Handling class imbalance by GAN based Data Augmentation in Medical Images

Amitkumar M Maheshwari

Mid Thesis Report

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. 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. To reduce the network complexity, computational cost, and bring robustness an RL powered autoencoder is incorporated along with the GAN. 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 where input to GAN is generated with the help of autoencoder and reinforcement learning. 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
FCGAN	Face conditional GAN
SNGAN	Spectrally normalized GAN
SPGGAN	Self-attention progressive GAN
CESRGAN	Cascade ensemble super resolution GAN
TTUR	Two timescale update rule
AE	Autoencoder
RL	Reinforcement Learning
GFV	Global feature vector
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. 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: Class distribution in ISIC 2020 dataset show the distribution of different cases of skin lesions.

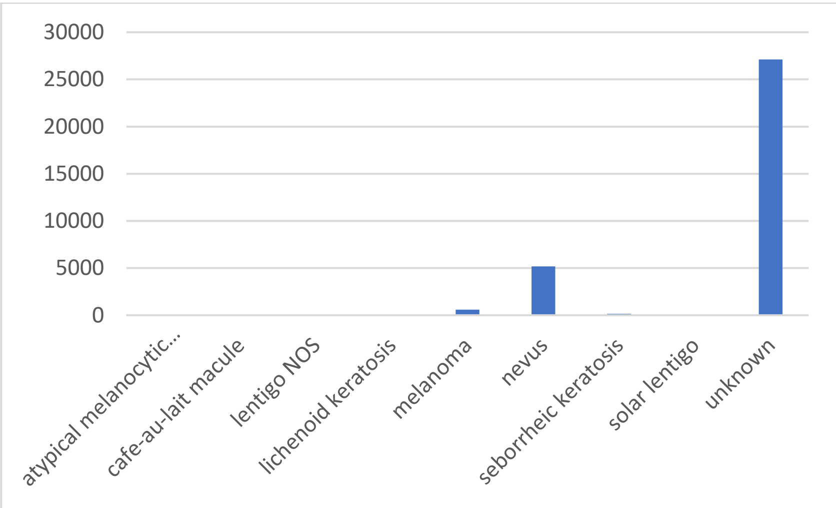

Figure 1.1: Class distribution in ISIC 2020 dataset

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.

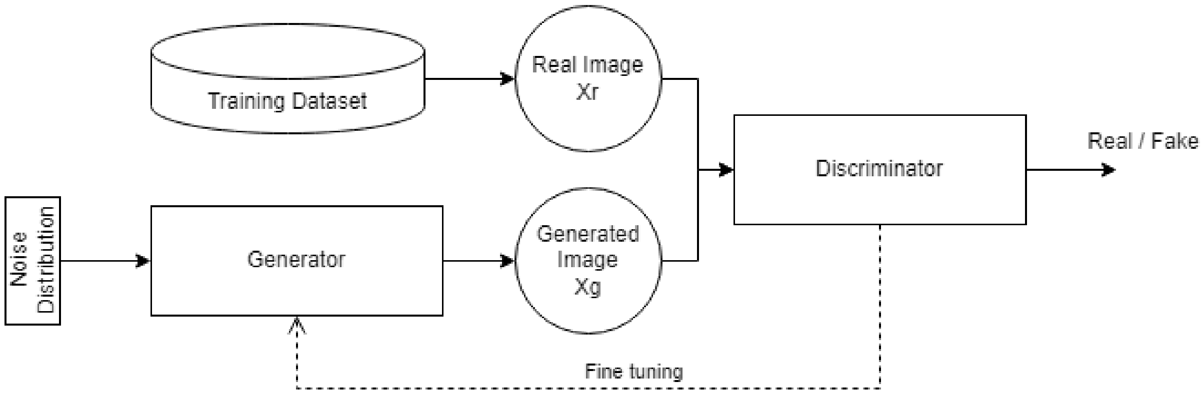

Figure 1.2: Basic architecture of GAN

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:

- To load and analyze the dataset to identify and eliminate any error/impurity in the dataset as well as to perform the image preprocessing to normalize the images and bring them to a uniform size
- To train the RL powered autoencoder system that can pick the best input latent representation for GAN network

- To generate GAN models using different techniques to identify the most suitable GAN with early feedback loop breaker mechanism, 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

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 in order 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 great 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.

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).

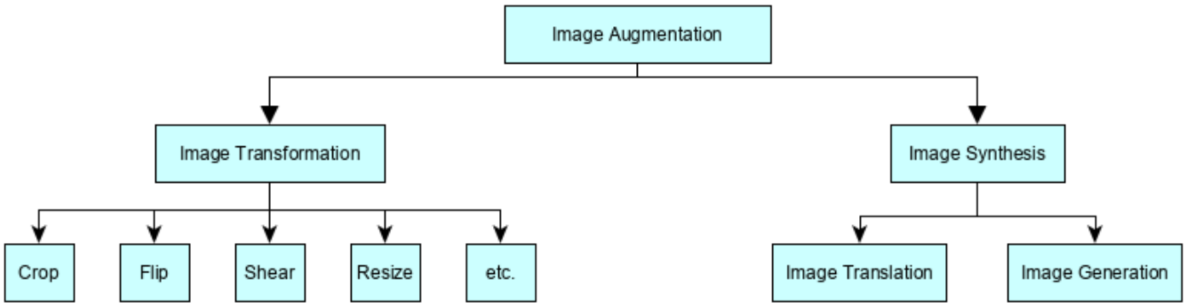

Figure 2.1: Traditional(left) and Generative(right) techniques of Images Augmentation used in (Dumagpi and Jeong, 2021)

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.

