ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Ethics in Artificial Intelligence

BIAS MITIGATION IN SKIN DISEASE CLASSIFICATION

CANDIDATE
Chiara Bellatreccia

SUPERVISOR
Prof. Roberta Calegari

CO-SUPERVISOR
Andrea Borghesi

Academic year 2023-2024
Session 5th

I would rather be a cyborg than a goddess.
Donna Haraway

Abstract

The use of artificial intelligence (AI) in diagnosing skin diseases presents a significant opportunity to enhance healthcare accessibility. However, the effectiveness of AI-based diagnostic systems is often compromised by several challenges, particularly those related to fairness and representation. One prominent issue is the limited diversity in real-world datasets, which can lead to substantial classification biases. This study addresses these challenges by analyzing a dataset collected from an Italian hospital. The dataset exhibits limited data availability, resulting in inadequate representation—especially for darker skin tones. Furthermore, the dataset primarily consists of non-dermoscopic, consumer-grade images, which often suffer from quality issues such as inconsistent lighting and blurriness. These factors collectively complicate the development of accurate and fair AI models for skin disease diagnosis.

To address these issues, this research proposes a novel diagnostic pipeline designed to improve both accuracy and fairness in real-world scenarios. The proposed pipeline consists of two main stages: (1) data pre-processing and augmentation, wherein images that better represent darker skin tones are generated using a state-of-the-art diffusion model, and (2) disease classification through deep learning techniques. The efficacy of the proposed methodology is demonstrated through comprehensive validation on real-world data, highlighting significant improvements in both reliability and fairness across various skin disease classifications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter aims to provide a general introduction to the Thesis. Specifically, it outlines the problem under investigation, reviews previous approaches, and presents the adopted methodology, comparing it with prior work.

The Chapter is structured as follows:

- **Section 1.1** discusses the issue of AI fairness in dermatology from a general perspective, highlighting its significance and referencing relevant literature that substantiates the existence of this problem.
- **Section 1.2** provides an overview of previous approaches in the literature, analyzing their limitations and key strengths.
- **Section 1.3** details the methodology adopted in this study, explaining why it is well-suited for this specific use case and how it can serve as a generalizable pipeline.

1.1 Problem Description

Skin diseases are among the most prevalent human health conditions, affecting nearly 900 million people globally at any given time [33]. Early and accurate diagnosis is critical for effective treatment, yet access to dermatological care remains limited in many regions. Automating the diagnostic process through artificial intelligence (AI) offers the potential to make healthcare more accessible, especially for underserved populations.

Advances in deep learning (DL) have significantly improved diagnostic accuracy and reliability [57, 17, 47], but addressing biases and ensuring fairness in AI systems remain significant challenges.

A major hurdle is the use of non-dermoscopic images captured with consumer-grade cameras. While this approach democratizes access by relying on widely available tools, it introduces variability in image quality, lighting, and focus, complicating disease classification [36]. Additionally, diagnostic performance often varies across demographic groups, potentially disadvantaging underrepresented populations, such as those with darker skin tones. Ensuring fairness requires accurate evaluation of model performance for each demographic group and targeted methodologies for bias detection and mitigation [64].

To address these issues, this study proposes a pipeline to mitigate the bias detected in the classification of dermatological diseases within a non-dermoscopic dermatology dataset by augmenting it with synthetic images. Specifically, the dataset utilized in this study is imbalanced in terms of skin color, predominantly featuring images of diseases on caucasian skin, which introduces bias in classification. Another key characteristic of this dataset is the absence of skin color labels, necessitating an automatic measurement of skin tone to assess the presence of potential bias.

The issue of skin color imbalance in datasets is highly prevalent in dermatology, affecting both dermoscopic and non-dermoscopic datasets. For instance, Forero et al. [45] demonstrated through an in-depth exploration of the HAM10000 dataset [63] that fewer than 5% of images originate from black patients. Similarly, Alipour et al. [3] conducted a review of major dermatology datasets, emphasizing the necessity of greater efforts to ensure diversity within them.

As evidenced by several previous studies, an imbalanced dataset in terms of skin color induces bias in the classification of dermatological lesions and diseases, often favoring lighter skin tones at the expense of darker ones. For example, Bencevic et al. [10] found a significant correlation between image segmentation performance and skin color, highlighting a notable bias against darker skin tones. Daneshjou et al. [20] developed the Diverse Dermatology Images (DDI) dataset—the first publicly available, expertly curated, and pathologically confirmed image dataset with diverse skin tones—and evaluated state-of-the-art AI models on it. Their findings indicate that these AI models

exhibit substantial limitations on this dataset, particularly disadvantaging darker skin tones and less common diseases. Furthermore, Diaz et al. [22] conducted a systematic literature review to highlight the extent of bias present in clinical datasets. Their results reveal that many imaging datasets underrepresent certain skin tones, leading machine learning models to be trained primarily on images of individuals with lighter skin.

Additionally, robust skin tone estimation is crucial for assessing fairness in classification. The study by Kalb et al. [37] highlights inconsistencies in skin color estimation across prior works in the literature. Such inconsistencies can potentially compromise fairness verification, as demographic groups may not be accurately categorized as required.

In conclusion, **ensuring fairness in classification across different demographic groups is a critical and pressing issue in automated dermatological analysis**, particularly for clinical datasets, where models trained on dermoscopic datasets exhibit significant performance gaps [27]. This study proposes a potential pipeline to mitigate bias in the classification of a non-dermoscopic dataset, as introduced in the subsequent sections of this Chapter.

1.2 Related Work

Significant efforts have been made to mitigate bias in dermatological AI applications without compromising the privacy or integrity of demographic data. For instance, the study by Chiu et al. [16] introduces a method to ensure fairness by enhancing feature selection during the model training phase, purposely omitting sensitive demographic attributes. This technique relies on sophisticated feature entanglement strategies to focus solely on disease-relevant features, minimizing biases associated with non-disease attributes like skin tone. Moreover, the introduction of PatchAlign, as discussed in [1], marks a notable advancement in aligning skin condition image patches with corresponding clinical descriptions. Using a Masked Graph Optimal Transport (MGOT) algorithm effectively reduces noise and improves diagnostic accuracy and fairness across various skin tones by focusing on disease-relevant image regions. The work of Yuan et al. [67] presents EDGEMIXUP, a preprocessing technique that alters image data to diminish bias by manipulating colour saturation and integrating edge detection outputs. This

method has shown efficacy in decreasing the performance disparity between different skin tones while maintaining overall diagnostic accuracy. Similarly, the FairSkin framework introduced in [68] leverages diffusion models to generate synthetic medical images that represent various skin tones equitably. Through a resampling mechanism and class diversity loss, this approach ensures that the synthetic data aids in balancing dataset representation across demographic groups. Lastly, [31] and [42] propose innovative solutions to enhance fairness through structural model adjustments. The FairQuantize methodology employs weight quantization to adjust model performance across different demographics, and the channel pruning approach identifies and reduces bias by pruning channels that disproportionately affect specific demographic groups.

While the related works present innovative solutions for addressing bias and achieving acceptable accuracy in AI-based diagnostics for skin diseases, these solutions are still largely explorative and preliminary, rather than robust solutions to be applied in real-world scenarios. When applied to real-world scenarios, particularly employing non-dermoscopic images, they often yield unsatisfactory results [66]. When used in our specific scenario, the existing techniques still pose significant challenges that frequently lead to suboptimal outcomes if these techniques are applied in isolation [29].

1.3 Our Approach

It is worth emphasizing that the dataset used in our study introduces several unique challenges that must be responsibly addressed. The main challenges are related to (1) inherent dataset features (including its characteristics and variability), and (2) specific challenges related to the skewness of the available data, which significantly over-represent certain populations, thus inducing unfairness in the classification process (more details follow in the data description Section in Chapter 3). Failing to meticulously study and address these issues within the development pipeline could lead to misdiagnoses, which in turn may exacerbate existing healthcare inequalities and result in adverse outcomes for affected patients. Such oversight highlights the critical need for rigorous evaluation and refinement of AI diagnostic tools to prevent potential harm and ensure their reliability and fairness across all populations.

This work builds upon existing state-of-the-art foundational efforts, intending to address additional limitations in real-world, highly imbalanced datasets. We explicitly consider both classification performance and fairness metrics in our analysis. There are a few methods in the literature that aim at improving the fairness of non-dermoscopic image disease classification through the refinement of sophisticated Deep Learning (DL) models ([18, 42, 31, 16]) – this approach is *orthogonal* as it does not focus on the classification model itself but rather proposes a pipeline for image data pre-processing and data augmentation that can *complement* any existing DL model for classification of skin diseases. In particular, the pre-processing technique employs the Individual Typology Angle (ITA) metric along with a novel thresholding method based on a Gaussian Mixture Model to accurately measure the skin tone depicted in each image. For data augmentation, this study proposes a novel combination of stable diffusion with DreamBooth to address the challenge of data scarcity, which is particularly acute for darker skin tones. To the best of our knowledge, this is the first work to consider using DreamBooth for generating skin disease images for different skin shades. The pre-processing method can be affected by issues such as poor lighting and image blurriness, which may distort the perceived skin tone. To counteract these problems, the images used for training DreamBooth are carefully hand-picked, ensuring that they represent the skin tones targeted for augmentation. This meticulous selection process is especially crucial as only three out of the nine diseases catalogued in our dataset have examples of 'dark' and 'brown' skin, necessitating precise and representative training data to enhance model fairness and accuracy. The final step is the training of DL models for skin disease classification using pre-processed and augmented data. The current study opted for two of the most efficient models currently available, namely the Swin Transformer (ST) and the Convolutional Neural Network (CNN); potentially, other DL approaches could be plugged in, according to the available resources and desired outcomes. The overall pipeline is illustrated in Fig. 1.1 and consists of the previously discussed preprocessing steps, plus the comparison of enhanced results via data augmentation. Please note that the proposed pipeline requires co-design and co-creation phases (especially in the selection phase during the pre-processing), during which stakeholders (in this case, doctors) are involved to assist in the selection and validation processes.

This Thesis is organized as follows:

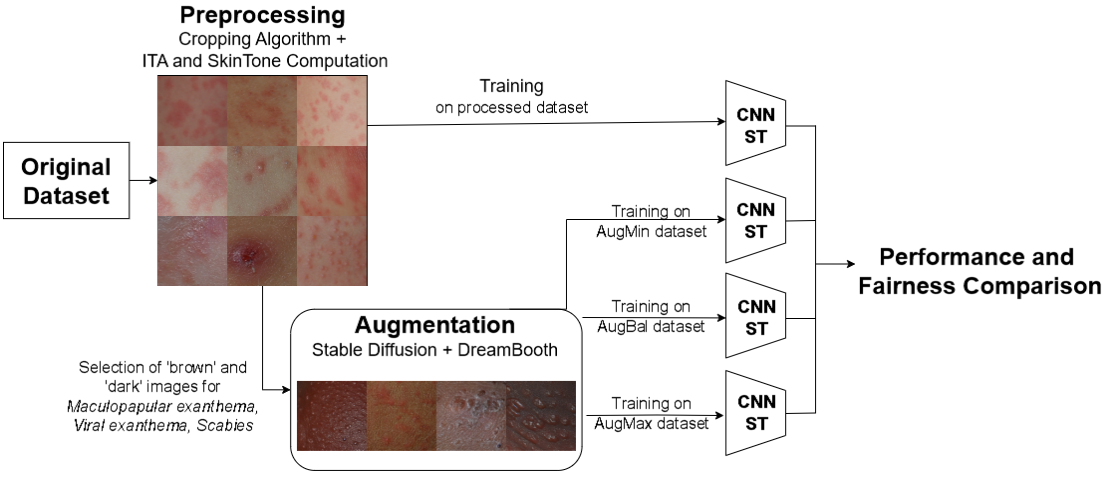


Figure 1.1: Diagram of our pipeline.

- **Chapter 2** introduces the theoretical background necessary for understanding this work. Specifically, it details the architecture of the Swin Transformer (ST) and explains the concept of Diffusion Models, followed by an introduction to Stable Diffusion and, finally, the DreamBooth fine-tuning technique.

- **Chapter 3** provides a detailed explanation of the examined dataset, highlighting its critical characteristics, explaining the adopted preprocessing methods, and describing the process of estimating the skin tone of the images using the ITA metrics.

- **Chapter 4** describes the classification process of various diseases within the dataset using a Convolutional Neural Network (CNN), explaining how it was trained on the given dataset. Additionally, this Chapter presents the classification results on the original dataset in terms of both performance and fairness.

- **Chapter 5** outlines the classification process of various diseases within the dataset using an ST, explaining how it was trained on the given dataset. This Chapter also presents the classification results on the original dataset in terms of both performance and fairness.

- **Chapter 6** details the image generation process through the fine-tuning of a Stable

Diffusion model using DreamBooth. It then explains the different approaches used for dataset augmentation and the classification results – both for the CNN and ST – on the augmented dataset with synthetic images, focusing on fairness and performance, with the aim of comparing these results to those on the original dataset from previous Chapters 4 and 5.

- **Chapter 7** presents the conclusions and discusses possible extensions or improvements to this work.

Chapter 2

Background

This Chapter provides a theoretical foundation for the key concepts necessary to comprehend the pipeline developed in this work. Specifically, the Chapter focuses on the theoretical explanation of the ST, employed for image classification, and Diffusion Models, first offering a general introduction and subsequently concentrating on Stable Diffusion and, finally, DreamBooth. At the end of each theoretical section, a brief discussion on its application in the dermatological field and its role in our work is provided.

This Chapter is structured as follows:

- **Section 2.1** presents previous studies on the use of CNNs in dermatology, highlighting their advantages over traditional automated diagnosis methods, as well as their limitations compared to other architectures such as Vision Transformers (ViTs) and STs.
- **Section 2.2** introduces the ST architecture in detail, explaining its advantages over traditional ViTs and discussing prior research that has employed STs or modified versions thereof in dermatology.
- **Section 2.3** provides a general introduction to Diffusion Models, explaining their forward process, reverse process, loss function, and typical architectural components.
- **Section 2.4** introduces Stable Diffusion and its architecture, emphasizing its advantages over traditional Diffusion Models and listing previous studies that have

utilized it in dermatology.

- **Section 2.5** presents the DreamBooth fine-tuning technique and its specialized loss function, concluding with a discussion of prior research that has applied this technique in combination with Stable Diffusion for dermatological applications.

2.1 Convolutional Neural Networks in Dermatology

Before the adoption of **Convolutional Neural Networks (CNNs)** in dermatology, automated diagnosis methods primarily relied on traditional machine learning techniques and image processing algorithms. These approaches required manual feature extraction, where handcrafted features such as texture, color, and shape were selected to characterize dermatological lesions. Among the most commonly employed techniques were **Support Vector Machines (SVMs)**, which were widely used for both binary and multiclass classification tasks, **Random Forests (RFs)**, which leveraged decision tree ensembles to improve robustness, and **k-Nearest Neighbors (k-NN)**, which classified lesions based on their proximity to labeled examples in the feature space. Additionally, rule-based systems and thresholding algorithms were used to segment and classify skin lesions based on predefined criteria, but these methods often struggled with variations in illumination, skin tone, and lesion morphology. Other approaches, such as **Principal Component Analysis (PCA)** and **Linear Discriminant Analysis (LDA)**, were occasionally employed for dimensionality reduction and feature selection, but their effectiveness was constrained by the quality of the manually extracted features. Overall, these early methods lacked the adaptability and scalability necessary for large-scale dermatological applications, as discussed in [38]:

- **Manual Feature Extraction:** Traditional methods often rely on manually extracted features such as texture, color, and shape, which are then used to classify skin lesions.

- **Rule-Based Systems:** These approaches use predefined rules and thresholds to identify and classify dermatological diseases, making them inflexible and unable to capture complex patterns.

- **Lower Accuracy and Human Error:** Since these methods depend on manual feature extraction, they often exhibit lower accuracy and are more susceptible to human error.
- **Limited Scalability:** Traditional techniques require human intervention for each analysis, making them inefficient and unsuitable for large-scale applications.

In 2017, Esteva et al. [24] utilized a CNN for the classification of skin cancer, demonstrating its potential in dermatological diagnosis. This study marked a pivotal moment in the field, as it became evident that Deep Learning models could achieve significantly higher reliability and accuracy compared to any previously employed automated techniques. The superior performance of CNNs in capturing complex patterns and features directly from raw image data underscored their transformative role in dermatological image analysis, setting the foundation for subsequent advancements in AI-driven skin disease classification. CNNs offer several advantages over traditional methods:

- **Automated Feature Learning:** CNNs learn features automatically from data during the training process, eliminating the need for manual feature engineering.
- **High Accuracy:** CNNs achieve superior accuracy in skin lesion classification due to their ability to capture intricate patterns. In some cases, their diagnostic performance is comparable to, or even exceeds, that of board-certified dermatologists, as further explained in the papers cited below.
- **Scalability:** CNNs can efficiently process large datasets, making them highly scalable and suitable for large-scale applications.
- **Adaptability:** CNNs can be trained on diverse datasets, allowing them to generalize to various skin diseases and improve over time with the inclusion of new data.

For instance, **Ganthya et al.** [28] explore the application of CNNs for automated skin cancer diagnosis, covering key aspects such as convolutional layers, pooling layers, activation functions, and backpropagation algorithms. Their study also addresses dataset preparation, including image preprocessing and augmentation techniques to enhance model performance.

Wong et al. [65] employ a CNN model trained on real-world smartphone images with histopathological ground truth for binary lesion classification (benign vs. malignant). Results indicate that the model's accuracy in predicting malignant lesions was comparable to that of board-certified dermatologists (71.31% vs. 77.87%, 69.88%, and 71.93%, respectively), validating the clinical utility of automated diagnosis.

Additionally, **Musthafa et al.** [44] employ a CNN with an optimized layer configuration and data augmentation for skin cancer diagnosis using the HAM10000 dataset, achieving a remarkable accuracy of 97.78%.

These studies highlight that CNN architectures exhibit performance comparable to, and in some cases exceeding, that of board-certified dermatologists for automated dermatological diagnosis. Future advancements in dermatological AI will likely involve a hybrid approach that integrates automated diagnosis with expert clinical evaluation, as supported by **Ba et al.** [7], which examines the impact of CNN-assisted diagnosis on dermatologists' performance. The study shows that dermatologists, particularly those with less experience, benefit from CNN assistance in terms of diagnostic accuracy.

Despite their widespread adoption in dermatology, CNNs have several **limitations** when compared to more recent models such as **Vision Transformers (ViTs)** and **Swin Transformers**. Firstly, CNNs primarily focus on local feature extraction and may fail to capture crucial global contextual information necessary for accurate diagnosis. In contrast, Vision Transformers—especially Swin Transformers—utilize self-attention mechanisms to integrate global contextual information, enabling superior pattern recognition, as discussed in [4]. Secondly, training CNNs on large-scale datasets is computationally expensive and time-consuming. Swin Transformers, in contrast, are designed for computational efficiency through hierarchical representations and window-based self-attention, making them more suitable for large-scale applications. Finally, multiple studies have demonstrated that Vision Transformers outperform CNNs on dermatological datasets, as shown in [56] and in [61].

In this study, a simple Convolutional Neural Network is initially used for the classification of nine different skin diseases, achieving an overall accuracy of 77.0% and an F1-score of 0.77 on the non-augmented dataset described in Chapter 4. To enhance performance, a Swin Transformer is subsequently employed, attaining an overall accuracy of 91.3% and an F1-score of 0.90 on the non-augmented dataset. These results further

confirm the superiority of Swin Transformers for this dermatological classification task.

2.2 Swin Transformer: Architecture and Applications in Dermatology

For a long time, modeling in computer vision has been dominated by CNNs, with architectures such as AlexNet demonstrating revolutionary performance on ImageNet. In contrast, the evolution of architectures in the field of natural language processing (NLP) has followed a different trajectory, where the predominant architecture today is the **Transformer**. Designed for sequential modeling, Transformers leverage attention mechanisms to capture **long-range dependencies** in data. Given their remarkable success in NLP, researchers have explored their application to computer vision, starting with ViTs and more recently with the **Shifted Window Transformer**, i.e. the Swin Transformer.

However, significant challenges arise when transferring the performance of Transformers from a text-based domain to an image-based one. The first major challenge is scale: unlike word tokens, which serve as the fundamental elements in NLP Transformers, the scale of visual elements can vary significantly, as is the case in object detection. Another challenge stems from the substantially higher pixel resolution of images compared to the word-level representations in text sequences. This issue is particularly critical for tasks such as semantic segmentation, which requires dense, pixel-level predictions. The quadratic computational complexity of self-attention with respect to image size makes it impractical for high-resolution images.

