2022

SkinCAN AI: A deep learning-based skin cancer classification and segmentation pipeline designed along with a generative model

Shivang Rana
University of Windsor

SkinCAN AI: A deep learning-based skin cancer classification and segmentation pipeline designed along with a generative model

By

Shivang Rana

A Thesis
Submitted to the Faculty of Graduate Studies
Through the Department of Electrical and Computer Engineering
In Partial Fulfillment of the Requirements for
the Degree of Master of Applied Science
at the University of Windsor

Windsor, Ontario, Canada

2022

SkinCAN AI: A deep learning-based skin cancer classification and segmentation pipeline designed along with a generative model

by

Shivang Rana

APPROVED BY:

_____

P. Moradian Zadeh

School of Computer Science

_____

B. Balasingam

Department of Electrical and Computer Engineering

_____

J. Wu, Co-Advisor

Department of Electrical and Computer Engineering

_____

H. K. Kwan, Co-Advisor

Department of Electrical and Computer Engineering

April 26, 2022

DECLARATION OF ORIGINALITY

This thesis incorporates the outcome of research under the co-supervision of Professor H. K. Kwan and Professor Jonathan Wu. I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged under the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office. This thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

The rarity of Melanoma skin cancer accounts for the dataset collected to be limited and highly skewed, as benign moles can easily mimic the impression of the melanoma-affected area. Such an imbalanced dataset makes training any deep learning classifier network harder by affecting the training stability. We have an intuition that synthesizing such skin lesion medical images could help solve the issue of overfitting in training networks and assist in enforcing the anonymization of actual patients. Despite multiple previous attempts, none of the models were practical for the fast-paced clinical environment. In this thesis, we propose a novel pipeline named SkinCAN AI, inspired by StyleGAN but designed explicitly considering the limitations of the skin lesion dataset and emphasizing the requirement of a faster optimized diagnostic tool that can be easily inferred and integrated into the clinical environment. Our SkinCAN AI model is equipped with its module of adaptive discriminator augmentation that enables limited target data distribution to be learned and artificial data points to be sampled, which further assist the classifier network in learning semantic features. We elucidate the novelty of our SkinCAN AI pipeline by integrating the soft attention module in the classifier network. This module yields an attention mask analyzed by DenseNet201 to focus on learning relevant semantic features from skin lesion images without using any heavy computational burden of artifact removal software. The SkinGAN model achieves an FID score of 0.622 while allowing its synthetic samples to train the DenseNet201 model with an accuracy of 0.9494, AUC of 0.938, specificity of 0.969, and sensitivity of 0.695. We provide evidence in our thesis that our proposed pipelines outperform other state-of-the-art existing networks developed for this task of early diagnosis.

DEDICATION

To

my parents Sarita Rana, Rajesh Rana, and my brother Vansh,

my Nani Tara, my aunt Shikha, my aunt Preeti,

my dearest uncles Deepak, Dinesh,

my buddies Idrish Vhora, Aman Sansowa, Akash Parmar & Bhavana Sharma,

my friends Jagpreet, Rudra, Hardik, Harsh, Umang, Jay, Martin Andres, Alex,

Ryan, Srijan, Yuvraj for their insurmountable support,

and my store managers Mark & Ella, without whom it would never have been possible.

ACKNOWLEDGEMENTS

I would like to convey my heartfelt thanks to Dr. H. K. Kwan, who thoroughly guided me with my thesis. He has been a constant source of motivation and a kind mentor to me. I can't imagine finishing up my thesis without his immense support. I am lucky that not only did I get a chance to learn from him but also to have him as my co-supervisor.

I would also like to convey my heartfelt thanks to Dr. Jonathan Wu, who gave me direction throughout my degree and provided me with this opportunity to learn from him. He had graced me with his knowledge whenever I wished to seek it. I am lucky that have him as my co-supervisor.

I want to express gratitude to both of my supervisors for supporting me during this challenging time in my life.

I am very grateful to Dr. Pooya Moradian Zadeh and Dr. Balakumar Balasingham for serving as my thesis committee members and providing me with insights while igniting my curiosities toward new research directions. I got a chance to learn from Dr. Pooya Moradian Zadeh by taking a course under him during my M.App.Sc. study that constantly helped me progress toward my research goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
GAN	Generative Adversarial Network
VAE	Variational Autoencoders
D	Discriminator
G	Generator
BCE	Binary Cross-Entropy
SGD	Stochastic Gradient Descent
DCGAN	Deep Convolutional Generative Adversarial Network
CADx	Computer-Aided Diagnosis
CADe	Computer Aided Detection
ANN	Artificial Neural Network
GPU	Graphic Processing Unit
CPU	Central Porcessing Unit
ISIC	International Skin Imaging Collection
kNN	K-Nearest Neighbors
SVM	Support Vector Machine
RF	Random Forest
GLCM	Gray Level Co-occurrence Matrix
PCA	Principal Component Analysis
CNN	Convolutional Neural Network
AUC	Area under the curve
SVM	Support Vector Machines
RCNN	Region-based Convolutional Neural Network
DB	Dense Blocks

CLIP	Contrastive Language-Image Pre-training
XAI	Explainable Artificial Intelligence
FID	Fréchet Inception Distance
IS	Inception Score
PGGAN	Progressive Growing of Generative Adversarial Network
AP	Average Precision
AC	Accuracy
SE	Senstivity
SP	Specificity
ReLU	Rectified Linear Unit
CLAHE	Constrast Limited Adaptive Histogram Equalization
t-SNE	t-distribution stochastic neighbor embedding
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
SCC	Squamous Cell Carcinoma
VASC	Vascular Lesion
AK	Actinic Keratosis
DF	Dermatofibroma
BCC	Basal Cell Carcinoma
BKL	Benign Keratosis
MEL	Melanoma
NV	Nevus
CCE	Categorical Cross Entropy
SOTA	State of the Art
FC	Fully Connected layer

UNIT	Unsupervised image-to-Image Translation
MUNIT	Multimodal Unsupervised image-to-Image Translation
AdaIN	Adaptive Instance Normalization
WGAN	Wasserstein GAN
KL divergence	Kullback-Leibler divergence
LAPGAN	Laplacian Generative Adversarial Network

CHAPTER 1
INTRODUCTION

1.1 Overview

Early diagnosis of skin cancer can lead to an effective treatment cycle while helping increase the overall survival rate and recovery and assist in predicting any recurrences [1]. This requirement for an early diagnosis tool creates an opportunity for artificially intelligent methods to take up the mantle of initial diagnosis deployed to understand any advent or even presence of malignant cells. The benefits of deploying such accurate, low-cost artificially intelligent (AI) based diagnostic aids for detecting skin cancer include better access to clinical care and lesser need for biopsies, resulting in lower hospital expenses and better survival rate, eventually reducing the mental pressure on patients and family. Implementing AI-based diagnostic methods to assist clinical environments where adequate clinical expertise and equipment are absent. Therefore such a clinical tool could serve the less financially established population. The growing number of various image modalities such as dermoscopy and histopathology available to capture the lesion's essence has helped create a catalog of publicly available datasets. Addressing the potential of AI in this domain, deep learning algorithms trained on such datasets requiring minimal assets and constantly striving to improve efficiency could optimize and streamline clinical workflows [2].

1.2 Motivation

Cancer is responsible for over 10 million deaths in 2020 worldwide, and it has been estimated to disrupt over 1.9 million lives in the US in 2021 [3]. According to National Cancer Institute, among all the deadly cancers present globally, the most common type of cancer is skin cancer, having more diagnosed cases than the sum of all the other cancer cases in the US annually [4]. It has been statistically estimated that more than two people die every hour cause of complications from skin cancer [5]. According to predictive statistics, an increment of newly diagnosed melanoma cases by 5.8% and 4.8% in deaths

caused due to melanoma is indicated for 2021 [5]. There is a growth of 44% in the number of invasive melanoma cases detected yearly in the previous decade, while also indicated to cause the death of more than 7000 lives in the year 2021 [5]. Non-Melanoma Skin Cancer is estimated to be blamed for more than 5400 deaths every month globally.

The occurrence of skin cancer is determined by the abnormal growth of cells into cancerous cells in the epidermis layer of the skin. There are various types of skin cancer depending on which layer of the epidermis is affected during the mutation. Typically, the cause of this abnormal skin growth can be blamed upon the mutation of DNA, underexposure to ultraviolet (UV) rays or ionizing radiations, and even immunosuppression. With an intent to categorize, the types of skin cancer are divided into Recurrent basal cell carcinoma (BCC), Squamous cell carcinoma (SCC), Melanoma, Merkel cell carcinoma, and other rare skin cancers, including Kaposi sarcoma, Keratoacanthoma, and cutaneous T-cell lymphoma [6]. Some of them are depicted in Figure 1. Basal cell carcinoma and Squamous cell carcinoma are estimated to be diagnosed in over 3.6 million cases and 1.8 million cases each year in the US [3]. Organ transplant patients are evaluated to be at a 100-fold higher risk of developing some form of skin cancer [7]. Small, itchy, rough-looking patches called Actinic Keratosis appearing as red or brown are among the common precancers and have been known to affect over 58 million lives in the US [8]. The medical treatment expenses projected for treating skin cancer in America are more than 8.1 billion USD yearly (4.8 billion USD for Non-Melanoma Skin Cancer treatment & 3.3 billion USD for Melanoma Skin Cancer treatment) [9].

Moreover, clinics around the globe have reported an increase in the cases of skin cancers around the world due to the rise in the penetration of solar ultraviolet rays caused by the depleted ozone layer. The Caucasian population or even population with Fitzpatrick skin type I-III, or people with blonde or red hair are more inclined to develop this kind of mutated growth, as pigmentation of skin plays a crucial part in the pathway of cancerous growth. Exposure to UV rays from the sun is attributed to cause about 90% of cases of Non-Melanoma Skin Cancer and about 86 % of Melanoma Skin Cancer cases. Research has indicated that when melanin pigment forming cells called melanocytes initiate pell-mell multiplying, it yields in the development of malignant tumors in the skin. Researchers

have claimed that the risk of developing melanoma in the average human body doubles if they have encountered more than five sunburns. In some cases, even one prolonged sunburn in their life has been known to double the risk of developing malignancy late in their life [10], [11]. Severe exposure to indoor tanning beds is also attributed as the cause of the increase in skin cancer cases, especially among women in the age group 45 or younger [12], [13]. Typically, the characteristics exhibited by such melanoma lesions include asymmetry in shape, irregularity around the borders, colorful features, diameter of more than 6 mm, and constantly enlarging tumor (depicted as an acronym ABCDE) [14]. Amongst which, color and structure are the crucial factors for diagnosing melanoma. The list of features that are taken into examination while performing diagnosis are enlisted below:

- Distribution of Pigment along with lesion's heterogeneity or homogeneity
- Asymmetry or Symmetry in the shape of the lesion
- Presence of keratin on the surface
- The intensity of oddity in vascular morphology
- Presence of Ulceration and burning sensation
- Irregularity of lesion border
- Colour: white, grey, black, yellow, or even brown

Figure 1: Basal Cell Carcinoma (top-left), Squamous Cell Carcinoma (top-right), Melanoma (middle), Kaposi Sarcoma (bottom-left), Skin Lymphoma (bottom-right) [15]

Among various types of skin cancer, melanoma skin cancer is notoriously lethal, as it has been assessed to have an overall five-year survival rate fatally depending on the stage at which the cancer is diagnosed, ranging from almost 100% when detected at an initial stage, while less than 20% if diagnosed too late for a proper treatment cycle. The disease's survival rate is also known to degrade to about 66 % if the tumor reaches the lymph nodes and can eventually fall to only 27 % if cancer metastasizes to other organs. The fact that there is a drastic drop in survival rate with the function of timely detection of melanoma highlights the crucial need for tools that can enable early diagnosis while also actively reducing its incidence rate [16]. Contrary to the misconception that melanoma can be primarily found in pre-existing moles, in a medical setting, only 20-30% of melanoma were diagnosed in existing moles, while about 70-80% arise from mutation in normal skin [17]. Individuals with a family history of any skin cancer variant, be it melanoma or non-melanoma, are more prone to develop lethal melanoma than those without any genetic history [18]. The melanoma survivors are at heightened risk of nine times more likely to develop melanoma than other individuals, as research has indicated cancer’s substantial recurrence possibility [19].

The graphical data, as showcased in Figure 2 and Figure 3, elucidate the risk factors of developing melanoma, involving individuals' characteristics, including their age, sex, and race [20]. Although the incidence rate of developing skin cancer among the non-Hispanic white population is much higher than in Black or Asian communities [3], the research strongly indicates the survival rate of patients with skin of color is much less than the Caucasian population [21]. This is because it’s harder to diagnose skin cancer among African American individuals, as it is prone to develop in areas that aren’t exposed to the sun, for example, on palms, soles, groin, or inner section of the mouth. Once it is detected adequately at its later stages, the study has shown that in around 25 % of cases, until then, melanoma has already spread to lymph nodes and even other organs [21], [22].

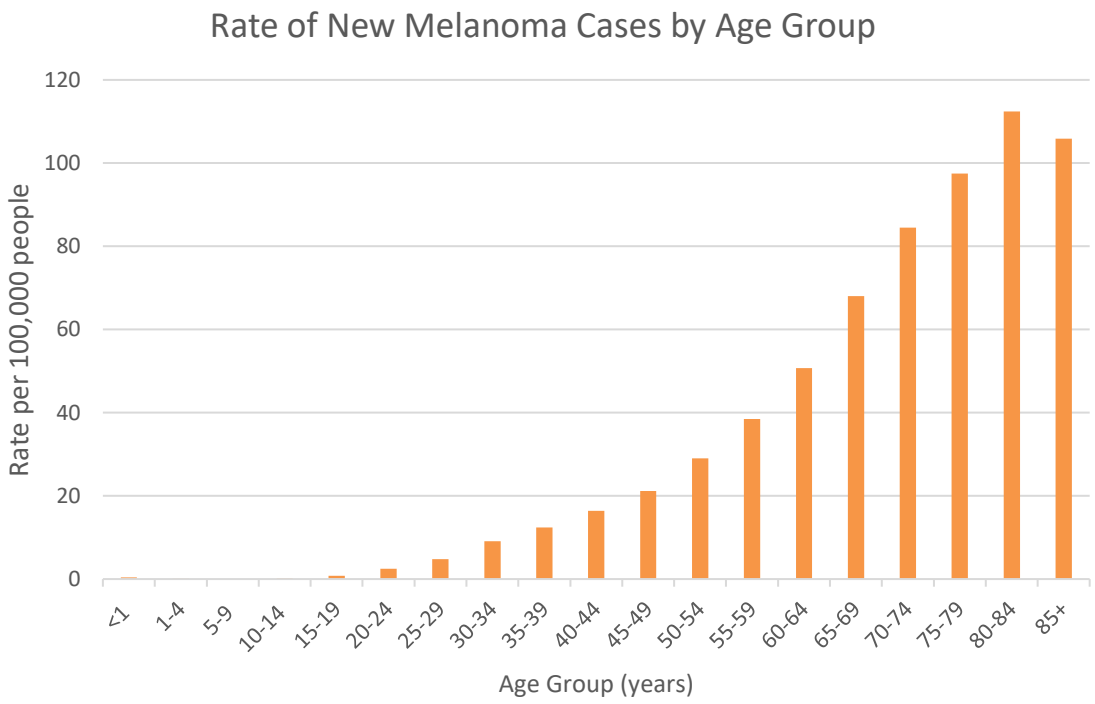

Figure 2: Bar Chart depicting the rate of new Melanoma cases by age groups [20]

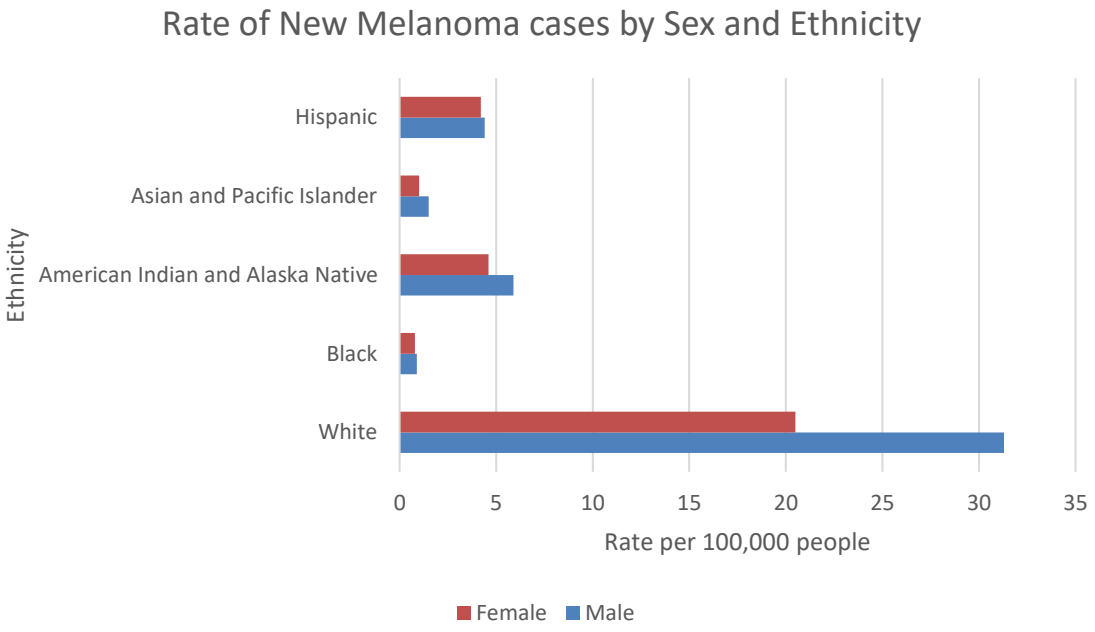

Figure 3: Graph of the rate of new Melanoma cases by sex and ethnicity [20]

Earlier studies have constantly affirmed that almost 50 % of melanoma tumors can be self-detected by visual inspection [23], [24]. American Academy of Dermatology Association has been vocal about the importance of self-examinations, usually once a month, for individuals who have either personal or family history of cancer [25]. But this method is proven to be not always reliable, as malignant ones can closely resemble benign lesions until the very later stage. Histopathology has proven to deliver the highest accuracy and precision for diagnosing skin cancer [26]. But taking a biopsy of a patient's skin comes with its difficulties. It is an expensive, invasive procedure and can cause infections [27] if not taken in a safe clinical environment. To overcome these issues, alternatives such as non-invasive techniques [28]–[30] have been developed to prevent foreseeable biopsy complications. Dermoscopy is one the most prevalent non-invasive procedure for facilitating skin lesion diagnosis cause of its effectiveness. Dermoscopy yields high-resolution imaging incorporating precisely applied pigmentation to gather information through visual inspection from multiple layers of skin by eliminating any reflections from the skin surface [31]. Medical research has shown that dermoscopy has achieved high diagnostic accuracy for a skin lesion compared to standard photography procedures [32]. Dermoscopy enables dermatologists to analyze the morphological features of the presence of tumors, which are not perceptible for bare eyes to inspect. A couple of techniques performed to enhance such morphological characteristics include Epiluminescence microscopy (ELM), Cross polarization Epiluminescence (XLM), side transillumination, and solar scans [33]–[37]. These techniques have assisted dermatologists by providing more intricate diagnostic criteria and further improving the diagnostic accuracy by 10-30% [38].

A typical procedure of dermoscopy starts with a skilled dermatologist observing a suspected lesion, then only performing a biopsy if required. This procedure could be accounted for consuming a lot of time, leading to the eventual advancement of fatal tumor growth in the patient. Even after that, studies have suggested that the diagnostic accuracy of most dermatologists can be even less than 80%, taking into consideration their skill and the clinical environment [39]. This leads to a much big-picture issue affecting the skin cancer crisis, like the scarcity of highly skilled dermatologists who are willing to work for the public healthcare system.

Attributing to all the problems discussed before, the intuition of this research leads to developing an automated or computer-aided diagnostic model that could not only provide assistance to dermatologists for making their decision quicker while improving diagnostic accuracy but also prove accessible diagnosis to the clinical environment that doesn't have the expense to involve such highly skilled dermatologist, at such an early stage of diagnosis.

1.3 Challenges Addressed

This research address challenges that fellow researchers frequently neglect. Even though some challenges are tackled, the research lacks a direction in which multiple problems are tackled efficiently. One of the most complex challenges that have become the force of hindrance to applying deep learning techniques in the field of medical imaging is the problem of dataset imbalance. It has been statistically proven that when the learning algorithm has been introduced with a skewed dataset, the algorithm is ineffective in performing for a class present in the minority [40]. This can also be concluded using the fact that improving the loss in a smaller set class has been proven more challenging than improving the loss value in the majority class. This challenge could be tackled using two techniques; the first includes balancing the ratio of training data by changing the overall proportion of categories, which could be attained by increasing the sample set in the weaker classes. In contrast, the second method involves designing and formulating the algorithm so that training is performed in a much-balanced fashion by re-weighting the losses that pertain to each class or even samples [41], [42].

Researchers have shown that defining the exact penalty cost for the majority and minority classes in domains such as medical diagnosis can be challenging [43]. This research proposes to develop a method that combines both by including a generative model, not only to increase the overall sample size through augmentation but also to improve the feature resolution of the images provided to the learning algorithm and proposes a pipeline along with a custom loss function explicitly designed to tackle imbalance in the dataset. Previous studies have shown high overhead while performing inferences to yield generative samples on the distinctly low-end graphic processing unit. This issue is also

addressed in this research while simultaneously optimizing the memory constraints and the time complexity.

1.4 Objective

In this thesis, a deep learning architecture pipeline called SkinCAN AI is proposed to provide an early diagnostic tool for the dermatologist to detect and segment skin cancer and improve overall diagnostic accuracy compared to other existing techniques. The proposed model involves generating synthetic samples for minority classes using SkinGAN architecture that is closely inspired by stylegan2-ADA [44] and develops its custom loss function and mini-batch logic on top of it for the classifier model. The classifier model is later integrated with an attention module that could help the learning algorithm focus on essential features in the lesion dermoscopic images. The proposed pipeline also includes an integrated module for data-specific manipulation tasks like removing hair, ink marks, or scale marking artifacts from the lesion images or even increasing the image's resolution without losing critical features. The whole pipeline is trained end-to-end to maximize results in diagnostic evaluation metrics while keeping in mind the pipeline's scalability and time complexity. The experiments conducted have shown that the proposed architectural pipeline can outperform the existing methods in skin lesion detection by deploying soft attention.

The pipeline presented in the thesis is developed, considering the feasibility of deploying them in a fast-paced clinical environment with a much lower computing requirement. This research also addresses the impact of such an artificially intelligent model where the medical cancer research is lagging in improving the health care diagnosis for communities like people of color and trans people.

1.5 Structure of the thesis

The rest of the thesis is cataloged as follows: Chapter 2 provides a brief on relevant studies conducted for skin lesion diagnosis and standardized datasets deployed for benchmarking in skin cancer classification. Subsequently, Chapter 3 elaborates on a survey of related deep

learning generative algorithms. Chapter 4 elucidates the proposed pipeline and its modules and relevant discussion on hyperparameter selection for optimal training. Experiments accompanying results, comparative analysis of the proposed pipeline against existing algorithms, and in-depth ablation studies are described in Chapter 5. The thesis explores the significance and impact of the proposed method and depicts the explainability of this AI model in Chapter 6. Finally, concluding remarks are pointed out in Chapter 7.