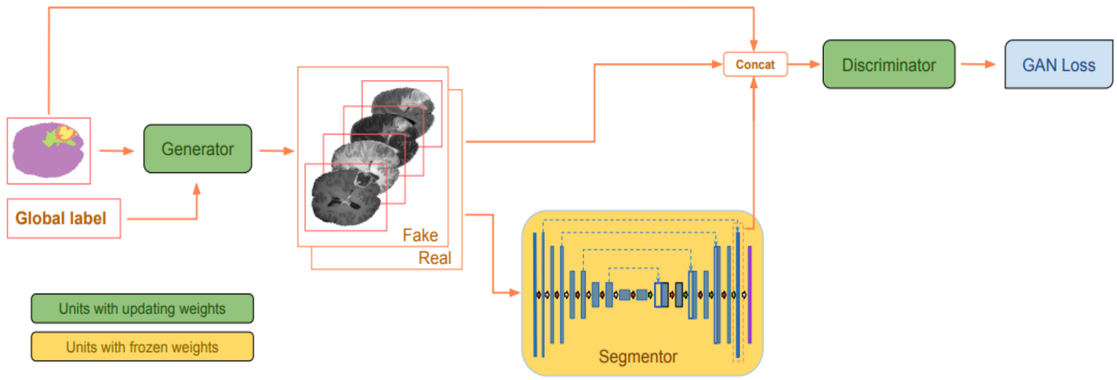

Figure 2.2: Basic representation of Red-GAN. Here we can observe the third component pre-trained "Segmentor" being introduced (Qasim et al., 2020).

(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.

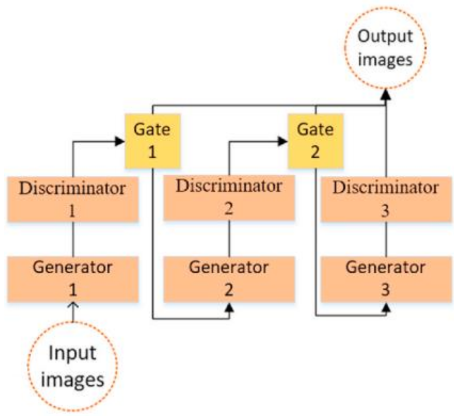

Figure 2.3: Cascading architecture of GANs as proposed in (Shahsavari et al., 2021)

Equation for the gates in CESR-GAN,

$$Q(x) = 1 \text{ if } D(x) > t \text{ else } 0, \text{ where } D(x) \text{ is discriminator's value and } t \text{ is threshold}$$

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 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 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.

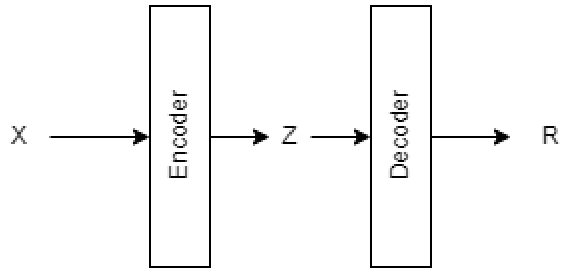

Figure 2.4: Basic structure of autoencoder(Zhai et al., 2019)

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 also 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	To generate synthetic chest x-ray images as the covid outbreak was fairly recent and relevant images were not widely available	Precision, Recall, F1-Score, Accuracy, Sensitivity, Specificity	Accuracy increased by 10%

		dataset initiative			
(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 Classificat ion 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),	GAN based anomaly detection: Bi-GAN,	SIX-ray. (Dumagpi and Jeong, 2021) has	To address an extreme class imbalance issue in	Precision, Recall, F1-Score	Overall fair enough improvement in

(Dumagpi and Jeong, 2021)	SVM	extended the dataset further	case of security x ray image dataset		classification task.
(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	Precision, Recall, F1-Score	Around 2-3% improvement in all the

			preserve the delicate textures in medical images to improve the classification		evaluation 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 inpaint 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: Similar to (Shubham et al., 2021).
(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.