To address these challenges, Liu et al., in their paper from which this introduction is adapted [43], proposed a general-purpose Transformer backbone known as the Swin Transformer. **This architecture constructs a hierarchical feature map that maintains computational complexity linear in image size**, enabling more efficient and scalable modeling for vision tasks.

2.2.1 Architecture

The Swin Transformer is a hierarchical architecture whose representation is computed through a **shifting window** mechanism. This design enhances computational efficiency by restricting self-attention computation to non-overlapping local windows while simultaneously enabling cross-window connections. The Swin Transformer constructs a hierarchical representation by starting from small-sized patches and progressively merging neighboring patches at deeper layers, providing the flexibility to model visual information at multiple scales. Due to the computation of self-attention within each local non-overlapping window, its computational complexity remains linear with respect to image size.

The architecture is illustrated in Figure 2.1: First, the input RGB image is split into non-overlapping patches using a **Patch Partition** module, similar to ViTs. Each patch is treated as a ”token,” whose feature representation consists of the concatenation of raw pixel RGB values. In its original implementation, a patch size of 4×4 is used, resulting in each patch having a feature dimension of 4×4×3=48. Next, multiple stages are applied to these patch tokens. The main component of these stages is the **Swin transformer Block**, shown in Figure 2.2.

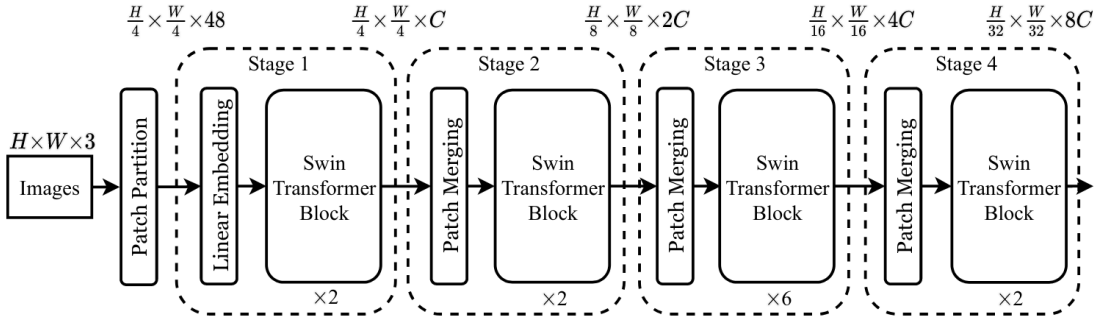


Figure 2.1: Swin Transformer Architecture [43].

Within a **Swin Transformer Block**, the standard multi-head self-attention (MSA) used in ViTs is replaced by a Window-based MSA (W-MSA) and a Shifted Window MSA (SW-MSA). Each block consists of two sub-units: each sub-unit includes a Layer Normalization (LN) + Attention module, followed by another LN + Multi-Layer Perceptron (MLP) layer, as shown in Figure 2.2. The first sub-unit utilizes W-MSA, while

the second sub-unit employs SW-MSA. Specifically:

- Unlike traditional ViTs, which apply global self-attention where each patch attends to all other patches—resulting in quadratic complexity with respect to the number of patches—**Window-based MSA** (Figure 2.3a) operates within fixed-size windows, where each window contains a fixed number of patches (M×M in the original paper). Self-attention is then computed only within each window, leading to linear complexity with respect to the number of patches. In particular, the computational complexity of W-MSA for an image with $h \times w$ is $\Omega(W\text{-}MSA) = 4hwC^2 + 2M^2(hw)C$, where M is the aforementioned number of patches and C is the third dimension of the feature map after the Linear Embedding in the first stage. This complexity is linear to patch number hw, when M is fixed. For the sake of comparison, it is worth noting that the computational complexity of a standard of the standard MSA in ViT is quadratic with respect to the number of patches, i.e. $\Omega(MSA) = 4hwC^2 + 2(hw)^2C$.

- However, if only W-MSA were used, relationships between different windows would be missing, which would be a limitation for the model. To address this, **Shifted Window-based Self-Attention (SW-MSA)** is introduced. SW-MSA takes the output of W-MSA, shifts all windows by (M/2, M/2) relative to the previous layer, and then applies W-MSA within the shifted windows. However, this shift results in the presence of "orphan" patches that do not belong to any window, as well as windows with incomplete patches. To handle this issue, the Swin Transformer employs a "Cyclic Shift" technique (Figure 2.3b), which moves orphan patches into windows with missing patches. After this shift, a window consists of patches that are no longer adjacent in the original feature map. To ensure attention is limited to adjacent patches, a masked MSA is applied during computation.

The original Swin Transformer architecture, represented in Figure 2.1, consists of four stages, each containing a Swin Transformer Block:

- **Stage 1:** Each input patch, initially of size 48 pixels, is projected into a feature dimension of C using a **Linear Embedding layer**. The resulting feature map is then processed through a Swin Transformer Block, producing an output with dimensions $\frac{H}{4} \times \frac{W}{4} \times C$.

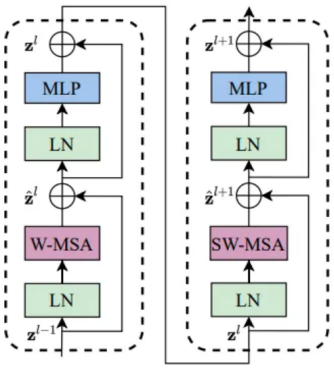
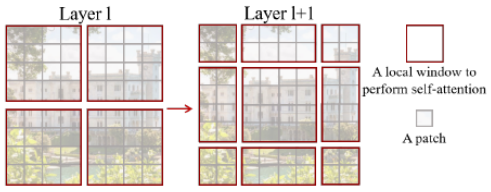
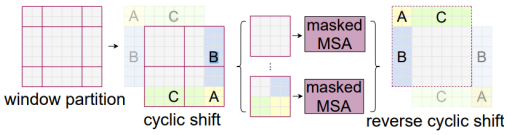


Figure 2.2: Swin Transformer Block [43].



(a) An illustration of the Window-based Multi-Head Self-Attention mechanism [43].



(b) Cyclic shift technique [43].

Figure 2.3

- **Stage 2:** The feature map of size $\frac{H}{4} \times \frac{W}{4} \times C$ is passed through a **Patch Merging layer** (Figure 2.4), which combines adjacent 2×2 patches into a single patch. This operation effectively downsamples the resolution by a factor of 2 while increasing the feature map dimension by a factor of 2. Unlike convolutional downsampling, the Patch Merging layer groups adjacent $n \times n$ patches and concatenates them depth-wise. Specifically, to downsample an input feature map by a factor of n, the input is first divided into groups, where each group consists of $n \times n$ adjacent patches. These groups are then concatenated along the channel dimension. As a result, an input feature map of size $H \times W \times C$ is transformed into $\frac{H}{n} \times \frac{W}{n} \times (2nC)$, as illustrated in Figure 2.4. In this case, the output dimension becomes $\frac{H}{8} \times \frac{W}{8} \times 2C$. After the Patch Merging layer, the feature map is processed by another Swin Transformer Block, maintaining the same output dimensions.

- **Stages 3 and 4:** The same procedure as Stage 2 is applied, further downsampling

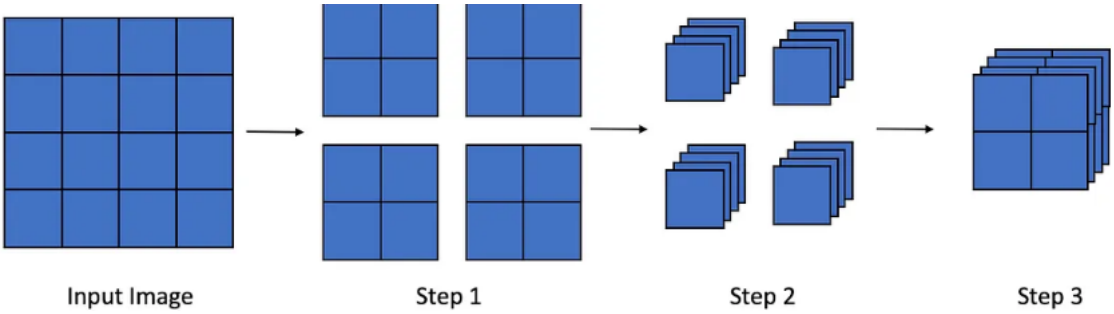


Figure 2.4: Downsampling operation in the Patch Merging layer of a Swin Transformer stage [39]. Assuming n=2, each group consists of 2x2 neighbouring patches. First, the input image is split into groups of 2x2. Later, the patches in each group are stacked depth-wise. Finally, the last step involves combining the stacked group.

the resolution, resulting in output dimensions of $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively.

These stages construct a hierarchical representation, maintaining a feature map resolution similar to convolutional neural networks such as VGGNet and ResNet. This characteristic allows the Swin Transformer to conveniently replace conventional backbone networks in existing vision tasks.

Finally, Swin Transformers include the **Relative Position Bias** B of size $(M^2 \times M^2)$ for calculating self-attention, i.e.:

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V, \tag{2.1}$$

where Q, K and V are the query, key, and value matrices; d is the query/key dimension, and M^2 is the number of patches in a window. Using relative position bias significantly improves performance over transformers that use absolute position embedding.

2.2.2 Swin Transformers in Dermatology

Swin Transformers have been increasingly utilized in dermatology due to their ability to capture multi-scale features, making them highly effective for skin disease classification and lesion segmentation. Their hierarchical structure and attention mechanisms enable superior performance compared to traditional Convolutional Neural Networks and Vision Transformers, particularly in tasks requiring fine-grained pattern recognition across

varying lesion sizes and shapes.

For instance, **Zhong et al.** [69], conducted empirical studies which demonstrate that the Swin Transformer Large model achieves the highest classification performance, with an accuracy of 93.5%, outperforming multiple ResNet variants.

Similarly, **Tang et al.** [62] introduces the **SkinSwinViT** methodology, which leverages cross-window attention within the Swin Transformer architecture for improved skin lesion classification. Experimental results highlight its superiority, achieving an impressive accuracy of 97.88%.

Moreover, modified versions of the Swin Transformer have shown remarkable results, surpassing both conventional Vision Transformers and classical CNNs. For instance, **Pacal et al.** [48] implement a modified Swin Transformer incorporating Hybrid Shifted Window-Based Multi-Head Self-Attention (HSW-MSA) and replacing the standard Multi-Layer Perceptron (MLP) with a SwiGLU-based MLP. This optimization enhances accuracy, training speed, and parameter efficiency. The modified Swin Transformer achieves an outstanding classification accuracy of 89.36% and an F1-score of 86.65% on the ISIC 2019 test set, surpassing all prior research and deep learning models documented in the literature.

In the study of **Paraddy et al.** [49], a Convolutional Swin Transformer (**CSwinformer**) is employed for precise skin lesion segmentation and classification. This approach introduces Swinformer-Net, which integrates a Swin Transformer with a U-Net architecture for accurate region of interest delineation. In the final classification phase, the segmented output is fed into a Multi-Scale Dilated Convolutional Neural Network meets Transformer (MD-CNNFormer) module. The model achieves an accuracy exceeding 95% across four benchmark dermatological datasets—HAM10000, ISBI 2016, PH2, and Skin Cancer ISIC.

These studies demonstrate that the Swin Transformer is particularly well-suited for dermatological applications, as it effectively captures features at multiple scales, outperforming convolutional architectures, ResNet models, and Vision Transformers for tasks such as lesion segmentation, disease classification, and feature extraction.

This work employs a Swin Transformer in its standard configuration, as implemented by Hugging Face. Additionally, the utilized model has been fine-tuned on a skin cancer

dataset, contributing to its superior performance on the specific dermatological task addressed in this study. The results indicate that even without synthetic data augmentation generated via Stable Diffusion, the Swin Transformer achieves a classification accuracy exceeding 90% and an F1-score of 0.90, further validating its suitability for this task.

2.3 A General Introduction to Diffusion Models

In recent years, **Diffusion Models** [59] have emerged as a powerful class of generative models within the field of Machine Learning. Also referred to as *diffusion probabilistic models* or *score-based generative models*, these techniques leverage a stochastic process to model complex data distributions. As latent variable generative models, Diffusion Models aim to establish a probabilistic relationship between observable data and underlying latent variables, enabling the **generation of new data samples that closely resemble the original dataset**.

A Diffusion Model operates through three fundamental components: the forward process, the reverse process, and the sampling procedure. The **forward process** systematically corrupts data by iteratively adding noise, transforming structured information into a distribution that approximates pure Gaussian noise. Conversely, the **reverse process**—parametrized by a neural network—learns to progressively denoise the corrupted data, effectively reconstructing realistic samples. The **sampling procedure** allows for the generation of new data points by initializing a noisy input and iteratively refining it through the learned reverse process.

Diffusion Models have demonstrated remarkable success in a variety of computer vision applications. By 2024, they have become the foundation for state-of-the-art techniques in image denoising, inpainting, super-resolution, and high-fidelity image and video synthesis. The core principle behind their success lies in their ability to learn the distribution of complex datasets and generate new instances with superior quality compared to previous generative approaches. Unlike Generative Adversarial Networks (GANs), which often suffer from mode collapse and training instability, diffusion models offer greater diversity in sample generation and improved training robustness.

Given their versatility and increasing adoption, diffusion models continue to push the boundaries of generative AI, finding applications in domains beyond computer vision,

including natural language processing, drug discovery, and medical imaging. Their ability to produce highly detailed and realistic outputs positions them as a leading approach in modern generative modeling. The following sections will delve into their theoretical foundations, training methodologies, and recent advancements, highlighting their growing impact on scientific research with a focus on the dermatological field.

2.3.1 Forward Process [55]

Given an initial image x_0, the forward process progressively adds stochastic noise to it to create a chain $x_1 ... x_T$ of noisy versions (i.e. latent vectors) of the image, as represented in Figure 2.5. The forward process is stochastic, but it does not have learnable parameters - only hyper-parameters. It is modelled as the blending between the (attenuated) previous latent and standard Gaussian noise, i.e.

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\epsilon_t \tag{2.2}$$

where $\epsilon_t \sim N(0; I)$. Sampling a vector x from a Gaussian distribution with mean μ and covariance matrix Σ can be realized by sampling $y \sim N(0; I)$ from a standard Gaussian and then setting $x = \mu + \Sigma^{1/2} y$. If $\Sigma = \sigma^2 I$, $\Sigma^{1/2} = \sigma I$. Hence, from a probability point of view, the latent vector at each time-step follows a multivariate Gaussian distribution with mean $\sqrt{1-\beta_t}\, x_{t-1}$ and variance β_t, i.e.

$$x_t \sim q(x_t|x_{t-1}) = N(\sqrt{1-\beta_t}\, x_{t-1}; \beta_t I) \tag{2.3}$$

which represents a Markov chain, since x_t depends only on x_{t-1}.
The hyper-params $\beta_1 ... \beta_T$ are the **noise schedule**. It is possible to directly generate a latent vector at each timestep t given x_0. Indeed, knowing that

$$x_1 = \sqrt{1-\beta_1}\, x_0 + \sqrt{\beta_1}\epsilon_1 \text{ where } \epsilon_1 \sim N(0; I)$$
$$x_2 = \sqrt{1-\beta_2}\, x_1 + \sqrt{\beta_2}\epsilon_2 \text{ where } \epsilon_2 \sim N(0; I)$$

one can compute

$$\begin{aligned}
x_2 &= \sqrt{1-\beta_2}\left(\sqrt{1-\beta_1}\,x_0 + \sqrt{\beta_1}\,\epsilon_1\right) + \sqrt{\beta_2}\,\epsilon_2 \\
&= \sqrt{1-\beta_2}\sqrt{1-\beta_1}\,x_0 + \sqrt{(1-\beta_2)\beta_1}\,\epsilon_1 + \sqrt{\beta_2}\,\epsilon_2 \\
&= \sqrt{1-\beta_2}\sqrt{1-\beta_1}\,x_0 + \sqrt{(1-\beta_2)\beta_1 + \beta_2}\,\epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0;I) \\
&= \sqrt{1-\beta_2}\sqrt{1-\beta_1}\,x_0 + \sqrt{(1-\beta_2)\beta_1 + \beta_2 - 1 + 1}\,\epsilon \\
&= \sqrt{1-\beta_2}\sqrt{1-\beta_1}\,x_0 + \sqrt{(1-\beta_2)\beta_1 - (1-\beta_2) + 1}\,\epsilon \\
&= \sqrt{(1-\beta_2)(1-\beta_1)}\,x_0 + \sqrt{1-(1-\beta_1)(1-\beta_2)}\,\epsilon.
\end{aligned}$$

In general, it holds that

$$x_t = \sqrt{\Pi_{i=1}^{t}(1-\beta_i)}\,x_0 + \sqrt{1-\Pi_{i=1}^{t}(1-\beta_i)}\,\epsilon \text{ where } \epsilon \sim N(0;I) \tag{2.4}$$

By setting $\alpha_t = \Pi_{i=1}^{t}(1-\beta_i)$, we obtain

$$x_t = \sqrt{\alpha_t}\,x_0 + \sqrt{1-\alpha_t}\,\epsilon \text{ where } \epsilon \sim N(0;I) \tag{2.5}$$

meaning that x_t follows a normal distribution, i.e. $x_t \sim q_{t|0}(x_t|x_0) = \mathcal{N}(\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$. Note that, since $\beta_t < 1$, $\lim_{t\to+\infty}\alpha_t = 0$ and only noise remains in the latent vectors. In practice, already with sufficient steps T (usually 1000), all traces of the original data are removed, and $q(x_T|x_0) = q_T(x_T) = \mathcal{N}(0;I)$. The forward process transforms an arbitrary complex distribution $q_0(x)$, e.g. the distribution of real images $p_{real}(x)$, into a standard Gaussian $q_T(x)$.

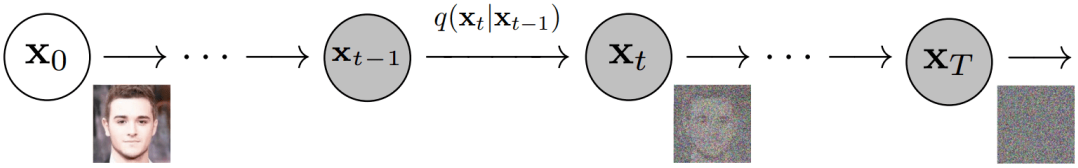


Figure 2.5: Forward process [34].

2.3.2 Reverse process [55]

To reverse the process, one could think of applying the Bayes rule

$$q(x_{t-1}|x_t) = q(x_t|x_{t-1})\frac{q(x_{t-1})}{q(x_t)} \tag{2.6}$$

However, a closed-form expression for $q(x_{t-1})$ and $q(x_t)$ is not available. Thus, to obtain a ground truth for training, it is possible to leverage the Markov chain properties to compute the conditional distribution with respect to x_0, available at training time

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\frac{x_{t-1}|x_0}{x_t|x_0} = q(x_t|x_{t-1})\frac{q(x_{t-1}|x_0)}{x_t|x_0} \tag{2.7}$$

Hence, with a few omitted steps, one can show that

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(\frac{1-\alpha_{t-1}}{1-\alpha_t}\sqrt{1-\beta_t}\,x_t + \frac{\sqrt{\alpha_{t-1}}}{1-\alpha_t}\beta_t x_0, \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t}I). \tag{2.8}$$

The learnable part of a diffusion model is the **reverse process** (Figure 2.6), i.e. the Markov chain of probabilistic mappings from latent x_T to the original image x_0. $p(x_{t-1}|x_t)$ will not, in general, be Gaussian, and will have a probability density function that depends on the real distribution $p_{real}(x)$. They will be Gaussian only in the limit of $\beta_t \to 0$. However, when β_t are small and T is large, they can still be approximated with Gaussians, by defining

$$p(x_T) =^{def} \mathcal{N}(0, I)p(x_{t-1}) =^{def} \tag{2.9}$$

$$\mathcal{N}(\mu_t(x_t;\theta_t), \sigma_t I)\ \forall t = T, ..., 1. \tag{2.10}$$

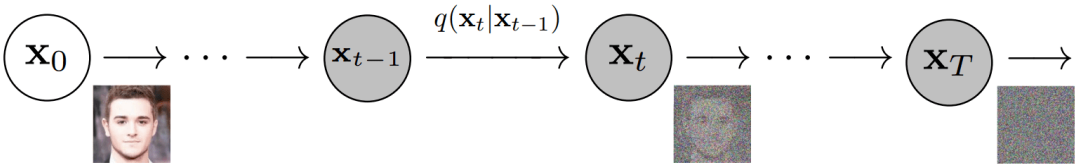


Figure 2.6: Reverse process [34].