CHAPTER 2
LITERATURE SURVEY

Humans often underestimate their ability to observe and understand the physical world-encompassing their lives, including how objects are visualized or how they interact. Although the advances in technology have made it possible to convert this enormous treasure trove of information about the physical world into the form of digital bits, the research community is still haunted by the arduous task of training or formulating algorithms that could understand the intricacy of the features present in these datasets. Deep learning practitioners have recently discovered that generative models can show promising results in creating data with close feature similarity with the physical real-world data.

With the growth of technology, the research community has ever since tried to integrate technology or computer-based aid into the medical domain so that high-level effectiveness could be achieved and earlier diagnosis could be made feasible. Usually, a computer-aided diagnosis (CADx) system comprises an algorithm, be it machine learning or deep learning behind the scenes trying to facilitate experts in the field to yield a precise diagnosis of a patient's current health condition. These high processing computational power-based techniques can effortlessly aggregate multiple data streams and provide an opportunity to develop robust information enriched diagnostic systems spanning from radiographic images to patients' historical data, pathology of cancer, genomics, and lastly, even involving the social network.

Under the hood of cancer imaging, AI has assisted in multiple clinical procedures by transforming the image interpretation methodology. Detection is the first step that can be elaborated with the localization of the object of interest in tumor imaging. This subfield developed is called computer-aided detection (CADe) [45]. Characterization is the second step involving diagnosis, segmentation analysis, recognizing the precise stage at which the tumor growth has already developed, and finally, prognostication to predict treatment trajectory correctly. The last step in which AI could be deployed is monitoring, in which pinpoint features are observed temporally to register the efficacy of current treatment on

the patient's health [46]. This thesis focuses on implementing AI in the detection and characterization step.

2.1 Datasets

Computer-aided tools require a knowledge bank of a dataset to learn the discrete features of the task at hand. While evaluating the performance of these diagnostic tools and ensuring that optimum and diverse information is distilled in the network, a feature-rich dataset plays a significant role. Cancer datasets have historically suffered not only in their quantitative abundance but also in their feature diversity and have eventually hampered the integration of artificial networks in the field. As it is impossible to increase the data size, AI networks must adapt themselves for few-shot learning or generate synthetic data to train and learn the nuances of diverse information about the tumor. Table 1 summarizes essential datasets that have led the development of AI networks from their neonatal stage. This thesis will mainly focus on architectures trained on the ISIC archive; therefore, they will be described in length moving forward in the section.

Table 1: Popular Skin Cancer Datasets

Name of Dataset	Year of Release and Updates	No. of Images
HAM10000 [18]	2018	10,015
PH^2 [47]	2013	200
ISIC archive [48]	2016-2020	25331
DermQuest [49]	1999	22082
DermIS [50]	-	6588
AtlasDerm [51]	2000	1024
Dermnet [52]	1998	23000

2.1.1 HAM10000

This publicly available dataset named “human-against-machine” of skin lesions contains 10,015 dermoscopy images cataloged from two sources, namely Cliff Rosendahl’s skin cancer practice in Queensland, Australia, and the Dermatology Department of the Medical University of Vienna, Austria. This dataset was initially just photographic scans of lesions compiled for 20 years, which was later digitalized with the help of a Nikon scanner and then converted to an 8-bit JPEG image. Lastly, the photos were resized to 800 x 600 pixels at 72 DPI. The dataset tries to tackle the issue of diversity by applying variegated acquisition functions and cleaning methods while collecting for eight different categories, as depicted in Figure 4. The datasets' acquisition devices vary from MoleMax HD, DermLite Foto (3Gen) camera, DermLite Fluid, DermLite DL3, and analog cameras [53]. Table 2 summarizes the HAM10000 dataset along with its subcategories. Figure 4 provides a glimpse inside the HAM10000 dataset.

Table 2: HAM10000 dataset [18]

Type of Skin lesion	No. of Images
Actinic Keratoses	327
Basal cell carcinoma	514
Benign Keratoses	1099
Dermatofibromas	115
Melanocytic Nevi	1113
Melanomas	6705
Vascular Skin Lesion	142

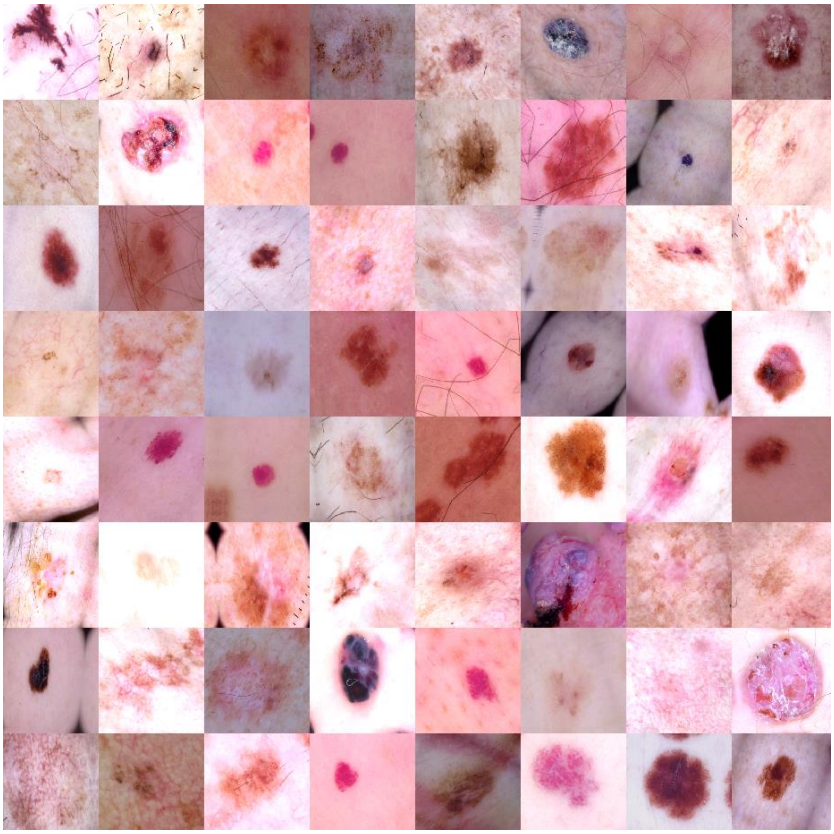

Figure 4: Images from HAM10000 [18]

2.1.2 ISIC Archive

International Skin Imaging Collaboration (ISIC) made this dataset publicly available at International Symposium on Biomedical Imaging (ISBI) challenge 2016. The original older dataset contained a much lesser number of samples, making it challenging to assist any AI networks in learning crucial diverse features from it. The ISIC has been known for increasing the size of its archive every year by expanding its categorical scope and its quality metadata. Table 3 explains the evolution of the ISIC dataset every year and its details. The recent ISIC datasets (depicted in Figure 5) have the most categorized classes among all the previous year's datasets while also including the metadata of patients such as demographics, age, and gender. The equipment used for acquisition in ISIC dataset compilation consists of the MoleMax HD dermatoscopy system. In contrast, some even used the Dermoscopic attachment connected to a digital single reflex lens camera system.

ISIC data archive contains not only the JPEG version of the lesion but also a feature-rich DICOM version.

Table 3: Summary of ISIC datasets through years [18], [54], [55]

ISIC 2016 [56]			
Classes of tumor	Training set		Testing set
Melanomas (MEL)	273		115
Benign Nevi.	627		264
Total	900		379
ISIC 2017 [57]			
Classes of tumor	Training set	Validation set	Testing set
Melanomas (MEL)	374	30	117
Seborrheic-Keratoses (SK)	254	42	90
Benign Nevi.	1372	78	393
Total	2000	150	600
ISIC 2018 [18], [58]			
Classes of tumor	Training set	Validation set	Testing set
Actinic Keratoses (AK)	327		
Basal Cell Carcinoma (BCC)	514		
Benign Keratoses (BKL)	1099		
Dermatofibroma (DF)	115		
Melanocytic Nevi. (NV)	6705		
Melanoma (MEL)	1113		
Vascular Skin Lesion (VL)	142		
Total	12,594	100	1000
ISIC 2019 [18], [54], [55]			
Actinic Keratoses (AK)	867		
Basal Cell Carcinoma (BCC)	3323		
Benign Keratoses (BKL)	2624		
Dermatofibroma (DF)	239		

Melanocytic Nevi. (NV)	12,875
Melanoma (MEL)	4522
Vascular Skin Lesion (VL)	253
Squamous Cell Carcinoma (SCC)	628
Total	25,331

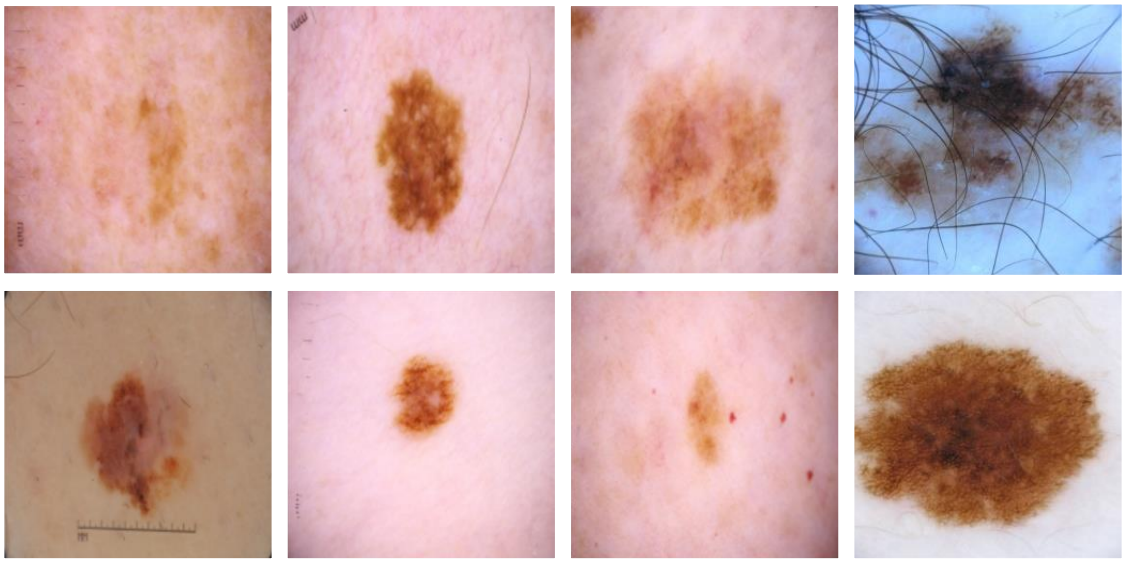

Figure 5: ISIC 2019 [18], [54], [55] dataset samples (bottom row examples are benign & top row examples are malignant)

2.2 Recent Trends of Deep Learning Techniques deployed for Skin lesion diagnosis

Deep neural networks were formulated to mimic the human brain's functionality by adopting its structure of interconnected neural nodes while using the power of computational acceleration provided by modern GPUs. Typically, these neural networks are trained on a massive amount of data, and knowledge is distilled into their weighted structure as they become experts in their field. The research community has deployed these deep neural networks to detect the presence of malignancy and distinguish among various types of skin lesions.

2.2.1 Artificial Neural Network-based techniques

ANN utilizes backpropagation to learn underlying intricate mathematical relationships and distill the information into its layers according to the pattern sequence observed in the training dataset. A set of extracted features are provided to ANN, which is tasked with classifying whether the feature containing the image is malignant or benign according to the visual patterns observed by the algorithm during training.

Xie et al. [59] proposed a pipeline of a classification system, which initially extracts lesions from a dataset using a self-generating neural network, and later extracts features like border, the texture of lesion, and pigmentation of the tumor. The precisely selected 57 features were reduced dimensionally using Principal Component Analysis (PCA) in the final step. They were deployed to train an ensemble of neural networks combining backpropagation neural networks and fuzzy neural network variants. This model achieved 91.11% accuracy and a performance boost of 7.5% in sensitivity compared to other existing classifiers.

Cueva et al. [60] proposed a model that extracts features based on the ABCDE rule using threshold parameters for pigmentation and diameter (>6mm malignant) while deploying Mumford-Shah Algorithm and Harris-Stephen Algorithm for evaluating lesion's asymmetry and border. This proposed model achieved 97.51% accuracy on a dataset of 31 images. Some of the methods submitted by the community even involved using a gray-level co-occurrence matrix (GLCM) to extract detailed features of the lesion [61], [62]. During the infancy stage of the computer-aided techniques, various machine learning algorithms were formulated for the diagnosis of skin lesions that includes k-nearest neighbors (kNN) [63], support vector machines (SVM) [64], random forest (RF) [65], and self-organizing maps [66].

However, it became evident that these algorithms cannot learn high-level features due to their dependency on pixel intensity space and the hand-crafted features, thus inculcating a priori of knowledge during prediction [67], [68]. Figure 6 below summarizes the flow of training an ANN for skin lesion detection.

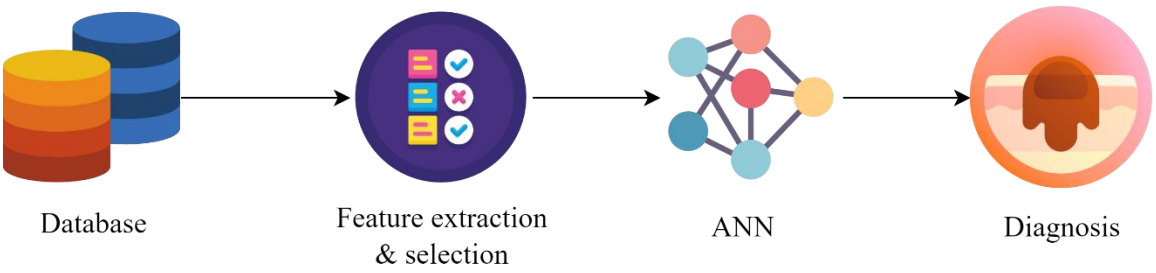

Figure 6: ANN pipeline for diagnosis

2.2.2 Convolutional Neural Network (CNN) based techniques

A convolutional neural network streamlines multiple convolutional layers, followed by non-linear pooling layers, which finally connect to a set of fully connected layers. CNN was made feasible to deploy for skin tumor diagnosis due to the advent of easy access to computational processing power and the increasing availability of data enriched with features to learn. To overcome the issues presented by hand-crafted features in classical machine learning, researchers have formulated CNNs and their variants like ResNet, Efficient Nets, NFNets, and Capsule networks which achieved significantly higher results in medical applications like detection, segmentation, and classification. These networks are implemented with a learning objective and loss function to discover and learn features for the given task [69]. Historically, in traditional computer-aided diagnosis, researchers have tried deploying hand-crafted image processing filters that yield features describing the characteristics of the cancerous tumor. But methods like the Harris Corner detector algorithm for detecting just the edge or corner in the image are memory-intensive and time-consuming, thereby causing more issues during training for such time-sensitive tasks.

Statistics have shown that less than 20% of patients are diagnosed with melanoma after a biopsy in the clinical environment. There are often cases where patients have opted out from even going for a biopsy [70]. This situation creates a difficult choice for health experts to either let the lesion progress without biopsy or get the patient to consider having a biopsy, which might become a burden to the patient and the hospital care services. The two strategies that academicians have come up with to solve this issue of limited data are either to perform data augmentation or perform transfer learning. The transfer learning strategy

has shown some great results, but the question of knowledge transfer of features observed between the non-medical and medical fields is still questionable [71]. This evidence has led the thesis to focus on developing new data augmentation techniques or even few-shot learning-based algorithms.

Lequan et al. [72] proposed a fully convolutional residual network comprising 16 residual blocks and took average results using both SVM and softmax classifier to yield 85.5% accuracy with segmentation and 82.8% without segmentation. Another study [73] deployed Inception v-3 architecture pre-trained on ImageNet and then using transfer learning. The architecture or model is trained on a standard dataset with abundant features to learn from in transfer learning. Then in the final step of training, only the last few layers are modified and learned on the target feature distribution. These modified layers were specifically fine-tuned to learn application-specific features. The study fine-tuned the pre-trained network on two different resolutions of skin lesion images: 1) coarser scale to learn contextual and geometric features of the dermoscopic image, 2) finer scale to learn the textual information about lesion.

A study from 2018 [74] formulated a Resnet-152 pre-trained and then finetuned to classify 12 classes of skin lesions which yielded a performance metric AUC score of 0.99. The dataset of 3797 lesions was used for training the network, and augmentations like scale transformations were also applied to create a robust algorithm. Dorj et al. [75] deployed AlexNet to extract features and then concatenated the pipeline to SVM, which served as a classifier trained to classify four variants of skin lesions. This study showed 95.1% sensitivity, 98.9% specificity, and 94.17% accuracy. A couple of studies [76], [77] also deployed ensemble learning for training deep CNN. In this technique, multiple versions of the same network are trained, and then averages of weights are taken to create a new model weight. Perez et al. [78] evaluated the utility of implementing 13 different data augmentation techniques for lesion classification while performing analysis on Inception-v4, ResNet, and DenseNet.

Mahbod et al. [79] formulated a pipeline for skin lesion classification by extracting deep learning features from pretrained complex convolutional networks and then fusing a machine learning classifier trained on those features to achieve the best performance. The

study's authors deployed AlexNet, ResNet-18, and VGG16 to generate information-rich features for their model. Later a multi-class classifier such as SVM and softmax is implemented to evaluate and yield 97.55% and 83.83% area under the curve (AUC) performance, respectively, for both classifiers, on lesion classification using the ISIC dataset.

Sagar et al. [80] deployed transfer learning on ResNet-50 architecture to classify melanoma and non-melanoma dataset of 3600 lesion ISIC dataset. This model achieved an accuracy of 93.5%, precision of 94%, and F1 score of 85%, which was better than all the popular models of CNNs at that time, including InceptionV3, DenseNet169, and MobileNet. Polap et al. [81] designed an AI model consisting of a cascade of convolutional layers for image feature extraction and a genetic algorithm for calculating the likelihood of a specific sample classified in a particular class. This method achieved 7.5% better results than transfer-learned algorithms. Ahmad et al. [82] formulated a triplet loss function and deployed that along with ResNet152 and Inception ResNetv2 models to achieve significant accuracy results. Adegun et al. [83] implemented a fully convolutional network with long and short-cut skip connections, designed to learn coarse features such as appearance and fine details. They achieved 98% accuracy, 98.5% recall, and 99% AUC score on the HAM10000 dataset when integrating their model with the Conditional Random Field (CRF) module to perform contour refinement and boundary localization of tumors using Gaussian Kernels. Al-Masni et al. [84] proposed a new approach where a full-resolution convolutional network is trained for the segmentation task of the lesion. Then segmented features are fed to a deep learning architecture to classify the lesion.

Another recent study [85] proposed a region-based CNN with ResNet152 trained on 2742 dermoscopic images (ISIC dataset). In the study, a mask and RCNN (region-based CNN) yielded interesting regions (ROIs) that were later fed to ResNet152 for classification. This model architecture achieved an accuracy of 90.4%, sensitivity of 82%, and specificity of 92.5%. Alzubaidi et al. [86] designed a CNN architecture that is pretrained on massive unlabeled medical data and then transferred learned to small labeled medical data, providing the network with better and faster-pretrained features augmentation. Their methodology achieved an F1 score of 98.53% and an accuracy of 97.51%. Liu et al. [87]

implemented a unique architecture of a convolutional neural network built in a multiscale ensemble fashion comprising three branches. Initially, the lesion area is selected by picking the maximum number of pixels belonging to the skin tumor. Then the site of interest is optimized using the model, while at the last stage, two scales were picked for input to other branches. Iqbal et al. [88] carefully designed a deep convolutional network with multiple specialized layers and filter sizes to provide maximum optimization for skin lesion classification tasks. Some studies in 2021 experimented with how data could be manipulated to make it better suited for training computer-aided skin lesion diagnosis tasks. This included Ali et al. [91], which investigated techniques like image normalization and noise artifacts removal followed by data augmentation. In contrast, other studies experimented by removing hair artifacts using specialized techniques like Dullrazor technology. Figure 7 shows various ideas and intuition adopted for applying CNN architecture in skin lesion diagnosis tasks.

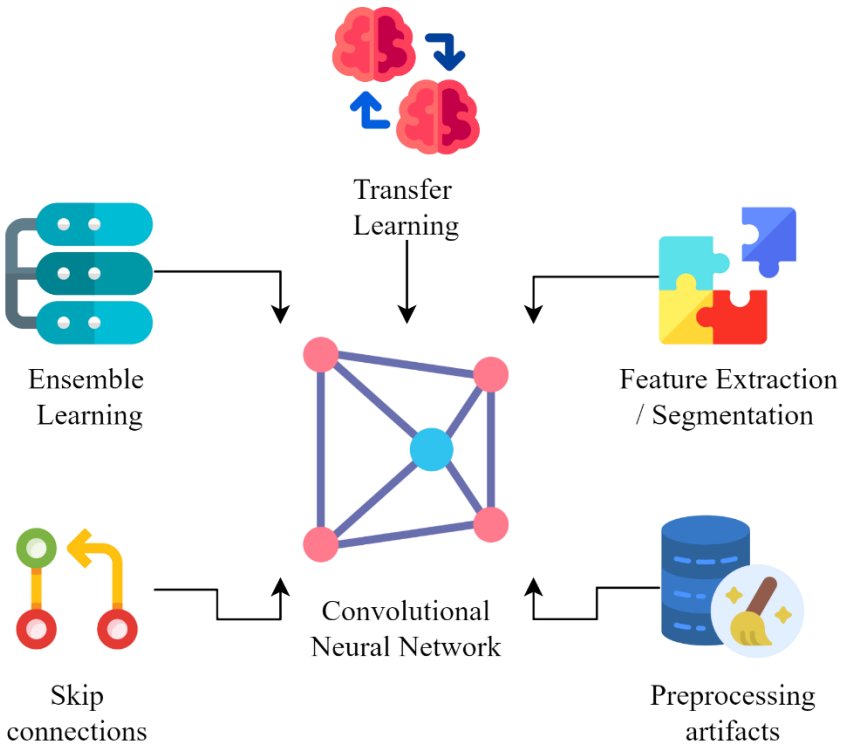

Figure 7: CNN pipeline for skin lesion diagnosis

2.2.3 Generative Adversarial Network (GAN) based techniques

Due to the limitation presented while training a CNN model with a skin lesion dataset, researchers started exploring ways to synthesize data to suffice the need for enormous data for training. Yi et al. [89] designed a generative adversarial network that worked on using the Wasserstein distance for learning optimization. This GAN would create 64 x 64 size samples of a particular data class. The model achieved decent precision with only 140 ISIC 2016 data samples. Baur et al. [90] proposed a deeply discriminated GAN which could generate high-resolution images up to the size of 256 x 256 and performed a comparative analysis between their model and DCGAN [91] & LAPGAN [92]. Another study by authors [93] investigated generating samples of skin lesions of dimensions 1024 x 1024 using progressive GAN (PGGAN) [94] and evaluated them using the Visual Turing Test and Sliced Wasserstein Distance.