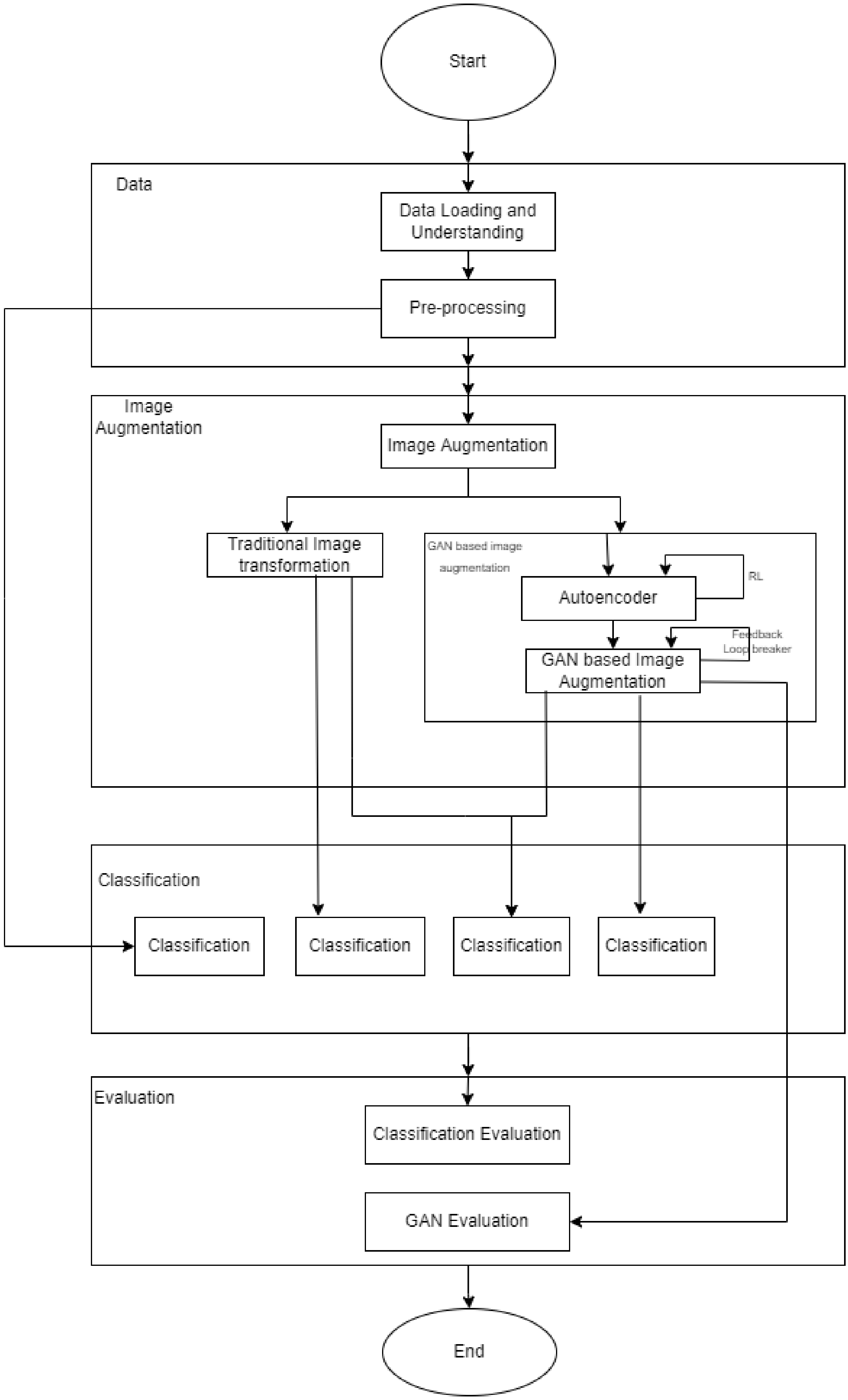

Figure 3.1: Flowchart of overall process execution

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.

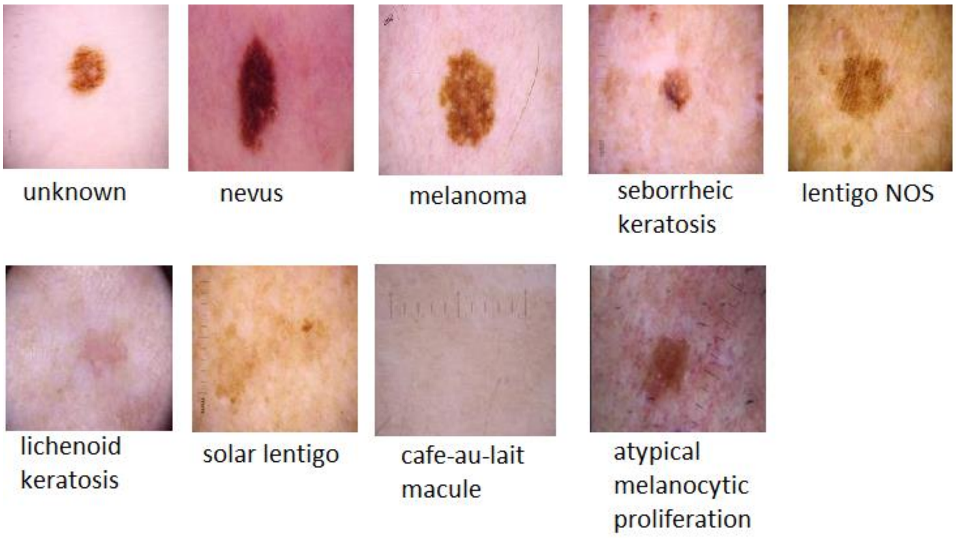

Figure 3.2: Sample images of different types of skin lesions

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. Where as “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.