2.3.3 Training and Loss [55]

In order to find the optimal parameter values, one could think of maximizing the log-likelihood of the real images $\{x_0^{(i)}\}_{i=1}^I$:

$$\theta_1^*, \ldots, \theta_T^* = \arg\max_{\theta_1,\ldots,\theta_T} \sum_{i=1}^{I} \log p(x_0^{(i)}|\theta_1, \ldots, \theta_T).$$

For each training sample, the joint distribution of the observed image x_0 and the latent vectors $x_1, \ldots, x_T$ is:

$$\begin{aligned}
p(x_0, x_1, \ldots, x_T|\theta_1, \ldots, \theta_T) &= p(x_0|x_1, \ldots, x_T, \theta_1, \ldots, \theta_T)\, p(x_1, \ldots, x_T|\theta_1, \ldots, \theta_T), \\
&= p(x_0|x_1, \theta_1)\, p(x_1, \ldots, x_T|\theta_2, \ldots, \theta_T), \\
&= p(x_0|x_1, \theta_1)\, p(x_1|x_2, \ldots, x_T, \theta_2, \ldots, \theta_T)\, p(x_2, \ldots, x_T|\theta_2, \ldots, \theta_T), \\
&= p(x_0|x_1, \theta_1)\, p(x_1|x_2, \theta_2)\, p(x_2, \ldots, x_T|\theta_3, \ldots, \theta_T), \\
&= \cdots = p(x_0|x_1, \theta_1) \textstyle\prod_{t=2}^{T} p(x_{t-1}|x_t, \theta_t) p(x_T).
\end{aligned}$$

To find the likelihood of x_0, one could, in principle, marginalize the joint probability:

$$p(x_0^{(i)}) = \int p(x_0, x_1, \ldots, x_T|\theta_1, \ldots, \theta_T)\, dx_1 \ldots dx_T,$$

but such marginalization is intractable.

However, it is possible to define a lower bound on the log-likelihood and optimize the parameters to maximize such a bound, which will, in turn, push the likelihood to high values. Such a bound is called the **Evidence Lower Bound (ELBO)**.

To derive the bound, it is sufficient to multiply and divide by the forward process distribution $q(x_1, \ldots, x_T|x_0)$:

$$\log p(x_0|\theta_1, \ldots, \theta_T) = \log \int p(x_0, x_1, \ldots, x_T|\theta_1, \ldots, \theta_T)\, dx_1 \ldots dx_T$$

$$= \log \int \frac{p(x_0, x_1, \ldots, x_T|\theta_1, \ldots, \theta_T)}{q(x_1, \ldots, x_T|x_0)} q(x_1, \ldots, x_T|x_0)\, dx_1 \ldots dx_T$$

$$= \log \mathbb{E}_{q(x_1,\ldots,x_T|x_0)} \left[\frac{p(x_0, x_1, \ldots, x_T|\theta_1, \ldots, \theta_T)}{q(x_1, \ldots, x_T|x_0)} \right]$$

and then to apply the Jensen's inequality [1] to the latter expression obtained. After a series of calculations, one can show a compact expression for the ELBO:

$$\text{ELBO}(\theta_1, \ldots, \theta_T) = \mathbb{E}_{q(x_1|x_0)} \left[\log p(x_0|x_1, \theta_1)\right] - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[D_{\text{KL}} \left(q(x_{t-1}|x_t, x_0) \,\|\, p(x_{t-1}|x_t, \theta_t)\right)\right].$$

By approximating expectations with sampling, writing explicitly the KL divergence and using the same explicit expression for the probabilities already obtained, the ELBO becomes the final loss to minimize:

$$L(\theta_1, \ldots, \theta_T) = \sum_{i=1}^{I} -\log \mathcal{N}\left(x_0^{(i)}; \mu_1\left(x_1^{(i)}; \theta_1\right), \sigma_1 I\right) +$$
$$\sum_{t=2}^{T} \frac{1}{2\sigma_t} || \frac{1-\alpha_{t-1}}{1-\alpha_t} \sqrt{1-\beta_t} x_t^{(i)} + \frac{\sqrt{\alpha_{t-1}}}{1-\alpha_t} \beta_t x_0^{(i)} - \mu_t\left(x_t^{(i)}; \theta_t\right) ||^2.$$

where the first addendum represents the reconstruction of the input from x_1 and the second addendum is the explicit expression of the KL divergence between two Gaussians with constant covariances, like in this case. μ_t represents the prediction of the network, while the first two elements in the norm represent the ground-truth mean of $q(x_{t-1}|x_t, x_0)$.

In practice, the actual loss computed during the training of Diffusion Models is this latter loss reparametrized to predict noise. After this step, the loss becomes

$$L(\theta_1, \ldots, \theta_T) = \sum_{i=1}^{I} \sum_{t=1}^{T} \frac{\beta_t^2}{\alpha_{t-1}(1-\beta_t)} || \left(\epsilon_t\left(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon_t; \theta_t\right) - \epsilon_t\right) ||^2. \quad (2.11)$$

The training algorithm is illustrated in Figure 2.7 by Prince's *Understanding Deep Learning*.

[1] $f(\mathbb{E}_{x\sim p(x)[x]}) \geq \mathbb{E}_{x\sim p(x)}[f(x)]$

Algorithm 18.1: Diffusion model training

Input: Training data $\mathbf{x}$
Output: Model parameters $\boldsymbol{\phi}_t$
repeat
  for $i \in \mathcal{B}$ **do** // For every training example index in batch
    $t \sim \text{Uniform}[1, \ldots T]$ // Sample random timestep
    $\boldsymbol{\epsilon} \sim \text{Norm}[\mathbf{0}, \mathbf{I}]$ // Sample noise
    $\ell_i = \left\| \mathbf{g}_t\left[\sqrt{\alpha_t}\mathbf{x}_i + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \boldsymbol{\phi}_t\right] - \boldsymbol{\epsilon} \right\|^2$ // Compute individual loss
  Accumulate losses for batch and take gradient step
until converged

Figure 2.7: Diffusion Model training algorithm by [50].

2.3.4 Architecture [23]

The architecture for Diffusion Models is a **modified U-Net architecture** [53]. The original U-Net architecture by Ronnemberg et al. is depicted in Fig. 2.8, while the modified version is represented in Fig. 2.9. While the baseline architecture is relatively straightforward, it becomes increasingly complex with advancements in diffusion models. For instance, later developments such as Stable Diffusion introduced an entire latent layer to embed image data. However, this section focuses on the initial versions of the Diffusion Models, as understanding these foundational designs facilitates comprehension of subsequent improvements.

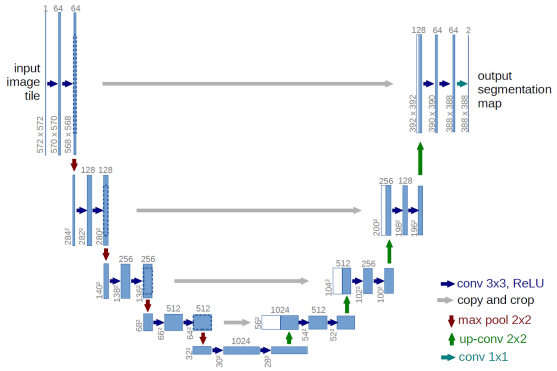


Figure 2.8: The original U-Net architecture by Ronneberger et al. [53].

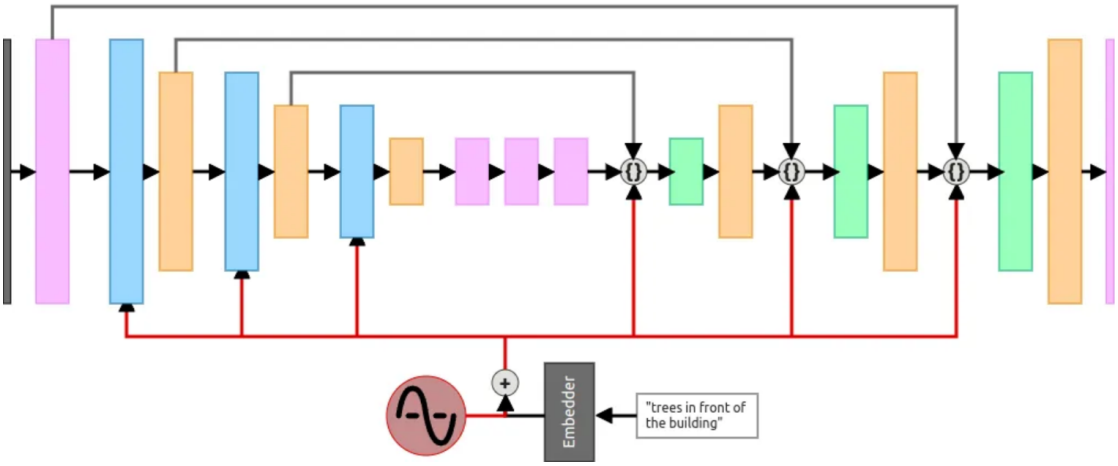


Figure 2.9: The modified U-Net architecture of early Diffusion Models [23]. Purple represents ResNet blocks, Blue represents Downsampling blocks, Orange represents Self-Attention blocks and Green represents Upsampling blocks.

The following paragraphs describe the main components of the modified U-Net architecture used in Diffusion Models.

Embeddings. At each timestep, the model integrates the initial noisy image embeddings that contain information about the current timestep and, if supported, the prompt. However, it is important to note that early diffusion models did not support prompting.

To encode the timestep t, a sinusoidal positional encoding is employed, a technique widely adopted in transformer architectures [15]. For the prompt, an embedding is generated using a suitable embedder.

Embedder. The embedder is responsible for transforming prompts into a format compatible with the network. In the early conditional diffusion models where prompts were supported, simple embedders were sufficient. Nowadays, more sophisticated embedders, such as CLIP, may be employed. These embedders enable the model to process arbitrary text prompts and generate corresponding images. However, such embedders must also be utilized during training to ensure consistency.

The outputs from the positional encoding and text-embedder are combined and injected into the downsampling and upsampling blocks of the network.

ResNet block. The ResNet block (Figure 2.10a) serves as a foundational component of the architecture and is utilized in both downsampling and upsampling operations. In the initial diffusion models, the ResNet block employed was a simple, linear design: one Conv2D layer followed by a GroupNorm and a GELU activation (Gaussian Error Linear Unit), another Conv2D layer and a final GroupNorm layer.

Downsampling Block. The downsampling block (Figure 2.10b) is the first component that processes both the image data and the embeddings containing timestep and prompt information. Functionally, it performs standard downsampling as in the U-Net architecture. The block receives input, downsamples it to match the resolution of the subsequent layer, and integrates the embedding information.

The downsampling process is implemented using a MaxPool2d layer with a kernel size of 2, halving the input dimensions (e.g., 64×64→32×32). This operation is followed by two ResNet blocks. Embeddings are processed through a Sigmoid Linear Unit (SiLU) and a linear layer to match the dimensionality of the ResNet block's output. The embedding tensor and the processed image tensor are then combined and passed to the next block.

Self-Attention Block. In the modified U-Net architecture, some ResNet blocks are replaced with self-attention blocks (Figure 2.10c) to enhance the model's ability to capture global dependencies. For the attention mechanism, Multi-Head Attention (MHA) is employed, where the embedding dimension is set to 128 and the number of attention heads is fixed at 4.

The self-attention block processes input tensors that have been downsampled (e.g., 128×32×32). To apply attention, the input tensor is reshaped to align with the requirements of the attention mechanism. Specifically, the last two dimensions are flattened and transposed, transforming the tensor from 128×32×32 to 1024×128. The reshaped tensor is normalized using layer normalization and is then used to compute the query (Q), key (K), and value (V) tensors for the attention operation.

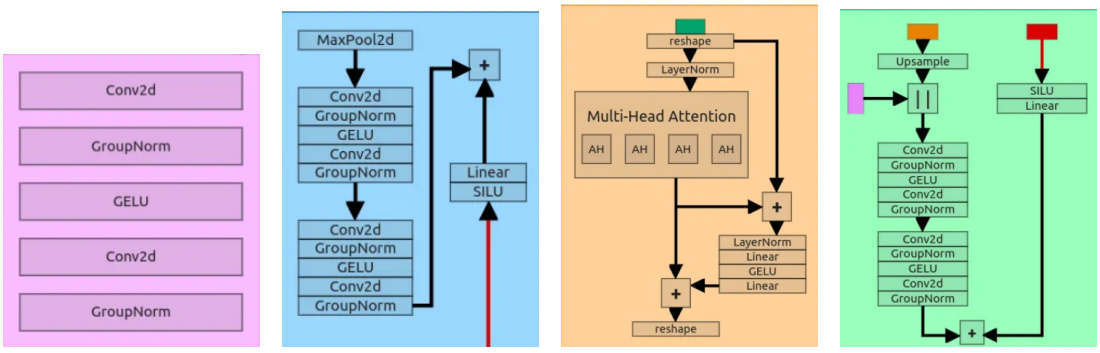
Two skip connections are incorporated within the block. The first adds the reshaped input directly to the output of the attention layer before passing it through a forward layer consisting of normalization, linear transformations, and a GELU activation. The

second skip connection adds the output of this forward layer back to the attention output. Finally, the tensor is reshaped back to its original dimensions (1024×128→128×32×32).

Upsampling Block. The upsampling block (Figure 2.10d) performs the reverse operation of the downsampling block, reconstructing higher-resolution representations. It integrates three inputs: (1) the output from the previous layer, (2) the residual connection from the corresponding downsampling block, and (3) the embedding tensor.

The primary input is upscaled using a simple upsampling layer with a scale factor of 2. The upscaled tensor is concatenated with the residual connection, ensuring both have compatible dimensions. For instance, after upsampling, the tensor from the 5th self-attention block may have dimensions 64×64×64, which matches the dimensions of the residual connection. The concatenated tensors are passed through two ResNet blocks.

The embedding tensor undergoes SiLU activation and a linear transformation before being added to the output of the second ResNet block. The final output of the architecture is obtained via a Conv2d layer with a kernel size of 1, reducing the tensor dimensions (e.g., 64×64×64) to match the target dimensions (e.g., 3×64×64). This output represents the predicted noise.



(a) The **ResNet block** used in Diffusion Models [23].

(b) The **Downsampling block** used in Diffusion Models [23].

(c) The **Self-Attention block** used in Diffusion Models [23].

(d) The **Upsampling block** used in Diffusion Models [23].

Figure 2.10

2.4 Stable Diffusion [5]

Traditional Diffusion Models rely on an iterative reverse diffusion process in which a full-sized image is passed through a U-Net architecture to obtain the final denoised result. While effective, this approach faces significant challenges in computational efficiency, especially when dealing with large image sizes and a high number of diffusion steps (T). The time required for sampling, which involves denoising the image from Gaussian noise, can become prohibitively long. To address these limitations, researchers proposed a novel approach known as Stable Diffusion, originally introduced as the **Latent Diffusion Model (LDM)** [52].

The Stable Diffusion framework introduces two key advancements over conventional diffusion models: the use of **latent space** for diffusion processes and the incorporation of **conditioning mechanisms**. These modifications significantly enhance both computational efficiency and the versatility of image generation.

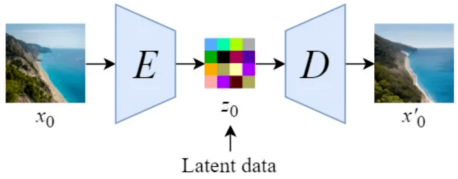
1. **Latent Space to reduce computational burden**: Stable Diffusion achieves its efficiency by performing the diffusion process in a compressed latent space instead of directly in pixel space. This approach involves encoding full-size images into lower-dimensional latent representations using a trained **encoder** (E). The diffusion process, including both forward and reverse diffusion, is then carried out within this latent space. Finally, the latent representation is decoded back to pixel space using a trained **decoder** (D). The encoder and decoder are typically components of a **Variational Autoencoder (VAE)**, which is trained independently, allowing for the decoupling of these components during the diffusion process. Figure 2.11a shows the concept of a Variational AutoEncoder: a full-sized image $x_0 \in \mathbb{R}^{CxHxW}$ is encoded into a latent representation $z_0 \in \mathbb{R}^{C'xH'xW'}$ by the encoder E. Here, $H' < H$ and $W' < W$, leading to significant dimensionality reduction. During the forward process, Gaussian noise is progressively added to z_0 over T steps to obtain z_T, the highly noisy latent representation. The reverse process begins with a noisy latent vector z_T. At each timestep t, the U-Net predicts the noise component, which is partially removed to obtain a less noisy representation z_{t-1}. After T steps, the final latent representation $\hat{z}_0$ is obtained. Finally, during the decoding step, the latent representation $\hat{z}_0$ is transformed back to pixel space

using the decoder D, yielding the generated image $\hat{x_0}$. By shifting the diffusion operations to latent space, Stable Diffusion dramatically reduces the computational costs of image generation. This allows for faster denoising and sampling while maintaining high-quality outputs. Furthermore, the approach enhances the overall stability and robustness of the training process.

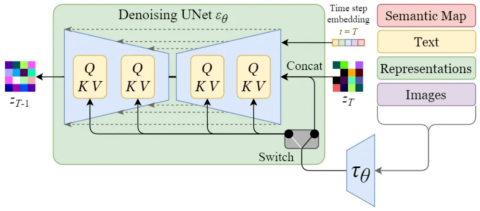
2. **Conditioning**: One of the most notable advancements introduced in Stable Diffusion is its ability to generate images conditioned on specific inputs, such as text prompts or spatial information. This is achieved through the integration of conditioning mechanisms within the diffusion process. The resulting framework supports diverse applications, including text-to-image synthesis, semantic image generation, and inpainting. Conditioning mechanisms include Classifier-Free Guidance and Cross-Attention.

    - **Classifier-Free Guidance:** Prior to Stable Diffusion, generating images of specific classes relied on classifier guidance, where a class label was incorporated into the model input. Stable Diffusion advances this concept by employing Classifier-Free Guidance (CFG) [35], which enables image generation conditioned on more complex inputs, such as textual descriptions.
    - **Cross-Attention Mechanism:** To incorporate conditioning information, the denoising U-Net employs a cross-attention mechanism. This mechanism aligns the conditioning inputs with the image features during the denoising process. In the case of **Textual Conditioning**, text inputs are first transformed into embeddings using pre-trained language models, such as CLIP or BERT. These embeddings are mapped to the U-Net using Multi-Head Attention (MHA), where the input tensors Q, K, and V represent the query, key, and value, respectively. In the case of **Spatial Conditioning**, other forms of conditioning, such as semantic maps or images, are integrated through concatenation with the intermediate feature maps of the U-Net. By incorporating these mechanisms, Stable Diffusion enables controlled and versatile image generation. Users can define specific attributes or styles in the generated images by crafting appropriate conditioning inputs. This functionality

also extends to advanced applications like prompt engineering, where careful design of text prompts results in highly customized outputs. Figure 2.11b shows the conditioning mechanism that was just introduced.



(a) AutoEncoder as proposed by the Stable Diffusion paper [60].



(b) Conditioning mechanism within Stable Diffusion's U-Net [60].

Figure 2.11

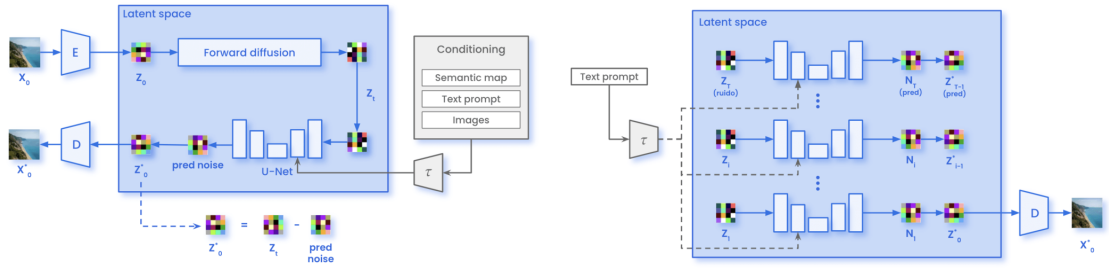
2.4.1 Architecture and Training

The Stable Diffusion architecture combines the principles of latent space diffusion and conditioning to achieve efficient and flexible image generation. The process can be divided into training and sampling phases, as outlined below.

