of data being extremely biased towards certain class(es) and the rest classes would rarely occur. Red-GAN (Qasim et al., 2020) had addressed this issue using highly imbalanced datasets, BraTS and ISIC. (Hammami et al., 2020) A cycled based GAN was used to generate synthetic MRI images to be used to train a multi organ detector mode.

Traditionally GAN are not designed to preserve all the textures which CBIS-DDSM screening images displays. To overcome this limitation, the TMP-GAN (Guan et al., 2022), basically an

image-to-image transaction GAN, has specifically designed to take care of the most delicate texture of images while pasting the segmented lesion part on target image. A progressive fusing mechanism makes sure that the synthetic lesion's continuity on the background to preserve the textures.

The other and more significant challenge in training deep learning models for medical images is the desired images are either very less to train the model on or they are extremely unbalanced as most cases would fall in normal/negative class.

A study, proposed in April and Published in May of 2020, merely a couple of months after covid was declared a worldwide pandemic and with an obvious heavy shortage of training images for positive cases, AC-GAN has been put in use for Synthesizing both Covid CXR and normal CXR images to train a classification model for covid detection (Waheed et al., 2020) . On other hand, instead of Image Translation (AC-GAN), (Srivastav et al., 2021) has achieved significant improvement in pneumonia detection by augmenting positive images using image generative GAN model – DC-GAN. However, both studies were not focusing on the "Class Imbalance" issue which is very common across the medical domain.

While GANs seem to be working good for radiographical medical images, they face difficulty in generating natural RGB images. Dermoscopic images of skin lesion images are different than gray scaled / radiographic images.

2.2.3 GAN used for skin lesion images

To obtain a reliable GAN based image synthesis on skin lesion images, a study, (Bissoto et al., 2021) reviewed 18 prominent research that claimed of gaining significant improvement in the model for classification or segmentation tasks that were trained on GAN based synthetic images. Further, their study has validated how different real: synthetic image ratio leads to a different outcome. Researchers tried four different GAN variants: SPADE, pix2pixHD, PGAN, and Style GAN to generate synthetic images and trained classification model Inception v4 with the generated training dataset using various real: synthetic image ratios. Researchers then went ahead and compared two basic techniques of utilizing the synthetic images in the classification model, Augmentation and Anonymization. However, in any terms, they could not achieve as good results as it was claimed in the referred papers.

One common trend that has been noticed in (Bissoto et al., 2021) and (Qasim et al., 2020) is that both were not able to perform well for the skin lesion dataset, while Red-GAN could perform reasonably okay for the brain tumor dataset. The concluded reason for these GANs' inability on performing better was, that "skin lesion images have a more visual appearance in comparison with brain tumor MRI images (or other radiographic images), thus image segmentation and mask to image mapping become more difficult in comparison with MRI images". And this opens a large gap for GAN based image synthesis for camera images and the reason given above, it should not be limited to skin lesion images but other medical images like surgical images or endoscopic images as well.

Other than radiographic images, studies had been carried out on rich in color and texture microscopic images of human protein where DC-GAN has been applied (Verma et al., 2020) and on dermoscopic skin images (Litjens et al., 2017; Rashid et al., 2019; Qin et al., 2020; Bissoto et al., 2021) where a different variant of GANs has been used for image augmentation. However, none of them focused on handling class imbalance, and only (Bissoto et al., 2021) tried and failed to improve the ultimate classification model. Although modified Style-GAN has provided promising results for skin lesion image generation (Qin et al., 2020)

Moving further from style-based GAN, some other studies show promising results on different skin lesion images. When evaluated (Abdelhalim et al., 2021) on HAM10000 dataset, the output 256x256 images for classification using Res-Net18, this variant of GAN showed higher sensitivity in comparison with other means of image augmentations. The other innovative study (Shahsavari et al., 2021) also demonstrated super resolution skin lesion image generation that resulted in improving SSIM, FSIM of the output. But this study only talked about enlarging the resolution of the images but not to use it in detection or segmentation task. A comparative study (Reddy Alasandagutti, 2021) that studied different techniques of image processing, machine learning models, feature extraction techniques comparing not only performance but also time and space complexity.

Using one VAE, two GANs, and auxiliary classifier, the study of Heavy-Tailed Student T-distribution in GAN, TED-GAN (Ahmad et al., 2021) has shown significant improvement in classification task by generating realistic looking skin lesion images. One major difference noted in (Ahmad et al., 2021) is instead of using complete random vector as latent space for generator's input, researchers has used variational encoder whose sole role is to prepare most suitable input for generator.

In further section, Auto Encoders are discussed in a brief.

2.2.4 Autoencoders

Autoencoding is way of learning the data representation. Auto Encoders, originally proposed in (Yann Lecun, 1987) are made of mainly two components, Encoder (E) and Decoder (D). The encoder component is responsible to map the input to a latent representation and on other hand decoder is used to re-construct the input from the latent representation generated by encoder. Encoder when convert the input into latent representation, it also reduces the dimensionality of an input making the whole process less resource efficient and more stable. The main aim of the autoencoder is to make the reconstruction error as low as possible.

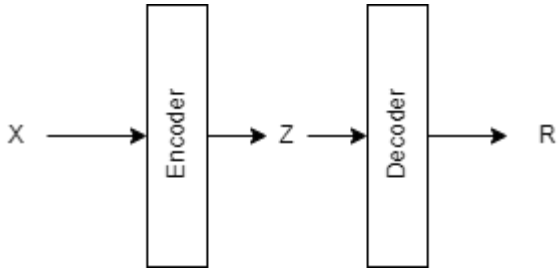


Figure 2.4: Basic structure of autoencoder(Zhai et al., 2019)

Basic structure of autoencoder is shown in the figure 2.4 where we can see two main components Encoder and Decoder also, we can see data flow in between them. X represents the input while Z is the reduced latent space generated by Encoder. Aim of Decoder is to reconstruct the original input form Z, i.e., R~X.

One among many researches done on top of original idea of autoencoder, in (Hinton and Zemel, n.d.), researchers have proposed an objective function based on “Minimum Description Length (MDL) to minimize the information required to describe the intermediate latent representation and the reconstruction loss. This will further reduce the computational cost of the overall operation. Fairly well researched and existing from long years, autoencoders independently are well used to learn data distribution, dimensionality reduction, and extracting features. VAE (Kingma and Welling, 2013) for an example. However, in recent years some innovative studies are seen where Auto Encoders are used beforehand with GANs, and results seems to be very promising any aspect of efficiency.

Normally in generative GANs, random noise is as input for generator and after learning from the loss for many iterations, generator becomes capable to generate an image which discriminator cannot differentiate from the original input images. However, in the study (Sarmad et al., n.d.), researchers have used an intermediate latent representation of the partial input point cloud to train the GAN for successfully generating the point could output that resembles the original dataset. Similarly, in (Ukwuoma et al., 2021) autoencoder is used to generate GFV from partially painted images. This GFV does have some amount of noise as the input images are not fully and properly painted. This noise GFV is used as input in the GAN and object of the GAN is to generate the clean most GFV which can be decoded back into fully painted images.

However, unlike (Sarmad et al., n.d.) and (Ukwuoma et al., 2021), in the case of generating synthetic medical images to address the issue of limited dataset, "partial" input won't be available to generate the latent representation for GAN input. A more relevant study (Ahmad et al., 2021), researchers have addressed this limitation by having a pre-trained autoencoder and swapped it while using it with GAN. In either way, researchers have found that using autoencoder helps in early convergence.

The main benefit of autoencoders with GAN is observed to reduce the complexity of the GAN architecture by reducing the dimensionality which ultimately results in faster convergence. One more technique that can be embedded with GAN is reinforcement learning.

2.2.5 Reinforcement Learning in GAN

Reinforcement learning is an area of machine learning in which the agent aims to learn to take the best possible action given in the state in environment. Meaning, the reinforcement learning is an ability to learn the decision making to obtain the best possible outcome (reward).

As in a way, (Reddy Alasandagutti, 2021) suggests that the GANs themselves are a functioning as RL and in GAN architecture, there isn't any area where it is needed to pick the best action among many possible actions, most of the research on GAN haven't applied of RL in any aspect of GAN. In fact, GAN being used in RL application is comparatively more intuitive than the other way around. Thus, RL applied in the GANs are very less explored. This makes it a relatively new area to research into.

Counting few studies where RL has been used with GAN, (Sarmad et al., n.d.) that mainly aiming to provide fast and robust control over GANs. Researchers has used an actor-critic based architecture RL-GAN-Net to learn the policy in continuous action space. Overall architecture of RL-GAN-Net consist of autoencoder, latent space GAN, and RL agent. From reinforcement learning perspective, environment is "shape completion framework" which composed of blocks as AE and GAN while actions are possible inputs to the generator. Instead of generating the whole image, generator will generate GFV which can be decoded.

With a similar aim as (Sarmad et al., n.d.), another study (Ukwuoma et al., 2021) with the aim of inpainting the image completely where the input to the network be partially painted images or recovering the distorted input images areas. The objective of RL in (Ukwuoma et al., 2021), to pick the correct GAN input to get image latent space representation that is most suitable for the input of the distorted/partially painted images. Study has applied LGAN with RL and AE has been used to generate input to the LGAN. RL in the architecture is responsible to pick right seed z that is used by LGAN to generate clean GFV. Unlike to normal GAN where generator targets to generate realistic images which resemble to the input image, with the whole different aim of manipulating the age of input face images to generate high fidelity face image with targeted age manipulation (Shubham et al., 2021) has used RL to learn nonlinear trajectory.

On other hand, (Rahmayanti et al., 2021) proposed a sketch generating application, a system to generate sketches from the real-world images using the proposed method "Doodle with stroke demonstration and deep Q-Network". As name suggest, a deep Q-Learning to pick the right actions has been included with conventional GANs.

All the discussed studies in this section served different purpose a major similarity in them is they all used RL to pick up the most optimum input for generator. Studies demonstrated that even though the RL wasn't directly applied into any component or execution flow of GAN itself, a slight scope of RL resides in autoencoders which can be used to generate the input for GAN.

Given this, further sections discuss the comparison in methods, data, evaluations strategies of different studies and a brief pros and cons of them w.r.t the current proposed research.

2.3 Discussions on prominent studies

Out of all the research works studied, in below tables some of the prominent studies are discussed which are more relevant to current research area and objectives.

2.3.1 General overview

Table 2.1: summarizes and compares different prominent and relevant studies by applied methodology, dataset domain, objectives, and outcome. Also, from the table a clear trend of GANs and their applications can be understood.

Table 2.1: Overview of studies that are more relevant to current research

Research work	Method	Dataset used	Objective	Evaluation strategy	Outcome
(Rashid et al., 2019)	Semi-Supervised GAN, ResNet, Dense Net	ISIC 2018	To obtain realistic dermoscopic images using GAN to augment into training dataset to enhance classification result	Precision, Recall, F1-Score	Slight improvement in classification accuracy obtained
(Waheed et al., 2020)	CovidGAN (based on ACGAN)	IEEE covid chest x-ray dataset, Covid19 Radiography dataset, Covid19 chest x-ray dataset initiative	To generate synthetic chest x-ray images as the covid outbreak was recent and relevant images were not widely available	Precision, Recall, F1-Score, Accuracy, Sensitivity, Specificity	Accuracy increased by 10%

(Qin et al., 2020)	Skin lesion style-based GAN, ResNet50 transfer learning	ISIC 2018	To improve skin lesion classification performance by addressing the issue of scarcity of labeled data and class imbalance	Accuracy, Sensitivity, Specificity, Avg. Precision, Balanced multiclass accuracy	Improved by Acc – 1.6% Sensitivity – 24.4% Specificity – 3.6% Precision – 23.2% Bal. Multi Acc – 5.6%
(Fu et al., 2020)	F-CGAN - two-staged conditional GAN	CUB Birds, Stanford Dogs dataset	To generate class dependent fined grained detailed images	IS FID	Comparing with ACGAN and SNGAN, IS increased, However, for FID, SNGAN works better.
(Verma et al., 2020)	DC-GAN, VGG16, NASNet Mobile, ResNet50, Inception V4	Human Protein Atlas Image Classification Kaggle Dataset	To generate synthetic samples to improve the classification models of human protein images as there is an extreme need of an automatic system to evaluate them.	Macro F1, Micro F1, Accuracy	Gradual and steady improvement (around 2-3%) in classification for different classification models
(Dumagpi et al., 2020), (Dumagpi and Jeong, 2021)	GAN based anomaly detection: Bi-GAN, SVM	SIX-ray. (Dumagpi and Jeong, 2021) has extended	To address an extreme class imbalance issue in case of security x ray image dataset	Precision, Recall, F1-Score	Overall fair enough improvement in classification task.

		the dataset further			
(Srivastav et al., 2021)	DC-GAN, VGG16	Labeled Optical Coherence Tomography (OCT), Chest X-ray Images	To augment synthetic images to oversample the training dataset to improve the classification model performance	Accuracy	A small improvement is observed in classification accuracy.
(Qasim et al., 2020)	Red-GAN – a SPADE based GAN with third component 'segmentor'	BraTS, ISIC	To mitigate the limitation of scares data regimes in segmentation task	Dice Score	On the fly segmentor component in Red-GAN didn't improve significant in Dice Score. However, the concept of oversampling has been proven.
(Hammami et al., 2020)	Cycle GAN Multi organ, detection: YOLO	Visceral anatomy benchmarks dataset	CT image augmentation using MRI images to enhance the dataset so that multi organ detection task can be improved	Mean average distance	With augmented dataset, significant better detection is observed
(Guan et al., 2022)	TMP-GAN	CBIS-DDMS	To be able to synthesize the images that can preserve the	Precision, Recall, F1-Score	Around 2-3% improvement in all the evaluation

			delicate textures in medical images to improve the classification		matrix for both datasets.
(Frid-Adar et al., 2018)	Comparative study of DC-GAN and AC-GAN	Liver lesions from Sheba Medical Center	To demonstrate the application of GAN for data augmentation and to improve classification performance	Accuracy, Specificity, Sensitivity	Augmented dataset is observed to perform better in classification. DC-GAN performed better than AC-GAN
(Abdelhalim et al., 2021)	SPG-GAN, TTUR	HAM10000	To apply GAN to generate realistic but completely different skin images	Sensitivity	13.8% of improvement is observed in sensitivity of melanoma class
(Shahsavari et al., 2021)	CESR-GAN	ISIC	To reconstruct the super resolution images from lower resolution images	SSIM, FSIM, PSNR	Significant improvement is observed in comparison with existing Variants
(Ahmad et al., 2021)	TED-GAN	HAM10000	To generate skin images that look realistic enough and help in improving the classification	Precision, Recall, F1 Score Accuracy	Improvement is observed in all evaluation matrix results in comparison

					with GAN, DeLiGAN
(Sarmad et al., n.d.)	RL-GAN-Net	ShapeNet Point Cloud	To train RL based GAN to validate if it can successfully learn to complete partially complete point cloud shapes	Chamfer Distance, Accuracy	In comparison with normal input and using AE, RL-GAN-Networks constantly better
(Shubham et al., 2021)	PGAN with RL	CelebA-HQ dataset	To manipulate the age of input face images to generate high fidelity face image with targeted age manipulation	Cosine similarity score	Better scores are observed in case of RL + PGAN
(Ukwuoma et al., 2021)	LGAN with RL, RLG Net	Celeb Faces Attribute, The street view house number, Stanford cars. ImageNet	To train RL to pick the correct GAN input that is most suitable for the GAN to successfully imprint distorted or partially painted images.	Accuracy	Even being real time, RLG-Net is observed to demonstrate the boost in accuracy

On other hand, Table 2.2: shows an overview of the review papers referred in this research.

Table 2.2: Overview of review papers

Research Work	Methods	Dataset	Evaluation Strategy
(Singh and Raza, 2020)	DC GAN, LAP GAN, Pix2pix, Cycle GAN, Unsupervised Image translation (UNIT)	Cancer Imaging Archive (TCIA), National Biomedical Imaging Archive (NBIA), Radiologist Society of North America (RSNA), Biobank	NA
(Bissoto et al., 2021)	PGAN – VGG19 Pix2pixBased – Mobile Net DC GAN – LeNet5, AlexNet, MUNIT – ResNet50 Pix2pixHD – InceptionV4	TCGA, OVCARE Private clinical images MICCAI 2016 BraTS 2016 ISIC 2018	GAN: FID Classification: AUC

2.3.2 Pros and Cons

Main advantage of GAN itself is also a motivation for the current research work and that is its ability to artificially synthesizing the data. However, different variants of GAN have their own set of advantages and disadvantages in comparison with each other. In this section, several prominent works are picked up and discussed their pros and cons keeping the context and motive of current research work in mind.

Talking about general plus points and pitfalls of GAN, GANs are proven to be very effective in oversampling the dataset and thus they help in reducing the biasness in dataset and prevent overfitting in the model that ultimately increase the performance of classification or detection models. On other hand, the amount of time required to train a GAN model is significant. In addition, they require higher computational power and are prone to mode collapse and instability as the network becomes complex. The higher the image resolution more the complexity of the GAN architecture and more the resources required and chances of instability.

Also, Due to no direct means of controls on output, GANs output need to be closely monitored due to dynamicity in the nature. Domain plays important role in deciding whether the output is acceptable or not. For example, colored dermoscopic images of skin lesions bring additional complexity, and such complexities require a customized changes in the GAN variants. In addition to that, GANs working on the dataset with multiple classes have another issue where generated images need to have less inter class similarity and more intra class similarity keeping the generated images unique as possible.

So far, several attempts are made to address one or more challenges of the GAN or exploit its capabilities in various domain and applications. In Table 2.3: pros and cons are discussed on individual level for some of the prominent and closely related to the current research as more relevant studies are more beneficial in current context.

Table 2.3: Advantages and Disadvantages of some of the prominent and relevant studies

Research work	
(Hammami et al., 2020)	Advantages: YOLO (as pre trained on normal dataset) style of classification model works faster and is more efficient in detecting multiple abnormalities in single images Disadvantages: Cycle-GAN consist of two GANs interacting with each other to produce better result, but that only fact increase the overhead. Also, as both GANs interacts their output heavily depend on each other.
(Qin et al., 2020)	Advantages: Proposed GAN variant – "SL-StyleGAN" is based on NVIDIA's proposed style-based GAN that was developed as an extension PGAN. Thus SL-StyleGAN demonstrate the benefits of Progressive GANs by default. In addition, "SL-StyleGAN" is specifically designed keeping skin lesion images in mind and dropped style mixing as it makes no sense in dermoscopic images.

	Disadvantages: Although it has made required changes to overcome the issues of style GAN, the resultant IS on output images' IS score stays less than normal Style GAN. Secondly, mode monotony is present for some diagnostic categories.
(Guan et al., 2022)	Advantages: "TMP-GAN" is based on PGAN, thus providing the advantages of Progressive GAN. In addition, it is specifically designed to preserve the delicate textures in lesion images. Disadvantages: Study shows that TMP-GAN works fine on grayscale lesion images, but no explicit experiment on RGB images (especially on skin lesion images)
(Qasim et al., 2020)	Advantages: Specifically designed GAN to be helpful in segmentation task. The GAN architecture itself has the third component called "segmentor" that can perform the segmentation task on the fly. Disadvantages: It demonstrated poor performance with ISIC dermoscopic images.
(Abdelhalim et al., 2021)	Advantages: TTUR mechanism decouples the learning rate of generator and discriminator, allowing both networks learn on their own rate. This opens the door to address an unhealthy competition between generator and discriminator. Disadvantages: Self-attention mechanism adds extra computational overhead. Also, researchers have noted unwanted bright spots on generated images.
(Shahsavari et al., 2021)	Advantages: Cascaded GANs motivate the generators to learn entire data distribution rather than the principal density distribution spots. Gates allow the flexibility of deciding whether to utilize cascaded GAN or not give flexibility against the computational overhead. Disadvantages:

	Although the innovative and flexible design, researchers have applied it into enlarging/enhancing the image resolutions instead of generating the images from scratch.
(Ahmad et al., 2021)	Advantages: AE allowed smarter input to GAN to reduce complexity and faster converge. Diverse images generated Disadvantages: Can't be fully automatic as the GAN produce the diverse images, the same can generate the images that technically don't fall in any of the possible classes. Thus, proper supervision is needed.
(Sarmad et al., n.d.)	Advantages: RL implementation can help in faster and robust image generation AE powered by RL can reduce the complexity and computational cost of the model. Disadvantages: The study is done using partial point cloud to demonstrate if GAN can produce proper point could shape or not. Thus, the study is more of a POC rather than a full application on generating the images (especially skin lesion images) from scratch.
(Shubham et al., 2021)	Advantages: All the advantages of AE and RL used in GAN as mentioned above can be yield using the approach mentioned in (Shubham et al., 2021) Disadvantages: Idea isn't generic enough to be picked as it is and apply in other objectives.
(Ukwuoma et al., 2021)	Advantages: In addition to benefits of RL and AE, instead of generating the image itself, a lower dimensioned GFV is generated. This further reduces the complexity of the architecture. Disadvantages: Idea isn't generic enough to be picked as it is and apply in other objectives.
(Singh and Raza, 2020)	Review Paper: Advantage: Multiple state of the art GANs are explained and compared

	Disadvantages: No implementation efforts made to improve any of the existing outcome.
(Bissoto et al., 2021)	Review Paper: Advantage: 18 significant researchers have been discussed. Several SOTA GAN variants have been studied. Included both radiographical images synthesis and dermoscopic image synthesis. Also discussed both applications, Classification and Segmentation for which GAN is applied. Disadvantages: Although good result was yield for radiographical images, GAN worked on dermoscopic images could not provide promising output.

2.4 Summary

There has been huge amount of research done on various ways of image augmentation, this chapter mainly explores and discusses some of the significant and prominent research work which has applied the GANs to address the scarcity of available dataset or to handle the extreme class-imbalance in the dataset. This chapter also discusses about autoencoders and reinforcement learning, several studies that has used them with GAN to further improve the performance of GAN.

Although all there exist many possible benefits of GANs, but some significant challenges and gaps also present which are reviewed as well.

3. Methodology

Oversampling can be helpful when dataset is very limited to address the model underfitting issue or when one or more classes in the dataset has very less samples in comparison with other dominant classes to prevent the model from being biased towards the dominating classes and being biased. The same situation is present in the dataset, which is used, and this study is attempting to address this issue.

3.1 Introduction

In this research, the primary focus is on developing a GAN model that can perform well on colored and textured medical camera images like dermoscopic skin lesion images rather than focusing more on the image classification model. The whole research is divided into four main parts: Data analysis and pre-processing, Image Generation, Image Classification, and Evaluation.

3.2 Overall Flow of execution (Flow Chart)

An overall flow of process execution is displayed in the figure 3.3. As mentioned, overall flow is divided into four major parts.

1. Data understanding and preprocessing

   This is the first step to perform. Here, data is loaded, analyzed, understood, and based on the understanding, pre-processed. Aligning with the objectives current study has kept the image data as center part of the entire study and experiments while from metadata only ground truth holds the equal importance.
2. Image Augmentation

   This part of execution flow is a central and most import part of the research work. Also, this part holds the maximum complexity and experimental efforts. This part is further divided into two parts, Image augmentation based on traditional image transformation techniques, and Artificial image synthesis using GAN and supporting deep networks.
3. Classification

   No innovations are aimed to be brought in this part. A common classification model(s) is shared between four different datasets obtained by different strategies.

4. Evaluation

   Two evaluation strategies are used in this research, evaluations of classification and evaluation of synthetic image generation by GAN. This part is a comparative study and will help in gathering the outcomes of the experiments.

Details of all these parts are discuss in further sections.

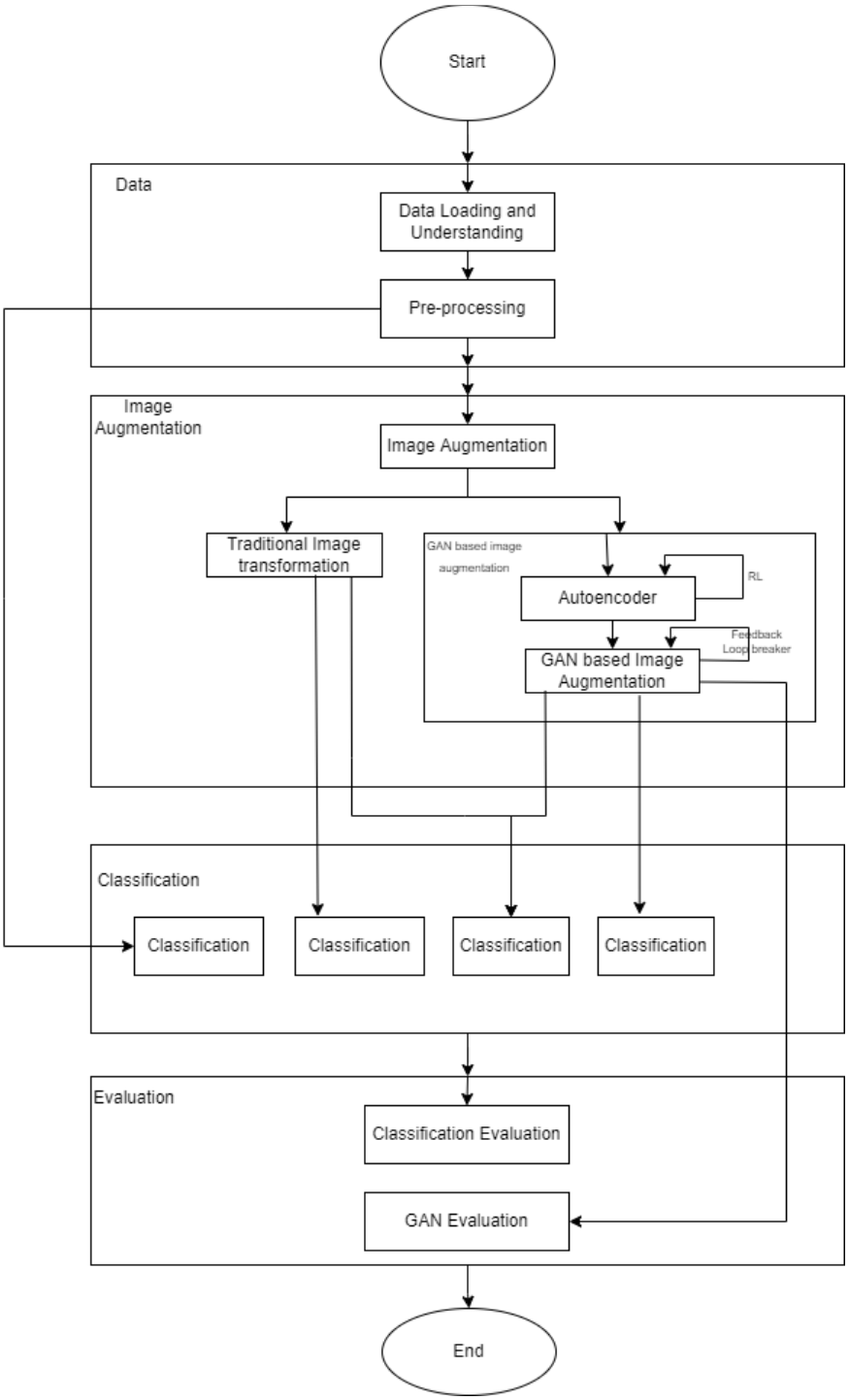


Figure 3.1: Flowchart of overall process execution

3.3 Data analysis and pre-processing

The dataset used in this research to train GAN based image augmentation architecture is (International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset., 2020). This dataset is about dermoscopic skin lesion images.

3.3.1 Understanding the data

ISIC 2020 dataset contains:

1. 33,126 JPEG and DICOM images
   Each image is 3 channeled RGB dermoscopic natural image. Images are one of 9 skin conditions. In general, all the images are high resolution and clearly showcase the skin condition.
2. Metadata containing information (patient ID, lesion ID, gender, age, and general anatomic site) for all 33,126 images
   A file with comma separated values about basic information associated with each image.
3. Duplicate images list
   A file contains information regarding duplicate entries in the dataset.
4. Ground truth of all 33,125 images
   Ground truth helps in knowing the actual class (the skin condition) associated with each image.

Figure3.1 shows different classes of skin lesions present in the ISIC 2020 dataset. It is clearly visible that inter class variation is very less among different types of skin lesions. Also, different tones of skin make background color and textures getting differ even within same class making intra class diversity high. Classifier model will have to deal with these challenges as well.

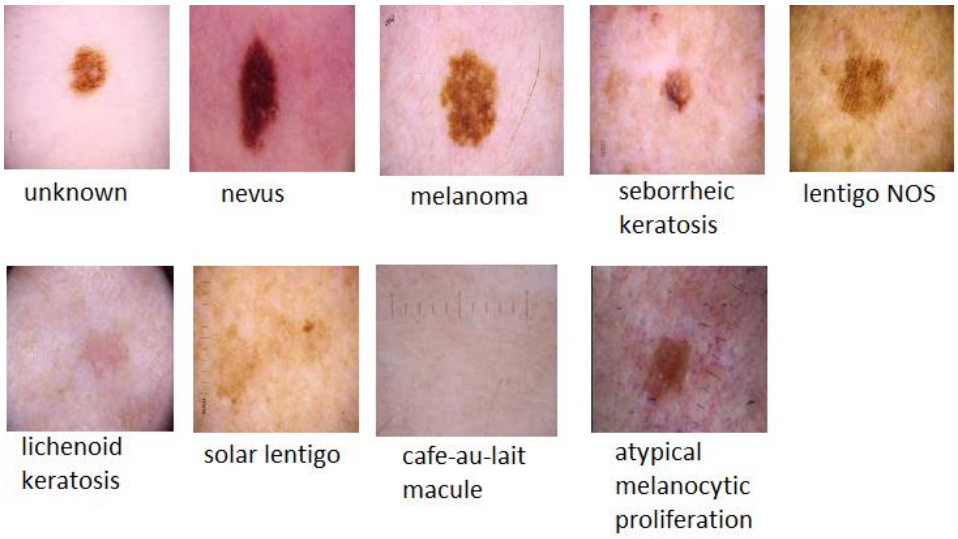


Figure 3.2: Sample images of different types of skin lesions

A general EDA on the metadata and ground truth information that comes alongside the skin lesion images, it is clear that some of the records are missing gender information, age, or the location of the lesion on the body. Although this information could have been used in machine learning classification model, this study is focused more over the skin lesion images and classification based on the deep learning model trained on the images. Thus, no imputation is needed for such missing records as the ground truth is available for every image.

EDA activity also confirms that there is no improper information given in the field of skin lesion diagnosis (i.e., ground truth). A cross verification on ground truth is done by checking if all the records that are marked as malignant falls under 'melanoma' class.

Looking at the class distribution one can see that out of 33,126 images, 81.88% of the images are labeled as 'unknown' making it a widely diverse class. This can lead to two major issues.

1. Huge Intra-class diversity. This issue can be a big challenge in classification as classifier model will find it difficult to get a common pattern.
2. Misleading result of "Accuracy" of the classifier. As even if classifier model classifies every image as "unknown" still, 81.88% of time the result will be accurate. However, that model is highly unacceptable.

If the "unknown" class is entirely dropped from the dataset, class distribution will be looked like below.

Table 3.1: Class Distribution of know skin lesion classes in ISIC 2020

Class	Percent distribution
nevus	86.52116
melanoma	9.73009
seborrheic keratosis	2.24925
lentigo NOS	0.733089
lichenoid keratosis	0.616461
solar lentigo	0.116628
cafe-au-lait macule	0.016661
atypical melanocytic proliferation	0.016661

As it is clearly seen, even in "known" classes, 86.52% of the images are of "nevus" class, making the dataset extremely biased towards "nevus" class. Whereas "melanoma" class is more critical to be detected correctly.