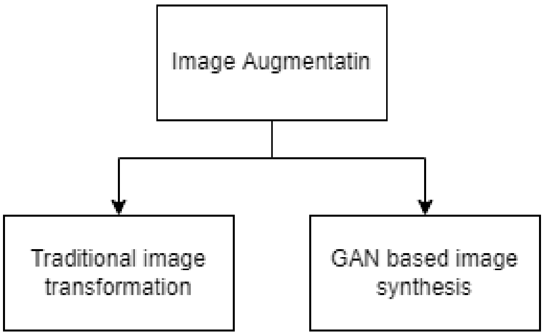

Figure 3.4: Image augmentation techniques

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 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.

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.

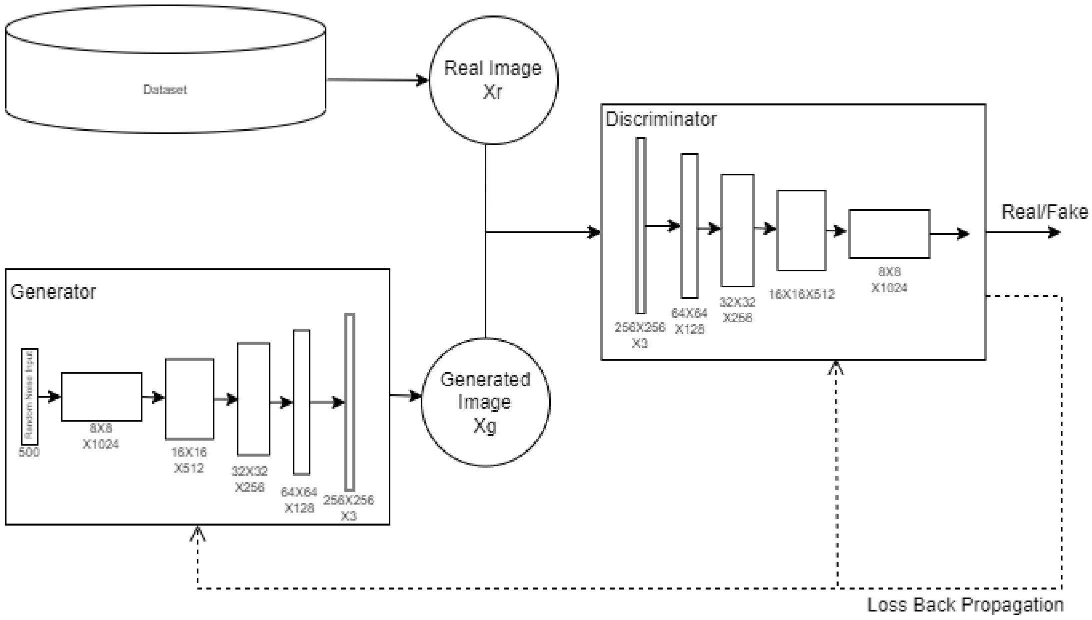

Figure 3.5: Basic architecture of DC-GAN

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. 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.

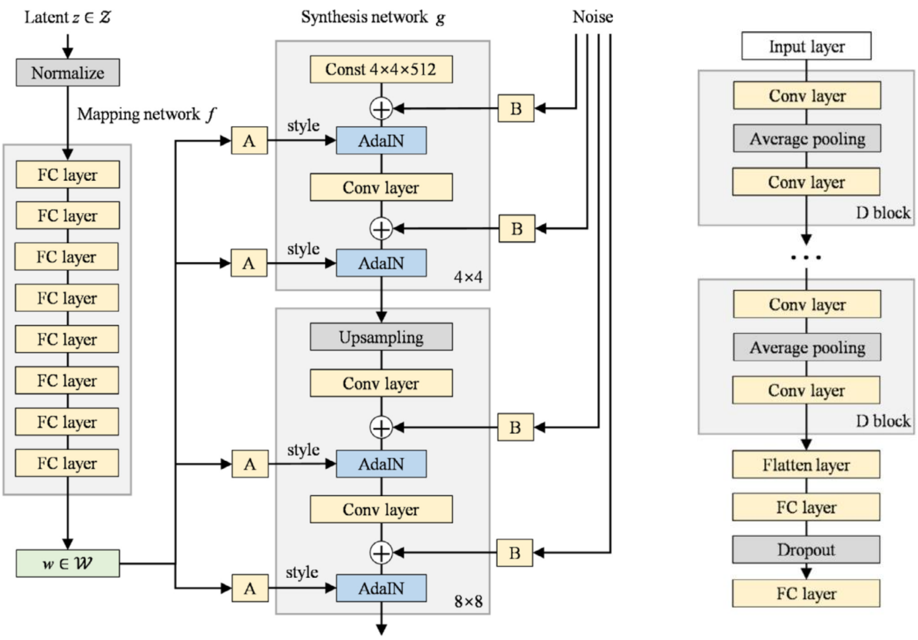

Figure 3.6: Architecture of Style-GAN (a) Generator, (b) Discriminator (Qin et al., 2020).