Training Phase - Fig. 2.12a. During training, input images $x_0 \in \mathbb{R}^{CxHxW}$ are encoded into latent representations $z_0 \in \mathbb{R}^{C'xH'xW'}$ using the encoder E. During the forward diffusion step, Gaussian noise is added to z_0 over T steps to generate z_T. Subsequently, the noisy latent representation z_T is passed through the U-Net, which predicts the noise component. This latter predicted noise is compared with the actual noise added during the forward diffusion process, and the resulting loss is used to update the U-Net parameters through backpropagation.

Sampling Phase - Fig. 2.12b. Sampling begins by initializing z_T as random Gaussian noise in the latent space. During the reverse diffusion step, the U-Net iteratively predicts and removes a fraction of the noise over T timesteps, refining z_T to obtain $\hat{z}_0$. The final latent representation $\hat{z}_0$ is decoded into pixel space using the decoder D, yielding the generated image $\hat{x_0}$. By leveraging latent space and conditioning, the Stable Diffusion

model achieves a balance between computational efficiency and flexibility, making it a powerful tool for high-quality, controlled image synthesis.



(a) (b)

Figure 2.12: Stable Diffusion architecture while training (a) and sampling (b) [5].

2.4.2 Stable Diffusion in Dermatology

Stable Diffusion has demonstrated significant potential in the dermatological field. By generating synthetic images of various dermatological lesions, Stable Diffusion enhances existing datasets, improving model training and increasing diagnostic accuracy. This technique has become a cornerstone in **data augmentation**, as evidenced by the **Derm-T2IM** study [25], where the inclusion of synthetic data generated with Stable Diffusion improved the robustness and adaptability of both CNNs and Vision Transformers. Furthermore, the ability to condition diffusion models through textual prompts enables the generation of high-quality and diverse dermoscopic images, leading to increased model accuracy, as observed by **Shavlokhova et al.** [58]. An additional critical advantage is **privacy preservation**: by leveraging synthetic data, researchers can circumvent ethical and legal challenges associated with the dissemination of sensitive medical information.

A particularly important application of Stable Diffusion, explored in this study, is its role in promoting **fairness**. The generation of synthetic images can be strategically designed to include underrepresented categories in dermatological datasets, such as patients with darker skin tones. It has been observed that this type of dataset imbalance induces bias in deep learning models used for classification [9]. To address this issue, in addition to conventional bias mitigation techniques, synthetic images of skin diseases on darker skin tones can be generated. For instance, in the **FairSkin** study, a novel diffusion

model framework mitigates these biases through a three-level resampling mechanism, ensuring a more equitable representation across different racial categories [68]. Similarly, **Borghesi and Calegari** [11] employ Stable Diffusion to create clinical images of skin diseases for underrepresented skin tones.

The objective of this work is to utilize Stable Diffusion—specifically in combination with the DreamBooth technique, which will be introduced in the following subsection—to generate images of skin diseases specifically on darker skin tones, with the aim of mitigating classification bias across different skin colors.

2.5 Fine-tuning Stable Diffusion Models via DreamBooth

Despite their versatility, pre-trained Diffusion Models often lack the capability to faithfully capture and reproduce specific, personalized visual concepts or subjects, limiting their direct application in tasks requiring high levels of customization.

DreamBooth is a fine-tuning technique introduced by Ruiz et al. [54] designed to address this limitation by enabling the personalized customization of pre-trained diffusion models. Originally proposed for fine-tuning text-to-image generative models, DreamBooth allows for the incorporation of a specific subject—such as a unique object, person, or artistic style—into a generative framework with a minimal set of subject-specific images. By introducing a **unique textual identifier** during the training process, the model learns to associate this identifier with the desired subject, while maintaining the diversity and generalization capabilities of the original model.

The process of fine-tuning with DreamBooth involves leveraging a combination of subject-specific data and a carefully balanced optimization strategy. This includes the use of a **class-preserving loss function**, which ensures that the model does not suffer from **catastrophic forgetting**[2]. In the context of Stable Diffusion, this methodology integrates seamlessly with the model's pre-trained architecture, allowing it to generate high-quality, customized images that maintain stylistic coherence and fidelity to the

[2]**Catastrophic forgetting**, also known as **catastrophic interference**, is a phenomenon in machine learning where a model loses previously learned information when it is trained on new data. This typically occurs in neural networks when sequential training is used, especially in the context of continual or lifelong learning.

specified subject.

2.5.1 Loss Function in DreamBooth Fine-Tuning

The fine-tuning process in DreamBooth relies on a carefully designed loss function to achieve two primary goals: (i) to train the model to generate images faithfully representing the specific subject, and (ii) to preserve the model's generalization capability to avoid catastrophic forgetting. The total loss combines multiple components, as described below.

Reconstruction Loss for Subject-Specific Training. The reconstruction loss ensures that the model learns to associate the unique identifier (e.g., "xyz") with the specific subject by conditioning on the provided images and textual descriptions. This is achieved by adapting the noise-prediction objective commonly used in Diffusion Models.

Let $\mathcal{D}_{\text{subject}} = \{x_i\}_{i=1}^{N}$ denote the dataset of N images of the specific subject. Each image x_i is paired with a textual prompt y_i containing the unique identifier for the subject (e.g., "a photo of xyz dog").

The reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{x_i \sim \mathcal{D}_{\text{subject}}, \epsilon \sim \mathcal{N}(0,\mathbf{I}), t} \left[\left\| \epsilon - \epsilon_\theta(z_t, t, y_i) \right\|^2 \right], \tag{2.12}$$

where:

- z_t is the noisy latent representation of the image x_i at diffusion timestep t, obtained via the forward diffusion process.
- ϵ_θ is the noise-prediction network of the diffusion model, parameterized by θ.
- $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is sampled Gaussian noise.
- y_i is the textual prompt describing the subject (e.g., "a photo of xyz dog").

This loss forces the model to learn the association between the unique identifier "xyz" and the subject in the training images.

Class-Preserving Loss. To prevent the model from forgetting its ability to generate diverse images of the subject's broader class (e.g., "dog"), a class-preserving loss is introduced. This loss ensures that the fine-tuning process retains the generalization capabilities of the pre-trained model.

Let $\mathcal{D}_{\text{class}} = \{x_j\}_{j=1}^{M}$ denote a dataset of generic images representing the same class as the specific subject (e.g., generic "dog" images). The class-preserving loss is defined as:

$$\mathcal{L}_{\text{class}} = \mathbb{E}_{x_j \sim \mathcal{D}_{\text{class}}, \epsilon \sim \mathcal{N}(0,\mathbf{I}), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, y_j)\|^2 \right], \tag{2.13}$$

where:

- z_t is the noisy latent representation of the generic image x_j.
- y_j is a generic prompt describing the class (e.g., "a photo of a dog").

This loss ensures that the fine-tuned model continues to generate images aligned with the broader class, avoiding overfitting to the specific subject.

Total Loss. The total loss used for fine-tuning combines the reconstruction loss and the class-preserving loss with weighting factors λ_{recon} and λ_{class}:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{class}} \mathcal{L}_{\text{class}}. \tag{2.14}$$

The weights λ_{recon} and λ_{class} are hyperparameters that control the balance between subject-specific fine-tuning and preservation of generalization. Typically, λ_{recon} is set higher to prioritize learning the new subject, while λ_{class} is chosen to prevent catastrophic forgetting.

Regularization Strategies. To further mitigate overfitting and ensure stability during training, additional regularization strategies are often employed:

- **Parameter-Freezing:** Fine-tuning is restricted to specific layers of the model, such as the cross-attention layers, to avoid excessive deviation from the pre-trained weights.

- **Noise Augmentation:** The input images in $\mathcal{D}_{\text{subject}}$ are augmented with varying levels of noise to enhance robustness and generalization.

By carefully balancing the reconstruction and class-preserving losses, DreamBooth achieves effective subject-specific fine-tuning while maintaining the versatility of the original diffusion model.

2.5.2 DreamBooth in Dermatology

The application of DreamBooth represents an innovation in the dermatological field. However, DreamBooth has already been employed in several studies. Specifically, its use in dermatology enables the generation of personalized images based on the unique characteristics of an individual, thereby improving diagnostic outcomes. Moreover, DreamBooth facilitates **few-shot learning**, as it requires only a limited number of images to generate high-quality synthetic images. These advantages have established DreamBooth as an effective and widely used tool in dermatological research [25]. For instance, DreamBooth can be particularly useful for generating feature-aligned synthetic data, as demonstrated from **Nair et al.** [46], where intermediate features of a diffusion model are aligned with expert output features, thereby enhancing generation accuracy.

This thesis explores the use of Stable Diffusion with the DreamBooth technique to generate highly faithful medical images that closely resemble the original ones. This approach leverages DreamBooth's ability to achieve effective training with a limited number of input images, making it particularly well-suited for medical imaging applications where data availability is often constrained. Overall, in our study, DreamBooth has proven to be an effective technique for synthetic image generation in scenarios with limited data availability. Specifically, it has allowed us to focus on underrepresented skin tones, ultimately improving the fairness of our model.

Chapter 3

Dataset description and Preprocessing

This Chapter provides a detailed explanation of the characteristics of the **non-dermoscopic dataset** used in this work, along with the preprocessing steps applied to it. In particular, since the images in the dataset were captured using **consumer cameras**, they exhibit low quality, which makes classification challenging. To address this issue, a sliding window-based algorithm was employed. Subsequently, it was necessary to automatically detect the skin color of each image to obtain a general understanding of the classification bias present in the dataset. Skin tone estimation was performed using the **Individual Typology Angle (ITA)**, a widely adopted formula in the literature for extracting skin tone from images.

The results indicate that some of the crops obtained through the preprocessing algorithm are of poor quality, suggesting room for improvement in this stage. Additionally, the low illumination conditions of the images make automatic skin color detection via ITA more challenging, although ITA remains a robust tool for this purpose.

The Chapter is organized as follows:

- **Section 3.1** provides a comprehensive description of the dataset and its main challenges;
- **Section 3.2** details the preprocessing steps applied to the dataset using the sliding window-based algorithm;
- **Section 3.3** explains the technique adopted to automatically measure skin color

in the images with the ITA, enabling the assessment of classification bias with respect to skin tone.

3.1 Data

The dataset consists of approximately 8,000 images of 273 pediatric patients at Sant'Orsola hospital in Bologna representing nine possible skin diseases: *drug-induced iatrogenic exanthema, maculopapular exanthema, morbilliform exanthema, polymorphous exanthema, viral exanthema, urticaria, pediculosis, scabies* and *chickenpox*. The images were captured using consumer-grade cameras by the hospital's doctors, meaning they are **non-dermoscopic**. The dataset exhibits several critical characteristics that complicate classification:

- **High variability in illumination:** Many images were taken under suboptimal lighting conditions and are therefore darker than in reality. In addition to complicating classification, poor illumination significantly affects skin tone classification, shifting skin tones to darker values that do not accurately reflect the patient's skin colour.
- **High variability in size and quality:** Since the images were taken with different consumer cameras, they lack a standard size or quality. As a result, a standardization process is required to enable the dataset to be processed by a neural network.
- **Inconsistent focus on affected skin areas:** Some images capture the whole body, others only a small body part, and some focus solely on the skin region where the disease is present.
- **Blurriness:** Some images are blurrier than others.
- **Imbalance in skin tones:** The dataset predominantly contains photos of patients with lighter skin tones, making classification more challenging for less-represented skin tone categories.
- **Imbalance in disease classes:** Some diseases (Figure 3.1) are overrepresented in the number of samples. Consequently, the network is expected to better classify

certain illnesses over others. While this is anticipated, addressing class imbalance is not the focus of this work.

3.2 Data preprocessing

The data preprocessing pipeline aims to standardize the dataset by generating uniformly sized image crops. The objective is to identify and extract regions of the images containing visible skin disease, using the binary mask associated with each image. The process follows a sliding-window approach and consists of the following steps [1]:

1. **Patch extraction:** The algorithm starts at the top-left corner and extracts a fixed-size patch (256×256 pixels).

2. **Disease coverage calculation:** For each patch, the binary mask is used to calculate the *disease coverage*, defined as the ratio of positive labels (1) to negative labels (0) within the mask. This metric is used to evaluate the presence of the skin disease based on contrast within the patch: if the disease coverage exceeds a predefined threshold (indicating sufficient contrast), the patch is retained, and discarded otherwise.

3. **Sliding window application:** The patch extraction process repeats as the sliding window moves across the image in set steps, generating overlapping patches. To reduce redundancy, a *non-maxima suppression* procedure discards patches with lower disease coverage when overlap exceeds a threshold.

4. Finally, patches exhibiting low contrast (e.g., due to poor illumination or blurriness) are removed to improve the overall quality of the dataset. This step ensures that only well-defined and informative regions are retained for further analysis.

As expected, the preprocessing step reveals an inherent imbalance in the dataset across the different disease classes. Figure 3.1 illustrates the distribution of retained

[1] The preprocessing algorithm was adapted from the one developed by Alessandro D'Amico, Riccardo Murgia and Mazeyar Moeini Feizabadi: `https://github.com/eskinderit/experiments-synthetic-generation-clinical-skin-images/blob/main/generate_images.ipynb`

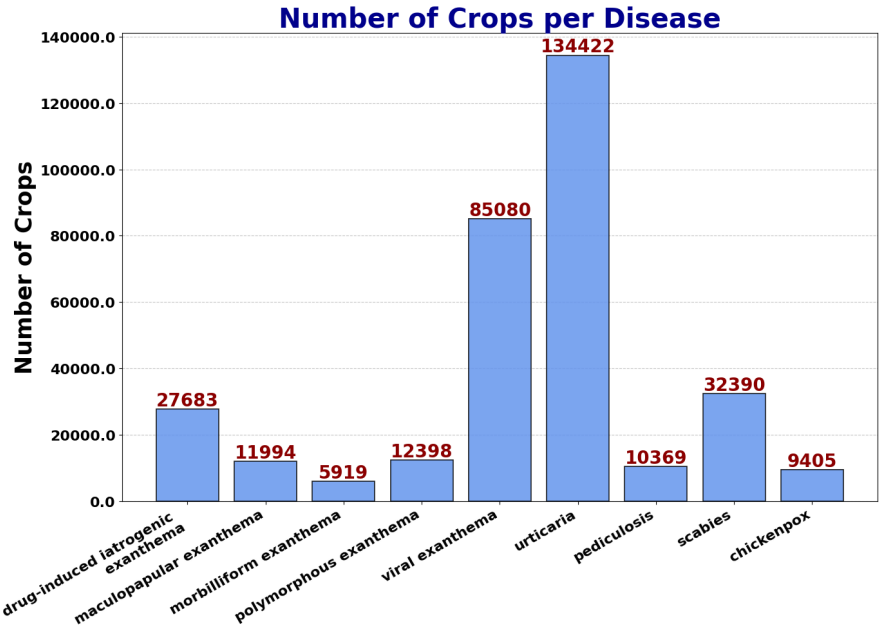


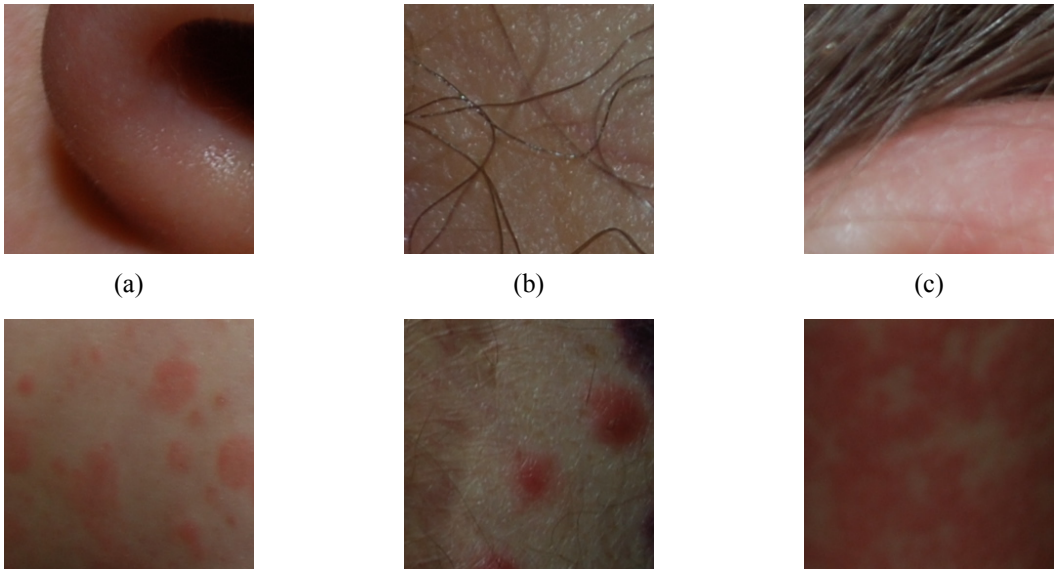
Figure 3.1: Number of generated crops for each disease.

crops for each disease. Notably, certain diseases are underrepresented, with fewer than ten thousand examples available after preprocessing. This imbalance is anticipated to impact model performance, particularly for less-represented diseases. Furthermore, the cropping algorithm is blind towards body parts and/or regions that do not display signs of skin disease but still exhibit high ”disease” coverage. This results in the generation of a variable number of crops - depending on the quality of the image - which may include portions of eyes, noses, lips, genitals, areas with body hair, or edges. A potential solution to address this issue is discussed in the Conclusions Chapter.

Other issues include poor illumination and blurriness, which cannot be resolved through cropping alone. Examples of blurry and poorly illuminated crops are shown in Figure 3.2.

3.3 Skin tone classification on non-dermoscopic images

In the literature, skin tone classification is commonly performed using the **Individual Typology Angle (ITA)**, a metric first introduced by Chardon et al. in 1991 [13], and



(a) (b) (c)

(d) (e) (f)

Figure 3.2: Examples of bad crops generated by the cropping algorithm: unnecessary body part (a), area with body hair (b), edges (c), poor illumination (d, e), poor illumination and blurriness (f).

widely adopted in subsequent studies for its simplicity and effectiveness [32, 30, 41]. ITA values are computed within the CIELAB colour space, leveraging the lightness (L*) and yellow-blue axis (b*) components to derive an angular value correlating with skin tone:

$$ITA = \tan^{-1}\left(\frac{L^* - 50}{b^*}\right) \times \frac{180}{\pi} \tag{3.1}$$

While this method has proven effective in controlled environments, such as dermoscopic datasets, it assumes uniform illumination and does not account for variations introduced by pathological changes in the skin or external artefacts.

In this work, we propose a modified ITA computation method tailored to our dataset, which includes images of skin conditions captured under non-standardized conditions with consumer-grade cameras. To address challenges such as altered pigmentation in the affected skin, inconsistent illumination, and shadows, we exclude disease-affected

regions from ITA computation using binary maps, ensuring only unaffected skin is analyzed. Unlike prior works relying on fixed thresholds from dermoscopic datasets [30, 41, 14], we classify ITA values into skin tone categories using a **Gaussian Mixture Model (GMM)**, which better handles dataset variability. This refined method provides a more accurate representation of skin tone, enabling a fairer evaluation of classification performance across diverse skin tones.

3.3.1 Our approach

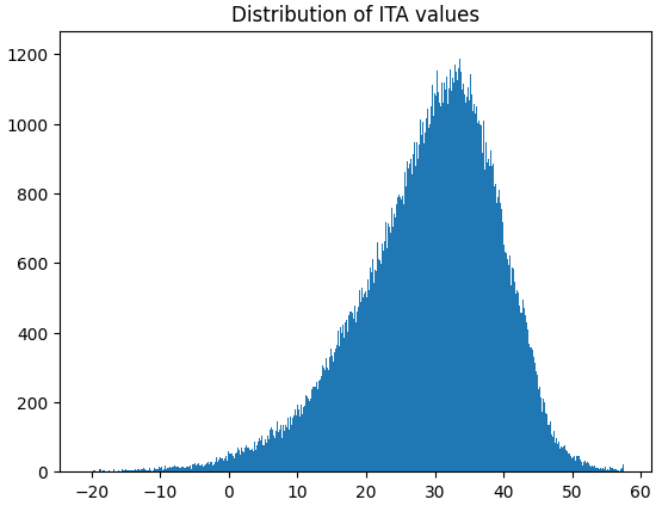


Figure 3.3: ITA dataset distribution.