Bissoto et al. [95] deployed pix2pixHD GAN [96] model to synthesize images from semantic maps and instance maps using annotated dataset. Rashid et al. [97] implemented a GAN model, where the discriminator later serves to become a trained classifier for classifying skin tumors. The study compared their model and fine-tuned versions of DenseNet and ResNet on the ISIC 2018 dataset and aftermath, concluded that their model achieves significant accuracy and performance gains.

Qin et al. [98] analyzed their model of the generative adversarial network, consisting of a generator and discriminator that was programmed to generate high-resolution images without any noisy artifacts observed in the original dataset, and later evaluation of the synthesized samples was performed to achieve maximum optimal classification results. Recently, Ding et al. [99] deployed a combination of segmentation masks and labels to generate a mapping of pathological markers of interest, and they later also utilized a novel technique of translating images to a numerical representation of matrix labels using conditional GAN (CGAN), thereby combining the shallow and deep features of lesion images. The study performed their analysis against several other GANs and found that their AUC values were significantly better for ISIC 2017 dataset.

Ahmad et al. [100] deployed a GAN that, instead of sampling noise from random noise Gaussian distribution, their model was trained by sampling noise vector from heavy-tailed student t-distribution. The authors showed that their model consisting of one VAE, two GANs, and one auxiliary classifier achieved an accuracy of 92.5%, attributing the success of their model to the increased diversity range of synthetic samples. The survey concludes that GANs have shown promising results in synthesizing realistic-looking data samples. The medical field needs such algorithms, so integrating data-hungry AI algorithms into the clinical diagnostic pipeline can be facilitated while maintaining the pace for early diagnosis of the lesion.

CHAPTER 3
EVOLUTION OF GENERATIVE NETWORKS

The immense potential of the Generative Adversarial Network to learn synthesizing images from random noise vector mapping has grabbed the attention of the deep learning community, and there has been an accelerated growth in the number of applications where GANs are deployed to achieve optimized results. The list of various domains where GANs are applied includes:

- Computer Vision task [101]–[104]
- Segmentation [105]–[107]
- Time Series prediction [108]–[110]
- Medical domain [111]–[114]
- Speech & Language processing [115]–[117]

While researching this thesis, an in-depth research analysis was performed to understand the underlying mathematical representation and architecture of existing generative models, which is elucidated at length in the following section.

3.1 Generative Models in Deep Learning

The spectacular growth in deep learning models deployed or integrated to achieve success in field-specific applications has inspired researchers to develop an algorithm that can synthesize data, which has replicated features of real-world information, later known as generative models. A couple of initial attempts at synthesizing images include the GLOW algorithm (flow-based generative model) [118] and even Variational Auto-encoder (VAE) [119]. Still, none of them were prominently successful in generating feature-rich images. The idea of generative adversarial models (GAN) was introduced by Ian Goodfellow [120] in 2014. The pioneering researchers proposed integrating a min-max two-player hostile game in an adversarial model that competes with itself to yield results. Recently, another type of generative model called diffusion models [121] has surprised the deep learning community by showing excellent results in a few generative tasks. An overview of the

different types of existing generative models predominantly in image synthesis applications is depicted in Figure 8, along with their brief functionality and limitations.

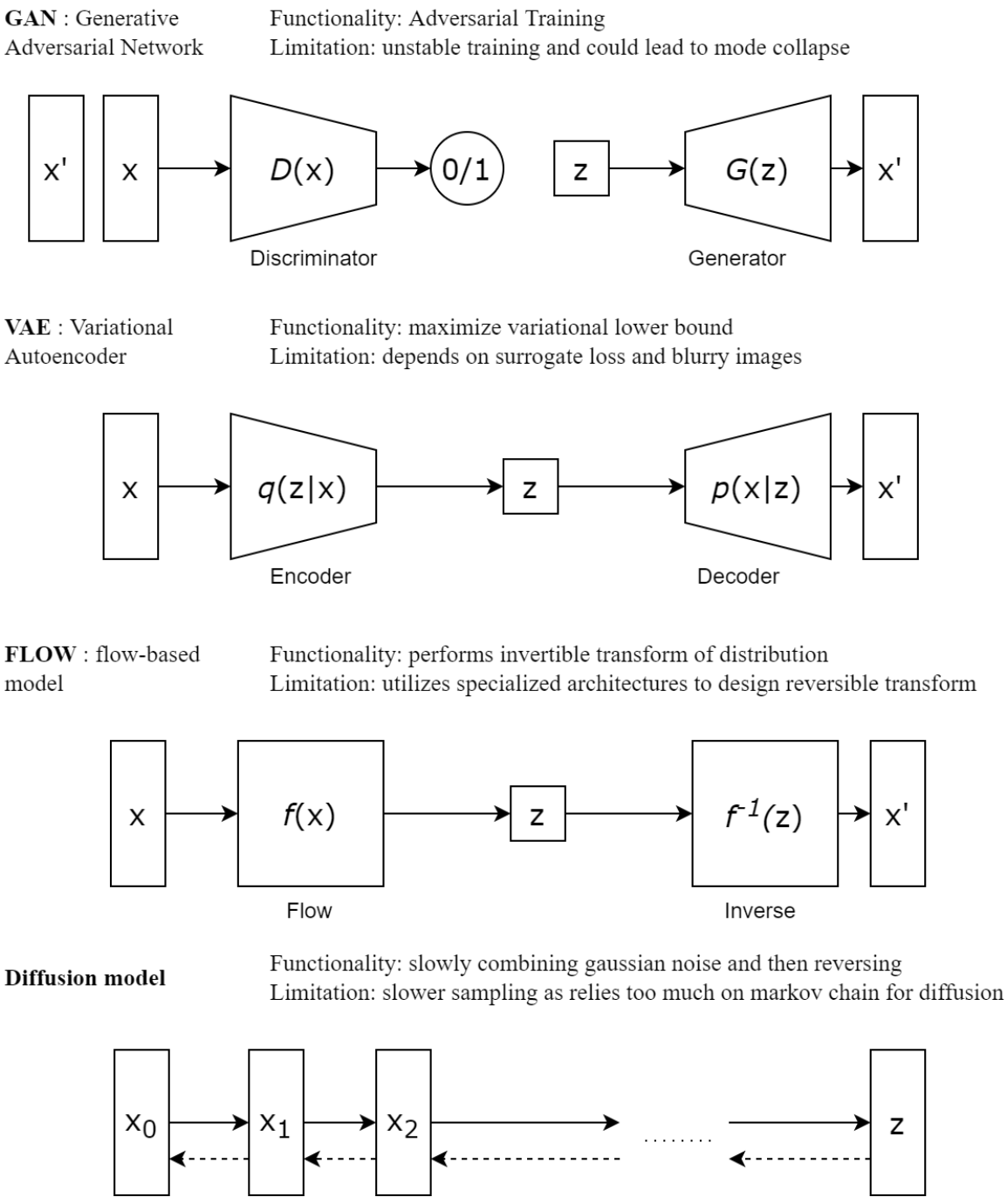

Figure 8: Types of Generative models in deep learning [122]

Generative algorithms can be categorized into three main modeling techniques depending on how they depict their probability distribution: 1) Likelihood-based models: these algorithms understand their probability function of distribution using the approximate or maximum likelihood. These models include flow models [123], [124], energy-based models [125], [126], variational autoencoders (VAE) [119], [127] and autoregressive models [128]–[130]. 2) Implicit generative models [131]: These algorithms implicitly describe their probability distribution using their sampling process. GANs are an example of implicit generative models. 3) Score-based models [132]: these models reverse the stochastic process of diffusing data into noise to generate samples, which is learned by minimizing score-based matching losses without requiring any adversarial optimization like that in GANs. Diffusion models are examples of the score-based generative model. From this point, the thesis focuses exclusively on the Generative Adversarial Network. The initial research survey concluded that GANs are far more superior in generating superior results than other generative models in current times.

3.2 Generative Adversarial Networks (GAN)

3.2.1 Network Architecture

A standard architecture of GAN (depicted in Figure 9) consists of two modules, namely the generator and discriminator. The discriminator (D) of the model consists of downsampling classifier layers, which take an input (real or fake sample) and yield a binary classification with a prediction value between 0 and 1. The generator (G) module comprises an upsampling path, where a fixed-length random noise is fed as input, and output of much higher resolution in the form of visual representation is expected. Vanilla GAN architecture is made from a stack of feed-forward neural network layers for both generator and discriminator modules [120].

After learning the generative process, the generator can map corresponding points from multidimensional noise vector space to unique feature vectors or visual entities in the target domain, thereby learning to form a compressed representation of problem data distribution. This compact vector space is defined as latent space, and the vector's dimensions are latent

variables. Each of these latent variables represents unique visual feature information. A batch of points can be drawn from this latent space and fed to a learned generator model to generate new fake samples.

The operation of the discriminator is like that of a classifier network; it takes information as input and transforms it into features. This extracted feature becomes the deciding factor on which classification task is performed at the final layer.

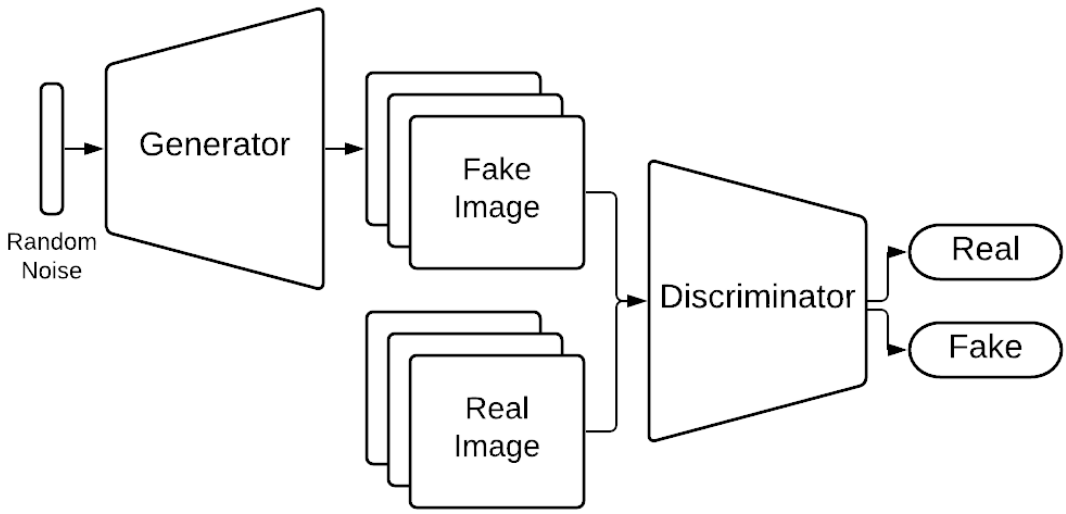

Figure 9: Vanilla Generative Adversarial Network Architecture [120]

3.2.2 Design Ideology

Both models (Generator and Discriminator) are targeted to learn simultaneously while achieving the optimization goal of Nash Equilibrium using the min-max game process. The individual objective functions of both segments of the model can be elucidated as follows:

1. Generator G aims to generate samples from random noise while fooling the discriminator into believing that the generated image samples belong to the actual dataset.

2. The discriminator D is designed to detect whether the image presented as input is either real or a fake one generated by the generator, simultaneously also providing feedback to the learning curve of the generator.

This training aims to optimize the model weights of both G & D networks while achieving a level of generalization ability and higher fidelity in generated images. The ideal state of the model after stable training would be such that the discriminator D identifies the realness of the generated samples of generator G with a probability of about 50%.

For accommodating the min-max game between generator G and discriminator D, a loss function can be described in the form of Binary Cross-Entropy loss. The loss function is formulated as follows:

$$J(\theta) = -\left(\frac{1}{m}\right)\sum_{i=1}^{m}\left[y^{(i)}\, log\left(h\left(x^{(i)},\theta\right)\right) + \left(1-y^{(i)}\right) log\left(1-h\left(x^{(i)},\theta\right)\right)\right] \qquad (III-1)$$

$$\min_G max_D\ J(D,G) = E_{x\sim \text{Pdata}\,(x)}[log(D(x))] + E_{z\sim Pz(z)}[log\ (1-D(G(z)))] \qquad (III-2)$$

Where x is input data, m is the number of samples in a mini-batch, θ_G is the generator model weights, θ_D is discriminator model weights, $J(\theta)$ is criterion or loss function with parameters, z is the dimensionality of latent noise vector, P_{data} is training data distribution and P_z is synthetic data distribution generated using a noise vector.

The equation $(III-1)$ describes the Binary Cross Entropy loss (BCE), and $(III-2)$ shows the transformed version of BCE loss for GANs. As seen in the equation, the loss function can be seen as two segments, one-part deals with original labels, and the second part deals with samples with fake labels. As depicted, the discriminator D is trying to maximize this loss function. In contrast, the generator G tries to minimize this function by fooling the discriminator D into yielding a value closer to one, making the second term in the equation approach negative infinity.

A subtle variation is introduced to the loss function. The original loss function tends to make GAN stuck early in training, especially when the discriminator has an easier task than the generator. Instead of the generator G trying to minimize the value of $log\left(1 - D\big(G(z)\big)\right)$ when target labels are zero, G is defined to maximize the value of $log\left(D\big(G(x)\big)\right)$ while making the target label as 1. This modification is called the non-saturating loss function. It eradicates the saturation region from the generator loss function in the early stage of training when high gradient values are much needed.

Optimizer is the function devised to update the model weights using the gradients generated by the loss function. Stochastic gradient descent (SGD) or ascend is deployed as an optimizer in the vanilla GAN model. The equation for SGD used to update the parameters θ is shown below in equation $(III - 3)$, where η is the learning rate (~0.0002):

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \qquad (III - 3)$$

3.2.3 Training and Sampling Algorithm

A subtle balance is required for training GAN, as it consists of scheduling the training of both networks, which have completely different tasks to achieve. For every epoch during training, firstly, a mini-batch of m noise vectors is sampled from Gaussian noise distribution. These m noise vectors of length z, are sent as input to the generator, which yields m fake samples. For training the discriminator, m samples of each fake and real data distribution are fed to the discriminator as input, along with their corresponding prediction label & weights are updated for optimization using stochastic gradient ascend. After finishing the training for the discriminator, the generator is trained for the batch of data. For generator training, m samples of fake images are generated and sent to the discriminator for estimating prediction values while keeping the parameters of the discriminator frozen. The loss is calculated keeping the target label as 1, and weight updates are performed using stochastic gradient ascend. Note that the generator can only see the results of the fake part of the loss function while using only the feedback when the discriminator is fooled into yielding a value closer to 1. The loss function deployed here for training is the binary cross-entropy loss, also known as the BCE loss function.

This alternating simultaneous training is crucial. Otherwise, if either of the models (G or D) becomes far better than the other, the learning will curb and result in poorer generated results. Traditionally, for every epoch of the discriminator, the generator is trained for a couple of more epochs, as generator G has a much more challenging job to perform compared to discriminator D. After training, during inference, the discriminator module is detached, and sampling can be performed using only learned generator G to output samples close to the target domain, by feeding noise vector sampled from noise distribution. Algorithm 1 shows the pseudocode for training and inferencing a traditional GAN model

---

Algorithm 1 Training and Inference on Vanilla generative adversarial network using backpropagation

---

Input: P_{data} , P_z , z

Output: (of **D**) Scalar value (between 0 and 1) (of **G**) Image of target domain dimensions with appropriate channels

Method: [Training]

1: **Initialization:** θ_G (parameters of Generator module), θ_D (parameters of Discriminator module), O_{SG}, z, e, $J(\theta)$ (BCE Loss).

2: **for the** number of training epochs **do**

3: Sampling m noise samples of size z

4: Sampling m data samples from P_{data}

5: Generate m images for P_z using **G**

6: Forward pass real and fake samples through **D**

7: Update **D** by stochastic gradient ascend:

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(x^{(i)}\right) + \log \left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right]$$

8: Sampling m noise samples of size z

9: Generate m images for P_z using **G**

10: Forward pass generated samples through **D**

11: Update the **G** by stochastic gradient ascend:

$$\nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \log \left(1 - D\left(G\left(z^{(i)}\right)\right)\right)$$

12: **end for**

Method: [Inference]

1: **Initialization:** z

2: **Load:** θ_G

3: Sampling desired m noise samples of size z

4: Generate m images for P_G using **G**

3.2.4 Issues with Vanilla GAN

A significant number of issues engender when the generative adversarial network is trained with the loss function mentioned in the previous section. The list of the problems faced includes the following:

- Gradient Disappearance while training
- Mode Collapse due to unstable training
- Poor Diversity of GAN generators
- Uncontrollable Training

The objective function of GAN is to optimize the weights in such a way that the two distributions of P_z (distribution of generated data) and P_{data} (distribution of actual data) coincide with each other. However, if the two distributions do not intersect or have a negligible intersection, it could lead to zero or disappearing gradients, thereby stopping the learning of the model. Another issue GAN faces is when the generator learns about a particular class that is miss-classified by the discriminator and is inclined to produce more and more examples of that class. Eventually, when the discriminator learns about the mistake of miss classification, the generator is left with no direction to learn anymore. This scenario affecting the diversity of the generated samples, caused by unstable training when the generator gets stuck in a local minimum, is called mode collapse. There are two types of mode collapse: 1) a subset of particular modes is present in the generated data 2) Almost none of the input data modes are learned by the generator to produce.

3.2.5 Evaluation metrics for image synthesis

For evaluating visual results of generative adversarial networks, two crucial properties of generated distribution are considered: a) Fidelity: which deals with the amount of realism and quality of the generated samples, and b) Diversity: which deals with capturing the essence of every class from the actual data distribution and enabling that the generator produces variegated results, thereby making sure there is no mode collapse in the model. The three most effective evaluation metrics that have been used for benchmarking assessment are presented as follows:

I. Comparing Images:

a. Pixel Distance: An absolute difference between the authentic and generated images is taken in this method. It's not reliable as it hinders the creativity of the model.
b. Feature Distance: This method looks at the features extracted from the generated data matching them with the features from the real data, thereby comparing the higher-level semantic information. For extracting features, an ImageNet pre-trained classifier has been deployed. The output vector will be taken from the final pooling layer for comparing the features, as it will have primitive feature information.

II. Inception Score (IS):

This technique [133] calculates the KL-Divergence metric for image quality inspection. This KL-Divergence means how different conditional distribution (Fidelity) is compared to the marginal distribution (Diversity), thereby estimating the relative entropy. Higher the KL-Divergence means a higher distance between the two distributions, and having a lower value means the quality of the generated images is better. The shortcoming of this technique is that it can be exploited by generating just a couple of realistic samples. The metric only works with the fake images, and no attention is provided to the actual authentic images. This metric also tends to miss out on the beneficial spatial relations among the features. The equation for calculating the Inception Score is elucidated below with $\hat{x}$ & y representing generated data & labels, respectively:

$$\text{IS}\ (G) = \exp\left(E_{\hat{x}\sim p_z} D_{KL}(p(y \mid \hat{x}) \parallel p(y))\right) \qquad (III-4)$$

III. Fréchet Inception Distance (FID):

FID [134] works on multivariate normal distributions of real and fake and functions as an improvement on the Inception score. FID calculates the distance between the statistics of the natural embedding distribution and that of the fake embedding distribution. Lower the FID means the closer the distributions are and the better results they produce. It has been observed that using a larger sample size has reduced noise and selection bias. An expression from [134] about Uni-variate Normal Distribution and Multi-variate Normal Distribution can be expressed below:

$$FID(G) = \|\mu_x - \mu_{\hat{x}}\|^2 + Tr\left(C_x + C_{\hat{x}} - 2\sqrt{C_x C_{\hat{x}}}\right) \qquad (III-5)$$

In the above expression, Tr represents the Trace of the matrix, μ represents the mean and C_x denotes covariance matrix of real distribution, while $C_{\hat{x}}$ denotes the covariance matrix of synthetic distribution. One of the shortcomings of FID is that it uses the features captured by the inception model and misses out on any other. It cannot be helpful for data-specific applications. It usually requires a larger sample size and is comparatively slower in computation. A limited number of stats like mean and covariance are used while missing out on skewness. It performs a decent job of considering any Gaussian noise, gaussian blur, black rectangles, swirl, salt, and pepper noises.

To overcome various issues presented in the vanilla GAN architecture, several solutions were proposed by the research community that can be summarized in Figure 10:

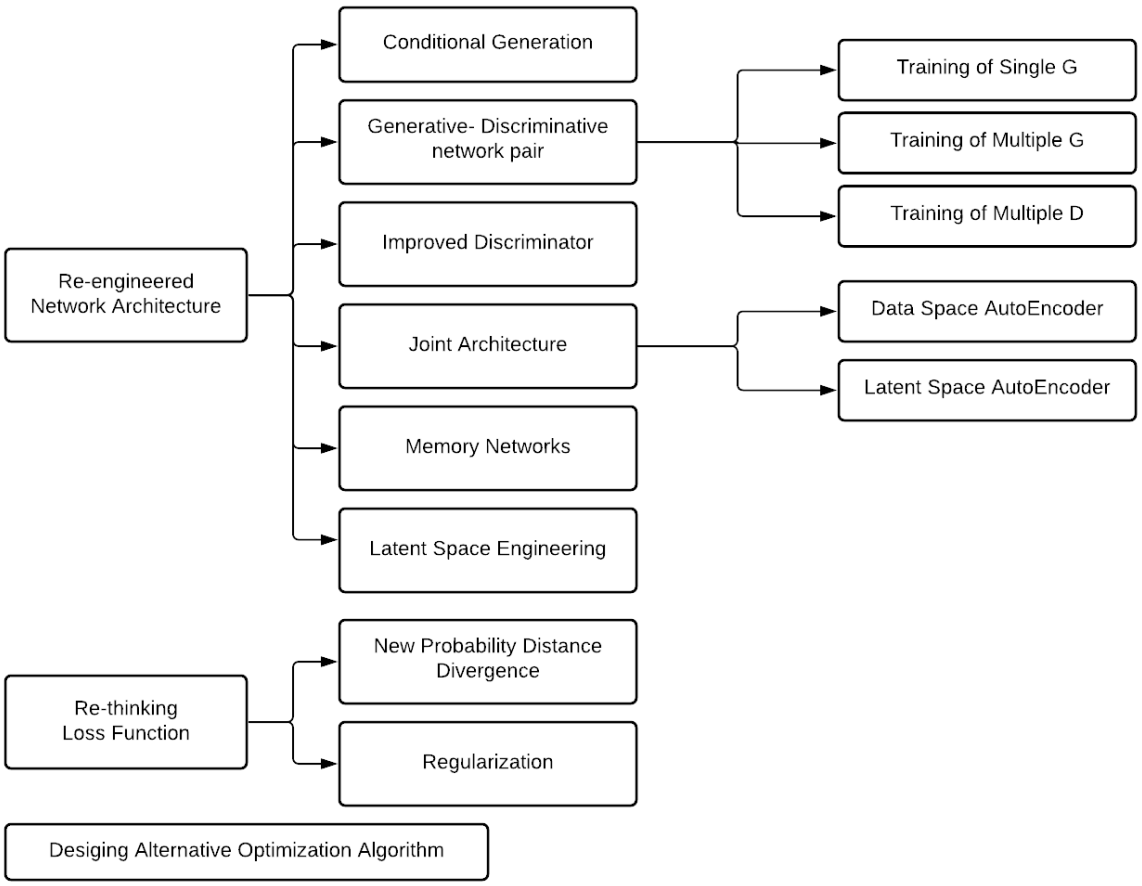

Figure 10: Taxonomy of Generative Adversarial Networks [135]

3.3 Wasserstein GAN

A solution was proposed to address the issue of the loss function, which deals with efforts to make generated distribution equal to the real distribution. The saturating nature of BCE loss causes the gradients to be close to zero if the two distributions are far apart. A unique distance is defined based on the analogy of dirt piles as distribution and distance to move those piles as Earth Mover's Distance. In this system, the loss function is defined so that the gradients keep on growing if the distance is more between the generated and natural distribution. This Earth Mover Distance (EMD), represented in equation $(III-6)$, is a function of the amount of distribution to be moved and the distance between them.

$$W(P_{data}, P_z) = \inf_{\gamma \sim \prod(P_{data}, P_z)} E_{(x,y)\sim\gamma}[\| x - y \|] \quad (III-6)$$

Where $\sim \prod(P_{data}, P_z)$ is the set of all the joint distributions whose marginals are P_{data} and P_z. The generator in WGAN tries to minimize the distance between the distributions, and the critic is tasked with increasing the distance between the real and fake distribution. Unlike BCE loss which has its output bounded between 0 and 1, the outcome of the critic is not bounded and can result in any real output. This enables WGAN to solve the issue of mode collapse and vanishing gradients. WGAN deals with the evergrowing numerical value of weight by pruning the weights in the model. Although this makes the training more stable, there is a chance that gradients can explode or disappear and limit the model's ability to learn.