As it is clear that vanilla 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. Figure 3.6 shows architecture of Style-GAN. 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).

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 3.7: shows the architecture of the autoencoder that is being used in this research.

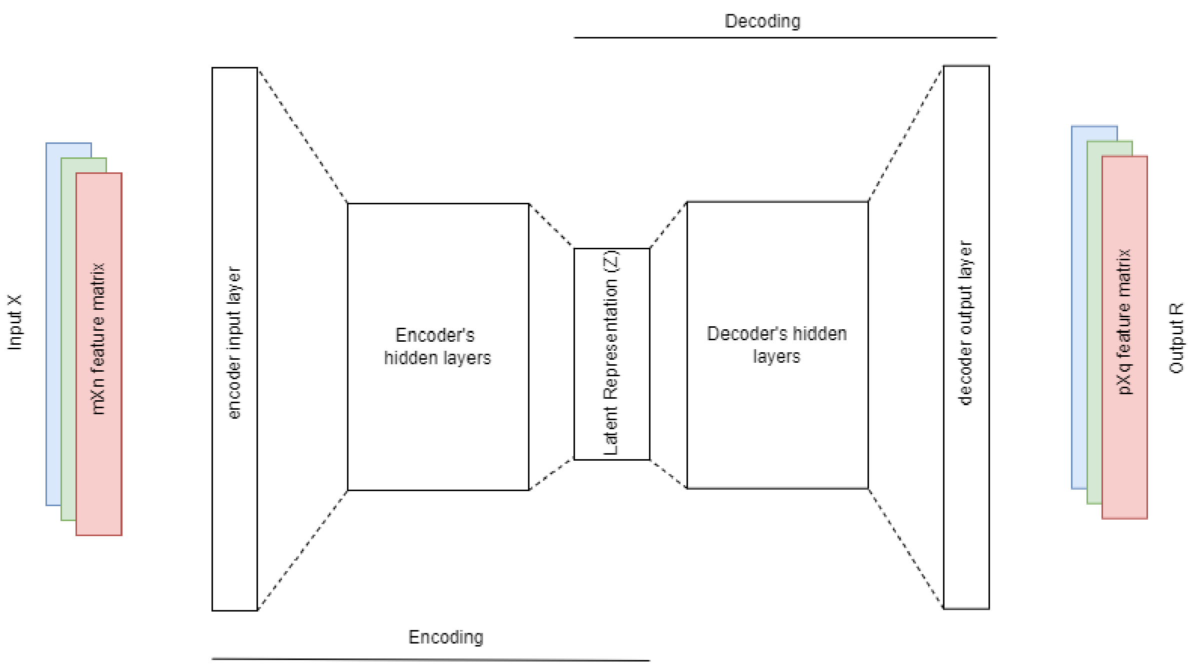

Figure 3.7: Autoencoder Architecture

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.

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

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

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 3.8 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.