The computation of the ITA must account for the fact that skin affected by disease often appears darker and reddish compared to healthy skin. To ensure reliable ITA values that represent the baseline skin tone, it is key to exclude disease-affected regions from the calculation. This was achieved by applying a **bitwise AND operation** between the original image crop and its corresponding segmentation mask, replacing disease-affected regions with black pixels. The ITA value was then computed exclusively for the non-black pixels in the crop. The resulting distribution of ITA values, shown in Figure 3.3, closely resembles a Gaussian distribution with a longer tail extending towards lower values. Following the computation of ITA values, ranges are required to classify skin tone according to the Fitzpatrick scale [32], which categorizes skin into **six types**. Various

thresholding schemes have been proposed to map ITA values to Fitzpatrick skin types [30, 41, 14]. However, these ranges were primarily designed for dermoscopic datasets, devoid of variability caused by illumination, angulation, or other artefacts. Given the non-dermoscopic nature of our dataset, these thresholds were deemed unsuitable.

Instead, we assumed that images with similar skin tones exhibit similar ITA values within a reasonable range of variation. To classify the ITA values, we fitted the distribution using a **Gaussian Mixture Model**[2] with six components, corresponding to the six skin tone categories in the Fitzpatrick scale. Each ITA value was assigned to a Gaussian component according to the thresholds given by the intersection of the Gaussian components (see the Appendix for further details). The resulting skin tone labels were categorized as *dark*, *brown*, *tan*, *intermediate*, *light*, and *very light*.

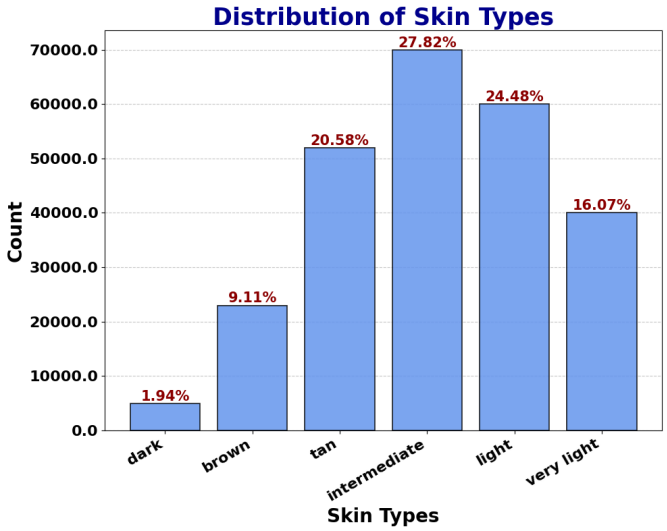


Figure 3.4: Six skin labels dataset distribution.

Examples of the automatic skin tone classification are presented in Figure 3.5. While the ITA value is generally robust, shadows and poor illumination can lower the ITA value, resulting in a darker assigned skin tone. Nonetheless, darker images—whether due to actual skin tone or suboptimal lighting—were correctly assigned a lower ITA value, whereas lighter images were assigned higher ITA values. Figure 3.4 shows the

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.GaussianMixture.predict`

distribution of skin tone labels across the dataset. Notably, the *dark* and *brown* skin tone categories are underrepresented, highlighting an imbalance in skin tone distribution within the dataset.

Figure 3.5: Examples of automatic skin tone classification based on the ITA values and the Gaussian mixture technique.

Despite the robustness of the ITA calculation, this labelling process is not entirely accurate. Poor illumination or other artefacts cause the computed ITA value to deviate from the expected value for the true skin tone for a non-negligible number of images. Future work could address this limitation by incorporating advanced correction techniques for artefacts such as shadows and uneven lighting.

Chapter 4

Classification with a Convolutional Neural Network

This Chapter addresses the classification of nine skin diseases from the original dataset using a **Convolutional Neural Network (CNN)**. Initially, a grid search was conducted to obtain the optimal combination of hyperparameters. Subsequently, the CNN was trained on the non-augmented dataset, and performance was measured using both traditional metrics, namely Accuracy and F1-score, as well as fairness metrics, specifically Disparate Impact (DI), Equalized Odds Ratio (EOR), and Predictive Rate Ratio (PRR).

The results demonstrate the presence of a skin color-dependent bias in the classification, with the CNN's performance varying significantly across different diseases. This variation highlights the need for a more stable and balanced performance across the different classes.

The Chapter is structured as follows:

- **Section 4.1** describes the hyperparameter search process and the training of the CNN on the non-augmented dataset.
- **Section 4.2** introduces the fairness metrics (DI, EOR, and PRR) in detail and discusses the classification results in terms of these fairness measures.

4.1 CNN Architecture and training

The task of dermatological classification necessitates the model's ability to effectively capture intricate features at multiple scales. To address this, a deep Convolutional Neural Network (CNN) architecture consisting of five convolutional layers, each with a kernel size of 3, was chosen. The first convolutional layer expands the number of channels from 3 (corresponding to the standard RGB input) to 64, allowing the network to extract detailed low-level features. In the subsequent layers, except for the final one, the number of channels is progressively doubled, enhancing the network's ability to represent more complex features. Following each convolutional block, a MaxPooling layer with a kernel size of 2 is applied to reduce the spatial dimensions and help mitigate overfitting. The final layer is a fully connected layer that outputs nine logits corresponding to the nine target classes. The architecture is shown in Figure 4.1.

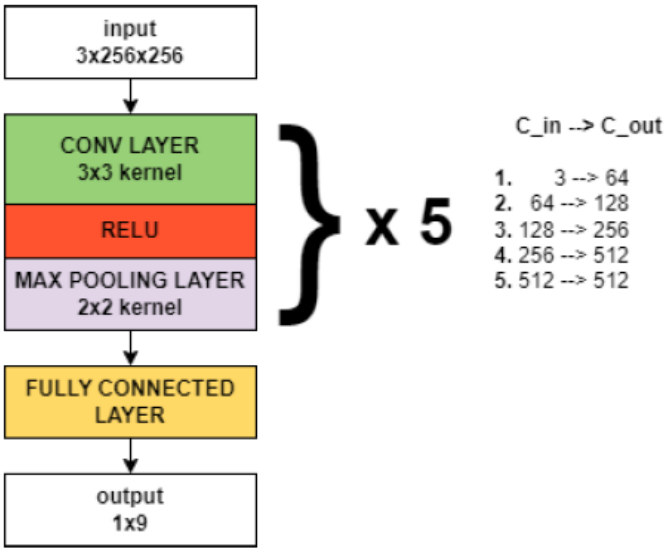


Figure 4.1: Diagram of the architecture of the Convolutional Neural Network.

The dataset of cropped images was divided into training (60% of the samples), validation (20%), and test sets (20%).

A **hyperparameter optimization phase** was carried out to fine-tune the batch size and learning rate. Six different combinations of these parameters were tested, with the model trained for 5 epochs using stochastic gradient descent (SGD) with momentum as

the optimizer. The optimal performance, as evidenced by trends in both the loss and F1 score in Figures 4.2 to 4.7, was achieved with a batch size of 128 and a learning rate of 0.01. It was observed that a learning rate of 0.01 combined with smaller batch sizes, such as 32 or 64, resulted in early overfitting, as shown in the F1-score plots in Figures 4.2 and 4.7. Additionally, a learning rate of 0.001 also led to overfitting, while a higher value of 0.05 caused divergence during training.

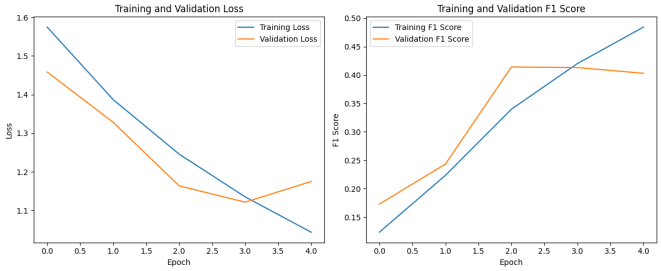


Figure 4.2: CNN grid search training using batch size = 32 and learning rate = 0.01.

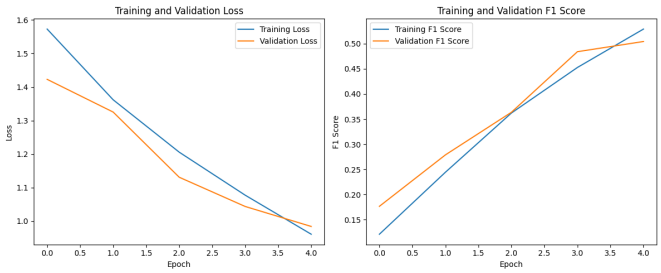


Figure 4.3: CNN grid search training using batch size = 64 and learning rate = 0.01.

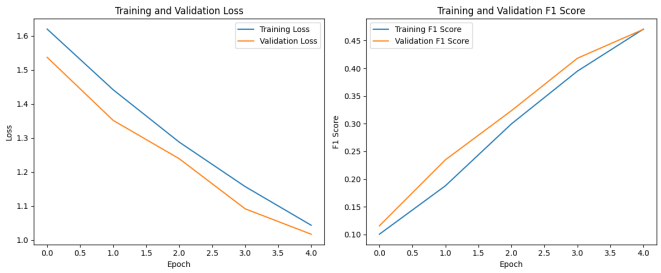


Figure 4.4: CNN grid search training using batch size = 128 and learning rate = 0.01.

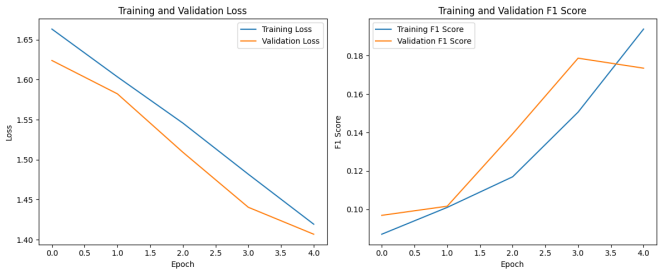


Figure 4.5: CNN grid search training using batch size = 64 and learning rate = 0.001.

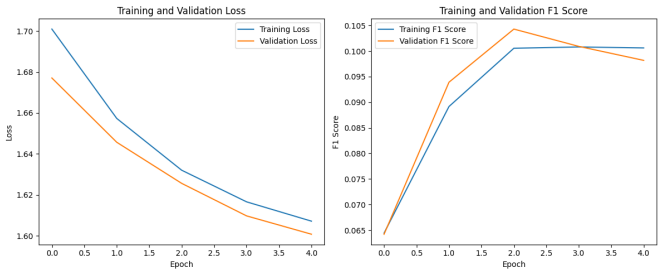


Figure 4.6: CNN grid search training using batch size = 256 and learning rate = 0.001.

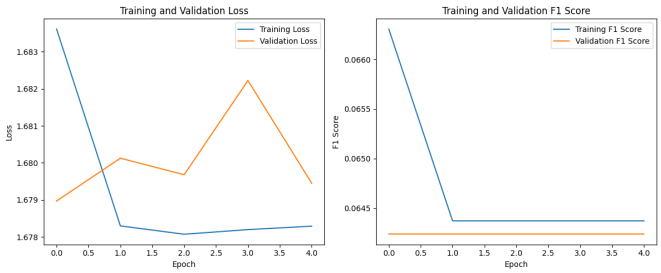


Figure 4.7: CNN grid search training using batch size = 128 and learning rate = 0.05.

For the final model training, the following configuration was adopted: **batch size** = 128, **learning rate** = 0.01, **number of epochs** = 15, and **optimizer** = SGD with momentum. To further enhance the training process, a cosine decay learning rate scheduler and an early stopping mechanism were employed to prevent unnecessary computational costs in cases of premature overfitting. Figure 4.8 illustrates the trends for both loss and F1 score, with the F1 score being the primary metric for saving model checkpoints during training.

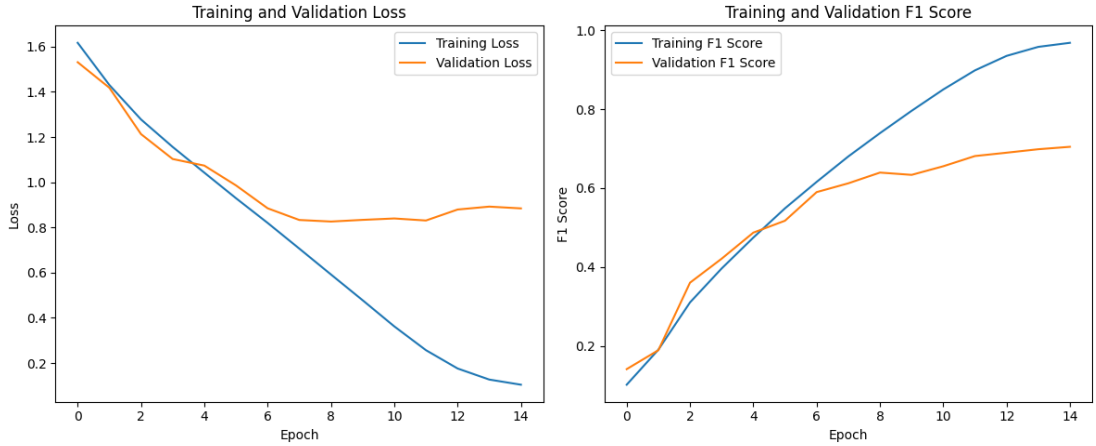


Figure 4.8: Loss and F1-score trends for the final training of the network with the optimal hyperparameter values.

4.2 Results and discussion

The model's evaluation was conducted using **standard performance metrics**, specifically Accuracy and the F1 score, computed for each skin tone category. The **Accuracy** results, as reported in Table 4.1, demonstrate relatively stable performance across most skin tones, except for the "very light" category, which achieved a significantly higher accuracy. This disparity may stem from the superior overall quality and illumination of images in this category, facilitating easier classification.

However, Accuracy alone does not capture the distribution of misclassifications among different skin tones. For example, despite comparable average Accuracy scores, discrepancies in true positive rates (TPRs) and false positive rates (FPRs) across groups may indicate biases in classification. The **macro F1 score** values, also presented in Table 4.1, align with the trends observed in Accuracy, providing additional insights into the model's performance across different skin tones. These findings highlight the necessity of incorporating multiple evaluation metrics to effectively detect biases and potential limitations in the model's performance across demographic groups.

While traditional performance metrics offer some insights into model behavior, they provide limited information regarding fairness in diagnostic predictions. To further investigate this aspect, the model was assessed using fairness metrics widely utilized in

		Accuracy	F1 score
Minority	dark	78.4%	0.74
	brown	78.2%	0.71
Majority	tan	75.5%	0.69
	intermediate	75.1%	0.68
	light	76.5%	0.68
	very light	82.0%	0.75

Table 4.1: Total CNN accuracy and F1 score across the different skin tones.

the literature for tasks related to skin disease classification. Specifically, we report the **Disparate Impact Ratio (DI)**, **Equalized Odds Ratio EOR)**, and **Predictive Rate Ratio (PRR)**, which will be later introduced. Since fairness metrics are conventionally designed for binary classification problems, their adaptation to this study's multi-class setting, which includes multiple demographic groups, requires careful consideration. To facilitate this adaptation, skin tones were grouped into two broader categories: a **Minority** group (comprising "dark" and "brown" skin tones) and a **Majority** group (including "tan," "intermediate," "light," and "very light" skin tones). This classification is motivated by two key factors: (1) the observed underrepresentation of "dark" and "brown" skin tones in the dataset, as illustrated in Chapter 3, and (2) methodological precedents established in prior research, such as Corbin et al. [19]. This grouping enables a meaningful application of fairness metrics while addressing the inherent complexities of a multi-class, multi-group classification task. Additionally, Accuracy and F1 score were calculated for each disease, considering the aforementioned skin tone aggregation, with the results summarized in Table 4.2.

The evaluation of the model's performance for individual diseases reveals considerable variability. Notably, for nearly half of the diseases, **Accuracy and F1 scores in the test set are higher for the Minority group than for the Majority group**, contradicting the expectation that the Minority group would be systematically disadvantaged. Two plausible explanations may account for this trend in conventional metrics:

1. Artifacts such as inadequate lighting, body hair, or image noise may lead to **errors in skin tone classification**, misassigning certain images to the Minority group instead of the Majority group. This misclassification complicates the accurate assessment of bias.

2. The **limited diversity between the training and test sets for the "black" and "brown" skin tone categories** (especially the former) may inadvertently inflate classification performance. For instance, in the case of maculopapular exanthema on "black" skin, the Minority group exhibits a substantially higher Accuracy and F1 score. Further qualitative examination reveals that the dataset contains only one individual with this skin tone and condition. As outlined in Chapter 3, the cropping algorithm generates multiple image crops from a single individual, distributing them across training, validation, and test sets. Consequently, during training, the model learns to recognize these crops effectively, and at test time, it encounters visually similar samples, leading to an artificially high Accuracy for this condition in "black" skin. We hypothesize that if the test set contained images of additional individuals with "black" skin who were absent during training, the model's performance would decline significantly. Conversely, the Majority group benefits from greater diversity in training samples, enabling the model to generalize more effectively when encountering unseen test images, resulting in more robust classification performance.

As previously mentioned, in addition to the traditional metrics, we have employed three widely used fairness metrics from the literature, which are outlined in detail in the following paragraphs.

Disparate Impact Ratio [26]. In the binary case, the DI is defined as

$$DI = \frac{Pr(\hat{Y} = 1 | X \in minority\_group)}{Pr(\hat{Y} = 1 | X \in majority\_group)} \tag{4.1}$$

This metric evaluates the proportion of individuals receiving a positive outcome between the *minority* group and the *majority* group. A value of 1 indicates perfect fairness (equal probabilities), while values below 1 suggest unfairness against the minority group. Conversely, values above 1 imply unfairness against the majority group.

The DI was calculated separately for each condition, where $\hat{Y} = 1$ represents the presence of the disease and $\hat{Y} = 0$ its absence. The results are presented in Table 4.2. A value between 0.8 and 1.2 is generally considered fair. Values below 0.8 indicate unfairness against the minority group, whereas values above 1.25 suggest unfairness against

the majority group. Notably, the model demonstrates significant bias against the minority group for diseases such as pediculosis and chickenpox, as evidenced by DI values below 0.8. This disparity may be attributed to the limited number of positive samples from the minority group for these conditions. In contrast, for diseases such as maculopapular rash, morbilliform rash, and scabies, there is a proportionally higher number of positive detections in the minority group compared to the majority group, resulting in DI values above 1.2. These observations highlight the varying degrees of fairness across different conditions and the impact of sample imbalances in fairness evaluations.

Equalized Odds Ratio [2]. A classifier satisfies Equalized Odds under a distribution over (X, A, Y) (where A indicates the sensitive feature) if its prediction $\hat{Y}$ is conditionally independent of the sensitive feature A given the label Y. As shown by [2], this is equivalent to $\mathbb{E}(\hat{Y}|A = a, Y = y) = \mathbb{E}(\hat{Y}|Y = y)\ \forall\ a, y$. Equalized odds requires that the true positive rate, $Pr(\hat{Y} = 1|Y = 1)$, and the false positive rate, $Pr(\hat{Y} = 1|Y = 0)$, are equal across groups. In our case, EOR was computed for each disease using the fairlearn library, in which EOR is defined as 'the smaller of two metrics: `true_positive_rate_ratio` and `false_positive_rate_ratio`. The former is the ratio between the smallest and largest of $Pr(\hat{Y} = 1|A = a, Y = 1)$, across all values of the sensitive feature(s). The latter is defined similarly but for $Pr(\hat{Y} = 1|A = a, Y = 0)$. The equalized odds ratio of 1 means that all groups have the same true positive, true negative, false positive, and false negative rates[1].

The EOR takes values between 0 and 1. An EOR value of 1 indicates fairness, meaning all groups exhibit identical true positive, true negative, false positive, and false negative rates. Consistent with the Disparate Impact Ratio analysis, skin tones were aggregated into two broader categories: a minority group and a majority group (Table 4.2). Notably, the EOR values for several diseases—iatrogenic drug-induced rash, morbilliform rash, polymorphic rash, pediculosis, scabies, and chickenpox—are significantly below 1. These findings suggest that the model's predictions $\hat{Y}$ are influenced by the skin tone attribute, even when conditioned on the true label Y. Such dependencies highlight potential biases in the classifier and underline the need for interventions to improve

[1] `https://fairlearn.org/main/api_reference/generated/fairlearn.metrics.equalized_odds_ratio.html#fairlearn.metrics.equalized_odds_ratio`

fairness across groups.