3.3.2 Data pre-processing

Metadata and ground truth information of the dataset is proper and doesn't need any explicit pre-processing. Images of the dataset are high resolution and captured neatly with dermoscopy. Thus, lighting and saturation is also proper in the images making them almost ready to use state. However, some required pre-processing steps are discussed further.

Increasing/Decreasing the dataset size:

ISIC 2020 dataset has more than 33,000 images. Training generative models like AE or GAN and classification models on the dataset that is as big as ISIC 2020 is computationally costly and extremely time consuming. In addition to that, training a good enough deep learning based model is possible with relatively smaller dataset as well.

For this research work,

1. Dropping extremely rare classes completely
   In the dataset, two classes 'cafe-au-lait macule' and 'atypical melanocytic proliferation' have only one image. One single image will not be sufficient for

generative models to learn the data distribution of the class to generate synthetic images. Neither it will be sufficient for classification models to learn the general pattern for classify the other images of these classes. Given this, these two classes are not contributing to the objectives of this research. Thus, they will be dropped from the dataset being used in the research work and experiments.

2. Reducing dominating classes to 1000 images each
   Classes 'unknown' and 'nevus' dominates the dataset. While the main objective of the research is to find the means of oversampling the deficient classes, the dominating classes are manually under sampled to certain limit.
3. Over sampling remaining classes to 1000 images per class.
   This will be done using image augmentation. More about this is discussed in upcoming sections of this chapter.

Resizing the images:

As neural network models are designed for specific input deamination, to input the image dataset into the network, all the must be in the same size. But looking at the dimension of the images in the dataset, this requirement doesn't seem to be fulfilled. Figure:3.2 shows there are as much as 88 different image dimension groups present in the dataset.

0	6000 X 4000	14703
1	1872 X 1053	7534
2	5184 X 3456	3418
3	2592 X 1936	674
4	4288 X 2848	729
...	...	...
83	2237 X 2237	1
84	2087 X 2087	1
85	1811 X 1811	1
86	1783 X 1783	1
87	1066 X 756	1

Figure 3.3: Different dimensions and their frequency in images of ISIC 2020 train set

Also, high resolution as 6000X4000 will drastically increase the computational power requirement if we feed them as they are in the GAN/AE network leaving classification network

aside. Given both situations, the first step in data pre-processing should be resizing all the images to common and lower resolution. However, resolution should not be lower as much that the features in images get compromised. In this research work, image resolution is kept 256X256.

Normalization of images being used in the generative models:

While utilizing the images within the generative models, the image pixel intensity values are normalized between 0 and 1.

3.4 Images Augmentations

Image augmentation is the main objective of this research. In this section, various ways of image augmentation are discussed.

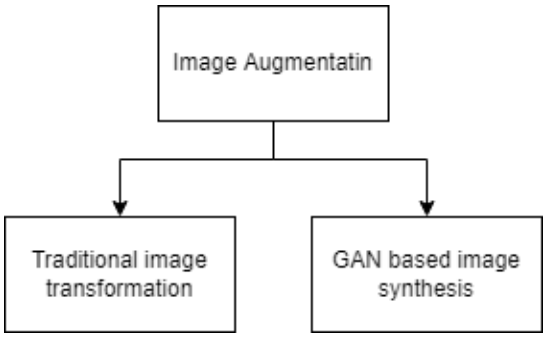


Figure 3.4: Image augmentation techniques

On the Assumption that present class imbalance in ISIC 2020 dataset will impact the classification model trained on this dataset and will be highly biased towards majority classes, image augmentation becomes critically important and thus it is the primary focus of this research. (Bissoto et al., 2021; Guan et al., 2022) extensively talks about different generative data augmentation techniques that include both image-to-image translation and noise-based image generation. However, fundamentally speaking two main ways of augmenting the images (shown in Figure 3.4: Image augmentation techniques) will be explored in this research, Traditional image transformation and GAN based image synthesis (Dumagpi and Jeong, 2021)

3.4.1 Traditional image transformation

Although less sophisticated, image transformation techniques like rotating, zooming, cropping, etc. have been used to up sample the images for any particular class(es). And in many studies (Verma et al., 2020; Waheed et al., 2020; Dumagpi and Jeong, 2021) , image transformation has either been used with image synthesis or compared with image synthesis concerning the effectiveness.

Given the nature of the images and the factors responsible for classification, a few techniques of transformation like thresholding, erosion, dilation, opening, closing, etc. cannot be used to augment new images as they might alter the color, contrast, texture of the image. Whereas linear transformation techniques like resizing/scaling, cropping, zooming in/out, rotating, and flipping can be safely used.

In the context of traditional image transformation techniques, this research will be a comparative study of the effectiveness of classification models trained on the dataset that included image transformation + GAN in data augmentation, only used GAN based synthetic images for data augmentation, and standalone usage of image transformation for data augmentation.

3.4.2 GAN based image augmentations

Mainly classified into two types, image to image translation model and noise-based image generation model, many variants of GAN based models are discussed (Singh and Raza, 2020; Bissoto et al., 2021).

Inspired by studies (Qin et al., 2020; Verma et al., 2020) with comparatively similar dataset and promising outcome, this research will explore and experiments with two widely accepted GAN variants, DC-GAN, and Style-GAN. DC-GAN is a relatively simpler GAN variant with both generator and discriminator comprising of the deep convolutional network. Unlike conditional GANs, DC-GAN doesn’t have external conditioning as the input and output layer of the discriminator network contains a single neuron and thus can’t produce probability distribution for the generated image.

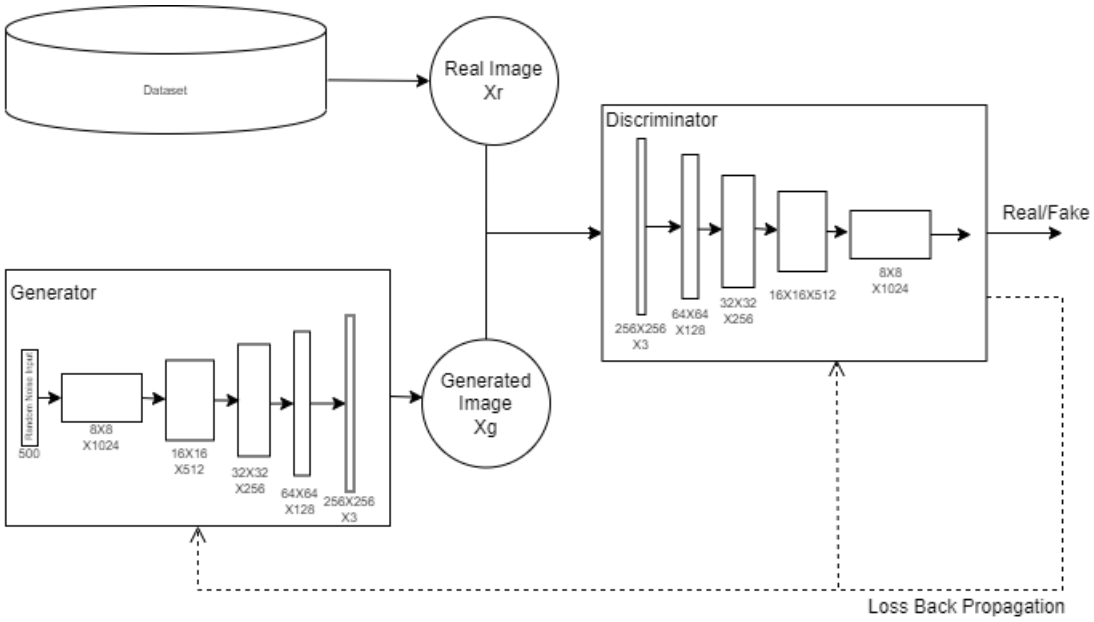


Figure 3.5: Basic architecture of DC-GAN

Although original DC-GAN was proposed with different set of neurons in hidden layers of generator and discriminator, Figure 3.5 shows DC-GAN architecture with parameter that are more suitable for 256X256 RGB images. As it is clearly seen, with higher resolution of input noise or output images or both, complexity and computation of the Architecture increases drastically.

$$MinMax\ Loss\ =\ E_x\left[\log\left(D(x)\right)\right]\ +\ E_z\left[\log\left(1-\ D\left(G(z)\right)\right)\right] \quad (2)$$

Loss function of the GAN is both, generator network and discriminator network's loss added together. Another bigger challenge with this architectural design is with higher number of neurons in the layers, back propagation of loss fails, and networks cannot learn the data distribution properly.

It is clear that vanilla GAN can produce realistic images but being stable, they cannot achieve high resolution.

3.4.3 Using Autoencoders with GAN

Purpose of autoencoder in this research is to support GAN network rather than acting as generative network itself. Figure 3.6 shows the architecture of the autoencoder that is being used in this research.

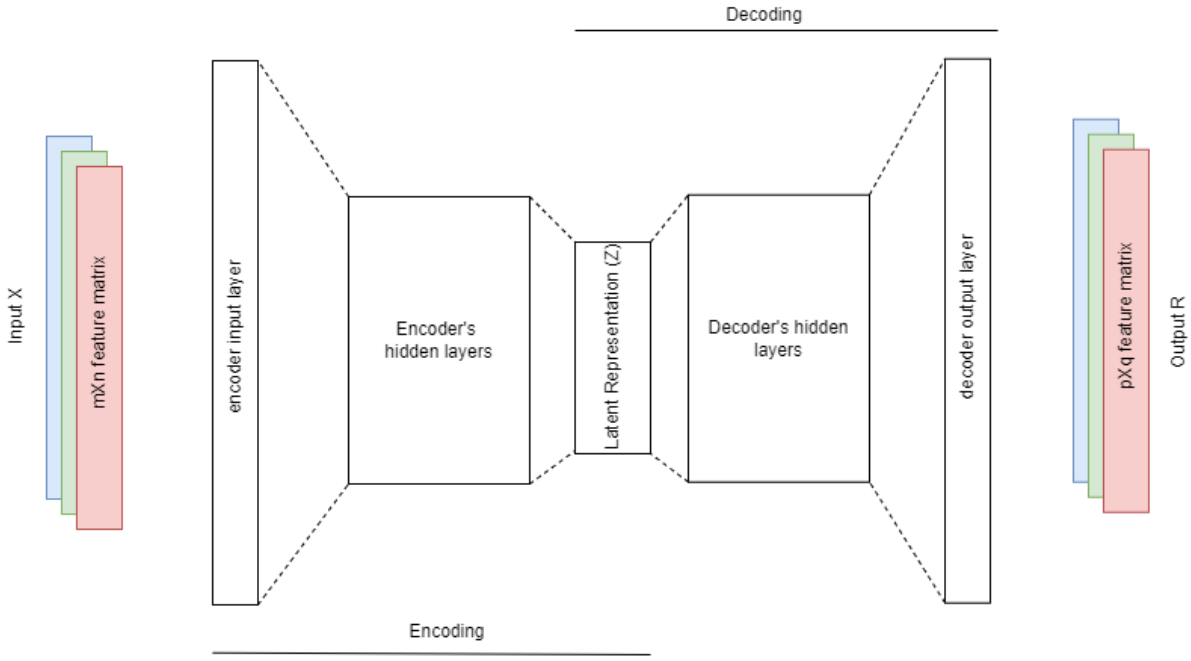


Figure 3.6: Autoencoder Architecture

In this research, input X for the encoder is mXn RGB image while output of decoder is pXq RGB image. In general, input and output dimensions of an independent autoencoders are not necessarily the same but to keep implementation clean and less complex to understand, in the context of this research pXq is the same as mXn. Thus, number of neurons in input layer and output layer are same and that is 3*m*n. Inner architecture of hidden layers and latent representation is a matter of hyper parameter tuning. The research will exploit different variation of the inner architecture.

Mathematically, Z can be seen as function of X (Z = f(X)), and R can be seen as function of Z (R = g(Z)) over some weights and biases associated.

$$Z = f(w_e, b_e; X)$$

$$R = g(w_d, b_d; Z)$$

(3)

The above equations can be understood as, within an autoencoder network, set of recognition weights ($w_{e)}$ are used to generate intermediate and reduced latent representation (Z) from the input data (X) and then with the set of generative weights ($w_{d)}$, the latent representation is converted into approximation (R) that is as close to the input data (X) as possible.

Figure 3.7 demonstrate the application of autoencoder (AE) with a basic GAN network. The autoencoder network used in this architecture to handle the input and output of generator is pretrained on the same input dataset.

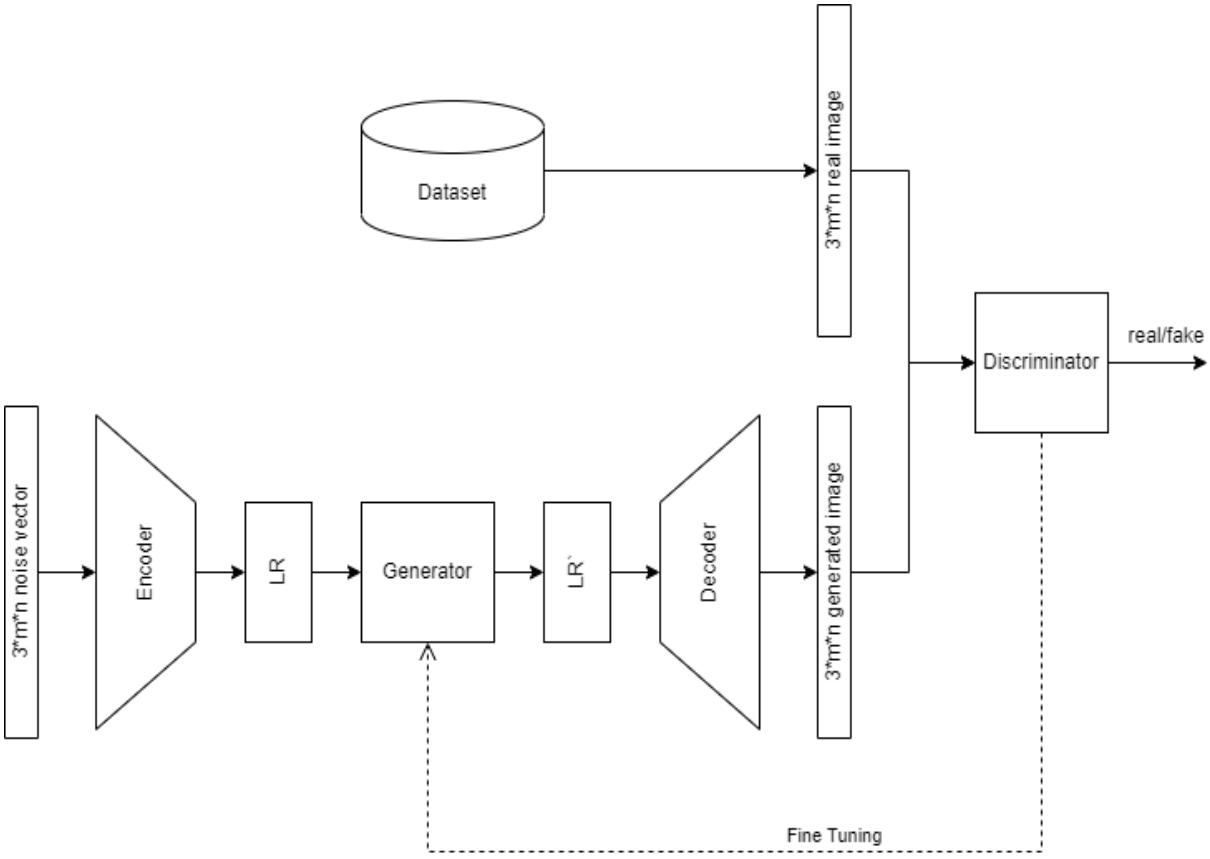


Figure 3.7: Autoencoder with GAN network

From the figure 3.7 it becomes quite clear that instead of training generator network with the random noise vector of size 3*m*n, encoder network has reduced the input dimension drastically to the size of LR vector (latent representation vector). Also in this architecture, instead of generating the image itself, generator is generating the approximation of most suitable latent representation (LR`) that can be decoded to real looking skin lesion images. This reduces the complexity of generator network drastically and make the generator network more stable.

3.4.4 Using Reinforcement learning with AE

Autoencoders being a generative network itself can be further supported by another promising technique, Reinforcement Learning. Similar to the challenges of incorporating RL with GAN,

with AE as well there isn't any direct flow or execution in AE network that is performing any action out of many possible actions on the environment. However, noise GFV can be produced as the output of AE when the input provided is random noise and this GFV can be used as a state vector for RL to learn the action.

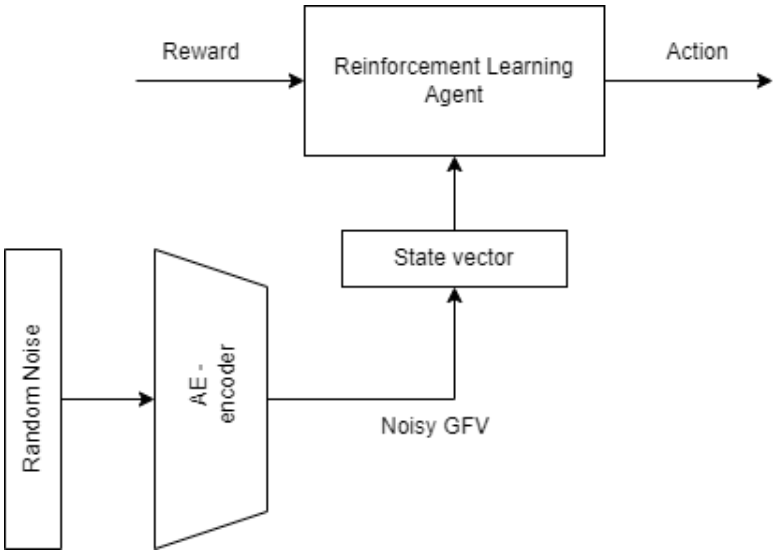


Figure 3.8: Autoencoder used with RL

Figure 3.8: shows the proposed mechanism to utilize the RL with AE in order to obtain the optimum action. Reward to the RL agent is derived from the outcome of GAN. Here the purpose of the RL Agent is to learn that in given state, which action yields the maximum reward. In further section the whole integrated system is demonstrated where this mechanism of RL and AE can be seen in action.

3.4.5 An Overall Architecture

In Earlier sections all different components are discussed separately with their working flow and architecture in detail. In this section the integrated system is discussed in brief, showing how different components complement each other and work together for image synthesis. Autoencoders are pretrained on the same dataset beforehand.

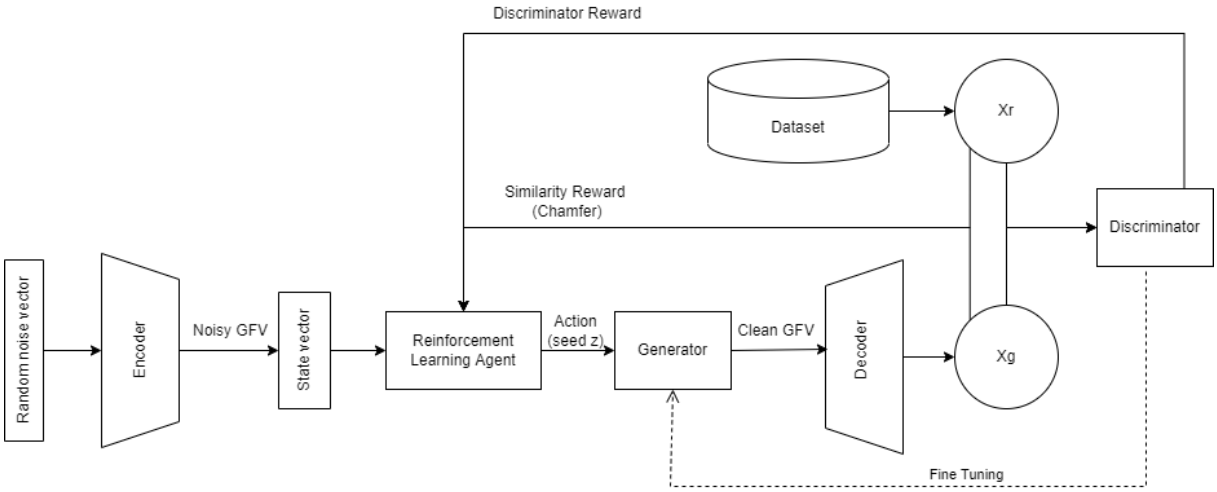


Figure 3.9: An integrated system of AE, GAN, and RL

As it is seen in the figure 3.9, although the input to the network stays a "random noise vector" but it doesn't go into generator directly but into the encoder. Encoder reduces the dimensionality and generate the noise GFV that is suitable to be a state vector for RL agent. Given this state, RL picks up the action (a seed z) that is an input vector for GAN's generator. Generator must learn to produce a clean GFV which can successfully be decoded back into approximately same image as input dataset (Xg ~ Xr). Thus, the role of generator is to produce low dimensioned clean GFV instead of high dimensioned RGB image itself.

The loss back propagation in GAN is already discussed in earlier sections. However additional thing in this architecture is feeding the rewards to RL agent based on the performance of the generator. Total reward is the function of discriminator reward and chamfer reward.

Total Reward

$$r = f(r_d, r_c) \tag{4}$$

This reward acts as feedback for RL agent suggesting that given the state which action yield maximum rewards (Q value) that ultimately shapes the policy.

3.5 Classification

The main aim of this research is limited to generating desired GAN model to overcome the class imbalance problem and thus this research doesn't focus on improving the image

classification models. These models will be used only for comparing the quality of the training dataset.

This research will use two classification models, basic CNN architecture and VGG Net using transfer learning. The image classification models will be trained on different datasets while keeping constant hyper-parameters, activation functions, and overall architecture. Once trained, these models will be evaluated on the same test dataset using the same evaluation metrics. Dataset generation is already discussed in the above sections.

3.5.1 Early Feedback loop breaker

This is highly experimental concept. The idea is to have a classification mechanism embedded into the generative model instead of an individual component of the execution flow. Purpose of this classification mechanism is to continuously validate the outcome of the GAN during the GAN training phase itself and provide the constructive feedback on the fly.

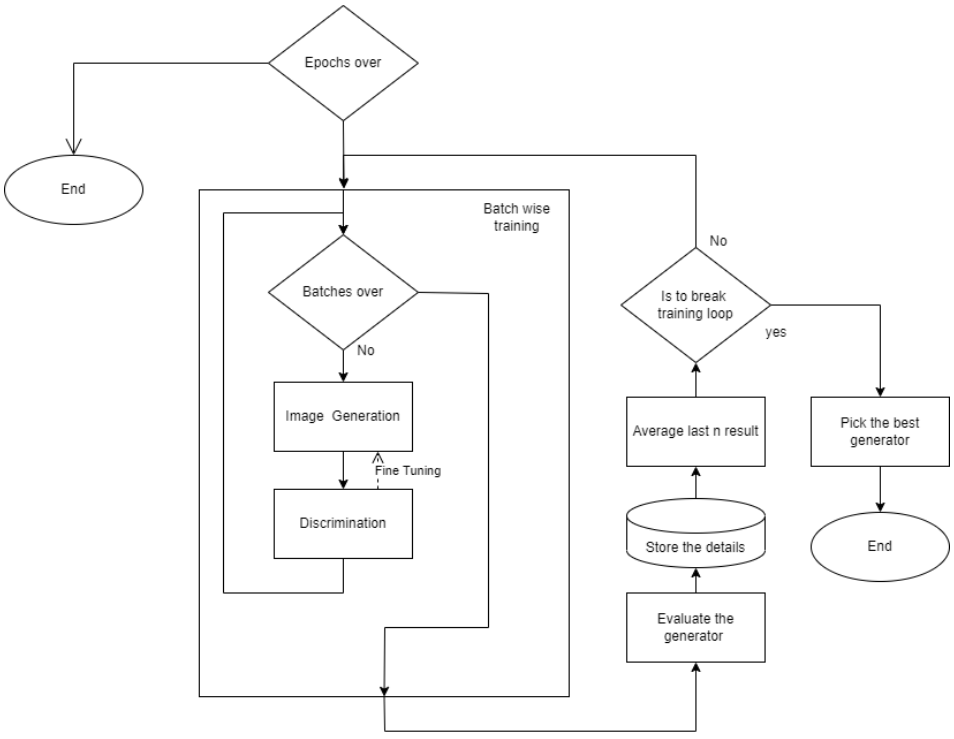


Figure 3.10: Execution flow of loop breaker mechanism in GAN

Figure 3.10 demonstrate the loop breaking mechanism placed in normal GAN workflow. Concept of moving average is being used in the loop breaker.

The last generator's output is being compared with the moving average where 'n' is a hyperparameter. Once the convergence is detected or performance of GAN is detected to be downgraded instead of completing the training for total number of epochs, this mechanism can act as loop breaker.

Loop breaker mechanism works on continuous evaluation of the classification task carried out on the dataset being augmented by the GAN, in further section the strategies of evaluating the GAN and classification model is discussed. At the end, it must be understood that this mechanism brings a great amount of computation with it, and it must produce the greater benefit than the cost it comes with. Thus, it is necessary to assess this mechanism with the architecture where it is not present, it terms of time taken, number of saved epochs, and betterment in final outcome.

3.6 Evaluation

Evaluation of this research will be a comparative study of the outcome of different models and experiments. As the development work in this research will be done in two parts, they both will be evaluated separately.

3.6.1 Evaluating GAN models

(Borji, 2019) talks briefly about the different measures to evaluate the performance of GAN models. In their study, they have proposed basic characteristics of a good GAN model evaluation measure

- Evaluation measures should favor the GAN model that can generate high-fidelity samples
- It should favor the GAN model that can generate a diverse sample
- It should favor the GAN model with controllable sampling
- It should favor the GAN model with well-defined bounds
- Evaluation measures should be sensitive to image distortion and image transformations.
- The evaluation measure's outcome should be in line with human perceptions.
- It should be less computational complexity.

This study will mainly be dependent on evaluating the classification model to determine whether the dataset generated using GAN is helping the classification process or not. However, there can be independent measures to evaluate GAN performance. Broadly speaking, there are two ways of evaluating the GAN, quantitive measures, and qualitative measures.

Inspired by, (Qin et al., 2020) this study will evaluate the GANs based on

- Quantitive Measures
    - Inception Score – IS

      $IS = exp(E_x [KL(p(y|x) || p(y))])$ (Qin et al., 2020)

      (5)

- Manually validating by visualizing the output of GAN.

3.6.2 Evaluating classification models

Given that the dataset is highly imbalanced and biased towards dominating class(es), the High Accuracy value is often misleading. For medical image classification, a high rate of false negatives cannot be accepted, on another hand high rate of false positives will require a continuous cross-checking mechanism, making the final diagnosis more time-consuming and expensive.

With this understanding, this research will evaluate the classification model using the following measures. The confusion matrix will be generated as one class is a positive case and the rest all being negative cases.

- Sensitivity (Recall, True Positive Rate): The number of positive cases that are correctly predicted out of the total positive cases
    - Sensitivity = True Positive / (True Positive + False Negative)
    - The value of sensitivity should be as high as possible
- Specificity: The number of negative cases that are correctly predicted out of the total negative cases
    - Specificity = True Negative / (True Negative + False Positive)
    - The value of specificity should be as high as possible
- Precision (Positive predictive rate): Rate of correctly predicted positive case our of total positive prediction.

- Precision = True Positive / (True Positive + False Positive)
- The value of precision should be as high as possible

Among Sensitivity, Specificity, and Precision, the metric sensitivity is more suitable in the context of this research and with respect to the nature of the data. Thus, research will follow sensitivity to determine the quality of the classification task. However, the conclusion of the research will not be comparative but quantitive. Based on the evaluation result, this research will try to propose the answers to the questions mentioned in the 'research question' section.

3.7 Summary

This chapters talks in good details about the various steps being involved to carry out the research work and related experiments. From understanding the dataset, utilizing it in a proper way in the generative model, to the usage of augmented images in the classification model and evaluations, this chapters also discuss the techniques and deep learning models that are being used.

Separately introducing the various components, its architecture, and the workflow, later the overall architecture and the complete execution flow are discussed where all the components are seen integrated as one unit. Classification model is not discussed in a great detail as it is deliberately put out of the main research focus. However, an innovative and experimental loop breaking mechanism is proposed in brief along with the purpose, expectations, and challenges. At last, this chapter talks about the evaluation strategies to be used.

4. Experiments and Analysis of implementation strategies

It becomes important to experiment with different approaches around the finalized methods to better understand and determine the actual nature of the progress in different given situation.

4.1 Introduction

Being aligned with and extending on top of the previous chapter, this chapter talks about different experiments being carried out around the methods discussed earlier. Also, this chapter analyses the implementation strategies that were being used in the experiments. Concept of moving average is used in this research to smoothen the trend of the progress and avoid the impacts of sudden fluctuations in overall analysis. Starting from the data analysis and preprocessing, this chapter will cover all major aspects of the generative model implementation and pre-requisites for the same. Later this chapter covers understanding of loop breaking mechanism that is integrated with classification task.

Each subsection of this chapter will have three main perspectives, Understanding the implementation of model or strategy and need of the experiments, going through basic algorithm around the model implementations, and if applicable, hyperparameters.

All the experiments carried out in this research work is done on the environment with below mentioned configurations.

- Operating System: Windows 10 Pro – 21H2 – 64-bit OS
- Physical Memory (RAM): 16 GBs
- Processor: Intel® Core™ i7 – 1050U CPU @ 1.80 GHz 2.30 GHz, 4 core – 8 logical processors.
- Fixed Disk Type: SSD
- Python version: 3.10.6
- Required python packages are kept up-to-date as of the date of experiments.

4.2 EDA and Data pre-processing

The ISIC 2020 is a dataset of dermoscopic images of different types of skin lesions. Along with the image data, this repository also contains the metadata associated with the images. For an example, patient details and skin condition details. Although the dataset that ISIC offers is almost cleaned and ready to consume, there are some basic steps of data pre-processing is

needed. However, a properly done EDA on the dataset gives better understanding on the data and helps in determining the actions to be taken during the data pre-processing.

4.2.1 Understanding Data

Talking about the metadata and ground truth, there are in total 33,126 records available alongside the equal number of skin lesion images. As shown in Table 4.1, this data comprises metadata and images of 2056 unique patients, nine different classes of skin lesions images, 18 different age groups, and images are captured from six different sites of patient's body.

Table 4.1 Unique records per field of ISIC 2020 dataset

Field	Number of Unique records
Image name	33,126
Patient ID	2056
Lesion ID	32,701
Gender	2
Approximate Age	18
Anatom Site	6
Diagnosis	9
Benign/Malignant	2
Target	2

The fields "Benign/Malignant" and "Target" are related to each other. Figure 4.1 shows the relation between these two fields.