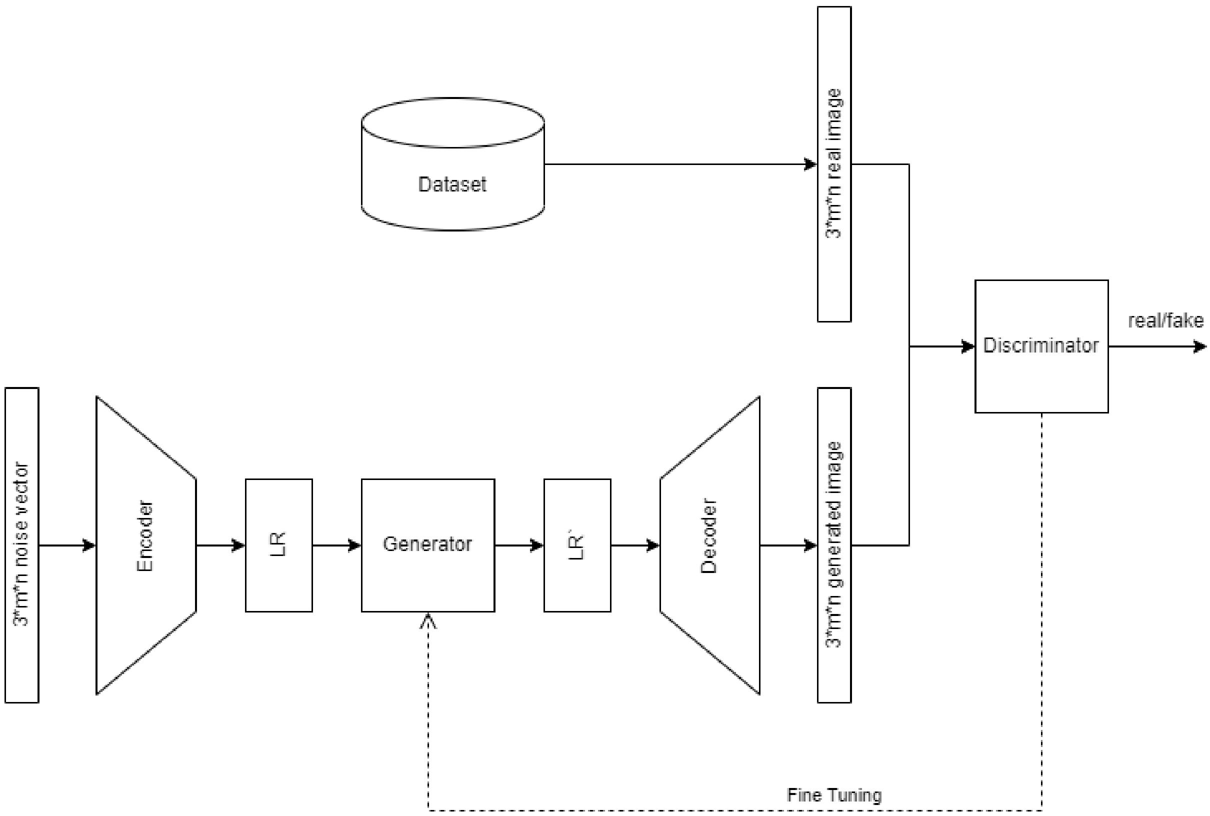

Figure 3.8: Autoencoder with GAN network

From the figure 3.8 it becomes quite clear that instead of training generator network with the random noise vector of size 3*m*n, encoder network has reduce 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.

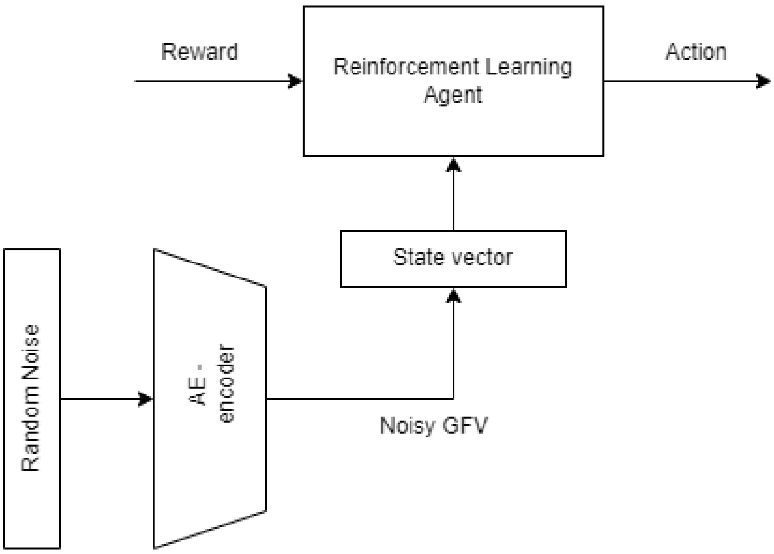

Figure 3.9: Autoencoder used with RL

Figure 3.9: 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.

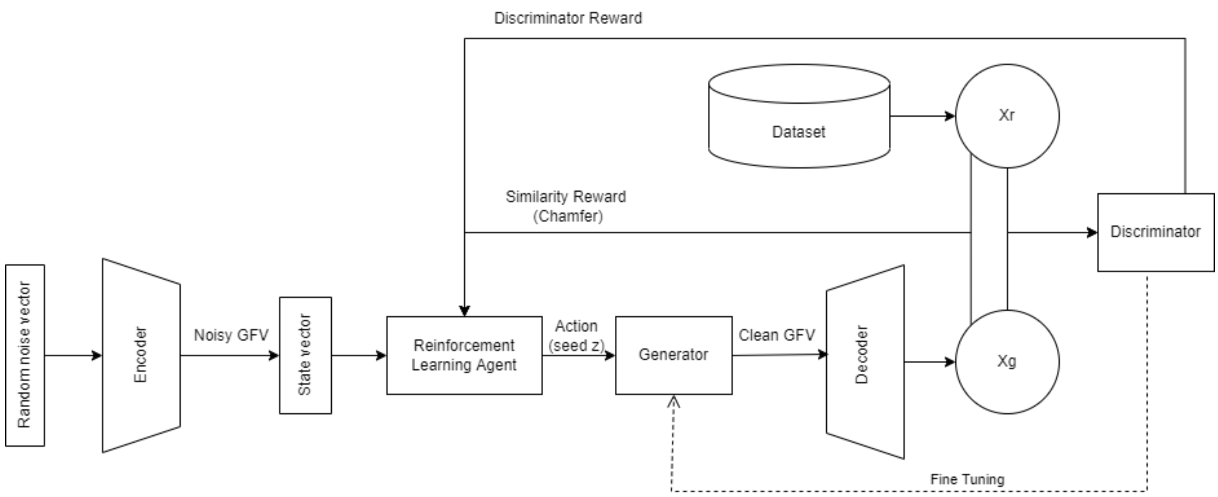

Figure 3.10: An integrated system of AE, GAN, and RL

As it is seen in the figure 3.10, 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 $r = f(r_d, r_c)$

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 matrices. 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.