	Accuracy		F1 score		DI	EOR	PRR
	Minority	Majority	Minority	Majority			
drug-induced i. exanthema	71.3%	76.8%	0.73	0.78	0.99	0.78	0.93
maculopapular exanthema	57.7%	49.03%	0.65	0.55	1.35	0.85	1.18
morbilliform exanthema	71.5%	74.12%	0.74	0.78	1.32	0.55	0.96
polymorphous exanthema	63.8%	52.28%	0.69	0.59	0.85	0.63	1.22
viral exanthema	80.9%	76.8%	0.79	0.75	0.98	0.82	1.05
urticaria	85.9%	85.6%	0.84	0.83	0.93	0.86	1.00
pediculosis	62.4%	65.4%	0.68	0.69	0.74	0.68	0.95
scabies	74.1%	69.8%	0.75	0.72	1.46	0.67	1.06
chickenpox	54.3%	57.0%	0.61	0.61	0.76	0.70	0.95
All	78.2%	76.8%	0.72	0.70	1.04	0.73	1.03

Table 4.2: Fairness and performance results for the CNN model.

Predictive Rate Ratio [6]. The predictive Rate Parity is achieved when the Positive Predictive Value (PPV), also known as *precision*, is the same across all skin tone groups. The formula for precision is

$$PPV = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \tag{4.2}$$

The PRR compares the precision between the two groups. It is computed by taking the ratio of the PPV of the minority group to the PPV of the majority group:

$$PRR = \frac{PPV_{minority\_group}}{PPV_{majority\_group}} \tag{4.3}$$

In our implementation, this value is calculated for each disease, using the common division of skin tones into minority and majority groups. The results are presented in Table 4.2. We observe that PRR values are fair across all diseases, in contrast to the unfair values observed for the DI and the EOR. To understand this difference, it is important to note that the PPV relative to a group measures the frequency with which the model correctly predicts the positive class for that group. In this sense, PPV serves as a measure of the *quality* of predictions. On the other hand, the Disparate Impact Ratio focuses on the probability of a positive prediction for each group, regardless of its correctness, making it a measure of *quantity*. Similarly, the EOR evaluates the True Positive Rate (recall) and the False Positive Rate, which also reflect the *quantity* of True Positives and

False Positives identified by the model. Therefore, while PRR compares the model's precision across groups, the other metrics (DIR and EOR) assess the distribution and balance of predictions among groups.

Chapter 5

Classification with a Swin Transformer

This Chapter addresses the classification of the nine diseases from the non-augmented dataset using a **Swin Transformer (ST)**. As discussed in Chapter 2, the Swin Transformer has recently gained significant popularity in the medical domain, particularly in dermatology, due to its scalability and ability to capture long-range dependencies. In this work, a pre-trained Swin Transformer is employed, and a grid search is performed to find the optimal fine-tuning hyperparameters. Subsequently, the performance is evaluated using the same metrics applied to the CNN case, to provide a comparison.

The results show that bias remains present in the classification, but the Swin Transformer achieves significantly higher and more consistent performance compared to the CNN. This improvement is likely attributed to the model's greater capacity and the fact that it was pre-trained.

The chapter is structured as follows:

- **Section 5.1** describes the grid search process to find the optimal hyperparameter configuration and the final training on the non-augmented dataset;
- **Section 5.2** presents the classification results in the same format as the CNN and discusses them in terms of fairness.

5.1 Swin Transformer Training

During the training phase, the cropped dataset was partitioned in the same manner as for the CNN: 60% for training, 20% for validation, and 20% for testing. To effectively capture both skin texture and disease-specific features, a model variant pre-trained on ImageNet-1k and fine-tuned on a skin cancer dataset was employed [1]. The fine-tuning process involved updating the parameters of the last three stages of the Swin Transformer while keeping the weights of the initial stage frozen, resulting in a total of **26 million trainable parameters**.

To facilitate model convergence and systematically explore the parameter space, hyperparameter tuning was primarily conducted on the learning rate, evaluating three values: 1e-4, 1e-3, and 1e-2. All experiments were carried out with a batch size of 512, as larger batch sizes are expected to yield more stable gradient estimates. Each learning rate configuration was tested over 20 training epochs, with an Early Stopping mechanism in place to halt training if the validation loss failed to improve within a predefined tolerance over a given number of epochs, thereby preventing unnecessary resource consumption. The results of the hyperparameter tuning process are illustrated in Figures 5.1 to 5.3: while lower learning rates promoted stable convergence, they frequently caused the model to become trapped in local minima, leading to suboptimal accuracy and F1 scores. Consequently, a learning rate of 1e-2 was selected, as it allowed for more substantial parameter updates. To enhance stability in later epochs, the learning rate was reduced by a factor of 100 after nine epochs, based on the observed loss patterns. The evolution of loss and F1 scores throughout training is presented in Figure 5.4.

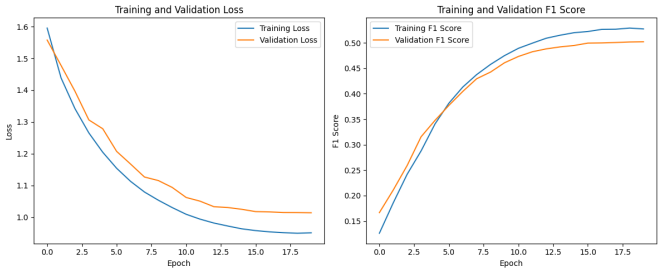


Figure 5.1: ST grid search training using batch size = 512 and learning rate = 0.0001.

[1] https://huggingface.co/gianlab/swin-tiny-patch4-window7-224-finetuned-skin-cancer

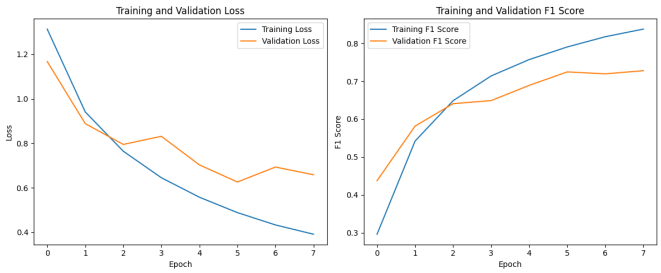


Figure 5.2: ST grid search training using batch size = 512 and learning rate = 0.001. Note how the training was stopped early due to lack of performance improvement on the validation set.

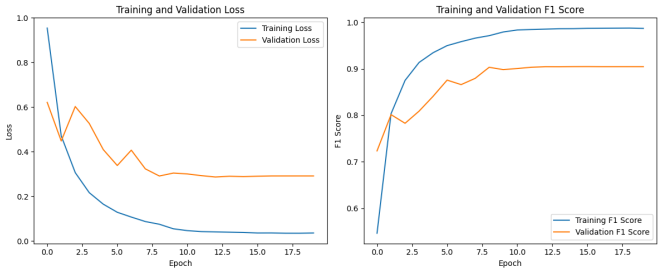


Figure 5.3: ST grid search training using batch size = 512 and learning rate = 0.01.

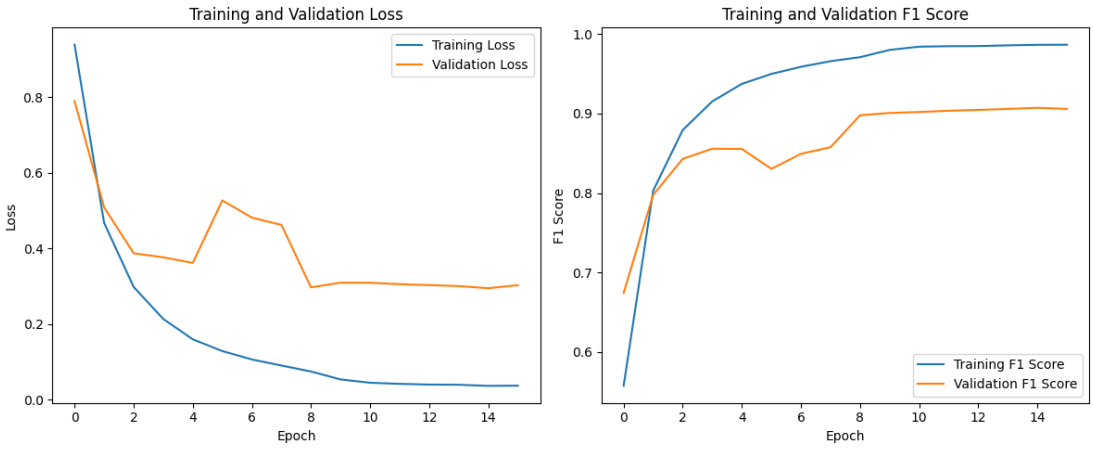


Figure 5.4: Loss and F1-score trends for the final training of the Swin Transformer with the optimal hyperparameter values.

5.2 Results and discussion

To assess the model's performance, we adopted the same metrics specified for the CNN. The final results are shown in Table 5.2. First, we notice a **significant performance improvement**, likely due to the remarkably higher capacity of the Swin Transformer (∼26

million trainable parameters versus the ∼4 million of the CNN). Moreover, examining the Accuracy and F1 score columns in Table 4.2 and Table 5.2, it is evident that, while for the CNN the Accuracy and F1 scores vary significantly between demographic groups depending on the disease, the Swin Transformer shows more consistent Accuracy and F1-score values. The DI values of the Swin Transformer indicate that, on average, it demonstrates greater fairness compared to the CNN. Specifically, the Swin Transformer exhibits only three out of nine instances of unfair values, in contrast to the CNN, which presents five out of nine instances of unfair values. On the other hand, *EOR* values are systematically lower in the Swin Transformer results compared to those of the CNN, once again indicating that the model's predictions are strongly influenced by the *skin tone* attribute.

		Accuracy	**F1 score**
Minority	dark	90.8%	0.91
	brown	91.2%	0.91
Majority	tan	91.9%	0.92
	intermediate	91.7%	0.92
	light	90.3%	0.90
	very light	91.5%	0.92

Table 5.1: Swin Transformer accuracy and the F1 score.

	Accuracy		**F1 score**		**DI**	**EOR**	**PRR**
	Minority	**Majority**	**Minority**	**Majority**			
drug-induced i. exanthema	94.7%	96.8%	0.90	0.93	1.01	0.72	0.98
maculopapular exanthema	88.4%	88.6%	0.84	0.84	1.36	0.74	1.00
morbilliform exanthema	89.7%	95.1%	0.89	0.92	1.21	0.76	0.94
polymorphous exanthema	91.7%	87.1%	0.92	0.89	0.81	0.73	1.05
viral exanthema	95.1%	95.6%	0.90	0.90	0.96	0.91	0.99
urticaria	91.9%	88.8%	0.93	0.93	0.99	0.74	1.03
pediculosis	84.0%	88.2%	0.89	0.92	0.78	0.61	0.95
scabies	81.9%	89.0%	0.90	0.93	1.23	0.10	0.92
chickenpox	88.4%	91.0%	0.89	0.86	0.72	0.35	0.97
All	91.1%	91.3%	0.90	0.90	1.01	0.63	0.98

Table 5.2: Results for the Swin Transformer model.

Chapter 6

Synthetic generation of skin images

Chapters 3, 4, and 5 have revealed that the dataset used in this study exhibits an underrepresentation of dark skin categories. This underrepresentation introduces a bias in classification, as observed in the cases of both the CNN and the Swin Transformer (ST). Chapter 2 explained how Stable Diffusion can be leveraged to balance datasets, a capability that may prove useful in improving classification fairness by generating images of underrepresented categories, such as skin diseases on dark skin.

This Chapter introduces a methodology for generating synthetic images of skin diseases on dark skin using Stable Diffusion in conjunction with DreamBooth. The first step involved the manual selection of images for model training, constructing mini datasets. Subsequently, a grid search was conducted to identify optimal hyperparameters, exploring various hyperparameter combinations in the final training phase to enhance diversity. Finally, the synthetic images were incorporated into the dataset using three different numerical approaches, and both the CNN and the Swin Transformer were retrained on the augmented dataset. The results obtained were then compared with those from the non-augmented dataset.

The findings demonstrate that synthetic data significantly improved fairness metric values, bringing them within fair thresholds in many cases for both the CNN and the Swin Transformer. Furthermore, for the CNN, synthetic data likely had a strong regularization effect, as evidenced by a substantial increase in performance across Accuracy and F1-score metrics.

In conclusion, the generation and utilization of synthetic data had an overall positive

impact, increasing classification fairness. This confirms that DreamBooth can serve as a valuable tool for generating dermatological images that faithfully replicate real cases and highlights how synthetic data can contribute to more equitable classification in the medical domain by facilitating the creation of balanced datasets.

This Chapter is structured as follows:

- **Section 6.1** describes the fine-tuning process of a Stable Diffusion model using DreamBooth, aimed at generating realistic images of dark-skinned patients.
- **Section 6.2** outlines the three different numerical data augmentation approaches employed to construct three datasets with varying numbers of synthetic images, enabling an analysis of the extent to which synthetic image addition enhances fairness.
- **Section 6.3** presents the results of data augmentation using the three approaches on the CNN, along with a discussion of the findings.
- **Section 6.4** reports the results of data augmentation using the three approaches on the Swin Transformer (ST), followed by their discussion.

6.1 Image Generation Via DreamBooth

The dataset utilized in this study contains a limited number of examples of skin diseases affecting individuals with black skin, with at most 4 or 5 individuals with dark skin for each disease. The preprocessing pipeline described in Chapter 3 generates a large number of image crops from photographs of the same individual. However, **using all these crops to train an image generation model would be redundant**, as the crops originating from the same individual are highly similar to one another. Consequently, it is sufficient to select only a few representative crops (3 or 4) per individual with dark skin and construct a small, curated dataset comprising multiple individuals with dark skin exhibiting the specific disease of interest. This curated dataset can then be used to train an image generation model.

One technique well-suited for training with such a limited number of examples is **DreamBooth**, introduced by Ruiz et al. in 2023 [54] and further explored in theory

in Chapter 2. DreamBooth is a fine-tuning technique for generative models, such as diffusion-based models, that enables the generation of high-quality, subject-specific images. By leveraging only a few samples of a subject, DreamBooth personalizes the model while preserving its ability to generate diverse and photorealistic outputs. This makes it particularly effective for scenarios where data scarcity is a critical limitation.

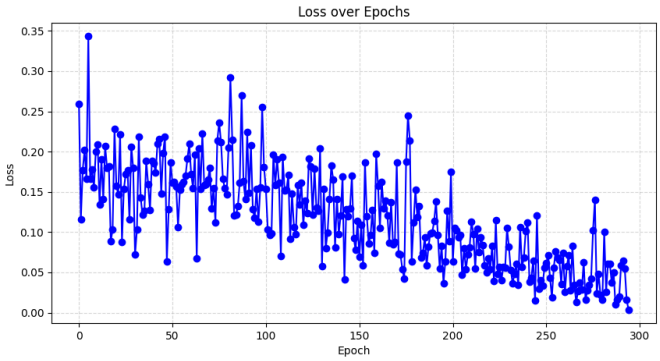


Figure 6.1: Example of loss plot obtained while fine-tuning Stable Diffusion with DreamBooth.

In this study, DreamBooth was employed to fine-tune a pre-trained Stable Diffusion model, specifically, the version provided by *RunwayML*, which was pre-trained on images of size 512×512. The fine-tuning process was divided into the following stages:

1. **Manual Exploration of the Dataset.** A manual inspection of the dataset was conducted to identify the skin diseases for which images of 'dark' or 'brown' skin types were available. This exploration revealed that only three out of the nine diseases—*maculopapular exanthema, viral exanthema*, and *scabies*—contained images of individuals with 'dark' and 'brown' skin. Consequently, image generation was feasible only for these three diseases.

2. **Construction of Mini-Datasets.** Mini-datasets were manually constructed for each of the three diseases, separately for 'brown' and 'dark' skin types. This process resulted in six datasets (two for each disease: one for 'brown' skin and one for 'dark' skin), with each dataset containing between 14 and 29 images.

3. **Fine-Tuning with DreamBooth.** For each of the six mini-datasets, the Stable Diffusion model was fine-tuned using the DreamBooth technique. A grid search was conducted to identify optimal hyperparameter configurations, exploring the following parameters:

    - *Learning rate*: Values of 5e-7, 2e-6, 5e-6, and 1e-5 were tested. This wide range ensured adequate exploration of the parameter space, as the learning rate is a critical factor for convergence.
    - *Maximum training steps*: For mini-datasets with fewer than 15 images, values of 1000, 2000, 3000, and 4000 steps were tested. For mini-datasets with more than 15 images, values of 2000, 3000, 4000, and 5000 steps were tested. This choice was guided by a commonly applied rule of thumb in DreamBooth, which recommends fine-tuning with at least 100 training steps per image.
    - *Instance prompt*: The instance prompt in DreamBooth plays a key role in both the training and image generation phases. During training, a unique identifier (e.g., `<unique_ID>`) is included in the prompt alongside descriptive context (e.g., "human skin" or "a person with a skin condition") to associate the fine-tuned model with the specific features of the training images. This enables the model to learn how to reproduce those features while maintaining its broader generative capabilities. During image generation, the instance prompt is used to guide the model in synthesizing new images that reflect the characteristics of the fine-tuned training data. By combining or modifying the instance prompt with additional textual descriptions, keeping the `<unique_ID>` in the text, it is possible to control the specific details of the generated images, ensuring alignment with the desired output while retaining diversity and realism. In our case, both the prompt "`<unique_ID>`" and "`<unique_ID>` human skin" were evaluated. Including 'human skin' in the prompt was hypothesized to provide context and aid in accurately reproducing skin texture. However, this inclusion might also reduce image diversity, as the model could exhibit a monotonic and non-diverse interpretation of 'human skin.'

A batch size of 1 was selected, as experiments revealed that smaller batch sizes promoted greater diversity when other hyperparameters were held constant. Additionally, both the U-Net and the text encoder were fine-tuned. Images in the mini-datasets were resized to 512×512 prior to fine-tuning to fully leverage the capabilities of the pre-trained Stable Diffusion model.

4. **Selection of Fine-Tuned Models.** For each of the six mini-datasets, the fine-tuned models with the most promising hyperparameter configurations were selected based on an **empirical evaluation** of the generated images. The evaluation prioritized diversity, accuracy, faithful representation of skin texture and colour, and fidelity to the real images in the mini-dataset.

6.2 Data Augmentation Approach

After determining the number of synthetic images required for each of the three diseases and each of the two skin colours ('dark' and 'brown'), this total was distributed among the fine-tuned models selected for that disease and skin colour. This distribution ensured that the synthetic images were generated by models fine-tuned with different hyperparameter combinations, thereby enhancing the diversity of the generated dataset. Fine-tuned models with varying hyperparameters tend to produce images with unique characteristics that reflect differences in the mini-dataset used for training. Additionally, the generation seed was frequently altered to further increase diversity in the synthetic images. Figure 6.2 presents examples of generated images compared to real ones for both skin tones. It can be observed that the generated images closely resemble the original ones, accurately replicating both the texture of the disease and its localization.

To incorporate the synthetic images into the original training set (while leaving the test and validation sets untouched to ensure they contained only real images), thereby providing more examples of diseases on darker skin tones, three distinct numerical approaches were followed:

1. **AugMin** – in this approach, synthetic images of 'dark' and 'brown' skin are added such that the total number of images (real + synthetic) for each of these two skin

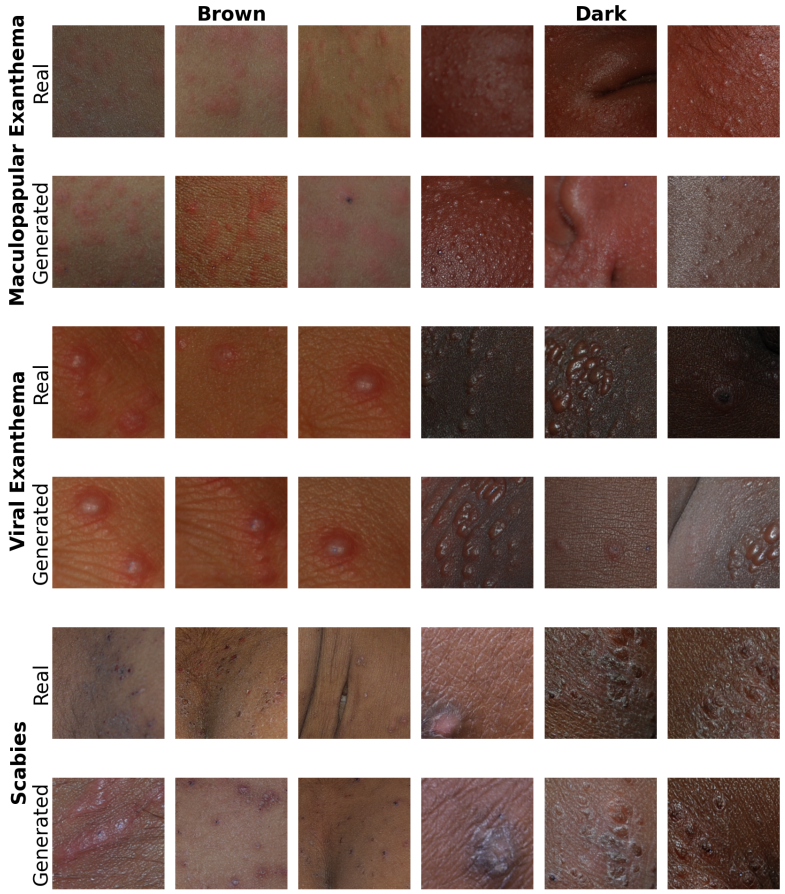


Figure 6.2: Real vs. generated images for each of the three target diseases using the DreamBooth technique.

colours for a specific disease was equal to the smallest number of images among the other four skin colours ('very light,' 'light,' 'intermediate,' 'tan') for that disease[1].