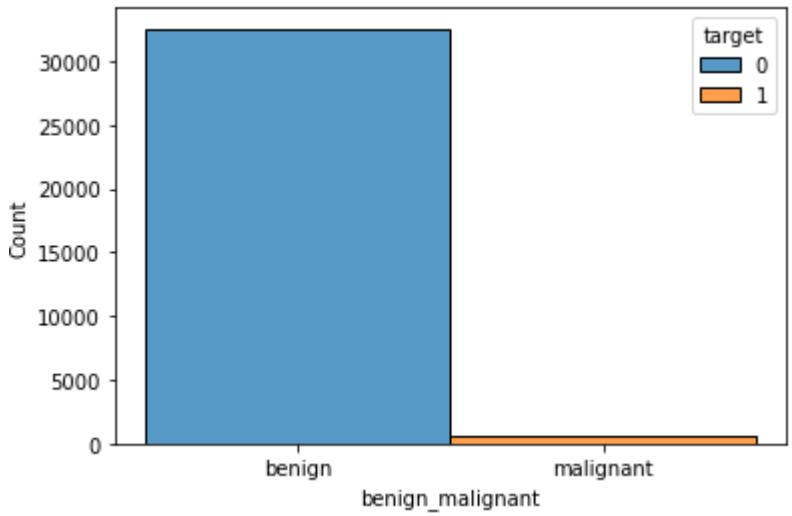


Figure 4.1: Benign/Malignant vs Target

It is clearly seen that all the records that have “Benign/Malignant” field set as “Malignant” are also having the “Target” field set as 1. This plot also revels the proportion of record being malignant vs benign, and that proportion is 584/32542 and that is 0.017 i.e., 1.79%. This gives a clear representation of data being highly imbalanced. Further checking, Figure 4.2, it is also clear that all the malignant records are having same diagnosis while benign records have rest other diagnosis. This information makes the field “benign/malignant” a central and important field of the metadata.

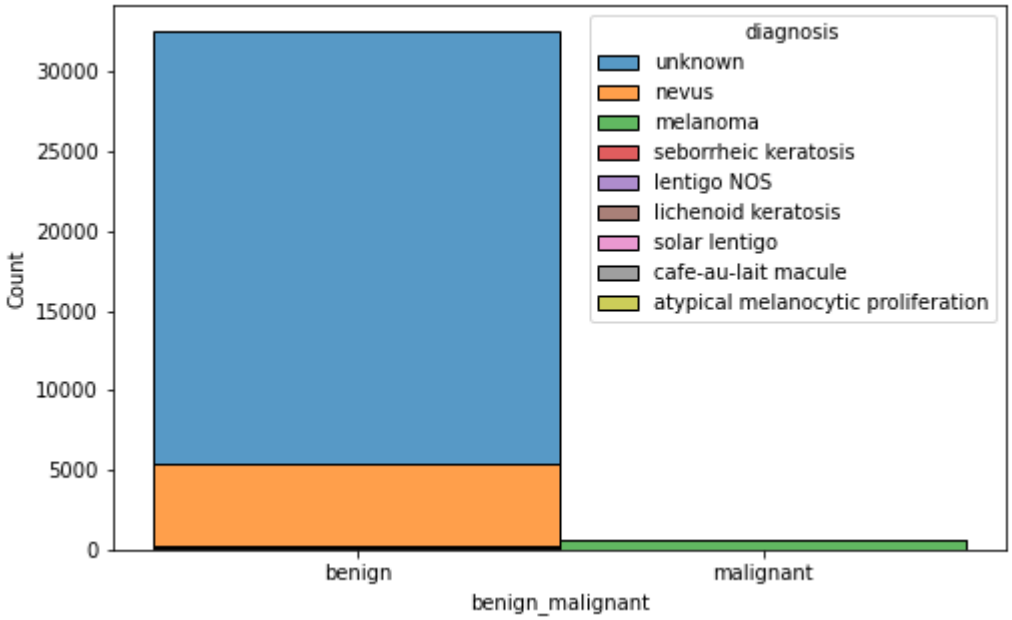


Figure 4.2: Benign/Malignant vs Diagnosis

However, by the nature of this research, the fields in metadata are less significant than the images associated with metadata. Thus, more focus has been put in understanding the image data and performing required pre-processing steps on them.

Figure 3.2 showcases the sample images of ISIC 2020 dataset. However, this research work is more focused on experimenting on “melanoma” images. Figure 4.3 explicitly shows the sample images of “melanoma” lesion.

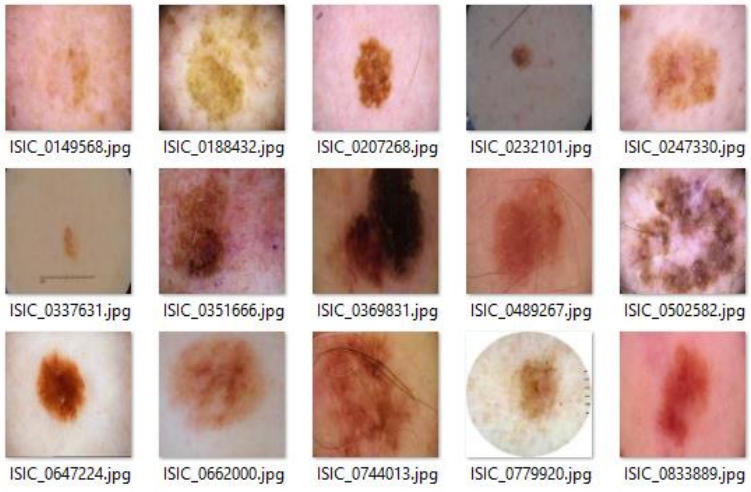


Figure 4.3: Sample images of Melanoma lesion

As it can be seen from the figure 4.3, there are clearly visual differences in both size and shape of the lesions. In addition, the skin tone also varies from person to person. Some images also contain body hair on and around the lesion. While most images are square there are few images that are round shaped. All these differences combined makes both image synthesis and classing task more challenging. Properly performed data pre-processing steps can help in addressing some of the challenges.

4.2.2 Data pre-processing

Several steps that can be performed to prepare the data for model training and validation. Further in this section, the data pre-processing steps are discussed on their importance and variations based on the parameters.

Removing unnecessary data

ISIC 2020 dataset also contains duplicate image list that contains total 426 records. These records can be safely removed from the final dataset. In addition, there are only one images

each for ‘café-au-lait macule’ and ‘atypical melanocytic proliferation’ classes which is certainly not significant. Especially due to the fact that they are ‘benign’ typed images.

The biggest category in the data set is “unknown”. Thus, majority of the images are not tagged with proper category. This research will experiment with and without the unknown category. Figure4.4 shows data distribution with and without “unknown” category.

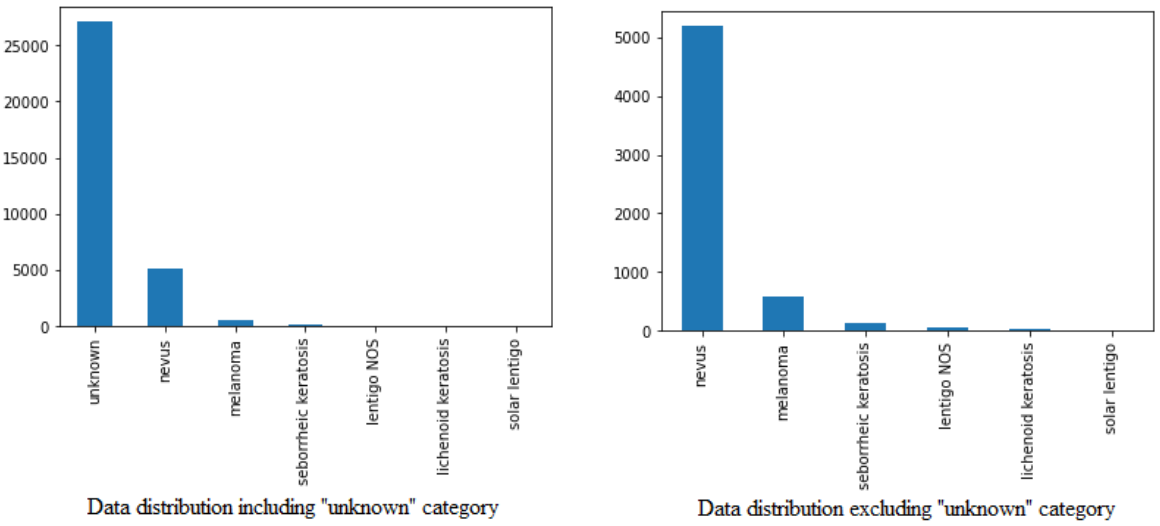


Figure 4.4 Data/Class distribution with and without “unknown” category

While the aim of this research is untimely determining the approach of up sampling the dataset with less occurring class using the synthetic image data, However, the classes “unknown” and “nevus” need to down sampled – 1000 images each.

Refining the data

Referring the figure 4.4 and table 3.1, even after removing the two least occurring categories from the dataset, there are still few categories doesn’t contribute any significant quantity in the dataset. Thus, such categories are combined and form “others_benign” class.

As it has been clear in the previous chapter itself, “Classification Task” is not the primary focus of this research, it focuses more on the image augmentation and preparing the data for the same.

The first step in preparing any image generative model or image classification model is to bring all the training data to common size. The quality of the training images should be sufficient that the images do not lose the important features within them. On other hand as the size of the

images directly controls the complexity of the models. Unnecessary high-resolution images do more harm than the good.

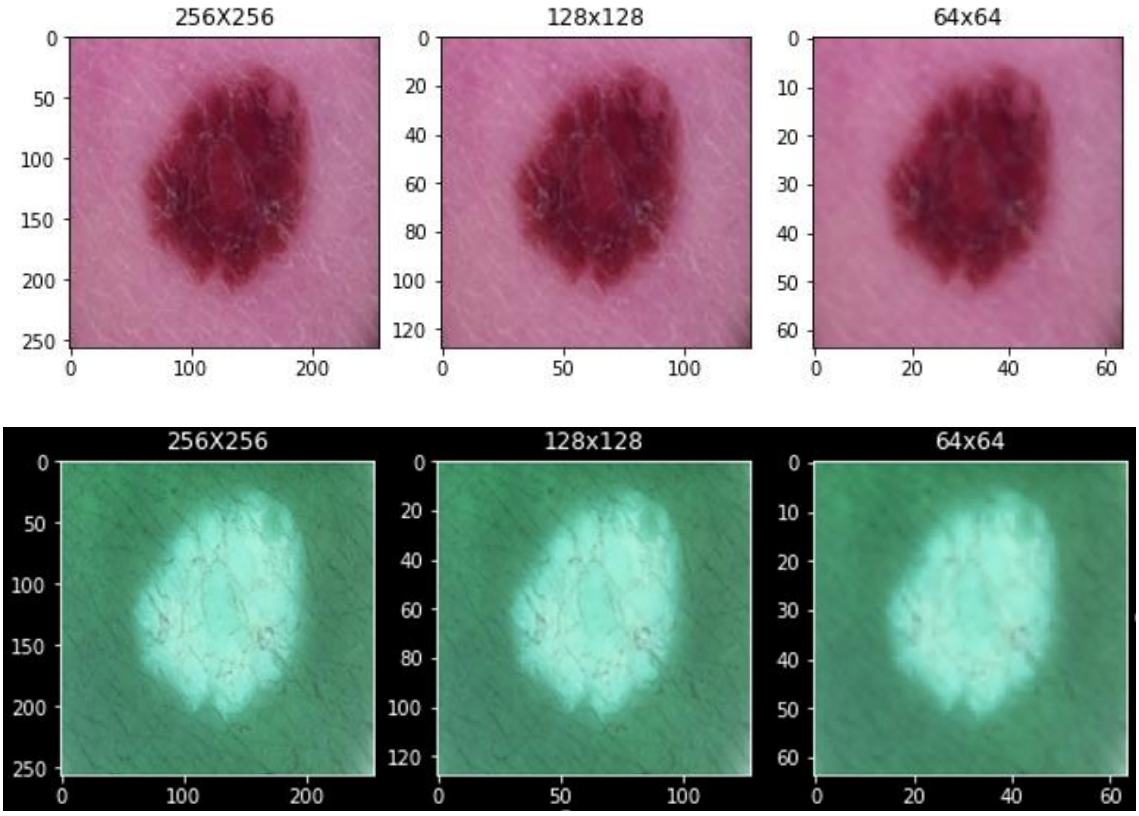


Figure 4.5: Sample skin lesion image in different resolution

Image resolution is considered as a hyper parameter in the process of image re-sizing during the data pre-processing. Figure 4.5 shows the same image in three different resolutions and the second part of the figure shows the same images in negative form so that the features of the images can be clearly seen. It is quite clear that as the resolution goes down, the features of the images are compromised. During the phase of experiments, this research has taken these three resolutions and carried out different experiments to understand the difference in outcome.

All the above-mentioned steps are performed in prior to actual training the generative models while normalizing the images is performed during the AE and GAN training.

4.3 Moving Average

Moving Average is calculating the average of defined window in the series of datapoint. Instead of referring the datapoints directly a moving average offers dynamic and statistically more stable representation of the trend in which the datapoints are moving.

$$MA_i = \left[\sum_{j=i}^{i-t} X_j\right]/t$$

(6)

The main advantage of having moving average is, it prevents an impact of sudden or abrupt change in the datapoint by producing smoother trend. Moving averages are more reliable in decision making as it:

- Doesn't get impacted by an abrupt/false positive/outliers data point. Instead, it calculates the average of the current and last n-1 datapoints to generate the representative datapoint for current record. These smoothen the trend and also make sure that the current status is being supported by last n-1 records.
- Unlike the normal mean, it is parameterized so it always in control and put the focus on recent changes in the trend rather than carrying out the whole historical data which might not be relevant anymore. Also, as it is parameterized, tuning the moving average is possible while a simple 'mean' cannot be tuned.

However, in this study, basic customization has been done in implementing the Moving average.

$$MA = \begin{cases} if\ len(X) < t, & avg(X) \\ else, & MA(X,t) \end{cases}$$

(7)

In the given equation and the implementation of the logic for this equation, the term "t" is hyper parameter, and it is tuned differently for in different application. Figure 4.6 demonstrates how moving averages helps in smoothing and stabilizing the trends where data points can have sudden outliers.

As the decision making in this research depends how better or worse any datapoint performing, such sudden and expected datapoint can lead to unwanted decision. So, instead of considering the datapoints directly, moving average representation is used.

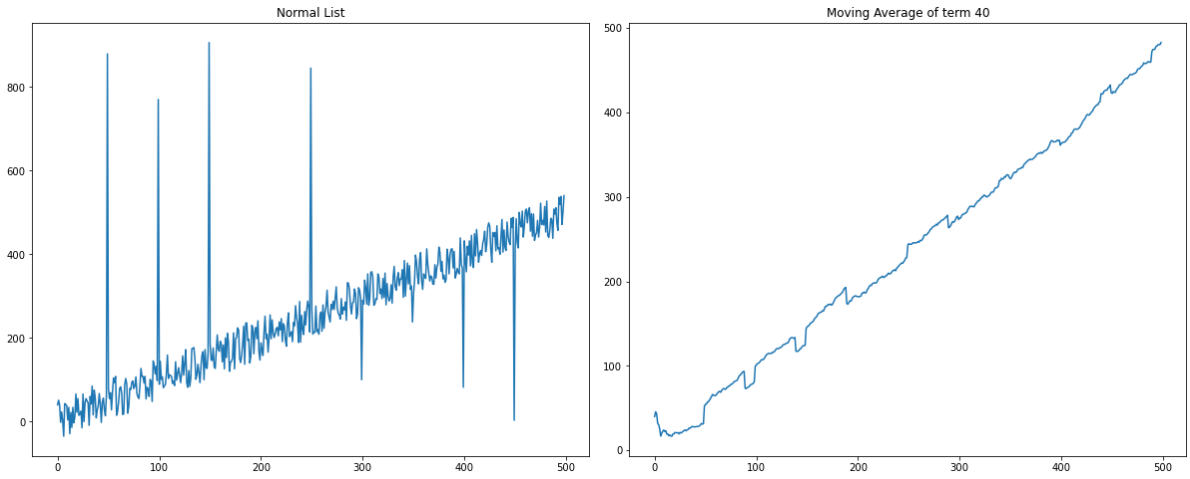


Figure 4.6: Sample list of datapoints and its Moving Average representation

Further in coming sections, it is discussed how this moving average is integrated in generative models and classification models.

4.4 Traditional image transformation

Image transformation is basically function of changing the coordination of image. Traditional image transformation is most basic way to augment images for up-sample the dataset. Though this mechanism may lack of diversity in the output images, in many cases of classification task, only traditional ways of image transformation has been proven to be helping thus it is in this research.

Image translation

Image translation is the process of shifting the image from one location on the plane to other location. During the image translation rectilinear shift of the coordinates is carried out with no alteration on the image scale or pixel intensity. Figure 4.7 demonstrates the image translation performed on the skin lesion image.

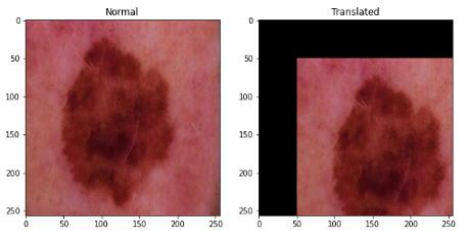


Figure 4.7: Image Translation

Image scaling

Image scaling is resizing the image. Scaling can be both enlarging or reducing the scale keeping the resolution same. Meaning either some part of the image will be lost in case of enlarging the image or a blank corners and sides will be added in case of reducing the scale of the image. Thus, it can be said, enlarging the image will serve same as cropping and re-sizing the images whereas in case of reducing the image, explicitly cropping and re-sizing is needed.

Scaling can also be done off the height/width ratio. Figure 4.8 demonstrates image scaling operation.

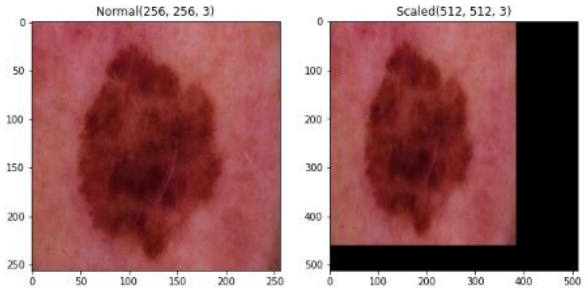


Figure 4.8: Image Scaling

Image shearing

Image shearing is a linear mapping of each coordinate of the image being displaced towards the fixed direction. Shearing can be done two ways, horizontal and vertical. Shearing shifts every points horizontally or vertically with specific points with proportion to original coordinates. Point to note in shearing is, it adds blank areas in the remaining part of the image which becomes hard to crop out unlike scaling. Figure 4.9 demonstrates image shearing operation.

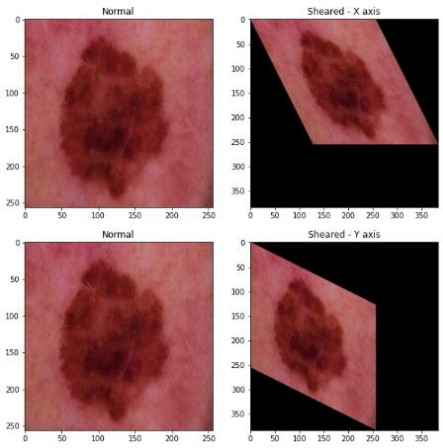


Figure 4.9: Image Shearing

Image reflection

As the name suggest, reflection is mirroring the image. Similar to shearing, reflection is also done horizontally or vertically. Image reflection is a way to flip the image horizontally and vertically and is one of the most useful techniques as it doesn't alter the size, shape, and colors of the image still provides some level of diversity. Figure 4.10 demonstrates image reflection operation.

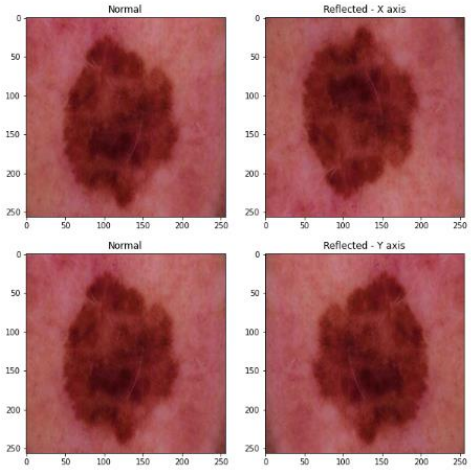


Figure 4.10: Image Reflection

Image rotation

Image rotation is the process of rotating the image on the plane for specified degree. In general image rotation is used when the input image is out of alignment. Image rotation can be done

either clockwise or anti-clockwise. Similar to image reflection, image rotation is extensively used for image augmentation. Figure 4.11 demonstrates image rotation operation.

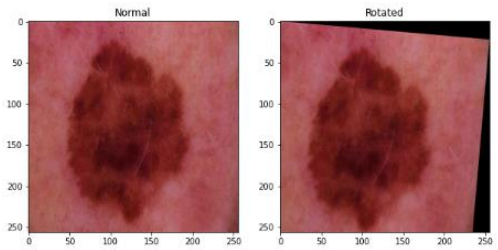


Figure 4.11: Image Rotation

Image cropping

Image cropping is most common image transformation technique. Image cropping is dropping particular portion of image and considering the rest as resultant image. It is used for removing unwanted part of the image. Some of the above mentioned techniques introduce the blank part in the resultant image, such parts are usually unwanted and can be removed by cropping. Figure 4.12 shows cropped image alongside with original image.

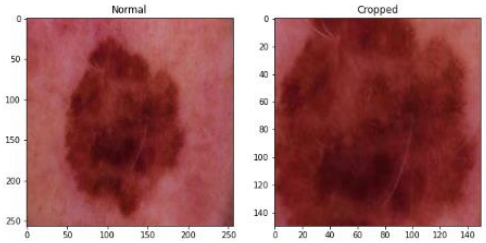


Figure 4.12: Image Cropping

4.5 Generative Models

Generative models are the most experimented during this research work. Several strategies are tested with different sets of hyper parameters. Mainly two types of generative models are explored in this research, Autoencoders and GANs.

A general step is performed in addition to data pre-processing and for a pre-requisite for generative models is separating the training images into multiple subsets based on the skin tone colour, shape, and size of the lesions and whether the image is square or round. This is to make sure that the models can capture the details regarding the background colours and lesions more

accurately. Different instances of the generative models are trained on these subsets. However, after every epoch, the resultant generated images are taken together for validation purpose.

Figure 4.13 showcases training images segregated into different subsets. These sets are being segregated based on the skin tone. In case of need in any subset, manual over sample has performed only during the training of generative models. While during the classification task, oversampling has been done by generated images only.

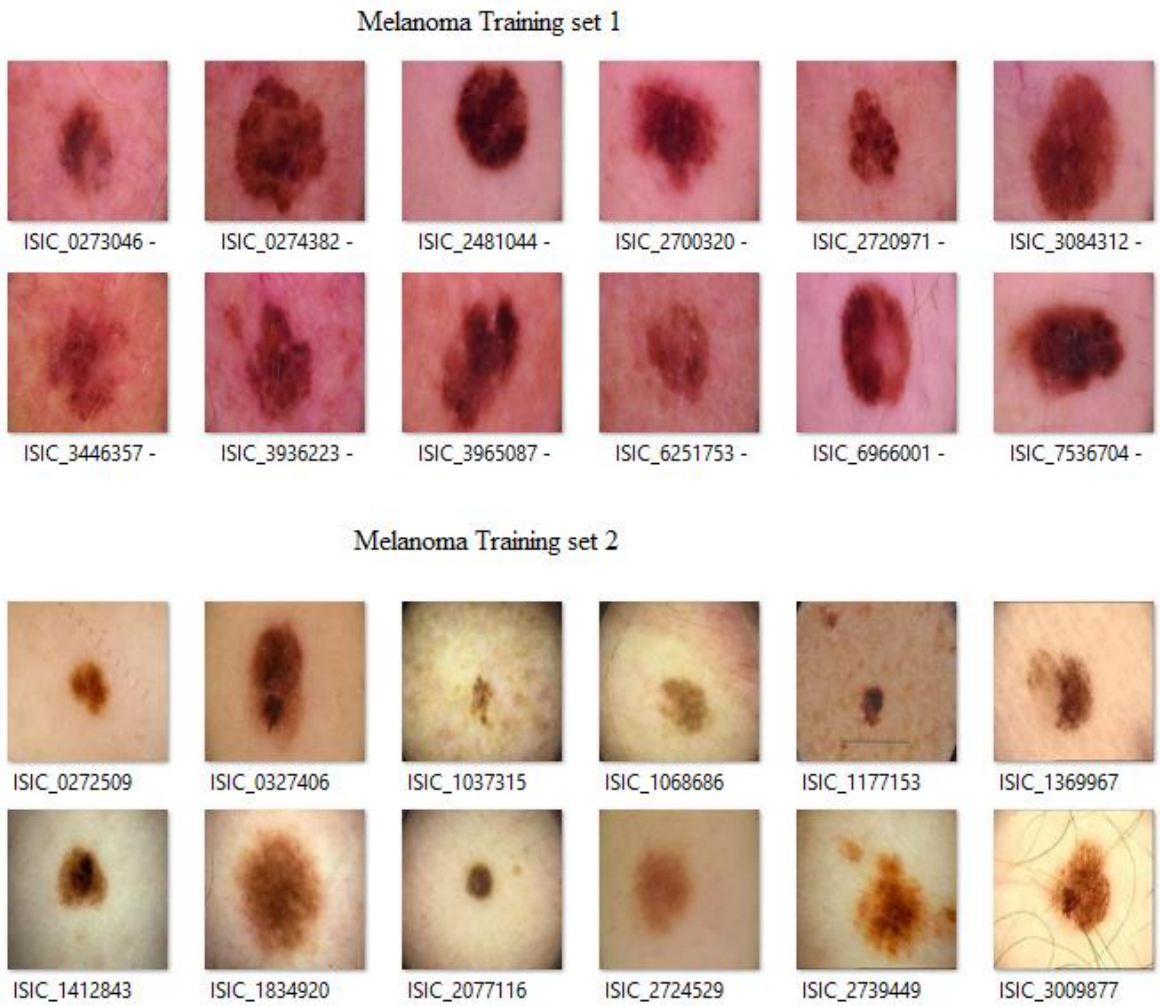


Figure 4.13: Different subsets of melanoma skin lesions

4.5.1 Autoencoders (AE)

Autoencoders are basic generative models with two components, Encoder and Decoder. As names suggest, encoder encodes the input to lower dimension global feature vector (GFV),

whereas decoder uses this GFV to decode it back into the output that is similar to the input. The more output is similar to input, the better AE is performing.

In simple AE, both encoder and decoder are trained together on single loss mechanism, unlike GAN. Thus, these two are not tuned separately and doesn't hold separate pairs of hyper parameters. Instead, Experiments with standalone Autoencoders has been performed based on variations in the hidden layer and input image resolution.

Hidden layers variations are done with respect to the type of hidden layers – be it a convolution layer or linear hidden layers as well as number of neurons in the hidden layers. Figure 4.14 shows different hidden layers configurations keeping the input image resolution constant at 128x128.

As it is seen in figure 4.14, Conv2d, ConvTranspose2d, and Max-pool layers changes the output size. This change is outsize can be calculated by the mathematical formulas,

Conv2d,

$$H_{out} = [(H_{in} - k + 2p)/s] + 1 \quad (8)$$

ConvTranspose2d,

$$H_{out} = (H_{in} - 1)s - 2(p) + dialation(k-1) + output\ padding + 1 \quad (9)$$

Max-pool,

$$H_{out} = \{[H_{in} + 2(p) - dialation(k - 1) - 1]/s\} + 1 \quad (10)$$

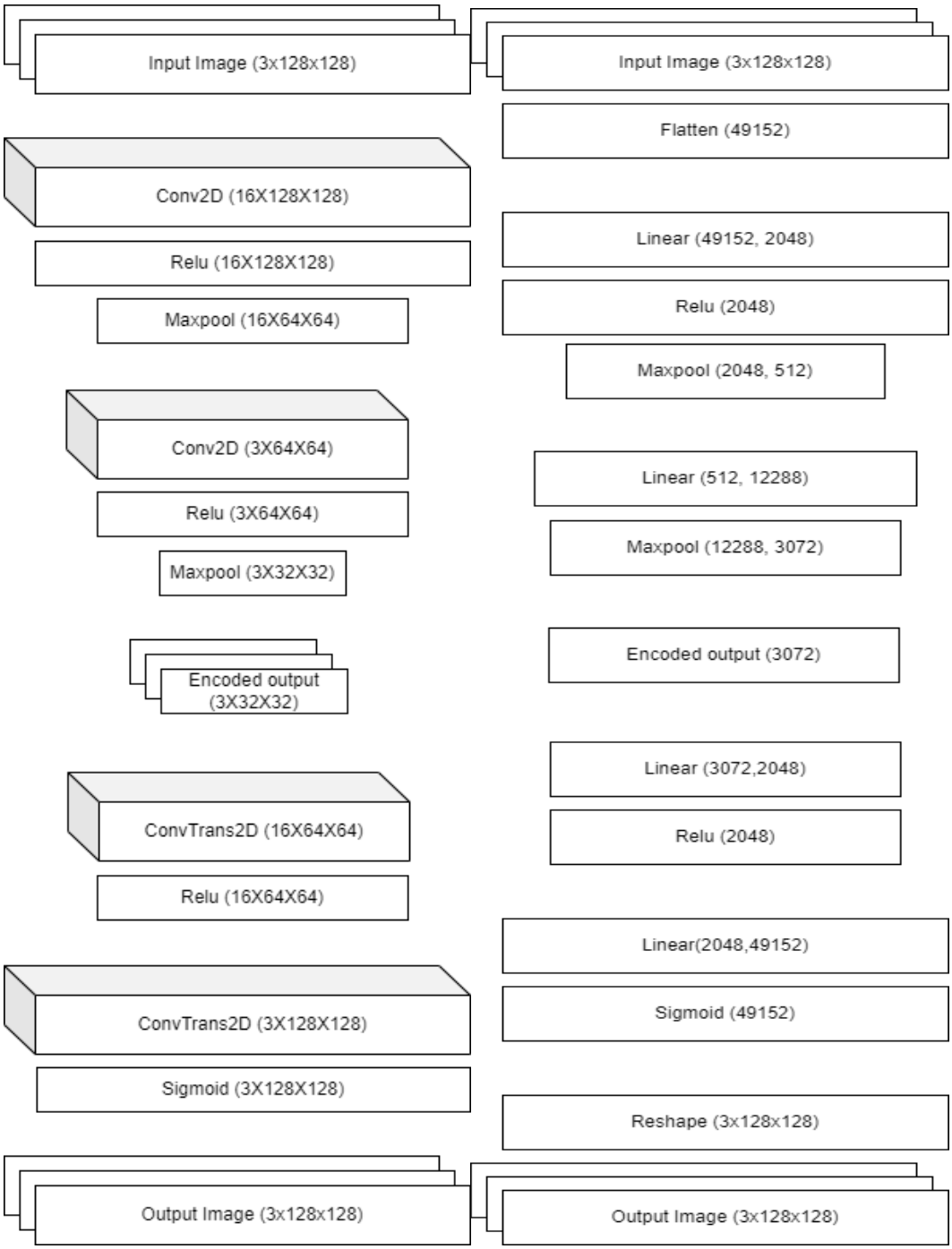
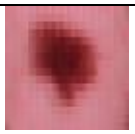
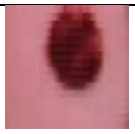
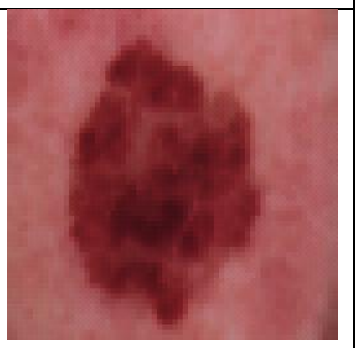
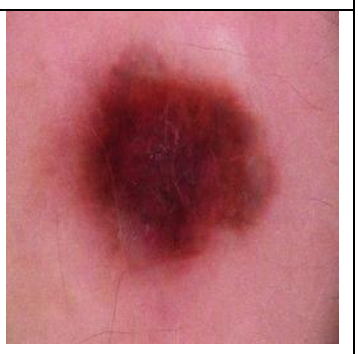


Figure 4.14: AE architecture with different types of hidden layers

Other than they type of hidden layers, number of neurons, input and output image size, encoded representation (GFV), number of epochs, and learning rate can be tuned to get the fine result. Table 4.2 shows the resultant output image across various hyperparameter tuning.