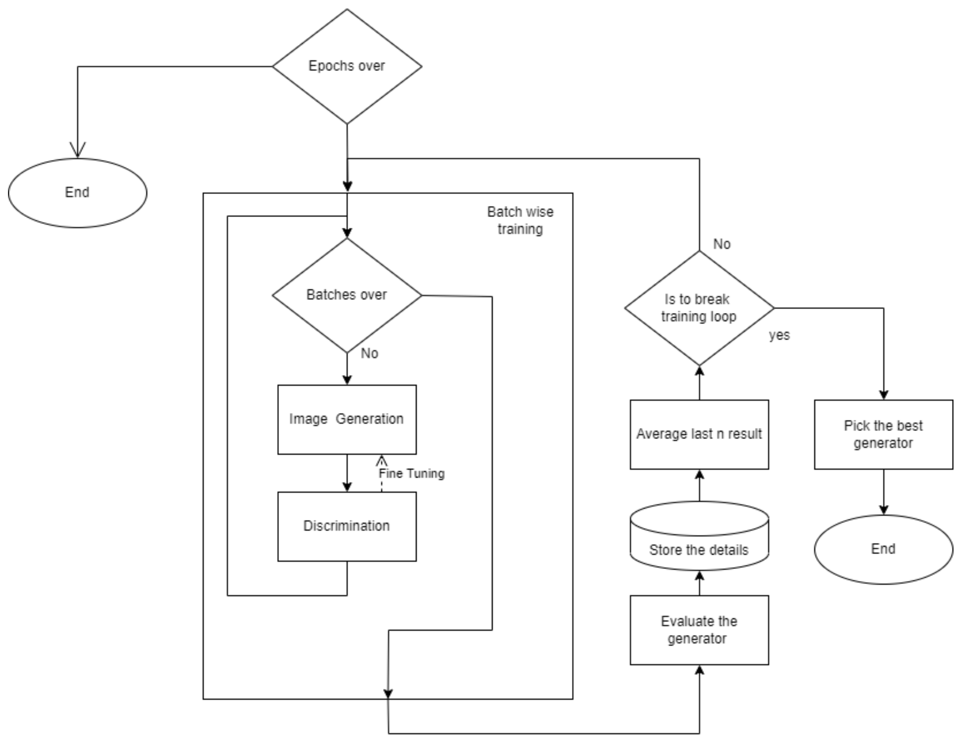

Figure 3.11: Execution flow of loop breaker mechanism in GAN

Figure 3.11 demonstrate the loop breaking mechanism placed in normal GAN work flow. Concept of moving average is being used in the loop breaker.

MA = ($\Sigma_{(i \text{ to } i-n)} R_i$) / n

Where n is the number of the terms to be considered in averaging, 'i' is the last index of the collection R where generator's evaluation data is being stored.

The last generator's output is being compared with the moving average and '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(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.

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
- 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.

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.

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.

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: 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

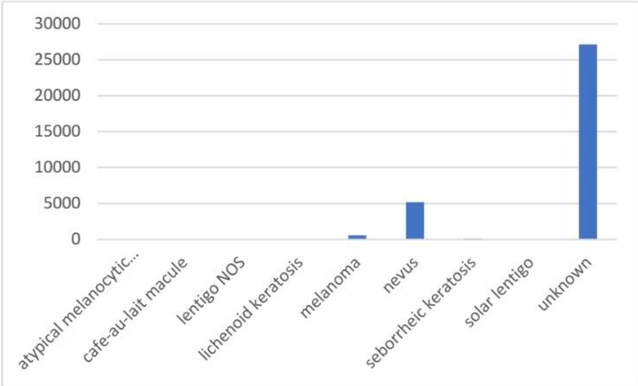

Figure 1.1: Class distribution in ISIC 2020 dataset

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.

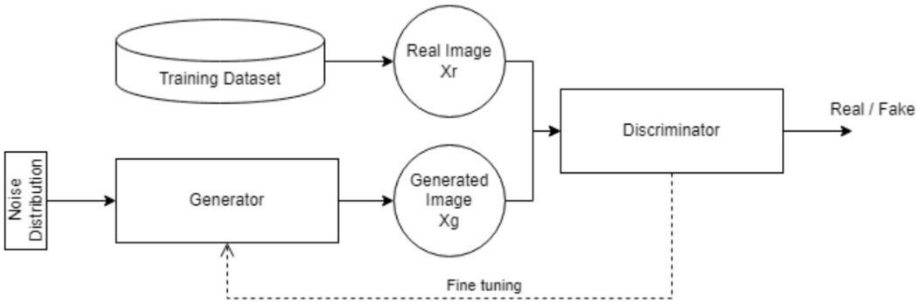

Figure 1.2: Basic architecture of GAN

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.