2. **AugBalanced** – synthetic images are added for the 'dark' and 'brown' skin colours of each disease such that, for that specific disease, the number of images for each of the two colours ('dark' and 'brown') represented one-sixth (approximately 17%) of the total number of images for that disease. This approach was based on the

[1]For example, if, for the disease *scabies*, the skin colour 'tan' has the smallest number of examples among the other four colours, with x images, then synthetic images of scabies were added for 'dark' and 'brown' skin to ensure that each of these two colours also had a total of x images.

rationale that there are six possible skin colours, and ideally, the number of images for each colour should constitute roughly one-sixth of the total number of images for any given disease. Consequently, the combined number of images for 'dark' and 'brown' skin colours accounted for approximately one-third (34%) of the total images for that disease.

3. **AugMax** – the third approach is analogous to the first but differs in that the total number of images (real + synthetic) for each disease and each of the two skin colours ('brown' and 'dark') was made equal to the largest number of images among the other four skin colours ('very light,' 'light,' 'intermediate,' 'tan') for that disease.

The introduction of these three distinct approaches aims to evaluate the impact of each proportion of 'dark' and 'brown' skin images on various fairness metrics, as well as on accuracy and F1 score. This approach allows for a **constructive analysis of how the representation of underrepresented skin tones in the training set influences model performance and fairness outcomes**. Comparing the three proportions is particularly valuable for understanding the trade-off between fairness and performance, as well as for identifying the proportion that best balances equitable representation across skin tones with high predictive accuracy. Such a comparative study provides insights into the relationship between data distribution and model behaviour, highlighting the potential benefits or drawbacks of increasing diversity within a dataset.

6.3 Results of Synthetic Augmentation on the CNN

The addition of synthetic images generally resulted in a significant performance improvement across all diseases[2], as evidenced by the values of Table 6.1. This effect may be attributed to the **regularizing impact** of these new data on the dataset, which benefited all diseases. Furthermore, Accuracy and F1-score also improved across individual skin tones, including both darker and lighter tones, for which no synthetic images were generated. Overall, except for the 'very light' skin tone, the addition of synthetic data

[2]Including those for which no synthetic data was added

helped equalize performance across skin tones, raising metrics for lighter tones (which previously had lower Accuracy and F1 scores compared to 'dark' and 'brown' tones) more than it did for darker tones. Regarding fairness metrics, synthetic data also benefited diseases for which no synthetic images were generated. We provide now a more detailed analysis of each augmentation technique.

AugMin adds the fewest synthetic images and provides the smallest improvement in terms of both traditional and fairness metrics, suggesting that more synthetic data could be beneficial. As shown in Table 6.1, the Accuracy and F1-score improvements for individual diseases sometimes narrowed the performance gap between Minority and Majority groups (e.g., in the case of *drug-induced iatrogenic exanthema, morbilliform exanthema, polymorphous exanthema, viral exanthema, urticaria*, and *scabies*), while in other cases, the performance gap widened. Regarding fairness metrics (Table 6.2), no significant improvement was observed for the three diseases targeted with synthetic images in 'dark' and 'brown' tones, and in some cases, a decline was noted. Interestingly, however, certain diseases for which no synthetic images were generated showed counterintuitive fairness improvements. In summary, this approach appears to function as a regularizer that enhances overall performance and improves the homogeneity of model performance across skin tones. However, it is not effective in improving classification fairness, particularly for the targeted diseases (i.e., *maculopapular exanthema, viral exanthema*, and *scabies*).

AugBalanced outperformed the previous approach in terms of both Accuracy and F1-score. In this case, fairness outcomes for the three targeted diseases were very similar to those observed without synthetic data. However, for most other diseases, fairness appeared to improve, particularly for *EOR* values. Overall, this demonstrates that a greater presence of synthetic data has a stronger regularizing effect on performance, benefiting nearly all diseases and all skin tones.

AugMax yielded the best trade-off between fairness and performance. Accuracy and F1-score metrics remained higher than the model trained on the original dataset, while the average fairness metrics for each disease fell within ranges considered fair. Significant improvements were observed for both *DI* and *EOR* values in two of the three diseases for which synthetic images were generated (i.e., *viral exanthema* and *scabies*). Additionally, for most other diseases, fairness metrics also improved. For the CNN,

generating synthetic images for the three targeted diseases and incorporating them into the dataset proved beneficial for both overall model performance and classification fairness. The more synthetic images, the merrier: **AugMax achieves the best trade-off between fairness and performance.**

	No synthetic augmentation				**AugMin**				**AugBalanced**				**AugMax**			
	Accuracy		**F1 score**		**Accuracy**		**F1 score**		**Accuracy**		**F1 score**		**Accuracy**		**F1 score**	
	Min	**Maj**	**Min**	**Maj**	**Min**	**Maj**	**Min**	**Maj**	**Min**	**Maj**	**Min**	**Maj**	**Min**	**Maj**	**Min**	**Maj**
DII ex.	71.3%	76.8%	0.73	0.78	76.5%	80.5%	0.77	0.80	**77.8%**	**82.6%**	**0.79**	**0.83**	76.5%	81.2%	0.78	**0.83**
MP ex.	57.7%	49.0%	0.65	0.55	**70.5%**	57.2%	**0.73**	0.61	69.3%	58.7%	**0.73**	0.63	68.5%	**59.0%**	**0.71**	0.63
MF ex.	71.5%	74.1%	0.74	0.78	75.2%	77.5%	0.79	0.82	**76.4%**	80.0%	**0.82**	**0.84**	73.3%	**80.7%**	0.79	**0.84**
PM ex.	63.8%	52.3%	0.69	0.59	69.9%	59.1%	0.73	0.64	**74.7%**	**62.6%**	**0.77**	0.66	71.2%	61.6%	0.75	**0.67**
V ex.	80.9%	76.8%	0.79	0.75	81.4%	78.8%	0.82	0.79	80.5%	79.4%	0.82	**0.80**	**83.2%**	**79.8%**	**0.83**	**0.80**
urticaria	85.9%	85.6%	0.84	0.83	88.4%	88.4%	0.85	0.86	**89.2%**	89.2%	**0.86**	**0.87**	88.9%	**89.4%**	**0.86**	**0.87**
pediculosis	62.4%	65.4%	0.68	0.69	57.7%	70.9%	0.63	0.74	**68.6%**	**74.8%**	0.71	**0.76**	67.5%	72.0%	**0.73**	**0.76**
scabies	74.1%	69.8%	0.75	0.72	79.1%	75.3%	0.79	0.77	**81.3%**	**77.6%**	**0.81**	**0.79**	78.4%	75.5%	**0.81**	0.78
chickenpox	54.3%	57.0%	0.61	0.61	50.9%	60.7%	0.59	0.66	52.0%	62.3%	0.61	0.69	**62.4%**	**65.5%**	**0.66**	**0.70**
All	78.2%	76.8%	0.72	0.70	81.1%	80.2%	0.75	0.74	82.0%	**81.5%**	**0.77**	**0.76**	**82.1%**	81.4%	**0.77**	**0.76**

Table 6.1: CNN accuracy and F1-score: disease aggregation.

	No synthetic augmentation			**AugMin**			**AugBalanced**			**AugMax**		
	DI	**EOR**	**PRR**	**DI**	**EOR**	**PRR**	**DI**	**EOR**	**PRR**	**DI**	**EOR**	**PRR**
DII ex.	**0.99**	0.78	0.93	0.97	**0.93**	**0.95**	0.98	0.83	0.94	**0.99**	0.76	0.94
MP ex.	**1.35**	0.85	1.18	1.44	0.81	1.23	1.46	0.85	1.18	1.44	**0.86**	**1.16**
MF ex.	1.32	0.55	**0.96**	1.26	0.68	0.97	1.21	**0.82**	0.95	**1.19**	0.71	0.91
PM ex.	**0.85**	0.63	1.22	**0.85**	**0.65**	1.18	0.82	0.56	1.19	0.83	0.61	**1.16**
V ex.	0.98	0.82	1.05	0.95	0.73	1.03	0.95	0.80	**1.01**	**0.99**	**0.86**	1.04
urticaria	0.93	0.86	**1.00**	**0.95**	**0.97**	**1.00**	**0.95**	**0.97**	**1.00**	0.94	0.93	0.99
pediculosis	0.74	0.68	**0.95**	0.73	0.81	0.81	**0.75**	**0.84**	0.92	0.74	0.73	0.94
scabies	1.46	0.67	1.06	1.44	0.68	1.05	1.43	0.69	1.05	**1.35**	**0.91**	**1.04**
chickenpox	0.76	0.70	**0.95**	0.71	0.75	0.84	0.73	0.83	0.83	**0.84**	**0.95**	**0.95**
All	1.04	0.73	1.03	0.93	0.78	**1.01**	**1.03**	**0.88**	**1.01**	**1.03**	0.81	**1.01**

Table 6.2: CNN DI, EOR and PRR: disease aggregation.

		No synthetic augmentation		**AugMin**		**AugBalanced**		**AugMax**	
		Accuracy	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
Minority	dark	78.4%	0.74	81.6%	0.77	**82.7%**	**0.79**	81.9%	**0.79**
	brown	78.2%	0.71	81.0%	0.74	81.8%	**0.76**	**82.1%**	**0.76**
Majority	tan	75.5%	0.69	79.3%	0.73	**80.5%**	0.75	**80.5%**	**0.76**
	intermediate	75.1%	0.68	78.7%	0.73	**80.0%**	**0.75**	79.6%	**0.75**
	light	76.5%	0.68	79.7%	0.73	**81.0%**	**0.74**	80.9%	**0.74**
	very light	82.0%	0.75	84.9%	0.79	86.2%	**0.81**	**86.3%**	**0.81**
All		77.0%	0.77	80.3%	0.80	**81.6%**	**0.81**	81.5%	**0.81**

Table 6.3: CNN Accuracy e F1-score: skin tones aggregations.

6.4 Results of Synthetic augmentation on the ST

The Swin Transformer has already demonstrated very high Accuracy and F1-score values across all diseases and skin tones, while EOR values were notably problematic, particularly for the latter diseases. Even in this case, different trends can be observed depending on the data augmentation approach:

AugMin improved the Accuracy and F1-score for several diseases in the minority categories (i.e., 'dark' and 'brown' skin tones) at the cost of a slight and acceptable decrease in Accuracy and F1-score for the majority category. Overall, however, Accuracy and F1-score remained high and comparable to the baseline across skin tones and diseases. Regarding fairness metrics, AugMin produced significant improvements, particularly for the EOR metric: the addition of synthetic data improved seven out of nine EOR values (with only one deteriorating compared to the baseline), bringing four of these values into the fairness range. Additionally, DI values improved for two of the three target diseases. Furthermore, overall PRR values, which were already good, also showed improvement. On the whole, this approach provides the best trade-off between performance and fairness.

AugBalanced resulted in a more substantial decrease in Accuracy and F1-score compared to the other approaches, although these metrics remained high. In terms of fairness metrics, AugBalanced led to notable improvements in EOR values, with six out of nine improving (at the cost of two deteriorating compared to the baseline). However, it was not as effective as AugMin, bringing only two EOR values within the fairness threshold. Finally, this approach worsened DI values overall but improved PRR values.

AugMax also delivers good Accuracy and F1-score values, albeit distributed differently compared to AugMin. Regarding fairness metrics, this approach performs well for DI and PRR values (though without substantial improvement over the model trained on the original dataset, except for pediculosis). However, it is less effective for EOR values, with the exception of two diseases—*urticaria* and *scabies*—for which the values were brought within the fairness range.

In conclusion, **for the Swin Transformer, AugMin proved to be the most effective.** This highlights how the model's fairness (and, in some cases, classification accuracy) benefited from the addition of synthetic images. Nonetheless, the Swin Transformer exhibited fewer improvements in Accuracy and F1-score compared to the CNN when the same number of synthetic images were added. This outcome can be explained by at least two factors:

1. Although the Swin Transformer has significantly more trainable parameters than the CNN, it is a pre-trained model, making it inherently more resistant to substantial changes. In contrast, the CNN is entirely retrained from scratch, allowing for greater flexibility in performance improvements.

2. The Swin Transformer had already achieved high Accuracy and F1-score during training without synthetic data, showing far fewer signs of overfitting compared to the CNN. Consequently, **synthetic images did not have the same regularizing effect on the Swin Transformer as they did on the CNN**, as there was less room for improvement. This also explains why the Swin Transformer benefited more from the approach involving fewer synthetic data: adding more data likely exceeded the model's "saturation point," limiting the desired improvements in fairness, though it still delivered better results than the model trained on the original dataset in terms of fairness.

Overall, **the generation of synthetic images proved to be a successful strategy for both the CNN and the Swin Transformer**. In both cases, all approaches achieved better fairness outcomes compared to the respective models trained on the original dataset, and for the CNN, significant gains were observed in Accuracy and F1-score as well. The use of the DreamBooth technique, which enables lightweight fine-tuning of the Stable Diffusion model and allows for multiple training iterations with differently fine-tuned models, was critical in ensuring diversity and equitable representation. Our findings advocate for the continued use of synthetic data augmentation to enhance fairness and performance in dermatological AI applications, paving the way for more equitable healthcare solutions.

	No synthetic augmentation				AugMin				AugBalanced				AugMax			
	Accuracy		F1 score		Accuracy		F1 score		Accuracy		F1 score		Accuracy		F1 score	
	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj	Min	Maj
DII ex.	94.7%	96.8%	0.90	0.93	95.8%	94.8%	**0.94**	**0.94**	**96.3%**	**97.1%**	0.90	0.91	94.4%	96.2%	0.93	0.93
MP ex.	88.4%	88.6%	**0.84**	**0.84**	87.5%	**89.5%**	**0.84**	0.83	88.1%	87.4%	**0.84**	**0.84**	**91.8%**	86.5%	**0.84**	**0.84**
MF ex.	89.7%	**95.1%**	0.89	**0.92**	**96.4%**	93.9%	0.89	0.86	95.2%	93.9%	0.89	0.86	**96.4%**	91.8%	**0.91**	0.85
PM ex.	**91.7%**	87.1%	**0.92**	**0.89**	**91.7%**	**87.5%**	**0.92**	0.88	84.3%	84.4%	0.87	0.86	**91.7%**	86.5%	0.89	0.86
V ex.	95.1%	**95.6%**	0.90	**0.90**	95.9%	94.9%	0.90	**0.90**	95.1%	95.5%	0.88	0.88	**96.1%**	95.5%	**0.91**	0.89
urticaria	**91.9%**	**88.8%**	**0.93**	**0.93**	90.4%	88.1%	**0.93**	0.92	86.3%	84.7%	0.90	0.90	90.2%	86.9%	**0.93**	0.91
pediculosis	84.0%	88.2%	0.89	0.92	**86.6%**	**89.4%**	**0.90**	**0.93**	83.5%	89.1%	0.88	**0.93**	85.6%	87.3%	0.89	0.92
scabies	81.9%	**89.0%**	0.90	**0.93**	84.1%	87.6%	**0.91**	0.92	85.2%	88.2%	**0.91**	0.91	**85.7%**	87.8%	**0.91**	0.92
chickenpox	**88.4%**	91.0%	**0.89**	0.86	88.4%	**92.7%**	**0.89**	**0.87**	85.0%	88.6%	0.88	0.86	83.2%	87.7%	0.87	**0.87**
All	91.1%	**91.3%**	**0.90**	**0.90**	91.3%	90.7%	**0.90**	0.89	89.3%	89.4%	0.89	0.89	**91.4%**	90.1%	**0.90**	0.89

Table 6.4: ST accuracy and F1-score: disease aggregation.

	No synthetic augmentation			AugMin			AugBalanced			AugMax		
	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR	DI	EOR	PRR
DII ex.	1.01	0.72	0.98	1.01	0.87	**1.01**	**1.00**	**0.88**	**0.99**	0.94	0.70	0.98
MP ex.	1.36	0.74	**1.00**	**1.28**	**0.93**	0.98	1.41	0.65	1.01	1.52	0.56	1.06
MF ex.	1.21	0.76	0.94	**1.17**	0.91	0.98	1.22	**0.95**	**1.01**	1.21	0.77	1.05
PM ex.	0.81	0.73	1.05	0.79	0.54	1.05	0.76	0.58	**1.00**	**0.83**	**0.79**	1.06
V ex.	0.96	0.91	0.99	**0.98**	0.91	1.01	0.97	**0.95**	**1.00**	0.95	0.76	1.01
urticaria	**0.99**	0.74	1.03	0.98	0.78	1.03	0.97	0.75	**1.02**	0.97	**0.93**	1.04
pediculosis	0.78	0.61	0.95	0.78	**0.88**	0.97	0.76	0.76	0.94	**0.82**	0.43	**0.98**
scabies	**1.23**	0.10	0.92	1.28	0.40	0.96	1.28	0.45	0.96	1.32	**0.91**	**0.98**
chickenpox	0.72	0.35	**0.97**	0.71	0.40	0.95	0.72	0.35	0.96	**0.73**	**0.42**	0.95
All	1.01	0.63	0.98	**1.00**	**0.74**	0.99	1.01	0.70	**1.00**	1.03	0.70	1.01

Table 6.5: ST DI, EOR and PRR: disease aggregation.

		No synthetic augmentation		AugMin		AugBalanced		AugMax	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Minority	dark	90.8%	0.91	**91.5%**	**0.92**	89.9%	0.90	**91.5%**	0.91
	brown	91.2%	**0.91**	91.3%	0.90	89.2%	0.88	**91.4%**	0.90
Majority	tan	**91.9%**	**0.92**	91.7%	0.90	90.3%	0.89	91.7%	0.90
	intermediate	**91.7%**	**0.92**	91.0%	0.90	90.2%	0.89	90.5%	0.89
	light	**90.3%**	**0.90**	89.4%	0.87	88.2%	0.87	88.1%	0.86
	very light	**91.5%**	**0.92**	91.1%	0.90	88.5%	0.89	90.1%	0.89
All		**91.3%**	0.90	90.8%	**0.91**	89.4%	0.89	90.2%	0.90

Table 6.6: ST Accuracy and F1-score: skin tones aggregation

Chapter 7

Conclusions and Further Work

This study demonstrated the effectiveness of using advanced image generation techniques, like DreamBooth combined with Stable Diffusion, to enhance the representation of underrepresented skin tones in medical datasets. Our methods significantly improved fairness metrics, balancing performance and fairness effectively. Incorporating synthetic images, especially in the training sets for diseases affecting 'dark' and 'brown' skin tones, addressed data scarcity issues and reduced bias in medical image analysis. The comparison of different data augmentation strategies (AugMin, AugBalanced, AugMax) helped us understand the trade-offs between dataset diversity and predictive accuracy. While the CNN showed more significant improvements due to its flexibility, the pre-trained nature of the ST limited its adaptability to synthetic data enhancements. However, both models benefited from our approach, underscoring the potential of synthetic data to improve diagnostic tools across diverse skin tones. We also noticed that although synthetic images are produced only for specific diseases, the experimental results demonstrate enhanced performance across all nine diseases catalogued in our study. Our findings advocate for using synthetic data augmentation to enhance fairness and performance in dermatological AI applications, paving the way for more equitable healthcare solutions.