Along with some significant advantages, there is one condition for WGAN to function properly, and that condition is defined as 1-Lipchitz Continuity. This condition requires the critic's function to be always 1-Lip continuous, meaning the gradient should be at most 1 for every point. This L-continuous condition makes sure the training is stable and the W-loss is valid. For enforcing the Lipchitz continuity, a gradient penalty term is introduced as proposed by [136]. This replaces the task of weight clipping to implement Lipchitz continuity by introducing a regularization term in the loss function. This regularization term penalizes the critic if the gradient is more than 1. It has been proven that WGAN GP (Wasserstein GAN with Gradient Penalty), whose equations are presented in $(III - 7)$ and $(III - 8)$ generates much better samples visually while making sure the training is stable and successful. Although, it has been observed that the time required for convergence is more with gradient penalty as there is an added computational overhead compared to the typical WGAN.

$$J(\theta_D) = -E_{x\sim p_d}[D(x)] + E_{\hat{x}\sim p_g}[D(\hat{x})] +$$
$$\lambda E_{\hat{x}\sim p_g}[(\| \nabla D(\alpha x + (1-\alpha\hat{x}) \|_2 - 1)^2] \quad (III - 7)$$
$$J(\theta_G) = -E_{\hat{x}\sim p_g}[D(\hat{x})] \ (III - 8)$$

3.4 Least Square GAN

Another variant of modification in loss function is proposed by using least square metrics. This loss function, Least square GAN loss [137] described below, has proven to have more stable training gradients and is less prone to exploding or vanishing gradient problems:

$$J(\theta_D) = -E_{x\sim p_d}[(D(x)-1)^2] + E_{\hat{x}\sim p_g}[D(\hat{x})^2] \qquad (III-9)$$

$$J(\theta_G) = -E_{\hat{x}\sim p_g}[(D(\hat{x}-1))^2] \qquad (III-10)$$

3.5 Controllable and Conditional GAN

The actual intuition behind conditional generation means that the class of examples sampled is based on some queries provided to the generator model. The usual task of training such a model requires labeled training data to be available. One easy way to implement conditional generation is by concatenating a one-hot vector where this vector encodes the information about the class. The discriminator of such a model is also provided with class information. This additional input data means that the information about data classes becomes significantly crucial for an accurate conditional generation. The discriminator will reject the input if there is a mismatch between the category of a particular sample and the actual sample from some other class. The information about the one-hot encoded vector can be incorporated with the sample by concatenating matrices of zero and one matrices with the input image in the discriminator. Typically, the size of one-hot vectors depends on the number of classes in the output domain.

The researcher's community investigated the methods to change specific features in the visual output for a controllable generation by tweaking the input noise vector to get different visual artifacts in the desired output manner. Here, unlike conditional generation, the training does not require class labels. The objective of controllable generation is to nudge the training towards generation, which can be easily manipulated by changing the z noise vector. To achieve desirable output features, it is essential to understand how an image morphs while interpolating in the z-space. The goal is to find the right direction for

modifying the desired characteristics in the generated image. The main obstacle while performing controllable synthesis is that the z-space is entangled. The underlying features are often correlated, making it difficult to control one specific feature without changing the other one. For example, if the features are not correlated, then adding the beard should be easy, but as they are correlated, adding the beard adds masculinity to the image, which might not be the desired output, enforcing that the beard feature and masculinity feature are entangled. This entanglement results from z-space not having enough dimensions to provide individual mapping for every feature in the visual representation. One method to achieve controllable generation is by using classifier gradients. In this technique, the generator's weights are frozen, and the generator's output is sent to a pretrained classifier tasked with detecting specific features. This classifier provides gradients using softmax probabilities to the noise vector z and manipulates it. Although a classifier is deployed here to find directions in the z-space, the computational methodology seems lazy and messy.

Another way to achieve controllable generation is by creating disentanglement in the z-space. The latent factors of variations in the noise vector do not depict anything but impact the output image. If the z-space is disentangled, then it means that there will be a presence of specific indices on the noise vectors that can affect the features of the output image. Using class vectors during generation encourages disentangled z-space creation by supervision. Instead of using a one-hot encoder, these class vectors are deployed as embeddings in the noise vector. A different way to achieve disentanglement is by adding a regularization term that encourages the generator to associate each index of the noise vector with a distinct feature.

3.6 High Fidelity Image Synthesis GAN

Intending to achieve high fidelity and diversity in image synthesis tasks, researchers proposed a progressive, growing technique to implement an architecture called ProGAN [94]. The core idea of progressive growing is to gradually increase the more nuanced layers in the generator and discriminator by initially training with only low-resolution layers. The hypothesis behind the method was that this progressive growth (gradually doubling the image resolution after scheduled training intervals) would speed up the training and

increase the training stability. It had shortcomings like semantic sensitivity and constraint dependency of the dataset.

Another attempt at achieving super-resolution was by [138], where the authors designed a large-scale architecture that was trained to generate high quality using ImageNet samples. This architecture named BigGAN has implemented orthogonal regularization in the generator model to tackle the instability caused due to such a scale and truncated latent space to achieve fidelity and diversity among samples.

Recently, the crown of a most successful attempt at solving these issues and achieving the highest fidelity goes to StyleGAN architecture by authors from the Nvidia team [139]. The authors have reconfigured the generator architecture, so the model can generate a specific style of an image using latent space information in each convolution layer. The synthesis process starts with low-resolution images, which progressively grow towards higher resolution images. The input of every layer in the network is modified to achieve a handle over the visual features available in the actual data distribution. This technique implements automatic learning, unsupervised high-level attribute separation, and stochastic variation of generated images, thereby enabling intuitive, scale-specific control synthesis of composition. Figure 11 depicts the internal structure of the Generator module of the StyleGAN network consisting of a mapping & synthesizing network.

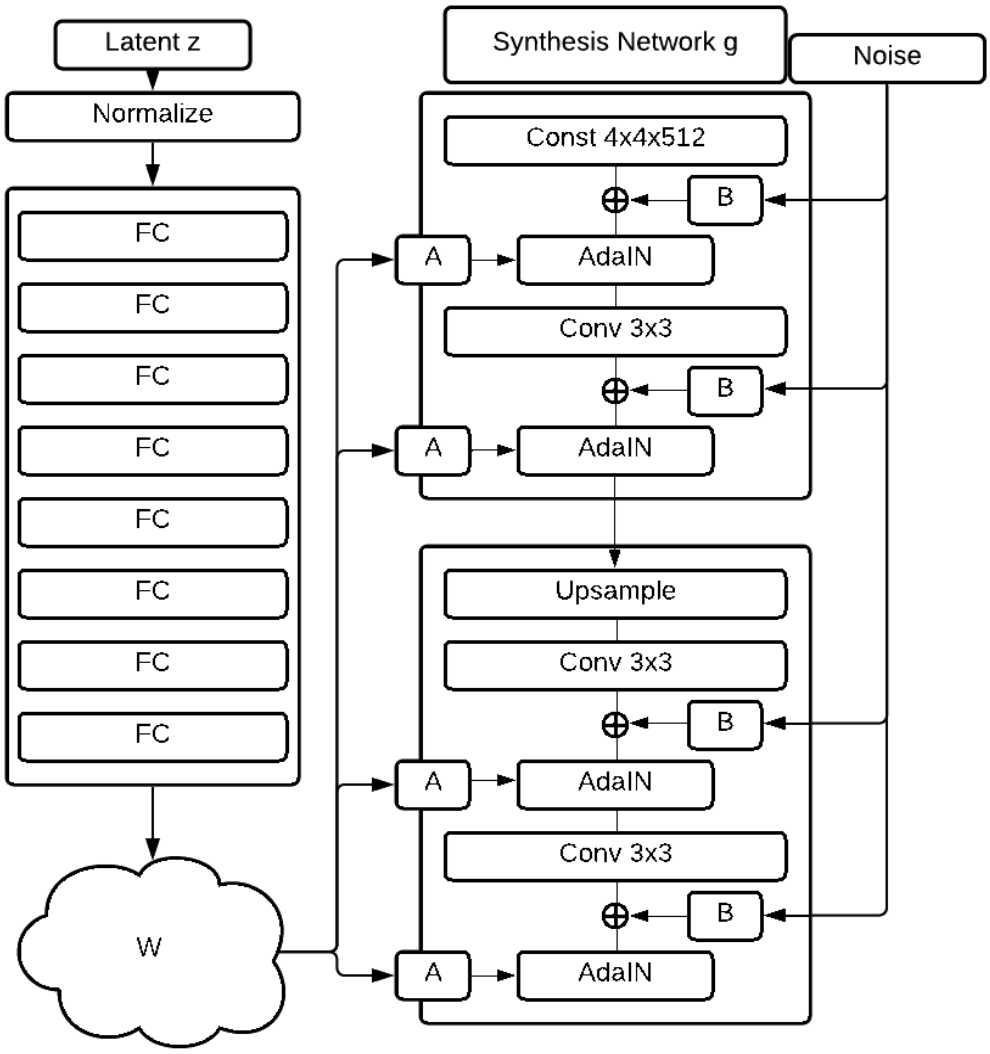

Figure 11: StyleGAN architecture [139]

This architecture works by feeding the noise input vector into a mapping network having learnable parameters, which transforms the input into an intermediate noise *w* vector. Coarse information about image styles will be passed in the first initial layers, followed by more refined styles. These styles are extracted from multiple *w* vector noise and inserted into various generator segments. Here, the *w* noise vector determines the statistics of the image. It has been observed that this mapping network tends to improve disentanglement and provide more control over visual features. The mapping network usually comprises several fully connected layers. Along with progressive growth to improve the results, Adaptive Instance Normalization is implemented, in which *w* vector, after generated from

z noise, is passed to a fully connected layer to engender scaling and bias adaptive styles. This step in the generator transfers the information about the style onto the generated sampled image.

$$AdaIN(x_i, y) = y_{s,i}\frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i} \qquad (III - 11)$$

This StyleGAN architecture has more control over features by having a style mixing procedure. The noise samples from a normal distribution are concatenated to the convolution output before the Adaptive Instance normalization layer to enable the injection of stochastic variation.

3.7 GAN for Data Augmentation & Privacy

GAN has these excellent characteristics of an ideal candidate for data augmentation application. It can produce samples immensely close to the actual data distribution and supplement the real data for classification purposes. This can be crucial when the actual data is expensive and rare. The classic use case of this is in the domain of medical image analysis, like brain tumor or liver lesions diagnosis. The study presented in [140] has proposed a GAN-based method to synthesize medical images for data augmentation tasks in rare diseases.

A study by [141] has designed an approach called fast AnoGAN, which can detect the anomalous nature of features in a variety of biomedical image datasets. Pros of this augmentation technique include better quality and highly refined images (compared to handcrafted examples). It can even generate labeled class sample sets, thereby improving the downstream task of generalization. The list of disadvantages of this method includes the diversity of the generated samples is limited to the diversity of the training examples. And if overfitting to actual data or highly prone mode collapse happens, GAN is rendered useless for augmentation.

GAN can find its use case in privacy, as it can help mask the actual data by generating close to the natural distribution. Such GANs can enforce patient privacy in a clinical environment and encourage data sharing among esteemed institutions. Although, there is

the issue of data leakage, where the GAN sample is almost identical to the real data point but can be addressed by using a couple of postprocessing techniques. GAN can also help provide anonymity and ensure a safe environment for criminal case witnesses and assault victims. One colossal con of this application is Deepfakes, where media can be manipulated for wrongful intentions.

3.8 GAN for Image-to-Image Translation

For transforming styles, generative adversarial networks can be deployed for image/video, text, and facial landmarks to image translation. Recently, it has become a crucial application to deploy algorithms for image-to-image translation. The objective of this image translation task is to learn the mapping from the source image domain to the target domain. This method tries to keep the coarser feature or content of the image intact while changing just the properties of the target domain image to look like the objective domain. Many GAN variants are proposed, which achieve good results in image-to-image translation tasks. Typically, a style transfer method adopts an encoder-decoder discriminator (EDD) architecture and can produce diverse outputs. However, this architecture may tend to generate artifacts in the visually rendered samples. And usually, image-to-image translation GAN models are prone to unstable training, and mode collapses are frequent. Comparing the task of noise to image and image to image, the later model has a more onerous duty of generating an image, using the other image as a reference. Stochasticity is introduced in image-image models using dropout, which drops nodes randomly. Translation from image to image can be either paired or unpaired depending on the requirement and availability of data in a specific application.

3.8.1 Pix-2-Pix

In this model [142], the conditional generator structure is designed to be a U-Net architecture, and the discriminator structure is intended to be a PatchGAN network. The PatchGAN outputs a matrix of classification probabilities according to a patch while still incorporating BCE as a loss function. As a generator having an encoder and decoder framework, the U-Net structure transforms the information into a latent space embedding

using an encoder. Later, using transposed convolutions parses the instructions to generate an appropriate image at the output of the decoder. There is this presence of skip connections between encoder level layers and decoder level layers of the same resolution, which enables transmission of information during forward propagation and increases the flow of gradient during the backward pass while allowing deeper architectures. The input size of the encoder is 256 x 256 x 3, followed by eight encoder blocks, and each encoder block contains convolution, batch norm, and Leaky ReLU layers. This later outputs like a bottleneck layer with output dimensions of 1 x 1 x 512. The decoder segment of the model has eight decoder blocks, each comprising transposed convolution, batch norm, and ReLU layers, and outputs an image of dimensions: 256 x 256 x 3. Dropout is usually inculcated in the first three decoder blocks, which add noise to the network, enabling more diversity in the output. The layer of dropout is only present during training. In contrast, dropout is eliminated during inference, and to consider dropout during training, the inverse dropout probability scales neurons. For a paired image to image translation, pixel distance loss is formulated using L1 loss, which encourages the generator to be softly encouraged to produce more identical outputs to the target domain. For discriminator, loss gradients are generated by comparing the output classification matrix with the fake matrix (full of zeros) or the real matrix (full of ones).

3.8.2 CycleGAN

Often, there is no availability of paired datasets. However, unpaired translation can still be performed on an unpaired image pile of two domains, and still, a decent level of translation results can be achieved using a GAN variant. In this application of unpaired translation, mapping between two piles of datasets is performed. At the same time, the content is preserved, and only the stylistic features are modified to match the target domain images. Here, cycle consistency loss [143] is introduced to perform this task. The discriminator is a PatchGAN, and the generator combines U-Net, DCGAN (acting as a bottleneck), and a ResNet.

In this model, two GANs are deployed simultaneously with reversible roles. One GAN is trying to learn the mapping function from the source domain to the target domain. The

second one is aimed to find the transformation between the target domain and the source domain again. There is no real target outputs in this application. In this CycleGAN model, each discriminator oversees one pile of images. Cycle-consistency loss is the sum of adversarial losses from both the directions of the two GANs, while there is the presence of only one optimizer, which is shared between both the GANs.

$$\begin{aligned} G^*, F^* &= \arg \min_{G,F} \max_{D_X, D_Y} L(G, F, D_X, D_Y) \\ &= L_{GAN}(G, D_Y, X, Y) \\ &\quad + L_{GAN}(F, D_X, Y, X) \\ &\quad + \lambda L_{cyc}(G, F) \end{aligned} \qquad (III - 12)$$

The authors of the papers conducted multiple ablation studies and concluded the following things: 1) using only Cycle loss did not result in good samples 2) using only GAN loss resulted in frequent mode collapse 3) Deploying a combination of GAN loss and Cycle loss helped CycleGAN model in transferring the unique style feature elements while maintaining a shared content between the images of both domains. They also found that using Least Square Loss as adversarial loss helped with vanishing gradients and refraining the model from going into mode collapse. Usually, the GAN loss functions are chosen empirically, and the decision depends on training time availability. An optional loss called identity loss is introduced to preserve color from the original data. This loss dictates that if an opposite generator is deployed, the output should remain almost like the input image when checked using pixel distance loss, thereby expecting identity mapping. One identity loss is introduced for each GAN in the architecture. The loss function will have the least square loss, cycle consistency loss, and identity loss while training a GAN structure. This innovative unsupervised image translation model works excellent without the requirement of aligned image pairs in the dataset while achieving incredible results on translation tasks involving color and texture but fails to translate when geometric changes are required.

3.8.3 UNIT

This technique of Unsupervised image-to-image Translation (UNIT) was proposed by [144], which deals with the concept of shared latent space. This architecture is inspired by a combination of generative adversarial networks and variational autoencoders. The

adversarial training objective interacts with weight-sharing constraints during the generation of images of the corresponding two domains to enforce the shared latent space. Using a variational autoencoder, the architecture connects the translated image to the input image in respective domains. The model efficiently generates street scene image translation and face image translation. There are two limitations of this architecture: a) the presence of saddle point searching problem, making the training unstable b) the translation model is unimodal because of the Gaussian latent space assumption.

$$Loss\ (UNIT) = \begin{matrix} \min\limits_{E1,E2,G1,G2} \max\limits_{D1,D2} L_{VAE1}(E1,G1) + \\ L_{GAN1}(E2,G1,D1) + L_{CC2}(E1,E2,G1,G2) + \\ L_{VAE2}(E2,G2) + L_{GAN2}(E1,G2,D2) + \\ L_{CC2}(E1,E2,G1,G2) \end{matrix} \qquad (III-13)$$

3.8.4 MUNIT

To overcome the issues of unimodality present in UNIT architecture, researchers came up with a multimodal architecture [145], which produces diverse results from a multimodal conditional source distribution. In this technique, two autoencoders are trained simultaneously, while one encodes the image's content, which is domain invariant. The second is encoding the style of the image, which envisions the domain-specific properties. The model's objective is to recombine the content code with a random style code sampled from the style latent space of the output domain to translate. The method has successfully achieved high fidelity and diverse images while controlling the style of translation output.

$$\begin{matrix} \min\limits_{E1,E2,G1,G2} \max\limits_{D1,D2} L(E1,E2,G1,G2,D1,D2) \\ = L_{x1}^{GAN} + L_{x2}^{GAN} + \lambda_x(L_{x1}^{\text{recon}} + L_{x2}^{\text{recon}}) + \\ \lambda_c(L_{c1}^{\text{recon}} + L_{c2}^{\text{recon}}) + \lambda_s(L_{s1}^{\text{recon}} + L_{s2}^{\text{recon}}) \end{matrix} \qquad (III-14)$$

3.9 Diffusion Model

Diffusion models perform image generation by taking input of noisy images and conducting specifically trained denoising operations that gradually remove the noise and unveil the image close to the training data distribution. Recently two researchers from Open

AI [121] proposed their diffusion model, which beats the SOTA GANs with a much better FID score of 7.72 on ImageNet 512x512 [146] and requires fewer training epochs. They incorporated several upgrades to the existing diffusion model, like attention mechanism, adaptive group normalization, and class label conditioning. The concept of this model described in the paper is new, and more ablation studies are required to be performed to prove its potential in image synthesis

CHAPTER 4
PROPOSED MODEL

4.1 Design Objective and Intuition

After conducting the literature survey, we identified several existing challenges in the state-of-the-art models & pipelines. We took the existing architecture as the baseline while designing our pipeline. The challenges we aimed at tackling included a lack of proper focus mechanism that distinguishes skin lesion features from their surrounding less valuable features, lack of robustness and reproducibility while keeping the training time optimized, low diversity among generated samples, and limited dataset samples limiting the learning ability of the algorithm. Usually, deep convolutional models in our survey were trained without a decent augmentation technique and evaluated against a minimal sample size of data, leading to overfitting in the model. We designed a novel model pipeline that augments a synthetically generated dataset to increase the learned feature diversity and fidelity, enforces segmenting lesions from healthy skin and artifacts present in the lesion dataset, localizes features, and extracts them later used by the classifier to make its decision. This pipeline includes a generative adversarial network that can be trained in a few shot learning fashion. The pipeline is further followed by a deep learning architecture that can localize on the most critical features of the lesion. Before passing through various learning algorithms, the input database of images was preprocessed by noise removal, artifact removal techniques, and duplication removal. The thesis later investigates the impact of such preprocessing techniques on our pipeline.

Traditional GAN networks were designed for images with continuous patterns in pixels with observable changes. But skin lesion images are visually different from face images because the patterns and styles are far less diversified and notable. To address these existing issues, the thesis modifies StyleGAN architecture. We propose a version of the style controlling technique designed specifically for medical lesion images with a more straightforward generator and discriminator modules. In our design, a custom StyleGAN2 ADA architecture (SkinGAN) is deployed to synthesize samples of a skin lesion in a few

shots learning manners, to incorporate faster training time while also keeping the model size smaller for clinical server deployment. The SkinGAN model was able to generate highly diversified data while retaining its fidelity as close to the training data as possible. Next in our pipeline, the thesis investigated multiple deep learning architectures and concluded that DenseNet201 [147] architecture was most suitable for feature extraction. This DenseNet201 network was embedded with a novel soft attention module to enforce the model to focus on the input lesion image's salient and relevant region of interest. The soft attention module is better suited for this task than traditional complicated segmentation techniques, which can ignore important characteristics of lesion images. Thereby the hard attention might not end up using the full feature capability of the input dataset for learning, while even leading to misclassification during inference. We also proposed a custom loss function along with mini-batch logic specific to the skin lesion classification task. A unique augmentation pipeline is also introduced to help model in learning feature diversity and prepare itself better for unseen future samples by improving generalization capability. The model pipeline formulated and elucidated in the thesis is named SkinCAN AI. The model pipeline also incorporated features from the metadata of patients, and the thesis later investigates the impact of metadata on the learning process.