Table 4.2: Outcome of Autoencoder with different set of hyperparameters

Image size	Learning rate	Layer	Epochs	GFV Size	Time taken per epoch	Output
64x64	0.001	Linear	50	3X16X16	15.78 sec	
64X64	0.001	Non Linear	100	3X16X16	1.99 sec	
64X64	0.05	Linear	50	3X16X16	15.36 sec	
64X64	0.01	Non Linear	50	3X16X16	2.04 sec	
128x128	0.001	Non Linear	100	3X32X32	4.7 sec	
128X128	0.01	Non Linear	75	3X32X32	4.69 sec	
128X128	0.01	Linear	50	3X32X32	37.48 sec	

128X128	0.05	Non Linear	75	3X32X32	4.8 sec	
256x256	0.001	Non Linear	100	3X64X64	9.9 sec	
256X256	0.05	Non Linear	75	3X64X64	10.26 sec	
256X256	0.01	Non Linear	75	3X64X64	10.24 sec	
256X256	0.01	Linear	10	3X64X64	162 sec	

An important point to note here is that the way AE architecture is designed that it can work with different image resolution without changing the AE model itself. With increasing the image

resolution, the GFV size is increasing. This helps in keeping the required time to train the model constantly low. Meaning increasing the image resolution, doesn't drastically impact on the computational cost and time, rather it impacts on the memory required to hold and process the GFV.

In the integrated system, The AE being used in, is pre-trained and thus it doesn't add much time and processing during the training of GAN. However, it is important to get to know the time required to train the AE in first place. Table 4.2 also talks about the time taken by AE training with different set of hyperparameters.

This research is not aiming to use Autoencoders standalone but to be used with GAN as input to GAN. Further in coming sections, first experiments with GAN are discussed independently and then different ways of integrating AE with GAN. Figure 4.15 shows the progressive image generation across 100 epochs. The development of fine grained images can be seen as number of epochs increases.

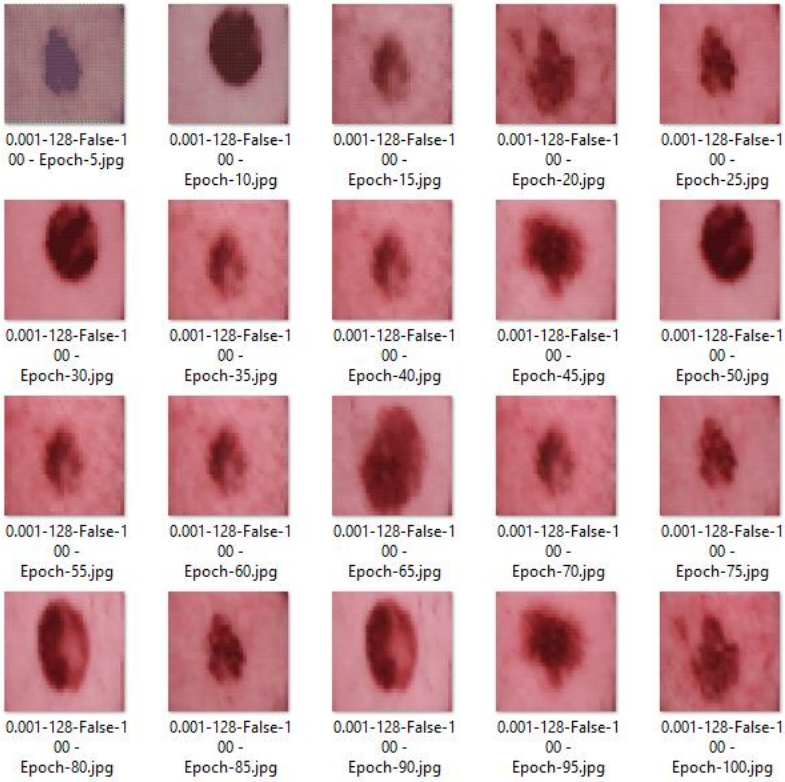


Figure 4.15: Progress of image generation using AE for 128X128 image through 100 epochs

4.5.2 Generative Adversarial Network

Generative Adversarial Network is more sophisticate deep generative model. It is made of two networks, Generator and Discriminator. Role of the generator network is to learn the data distribution in the training images and mimic them to generate synthetic real looking images out of random noise provide as input. Whereas discriminator network has to correctly distinguish the fake images that are generated by generator from real images. In a way, both generator and discriminator work as competitor during the training period.

Inverse of discriminator's loss is loss of generator; thus, both the nets are being trained of the loss of discriminator. Other than this little coupling, unlike the AE, in GAN, generator and discriminator can be trained on different parameters and with different speed. Also, once the training is done, only generator network is used to generate the synthetic images.

Similar to the experiments done with AE, GAN hidden layers variations are done with respect to the type of hidden layers. Figure 4.16(a) and Figure 4.16(b) shows different experimental configurations of GANs. In this diagram as well, the image resolution is kept constant at 64x64. Although the experiments have been performed for other image resolution size as well.

The architecture of both Generator and Discriminator networks can be seen in the figure 4.16(a). It also expose the hidden layer and neuron configuration. This architecture is set to run on the training images with the resolution 64X64. However, examining the architecture properly, it is clear that the same architecture can be extended further for 128X128 images as well as 256X256 images by simply extending the generator network at bottom and discriminator network at the start point. This is possible as the number of channels, i.e., the depth of the layer to increase or decrease is done separately, and the size of the channel is directly being calculated by the mathematical formulas.

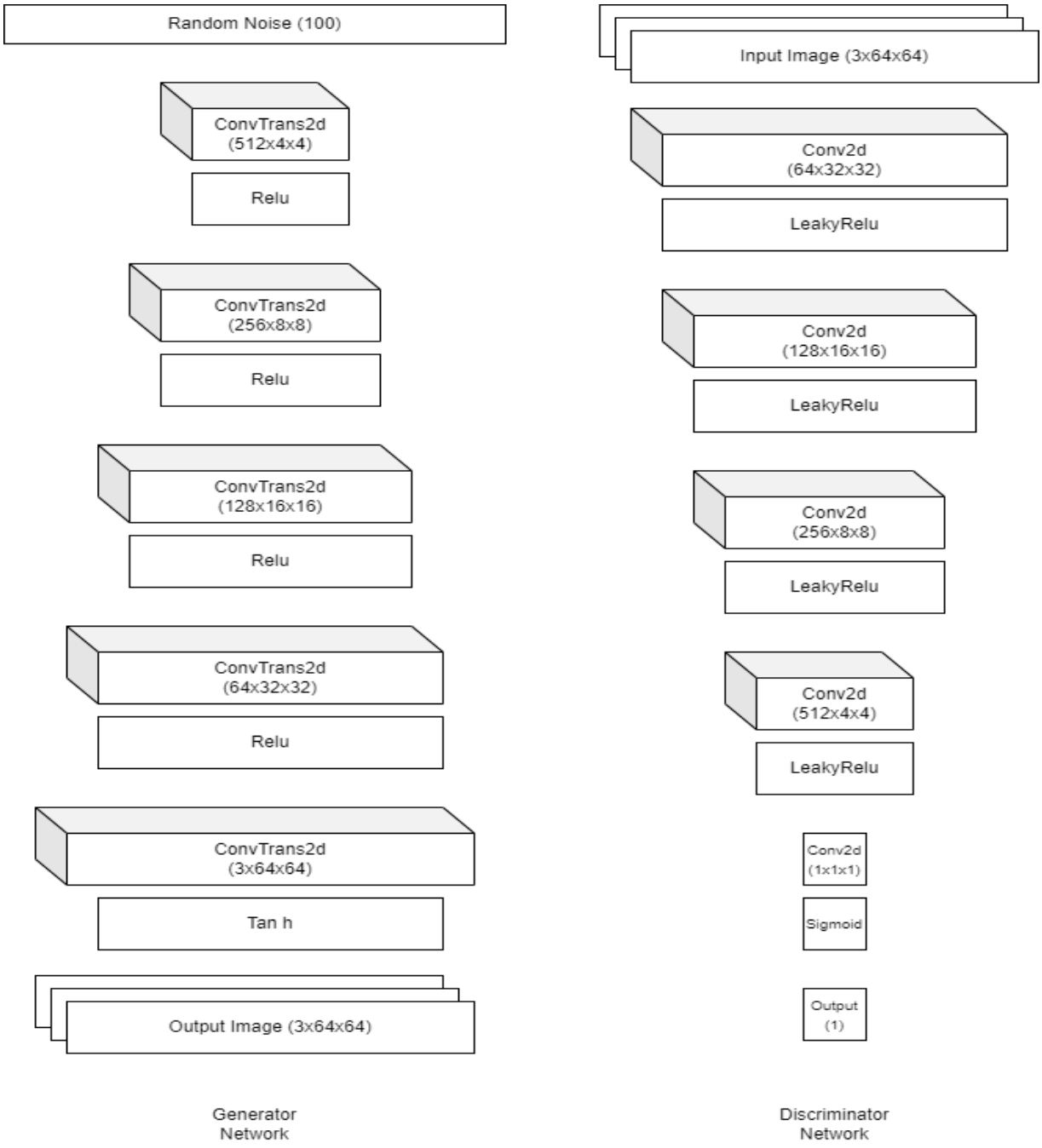


Figure 4.16(a): Experimental setup of GAN with Convolution and Trans-Convolution layers

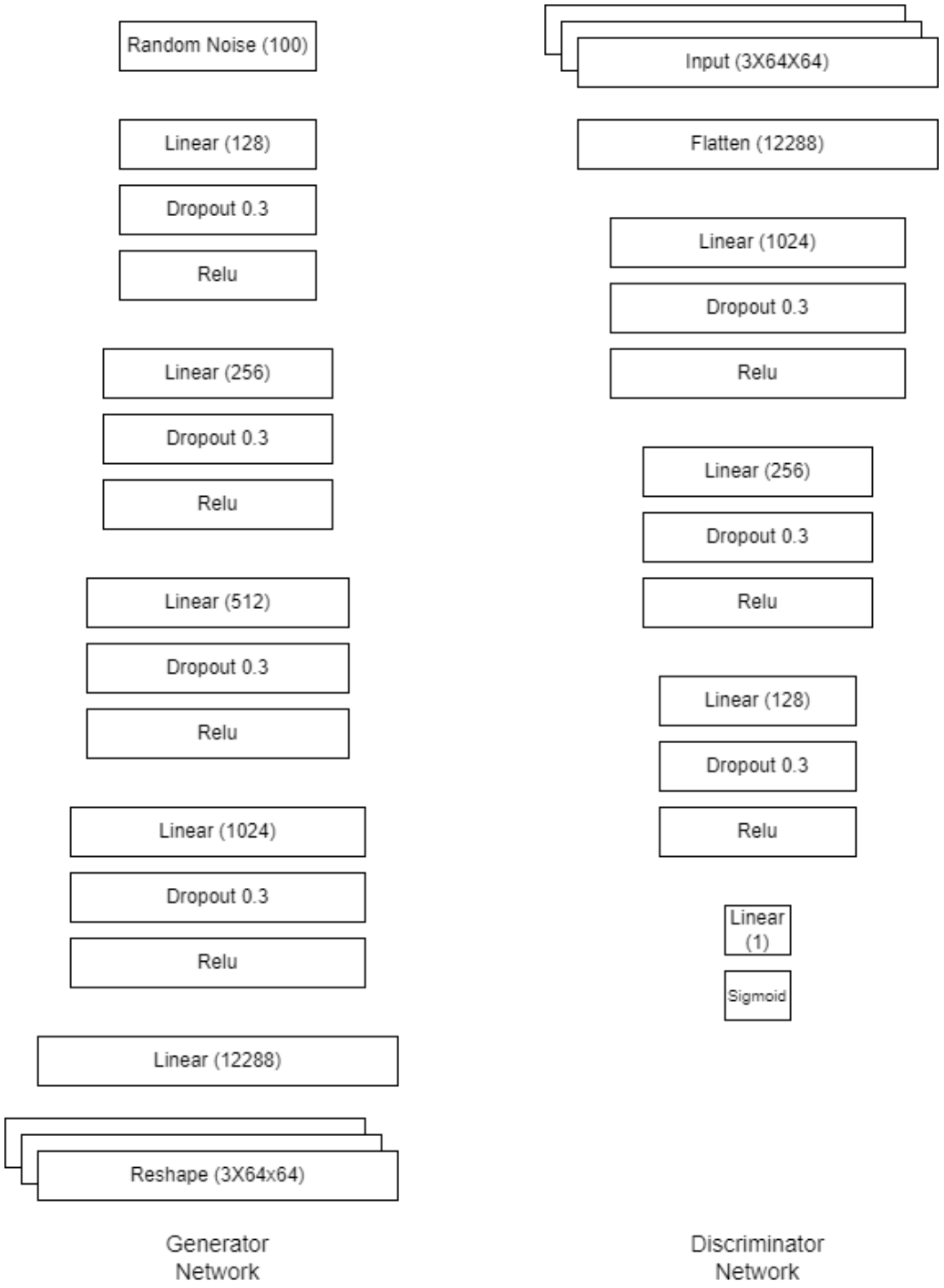


Figure 4.16(b): Experimental setup of GAN with Linear layers

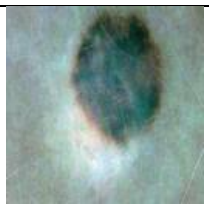
Figure 4.16(b) on other hand demonstrates the GAN architectures developed using Linear hidden layers. Unlike the architecture with Conv Layers, liner layer architectures are not

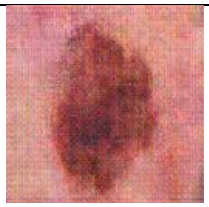
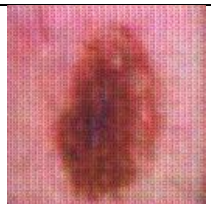
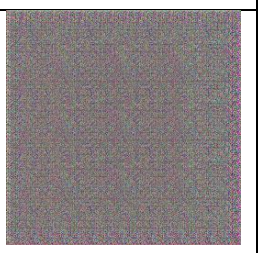
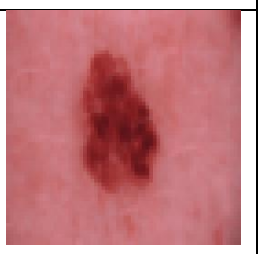
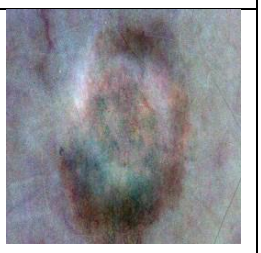
directly extendable but proper calculations are needed while adding the layers at the beginning or at the end of the layers.

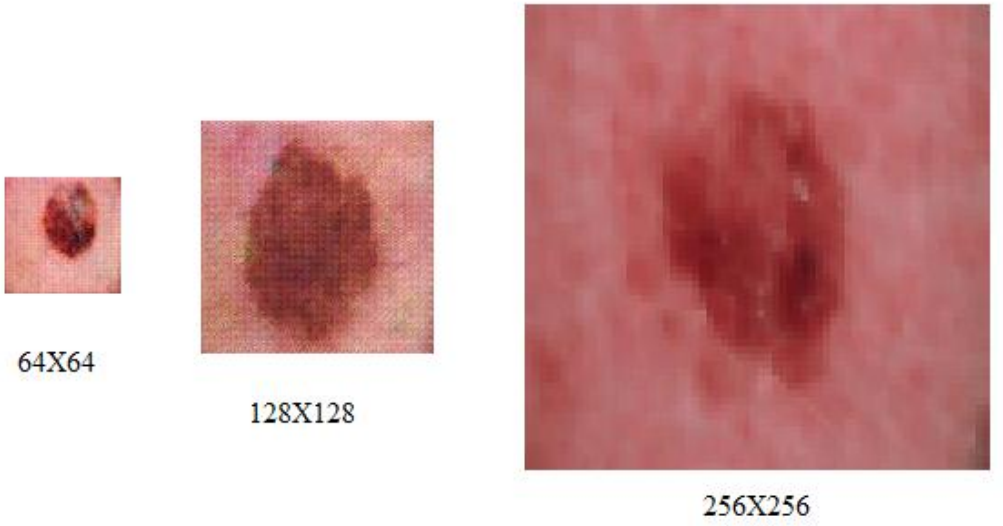
There is one more draw back of a model with linear hidden layers is, it requires significant more amount of memory for the computation as well as it takes more time as it is computationally more complex. However, linear models need less epochs to converge to lower loss and produce better quality images. These characteristics make them worth considering in experiments.

Table 4.3 shows various experiments carried out with different setups of GAN and outcome images of trained generator.

Table 4.3: Outcome of GAN with different set of hyperparameters

Generator Input size	Generator Output size	Gen. Learning rate	Disc. Learning rate	Gen. layers	Disc. layers	Epochs	Time taken per epoch	Generated image
100	64	0.0001	0.0001	Linear	Linear	150	14.22 sec	
100	64	0.005	0.001	Linear	Linear	150	10.08 sec	
100	64	0.0001	0.0001	Non Linear	Non Linear	150	26.00 sec	
100	64	0.001	0.001	Non Linear	Non Linear	150	21.6 sec	
300	64	0.005	0.001	Non Linear	Non Linear	150	23.02 sec	
100	128	0.005	0.001	Linear	Linear	100	33.56 sec	

100	128	0.001	0.001	Non Linear	Non Linear	150	38.4 sec	
100	128	0.005	0.001	Non Linear	Non Linear	100	37.8 sec	
300	128	0.0001	0.0001	Non Linear	Non Linear	150	38.8 sec	
300	128	0.005	0.001	Non Linear	Non Linear	100	36.6 sec	
300	256	0.0001	0.0001	Non Linear	Non Linear	150	96.8 sec	
300	256	0.05	0.01	Non Linear	Non Linear	100	95.7	
300	256	0.005	0.001	Linear	Linear	100	94.8 sec	



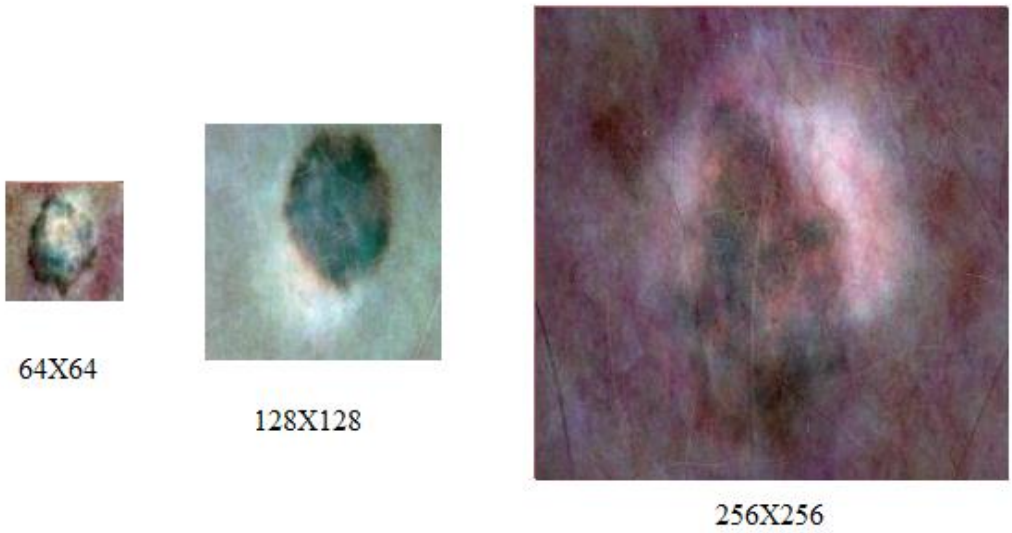


Figure 4.17: (a)Output of Non Linear GAN on different resolution of training images

(b) Output of Linear GAN on different resolution of training images

As it is seen in the table 4.3, the experiments carried on lower learning rate doesn't show any progress whereas comparatively higher learning rate shows good progress. Unlike the AE, increasing the image resolution that generator have to produce, and discriminator have to consume, more hidden layers need to be added or existing layers need to be adjusted with the

current size of the image. Meaning, the computational cost in GAN is more sensitive towards the image resolution in comparison with AE.

In addition to table 4.3, figure 4.17 properly demonstrates the generated images by the GAN in different image resolution. In case of normal AE, the generated images were exact same as the training images. This behaviour is good in many terms but in the context of this research, it limits the diversity of the dataset. However, with GANs, the generated images are not exact same as the training images and hold the degree of diversity. It is also important to note that, the diverse images should stay realistic else it will end up spamming the dataset with unwanted and unreal images, that can cause the performance of detection model even further degrade. While the generated images during the experiments are both diverse and realistic. Especially in case of the images generated from the linear GANs have preserved the textures and details significantly better.

4.5.3 Integrating AE with GAN

In the previous sections of this research, AE and GAN are discussed separately and also what all tuning is carried out on both of them separately. This section talks explicitly on the integration of AE with GAN. Several ways of integrating the AE in GAN are discussed with the possible tuning and outcome.

The motive behind leveraging the ability of AE with GAN is to liberate the GAN from the limitation of resources. It is understood that GAN holds to independent networks that are trained with separate set of weights and biases. Thus, resource requirement of the GAN for the training image dataset with the same resolution is high in comparison with AE. This is explained in detail in the next chapter with proper data gathered during the experiments and research implementations.

Talking about approaches of utilizing the pre-trained GAN, figure 3.8 describes basic flow of possible integration of AE with GAN. In the figure 3.8, it is demonstrated that a noisy image is fed into the pre-trained encoder that generates noisy GFV. And GAN is trained on that noisy GFV to generate clean GFV that can be fed into decoder to obtain the real looking image which goes into discriminator for validation. Figure 4.17 provides details insight on this approach. For making the approach simpler to understand, the Figure 4.17,4.18, and 4.19 have considered the

image resolution as 64X64. However, image resolution for this experiment has been taken from the results of prior experiments carried out on AE and GAN independently.

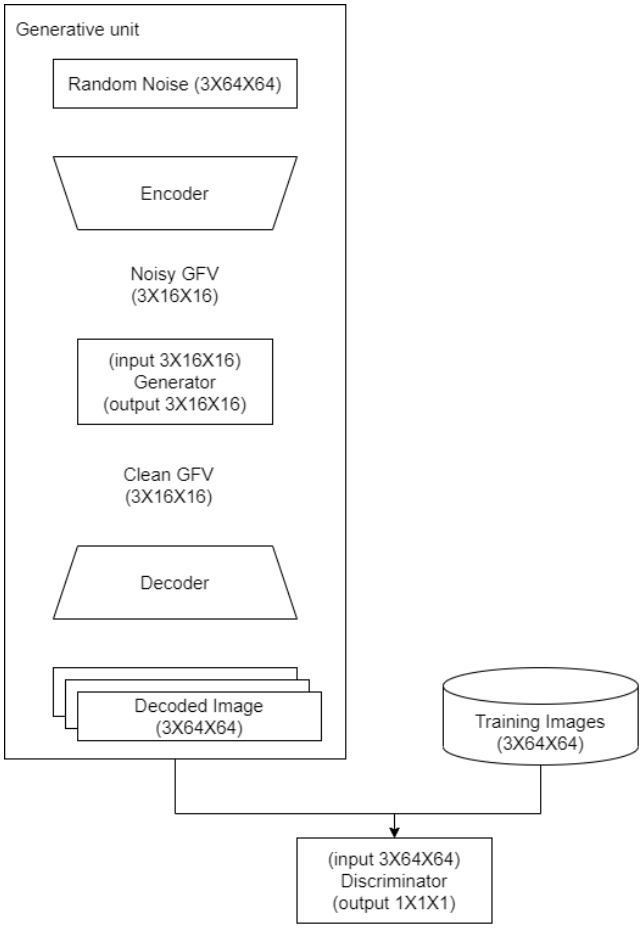


Figure 4.18: Integrating AE with GAN – Approach 1 – Generator to produce clean GFV from noise GFV

However, this isn't the only way to integrate AE with GAN, there are several other possible ways to achieve the same. Figure 4.18 demonstrate another approach where, similar to conventional GAN design, the generator is fed with random noise. The difference here is that the task of generator is not to produce the fine grained image, but the clean GFV which can be fed into decoder part of the AE. Once the decoder decodes the input GFV into the image, that resultant image is then supplied into the discriminator network. Based on the loss of discriminator network, the generator is fine tuned to generate more accurate GFV. Here the performance of AE is important, and it is assumed that AE is able to produce the best quality images if provided appropriate clean GFV. The main benefit of this approach is as generator is

having random noise as input instead of high dimensional noisy GFV, this makes the overall architecture less complex, Also, it keeps the GAN performing the tasks that are similar to original GAN architecture.

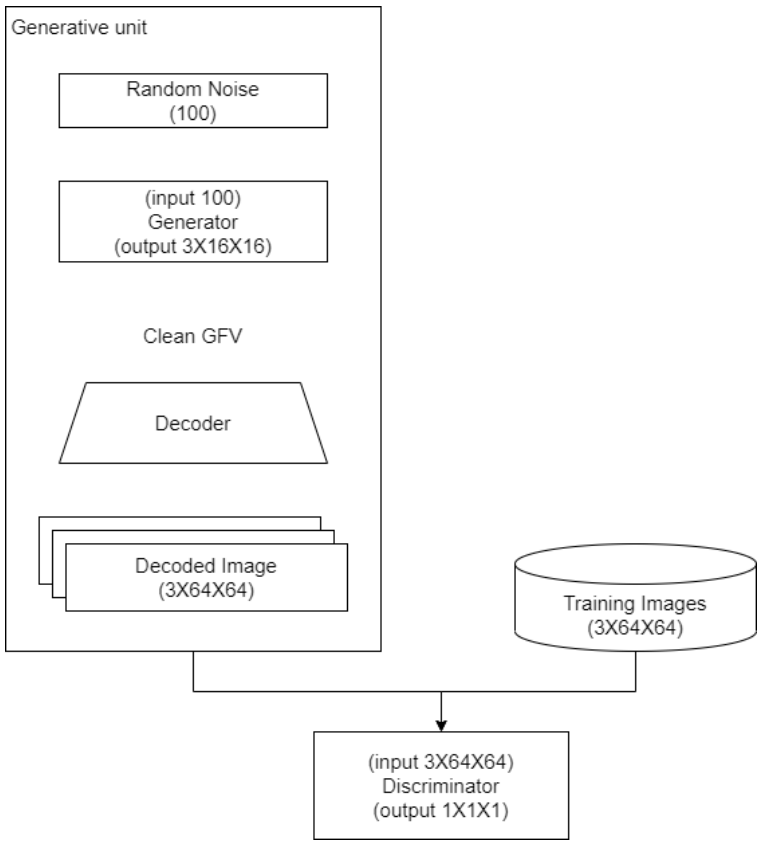


Figure 4.19: Integrating AE with GAN – Approach 2 – Generator to produce clean GFV from random noise

The GANs are not necessarily always designed to generate images. The simple idea behind the GAN is, generator learns the data distribution pattern in the training data and discriminator validate how good the generator is able to mimic the training data. Given this, it is also possible to utilize the GAN to learn the GFVs only rather than learning the image data itself. Figure 4.19 shows the mentioned approach. Here as well similar to conventional GAN design, the generator is fed with random noise. Also, similar to the approach mentioned in figure 4.18, the task of generator is the clean GFV.

However, unlike the first two approaches, here the clean GFV isn't fed into the decoder but, it is directly fed into the discriminator. Ground truth for the discriminator is the encoded GFVs of the training images. This GFV are generated from the pre-trained AE network. Similar to an earlier approach, based on the loss of discriminator network, the generator is fine tuned to generate more accurate GFV. Computationally, this approach is the most optimum in comparison with the other two. In this approach, Generator does not have to learn the data distribution of the training images as well as, the discriminator does not have to discriminate the higher dimensional images from real to fake but the fake GFVs from the real ones. However, the overall performance is highly dependent on the performance of AE.

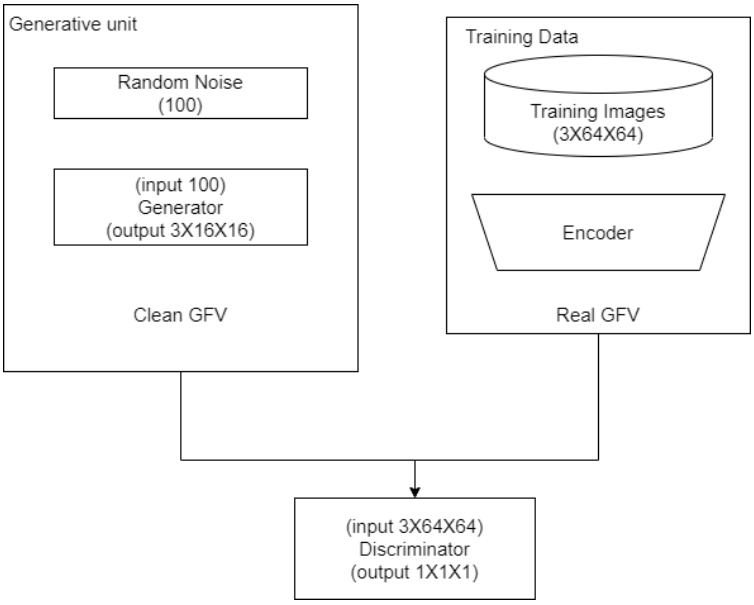
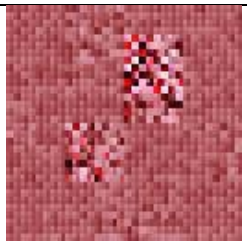
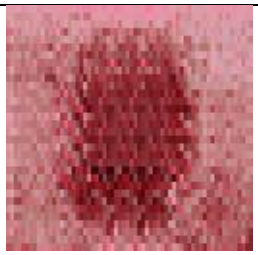


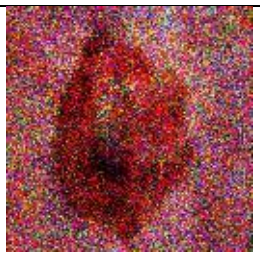
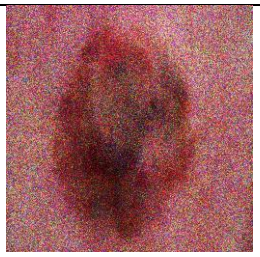
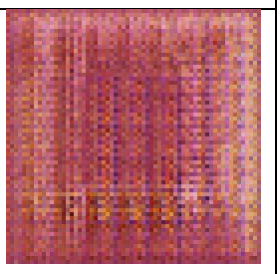
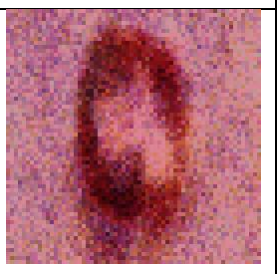
Figure 4.20: Integrating AE with GAN – Approach 3 – Generator to produce clean GFV Discriminator to consume the GFV directly.