There are several improvements that can be applied to the pipeline of this work to enhance its robustness and generalizability:

- One key aspect to address is the **preprocessing stage**, particularly in terms of better isolating disease-affected areas and ensuring that only high-quality images are included. An important step in this direction could be the **removal of hair from the images**. For instance, Delibasis et al. [21] leverage machine learning-based approaches to segment hair pixels, subsequently employing an inpainting algorithm to replace hair pixels with values derived from the surrounding image structure. Variational autoencoders can also be employed for this purpose, as demonstrated by the study of Bardou et al. [8]. Another relevant preprocessing improvement involves **removing facial features** such as the nose, eyes, and mouth, replacing them with black pixels and constraining the cropping algorithm to avoid generating crops containing these pixels. This approach would enhance both classification performance and data anonymization. Existing tools like Eczemaless[1] are designed for this specific purpose, while OpenCV offers built-in functions that can also be utilized to achieve similar results. Furthermore, preprocessing can be enhanced by **imposing stricter criteria for crop acceptance**, including brightness and disease coverage. In particular, implementing a more effective automatic detection method for the Region of Interest (ROI) could improve data quality. For instance, Kim et al. [40] propose a semi-supervised method for acne segmentation, which may prove useful for identifying regions of interest in the context of skin rashes.

- Another critical improvement concerns achieving a **more robust labeling of skin tone**. As discussed in Chapter 3, the ITA metric is highly susceptible to image illumination, and despite employing a Gaussian Mixture Model (GMM) for automatic classification, a significant number of errors persist. Unsupervised and machine learning-based methods tend to be ineffective when dealing with shaded images, as they consistently produce distorted results, even when illumination-enhancing processing is applied. A potential solution would involve manual classification of skin tone by certified dermatologists, although this process is resource-intensive. Alternatively, automated classification could leverage a deep learning model specifically trained to identify skin tone from non-dermoscopic images. However, this

[1] `https://eczemaless.com/dermatology-image-anonymization/`

would necessitate training the model on a sufficiently diverse dataset of non-dermoscopic images to ensure reliable classification even under varying lighting conditions.

- A further enhancement could involve **generating disease images on dark and brown skin by transforming images of light skin conditions**. For example, a straightforward approach could employ a pix2pix model to colorize light skin images after converting them to grayscale to simulate dark skin (e.g., using the instruct pix2pix model by Brook et al. [12]). However, empirical observations during this study revealed that pretrained image-to-image models often fail to follow prompts requiring dark skin colorization, even after fine-tuning on dark skin datasets. This limitation likely arises from pretraining biases favoring light skin examples. Moreover, it is crucial to consider that the texture of certain skin conditions may significantly vary across skin tones, implying that simple colorization might not adequately capture the differences. An alternative to colorization would be using *Style Transfer* or *Deep Blending* techniques through deep learning methods, as suggested by Rezk et al. [51].

- **Incorporating certified dermatologists into the pipeline** is also essential, particularly in critical stages such as selecting the most promising models for image generation. In the current study, all synthetic images generated by the selected model were utilized to augment the dataset. However, in a real-world scenario, synthetic images should undergo expert evaluation to discard unrealistic samples. As previously mentioned, dermatologist involvement could also benefit the skin tone classification step or the definition of a Region of Interest (ROI).

These improvements would make the pipeline more robust and generalizable. Nevertheless, it is important to acknowledge the limitations imposed by the dataset, which presents considerable challenges due to the low quality of the images. Overall, despite these challenges, the proposed pipeline has proven effective in enhancing fairness in classification, thereby fulfilling its primary objective.

Appendix

- ## Analytical computation of the thresholds for the ITA values

  Each Gaussian distribution is given by the probability density function:

  $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  Where:

  - μ is the mean of the Gaussian,
  - σ^2 is the variance.

  The decision boundary x between two Gaussian distributions occurs where the probability densities of the two distributions are equal:

  $$\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

  Taking the natural logarithm of both sides, we get:

  $$-\frac{(x-\mu_1)^2}{2\sigma_1^2} = -\frac{(x-\mu_2)^2}{2\sigma_2^2} + \log\left(\frac{\sigma_1}{\sigma_2}\right)$$

  This simplifies to the quadratic equation:

  $$ax^2 + bx + c = 0$$

Where:

$$a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$$

$$b = -2\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)$$

$$c = \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} + 2\log\left(\frac{\sigma_1}{\sigma_2}\right)$$

The solution to this quadratic equation is:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The correct decision boundary is the root that lies between the means μ_1 and μ_2.

Bibliography

[1] Aayushman, H. Gaddey, V. Mittal, M. Chawla, and G. R. Gupta. Fair and accurate skin disease image classification by alignment with clinical labels. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 394–404, Cham. Springer Nature Switzerland, 2024. isbn: 978-3-031-72378-0.

[2] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, October 2018. url: https://proceedings.mlr.press/v80/agarwal18a.html.

[3] N. Alipour, T. Burke, and J. Courtney. Skin type diversity in skin lesion datasets: a review. *Current Dermatology Reports*, 13:1–13, August 2024. doi: 10.1007/s13671-024-00440-0.

[4] J. Archpaul, E. V, V. Ravi, P. S, T. Alahmadi, T. Stephan, P. Singh, and M. Diwakar. Deepscan: integrating vision transformers for advanced skin lesion diagnostics. *The Open Dermatology Journal*, 18, March 2024. doi: 10.2174/0118743722291371240308064957.

[5] I. Aristimuño. An introduction to diffusion models and stable diffusion, 2023. url: https://blog.marvik.ai/2023/11/28/an-introduction-to-diffusion-models-and-stable-diffusion/.

[6] A. Ashokan and C. Haas. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management*, 58(5):102646, 2021. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102646. url: https://www.sciencedirect.com/science/article/pii/S0306457321001369.

[7] W. Ba, H. Wu, W. Chen, S. Wang, Z. Zhang, X. Wei, W. Wang, L. Yang, D. Zhou, Y. Zhuang, Q. Zhong, Z. Song, and C. Li. Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images. *European Journal of Cancer*, 169:156–165, July 2022. doi: 10.1016/j.ejca.2022.04.015.

[8] D. Bardou, H. Bouaziz, L. Lv, and T. Zhang. Hair removal in dermoscopy images using variational autoencoders. *Skin Research and Technology*, 28, March 2022. doi: 10.1111/srt.13145.

[9] M. Benčević, M. Habijan, I. Galić, D. Babin, and A. Pizurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245:108044, March 2024. doi: 10.1016/j.cmpb.2024.108044.

[10] M. Benčević, M. Habijan, I. Galić, D. Babin, and A. Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245:108044, 2024. issn: 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2024.108044. url: https://www.sciencedirect.com/science/article/pii/S0169260724000403.

[11] A. Borghesi and R. Calegari. *Generation of clinical skin images with pathology with scarce data*. In *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health*. A. Shaban-Nejad, M. Michalowski, and S. Bianco, editors. Springer Nature Switzerland, Cham, 2024, pages 47–64. isbn: 978-3-031-63592-2. doi: 10.1007/978-3-031-63592-2_5. url: https://doi.org/10.1007/978-3-031-63592-2_5.

[12] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: learning to follow image editing instructions, 2023. arXiv: 2211.09800 [cs.CV]. url: https://arxiv.org/abs/2211.09800.

[13] A. Chardon, I. Cretois, and C. Hourseau. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13, 1991. url: https://api.semanticscholar.org/CorpusID:25650931.

[14] M. Charlton, S. A. Stanley, Z. Whitman, V. Wenn, T. J. Coats, M. Sims, and J. P. Thompson. The effect of constitutive pigmentation on the measured emissivity of human skin. *PLOS ONE*, 15(11):1–9, November 2020. doi: 10.1371/journal.pone.0241843. url: https://doi.org/10.1371/journal.pone.0241843.

[15] P.-C. Chen, H. Tsai, S. Bhojanapalli, H. W. Chung, Y.-W. Chang, and C.-S. Ferng. A simple and effective positional encoding for transformers, 2021. arXiv: 2104.08698 [cs.CL]. url: https://arxiv.org/abs/2104.08698.

[16] C.-H. Chiu, Y.-J. Chen, Y. Wu, Y. Shi, and T.-Y. Ho. Achieve fairness without demographics for dermatological disease diagnosis. *Medical Image Analysis*, 95:103188, 2024. issn: 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103188. url: https://www.sciencedirect.com/science/article/pii/S1361841524001130.

[17] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. Smith. Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In October 2015. isbn: 978-3-319-24887-5. doi: 10.1007/978-3-319-24888-2.

[18] A. Corbin and O. Marques. Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques. *IEEE Access*, 11:78339–78352, 2023. doi: 10.1109/ACCESS.2023.3289320.

[19] A. Corbin and O. Marques. Assessing bias in skin lesion classifiers with contemporary deep learning and post-hoc explainability techniques. *IEEE Access*, PP:1–1, January 2023. doi: 10.1109/ACCESS.2023.3289320.

[20] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. A. C. Allerup, U. Okata-Karigane, J. Zou, and A. S. Chiou. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32), August 2022. issn: 2375-2548. doi:

10.1126/sciadv.abq6147. url: http://dx.doi.org/10.1126/sciadv.abq6147.

[21] K. Delibasis, K. Moutselos, E. Vorgiazidou, and I. Maglogiannis. Automated hair removal in dermoscopy images using shallow and deep learning neural architectures. *Computer Methods and Programs in Biomedicine Update*, 4:100109, 2023. issn: 2666-9900. doi: https://doi.org/10.1016/j.cmpbup.2023.100109. url: https://www.sciencedirect.com/science/article/pii/S2666990023000186.

[22] M. Diaz, B. Lucke-Wold, S. Batchu, and G. Kleinberg. Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *Journal of Biomed Research*, 3:42–47, January 2022.

[23] K. Erdem. Step by step visual introduction to diffusion models. *Medium*, 2023. url: https://medium.com/@kemalpiro/step-by-step-visual-introduction-to-diffusion-models-235942d2f15c.

[24] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, January 2017. doi: 10.1038/nature21056.

[25] M. A. Farooq, W. Yao, M. Schukat, M. A. Little, and P. Corcoran. Derm-t2im: harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE, July 2024. doi: 10.1109/embc53108.2024.10781852. url: http://dx.doi.org/10.1109/EMBC53108.2024.10781852.

[26] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact, 2015. arXiv: 1412.3756 [stat.ML]. url: https://arxiv.org/abs/1412.3756.

[27] F. Filho, E. Santos, R. Mota, K. Cunha, F. Papais, A. Arruda, M. Baltazar, C. Vieira, J. G. Tavares, R. Barros, O. Souza, T. Bezerra, N. Lopes, É. Moutinho, J. Guido, S. Cruz, P. Borba, and T. I. Ren. An analysis of data variation and bias

in image-based dermatological datasets for machine learning classification, 2025. arXiv: 2501.08962 [cs.CV]. url: https://arxiv.org/abs/2501.08962.

[28] M. Ganthya. Convolutional neural networks in dermatology: skin cancer detection and analysis, August 2024. doi: 10.21203/rs.3.rs-4833522/v1.

[29] E. R. Gordon, M. H. Trager, D. Kontos, C. Weng, L. J. Geskin, L. S. Dugdale, and F. H. Samie. Ethical considerations for artificial intelligence in dermatology: a scoping review. *British Journal of Dermatology*:ljae040, 2024.

[30] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, Los Alamitos, CA, USA. IEEE Computer Society, June 2021. doi: 10.1109/CVPRW53098.2021.00201. url: https://doi.ieeecomputersociety.org/10.1109/CVPRW53098.2021.00201.

[31] Y. Guo, Z. Jia, J. Hu, and Y. Shi. FairQuantize: Achieving Fairness Through Weight Quantization for Dermatological Disease Diagnosis. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15010. Springer Nature Switzerland, October 2024.

[32] V. Gupta and V. K. Sharma. Skin typing: fitzpatrick grading and others. *Clinics in Dermatology*, 37(5):430–436, 2019. issn: 0738-081X. doi: https://doi.org/10.1016/j.clindermatol.2019.07.010. url: https://www.sciencedirect.com/science/article/pii/S0738081X1930121X. The Color of Skin.

[33] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J.L. Murray, and M. Naghavi. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *Journal of Investigative Dermatology*, 134(6):1527–1534, 2014. issn: 0022-202X. doi: https://doi.org/10.1038/jid.2013.446. url: https://www.sciencedirect.com/science/article/pii/S0022202X15368275.

[34] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. arXiv: 2006.11239. url: https://arxiv.org/abs/2006.11239.

[35] J. Ho and T. Salimans. Classifier-free diffusion guidance, 2022. arXiv: 2207.12598 [cs.LG]. url: https://arxiv.org/abs/2207.12598.

[36] S. Inthiyaz, B. R. Altahan, S. H. Ahammad, V. Rajesh, R. R. Kalangi, L. K. Smirani, M. A. Hossain, and A. N. Z. Rashed. Skin disease detection using deep learning. *Advances in Engineering Software*, 175:103361, 2023. issn: 0965-9978. doi: https://doi.org/10.1016/j.advengsoft.2022.103361. url: https://www.sciencedirect.com/science/article/pii/S0965997822002629.

[37] T. Kalb, K. Kushibar, C. Cintas, K. Lekadir, O. Diaz, and R. Osuala. Revisiting skin tone fairness in dermatological lesion classification, 2023. arXiv: 2308.09640 [eess.IV]. url: https://arxiv.org/abs/2308.09640.

[38] A. Kallipolitis, K. Moutselos, A. Zafeiriou, S. Andreadis, A. Matonaki, T. G. Stavropoulos, and I. Maglogiannis. Skin image analysis for detection and quantitative assessment of dermatitis, vitiligo and alopecia areata lesions: a systematic literature review. *BMC Medical Informatics and Decision Making*, 25(1):10, January 2025. doi: 10.1186/s12911-024-02843-2. url: https://doi.org/10.1186/s12911-024-02843-2.

[39] C. T. Kien. Explanation: swin transformer, 2023. url: https://chautuankien.medium.com/explanation-swin-transformer-93e7a3140877. Accessed: February 7, 2025.

[40] S. Kim, H. Yoon, and J. Lee. Semi-supervised facial acne segmentation using bidirectional copy–paste. *Diagnostics*, 14:1040, May 2024. doi: 10.3390/diagnostics14101040.

[41] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. F. Codella, R. Panda, P. Sattigeri, and K. R. Varshney. Fairness of classifiers across skin tones in dermatology. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 320–329, Cham. Springer International Publishing, 2020. isbn: 978-3-030-59725-2.

[42] Q. Kong, C.-H. Chiu, D. Zeng, Y.-J. Chen, T.-Y. Ho, J. Hu, and Y. Shi. Achieving fairness through channel pruning for dermatological disease diagnosis. In M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 24–34, Cham. Springer Nature Switzerland, 2024. isbn: 978-3-031-72117-5.

[43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: hierarchical vision transformer using shifted windows, 2021. arXiv: 2103.14030 [cs.CV]. url: https://arxiv.org/abs/2103.14030.

[44] M. M M, M. T R, V. V. Kumar, and S. Guluwadi. Enhanced skin cancer diagnosis using optimized cnn architecture and checkpoints for automated dermatological lesion classification. *BMC Medical Imaging*, 24, August 2024. doi: 10.1186/s12880-024-01356-8.

[45] A. Morales Forero, L. Rueda Jaime, S. Gil-Quiñones, M. Barrera Montañez, S. Bassetto, and E. Coatanéa. An insight into racial bias in dermoscopy repositories: a ham10000 data set analysis. *JEADV Clinical Practice*, 3:n/a–n/a, June 2024. doi: 10.1002/jvc2.477.

[46] L. Nair. Improved generation of synthetic imaging data using feature-aligned diffusion. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, MM '24, pages 25–30. ACM, October 2024. doi: 10.1145/3689096.3689460. url: http://dx.doi.org/10.1145/3689096.3689460.

[47] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. Soroushmehr, M. Jafari, K. Ward, and K. Najarian. Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1373–1376, 2016. doi: 10.1109/EMBC.2016.7590963.

[48] I. Pacal, M. Alaftekin, and F. Zengul. Enhancing skin cancer diagnosis using swin transformer with hybrid shifted window-based multi-head self-attention and

swiglu-based mlp. *Journal of Imaging Informatics in Medicine*, 37, June 2024. doi: 10.1007/s10278-024-01140-8.

[49] S. Paraddy and V. Patil. Addressing challenges in skin cancer diagnosis: a convolutional swin transformer approach. *Journal of imaging informatics in medicine*, October 2024. doi: 10.1007/s10278-024-01290-9.

[50] S. J. Prince. *Understanding Deep Learning*. MIT Press, 2023. url: http://udlbook.com.

[51] E. Rezk, M. Eltorki, and W. El-Dakhakhni. Improving skin color diversity in cancer detection: deep learning approach. *JMIR Dermatology*, 5:e39143, August 2022. doi: 10.2196/39143.

[52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022. arXiv: 2112.10752 [cs.CV]. url: https://arxiv.org/abs/2112.10752.

[53] O. Ronneberger, P. Fischer, and T. Brox. U-net: convolutional networks for biomedical image segmentation, 2015. arXiv: 1505.04597 [cs.CV]. url: https://arxiv.org/abs/1505.04597.

[54] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation, 2023. arXiv: 2208.12242 [cs.CV]. url: https://arxiv.org/abs/2208.12242.

[55] S. Salti. Generative models. Lecture slides, University of Bologna, 2024. Accessed: February 15, 2024.

[56] V. A. Saputra, M. S. Devi, Diana, and A. Kurniawan. Comparative analysis of convolutional neural networks and vision transformers for dermatological image classification. *Procedia Computer Science*, 245:879–888, 2024. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2024.10.315. url: https://www.sciencedirect.com/science/article/pii/S1877050924031235. 9th International Conference on Computer Science and Computational Intelligence 2024 (ICCSCI 2024).

[57] A. K. Sharma, S. Tiwari, G. Aggarwal, N. Goenka, A. Kumar, P. Chakrabarti, T. Chakrabarti, R. Gono, Z. Leonowicz, and M. Jasiński. Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network and handcrafted features based deep neural network. *IEEE Access*, 10:17920–17932, 2022. doi: 10.1109/ACCESS.2022.3149824.

[58] V. Shavlokhova, A. Vollmer, C. Zouboulis, M. Vollmer, J. Wollborn, G. Lang, A. Kübler, S. Hartmann, C. Stoll, E. Roider, and B. Saravi. Finetuning of glide stable diffusion model for ai-based text-conditional image synthesis of dermoscopic images. *Frontiers in Medicine*, October 2023. doi: 10.3389/fmed.2023.1231436.

[59] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. arXiv: 1503.03585 [cs.LG]. url: https://arxiv.org/abs/1503.03585.

[60] Steinsfu. Stable diffusion clearly explained, 2023. url: https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e.

[61] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, N. Shinkai, H. Machino, K. Kobayashi, K. Asada, M. Komatsu, S. Kaneko, M. Sugiyama, and R. Hamamoto. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. *Journal of Medical Systems*, 48:84, September 2024. doi: 10.1007/s10916-024-02105-8.

[62] K. Tang, J. Su, R. Chen, R. Huang, M. Dai, and Y. Li. Skinswinvit: a lightweight transformer-based method for multiclass skin lesion classification with enhanced generalization capabilities. *Applied Sciences*, 14:4005, May 2024. doi: 10.3390/app14104005.

[63] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), August 2018. issn: 2052-4463. doi: 10.1038/sdata.2018.161. url: http://dx.doi.org/10.1038/sdata.2018.161.

[64] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15, 2024.

[65] S. C. Wong, W. Ratliff, M. Xia, C. Park, M. Sendak, S. Balu, R. Henao, L. Carin, and M. K. Kheterpal. Use of convolutional neural networks in skin lesion analysis using real world image and non-image data. *Frontiers in Medicine*, 9, 2022. issn: 2296-858X. doi: 10.3389/fmed.2022.946937. url: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.946937.

[66] Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7(1):286, 2024.

[67] H. Yuan, A. Hadzic, W. Paul, D. V. de Flores, P. Mathew, J. Aucott, Y. Cao, and P. Burlina. Edgemixup: improving fairness for skin disease classification and segmentation. *arXiv preprint arXiv:2202.13883*, 2022.

[68] R. Zhang, Y. Yao, Z. Tan, Z. Li, P. Wang, H. Liu, J. Hu, S. Liu, and T. Chen. Fairskin: fair diffusion for skin disease image generation, 2024. arXiv: 2410.22551 [cs.CV]. url: https://arxiv.org/abs/2410.22551.

[69] F. Zhong, K. He, M. Ji, J. Chen, T. Gao, S. Li, J. Zhang, and C. Li. Optimizing vitiligo diagnosis with resnet and swin transformer deep learning models: a study on performance and interpretability. *Scientific Reports*, 14, April 2024. doi: 10.1038/s41598-024-59436-2.