4.2 Network Architecture

4.2.1 SkinGAN architecture

The proposed generative adversarial network SkinGAN closely follows StyleGAN2-ADA [44] network by Nvidia researchers but develops on it by modifying the pipeline in such a way that it becomes more suitable for the synthetic generation of skin lesions. Instead of using style mixing in the traditional StyleGAN network by combining two latent vectors, the proposed SkinGAN model only incorporates one latent vector used for synthetic generation to overcome any issues caused by unobservable similarities present in synthetic lesion images.

The generator architecture of SkinGAN is a modified version, as the last layer module is required to generate a pixel dimension of 224, so it can be smoothly passed on to the final

deep learning classifier layer without any dedicated resizing step. The discriminator of SkinGAN is modified to make use of residual connections, along with the technique of Freeze-D (Freeze Discriminator) explained further in detail. The proposed discriminator comprises residual connections to enable faster optimization. We have kept residual connections only in the discriminator, as their presence in the generator has shown negligible to zero improvements in synthetic results, sometimes even leading to longer training times. The proposed model also implements techniques like weighted demodulation and adaptive discriminator augmentation and specialized regularization methods while eliminating any modules or practices that have been shown to hamper the performance that an ideal optimized generative adversarial pipeline can achieve.

4.2.1.1 Adaptive Discriminator Augmentation

In the presence of small training data such as skin lesion diagnosis one, the discriminator characteristically having an abundance of parameters will tend to statistically show poor performance by overfitting, which can be observed by plotting validation accuracy. One more concrete evidence of the overfitting of such a GAN model is that the discriminator shows similar poor performance on synthetic samples. Suppose this is the scenario observed in graphs during training. In that case, it becomes evident that the discriminator is not learning any underlying semantic features of the original data distribution, instead is lazily focusing only on high-frequency patterns in the dataset.

In practice, to address the issue of overfitting, usually in the case of any computer vision application, augmentation transformations are introduced while preserving the label representing features in the image. This augmentation could be performed in several ways, including rotation, image scaling, random crops, or even color transformation. The research community has experimented with the number of techniques to implement this augmentation transformation in a pipeline of the generative network. Attempts to achieve this include contrastive loss or consistency regularization, but the most successful is the dynamic augmentation method. The simplest way to integrate the augmentation module in GAN is by sending the samples of the real dataset and synthetic ones alike in the augmentation pipeline and then letting the discriminator train on them. But this creates the

issue in the generator module, which is supposedly learning simultaneously, to start focusing on those augmented features and generating them with an intent to fool the discriminator. Eventually, the generator module becomes confused between natural semantic features and augmented features; later, after training for a longer time couldn't distinguish between them.

To prevent this issue, a strategy needs to be forged that implements augmentation for regularization in the discriminator but doesn't end up leaking in the semantic features learned by the generator. This leads to the idea of invertible augmentation. These are the augmentations that the generator module can't learn. However, it essentially retains the input data's contextual and overall semantic concept. Thereby avoiding leaking augmented features in the latent space of the generative network. An example of such non-leaking augmentation would include performing rotational transformation between {0°, 90°, 180°, 270°} with a probability chosen from a uniform distribution. The generator would not have any way to distinguish between them, as their frequency is uniform, making the augmentation principally equally embedded in the input data. But the situation is changed if the same transform is happening with non-uniformly distributed probabilities, which might lead to a higher occurrence of one orientation in the data presented to the discriminator and end up confusing the generator later.

In the deployment of SkinGAN architecture, a specialized augmentation pipeline is created consisting of 18 different types of transformation functions. These 18 augmentation techniques include pixel blitting, general geometric transformations, color transformations, image space filtering, and image space corruptions. The proposed model has modified the augmentation pipeline to include only pixel blitting and general geometric transformations. Any other augmentations like color transformation on the skin lesion dataset could completely distort the semantic features that the discriminator should learn. Pixel blitting comprises x-flip, 90° rotations, and integer translation, while geometric transformations consist of fractional translation, anisotropic scaling, arbitrary rotation, and isotropic scaling. These augmentations should be strictly differentiable during training, as it also affects the generator training. A probability scalar p ($p \in [0,1]$) is deployed to control the strength of augmentation. It can be observed that the chances of the discriminator observing

an image straight from the dataset is still unlikely even when the p-value is significantly low due to the presence of several augmentation techniques in the pipeline. At the same time, the generator can keep on yielding images close to the input dataset, given the condition that the value of p doesn't cross any prescribed limits or does not enable leaking in the generator. Higher values of p, especially when they reach 1, can increase the leakage into the generator module because it allows for extra augmentations to be performed while feeding the discriminator.

Stochastic Discriminator augmentation is performed using a technique called Discriminator Googles. In this technique, the discriminator sees the real and generated data as augmented, while the strength of that augmentation is controlled by parameter p. This parameter p enables recovering the semantic idea of original data distribution. The discriminator is guiding the generator while wearing goggles by seeing through the probability masks of each augmentation and is unable to see even a single true clean real data sample.

4.2.1.2 Adaptive Control Scheme for parameter p

The value of p is a crucial hyperparameter to guide the discriminator and generator module. Instead of keeping the p constant, a dynamic scheduling strategy is adopted. Usually, in the practice of computer applications, augmentation is only required when there is an issue of overfitting. So, the same principle is adopted to create a strategy in which the value of p is dynamically varied according to the detection of overfitting in the discriminator module. For measuring the overfitting in discriminator, a unique set of heuristics are deployed, which includes as shown in the equation below:

$$r_v = \frac{E[D_{\text{train}}] - E[D_{\text{validation}}]}{E[D_{\text{train}}] - E[D_{\text{generated}}]} \qquad (IV-1)$$

$$r_t = E[\text{sign}(D_{\text{train}})] \qquad (IV-2)$$

In the above equation, D_{train} , $D_{\text{validation}}$ and $D_{\text{generated}}$ represents the output of the Discriminator module from training data, validation data, and generated datasets,

respectively. While E represents the mean taken over M consecutive mini-batches (M = 4 in practice). Whenever the value of heuristics reaches zero, it depicts the absence of overfitting, while when their value moves closer to 1, that indicates overfitting. The value of heuristic r_v represents the performance of discriminator on validation samples compared to when generator samples are fed to discriminator, with an assumption of the existence of a validation set. The second heuristic r_t works on the principle that how many samples of the training set are considered a positive sample by the discriminator cause if almost all the training samples are considered positive by the discriminator, it becomes clear indicative of overfitting.

In deployment, the value of p is initialized to a value of zero to ensure that no augmentation is implemented in the initial learning stage of the generator. Whenever the heuristic indicates overfitting, the value of p is increment by a constant value to counter that. It is crucial to note that parameter p is changed only once in every four mini-batches because if the value is changed at a much faster or slower rate, that could lead to unstable training and leaking.

4.2.1.2 Freeze-D

In the study [148], the discriminator is split into a feature extractor network and classifier network, focusing only on fine-tuning the classifier for a specific application (here, we will be focusing on skin lesion synthetic generation). This is implemented by freezing the high-resolution lower layers of the discriminator and then performing transfer learning to fine-tune the low-resolution application-specific layers. This has been found to significantly reduce the training time while providing an excellent prior to start from in discriminator, especially in medical applications where the dataset tends to be limited.

4.2.1.3 Weight Demodulation

The older normalization techniques like the adaptive instance normalization tend to embed noisy artifacts like water droplets and phase artifacts, where features are stuck in one local space during latent space interpolation. For reconstructing such a normalization technique,

the noise needs to be outside the style box. Otherwise, too much noise addition introduces speckle artifacts in the generated image, which couldn't be allowed in feature-sensitive medical datasets like that of skin lesions. The process of weight demodulation is performed by scaling the weights in the convolution layers according to the latent vector. After demodulation is accomplished by forcing features to have unit variance. The weight demodulation is implemented using the equation described below:

$$w'_{ijk} = s_i \cdot w_{ijk} \qquad (IV-3)$$

$$w''_{ijk} = \frac{w'_{ijk}}{\sqrt{\sum_{i,k} {w'_{ijk}}^2 + \epsilon}} \qquad (IV-4)$$

Here, w, w', and w'' are original, modulated, and demodulated weights, respectively, $i, j,$ and k represent input feature map, output feature map, and spatial location during convolution, s_i represents the scale decided by ith input feature map, ϵ for numerical stability.

Although the concept of replacing instance normalization with a version of weight normalization is not precisely similar from a mathematical point of view, this change achieves the target of high fidelity generated images with no deterioration of FID by normalizing using unit standard deviation.

4.2.1.4 Perceptual Path Length Regularization

In an ideal GAN network, it is crucial that subtle changes in the latent vector z should map to only slight smooth feature changes in the synthetic image, rather than causing drastic visual changes in the generated image. Perceptual path length regularization is implemented to enforce such requirements.

Studies have established that metrics used for evaluating generative networks like FID score are biased towards generating correct texture rather than generating proper shape. This gap can also be addressed by using a specialized metric called the perceptual path length metric. It has been established that the lower the value of the perceptual path length

score, the better the results would be in maintaining the semantic shape information from the target dataset. The perceptual path length regularization term is introduced in the generator loss function to address mentioned issues. To implement perceptual path length regularization, the following equations are utilized:

$$J_w = \frac{\partial g(w)}{\partial w} \qquad (IV-5)$$

$$J_w^T y = \nabla_w (g(w) \cdot y) \qquad (IV-6)$$

$$E_{w,y\sim N(0,I)} \left(\| J_w^T y \|_2 - a \right)^2 \qquad (IV-7)$$

Here, $w \in W$ is latent vector from latent space W , $g(w): W \mapsto Y$ is the generator mapping from latent space W to image space Y. J_w is the Jacobian matrix describing the small changes, y are synthetic images with normally distributed pixel intensities. $J_w^T y$ is introduced to avoid heavy explicit computation. The constant a depicts the long-running exponential moving average of the L2 norm of $J_w^T y$. Equation $(IV-7)$ becomes the regularization term added to the generator loss function to enforce smooth mapping between latent space and image space.

A study by [149] found that performing regularization using R1 regularization and previously mentioned path length regularization has proved to be a heavy computational burden while training the GAN network on a GPU and can even lead to high computation costs. So, the method proposed by the authors to counter this issue was to perform regularization only once in every 16 mini-batches for the discriminator and every 8 for the generator, as they found no significant impact on performance but decreased the training time and cost. Here in the SkinGAN model, the same methodology is adopted but modified. Both R1 and perceptual path length regularization are only performed every 16 mini-batches for the generator, and 24 mini-batches for the discriminator, as the presence of adaptive augmentation is already significantly helpful in avoiding model overfitting.

4.2.1.5 Residual connections

Progressive growth techniques introduced with the earlier StyleGAN model have produced phase artifacts in the resulting image while also causing a computational overhead during training. Inspired by investigations in a study [150], modifications in discriminator and generator modules are proposed by adding skip connections or residual connections to the architecture while abandoning progressive growth training.

Although our investigation during the thesis found that allowing residual connections in the generator could lead to less knowledge distillation into the weights learned. So, we have implemented residual connections only in the discriminator module, allowing a better flow of information between the layers. The presence of residual links has shown significant improvement in the perceptual path length score while also benefiting the FID score. In a residual connection during implementation, two paths are created, one where the feature maps are upscaled and another is an identity transformation. Later these paths are combined to estimate the output by elementwise addition.

4.2.2 Soft Attention Module

While observing the skin lesion dataset, it becomes visually clear that only a specific portion of the dataset image is the pixels containing semantic feature information about the identity of the lesion. Looking through the data samples, it can be established that images are filled with artifacts such as veins, hair, or even grid-scale marking, which doesn't serve any semantic knowledge for any AI network to learn from. We propose deploying an attention mechanism that addresses the problem of hard segmenting skin lesions before classification and helps eliminate artifacts that can introduce bias in the network. This attention mechanism is implemented with a soft logic, as we don't want to discredit any useful diagnostic indicator while classifying. This is achieved by giving fewer weights to feature maps that correspond to areas in the skin lesion image with low to no semantic information. The equation for soft attention is given below:

$$f_s = \beta f\left(\sum_{n=1}^{N} \text{softmax}\,(W_n * f)\right) \qquad (IV-8)$$

Here $f \in R^{w\times h\times d}$ is a tensor which fed to a 3D convolution layer having weights $W_n \in R^{w\times h\times d\times N}$, N in the above equation represents the number of attention maps or even the size of the 3D weights vector. The Soft Attention mechanism module is depicted in Figure 12.

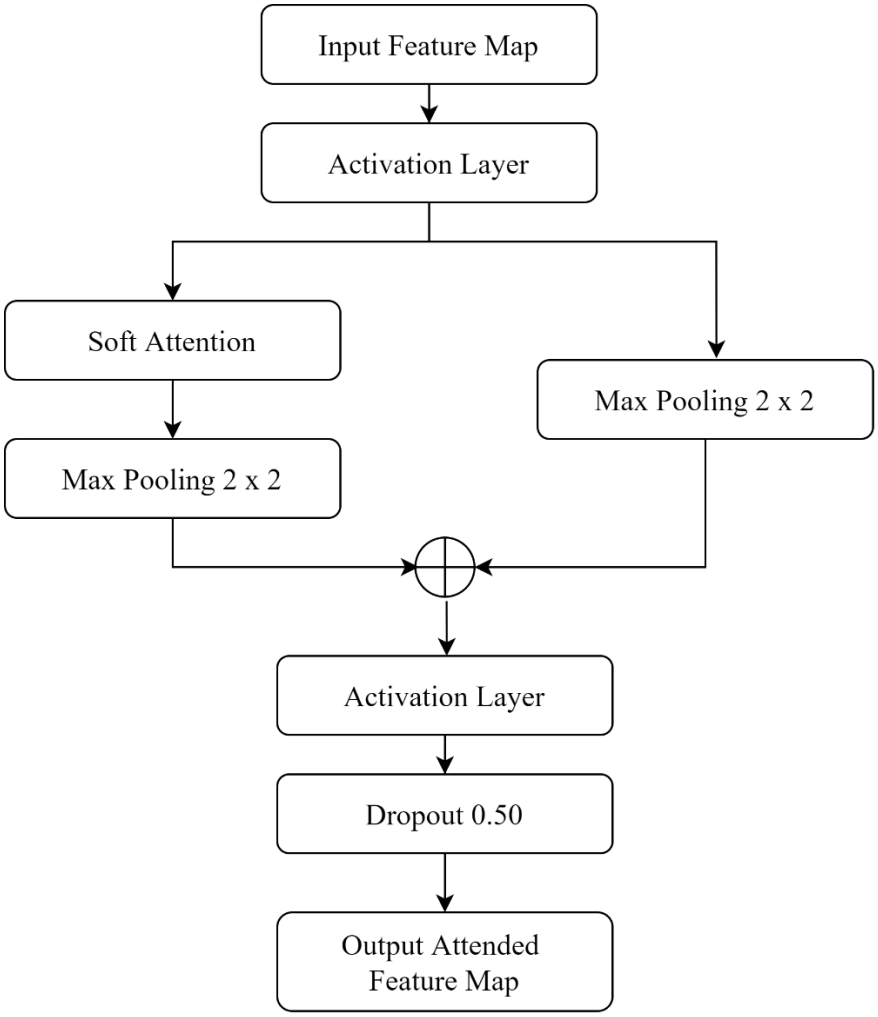

Figure 12: Soft Attention Module

The 3D convolution operations and the softmax layer yield the attention maps, which are concatenated to generate one single attention map. This aggregated attention map is then multiplied with the feature tensor f, to enable variable attention to each feature and scale

them accordingly. These attended features were later scaled by learnable parameter β. Then f_s is concatenated with f using a residual connection, allowing the network to train in an optimized way. Usually, in deployment practice, the value of β is kept low at the start of training, but the network gradually learns to implement an attention map mechanism. This technique eliminates the use of methods like GradCAM [151], which has shown issues such as ignoring critical features in the lesion image or even, in some cases focusing on irrelevant areas. The soft attention mechanism inherently focuses on the classifier's sensitivity, allowing it to focus on correctly classifying positive ones. This is crucial because misclassifying a malignant patient as benign is far more severe than vice versa and could, unfortunately, lead to fatality or cancer metastases in a clinical environment.

4.2.3 DenseNet201

In our proposed methodology, we proposed integrating DenseNet over other networks such as ResNet because it uses the full learned feature map size, instead of using zero-padding like in ResNet where noise is introduced, or using stride 2 in the 1D convolution layer, which eventually loses a few critical learned features. According to investigations performed in study [152], it becomes evident that the list of good candidates for task of skin lesion classification include DenseNet201 [147], InceptionResNetV2 [153], and SE_ResNet150 [154]. But taking into consideration the AUC score and computational processing power utilized during training and inferencing, DenseNet201 proves to be an ideal deep learning architecture.

This architecture was introduced to overcome issues of vanishing gradients caused by the increase in the depth of layers in the model. DenseNet architecture work on the principle that input of the current layer is the concatenation of feature map input of all the previous layer. Dense blocks are tiny modules having dense connections among them. A combination of a 1x1 convolution layer and pooling layer connects these Dense blocks. This 1x1 convolution layer allows shrinking the depth of the feature map while preserving the spatial dimensions. While pooling layers with stride 2 reduces both feature maps and half the spatial dimensions. DenseNet architecture has proven to be highly parameter efficient, as it doesn’t require extra parameters to preserve feature information while also

producing significantly diversified feature maps. It is critical to note that there is no change in feature map dimension inside the Dense block to allow easy concatenation among them. While transition layer connecting Dense blocks consists of a pooling layer only having the control to reduce the feature map dimensions.

The architecture of the Dense block containing the non-linear transformation and bottleneck layer is depicted in Figure 13. The presence of a transition layer between the Dense blocks allows reducing feature map dimensions, which is illustrated in Figure 14. Advantages of DenseNet include a) strong gradient flow, b) parameter and computational efficiency, and c) maintaining low complexity features. The entire architecture of DenseNet201 is shown in Figure 15.

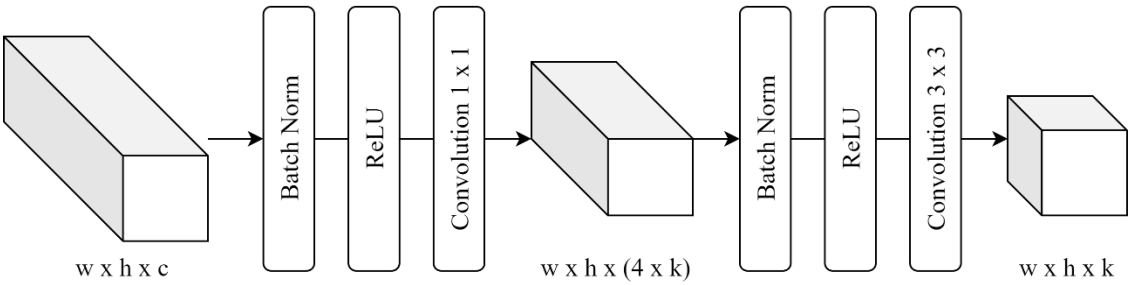

Figure 13: Dense Block comprising of bottleneck layer and non-linear transformation [155]

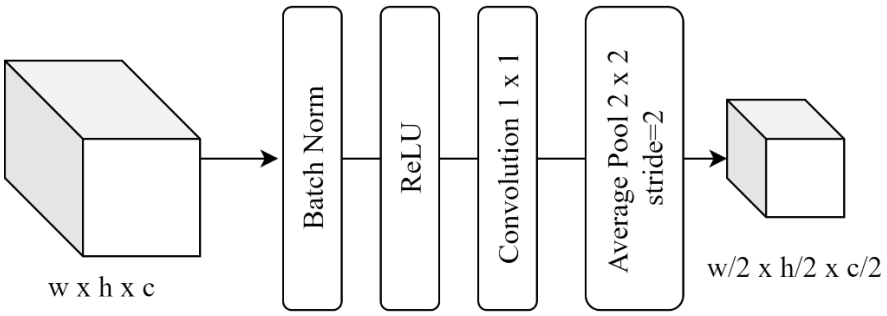

Figure 14: Transition Layer of DenseNet [155]

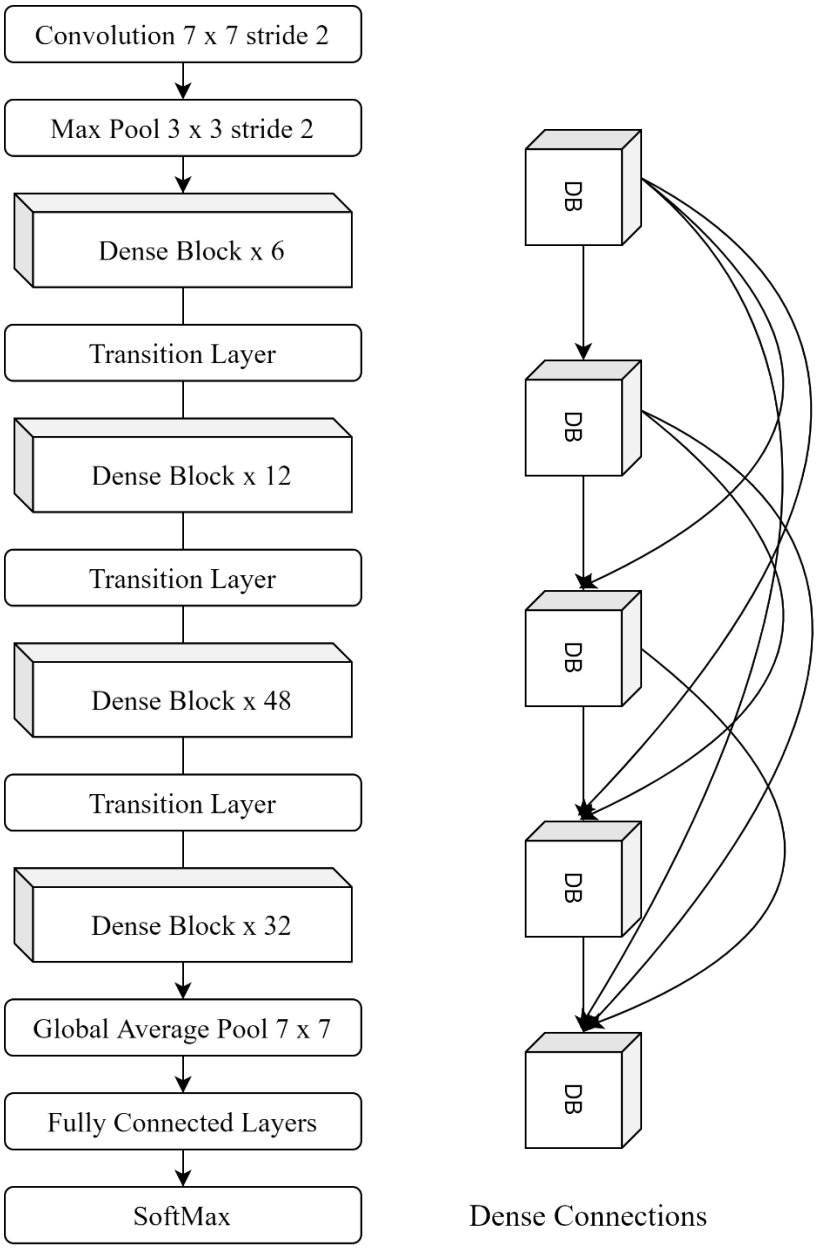

Figure 15: DenseNet201 Architecture along with Dense Connections [155]

4.3 SkinCAN AI Model Setup

This section explains how every module explained before is combined to form a pipeline. The proposed SkinGAN architecture would comprise of generator and discriminator module that allows the generation of high fidelity and high diversity synthetic samples of skin lesions. The generator module will contain a mapping network and synthesis network.