All three approaches have some advantages and some disadvantages. Thus, it becomes important to perform fair enough different experiments on them to validate their performance in different given hyperparameters. Table 4.4 shows various experiments carried out with different set of hyperparameters and approaches. Table 4.4 also shows the time taken in each experiment and the resultant generated image. Here only output images are not taken in consideration, but the space and processing cost are also considered. From the tests carried out independently on the AE and GANs, now the better performing standalone models are known.

Integrated system of AE and GAN are tested with different setups and the selected AE and GAN models.

Table 4.4: AE and GAN integration results with different set of parameters and approaches

Appro-ach	AE layers	GAN layers	epochs	Gen. input size	Gen. output size	Time taken per epoch	Output image
1	Non linear	Non linear	100	3X32X32	3X32X32	38 sec	
3	Non linear	Non linear	100	3X10X10	3X32X32	24.6 sec	
3	Non linear	Linear	75	300	3072	5.6 sec	
1	Linear	Linear	75	3072	3072	24.8 sec	
2	Linear	Linear	75	300	3072	20.8 sec	

3	Linear	Linear	50	300	3072	8.8 sec	
3	Linear	Linear	150	150	12288	20.8 sec	
3	Non Linear	Non Linear	150	300	12288	26 sec	
3	Non Linear	Linear	150	300	12288	12.4 sec	

As it is clear from the observations collected in the table 4.4, the approach three is more suitable for integration of AE with GAN. As in approach one, the input of Generator is significantly higher and in approach one and two, generated GFV needs to pass through the decoder to get the resultant image because the discriminator is designed to consume the images only, these two approaches take significantly more time in training.

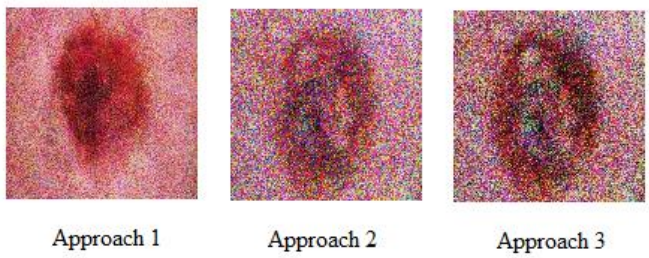


Figure 4.21: Generated images from different approaches of AE and GAN integration.

Figure 4.21 shows the generated images from different approaches. Even though there is no significant improvement noticed in the approach 3, however, as it is seen in Table 4.4, approach 3 is more efficient and still can produce the better images than approach 1 and 2. On other hand figure 4.22 clearly shows the quality of generated image using approach 3 if being generated for higher resolution.

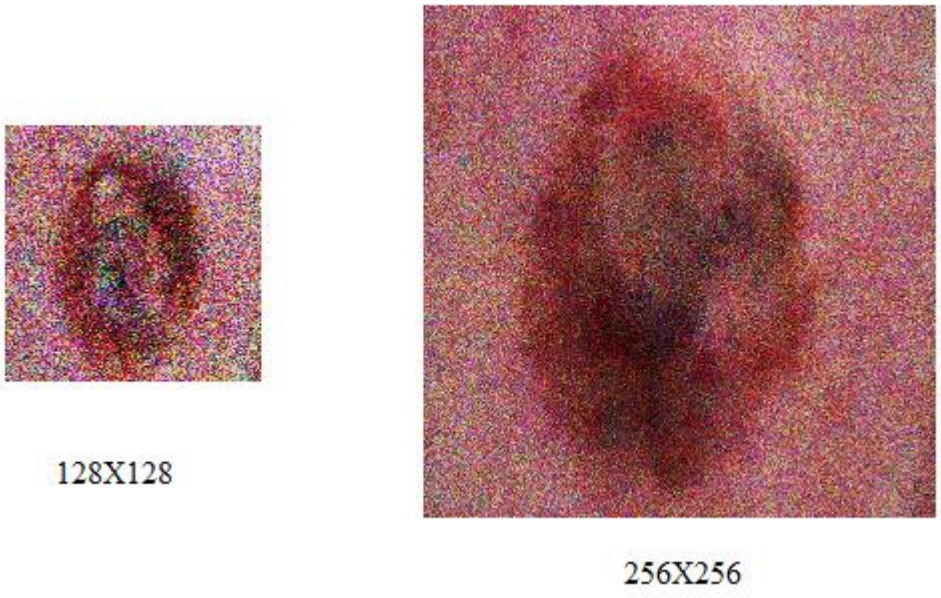


Figure 4.22: Generated images for different resolution using approach 3 of AE and GAN integration

4.6 Utilizing Reinforcement Learning

So far, the analysis and experiments of Autoencoders, GANs and integration of both are discussed. This section explicitly talks about the possible ways of utilizing the reinforcement learning in generating the input for the generator network. In conventional form of GAN, generator accepts the random noise as the input, and it only learns based on the loss factor or discriminator network. Meaning, in normal, the input of generator network doesn't hold the significance.

This section of the research explores the possibility of Instead of using random noise if a particular input vector (or matrix) is used as input of generator network and observes the

outcome of this test. It has been tested in the research (Ukwuoma et al., 2021) work that in case of portion recovery of the image, RL with GAN has been proven helpful. However, in the context of this research work, GAN needs to generate the whole image out of nothing.

For Reinforcement learning to be helpful in decision making, it requires huge number of occurrences of particular state so that it can tryout all possible actions from that state to come up with optimum policy. And as the state occurrence become rare and number of possible actions becomes more, the convergence of Q-learning becomes more and more difficult. The same challenge is present for the current research work as well.

To test the concept, instead of trying RL on 128X128 image and Generator accepting random noise of 3X10X10 size, the experiment with RL is carried out for 64X64 grayscale image with GFV size 8X8 and thus, the random noise size is kept 3X3. This configuration isn't suitable for the context of main objective of this research and is only to test concept.

Two possibilities of utilizing RL

1. With pre-trained GAN and AE, utilizing RL to further finetune the images generated by the GFV produced by generator.
   Considering 8X8 GFV as a state vector and only an integer numbers be possible in the state, in that case as well 64! different combinations are possible for a state vector. Although in a concept a proper Q-learning table is possible with each state-action pair has proper convergence on reward. But practically this is very challenging and is not feasible with the available resources.
2. During the training of GAN, utilizing RL to get the input for GAN.
   In this approach, the GAN is likely to get trained by itself before RL can have a proper Q-learning table populated with all the possible state-action pairs.

4.7 Classification and Early loop breaker

As improving at classification is not the direct objective of this research work, experiments on classification task is limited to obtain a stable classification model. The only experiment to be carried out with respect to building the classification model is to build two models, one which can classify for four classes and the other which can classify 5 classes. This is to include and exclude the "unknown" class. Further research is comparative study of the result obtained by this classification model using different datasets that were arranged by different means and

strategies. However, main focus on the experiment perspective is on the loop breaker mechanism. It is tested with several possible hyperparameters setups.

In general, the classification is performed after the GAN training completes. This prevents running the classification task multiple times. However, if the GAN is able to produce the fairly good enough images during the training itself, then too training continues till the hard break, i.e., till the end of the defined number of epochs. Also, if continuing the training degrades the performance of GAN, then classification task has to suffer the loss. To avoid such issues, early loop breaker mechanism is introduced as a soft break. Hyperparameters to this mechanism provides the flexibility as well as to keep running the classification for every epoch when the GAN is obviously has not been able to produce good enough images.

Figure 4.23 showcases the architectural diagram of the classification model that is being used to validate the quality of the augmented dataset by classifying the images into proper classes. As it can be seen in the figure, both classification models are designed to consume 128X128 images. This is highbred model that consist of both Convolutional layers as well as linear layers. Maxpool layers are used to reduce the dimensionality as the architecture goes deeper. After every set of hidden layers, “relu” activation function has been used to bring the output of the layers under certain limit whereas at the end “softmax” layer is used to generate probabilistic output.

Once trained, the classification model is kept constant across all the further experiments and evaluation. However, loop breaker mechanism needs to be tuned to finalize.

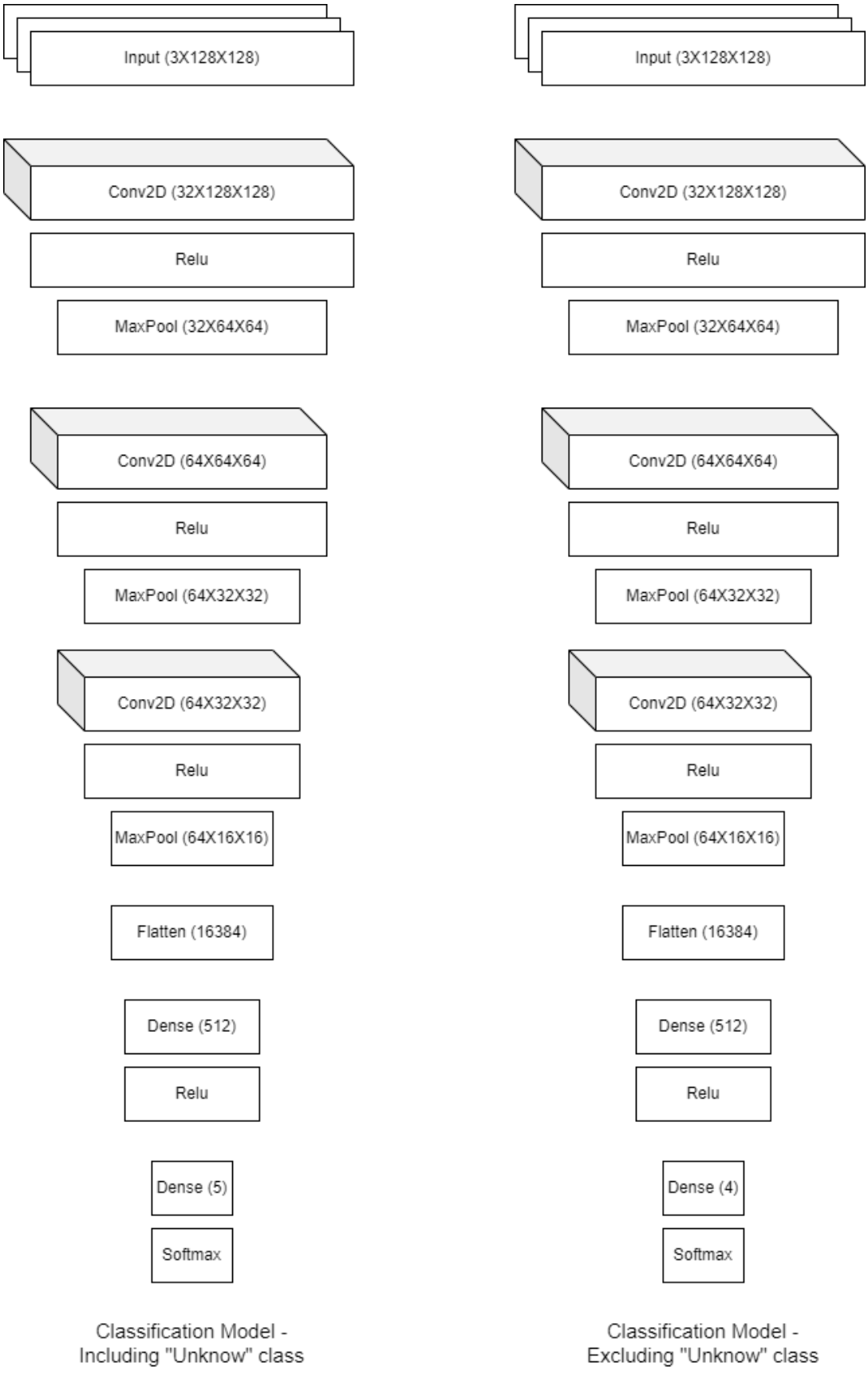


Figure 4.23: Architecture of Classification model.

Early Loop Breaker

As demonstrated in figure 3.11, the loop breaker mechanism is proposed originally in this research to prevent unnecessary and unwanted drag in training of GAN network. Extending the explanation, figure 4.24 describes how early loop breaker mechanism attaches the GAN network with image classification network.

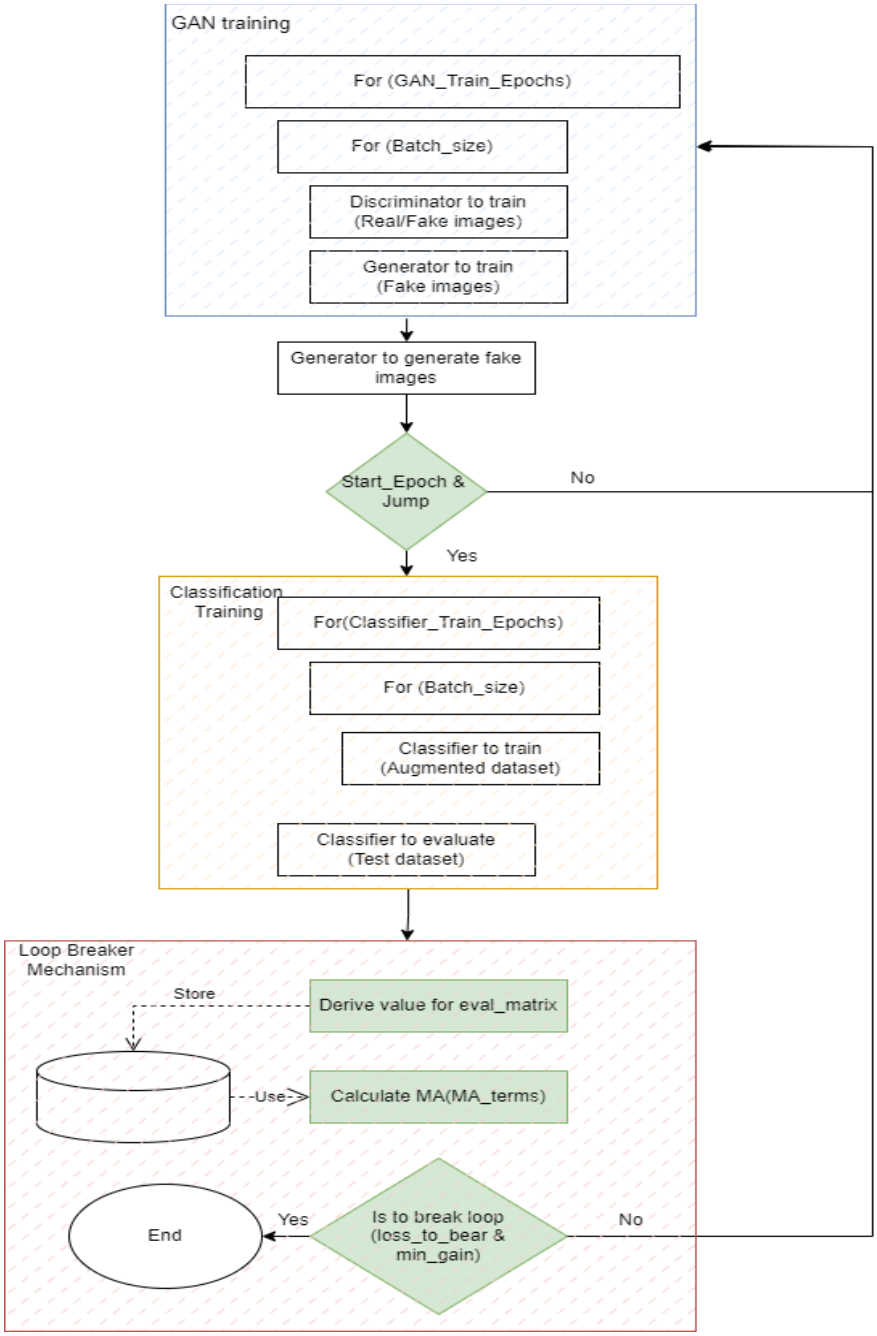


Figure 4.24: Integration of loop breaker mechanism in GAN and classification model training.

Hyperparameters for early loop breaker mechanism:

In figure 4.23, the attributes in the boxes that are with solid green, are acting hyperparameters for loop breaking mechanism. It is clear to notice that those attributes are key factors in decision making and thus understanding them and tuning them as per need is important.

- Start epoch:

Start epoch defines the number of epochs in GAN training from which the loop breaker mechanism should start checking the performance of the GAN. It is obvious that in initial stage of GAN training, the GAN won't be able to generate good quality images. This parameter is to avoid running classification at this stage.

- Jump:

Jump suggests the number of subsequent epochs to skip in taking into consideration for the loop breaker mechanism after one successful check. This is to prevent the overhead of performing the classification task after every epoch.

- Moving average term:

This parameter, as its name suggests, is fed into moving average function. This is to set the configuration for moving average that how many recent datapoints are to take to calculate the average value that needs to be checked against the current observation.

- Loss to bear:

This parameter acts as percentage stop-loss function for the loop breaker mechanism. It defines the stopping condition. Meaning, if the current evaluation of classification on GAN's augmented dataset is significantly lower than the moving average, then it will be considered as minima and GAN training loop will be stopped. Because of its criticality, this term needs to be tuned properly as per the nature of the training images.

- Minimum required gain:

This parameter is also for stopping condition similar to “Loss to bear”. However, instead of checking if the trend is going downwards, it is to make sure, the trend is going upwards. This parameter suggests the minimum amount of percentage gain that is needed for the GAN training to continue. If the GAN augmented dataset fails to outperform the previous dataset, it is assumed that the dataset quality has reached to saturation and no further improvement is obtained by continuing the GAN training. This parameter is optional.

- Evaluation Matrix:

This parameter takes two possible values, “accuracy” or “sensitivity”. Based on the value passed the classification model’s outcome is evaluated and the result of the evaluation is stored to compare with subsequent evaluations.

4.8 Summary

This chapter extensively talks about the different techniques and tests that are carried. In order to get better understanding of the data, and models to be used as a final implementation. It also discusses the outcome of the tests.

Chapter starts with the discussions around various steps to understand the data. Based on the understanding and EDA findings of the dataset, actions to perform for data preprocessing is mentioned. As being in a center of the research, techniques of images augmentation are tested though different experiments and hyperparameter tuning. Overall experiments examined four different methods of image augmentations in detail, traditional image transformation, image generation using autoencoders, image generation using GANs, and a proposed method of image generation using integrated system of AE and GAN. All the experiments are carried out on different sets of hyperparameters and on different image resolution. The chapter then talks about the challenges presents in the proposed method of utilizing the reinforcement learning on the top of the integrated system.

Later in this chapter, it talks about finalized architecture of classification model and how the proposed mechanism of “Early loop breaker” integrates the image generation and image

classification as single unit. At the end, a brief idea of different parameters of loop breaker mechanism is mentioned.

In the next chapter, the findings of the experiments, the intuition being made out of it, and the finalized approach and its results are discussed in a grate detail.

5. Results and Discussions

So far in this research, the motive and the background for the current research work, related works in the area, proposed methods and techniques, and experiments on the proposed methods are discussed in detail.

5.1 Introduction

While proposing, various research questions were formed to be answered as the outcome of this research work and various objective were defined to the direction of achieving the end goal of the research.

This chapter critically analyses the outcome of the experiments, examine, and discuss the results of the experiments on finalized set of parameters. This chapter also discuss if the end result is aligned with the defined objectives or not and if the research questions are answered.

5.2 Image Augmentation

Majority of focus of this research is on the various approaches of Image Augmentation. Various experiments have been carried out on different means of augmenting the images. In this section, results of the various approaches are compared visually as well as on similarity indices. Two major ways of image augmentation are discussed in the previous chapters. Traditional Image transformation, and Image synthesis by generative models.

5.2.1 Traditional Image transformation

Morphological image transformation techniques like Erosion and Dilation alters the thickness of the boundaries of the foreground features in the image and thus it alters the shape of the skin lesions and enhances or reduces the presence of the texture present in the image. On other hand, techniques like opening and closing, are extension of erosion and dilation operations performed in specific order.

If the dilation is performed on the image to enhance the almost hidden features and then erosion is performed to make the shape similar to the original image, it is called closing. Whereas opening is the opposite operation than the closing, where erosion is performed first to eliminate the noise in the image, and then dilation is performed to undo the shape change done by dilation.

The same way other techniques like morphological gradient, hats also do the combination of erosion and dilation. Point to understand here is, although these techniques are useful for image editing purpose and enhancing the feature of the images, but such techniques do more harm than the benefit in the context where the output images should be used in the classification. Because in that context, the output images are expected to look like the original images in terms of noise, color, feature etc.

So, it is understood that the morphological image transformations cannot be used in this context, However, some of the most basic image transformation techniques are still to be useful in the process of augmenting the image dataset.

For this research, after performing tests on various image transformation techniques, “rotation”, and “flipping” is used, whereas resizing the images to make all the images into uniform size is done as well.

Figure 5.1 compares the output images produced by image rotating and flipping with the original image and figure 5.2 demonstrate the pixel intensity distribution between the original and transformed image. It is clearly seen that even though the image transformation techniques help in up-sampling the dataset, it fails to introduce a degree of diversity in the output image in comparison with original image.

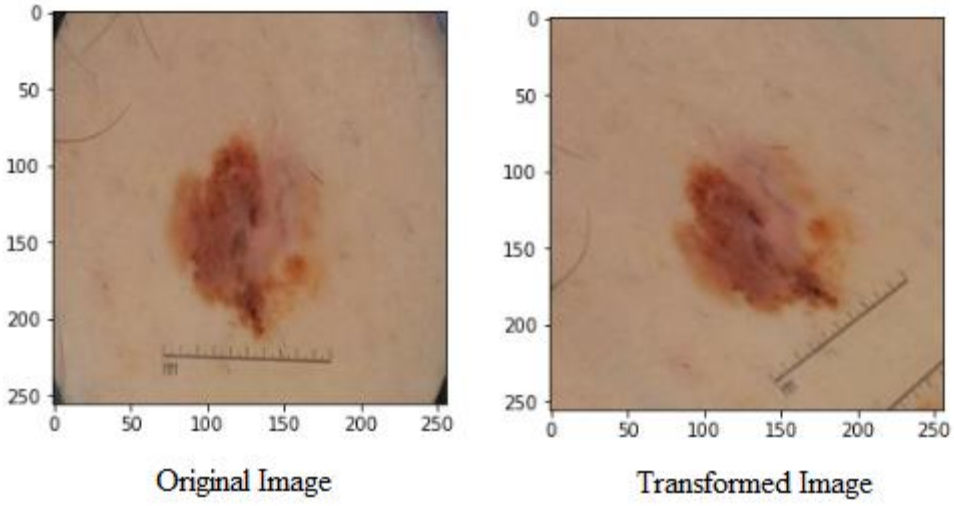


Figure 5.1: Image Transformation

As it is seen in the figure 5.1, the transformed image is both flipped and rotated in comparison with the original image. Due to rotation, the black parts on the corner in the original image is now missing in the transformed image. However, zooming in the image and checking the pixel density, it is clear that, not much change is carried out in the image transformation process.

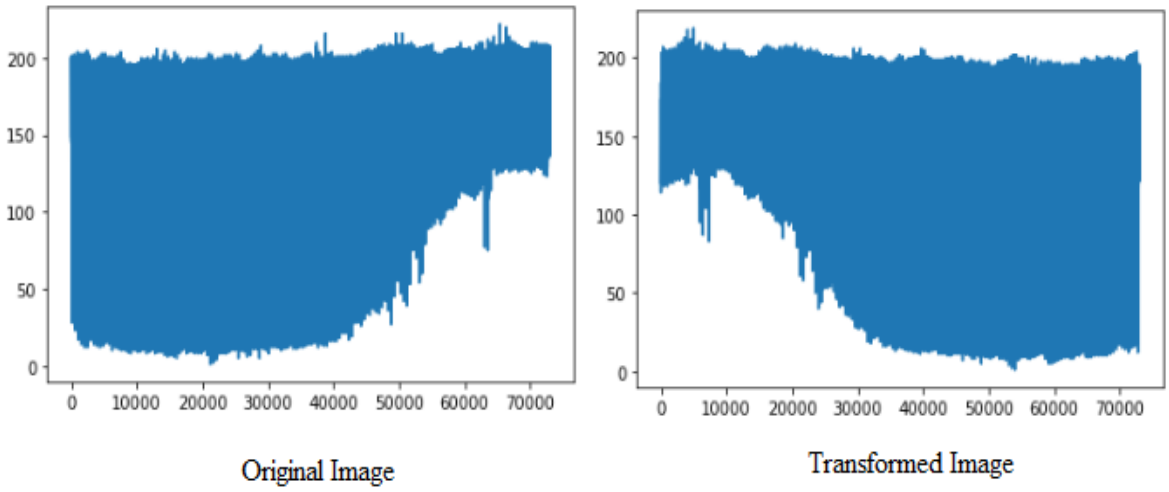


Figure 5.2: Pixel data distribution in original and transformed image

A preferred image to augment the dataset is the one which carries the general pattern of the data distribution but not be exact same. Such images can lead the model to overfit for certain images while during validation or test if the image contains variations, the model will fail to correctly classify.

5.2.2 Image synthesis using Generative Models

On other hand, as seen in previous chapter, image augmentation is possible using deep learning based generative models. Unlike the image transformation techniques, these models learn the data distribution of the training images and produce the synthetic images based on the learning. Thus, these models are capable of generating the images with variety of features learnt from multiple training images. In this research two types of generative models are explored, autoencoders and GANs.

Autoencoders (AE)

Autoencoders are the most basic deep learning based architecture that can be used in both noise and dimensionality reduction. From the experiments on different image resolution, it is observed that AE is able to generate the images with very high degree of similarity. And to achieve that, AE doesn't need large number of epochs for training. Figure 5.3 and figure 5.4 demonstrate the AE training progress throughout the epochs.

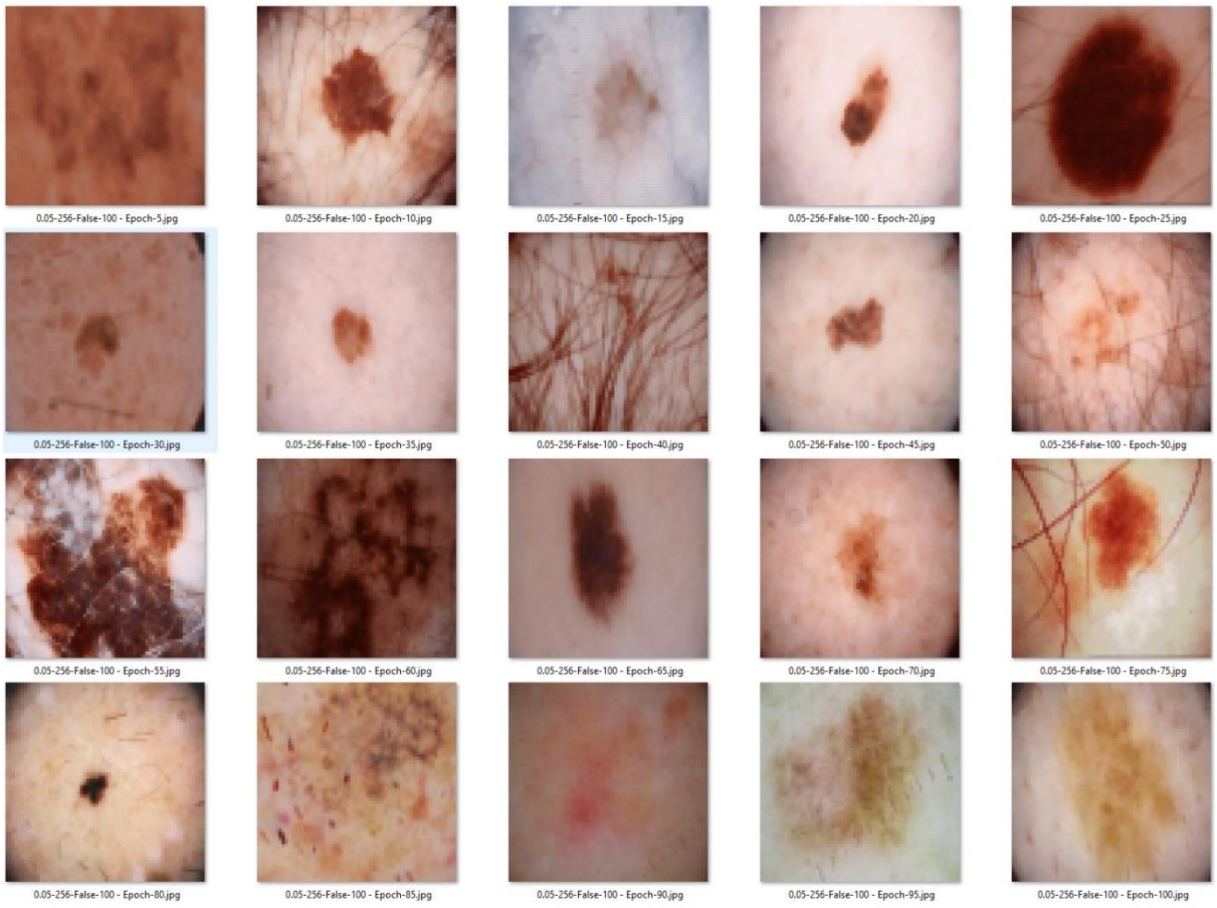


Figure 5.3: Nonlinear AE training progress for 256X256 image

It is clearly seen that from early epochs, the generated images are holding good clarity, clear features, and no noise. This suggests that even when training AE is an overhead in integrated system of AE and GAN, but it doesn't add much additional time. Further talking about Linear AE, it in fact takes even less epochs to be fully trained.

Figure 5.4: Linear AE training progress for 256X256 image

However, Figure 5.5 shows comparison of original image with the generated image from AE in different image resolution.

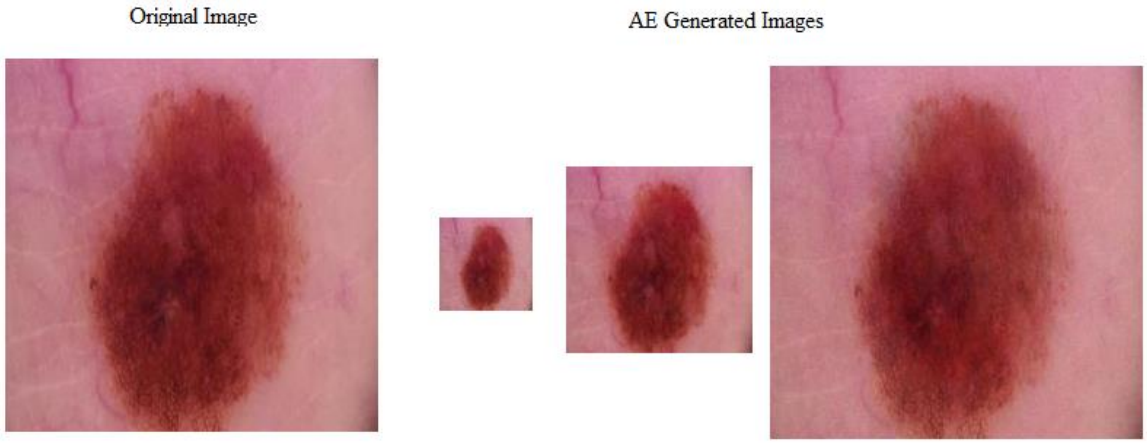


Figure 5.5: Comparison of AE generated images with original image

As it can be seen in the figure 5.5, the generated images from AE are visually very much similar to the original image. And this is true for all three image resolutions. Though it is benefit of AE that it can store the original image as it is in much lower dimensions, but in the context of image augmentation, using AE to synthesize the images results same as using the original image itself multiple times. And this certainly leads to overfitting.

In addition to visual analysis of the generated image, figure 5.6 shows pixel data distribution comparison between original image and AE generated 256X256 image. The data distribution in both of the images are same excepting some minor differences at the beginning of the

distribution. This observation supports the conclusion which was made out from visual analysis of figure 5.5.