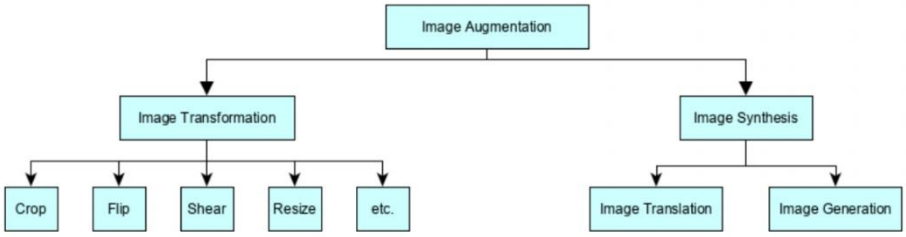

Figure 2.1: Traditional(left) and Generative(right) techniques of Images Augmentation used in (Dumagpi and Jeong, 2021)

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.

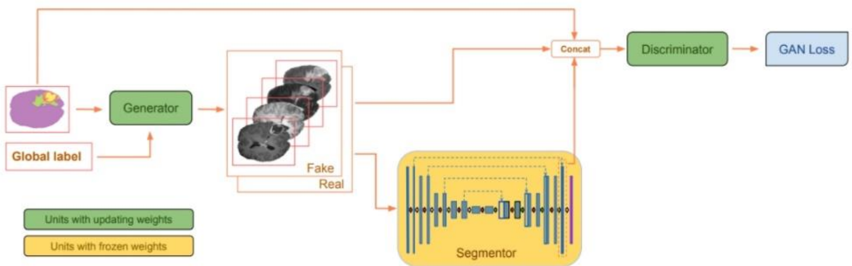

Figure 2.2: Basic representation of Red-GAN. Here we can observe the third component pre-trained "Segmentor" being introduced (Qasim et al., 2020).

9

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.

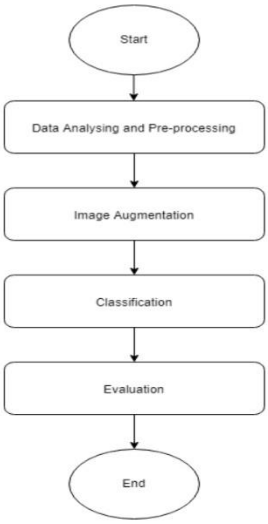

Figure 7.1: Methodology flow

13

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

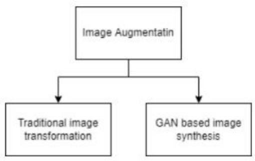

Figure 7.2: Image augmentation techniques

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

    $FID(r,g) = \|\mu r - \mu g\|2$^2 + $Tr[\Sigma r + \Sigma g - 2(\Sigma r \Sigma g)$^(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.

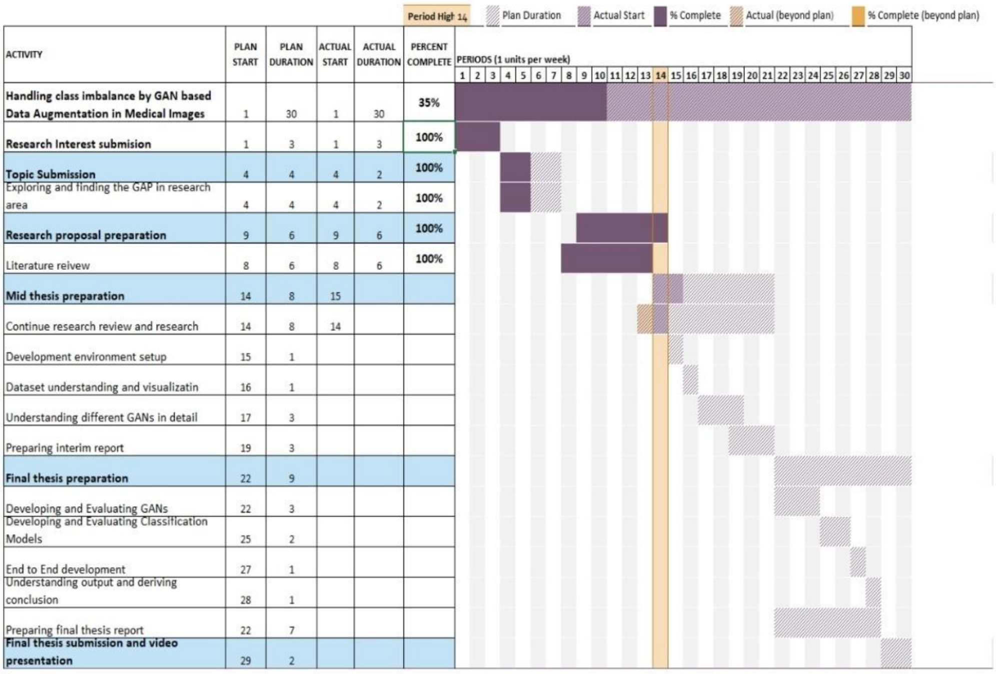

Figure 8.1: A brief research plan

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.

19

- 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.