The mapping network transforms the latent vector z into an intermediate latent vector w through normalization and fully connected layers. The synthesis network consists of weight demodulation segments, convolution 3x3, and upsampling layers. The discriminator module includes an adaptive discriminator augmentation module, residual connections, and bilinear downsampling modules in every discriminator block. A minibatch standard deviation layer and a fully connected linear layer are added at the end of the discriminator. At the same time, the classifier deep learning network is identified to be DenseNet201 with a Soft Attention module embedded at the end to enable optimized skin lesion detection. The whole SkinCAN AI pipeline is depicted in Figure 16 at the end of this chapter.

4.4 SkinCAN Loss Function

In this thesis, we propose a modified and improved loss function designed specifically for the task of skin lesion diagnosis. Traditionally for any machine learning task of multiclassification, cross-entropy loss along with softmax is deployed. The softmax function derives the class probabilities of each class for the multiclassification job. The equation for the softmax function is given below:

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad (IV-9)$$

Here the term s_j denotes the scores of each class in C. It is crucial to note that the softmax activation value of a particular s_i will depend on all the scores present in s. The equation of cross-entropy loss (CE) is given below:

$$CE = -\sum_i^C s'_i \log(s_i) \qquad (IV-10)$$

Here s'_i represents the actual value of class i in C, while the s_i depicts the predicted class score by the network. In practice, the softmax activation function is deployed before implementing cross-entropy loss. If the softmax function is deployed before deriving cross-entropy loss, then the loss function is known as categorical cross-entropy loss (CCE) or Softmax loss. In this thesis, we deployed Focal loss, which is an extension of the cross-

entropy loss function but does a better job when there is an imbalance in the class distribution of the training dataset. The focal loss enables a sense of freedom to model in giving a prediction for classes about which the model is not 100% sure. In cancer diagnosis, even the slightest chance of a positive case must be detected, even if it is a false positive. The focal loss is also functional when the information that will decide for classification is sparsely present, and most of the background information of pixels might not be helpful for classification. This situation is very relevant in the case of skin lesion diagnosis, as most pixel information surrounding the actual lesion features is not that useful. Focal loss is also beneficial to the model by making it focus on learning distinguishing features between difficult classes. The equation for implementing focal loss is given below:

$$\text{softmax}(y_i) = \frac{e^{y_i}}{\sum_{i=1}^{C} e^{y_i}}$$

$$L_{\text{focal loss}}(y, y') = -\frac{1}{n}\sum_{i=1}^{n} y'_i \cdot \alpha_i \left(1 - \text{softmax}(y_i)\right)^{\gamma} \times \log\left(\text{softmax}(y_i)\right) \qquad (IV-11)$$

Here the term y is the predicted label, while the term $y'$is the ground truth from the dataset. The term gamma γ is called the modulation factor. The higher the modulation factor, the lower the loss for well-classified samples. Thereby the learning model could focus more on the "difficult-to-classify" samples. The term α is known as the equilibrium weighing factor. Higher weights could be assigned to classes with low samples and smaller weights to dominating categories. At the modulation factor equal to 0 and equilibrium weighing factor equal to 1, the equation of focal loss becomes the equation of cross-entropy loss. Our research experimented with softmax, weighted cross-entropy, angular softmax, and even a mix. Still, it concluded that the model achieves the best performance on the validation set with a focal loss function. With our experimentation, we found the best results with a gamma value of 2 and an alpha value of 0.75.

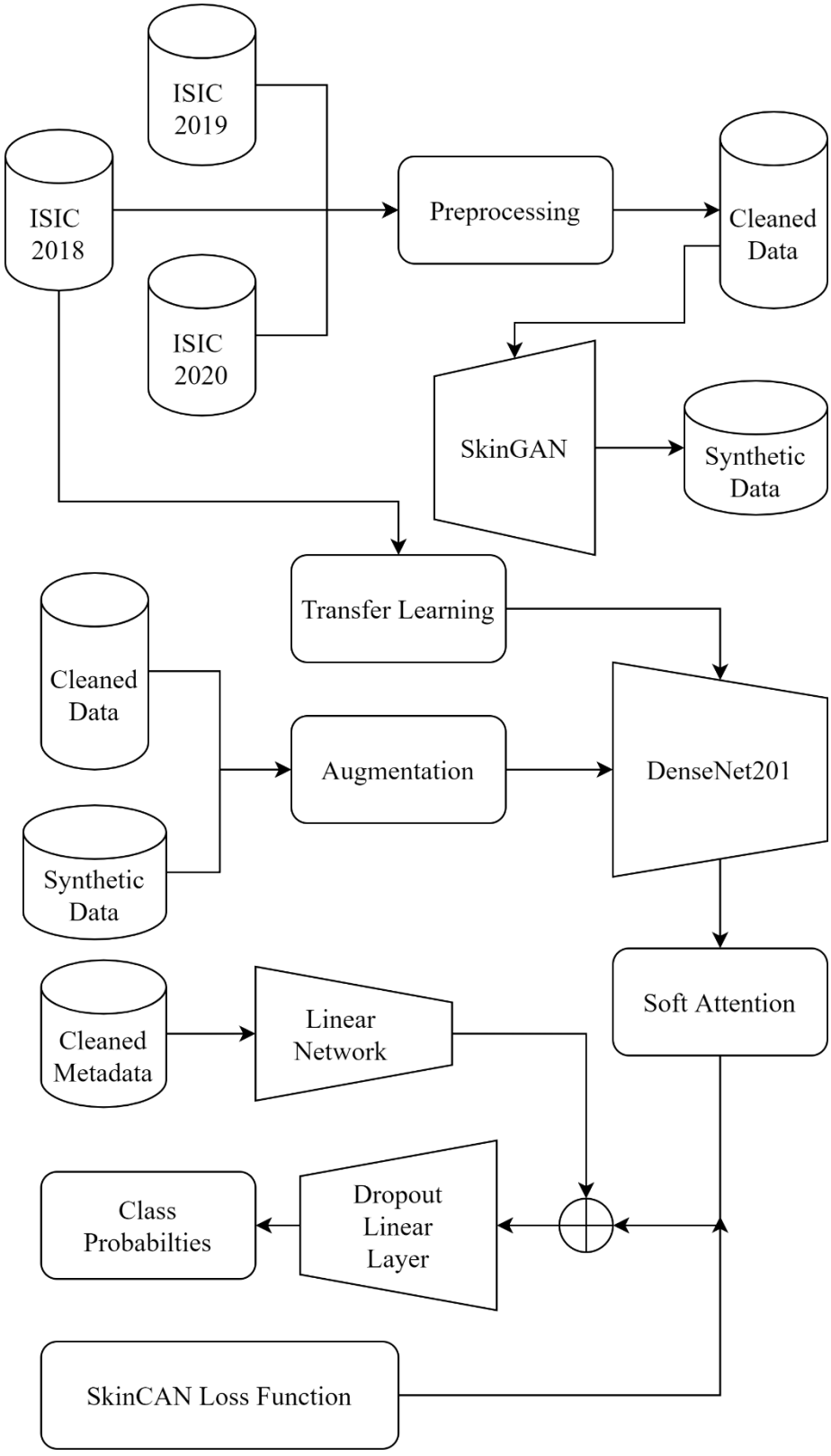

Figure 16: SkinCAN AI Pipeline

CHAPTER 5
EXPERIMENTAL RESULTS & ANALYSIS

5.1 Experimental Setup

While developing this thesis, experiments were performed in the Jupyter notebook environment on Google Colab Pro using Python 3.6.9 with 16 GB P100 GPU with Pascal architecture accompanied with 2 x vCPU having 25 GB RAM and Intel® Xeon ® CPU @ 2.30 GHz. To approach the challenge of diagnosing skin cancer, the experimentation pipeline adopted while designing the model is elucidated in Figure 17.

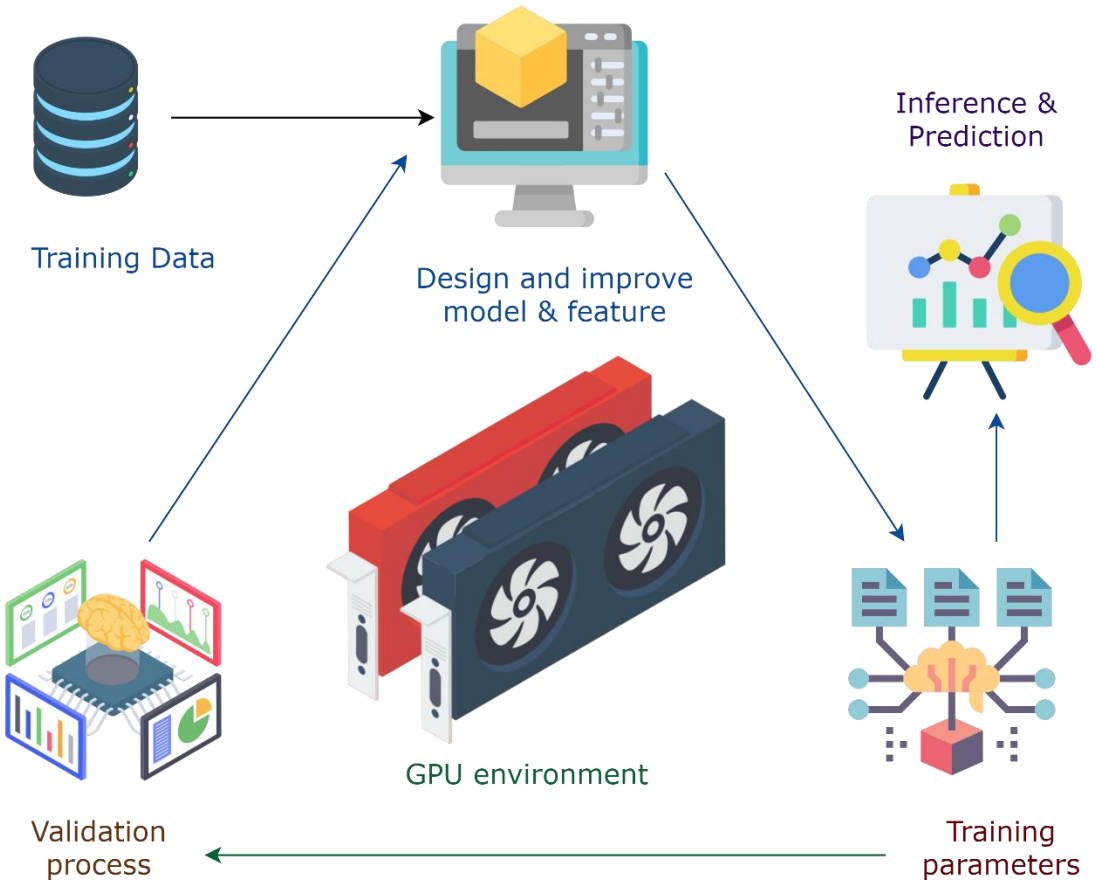

Figure 17: Experimentation pipeline for developing deep learning AI model

5.1.1 Software libraries deployed

The code for the experimentation is built using Pytorch open-source deep-learning library as it has shown significantly cleaner code and faster computation. Few earlier experiments in the study were also conducted using Keras API with Tensorflow-CPU. The software stack used for developing the code includes Rapids CUML, matplotlib, pandas, NumPy, SciPy, and sci-kit learn toolkit. For deep learning computations, libraries utilized for the proposed model have albumentations, hugging face, OpenCV, pretrainedmodels, and weights & biases.

5.1.2 Dataset

The dataset of ISIC 2020 [48] has a total of 33,123 training images, from which 32,542 images are classified as benign, while only 584 samples are malignant melanoma. For the patient, the most critical information is early diagnosis, questioning if the cancer detection is malignant or not. As the dataset had such a skewed class imbalance with only 1.8% of positive cases, we explored external databases that could be amalgamated with this to improve the class ratio. For training our proposed model, ISIC 2020 [48] (having total about 44k samples) along with external data from ISIC 2019 [18], [54], [55] & ISIC 2018 [18], [58] (about 60k samples) datasets are combined (making in total of +100k) and utilized to train the SkinCAN AI pipeline. The older ISIC datasets (2019, 2018) have about over 5k samples in a total of positive cases. Having a much higher percentage of melanoma samples in the training dataset enabled the deep learning pipeline to learn more feature diversity and information. Therefore helps the algorithm to generalize better and prepare well for the real data distribution. The graph below in Figure 18 and Figure 19 shows the distribution of diagnosis classes in the dataset.

As the categories of ISIC 2019 and ISIC 2020 were vastly different, a diagnosis mapping between their classes was established. The rare classes like seborrheic keratosis, lichenoid keratosis, solar lentigo, and lentigo NOS from ISIC 2020 are combined under one classification of benign keratosis, as they all had significantly lower data samples making it impossible to learn them. We also classified café-au-lait macule and atypical melanocytic

proliferation under unknown. At the same time, we were keeping all other classes as it is after processing a dataset with nine categories.

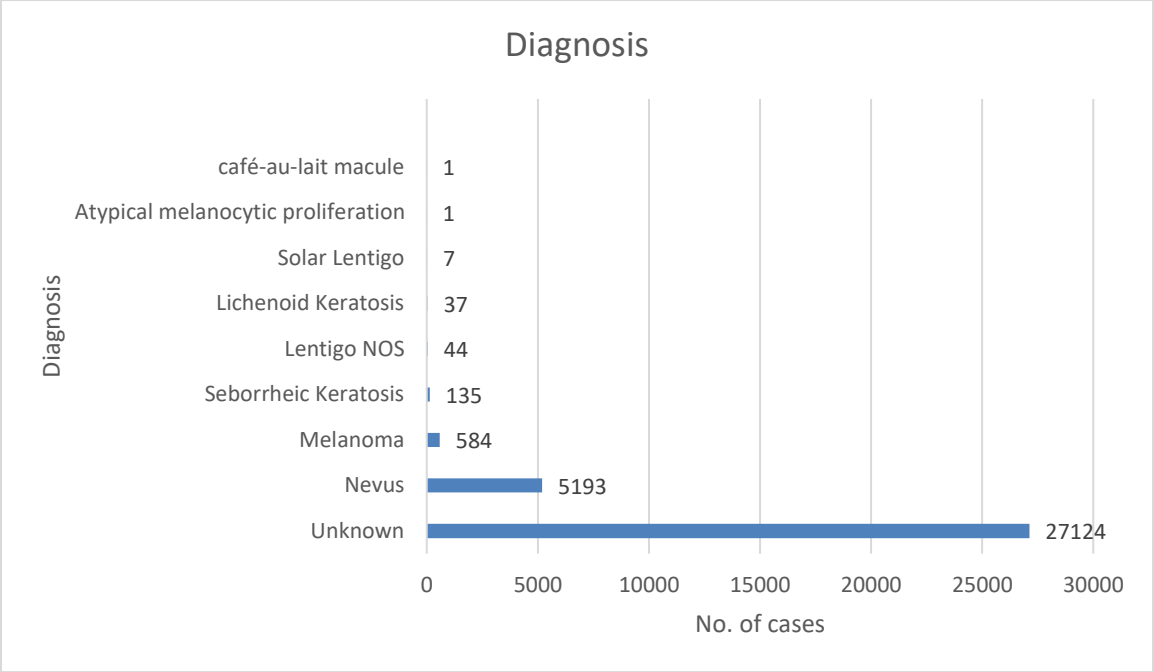

Figure 18: Cases of Diagnosis in the ISIC 2020 dataset

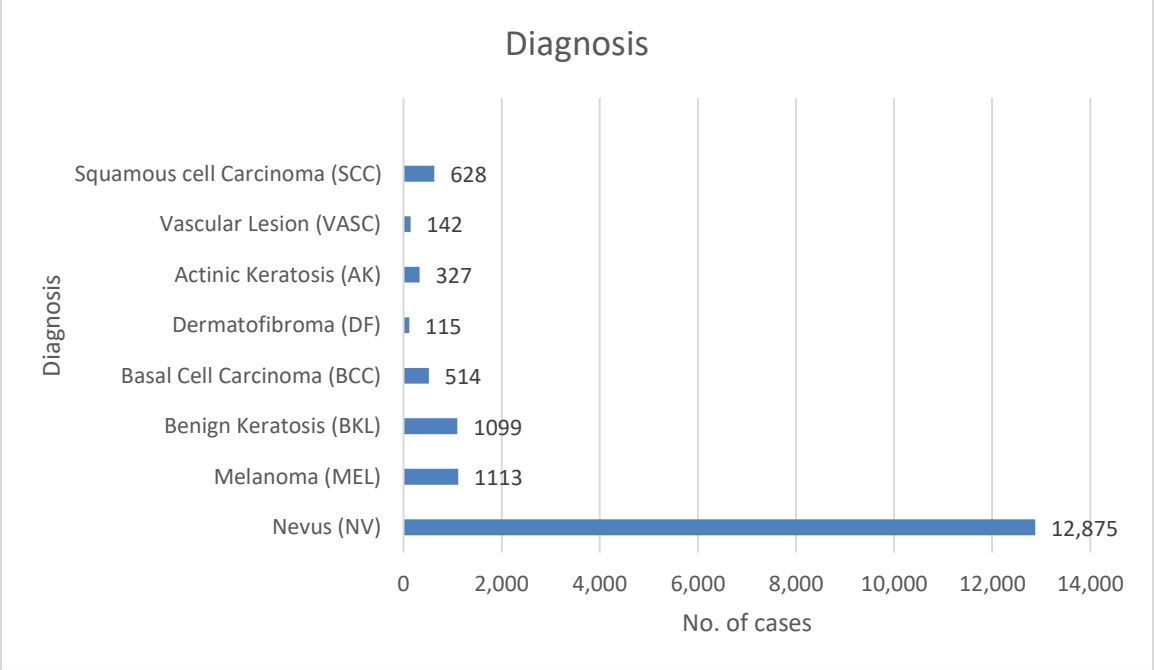

Figure 19: Cases of Diagnosis in the ISIC 2019 dataset

5.1.3 Evaluation Metrics deployed

For evaluation of skin lesion detection from DenseNet201 architecture, the following metrics are deployed in practice:

- True Positive: TP
- False Positive: FP
- True Negative: TN
- False Negative: FN

1) Accuracy (AC)

$$AC = \frac{TN + TP}{TP + FP + TN + FN} \qquad (V - 1)$$

2) Specificity (SP)

$$Specificity = \frac{TN}{TN + FP} \qquad (V - 2)$$

3) Sensitivity (SE)

$$Sensitivity = \frac{TP}{TP + FN} \qquad (V - 3)$$

4) Area under the curve (AUC)

$$AUC = \int_a^b f(x)dx \qquad (V - 4)$$

While for evaluating the performance of the generative adversarial network, the FID score (discussed in the previous section) is utilized in this research.

5.2 Implementation Details

5.2.1 Preprocessing

The dataset consists of images of multiple various resolutions. Some could even be as high as 4000-pixel dimensions and contain too much information about visual features than even needed for computation. In the preprocessing, all the lesion images were center square cropped to focus on the lesion image and then resized to 512 x 512. This step reduced to

overall memory requirement of the dataset from 32GB to 3GB, therefore lowering the budget allocation needed for computational storage.

A triple stratified k-fold strategy was adopted to perform the cross-validation split and maintain the distribution in each split as close as possible to the actual data. The first stratification was that images from the same patient were preserved in the same fold. The second stratification was applied to maintain class distribution in each fold. This depicted strategy ensures that the classifier is not learning too much about a single class in one-fold, later leading to overfitting towards that specific class and causing unstable training steps. The third stratification was implemented to balance patients with more images and others with fewer images in every fold. We found that a plethora of images were duplicated when we combined ISIC of different years. We suspect this might cause a data peeking problem if the training is performed on a particular image and the same image is present in the validation set. We can't simply perform pixel comparison to eliminate duplicates, as some duplicate photos might be slightly tilted or scaled. To address this, a principle was established that similar images would have similar numeric representation in the embedding space once passed through a feature extractor deep learning network. So, we fed all 100k+ images in a pretrained EfficientNet, and a numeric vector of 1000 length was yielded. This vector was reduced dimensionally and plotted on a t-SNE plot. Later, Rapids CUML KNN was deployed in embedding space to observe the duplicates and eliminate them. We found 493 duplicates in the training data, which were removed before further processing, as illustrated in Figure 20.

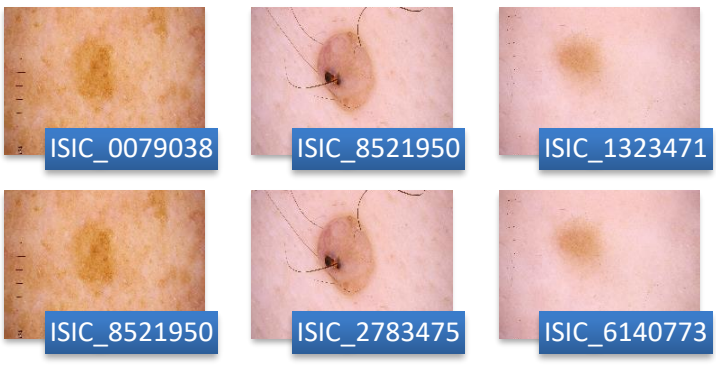

Figure 20: Duplicates in the ISIC dataset

Another important preprocessing step involves removing the hair artifacts, as they might cause difficulty for the soft attention map mechanism to generate accurate maps. The image is passed through a sequence of morphological transformations to remove these hairlike features. Then, replacing the pixel values of hair artifacts with the closest neighboring pixels is implemented using an inpainting algorithm. The grayscale version of the image is passed through a specific black top-hat filter BH, which is yielded by subtracting the O skin lesion image with the C closing of the input skin lesion image.

$$\text{Black Hat } (O) = O_{BH} = (O \cdot C) - O \qquad (V - 5)$$

There has been the presence of missing data in metadata available for patients. These are imputed with the most frequent ones for age and sex. In contrast, the other missing data, like body part location, is set to be unknown. The categorical metadata was one-hot encoded and was combined with numeric metadata. This created an entire set of 14 metadata on which computational training will be performed.

The following preprocessing step is performed only during training DenseNet, which includes augmenting the input data point with multiple various transformation functions, enlisted below:

1. Transpose
2. Flip
3. Rotate
4. Random Brightness
5. Random Contrast
6. Motion Blur
7. Median Blur
8. Gaussian Blur
9. Gaussian Noise
10. Optical Distortion
11. Grid Distortion
12. Elastic Transform
13. CLAHE

14. Hue Saturation Value
15. Shift Scale Rotate
16. Cutout
17. Circular Crop

These augmentations are performed to input data from the original and synthetic datasets alike. These augmentations were performed with specific tuning parameters like the probability of applying them, according to the need of the application, here for skin lesion detection. A library called albumentations was used in the code for deploying such augmentation functions.