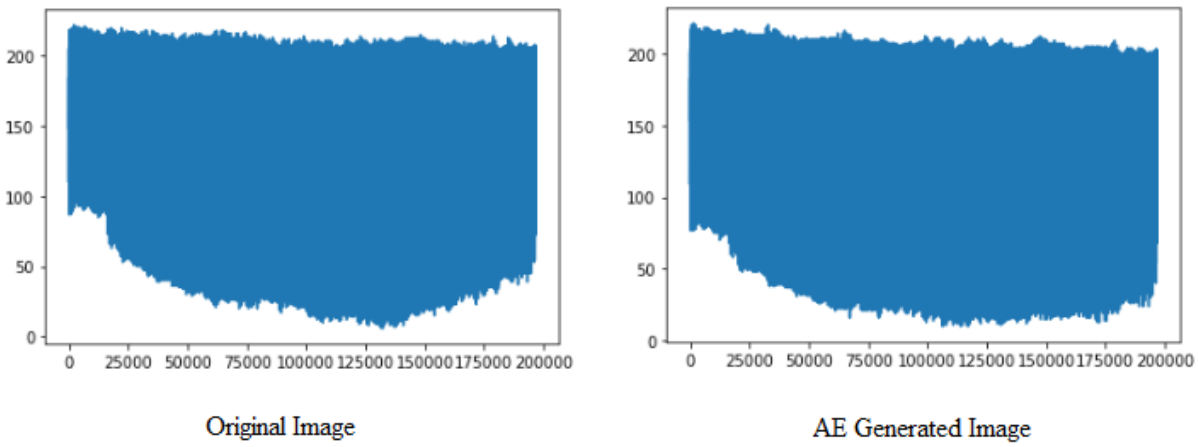


Figure 5.6: Pixel data distribution in original and AE Generated image

Further in coming sections, Images generated using simple GAN are analyzed

GAN

Unlike Autoencoders, GANs are more likely to learn the data distribution from the entire training set and be able to introduce the diversity in image generation based on its learning. The more variation is present in the training set is, the more diverse the GAN generated images should be. By these characteristics, GANs are able to generate the images that are not even existing in the training set itself.

Figure 5.7 showcases the generated image from the GAN and compare it with most similar image present in the training dataset. The GAN generated images shown in the figure 5.7, are not identical with the training images. It is clearly visible that the skin tone is completely different in the generated images in the comparison with the original images.

Also, an interesting thing to note is, in the 256X256 generated images, textures like a body hair is also slightly visible which was not present at all in the training image from which maximum features and shape has been taken. This simply suggests that these features are been taken from other images of the training set and this ability makes GAN able to produce high diverse yet realistic images.

Another important thing to note is, the same image generated in 256X256 resolution is different in resolution 128X128. This is not only true in different resolution, but also in the same resolution, such verity is possible. Meaning, GAN is able to produce multiple different looking images with the reference of the single training image. This ability helps significantly when the training data availibility is extremely low.

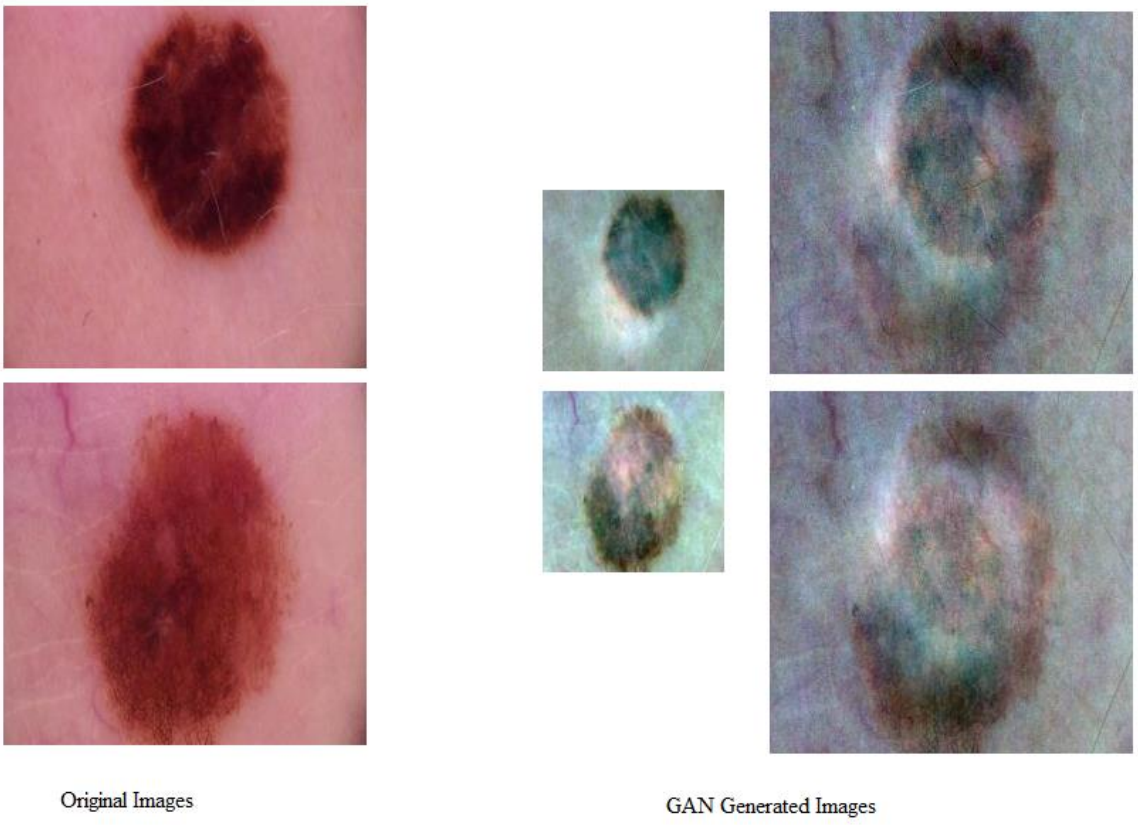


Figure 5.7: Comparison of GAN generated images with most similar images available in the training set

Continuing on the point of being able to produce the diverse dataset, figure 5.8 demonstrate the different images generated by the same GAN model. This GAN was training on the limited set of images to make it concentrate on limited features, shape, and colors under certain experiment. The images shown in the figure 5.8 are generated using the GAN model with the linear hidden layers. On other hand, figure 5.9 showcases the output of the GAN that consists Nonlinear hidden layers. It is observed that similar to AE, the nonlinear GAN is producing the images that are more similar to the training images and thus, showing less diversity in the output.

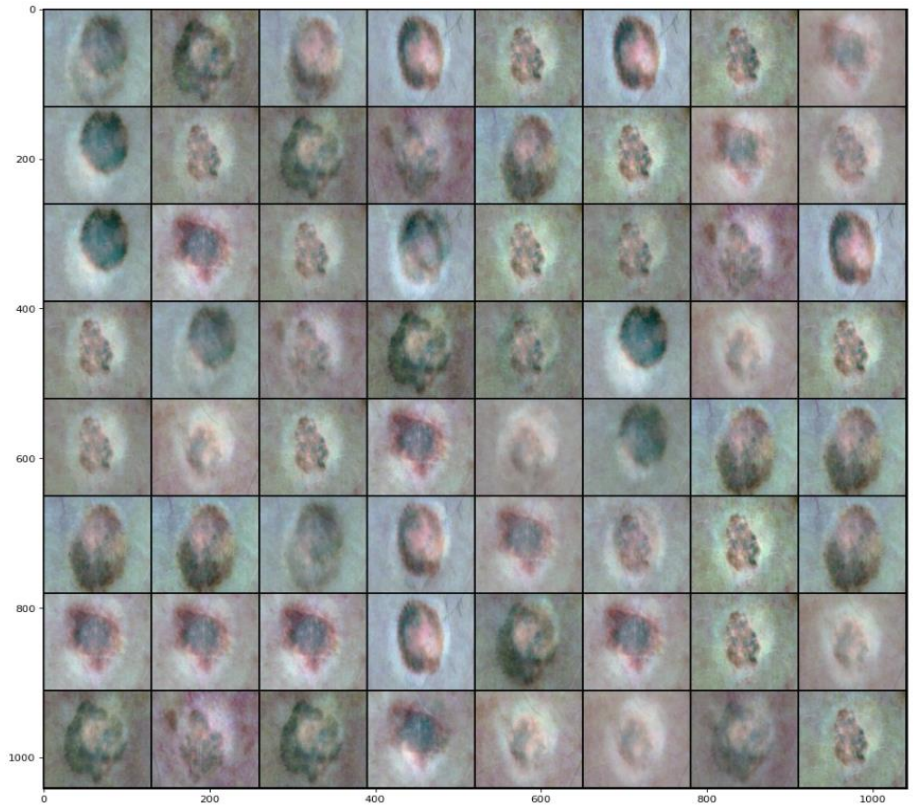


Figure 5.8: Images generated by Linear GAN

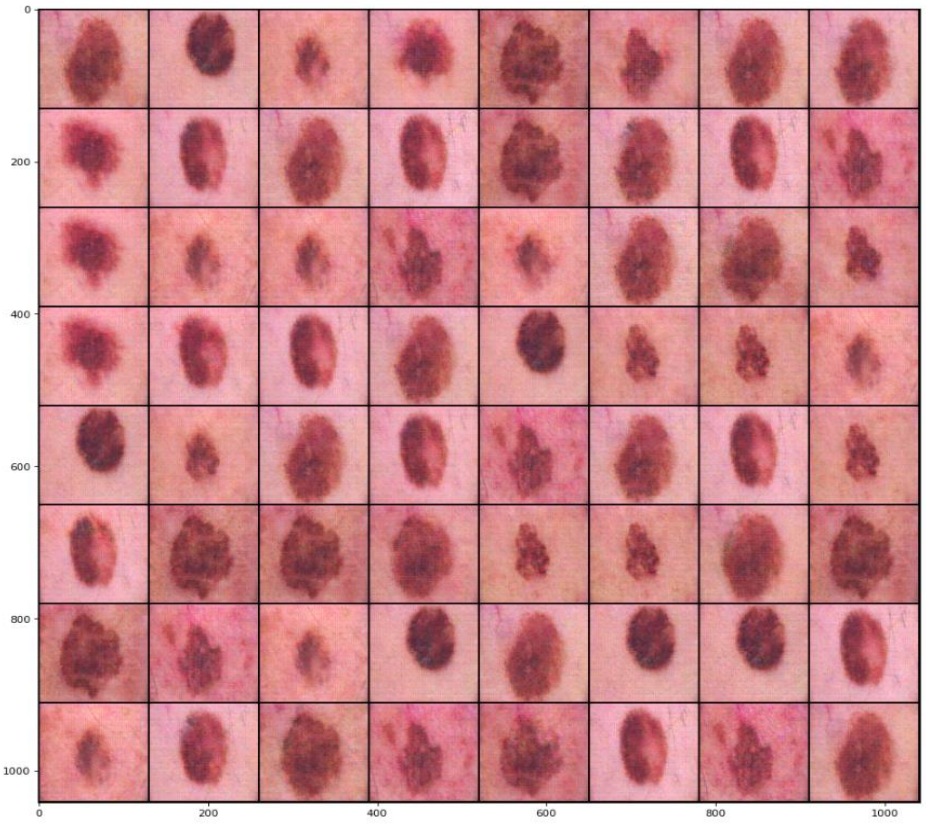


Figure 5.9: Images generated by Nonlinear GAN

Unlike AE and Image transformation, as shown in figure 5.10, the GAN model learns the basic pattern of the data distribution as well as adds the features learnt from other images of the training dataset to generate the image which is not similar to existing images yet stays valid and realistic in the context of the training images. The reason behind the new image stays valid is, no matter how diverse the images is being generated, but all the features in the image are learnt from the images in the training dataset only. Thus, the diversity comes from the dataset only.

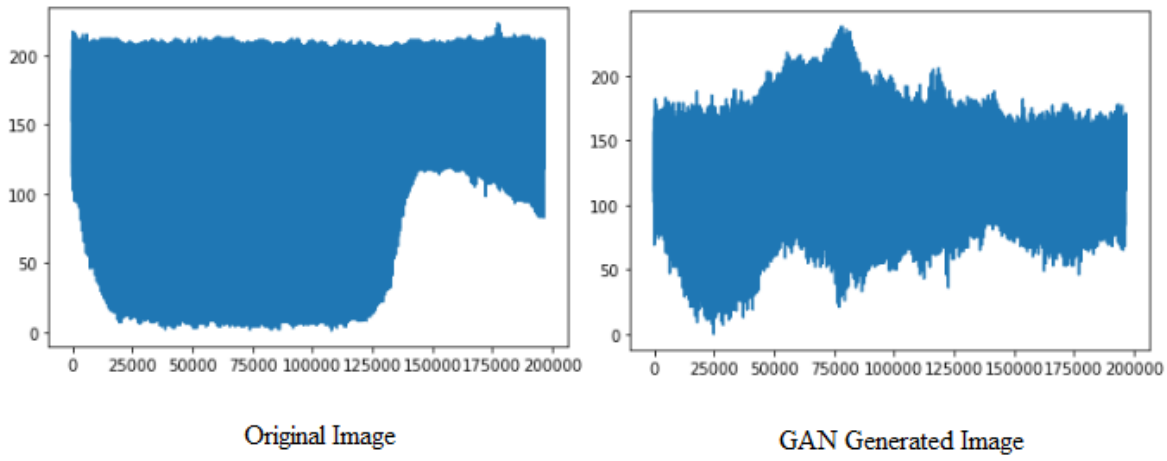


Figure 5.10: Pixel data distribution in original and GAN Generated image

Figure 5.10 shows the pixel data distribution of original and GAN generated image, this is for the same images which is shown in the figure 5.7. By referring figure 5.7, it becomes clearer that how this difference in the data distribution tuns out to be the difference in the image. It is clear in the figure 5.7, that even though the image is different than the most similar image available in the training dataset, the generated image stays valid.

However, the main drawback in having the GAN is the computational requirement to train and generate the images. This drawback is discussed in detail in coming sections. In attempt to overcome this drawback, the integrated system of AE and GAN is proposed in this research work.

Integrated system of AE and GAN

By now, it is understood that AE is helpful in reducing the dimensionality of the image to be generated, but the generated images are same as the training image. So, AE itself, cannot

provide the needed diversity. Whereas, the GANs can produce the rich and diverse images, but in case of GAN, it needs to be training on higher dimensional data, which is both, time consuming as well as space and computationally expensive.

Thus, an integrated system of AE and GAN is proposed where AE is utilized to reduce the dimensionality of training data for GAN training and GAN is to generate the images for augmentation task. In previous chapter three possible approaches to integrate the AE with GAN are discussed and also their results are observed. In this chapter, the finalized approach is further discussed in detail.

Figure 5.11 demonstrates the sample images being generated through all the so far discussed means.

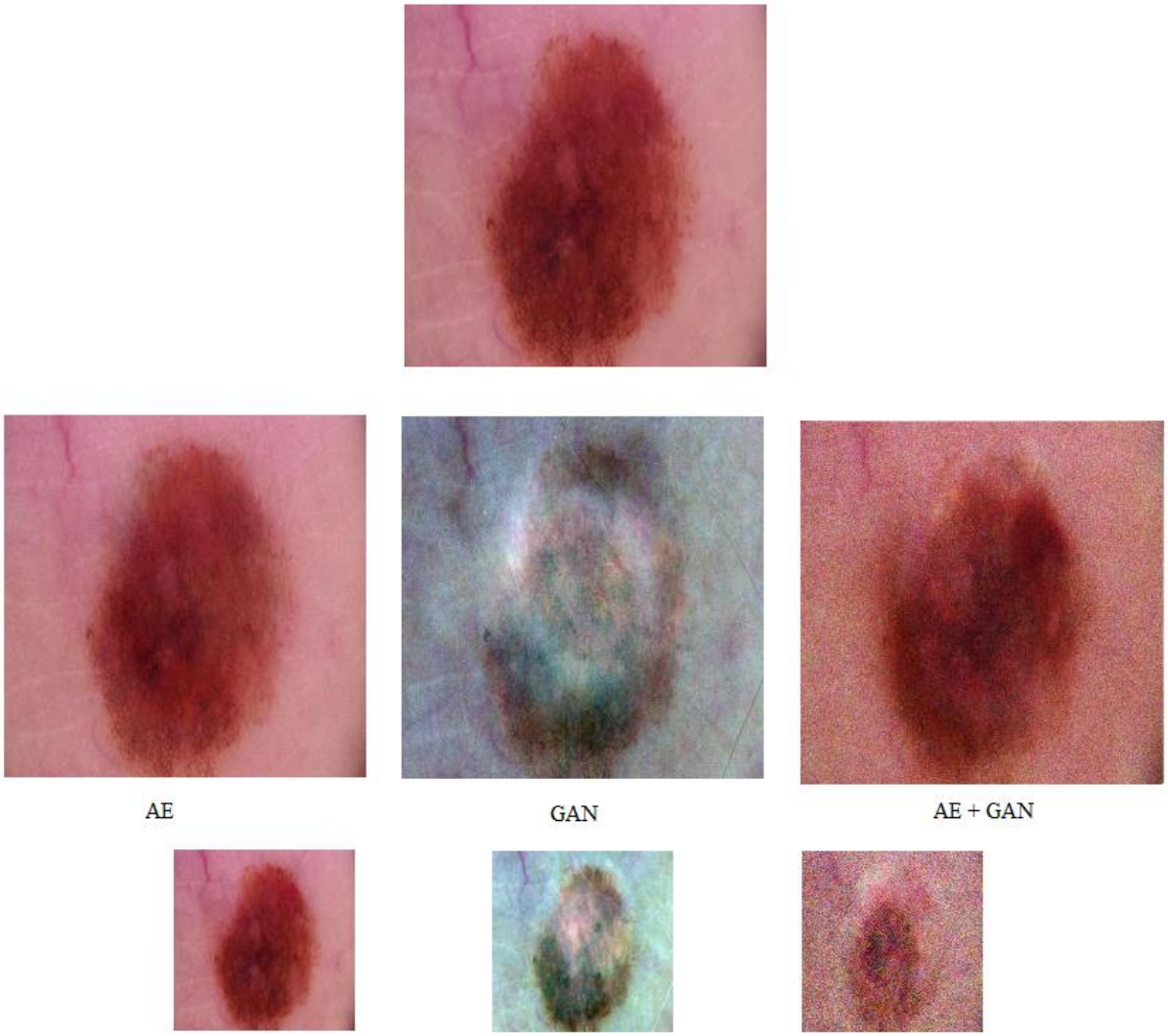


Figure 5.11: Comparison of generated images of AE, GAN, and AE + GAN

All the models used to generate the output images shown in figure 5.11 are linear as it has been observed that in comparison with nonlinear models, linear models can produce images with less noise, and it converges earlier. One thing to note in the figure 5.11, integrated system of AE and GAN works better as the image dimension of output of AE goes high.

Even though the GAN model being used in the system working standalone directly to generate the images isn't able to generate the image exact same as input or as good as the AE network, but as in the integrated system, GAN network works on much smaller dimension, it can produce the suitable input for decoder of AE that can generate the images same as AE does when works standalone.

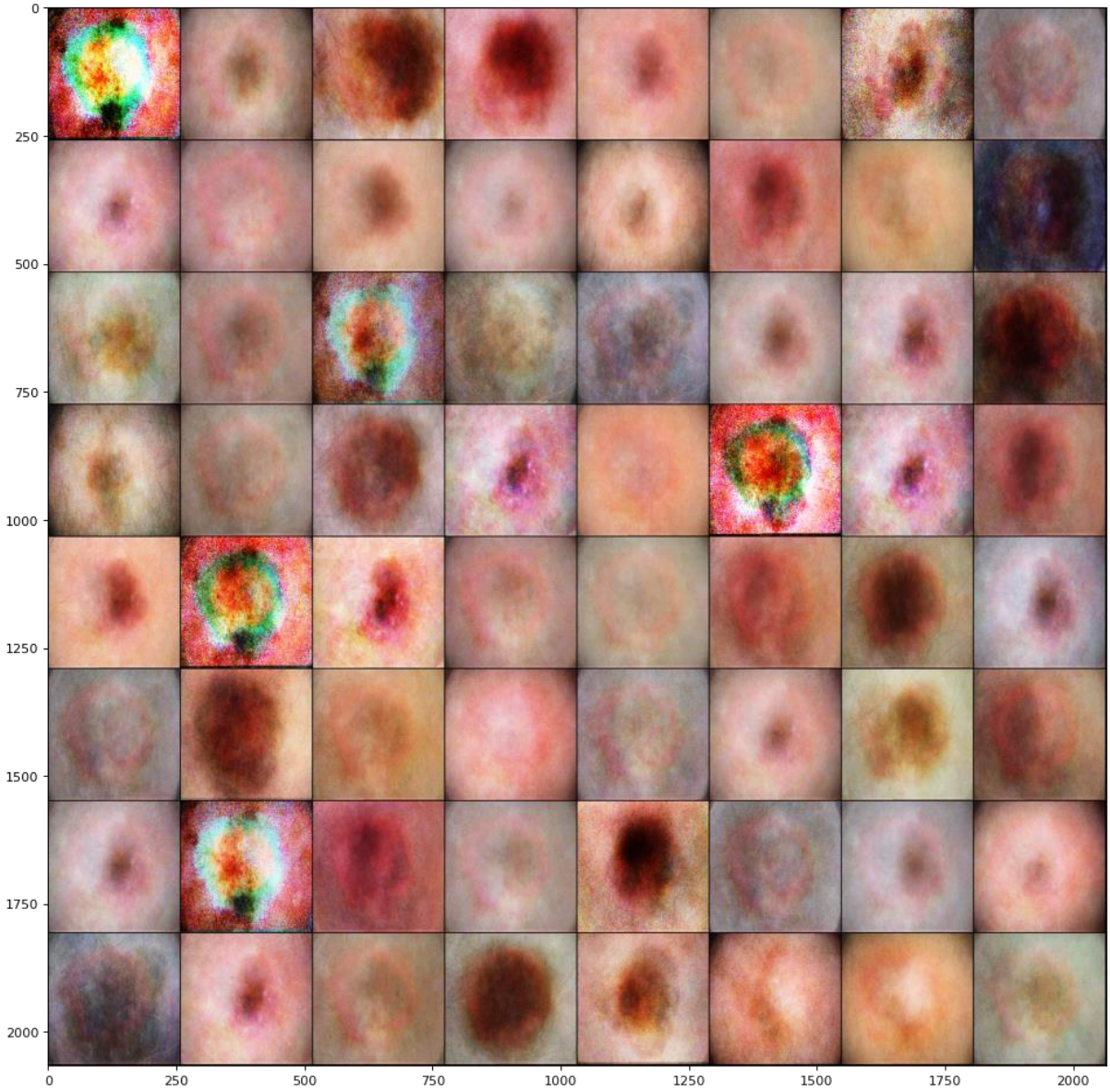


Figure 5.12: Images generated for entire training set using integrated system of AE and GAN

Meaning, with integrated system, fined grained images are produced as well as it still holds the ability to learn the features of different images and combining them in output images a degree of diversity is maintained. Figure 5.12 shows the generated images using integrated system of AE and GAN for entire training set. It is seen in the samples that they aren't exactly same as the images of training set, they hold good enough clarity, and most of them are valid images for melanoma skin lesions.

5.3 Inception Score analysis

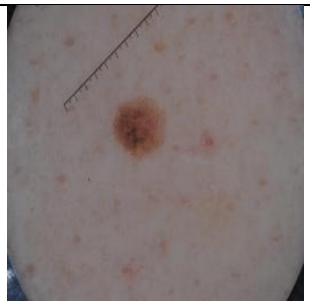
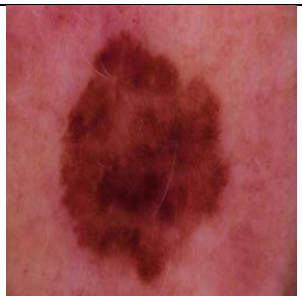
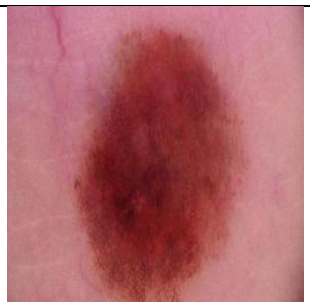
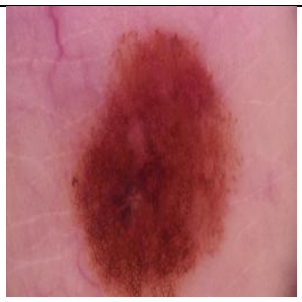
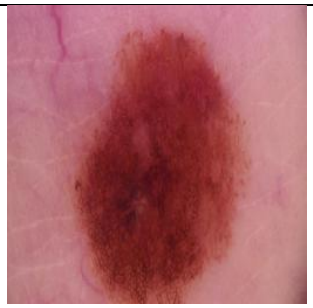
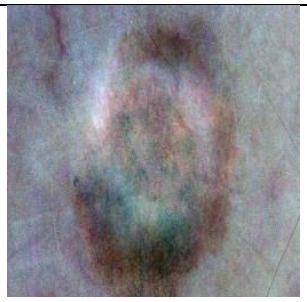
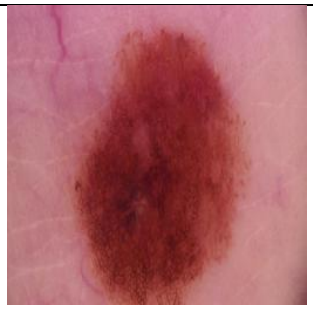
Inception score is the quantitative metric that is used to evaluate how good images the generative models are producing (Salimans et al., 2016; Qin et al., 2020). The inception score calculates the KL-Divergence the logic response of one image and average response of all the generated images. The fundamental researches on GANs like have extensively used this metric to assess the output of the models.

The lowest possible value of the inception score is one where all the provided images are exactly same. Thus, in general lower the value of inception score, better the generated image. But the purpose of the GAN is to produce the image with diversity so that it can cover variety of the features across the dataset. Meaning an ideal inception score for this context is not one but higher. On other hand, highest possible score for inception score can be the number of classes present. However, in the custom algorithm as there is only one class present in the training, the inception score of totally different images can be taken as a base and all generated images should score less than that.

In addition, using inception score, the intra class diversity can be calculated and generated images should show similar inception score as the training images in order to justify that the diversity among generated images are similar to the real set of images.

Table 5.1 contains the resultant inception scores for different comparisons or cases. This table has observation for 256X256 images and contains basically two types of cases, image to image comparison that suggests how similar one image is with other image, and a resultant inception score for the group of images, that suggests how similar the images are within the group. Meaning intra class diversity.

Table 5.1: Inception scores

Case / Comparison	Image 1	Image 2	Inception Score
Two totally different images			2.6033
Training image and AE generated image			1.4567
Training image and GAN generated image			2.1675
Training image and image generated using integrated system of AE and GAN			1.8990
Training images intra class Inception score			2.4623
Generated images intra class Inception score			2.1433

As shown in the table 5.1, two totally different images, which belongs to different classes, have inception score of 2.60. These images are selected by visual analysis on available images on

training dataset. Thus, 2.60 inception score is considered as base line for image to image comparison. Given this information, the other three cases listed in the table is now inter comparable.

Supporting to the analysis made for AE generated images in the previous section, inception score of AE generated image and original score came out to be the 1.47. That suggests that AE generated images are more similar to original images and no significant diversity present in these images.

On other hand, the images generated by standalone GAN, when checked with original or most similar image, the inception score resulted in 2.16. This suggests it has higher degree of variations in comparison with the original images. While the same experiment results in 1.89 with the images generated by integrated system of AE and GAN. This suggests the proposed system is able to generate images that are having balanced diversity.

The inception score for melanoma intra class training images of the subset is 2.4623 which is similar to inception score 2.1433 of generated images by the integrated system of AE and GAN trained on the same subset of the training images. This suggests the generated images are having similar intra class variations.

5.4 Computational Cost

Computational cost covers both the aspects, space requirement and time requirement. Any good model should produce good results. However, it must stay feasible in training as well as during utilizing. The experiments carried out throughout this research are performed on the consistent and limited resources

When calculating the time required for the model to be trained, only considering the time taken per epoch isn't an accurate way as different models needs different number of epochs to get fully trained. Thus, proper formula for time taken is time taken per epoch times number of epochs required. Table 5.2 and figure 5.13 shows the time taken by different generative models to get train.

Table 5.2 Time taken by the model to get train

	Time per epoch (seconds)	No of epochs required	Total time taken (~minutes)
AE	162	10	27
GAN	94	100	156
Integrated system of AE and GAN	20	100	3.33

Figure 5.13 provides visual representation of the data of table 5.2.

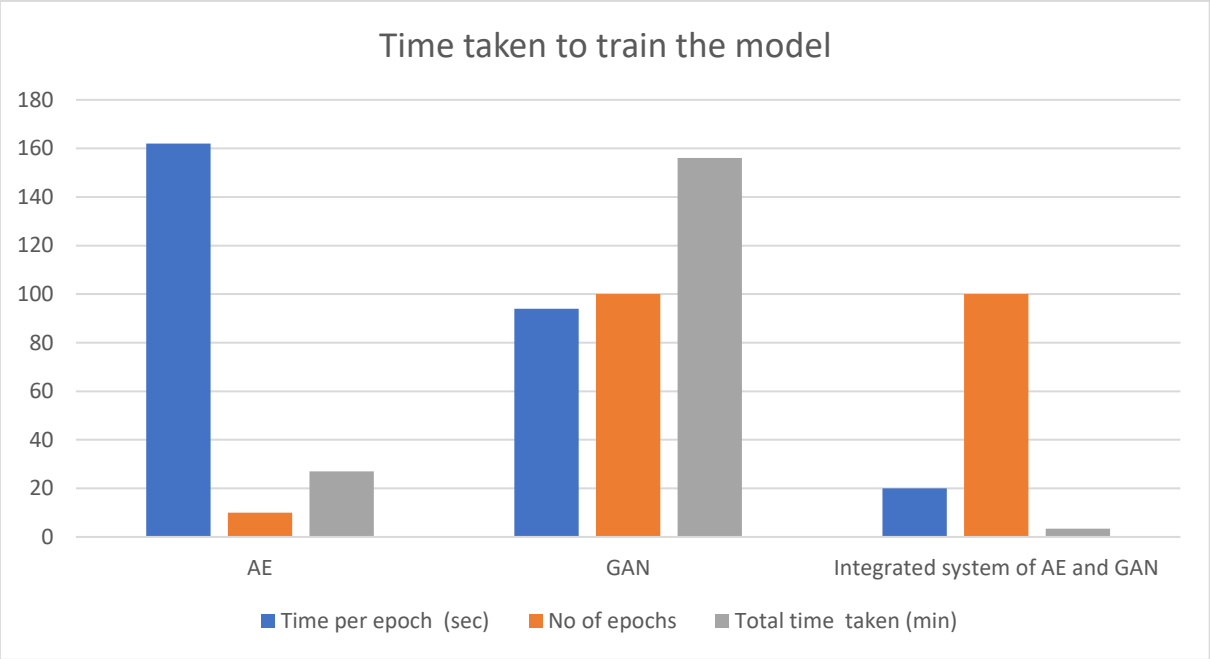


Figure 5.13: Chart of the time taken to train the model

Point to note in the above chart and table is, when it talks about the time required to train the model of integrated system of AE and GAN, it doesn't count the time required for get the AE model ready as pre-requisite. Thus, to calculate the actual time required for get entire system to get trained, it is needed to add both of them together, i.e., 27 minutes + 3.33 minutes, that is 30.33 minutes. While standalone GAN needs 156 minutes to get fully trained.

As mentioned, computational cost also includes then space requirement, though the exact memory requirement might not be the accurate figure in terms of model requirement as it can depend on the quality of the code, better and more accurate way to comparative idea of the space requirement for different models is to check the number of trainable parameters. The higher the number, more complex and heavier the model.

Table 5.3: Number of trainable parameters in different models

Model	No of trainable parameters
AE	85,59,94,368
GAN (Generator + Discriminator)	60,72,85,633
Integrated system of AE and GAN	1,56,83,546

Number of trainable parameters in each finalized models are shown in the table 5.3. It gives an idea how bulky and complex a model is for training and storing in the memory.

Training a complex model is a major challenge as it is prone to make whole training process instable due to gradient explosion or mode collapse etc., while in normal scenarios, storing a model isn't a problem, transfer learning does that all the time. From the table 5.3, it becomes clear that stand alone AE holds largest number of trainable parameters while the GAN in integrated system of AE and GAN has the lowest number of trainable parameters. Whereas in standalone GAN as well parameters are significantly high.

Unlike the time calculation, GAN of an Integrated system of AE and GAN should not include AE's individual parameters, and during the training of GAN, AE is given pre-trained and constant. Also, with the holistic information of table 5.2 and table 5.3, it is clear that cost involved in training standalone GAN is greater than AE.

5.5 Classification

All the experiments discussed till now, are done on the subset of the training images and are meant to be a proof of the concepts. While based on the outcome of the experiments and finalized the methods different augmented datasets for entire training set are prepared and been tested in the constant classification model to obtain the observation.

One thing to keep in mind here is, the results of the individual classification are not considered as any indicator, rather it is comparative study of the different means of image augmentations and are considered as relative outcome with each other.

Even though the classification task is kept out of the scope of this research work, a primary aim of this research is to develop a mean to synthesize the dermoscopic images that can lead into better classification results. Thus, it becomes important to analyze and infer the results of classification task. Table 5.4 contains the measures of evaluation metrics of classification models that are trained on the dataset with augmented images produced by different models or mechanisms.

In the highly imbalance dataset, the metric "accuracy" can be highly misleading, thus evaluation cannot rely only on accuracy. On other hand, the nature of the classification task is such that, missing out the positive cases is more harmful even if model can predict negative cases accurately. So, in this context the metric "sensitivity" makes more sense as it provides the rate of correctly identified positive cases out of all the positive cases present in the data.

Table 5.4: Classification results including 'unknown' class

Case	Accuracy	Sensitivity
With no augmentation	Train: 0.9646 Val: 0.9405	0.06
With down sampling the dominant classes to 1000 images in each category (still class imbalance is present)	Train: 0.8827 Val: 0.7186	0.2456
With traditionally transformed images augmented for positive (melanoma) case in dataset	Train: 0.8275 Val: 0.7446	0.2214
With AE+GAN generated (synthesized) images augmented for positive (melanoma) case in dataset	Train: 0.8647 Val: 0.7685	0.3639
With combination of transformed and synthesized images augmented for positive (melanoma) case in dataset	Train: 0.8767 Val: 0.7630	0.2897

In the context of this research, positive case is the case of "melanoma" and rest all the cases are "negative", this leads the class imbalance even more severe as the model predicted all other classes too is considered as negative prediction. Thus, if there are four classes and probability

of every class being predicted is 25%, in this context, the probability of positive prediction becomes 25% and negative prediction becomes 75%.

From table 5.4, it is clear that even though holding the high accuracy, model cannot perform well in case of no augmentation has been performed. Meaning, in case of class imbalance is present in the data, the model becomes extremely biased towards the dominating class, and it is predicting positive cases too as a negative. Whereas difference between train accuracy and test accuracy decreases when dataset is augmented to handle the class imbalance, this strongly suggests that the model trained on imbalance dataset is biased and overfitted during the training and could not perform well on test dataset.

Interesting thing to note is, even for different trials of augmenting the data for melanoma class, the sensitivity of the model is having limited improvement. This observation brings the attention towards the "unknown" class of the dataset.

There can be two reasons behind the images marked as "unknown" impacting the model performance. One explanation is 'unknown' class is the most dominating class in the dataset, removing that makes data imbalance to get settled at certain level. And second, 'unknown' category might have the images that are not properly tagged and thus chances are there that legit melanoma images be present in that category which makes it hard for the model to decide if a melanoma image should be classify as 'melanoma' or 'unknown'. Due to the second reason, the result of classification without unknown class is more accurate and better indicator for the augmentation performance too.

Thus, it becomes important to try out the same classifications tests without the presence of "unknown" class to eliminate any additional disturbance this class is introducing.

Table 5.5: Classification results excluding 'unknown' class

Case	Accuracy	Sensitivity
With no augmentation	Train: 0.9437 Val: 0.8895	0.0847
With down sampling the dominant classes to 1000 images in each category (still class imbalance is present)	Train: 0.8710 Val: 0.7583	0.3333

With traditionally transformed images augmented for positive (melanoma) case in dataset	Train: 0.8632 Val: 0.7743	0.4857
With AE+GAN generated (synthesized) images augmented for positive (melanoma) case in dataset	Train: 0.9119 Val: 0.8245	0.5734
With combination of transformed and synthesized images augmented for positive (melanoma) case in dataset	Train: 0.8946 Val: 0.8057	0.5477

Table 5.5 shows the result of same activities excluding the 'unknown' class. Excluding the 'unknown' class without doing any other augmentation activities also show cases betterment in the model performance. From the table 5.5, it is clear that the cases where augmentation has not been performed or performed with images that lacks the diversity, the higher accuracy is achieved but at the same time, sensitivity is not improving. This also indicates that the model is either biased towards dominant class and when dataset was augmented with the transformed images, the model becomes overfitted for the training images and fails to perform better on the test dataset. While in case of the dataset augmented with the images generated using the proposed integrated system, validation accuracy is comparable with train accuracy and also the sensitivity on test dataset is significantly improved which suggests that the model is able to perform better on both train dataset as well as test dataset.

5.6 Early Loop Breaking Mechanism

Before understanding what the "loop breaker" is supposed to do and how it does that, first it is important to understand the problem itself. How many epochs are needed to get the generative models get trained properly depend on the number of images, diversity in the training images, quality of models and its parameters. Thus, at any pre-fixed number of epochs might not be efficient approach.

A conventional way to address this issue is to treat the number of epochs as hyperparameter and train the model into multiple folds. However, this approach requires significant higher number of experiments to be carried out for training single model and the finalized parameters are not applicable if images dataset is different.

On other hand, in proposed mechanism of loop breaker, instead of having fixed number of epochs, a soft break is introduced in generative model's training. It is important to understand that in this mechanism too, hyperparameters are needed to be tuned, but instead of tuning the

parameters on generative model, parameters of this mechanism targets on classification task that is usually less complex and time consuming than the generative models like AE or GAN.

In this mechanism, loop can be broken in two possible scenarios if the generative models have missed the minima and started producing degraded images or if the generative model has reached to saturation and no more improvement is observed in the generated images.

Figure 5.14 demonstrate the first scenario where the loop of GAN training should have been broken. To magnify the impact of the issue and to better explain the use case of the loop breaking mechanism, lower dimensioned images are used. The figure shows resultant generated images using the integrated system of AE and GAN on 128X128 images and it also explain the need of loop breaking mechanism to be in place

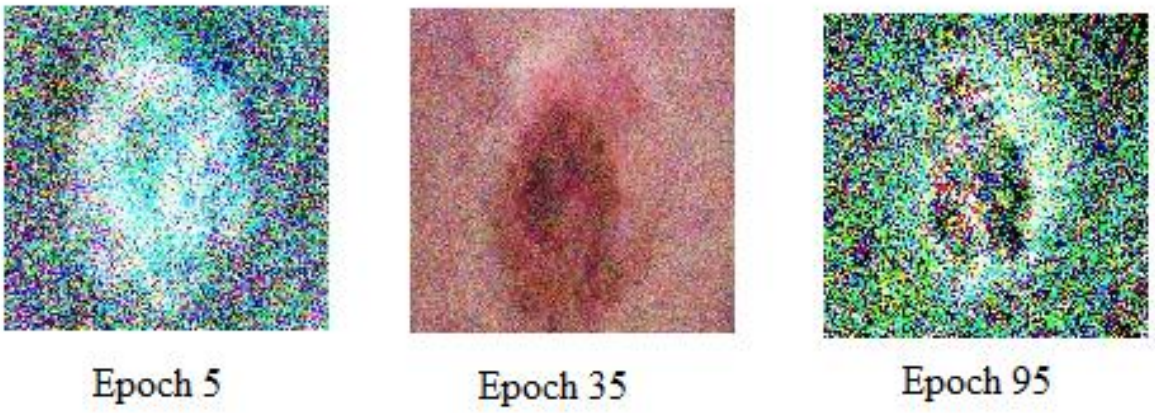


Figure 5.14: Comparison of generated image by an integrated system through different epochs

It is clearly visible that during the 35th epoch, the GAN was able to produce the better results than the 95th epoch.

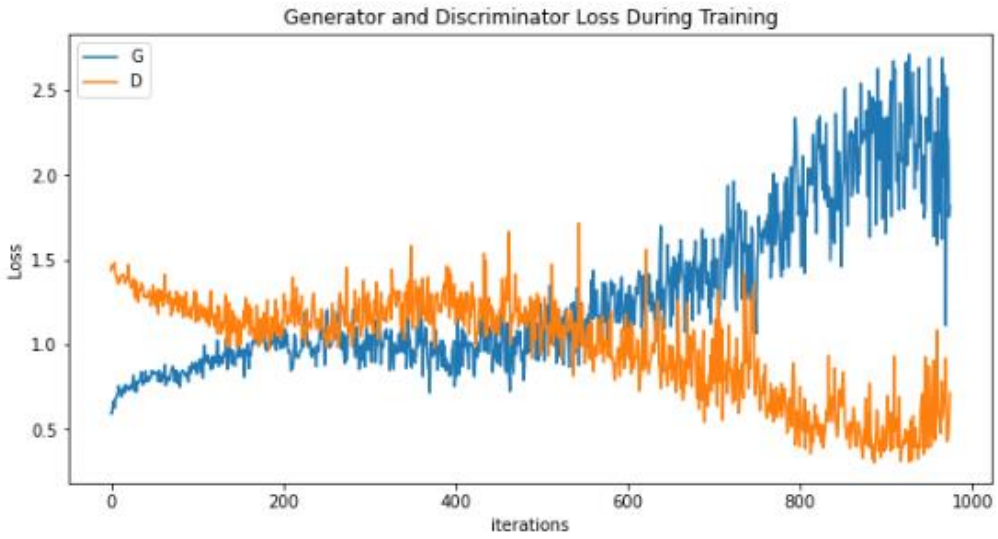


Figure 5.15: Loss graph of generator and discriminator networks

The loss graph of generator and discriminator network shown in figure 5.15 too supports the same observation. The graph showcases that around the index 350 to 400, i.e., 35th to 40th epoch (roughly calculating 10 batches per epoch), the loss of generator and discriminator have been converged. However, from 500th index, the loss of generator has started going up which suggests that quality of image getting generated is decreasing at that phase.

Figure 5.16 showcases the loop breaking mechanism in action with start point is set to 100, loss to bear is set to 0.5% and minimum required gain is set to 0.1%. This mechanism is applied on the moving average of sensitivity score of the model that was trained on 400 different training datasets.

As it is seen in the chart, the loop breaking mechanism has indicated that the GAN training should not be continued post 171st epoch as the sensitivity of the model trained on the images generated by GAN is either degrading or not improving.

Figure 5.16 Demonstration of loop breaking mechanism

5.7 Summary

This chapter has discussed the result of different aspects of finalized approaches in detail. It is found that proposed mechanism of integrated system of autoencoder and GAN has outperformed the other approaches in terms of quality of augmented dataset, computational cost, model complexity, and classification task.

Beginning from the qualitative analysis, it is seen that traditionally transformed images cannot add the diversity in the result. Autoencoder generated images too are found similar to the training images whereas GAN based mechanism could introduce the required diversity in the images.

As in quantitative analysis, inception score is discussed for the various comparisons and cases to determine the quality of the generated images. Also, classification metrics like accuracy and sensitivity are discussed and inferred. At the end, applications of loop breaker mechanism are demonstrated.

6. Conclusions And Recommendations

This research has aimed to address the problem of extreme data imbalance present in the dataset of dermoscopic images or to be able to come up with well experimented findings on the various research questions.

6.1 Introduction

So far in this research has discussion on proposed method and the subcomponents that are going to support it. Later various experiments are carried on the defined methods and modules, analyzed the outcome of the experiments, and finalized the approaches to be implemented to prove the defined concepts.

The finalized approaches are then tested to get the answers to the defined research questions and get the data which helps in concluding the research. Further, this chapter provides a brief inference on the obtained answers and conclude the research. Also, this chapter discusses the findings gathered that contributes to the knowledge base in the domain of image synthesis using generative models and provides an open door to extend on the top of this research work.

6.2 Conclusion

This research can be concluded from two fronts, Classification improvement and Computational cost reduction.

Classification

From the observed outcomes of the classification tasks ran on different set of datasets, it becomes clear that the presence of the class imbalance issue in the ISIC 2020 dataset leads the skin lesion classification to be biased towards dominant classes and thus, both accuracy of the model and sensitivity of the model suffers due to it.

First thing to understand in this context is the “Accuracy” can be highly misleading metric if the causative factors are not properly analyzed. For the model that is trained on the untreaded dataset has accuracy score 0.94 and at the same time the sensitivity score is 0.08, that suggests the biasness of the model as the model tends to classify all the cases as the dominant (negative) class. While down sampling the data for dominant classes gets the gap between accuracy and sensitivity lower where the accuracy is 0.87 and sensitivity is 0.33. However, the dataset is still imbalanced, as the positive cases are still significantly less than the negative cases.

With the properly done image augmentation using the synthetic images generated by the proposed integrated system of autoencoder and GAN, the model became significantly less overfitted and less biased. The difference between training accuracy and validation accuracy is around 9%. Also, the sensitivity of the model on test dataset has increased from 0.33 to more than 0.57 which is 72.7% improvement.

In addition, As the sensitivity of the model trained on traditionally transformed images is 0.48, it can also be concluded that the classification using the images generated by the proposed method performs better than the of the images augmented using traditional image transformation techniques. Also, data augmentation works better than the data anonymization.

This concludes that the images synthesized using GAN holds good amount of diversity and stay realistic which in terms is proven to be good enough mean of image augmentation where the original dataset is highly imbalance and biased towards the dominant class.

Computational Cost

While the GAN was proposed to improve the dataset to improve the classification task, the proposed method of integrated system of AE and GAN is aiming to reduce the computational cost required by the GAN. Form the recorded time taken and the number of neurons present in the method, it is observed that the significant computational cost can be saved by using the proposed method of integrating the AE with GAN.

The number of trainable parameters in the simple GAN model for 256X256 image generation is 60,72,85,633 while to be able to generate the GFV for the same sized image the number of trainable parameters required in proposed method is 1,56,83,546. And as a result, the overall time required for training process can be reduced from 156 minutes to around 33 minutes, that is 78.8% less time consumption.

Given these figures, it can be concluded that the proposed method of integrating the AE with GAN is more efficient and faster in comparison with standalone GAN networks.

In addition to the computational time saving using relatively less bulky GAN, the loop breaker mechanism also saves the time by enabling the soft breaking. However, the amount of time saved by this mechanism depends on the set hyperparameters.

Apart from the improvement in classification and reduction in computational cost, from qualitative analysis it is clear that proposed method of integrated system of AE and GAN is able to generate divers and realistic images with inception score around 1.8 which is well balanced and lay approximately at the equal distance from the score of exact same images and totally different images.

6.3 Contribution to knowledge

The major learning from this research work is understanding the difference between conventional GAN and the proposed method. The conventional GAN is trained on random noise to generate the image. And thus, the model becomes unstable and complex as the dimension of the image increases. On other hand, when a pre-trained autoencoder is integrated with the GAN, the GAN is required to learn to produce the GFV that can be decoded back to the image. It has been observed that by training the GAN to learn a GFV rather than the image, good amount of training time can be saved as GFV tends to be significantly lower dimensioned than the image itself. Also due to reduction in dimension, GAN too become less complex and overfitted.

In addition to that, it is also observed that GAN generated images can degrade or hit the saturated quality in case of keep the training process going on even after good convergence of generator and discriminator's performance. Post that the quality of images either degrades or doesn't improve. Keeping the track on the improvement can help in terminating the training even before the defined number of iterations and that can save the computational cost on GAN training keeping the performance unimpacted.

6.4 Future Recommendations

ISIC dermoscopic images contains the skin lesion images of different skin tones as well as either square or round in shape. This is one of the major challenges in training the generative model as the different skin tones are also considered as difference in the images which technically isn't accurate. This issue can be properly address by using combination of "Image segmentation" and "Image to Image" translation. Where the skin lesions are segmented and then translated on to uniform background. This approach needs to be researched well and tested on the POCs.

As per the context of this research, only “melanoma” class images are synthesized. However, there are other classes as well for which the number of images are less. In order to be able to augment the whole dataset, training separate GANs is inefficient and time consuming. In such cases this study can be extended by modifying the GAN that can be conditioned to learn data distribution for multiple classes and generate the images of different classes based on the condition.

Another important area in which this study can be extended is to utilize the reinforming learning in the integrated system of GAN and AE. This research has tried out the feasibility of two possible approaches to achieve this objective and found that given enough resources, the reinforcement learning can be used with pre-trained GAN in order to obtain the best suitable input for the generator network. Due to limitation of the resources, this concept is not practically proven in this research work and thus it is an opportunity to extend the research in the given direction.

References

Abdelhalim, I.S.A., Mohamed, M.F. and Mahdy, Y.B., (2021) Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Systems with Applications*, 165.

Ahmad, B., Jun, S., Palade, V., You, Q., Mao, L. and Zhongjie, M., (2021) Improving skin cancer classification using heavy-tailed student t-distribution in generative adversarial networks (Ted-gan). *Diagnostics*, 1111.

Anon (2020) *International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset. Creative Commons Attribution-Non Commercial 4.0 International License*.

Bissoto, A., Valle, E. and Avila, S., (2021) GAN-Based Data Augmentation and Anonymization for Skin-Lesion Analysis: A Critical Review. [online] Available at: http://arxiv.org/abs/2104.10603.

Borji, A., (2019) Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179, pp.41–65.

Brock, A., Donahue, J. and Simonyan, K., (2018) Large Scale GAN Training for High Fidelity Natural Image Synthesis. [online] Available at: http://arxiv.org/abs/1809.11096.

Dumagpi, J.K. and Jeong, Y.J., (2021) Evaluating gan-based image augmentation for threat detection in large-scale xray security images. *Applied Sciences (Switzerland)*, 111, pp.1–21.

Dumagpi, J.K., Jung, W.Y. and Jeong, Y.J., (2020) A new GAN-based anomaly detection (GBAD) approach for multi-threat object classification on large-scale x-ray security images. *IEICE Transactions on Information and Systems*, E103D2, pp.454–458.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H., (2018) GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, pp.321–331.

Fu, Y., Li, X. and Ye, Y., (2020) A multi-task learning model with adversarial data augmentation for classification of fine-grained images. *Neurocomputing*, 377, pp.122–129.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., (n.d.) *Generative Adversarial Nets*. [online] Available at: http://www.github.com/goodfeli/adversarial.

Guan, Q., Chen, Y., Wei, Z., Heidari, A.A., Hu, H., Yang, X.H., Zheng, J., Zhou, Q., Chen, H. and Chen, F., (2022) Medical image augmentation for lesion detection using a texture-constrained multichannel progressive GAN. *Computers in Biology and Medicine*, 145.

Hammami, M., Friboulet, D. and Kechichian, R., (2020) *CYCLE GAN-BASED DATA AUGMENTATION FOR MULTI-ORGAN DETECTION IN CT IMAGES VIA YOLO*. *2020 IEEE International Conference on Image Processing (ICIP)*.

Hinton, G.E. and Zemel, R.S., (n.d.) *Autoencoders, Minimum Description Length and Helmholtz Free Energy*.

Kingma, D.P. and Welling, M., (2013) Auto-Encoding Variational Bayes. [online] Available at: http://arxiv.org/abs/1312.6114.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B. and Sánchez, C.I., (2017) *A survey on deep learning in medical image analysis*. *Medical Image Analysis*, .

M E Vestergaard, S W Menzies, P Macaskill and P E Holt, (2008) Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting.

Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H. and Menze, B., (2020) Red-GAN: Attacking class imbalance via conditioned generation. Yet another perspective on medical image synthesis for skin lesion dermoscopy and brain tumor MRI. [online] Available at: http://arxiv.org/abs/2004.10734.

Qin, Z., Liu, Z., Zhu, P. and Xue, Y., (2020) A GAN-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 195.

Rahmayanti, S.R., Fatichah, C. and Suciati, N., (2021) Sketch Generation from Real Object Images Using Generative Adversarial Network and Deep Reinforcement Learning. In: *Proceedings of 2021 13th International Conference on Information and Communication Technology and System, ICTS 2021*. Institute of Electrical and Electronics Engineers Inc., pp.134–139.

Rashid, H., Tanveer, M.A. and Aqeel Khan, H., (2019) Skin Lesion Classification Using GAN based Data Augmentation. *Conference proceedings : ... Annual International Conference of the*

IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2019, pp.916–919.

Reddy Alasandagutti, A., (2021) *Using Deep Learning to Automate the Diagnosis of Skin Using Deep Learning to Automate the Diagnosis of Skin Melanoma Melanoma*. [online] Available at: https://egrove.olemiss.edu/hon_thesis/1928.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X., (2016) Improved Techniques for Training GANs. [online] Available at: http://arxiv.org/abs/1606.03498.

Sarmad, M., Korea, S., Lee, H.J. and Kim, Y.M., (n.d.) *RL-GAN-Net: A Reinforcement Learning Agent Controlled GAN Network for Real-Time Point Cloud Shape Completion*.

Shahsavari, A., Ranjbari, S. and Khatibi, T., (2021) Proposing a novel Cascade Ensemble Super Resolution Generative Adversarial Network (CESR-GAN) method for the reconstruction of super-resolution skin lesion images. *Informatics in Medicine Unlocked*, 24.

Shubham, K., Venkatesh, G., Sachdev, R., Akshi, Jayagopi, D.B. and Srinivasaraghavan, G., (2021) Learning a Deep Reinforcement Learning Policy over the Latent Space of a Pre-trained GAN for Semantic Age Manipulation. In: *Proceedings of the International Joint Conference on Neural Networks*. Institute of Electrical and Electronics Engineers Inc.

Singh, N.K. and Raza, K., (2020) *Medical Image Generation using Generative Adversarial Networks*.

Srivastav, D., Bajpai, A. and Srivastava, P., (2021) Improved classification for pneumonia detection using transfer learning with GAN based synthetic image augmentation. In: *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*. Institute of Electrical and Electronics Engineers Inc., pp.433–437.

Ukwuoma, C.C., Belal Bin Heyat, M., Masadeh, M., Akhtar, F., Zhiguang, Q., Bondzie-Selby, E., Alshorman, O. and Alkahtani, F., (2021) Image Inpainting and Classification Agent Training Based on Reinforcement Learning and Generative Models with Attention Mechanism. In: *Proceedings of the International Conference on Microelectronics, ICM*. Institute of Electrical and Electronics Engineers Inc., pp.96–101.

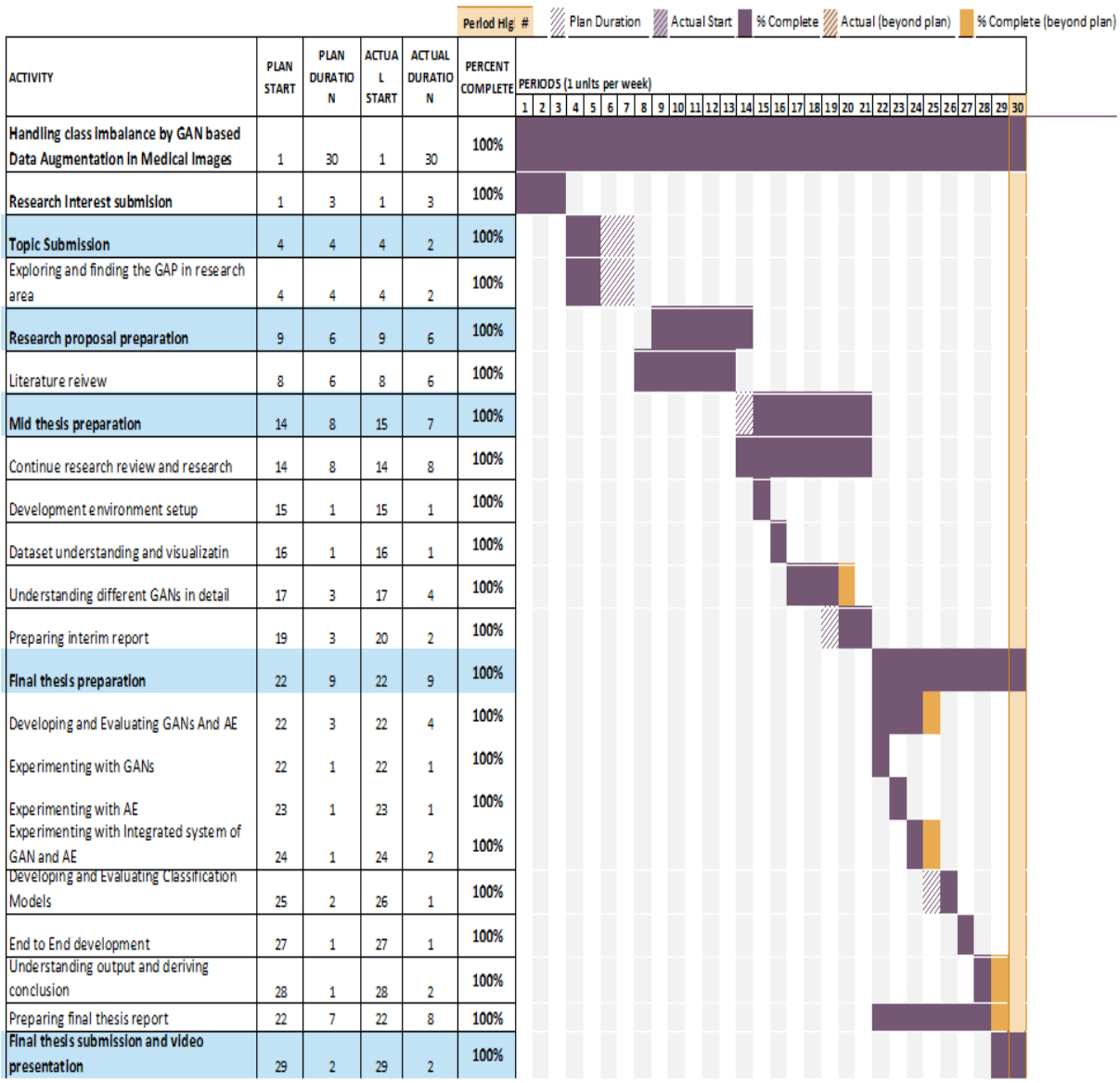
Verma, R., Mehrotra, R., Rane, C., Tiwari, R. and Agariya, A.K., (2020) Synthetic image augmentation with generative adversarial network for enhanced performance in protein classification. *Biomedical Engineering Letters*, 103, pp.443–452.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F. and Pinheiro, P.R., (2020) CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access*, 8, pp.91916–91923.

Yann Lecun, (1987) *PhD thesis: Modeles connexionnistes de l'apprentissage (connectionist learning models)*.

Zhai, J., Zhang, S., Chen, J. and He, Q., (2019) Autoencoder and Its Various Variants. In: *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*. Institute of Electrical and Electronics Engineers Inc., pp.415–419.

Appendix A: Final Research Plan



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Handling class imbalance by GAN based Data Augmentation in Medical Images	1	30	1	30	100%
Research Interest submision	1	3	1	3	100%
Topic Submission	4	4	4	2	100%
Exploring and finding the GAP in research area	4	4	4	2	100%
Research proposal preparation	9	6	9	6	100%
Literature reivew	8	6	8	6	100%
Mid thesis preparation	14	8	15	7	100%
Continue research review and research	14	8	14	8	100%
Development environment setup	15	1	15	1	100%
Dataset understanding and visualizatin	16	1	16	1	100%
Understanding different GANs in detail	17	3	17	4	100%
Preparing interim report	19	3	20	2	100%
Final thesis preparation	22	9	22	9	100%
Developing and Evaluating GANs And AE	22	3	22	4	100%
Experimenting with GANs	22	1	22	1	100%
Experimenting with AE	23	1	23	1	100%
Experimenting with Integrated system of GAN and AE	24	1	24	2	100%
Developing and Evaluating Classification Models	25	2	26	1	100%
End to End development	27	1	27	1	100%
Understanding output and deriving conclusion	28	1	28	2	100%
Preparing final thesis report	22	7	22	8	100%
Final thesis submission and video presentation	29	2	29	2	100%

Appendix B: Research Proposal

Handling class imbalance by GAN based Data Augmentation in Medical Images

Amitkumar M Maheshwari

Research Proposal
Master of Science in Machine Learning and Artificial Intelligence

April 2022

Abstract

Deep learning based models have proven their strength in medical fields, especially working with medical images. In recent times, many open source platforms collaborated with medical institutes and experts had attempted to address the fundamental obstacle of the lack of reliable training datasets by making the data available to the community with proper annotation. However, this attempt doesn't solve the other significant problem which is the lack of particular class(es) in the available training dataset. It is generally observed in medical images that some anomaly/abnormality/condition would occur very rarely in comparison with other cases. Such class imbalance impacts the performance of the models by leading the output to be biased towards the dominating class(es). The class imbalance issue isn't hidden from the research community and there has been fair enough research has been done to address the lack of training image by synthetically augmenting. Although in many cases of radiographic image datasets, successful image augmentation has been presented still in the case of camera-based or natural medical images that contain a high degree of variance in visual appearance and colors, the performance of synthetical augmentation is still not satisfactory. This research is aimed to further improve image augmentation for camera-based medical images by using GAN-based image synthesis. This research will utilize skin lesion dermoscopic images to train and validate image augmentation carried out using GAN variants like DC-GAN and Style-GAN. The augmented dataset will be independently evaluated as well as the classification models trained on the dataset.

Table of content

List of Figures

List of Tables

List of Abbreviations

GAN	Generative Adversarial Nets
AC GAN	Auxiliary GAN
DC GAN	Deep Convolutional GAN
PG GAN	Progressive GAN
TMP GAN	Texture-constrained Multichannel Progressive GAN
CNN	Convolutional Neural Network
VGG NET	Visual Geometry Group Net
YOLO	You Only Look Once
ISIC	International Skin Imaging Collaboration
BraTS	Brain Tumor Segmentation
CBIS	Curated Breast Imaging Subset
DDSM	Digital Database for Screening Mammography
CT	Computed Tomography
MRI	Magnetic resonance imaging
VAE	variational autoencoders

1. Background

Machine learning, especially deep learning based models and AI is continuously making their prominent place in modern-day medical science. From routine checks, to assisting in complex surgical operations AI solutions have been established as digital assistance to doctors and other medical staff. However, for better-performing models, a better training dataset is needed. An ideal training dataset should have sufficient and diverse enough training data. But in the medical domain, there are often cases of unavailability of training data, or even if the data is available, the number of positive cases of rare anomalies is very less in comparison with the number of negative cases which results in either overfitted or extremely biased detection/classification model. Often misclassification of any medical condition can be as bad as fatal, so it is important to develop an unbiased and reliable classification model. Additionally, medical experts are required to get the training data reviewed to label them. This process is manual, time-consuming, and cost inefficient. On top of that, it is highly dependent on the expertise of the medical professional and prone to human error.