5.2.2 Training and Testing Strategy

In our proposed SkinGAN AI pipeline, the generative adversarial network is initially trained to produce high fidelity and diverse set of images. In this step, multiple SkinGANs are deployed to learn the target distribution of each class in the dataset and the synthetic generation of 1000 images for each type of skin lesion diagnosis. This training segment of GAN involves performing adaptive discriminator augmentation while scheduling to change the value of the p parameter depending on the heuristics of overfitting. Also, the regularization of R1 and perceptual path length is performed only every 24 minibatch for the discriminator and 16 for the generator. The high-resolution layers of the discriminator are frozen during GAN training, and it is only fine-tuned in the fully connected layers to the task during training. Before the start of training, DenseNet201 is pretrained on ISIC 2018 dataset. Later DenseNet is trained to classify the lesion correctly by learning the β parameter of the soft attention layer in this process while also using accurately defined soft attention maps on the feature space of the skin lesion. During deployment, all the input training data is preprocessed and transformed into 224 x 224 x 3 while also divided into a minibatch size of 32 [156]. The optimizer used for efficient training is ADAM, while activation chosen for default is ReLU in DenseNet201. Table 4 depicts the training details adopted for the SkinCAN AI pipeline

Table 4: Training Details

Training Details	
Training Fold	5 Stratified Fold
Optimizer	ADAM with tuned learning rate
Computation	Mixed Precision
Batch Size	32
GPU deployed	Single GPU training
Epochs with learning schedule	1 Warmup Epoch with a minimal learning rate + Multiple Cosine decay-based L.R epochs
Categories	9 Classes

5.3 Results and Ablation Studies

The results of SkinGAN AI performance under every diagnostic category are depicted in Table 5. Where AP represents Average Precision, AC represents Accuracy, SE represents Sensitivity, and SP represents Specificity.

Table 5: Diagnosis classification result of the proposed model

Classification Result of SkinCAN AI										
Metrics category	Mean Value	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
AC	**0.949**	0.959	0.945	0.949	0.923	0.953	0.985	0.989	0.991	0.851
AP	**0.602**	0.750	0.936	0.744	0.440	0.606	0.592	0.594	0.354	0.408
AUC	**0.938**	0.933	0.970	0.954	0.953	0.921	0.999	0.9753	0.971	0.768

AUC (SE > 80%)	**0.856**	0.832	0.942	0.896	0.897	0.831	0.986	0.942	0.911	0.469
SP	**0.969**	0.974	0.972	0.936	0.933	0.964	0.986	0.994	0.978	0.992
SE	**0.695**	0.687	0.796	0.849	0.788	0.687	0.913	0.706	0.708	0.123

We performed a comparative study between our model and the state of art models present in the classification tasks, and the results are shown in Table 6.

Table 6: Comparative Analysis of performance with other skin lesion classification task models, keeping the same improved loss function

Models	AC	AUC	SE
Inception-v3	0.857	0.737	0.555
Transfer learning	0.866	0.804	0.620
ResNet50	0.878	0.797	0.597
Inception-v3-LSTM	0.889	0.847	0.630
ResNet50-LSTM	0.867	0.870	0.641
Faster RCNN	0.891	0.888	0.640
DenseNet169	0.899	0.913	0.654
DenseNet201	0.920	0.930	0.666
SE ResNeXt101 32x4d	0.943	0.928	0.613
SkinGAN AI	**0.949**	**0.938**	**0.695**

Next step in our investigation, we analyzed the FID scores of our proposed generative network from other existing generative adversarial networks on combined data of ISIC 2020 and 2019, shown in Table 7.

Table 7: FID score comparison between the generative adversarial networks on synthetic skin lesion image generation

Method	SkinGAN AI	StyleGAN	PGGAN	DCGAN
IS	**2.13**	3.56	4.72	7.91
FID	**0.62**	3.55	5.34	10.41

In our ablation studies, we performed an accuracy analysis to observe the impact of each module proposed by us involved in training, as shown in Table 8.

Table 8: Impact of individual modules on the accuracy of the model

Model	Transfer learning	StyleGAN2-ADA	Improved loss function	Soft Attention	AC
Model 1	-	-	-	-	0.886
Model 2	Yes	Yes	-	-	0.905
Model 3	Yes	-	Yes	-	0.920
Model 4	Yes	Yes	Yes	-	0.927
SkinGAN AI	Yes	Yes	Yes	Yes	**0.949**

As depicted in Figure 21, the synthetically generated skin lesion samples from the SkinGAN generative pipeline are far superior in visual comparison with other models. Models like PGGAN and DCGAN tend to yield noise artifacts like blurriness, checkerboard patterns, or even hair features mesh. Thereby not only lacking in the fidelity of generated image but also in the diversity of samples synthesized. From the visual comparison, it can be quickly established that our proposed SkinGAN can capture the nuances of semantic features while keeping the training requirements minimal and optimal.

Figure 22 illustrates the heatmaps generated by the pipeline's soft attention module compared to the heatmaps of GradCAM, a well-established technique for lesion segmentation. GradCAM has failed to capture the attention map. In some cases, the

GradCAM maps are overshooting to capture irrelevant features, while in others, it can be observed to ignore some vital semantic features.

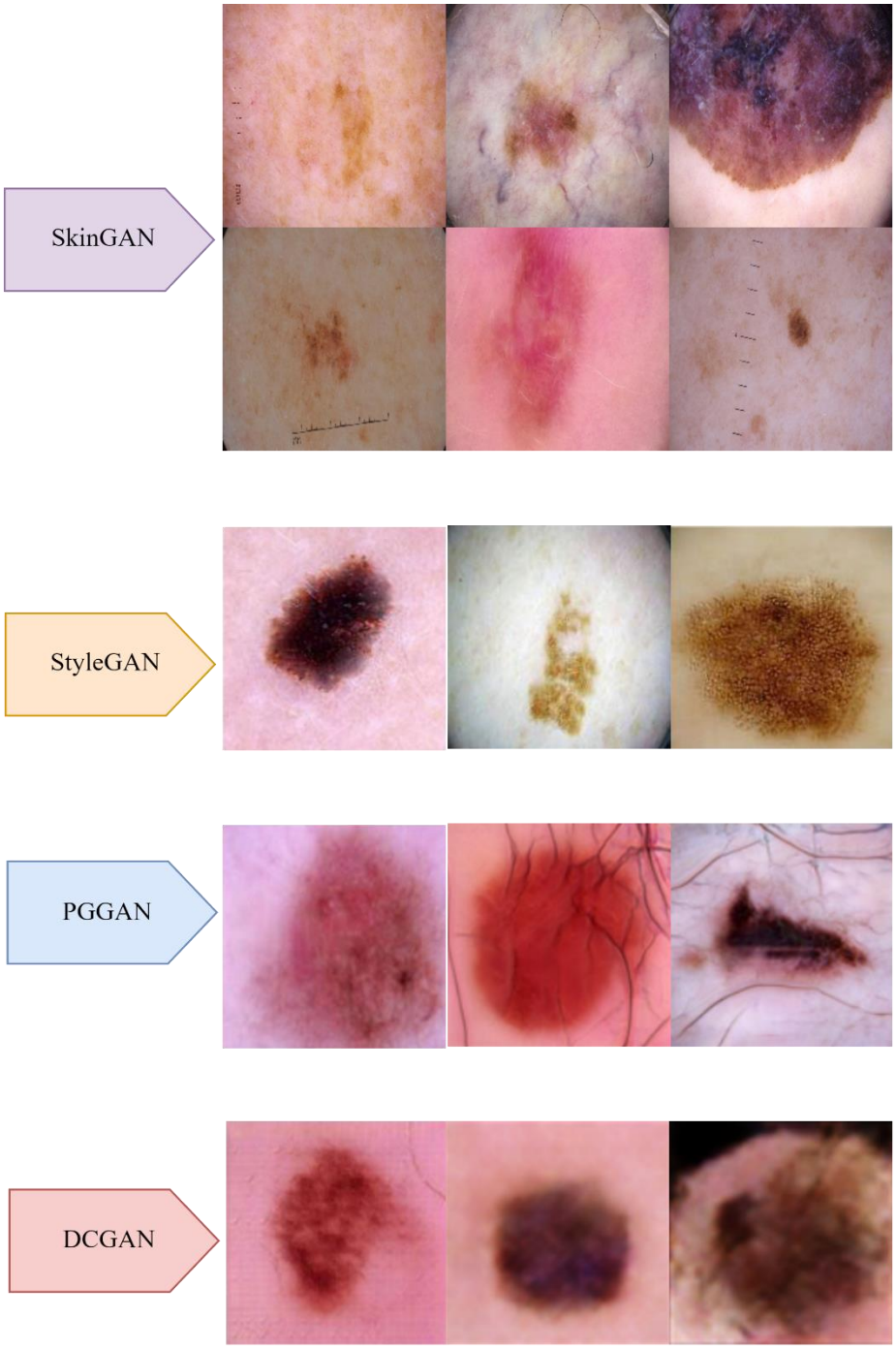

Figure 21: Visual comparative analysis of generated samples of various GAN models

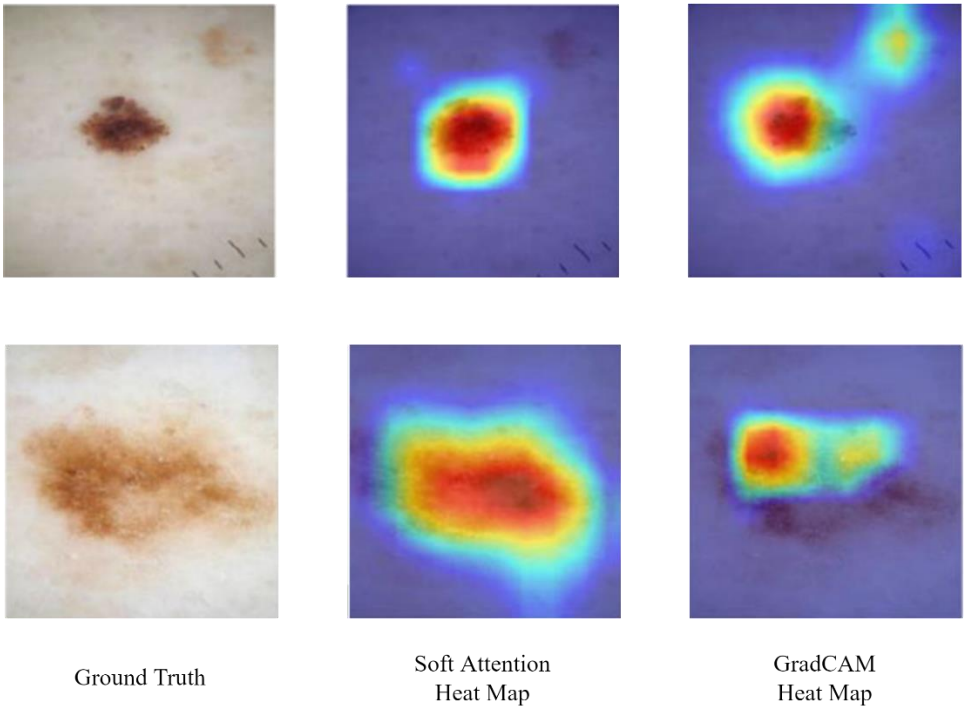

Figure 22: Comparison of heatmaps generated by soft attention module of pipeline and GradCAM

Of course, we would like to accentuate that; the proposed model has its shortcomings as elucidated moving forward. The interpolated overlap in lesion latent space while generation can lead to misclassification error. Although our model is outperforming most models in practice, it doesn't imply that the same result can be exactly expected from a completely unknown set. And more in-depth ablation studies are required to be performed before algorithms like such can be embedded in a clinical environment.

CHAPTER 6
SIGNIFICANCE & FUTURE DIRECTION

6.1 Significance

GANs have several excellent characteristics to be an ideal candidate for the task of data augmentation. GANs can produce synthetic samples close to the real data distribution, and those samples can supplement the real data for classification purposes. This can be crucial when accurate actual data is expensive and rare in nature. The classical use case of this is in the domain of medical image analysis, like brain tumor or liver lesions diagnosis. In the past couple of years, significant communities of researchers have proposed their GAN-based methods, which can synthesize medical images for data augmentation in various rare diseases classification tasks.

The current existing state of art GAN network can synthesize high fidelity images with a decent FID score but is unsuitable for medical application, as training requires quite large training data, which are often scarce for rare diseases, specifically cancer. Even though the existing GAN models have low inference latency, the network still requires hefty training even on fast computing processors like GPU. GAN architecture requires a tuned loss function that, even at a lower training span and lower memory requirement, can capture the fidelity & diversity of training distribution. So, this is where the significance of our proposed model SkinCAN AI comes into the equation. The novelty of the proposed pipeline lies in the simplicity of the architecture deployed for the task and its ability to capture the essential features from an even smaller dataset, thereby surpassing the performance of the existing trained model. The training time requirement and model size make it more suitable for a clinical environment.

6.2 Explainable Artificial Intelligence (XAI)

The explainability, interpretability, or reasoning skill of artificially intelligent models are currently under many questionnaires. As it has become increasingly clear, humans need to address these trust issues on the black-box nature of deep learning if we as humans continue

finding new ways to augment AI in our daily lives. Existing XAI approaches are primarily developed to evaluate a situation where AI is dealing with much simpler tasks. Research is still lagging for properly evaluating computer-aided medical diagnosis tools' interpretability.

One of the most commonly adopted methodologies for establishing confidence in AI models is visually observing saliency maps. A saliency map can be yielded by scoring which pixels influence the most in the decision-making process during the classification step. The pixel dependency value could be positive or even negative for deciding on critical semantic features. Several model explainers are being developed, including IntGrad [157] and GradCAM [158]. In our research, we have developed a module of soft attention to ensure that the classifier model focuses on the correct semantic feature segments. We tested our soft attention module maps with the segmentation masks provided during the ISIC 2018 task and visually found a significantly high image similarity score. For future work to better understand the generative model's interpretability, we wish to explore latent space explainers [159] while also including field experts for judging individual skin lesion classes' semantic features learned by the generative model.

6.3 Future Research Direction

While investigating the literature survey and performing ablation studies for the thesis, we created an intensive list of open research challenges that we would like to present here. We hope that this list can be used as a reference for future research and wish to explore it in our subsequent future research. These open research challenges are enlisted below:

- Long training periods: One of the significant challenges in training a neural network is quickly learning intricate skin lesion features in a few shots during training. This could be investigated by a Few Shot manner extensive network trained on multiple medical data, like CLIP [160] but for medical images.
- Lack of observable visual features in early-stage skin tumors: Medical research has established that when a lesion is at an earlier stage of being malignant, it is much

more difficult to detect during diagnosis. Other indicators instead of visual characteristics should be explored to address this challenge.

- Bias towards Caucasian population: Research community that investigates the bias in AI model has found a challenging bias in existing dermoscopic images towards light-skinned people, on which all state of the art medical algorithms is trained. These datasets are old, and not enough data is collected for dark-skinned people, making it more challenging for AI to detect or even learn such out-of-distribution lesion features.
- Less interclass feature variation among variants of skin cancer: Limited visual variation between malignant and benign skin cancer and even among classes of skin cancer leaves a lot of room for error. Some lesions are even tricky for expertly skilled dermatologists to identify. For such lesion diagnosis, clinically prescribed biopsy techniques for diagnosis are the only way currently.
- Lack of computational processing power: Usually, in a clinical environment, healthcare institutions are not willing to invest any budget in acquiring a dedicated GPU for deep learning diagnosis tasks. We believe that a central cloud computing service could be developed to address this.
- Investigation with genetic and environmental factors: The experts in medical science have identified a strong correlation between genetic factors and the risk of an individual developing Melanoma skin cancer. But such factors are still yet to be implemented or explored to supplement computer-aided diagnosis tools for increasing their performance.

CHAPTER 7 CONCLUSION

In this thesis, a novel pipeline called SkinCAN AI has been proposed to perform the challenging task of early detection of malignant skin lesions and create a soft segmentation mask to assist dermatologists in making critical decisions during skin tumor diagnosis. To address the current issue of limited availability of datasets, the proposed SkinGAN model enables the generation of synthetic data samples of skin lesions and improves the training metrics of the proposed DenseNet model for skin lesion classification. The previous attempts by the research community to embed generative adversarial networks have proven to have stability issues during training and end up requiring heavy computational resources, eventually making them not feasible to be deployed in the clinical environment. Overcoming these shortcomings, the proposed architecture was designed so that the algorithm can learn from a limited dataset without asking for high computational resources during both training and inference. The key feature of the SkinCAN AI pipeline is its ability to yield diagnostic results not only with high precision but also requiring less computational overhead, making it more suitable for early medical diagnosis in such a fast-paced clinical environment. The novel methodology of the SkinCAN AI adopts several proposed strategies like adaptive discriminator augmentation, weight demodulation, freeze D, soft attention module, and path length regularization to enable competitive results compared to other existing methods. Additionally, we trained our proposed pipeline with ISIC 2020 dataset and investigated the design choices made during the formulation of the model. The satisfactory performance of SkinCAN AI on such an imbalanced dataset with only 2% malignant proves its capabilities to be deployed as a computer-aided diagnosis tool for a clinical environment. The novel technique of deploying soft attention described in the thesis could help alleviate issues caused in training caused by artifacts in the dataset and allow the classifier network to focus only on the relevant regions for semantic information about a skin lesion. In future research work, we would like to explore deploying such end-to-end trained diagnosis pipelines for cellular-based applications or even test their performance on low-budget online cloud-based GPU for inference and testing its feasibility in-depth for the clinical environment. We wish to explore the possibility of deploying diffusion models or even self-supervised networks for skin lesion diagnosis.

BIBLIOGRAPHY

[1] J. E. Gershenwald, R. A. Scolyer, K. R. Hess, V. K. Sondak, G. V. Long, and M. I. Ross, "Melanoma staging: evidence-based changes in the American Joint Committee on Cancer eighth edition cancer stag-ing manual," *CA Cancer J Clin*, vol. 67, no. 6, 2017.

[2] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, R. H. Mak, R. M. Tamimi, C. M. Tempany, C. Swanton, U. Hoffmann, L. H. Schwartz, R. J. Gillies, R. Y. Huang, and H. J. W. L. Aerts, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA Cancer J. Clin.*, vol. 69, no. 2, pp. 127–157, Mar. 2019.

[3] "Cancer Facts & Figures 2021." [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2021.html. [Accessed: 22-Sep-2021].

[4] "Skin cancer (including melanoma)—patient version." [Online]. Available: https://www.cancer.gov/types/skin. [Accessed: 17-Sep-2021].

[5] "Skin cancer facts & statistics," 11-Mar-2019. [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/. [Accessed: 22-Sep-2021].

[6] "Types of skin cancer: Common, rare and more varieties," 08-Oct-2018. [Online]. Available: https://www.cancercenter.com/cancer-types/skin-cancer/types. [Accessed: 17-Sep-2021].

[7] UCSF Health, "Organ transplant and skin cancer risk," 14-Mar-2019. [Online]. Available: https://www.ucsfhealth.org/education/organ-transplant-and-skin-cancer-risk. [Accessed: 05-Oct-2021].

[8] H. W. Lim, S. A. B. Collins, J. S. Resneck Jr, J. L. Bolognia, J. A. Hodge, T. A. Rohrer, M. J. Van Beek, D. J. Margolis, A. J. Sober, M. A. Weinstock, D. R. Nerenz, W.

Smith Begolka, and J. V. Moyano, “The burden of skin disease in the United States,” *J. Am. Acad. Dermatol.*, vol. 76, no. 5, pp. 958-972.e2, May 2017.

[9] G. P. Guy Jr, S. R. Machlin, D. U. Ekwueme, and K. R. Yabroff, “Prevalence and costs of skin cancer treatment in the U.s., 2002−2006 and 2007−2011,” *Am. J. Prev. Med.*, vol. 48, no. 2, pp. 183–187, Feb. 2015.

[10] A. Pfahlberg, K. F. Kölmel, and O. Gefeller, “Timing of excessive ultraviolet radiation and melanoma: epidemiology does not support the existence of a critical period of high susceptibility to solar ultraviolet radiation-induced melanoma,” *Br J Dermatol*, vol. 144, no. 3, 2001.

[11] R. A. Lew, A. J. Sober, N. Cook, R. Marvell, and T. B. Fitzpatrick, “Sun exposure habits in patients with cutaneous melanoma: a case control study,” *J. Dermatol. Surg. Oncol.*, vol. 9, no. 12, pp. 981–986, Dec. 1983.

[12] W. Ting, K. Schultz, N. N. Cac, M. Peterson, and H. W. Walling, “Tanning bed exposure increases the risk of malignant melanoma,” *Int J Dermatol*, vol. Dec;46(12):1253-7, 2007.

[13] M. R. Wehner, M.-M. Chren, D. Nameth, A. Choudhry, M. Gaskins, K. T. Nead, W. J. Boscardin, and E. Linos, “International prevalence of indoor tanning,” *JAMA Dermatol.*, vol. 150, no. 4, p. 390, Apr. 2014.

[14] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer Statistics, 2019,”CA:A,” *CancerJournal for Clinicians*, vol. 69, pp. 7–34, 2019.

[15] “Skin Cancer Pictures & Photos.” [Online]. Available: https://www.cancer.org/cancer/skin-cancer/skin-cancer-image-gallery.html?filter=Basal. [Accessed: 05-Oct-2021].

[16] Z. V. Fong and K. K. Tanabe, “Comparison of melanoma guidelines in the U.S.A., Canada, Europe, Australia and New Zealand: a critical appraisal and comprehensive review,” *Br. J. Dermatol.*, vol. 170, no. 1, pp. 20–30, Jan. 2014.

[17] R. M. Cymerman, Y. Shao, K. Wang, Y. Zhang, E. C. Murzaku, L. A. Penn, I. Osman, and D. Polsky, “De Novo vs Nevus-Associated Melanomas: Differences in Associations With Prognostic Indicators and Survival,” *J. Natl. Cancer Inst.*, vol. 108, no. 10, Oct. 2016.

[18] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, p. 180161, Aug. 2018.

[19] P. T. Bradford, D. M. Freedman, A. M. Goldstein, and M. A. Tucker, “Increased risk of second primary cancers after a diagnosis of melanoma,” *Arch Dermatol*, vol. 146, no. 3, pp. 265–272, Mar. 2010.

[20] “USCS Data Visualizations.” [Online]. Available: https://gis.cdc.gov/Cancer/USCS/#/Demographics/. [Accessed: 20-Oct-2021].

[21] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2021,” *CA Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, Jan. 2021.

[22] O. N. Agbai, K. Buster, M. Sanchez, C. Hernandez, R. V. Kundu, M. Chiu, W. E. Roberts, Z. D. Draelos, R. Bhushan, S. C. Taylor, and H. W. Lim, “Skin cancer and photoprotection in people of color: a review and recommendations for physicians and the public,” *J. Am. Acad. Dermatol.*, vol. 70, no. 4, pp. 748–762, Apr. 2014.

[23] M. S. Brady, S. A. Oliveria, and P. J. Christos, “Patterns of detection in patients with cutaneous melanoma,” *Cancer*, vol. 89, pp. 342–347, 2000.

[24] M. Y. Cheng, J. F. Moreau, S. T. McGuire, J. Ho, and L. K. Ferris, “Melanoma depth in patients with an established dermatologist,” *Journal of the American Academy of Dermatology*, vol. 70, no. 5, 2014.

[25] “How to do a skin self-exam.” [Online]. Available: https://www.cancer.org/healthy/be-safe-in-sun/skin-exams.html. [Accessed: 20-Oct-2021].

[26] K. T. Tran, N. A. Wright, and C. J. Cockerell, “Biopsy of the pigmented lesion--when and how,” *J. Am. Acad. Dermatol.*, vol. 59, no. 5, pp. 852–871, Nov. 2008.

[27] S. Wahie and C. M. Lawrence, “Wound complications following diagnostic skin biopsies in dermatology inpatients,” *Arch. Dermatol.*, vol. 143, no. 10, pp. 1267–1271, Oct. 2007.

[28] M. O. Agnieszka Kardynal, “Modern non-invasive diagnostic techniques in the detection of early cutaneous melanoma,” *J. Dermatol. Case Rep.*, vol. 8, no. 1, p. 1, Mar. 2014.

[29] L. M. McIntosh, R. Summers, M. Jackson, H. H. Mantsch, J. R. Mansfield, M. Howlett, A. N. Crowson, and J. W. P. Toole, “Towards Non-Invasive Screening of Skin Lesions by Near-Infrared Spectroscopy,” *J. Invest. Dermatol.*, vol. 116, no. 1, pp. 175–181, Jan. 2001.

[30] C. Fink and H. A. Haenssle, “Non-invasive tools for the diagnosis of cutaneous melanoma,” *Skin Res. Technol.*, vol. 23, no. 3, pp. 261–271, Aug. 2017.

[31] M. E. Celebi, H. Iyatomi, G. Schaefer, W. V. J. C. Stoecker, and Graphics, “Lesion border detection in dermoscopy images,” vol. 33, no. 2. pp. 148–153, 2009.

[32] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, “Diagnostic accuracy of dermoscopy,” *Lancet Oncol.*, vol. 3, no. 3, pp. 159–165, Mar. 2002.

[33] S. W. Menzies, L. Bischof, H. Talbot, A. Gutenev, M. Avramidis, L. Wong, S. K. Lo, G. Mackellar, V. Skladnev, and W. McCarthy, “The performance of SolarScan: An automated dermoscopy image analysis instrument for the diag-nosis of primary melanoma,” *Arch. Dermatol*, pp. 1388–1396, 2005.

[34] G. Zouridakis, M. D. M. Duvic, and N. A. Mullani, “Transillumination Imaging for Early Skin Cancer Detection,” Department of Computer Science, University of Houston, Houston, TX, USA, Technol Report 2005;Biomedical Imaging Lab. , 2005.

[35] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, and H. Pehamberger, “Epiluminescence microscopy. A useful tool for the diagnosis of pigmented

skin lesions for formally trained dermatologists," *Arch. Dermatol.*, vol. 131, no. 3, pp. 286–291, Mar. 1995.

[36] H. Pehamberger, M. Binder, A. Steiner, and K. Wolff, "In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma," *J. Invest. Dermatol.*, vol. 100, no. 3, pp. 356S-362S, Mar. 1993.

[37] A. P. Dhawan, R. Gordon, and R. M. Rangayyan, "Nevoscopy: three-dimensional computed tomography of nevi and melanomas in situ by transillumination," *IEEE Trans. Med. Imaging*, vol. 3, no. 2, pp. 54–61, 1984.

[38] D. Piccolo, A. Ferrari, K. Peris, R. Diadone, B. Ruggeri, and S. Chimenti, "Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: a comparative study," *Br. J. Dermatol.*, vol. 147, no. 3, pp. 481–486, Sep. 2002.

[39] R. J. C. Marks, "Epidemiology of melanoma," *Clinical dermatology• Review article*, vol. 25, pp. 459–463, 2000.

[40] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Trans. Neural Netw.*, vol. 4, no. 6, pp. 962–969, 1993.

[41] B. Krawczyk and M. Woźniak, "Cost-sensitive neural network with ROC-based moving threshold for imbalanced classification," in *Intelligent Data Engineering and Automated Learning – IDEAL 2015*, Springer International Publishing, 2015, pp. 45–52.

[42] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 495–503, Feb. 2020.

[43] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, vol. 2, 2003.

[44] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, “Training Generative Adversarial Networks with Limited Data,” *arXiv [cs.CV]*, 11-Jun-2020.

[45] R. A. Castellino, “Computer aided detection (CAD): an overview,” *Cancer Imaging*, vol. 5, no. 1, pp. 17–19, Aug. 2005.

[46] C. C. Bennett and K. Hauser, “Artificial intel-ligence framework for simulating clini-cal decision-making: a Markov decision process approach,” *Artif Intell Med*, vol. 57, pp. 9–19, 2013.

[47] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, “PH2-A Dermoscopic Image Database for Research andBenchmarking,” in *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine andBiology Society (EMBC)*, Osaka, Japan, 2013, pp. 5437–5440.

[48] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvehy, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and H. P. Soyer, “A patient-centric dataset of images and metadata for identifying melanomas using clinical context,” *Sci Data*, vol. 8, no. 1, p. 34, Jan. 2021.

[49] A. Boer and K. Nischal, “Com: A Growing Online Resource for Learning Dermatology and Dermatopathology,” *IndianJ. Derm. Venereol. Leprol*, vol. 73, 2007.

[50] “DermIS,” *DermIS.net*. [Online]. Available: https://www.dermis.net/dermisroot/en/. [Accessed: 19-Jan-2022].

[51] G. Rgenziano, H. P. Soyer, V. D. Giorgi, D. Piccolo, P. Carli, M. Delfino, A. Ferrari, R. Hofmann-Wellenhof, D. Massi, and G. Mazzocchetti, “Interactiveatlas of Dermoscopy,” *EDRA Medical Publishing and New Media*, 2000.

[52] Dermnet, “Dermnet: Dermatology Pictures - skin disease Pictures.” [Online]. Available: http://www.dermnet.com/dermatology-pictures-skin-disease-pictures/. [Accessed: 19-Jan-2022].

[53] D. Wen, S. M. Khan, A. Ji Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. de Blas Perez, A. K. Denniston, X. Liu, and R. N. Matin, "Characteristics of publicly available skin cancer image datasets: a systematic review," *The Lancet Digital Health*, vol. 4, no. 1, pp. e64–e74, Jan. 2022.

[54] M. Combalia, C. F. Noel, V. Codella, B. Rotemberg, V. Helba, O. Vilaplana, A. C. Reiter, and S. Halpern, *Josep Malvehy: "BCN20000: Dermoscopic Lesions in the Wild."* 2019.

[55] C. F. Noel, D. Codella, M. E. Gutman, B. Celebi, M. A. Helba, S. W. Marchetti, A. Dusza, K. Kalloo, N. Liopyris, H. Mishra, and A. Kittler, *Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration*. 2017.

[56] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," *arXiv [cs.CV]*, 04-May-2016.

[57] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," *arXiv [cs.CV]*, 13-Oct-2017.

[58] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, 2018.

[59] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, “Melanoma classification on dermoscopy images using a neural network ensemble model,” *IEEE Trans. Med. Imaging*, vol. 36, no. 3, pp. 849–858, Mar. 2017.

[60] W. F. Cueva, F. Munoz, G. Vasquez, and G. Delgado, “Detection of skin cancer ”Melanoma” through computer vision,” in *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, Cusco, 2017, pp. 1–4.

[61] S. Choudhari and S. Biday, “Artificial Neural Network for SkinCancer Detection,” vol. 2014, pp. 147–153.

[62] J. A. Jaleel, S. Salim, and R. B. Aswin, “Computer aided detection of skin cancer,” in *2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT)*, Nagercoil, 2013, pp. 1137–1142.

[63] L. Li, Q. Zhang, Y. Ding, H. Jiang, B. H. Thiers, and J. Z. J. B. M. I. Wang, “Automatic diagnosis of melanoma using machine learning methods on a spectroscopic system,” vol. 14, pp. 1–12, 2014.

[64] S. Gilmore, R. Hofmann-Wellenhof, and H. P. J. E. D. Soyer, “A support vector machine for decision support in melanoma recognition,” vol. 19, pp. 830–835, 2010.

[65] M. Rastgoo, “Classification of melanoma lesions using sparse coded features and random forests,” *p. 97850C: International Society for Optics and Photonics*, vol. 2016, 2016.

[66] P. M. Sajid and D. A. Rajesh, “Performance Evaluation of Classifiers for Automatic Early Detection of Skin Cancer,” *J. Adv. Res. Dyn.Control. Syst*, vol. 10, pp. 454–461, 2018.

[67] X. Liu, X. Wang, and S. Matwin, “Interpretable deep convolutional neural networks via meta-learning,” in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 2018, pp. 1–9.

[68] J. Tang, L. Jin, Z. Li, and S. Gao, “RGB-D object recognition via incorporating latent data structure and prior knowledge,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1899–1908, Nov. 2015.

[69] J. Hu, Z. Tang, K. Wang, L. Zhang, and Q. J. P. R. Zhang, “Deep learning for image-based cancer detection and diagnosis− A survey,” vol. 83, pp. 134–149, 2018.

[70] X. Meng, J. Chen, Z. Zhang, K. Li, J. Li, Z. Yu, and Y. Zhang, “Non-invasive optical methods for melanoma diagnosis,” *Photodiagnosis Photodyn. Ther.*, vol. 34, no. 102266, p. 102266, Jun. 2021.

[71] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional Neural Networks forMedical Image Analysis: Full Training or Fine Tuning?,” *IEEE Trans. Med. Imaging2016*, vol. 35, pp. 1299–1312.

[72] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, “Automated Melanoma Recognition in Dermoscopy Images via Very Deep ResidualNetworks,” *IEEE Trans. Med. Imaging2017*, vol. 36, pp. 994–1004.

[73] T. DeVries and D. Ramachandram, “Skin lesion classification using deep multi-scale convolutional neural networks,” *arXiv [cs.CV]*, 04-Mar-2017.

[74] D. B. Mendes and N. C. da Silva, “Skin lesions classification using convolutional neural networks in clinical images,” *arXiv [cs.CV]*, 05-Dec-2018.

[75] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, and M. Lee, “The skin cancer classification using deep convolutional neural network,” *Multimed. Tools Appl.*, vol. 77, no. 8, pp. 9909–9924, Apr. 2018.

[76] B. Harangi, A. Baran, and A. Hajdu, “Classification of skin lesions using an ensemble of deep neural networks,” *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2018, pp. 2575–2578, Jul. 2018.

[77] L. Nanni, A. Lumini, and S. Ghidoni, “Ensemble of deep learned features for Melanoma Classification,” *arXiv [cs.CV]*, 20-Jul-2018.

[78] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "'Data augmentationfor skin lesion analysis,'' inOR 2.0 Context-Aware Operating Theaters," in *Clinical Image-Based Procedures,and Skin Image Analysis*, Granada, Spain: Springer, 2018, pp. 303–311.

[79] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 1229–1233.

[80] A. Sagar and D. Jacob, "Convolutional neural networks for classifying melanoma images," *bioRxiv*, bioRxiv, p. 2020.05.22.110973, 23-May-2020.

[81] D. Połap, "An adaptive genetic algorithm as a supporting mechanism for microscopy image analysis in a cascade of convolution neural networks," *Appl. Soft Comput.*, vol. 97, p. 106824, Dec. 2020.

[82] M. Ahmad, C.-M. Usama, K. Huang, and M. S. Hwang, "'Discriminative feature learning for skin disease classi-fication using deep convolutional neural network,'" *IEEE Access*, vol. 8, pp. 39025–39033, 2020.

[83] A. A. Adegun and S. Viriri, "FCN-based DenseNet framework for automated detection and classification of skin lesions in dermoscopy images," *IEEE Access*, vol. 8, pp. 150377–150396, 2020.

[84] M. A. Al-Masni, D. H. Kim, and T. S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmenta-tion and classification," *Comput. Methods Programs Biomed*, vol. 190, no. 0169–2607, p. 105351, 2020.

[85] M. F. Jojoa Acosta, L. Y. Caballero Tovar, M. B. Garcia-Zapirain, and W. S. Percybrooks, "Melanoma diagnosis using deep learning techniques on dermatoscopic images," *BMC Med. Imaging*, vol. 21, no. 1, p. 6, Jan. 2021.

[86] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and Y. Duan, "Novel transfer learning approach for

medical imaging with limited labeled data," *Cancers (Basel)*, vol. 13, no. 7, p. 1590, Mar. 2021.

[87] Y. P. Liu, Z. Wang, Z. Li, J. Li, T. Li, P. Chen, and R. Liang, "Multiscale ensemble of convolutional neural networks for skin lesion classification," *IET Image Process*, vol. 15, pp. 2309–2318, 2021.

[88] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Comput. Med. Imaging Graph.*, vol. 88, no. 101843, p. 101843, Mar. 2021.

[89] X. Yi, E. Walia, and P. Babyn, "Unsupervised and semi-supervised learning with Categorical Generative Adversarial Networks assisted by Wasserstein distance for dermoscopy image Classification," *arXiv [cs.CV]*, 10-Apr-2018.

[90] C. Baur, S. Albarqouni, and N. Navab, "MelanoGANs: High Resolution Skin Lesion Synthesis with GANs," *arXiv [cs.CV]*, 12-Apr-2018.

[91] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv [cs.LG]*, 19-Nov-2015.

[92] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," *arXiv [cs.CV]*, 18-Jun-2015.

[93] C. Baur, S. Albarqouni, and N. Navab, "Generating highly realistic images of skin lesions with GANs," *arXiv [cs.CV]*, 05-Sep-2018.

[94] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," *arXiv [cs.NE]*, 27-Oct-2017.

[95] A. Bissoto, F. Perez, E. Valle, and S. Avila, "Skin Lesion Synthesis with Generative Adversarial Networks," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, 2018, pp. 294–302.

[96] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," *arXiv [cs.CV]*, 30-Nov-2017.

[97] H. Rashid, M. A. Tanveer, and H. Aqeel Khan, "Skin lesion classification using GAN based data augmentation," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2019, pp. 916–919, Jul. 2019.

[98] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A GAN-based image synthesis method for skin lesion classification," *Comput. Methods Programs Biomed.*, vol. 195, p. 105568, Oct. 2020.

[99] S. Ding, J. Zheng, Z. Liu, Y. Zheng, Y. Chen, X. Xu, J. Lu, and J. Xie, "High-resolution dermoscopy image synthesis with conditional generative adversarial networks," *Biomed. Signal Process. Control*, vol. 64, p. 102224, Feb. 2021.

[100] B. Ahmad, S. Jun, V. Palade, Q. You, L. Mao, and M. Zhongjie, *Improving Skin Cancer Classification Using Heavy-Tailed StudentT-Distribution in Generative Adversarial Networks (TED-GAN)*. 2011.

[101] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two Transformers Can Make One Strong GAN," *arXiv [cs.CV]*, 14-Feb-2021.

[102] J. Lin, Y. Li, and G. Yang, "FPGAN: Face de-identification method with generative adversarial networks for social robots," *Neural Netw.*, vol. 133, pp. 132–147, Jan. 2021.

[103] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, and A. Mallya, "Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications," *arXiv [cs.CV]*, 06-Aug-2020.

[104] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement*, vol. 152, p. 107377, Feb. 2020.

[105] P. Polewski, J. Shelton, W. Yao, and M. Heurich, "Segmentation of single standing dead trees in high-resolution aerial imagery with generative adversarial network-based

shape priors," *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLIII-B2-2020, pp. 717–723, Aug. 2020.

[106] P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, and F. Scarselli, "Image generation by GAN and style transfer for agar plate image segmentation," *Comput. Methods Programs Biomed.*, vol. 184, p. 105268, Feb. 2020.

[107] H. Zhang, Y. Sun, L. Liu, X. Wang, L. Li, and W. Liu, "ClothingOut: a category-supervised GAN model for clothing segmentation and retrieval," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4519–4530, May 2020.

[108] Y. Choi, H. Lim, H. Choi, and I.-J. Kim, "GAN-Based Anomaly Detection and Localization of Multivariate Time Series Data for Power Plant," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020, pp. 71–74.

[109] J. B. Abraham, "Improving Stock Price Prediction with GAN-based Data Augmentation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 4, no. 1, pp. 9–14, Jan. 2021.

[110] T. Wang, D. Trugman, and Y. Lin, "SeismoGen: Seismic waveform synthesis using GAN with application to seismic data augmentation," *J. Geophys. Res. [Solid Earth]*, vol. 126, no. 4, Apr. 2021.

[111] S. Motamed, P. Rogalla, and F. Khalvati, "RANDGAN: Randomized generative adversarial network for detection of COVID-19 in chest X-ray," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Apr. 2021.

[112] S. A. Kamran, K. F. Hossain, A. Tavakkoli, S. L. Zuckerbrod, K. M. Sanders, and S. A. Baker, "RV-GAN: retinal vessel segmentation from fundus images using multi-scale generative adversarial networks," *arXiv preprint arXiv:2101. 00535*, 2021.

[113] J. K. Dumagpi and Y.-J. Jeong, "Evaluating GAN-Based Image Augmentation for Threat Detection in Large-Scale Xray Security Images," *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, vol. 11, no. 1, p. 36, Dec. 2020.

[114] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "MedGAN: Medical image translation using GANs," *Comput. Med. Imaging Graph.*, vol. 79, no. 101684, p. 101684, Jan. 2020.

[115] D. Croce, G. Castellucci, and R. Basili, "GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2114–2119.

[116] J. Lu, K. Zhou, B. Sisman, and H. Li, "VAW-GAN for Singing Voice Conversion with Non-parallel Training Data," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 514–519.

[117] Y. Diao, L. Yang, X. Fan, Y. Chu, D. Wu, S. Zhang, and H. Lin, "AFPun-GAN: Ambiguity-Fluency Generative Adversarial Network for Pun Generation," in *Natural Language Processing and Chinese Computing*, 2020, pp. 604–616.

[118] D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[119] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv [stat.ML]*, 20-Dec-2013.

[120] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014, vol. 27, pp. 2672–2680.

[121] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *arXiv [cs.LG]*, 11-May-2021.

[122] L. Weng, "What are diffusion models?," 11-Jul-2021. .

[123] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," *arXiv [cs.LG]*, 30-Oct-2014.

[124] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using Real NVP,” *arXiv [cs.LG]*, 27-May-2016.

[125] Y. Song and D. P. Kingma, “How to Train Your Energy-Based Models,” *arXiv [cs.LG]*, 09-Jan-2021.

[126] Y. Lecun, S. Chopra, R. Hadsell, M. ’. A. Ranzato, and F. J. Huang, “A tutorial on energy-based learning,” 2006. [Online]. Available: http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf. [Accessed: 26-Apr-2022].

[127] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” *arXiv [stat.ML]*, 16-Jan-2014.

[128] M. Germain, K. Gregor, I. Murray, and H. Larochelle, “MADE: Masked Autoencoder for Distribution Estimation,” *arXiv [cs.LG]*, 11-Feb-2015.

[129] “The neural autoregressive distribution estimator H. Larochelle, I. Murray. International Conference on Artificial Intelligence and Statistics,” pp. 29–37, 2011.

[130] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” *arXiv [cs.CV]*, 25-Jan-2016.

[131] S. Mohamed and B. Lakshminarayanan, “Learning in implicit generative models,” *arXiv [stat.ML]*, 11-Oct-2016.

[132] Y. Song and S. Ermon, “Improved Techniques for Training Score-Based Generative Models,” *arXiv [cs.LG]*, 16-Jun-2020.

[133] S. Barratt and R. Sharma, “A Note on the Inception Score,” *arXiv [stat.ML]*, 06-Jan-2018.

[134] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” *arXiv [cs.LG]*, 26-Jun-2017.

[135] Z. Wang, Q. She, and T. E. Ward, “Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy,” *arXiv [cs.LG]*, 04-Jun-2019.

[136] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved Training of Wasserstein GANs,” *arXiv [cs.LG]*, 31-Mar-2017.

[137] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 2813–2821.

[138] A. Brock, J. Donahue, and K. Simonyan, “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” *arXiv [cs.LG]*, 28-Sep-2018.

[139] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[140] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic data augmentation using GAN for improved liver lesion classification,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, 2018, pp. 289–293.

[141] P. Schlegl, S. M. Seeb ock, G. Waldstein, and U. Langs, “f-anogan: Fast unsupervised anomaly detection with generativeadversarial networks,” *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[142] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” *arXiv [cs.CV]*, 21-Nov-2016.

[143] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *arXiv [cs.CV]*, 30-Mar-2017.

[144] M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017, pp. 700–708.

[145] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal Unsupervised Image-to-Image Translation,” *arXiv [cs.CV]*, 12-Apr-2018.

[146] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[147] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” *arXiv [cs.CV]*, 25-Aug-2016.

[148] S. Mo, M. Cho, and J. Shin, “Freeze the Discriminator: a Simple Baseline for Fine-Tuning GANs,” *arXiv [cs.CV]*, 25-Feb-2020.

[149] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” *arXiv [cs.CV]*, 03-Dec-2019.

[150] A. Karnewar and O. Wang, “MSG-GAN: Multi-scale gradients for generative adversarial networks,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 7796–7805.

[151] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 618–626.

[152] X. Gong and W. Yao, “Dermoscopy image classification based on StyleGANs and decision fusion,” *IEEE Access*, vol. 8, pp. 70640–70650, 2020.

[153] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” *arXiv [cs.CV]*, 23-Feb-2016.

[154] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-Excitation Networks,” *arXiv [cs.CV]*, 05-Sep-2017.

[155] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 2261–2269.

[156] S. P. Nahata, *‘Deep learning solutions for skin cancer detection and diagnosis,’’ in Machine Learning With Health Care Perspective*. Cham, Switzerland: Springer, 2020, pp. 159–182.

[157] R. Hesse, S. Schaub-Meyer, and S. Roth, “Fast axiomatic attribution for neural networks,” *arXiv [cs.LG]*, 15-Nov-2021.

[158] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via Gradient-based localization,” *arXiv [cs.CV]*, 07-Oct-2016.

[159] C. Metta, A. Beretta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, and F. Giannotti, “Explainable Deep Image Classifiers for Skin Lesion Diagnosis,” *arXiv [cs.CV]*, 22-Nov-2021.

[160] A. Radford, I. Sutskever, J. W. Kim, G. Krueger, and S. Agarwal, “CLIP: Connecting Text and Images,” *OpenAI*, 05-Jan-2021. [Online]. Available: https://openai.com/blog/clip/. [Accessed: 29-Jan-2022].

VITA AUCTORIS

NAME:	Shivang Rana
PLACE OF BIRTH:	Vadodara, Gujarat
YEAR OF BIRTH:	1996
EDUCATION:	Reliance English Medium School, Vadodara, Gujarat, 2002 – 2012 Parth School of Competition, Vadodara, Gujarat, 2012 – 2014 Nirma University, Institute of Technology, Gujarat, 2014 – 2018