In this research, ISIC 2020 skin lesion images (International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset., 2020) are used to demonstrate the issue of class imbalance. (Table 1.1: Number of images per class in the ISIC 2020 dataset and Figure 1.1: Class distribution in ISIC 2020 dataset show) the distribution of different cases of skin lesions.

Diagnosis	Count of diagnosis
atypical melanocytic proliferation	1
cafe-au-lait macule	1
lentigo NOS	44
lichenoid keratosis	37
melanoma	584
nevus	5193
seborrheic keratosis	135
solar lentigo	7
unknown	27124
Total images	33126

Table 1.1: Number of images per class in the ISIC 2020 dataset

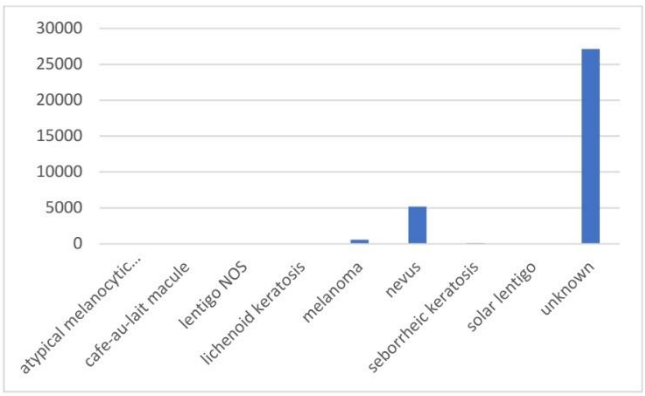


Figure 1.1: Class distribution in ISIC 2020 dataset

Class 'unknown' and class 'nevus' are highly dominating the entire distribution and it is obvious if this dataset is used to train the skin lesion classification model as is, the resultant model will be biased towards these two classes. The condition becomes too dangerous given the fact that 'melanoma' type skin lesion is critical to be detected especially when dermoscopy is the only reliable source of traditional detection as naked eye examination is proven to be less accurate (M E Vestergaard et al., 2008).

Two general approaches are there to handle class imbalance, under sampling and over sampling. Oversampling, the process of increasing the training data using data augmentation techniques (or just duplicating the data) is a more appropriate approach as just like the most cases of medical images, under-sampling of the two dominant classes to balance class distribution can't be the possible approach as it is observed in the Table , availability of the images in other classes are extremely less and an attempt to under-sample the dataset will result in underfitted model.

A combination of two independent deep learning based networks, one responsible for image generation and the other for image classification, interacting with each other can build an innovative image generation model (Goodfellow et al., n.d.). In their research, they proposed two deep learning models being trained parallelly, a Generative model G which learns the data distribution to produce the image as output and a Discriminative model D that takes the generated image as input and estimates the probability of the input image is from real training dataset rather than generated by G. Together both model can work as one unit that is capable of generating realistic synthetic images and it is known as generative adversarial nets (GAN). Figure 1.2: basic architecture of GAN shows the basic architecture of GAN.

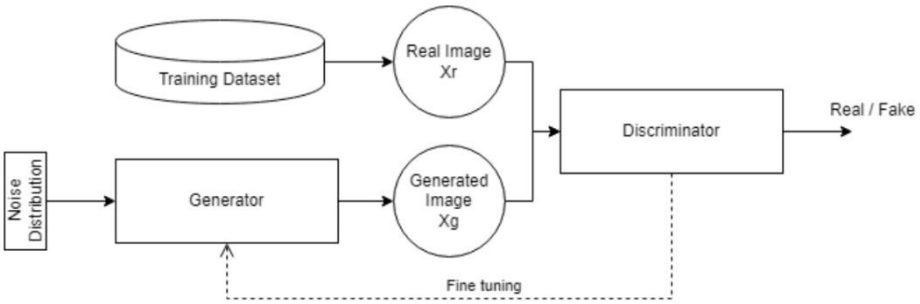


Figure 1.2: Basic architecture of GAN

Applications of GANs have a wide range in the computer vision field, there are many cases such as image augmentation, image registration, medical image generation, image reconstructions, and image-to-image translation where GANs are proven to be useful. Basic/Vanilla GAN has issues when working with high resolution images or more complex features like Mode collapse and gradient vanishing. Also, it performs limited on complex tasks such as image-to-image translations. Many researchers extensively worked on GAN to propose different variants of GAN to overcome the limitations of original GAN architecture like, AC-GAN to introduce the conditional operation, Progressive GAN to be able to progressively enhance the resolution of generated images, pix2pix GANs to be able to perform image to image translations and fusing segment of one image (or entire image) on other images to produce out of the box results.

2. Related Research Work

After Goodfellow and his team introduced the concept of Generative Adversarial Nets (GAN) (Goodfellow et al., n.d.), it had soon become an area of interest for many researchers working in the domain of computer vision, and deep learning, and a lot of work has been done in this field so far. Although it was introduced in 2014 a solid trend of using GAN variants to generate synthetic images to be used in other deep learning networks as input can be seen in recent years.

F-CGAN, a two-staged conditional GAN proposed in (Fu et al., 2020) works on image-to-image translation style instead of noised based image generation. F-CGAN showcased a significant improvement in generating fine-grained images when compared with previously acclaimed AC-GAN, and SNGAN and the classification models trained on the dataset generated by F-CGAN showed better accuracy than the standard model and SNGAN model. On the other side,

GANs (Dumagpi et al., 2020; Dumagpi and Jeong, 2021), have been put to generate synthetic images of positive threat X-ray images to balance an extremely unbalanced dataset. In (Dumagpi and Jeong, 2021) researchers have used DC GAN for image generation and Cycle-GAN for image translation in addition to traditional image transformation (shown in Figure 2.1: Traditional(left) and Generative(right) techniques of Images Augmentation). While evaluating they noticed that combining all three types of synthesized images can make the classification model generalized enough to bring significant improvement in average precision.

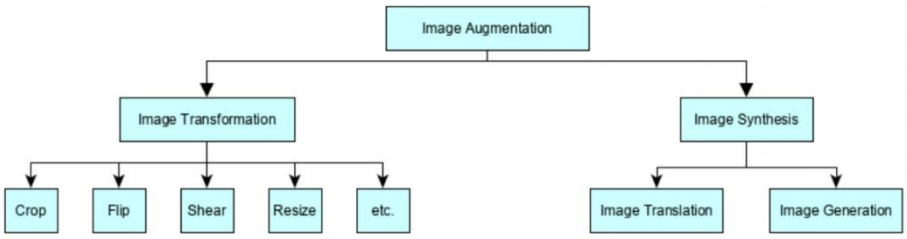


Figure 2.1: Traditional(left) and Generative(right) techniques of Images Augmentation used in (Dumagpi and Jeong, 2021)

This shows that GANs have applications from normal object classification to as critical as subway and airport security by improving the performance of the classification model. Talking about medical images, most research has been done on radiographic images like X-rays, CT, MRI, etc. while on natural or camera images we can see there was comparatively less focus.

There is a fundamental domain difference in medical images in comparison with other images be it camera images or radiographic images. Deep-learning based models like classification model or segmentation model, would, in general, look for certain types of anomalies and in many cases, such anomalies would display very delicate texture or color differences thus Image synthesis for medical images must be sensitive enough to learn such delicate distribution and produce images that contain due features properly. Where traditional GAN may not preserve all the textures of CBIS-DDSM screening images, (Guan et al., 2022) have proposed a method of GAN based image augmentation "texture-constrained multichannel progressive GAN (TMPGAN)". The objective was not to handle class imbalance but to generate synthetic images to overcome the issue of less training images available. TMP-GAN applies a progressive generation mechanism that improves image synthesis steadily. Foreground-Generation method is being used in it, which means the model will generate the synthetic lesions in selected areas

of normal/actual images to produce positive case images. A progressive fusing mechanism also makes sure that the synthetic lesion's continuity on the background to preserve the textures.

The other and more significant challenge in training deep learning models for medical images is the desired images are either very less to train the model on or they are extremely unbalanced as most cases would fall in normal/negative class.

A study, proposed in April and Published in May of 2020, merely a couple of months after covid was declared a worldwide pandemic and with an obvious heavy shortage of training images for positive cases, AC-GAN has been put in use for Synthesizing both Covid CXR and normal CXR images to train a classification model for covid detection (Waheed et al., 2020) . On other hand, instead of Image Translation (AC-GAN), (Srivastav et al., 2021) has achieved significant improvement in pneumonia detection by augmenting positive images using image generative GAN model – DC-GAN. However, both studies were not focusing on the "Class Imbalance" issue which is very common across the medical domain.

While most research related to data augmentation using GAN variants were focused to overcome the scarcity of the data itself, there were some researches focused on the challenge of data being extremely biased towards certain class(es) and the rest classes would rarely occur. In (Qasim et al., 2020) researchers talk about the class imbalance issue in the BraTS and ISIC datasets. To achieve the image segmentation task, unlike the traditional GAN where two components, Generator and Discriminator would compete, they introduced a SPADE based GAN with third component called "Segmentor" (Figure 2.2: Basic representation of Red-GAN.) which is fixed and pretrained on the same dataset to obtain the synthetic image segments on the fly.

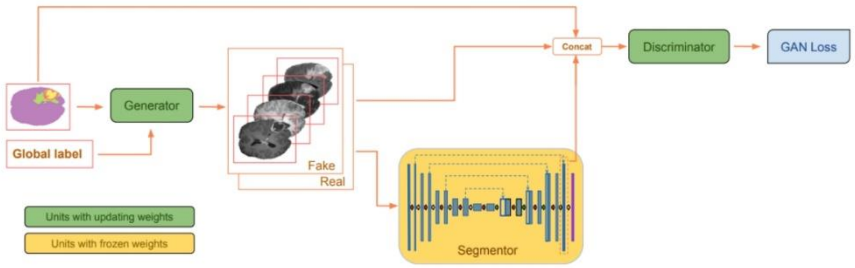


Figure 2.2: Basic representation of Red-GAN. Here we can observe the third component pre-trained "Segmentor" being introduced (Qasim et al., 2020).

A study (Frid-Adar et al., 2018) explored two very basic variants of GANs and those were DCGAN and ACGAN. Unlike DCGAN, ACGAN is a conditional GAN and as external conditional information, ACGAN provides class information in the GAN network. Trained on liver CT images for lesion segmentation, their study not only demonstrates the performance improvement but also compares the performance difference of classification when the model is trained on traditional image augmentation and GAN Based augmentation. Cycle-based GAN combined with YOLO (you only look once) architecture (Hammami et al., 2020) was used for generating synthetic MRI images to be used to train a multi-organ detector model. Instead of one set of generators and discriminator, Cycle GAN is made of two sets and works as bidirectional image translation. The output of the Cycle GAN is then fed into YOLO for detection.

To obtain a reliable GAN based image synthesis on skin lesion images, a study, (Bissoto et al., 2021) reviewed 18 prominent research that claimed of gaining significant improvement in the model for classification or segmentation tasks that were trained on GAN based synthetic images. Further, their study has validated how different real:synthetic image ratio leads to a different outcome. Researchers tried four different GAN variants: SPADE, pix2pixHD, PGAN, and StyleGAN to generate synthetic images and trained classification model Inception v4 with the generated training dataset using various real:synthetic image ratios. Researchers then went ahead and compared two basic techniques of utilizing the synthetic images in the classification model, Augmentation and Anonymization. However, in any terms, they could not achieve as good results as it was claimed in the referred papers.

One common trend that has been noticed in (Bissoto et al., 2021) and (Qasim et al., 2020) is that both were not able to perform well for the skin lesion dataset, while Red-GAN could perform reasonably okay for the brain tumor dataset. The concluded reason for these GANs' inability on performing better was, that "skin lesion images have a more visual appearance in comparison with brain tumor MRI images (or other radiographic images), thus image segmentation and mask to image mapping become more difficult in comparison with MRI images". And this opens a large gap for GAN based image synthesis for camera images and the reason given above, it should not be limited to skin lesion images but other medical images like surgical images or endoscopic images as well.

Other than radiographic images, studies had been carried out on rich in color and texture microscopic images of human protein where DC-GAN has been applied (Verma et al., 2020)

and on dermoscopy skin images (Litjens et al., 2017; Rashid et al., 2019; Qin et al., 2020; Bissoto et al., 2021) where a different variant of GANs has been used for image augmentation. However, none of them focused on handling class imbalance, and only (Bissoto et al., 2021) tried and failed to improve the ultimate classification model. Although modified Style-GAN has provided promising results for skin lesion image generation (Qin et al., 2020)

3. Research Questions

On the bases of reviewing the prominent works of literature so far, the below questions are formulated that the current research will ultimately explore.

- Does class imbalance present in the dataset affect the outcome of the classification of skin lesion images?
- Does GAN based data augmentation help in creating a synthetic dataset for camera/dermoscopic skin lesion images that can improve classification performance?
- Does the skin lesion dataset generated by GAN based data augmentation outperform the dataset generated by traditional image augmentation techniques?
- For the classification of skin lesion images, does the model train on data augmentation perform better than the model train on data anonymization?
- Does “image translation” based GAN perform better than “image generative” GAN?

4. Aim and Objectives

The main aim of this research is to develop a stable GAN model that can generate reliable synthetic medical images. The skin lesion dataset is highly imbalancedd and biased, the goal is to be able to generate synthetic images for a specific class(es) to handle the class imbalance present in the dataset that ultimately results in better trained and reliable classification models.

To achieve the aim following objectives are formulated:

- To load and analyze the dataset to identify and eliminate any error/impurity in the dataset
- To perform the image preprocessing to normalize the images and bring them to a uniform size

- To generate GAN models using different techniques to identify the most suitable GAN based on the nature of the given dataset
- To generate classification models being trained on the augmented dataset.
- To evaluate the performance of GAN and classification models

5. Significance of Study

This research is contributing to the synthetic medical camera image generation by using different variants of GAN models to handle the 'class imbalance' problem in dataset and scarcity of training images which leads to poor performance of classification models. Dermoscopic skin lesion images are selected to be used in this research as in this dataset, images are camera-based images and demonstrate extreme class imbalance. Among all types of skin cancers, 'melanoma' is the most lethal one thus it becomes very critical for medical science to have a stable and reliable melanoma detection mechanism as early diagnosis can greatly improve the survival rate of patients.

'melanoma' is one of the classes of skin lesions in the dataset which is being shadowed by the dominating class 'melanocytic nevus (nv)' the classification models benign trained on such biased datasets mostly perform poorly in melanoma detection. This research is aimed to overcome this issue by oversampling the minority class (here 'melanoma') with synthetic images of the melanoma class generated by using GAN.

In addition, a generic GAN model will not only help in balancing the skin lesion images but can also be utilized in generating other camera based medical images like surgical images of rare conditions or endoscopic images of anomalies found. This research will also open gates for further extended research to develop GANs that can be used domain agnostically.

6. Scope of the Study

To keep the research focused and feasible to be completed in given time duration, the scope of the research work has been limited as below:

- This research will explore only two approaches to image augmentation, traditional image transformation, and GAN based image synthesis. Image synthesis using "variational autoencoders (VAEs)" is included in the research

- Only noise-based Image generative GANs will be explored and only DC-GAN and Style-GAN variants will be further implemented for image augmentation. Image translation-based GAN techniques are not included in the research and so does the image segmentation.
- The classification models are only meant to evaluate the dataset balanced by image augmentation techniques and further improvements of the classification models are not in scope.
- Using reinforcement learning to improve the quality and speed of GAN models by introducing periodic feedback mechanisms in GAN architecture is not included in the scope of current research.

7. Research Methodology

In this research, the primary focus is on developing a GAN model that can perform well on colored and textured medical camera images like dermoscopic skin lesion images rather than focusing more on the image classification model. The whole research is divided into three main parts: Image Generation, Image Classification, and Evaluation.

The detailed flow of the entire research has been discussed in this section. The flowchart in the Figure 7.1: Methodology flow, shows the sequential order of different steps being performed to complete the objectives and achieve the main aim.

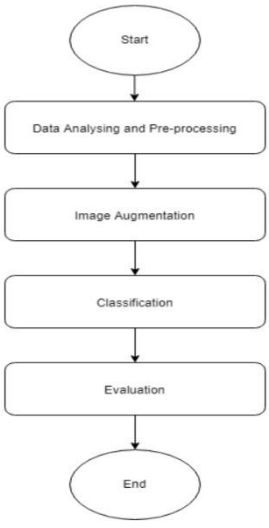


Figure 7.1: Methodology flow

7.1 Dataset analysis and pre-processing

A GAN model to be able to generate medical camera images, (International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset., 2020) is being used.

ISIC 2020 dataset contains:

1. 33,126 JPEG and DICOM images
2. Metadata containing information (patient ID, lesion ID, gender, age, and general anatomic site) for all 33,126 images
3. Duplicate images list
4. Ground truth of all 33,125 images

ISIC 2020 dataset is well organized and clean. However, a few basic steps will be performed as data pre-processing

1. EDA on the metadata of the images and ground truth information
2. Dropping the images
    a. Which were associated with dropped entries of metadata.
    b. Keeping the class ratio constant, reducing the dataset size to make further development feasible yet realistic.
3. Resizing the images to a uniform size
4. Normalizing the image pixel intensity values between (0,1)

7.2 Image Augmentation

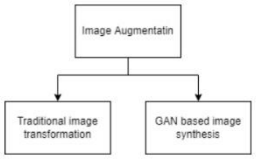


Figure 7.2: Image augmentation techniques

On the Assumption that present class imbalance in ISIC dataset will impact the classification model trained on this dataset and will be highly biased towards majority classes, image augmentation becomes critically important and thus it is the primary focus of this research. (Bissoto et al., 2021; Guan et al., 2022) extensively talks about different generative data augmentation techniques that include both image-to-image translation and noise-based image generation. However, fundamentally speaking two main ways of augmenting the images

(shown in Figure 7.2: Image augmentation techniques) will be explored in this research, Traditional image transformation and GAN based image synthesis (Dumagpi and Jeong, 2021)

7.2.1 Traditional image transformation

Although less sophisticated, image transformation techniques like rotating, zooming, cropping, etc. have been used to upsample the images for any particular class(es). And in many studies (Verma et al., 2020; Waheed et al., 2020; Dumagpi and Jeong, 2021) , image transformation has either been used with image synthesis or compared with image synthesis concerning the effectiveness.

Given the nature of the images and the factors responsible for classification, a few techniques of transformation like thresholding, erosion, dilation, opening, closing, etc. cannot be used to augment new images as they might alter the color, contrast, texture of the image. Whereas linear transformation techniques like resizing/scaling, cropping, zooming in/out, rotating, and flipping can be safely used.

In the context of traditional image transformation techniques, this research will be a comparative study of the effectiveness of classification models trained on the dataset that included image transformation + GAN in data augmentation, only used GAN based synthetic images for data augmentation, and standalone usage of image transformation for data augmentation.

7.2.2 GAN based image augmentation

Mainly classified into two types, image to image translation model and noise-based image generation model, many variants of GAN based models are discussed (Singh and Raza, 2020; Bissoto et al., 2021).

Inspired by studies (Qin et al., 2020; Verma et al., 2020) with comparatively similar dataset and promising outcome, this research will explore and experiments with two widely accepted GAN variants, DC-GAN and Style-GAN. DC-GAN is a relatively simpler GAN variant with both generator and discriminator comprising of the deep convolutional network. Unlike conditional GANs, DC-GAN doesn't have external conditioning as the input and output layer of the

discriminator network contains a single neuron and thus can't produce probability distribution for the generated image.

GAN can produce realistic images but being stable they cannot achieve high resolution. Style-based GANs can produce higher resolution output images where vanilla GAN might collapse. The low-resolution issue can be resolved by PGGAN too, PGGAN has limitation in effective control over the features of the image and style of the image during the image generation on the other hand Style-based GAN can generate high resolution images with good control over the features and style (Qin et al., 2020).

7.3 Image Classification

The main aim of this research is limited to generating desired GAN model to overcome the class imbalance problem and thus this research doesn't focus on improving the image classification models. These models will be used only for comparing the quality of the training dataset.

This research will use two classification models, basic CNN architecture and VGG Net using transfer learning. The image classification models will be trained on different datasets while keeping constant hyper-parameters, activation functions, and overall architecture. Once trained, these models will be evaluated on the same test dataset using the same evaluation matrices. Dataset generation is already discussed in the above sections.

7.4 Evaluation

Evaluation of this research will be a comparative study of the outcome of different models and experiments. As the development work in this research will be done in two parts, they both will be evaluated separately.

7.4.1 Evaluating GAN models

(Borji, 2019) talks briefly about the different measures to evaluate the performance of GAN models. In their study, they have proposed basic characteristics of a good GAN model evaluation measure

- Evaluation measures should favor the GAN model that can generate high-fidelity samples
- It should favor the GAN model that can generate a diverse sample
- It should favor the GAN model with controllable sampling
- It should favor the GAN model with well-defined bounds
- Evaluation measures should be sensitive to image distortion and image transformations.
- The evaluation measure's outcome should be in line with human perceptions.
- It should be less computational complexity.

This study will mainly be dependent on evaluating the classification model to determine whether the dataset generated using GAN is helping the classification process or not. However, there can be independent measures to evaluate GAN performance. Broadly speaking, there are two ways of evaluating the GAN, quantitive measures and qualitative measures.

Inspired by, (Qin et al., 2020) this study will evaluate the GANs based on

- Quantitive Measures
    - Inception Score – IS
      IS = exp(Ex[KL(p(y|x)||p(y))]) (Qin et al., 2020)
    - Frechet Inception Distance – FID
      $FID(r,g) = \|\mu r - \mu g\|2\text{^}2 + Tr[\Sigma r + \Sigma g - 2(\Sigma r \Sigma g)\text{^}(1/2)]$ (Borji, 2019)
- Manually validating by visualizing the output of GAN.

7.4.2 Evaluating classification models

Given that the dataset is highly imbalanced and biased towards dominating class(es), the High Accuracy value is often misleading. For medical image classification, a high rate of false negatives cannot be accepted, on another hand high rate of false positives will require a continuous cross-checking mechanism, making the final diagnosis more time-consuming and expensive.

With this understanding, this research will evaluate the classification model using the following measures. The confusion matrix will be generated as one class is a positive case and the rest all being negative cases.

- Sensitivity (Recall, True Positive Rate): The number of positive cases that are correctly predicted out of the total positive cases

    - Sensitivity = True Positive / (True Positive + False Negative)
    - The value of sensitivity should be as high as possible
- Specificity: The number of negative cases that are correctly predicted out of the total negative cases
    - Specificity = True Negativ / (True Negative + False Positive)
    - The value of specificity should be as high as possible
- Precision(Positive predictive rate): Rate of correctly predicted positive case our of total positive prediction.
    - Precision = True Positive / (True Positive + False Positive)
    - The value of precision should be as high as possible
- ROC curve plotting will be used to visualize the performance of the classification.

However, the conclusion of the research will not be comparative but quantitive. Based on the evaluation result, this research will try to propose the answers to the questions mentioned in the 'research question' section.

8. Required Resources

Below listed software and hardware will be required to carry out the research.

8.1 Software requirement

- Operating System: Windows 10 20H2 or above
- Language: Python 3.8
- For on-prem development work
    - Conda Package (Python packages) Manager: Anaconda Navigator 2.0
    - Notebook/IDE: Jupyter Notebook
- For online development work
    - Google collab
- Commonly used python packages for GAN and Classification model development, Data loading, and visualization, and for supporting development tasks.
- Microsoft Office 360 16.0.14

8.2 Hardware requirement

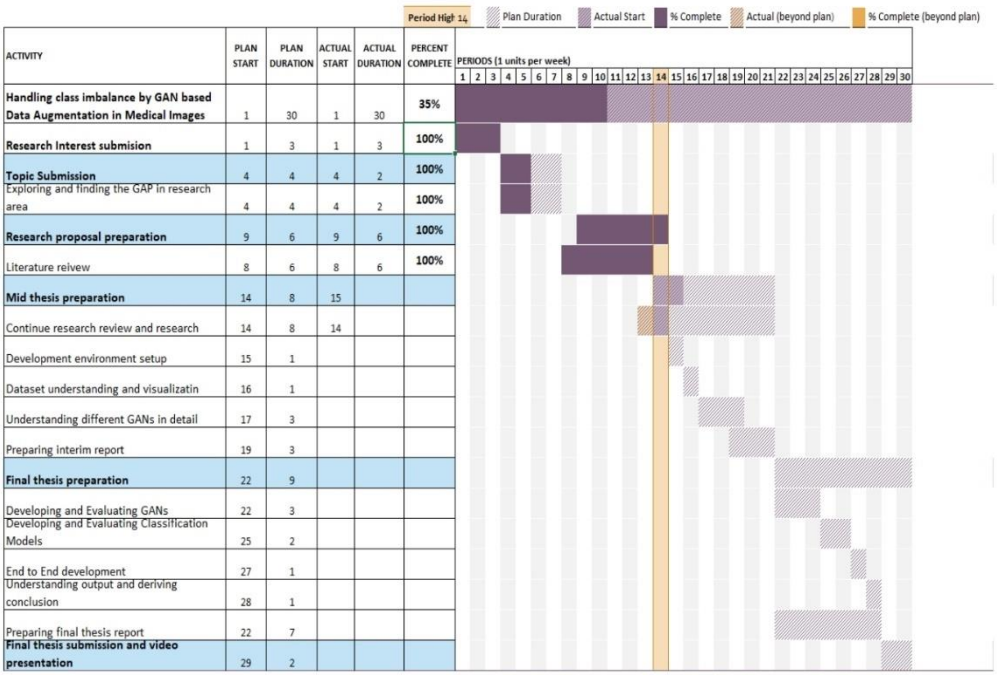
- Processor: Intel® Core ™ i7 – 10510U

- Clock rate: 1.80 GHz
- Cores: 4
- Logical Cores: 8

- RAM: 16 GBs
- Storage: (to support development environment, dataset, and development) 100 GBs

9. Research Plan

Below is shown a brief research plan that this research is following.



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Handling class imbalance by GAN based Data Augmentation in Medical Images	1	30	1	30	35%
Research Interest submision	1	3	1	3	100%
Topic Submission	4	4	4	2	100%
Exploring and finding the GAP In research area	4	4	4	2	100%
Research proposal preparation	9	6	9	6	100%
Literature reivew	8	6	8	6	100%
Mid thesis preparation	14	8	15		
Continue research review and research	14	8	14		
Development environment setup	15	1			
Dataset understanding and visualizatin	16	1			
Understanding different GANs in detail	17	3			
Preparing interim report	19	3			
Final thesis preparation	22	9			
Developing and Evaluating GANs	22	3			
Developing and Evaluating Classification Models	25	2			
End to End development	27	1			
Understanding output and deriving conclusion	28	1			
Preparing final thesis report	22	7			
Final thesis submission and video presentation	29	2			

Figure 8.1: A brief research plan

10. Risk and contingency plan

Followings are the potential risk factors that can affect the timeline and outcome of the research work. To maintain the resiliency in the research work, against each risk factor suitable mitigation has been planned.

- Risk: Hardware/Software issues
    - All the documents, references, and development work including the dataset are maintained on and in sync with cloud storage.
    - Windows operating system's image capturing the development environment has been taken as a backup
    - Google collab (or any other online development platform) can be utilized to continue development work in case of on-prem development environment is not available.
- Risk: Time constraint
    - The scope of the study has been planned according to the available time
    - However, scop is designed in a way that some buffer time should be available for additional experiments (e.g., working on classification model improvement). In case of critical time constraints, these extra experiments can be dropped to achieve the main goal by completing all the objective in proper manner.
- Risk: The research is not generating the expected outcome
    - Instead of waiting for the final outcome, be continuously in contact with the thesis supervisor and discuss the periodic progress and outcome and seek his guidance and if needed university professor's guidance and proceed accordingly.

References

Anon (2020) *International Skin Imaging Collaboration. SIIM-ISIC 2020 Challenge Dataset. Creative Commons Attribution-Non Commercial 4.0 International License*.

Bissoto, A., Valle, E. and Avila, S., (2021) GAN-Based Data Augmentation and Anonymization for Skin-Lesion Analysis: A Critical Review. [online] Available at: http://arxiv.org/abs/2104.10603.

Borji, A., (2019) Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179, pp.41–65.

Dumagpi, J.K. and Jeong, Y.J., (2021) Evaluating gan-based image augmentation for threat detection in large-scale xray security images. *Applied Sciences (Switzerland)*, 111, pp.1–21.

Dumagpi, J.K., Jung, W.Y. and Jeong, Y.J., (2020) A new GAN-based anomaly detection (GBAD) approach for multi-threat object classification on large-scale x-ray security images. *IEICE Transactions on Information and Systems*, E103D2, pp.454–458.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J. and Greenspan, H., (2018) GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, pp.321–331.

Fu, Y., Li, X. and Ye, Y., (2020) A multi-task learning model with adversarial data augmentation for classification of fine-grained images. *Neurocomputing*, 377, pp.122–129.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., (n.d.) *Generative Adversarial Nets*. [online] Available at: http://www.github.com/goodfeli/adversarial.

Guan, Q., Chen, Y., Wei, Z., Heidari, A.A., Hu, H., Yang, X.H., Zheng, J., Zhou, Q., Chen, H. and Chen, F., (2022) Medical image augmentation for lesion detection using a texture-constrained multichannel progressive GAN. *Computers in Biology and Medicine*, 145.

Hammami, M., Friboulet, D. and Kechichian, R., (2020) *CYCLE GAN-BASED DATA AUGMENTATION FOR MULTI-ORGAN DETECTION IN CT IMAGES VIA YOLO. 2020 IEEE International Conference on Image Processing (ICIP)*.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B. and Sánchez, C.I., (2017) *A survey on deep learning in medical image analysis*. *Medical Image Analysis*, .

M E Vestergaard, S W Menzies, P Macaskill and P E Holt, (2008) Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting.

Qasim, A.B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J.C., Sekuboyina, A., Kofler, F., Lipkova, J., Li, H. and Menze, B., (2020) Red-GAN: Attacking class imbalance via conditioned generation. Yet another perspective on medical image synthesis for skin lesion dermoscopy and brain tumor MRI. [online] Available at: http://arxiv.org/abs/2004.10734.

Qin, Z., Liu, Z., Zhu, P. and Xue, Y., (2020) A GAN-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 195.

Rashid, H., Tanveer, M.A. and Aqeel Khan, H., (2019) Skin Lesion Classification Using GAN based Data Augmentation. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2019, pp.916–919.

Singh, N.K. and Raza, K., (2020) *Medical Image Generation using Generative Adversarial Networks*.

Srivastav, D., Bajpai, A. and Srivastava, P., (2021) Improved classification for pneumonia detection using transfer learning with GAN based synthetic image augmentation. In: *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*. Institute of Electrical and Electronics Engineers Inc., pp.433–437.

Verma, R., Mehrotra, R., Rane, C., Tiwari, R. and Agariya, A.K., (2020) Synthetic image augmentation with generative adversarial network for enhanced performance in protein classification. *Biomedical Engineering Letters*, 103, pp.443–452.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F. and Pinheiro, P.R., (2020) CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access*, 8, pp.91916–91923.