Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University M'Hamed BOUGARA – Boumerdes

Institute of Electrical and Electronic Engineering
Department of Electronics

Final Year Project Report Presented in Partial Fulfillment of the Requirements for the Degree of

MASTER

In **Computer Engineering**

Title:

Fair Out-Of-Distribution Detection for Addressing Skin Tone Representation in Dermatology

Presented by:

- **Assala BENMALEK**

Supervisor:

- **Prof. Dalila CHERIFI**

Co-supervisor:

- **Dr. Celia CINTAS**

Registration Number:……/2024

Abstract

Addressing representation issues in dermatological settings is crucial due to variations in how skin conditions manifest across skin tones, thereby providing competitive quality of care across different segments of the population. Although bias and fairness assessment in skin lesion classification has been an active research area, there is substantially less exploration of the implications of skin tone representations and Out-of-Distribution (OOD) detectors' performance. Current OOD methods detect samples from different hardware devices, clinical settings, or unknown disease samples. However, the absence of robustness analysis across skin tones questions whether these methods are fair detectors. As most skin datasets are reported to suffer from bias in skin tone distribution, this could lead to higher false positive rates in a particular skin tone. This research presents a framework to evaluate OOD detectors across different skin tones and scenarios. We review and compare state-of-the-art OOD detectors across two categories of skin tones, FST I-IV (lighter tones) and FST V-VI (brown and darker tones), over samples collected from dermatoscopic and clinical protocols. We conducted a Gray-Level Co-Occurrence Matrix (GLCM) texture analysis on "Fitzpatrick17k dataset" samples from two main skin tone categories FST I-IV and FST V-VI, and compared statistical parameters across skin tone categories and nine skin conditions. This analysis indicates that FST V-VI textures are more heterogeneous and varied, while FST I-IV textures are more uniform and consistent. Our OOD detection experiments yield that in poorly performing OOD models, the representation gap measured between skin types is wider (from $\approx 10\%$ to 30%) up for samples from darker skin tones. Compared to better performing models, skin type performance only differs for $\approx 2\%$. Furthermore, this work shows that understanding OOD methods' performance beyond average metrics is critical to developing more fair approaches. We used the AIF360 tool to assess fairness in our OOD detectors and evaluated their performance with group fairness metrics. Our observations show that models with similar overall performance can have significant differences in representation gaps, with group fairness metrics correlating negatively with the representation gap. This indicates that increasing the representation of FST V-VI leads to improved group fairness resulting in fairer OOD detectors.

Keywords: *Algorithmic Fairness, Skin Tone Representation, Out-Of-Distribution detection, Texture Analysis, Dermatology.*

Dedications

”In the Name of Allah, the Most Merciful, the Most Compassionate All praise be to Allah, the Lord of the worlds. May prayers and peace be upon the Prophet Muhammad, His servant and messenger. This study is dedicated wholeheartedly to my beloved parents, who have been a constant source of inspiration, providing me with hope and encouragement when I faced moments of doubt and despair. They have always been there, guiding and supporting me.”

”ASSALA”

Acknowledgments

First and foremost, I would like to extend my deepest gratitude to God Almighty who provided me with his blessing and the opportunity to successfully conclude my project.

In the successful accomplishment of my final year project titled ”Fair Out-Of-Distribution Detection For Addressing Skin Tone Representation In Dermatology”, I would like to express my deepest gratitude to my supervisors, **Prof. Dalila CHERIFI** and **Dr. Celia CINTAS**. I am immensely grateful for their valuable guidance regarding my research and future career prospects. I deeply appreciate their patience, dedication, and unwavering belief in my abilities.

I also want to thank **IBM Research Africa** for giving me this research opportunity to ”Go Research!” and dive deeper into the research field which allowed me to grow my professional skills and patience towards the field. Lastly, I am immensely thankful to my parents, my mom for encouraging me to study since primary school, and my dad for his patience with me all these years. I would like to thank Dr. Muhammad Al-Zafar Khan for his continuous support along my journey. I immensely thank all my friends namely Wafa, Lamine, Belkacem, Chaima, Abdelghani, Meroua, Baha, Ihssene, and All the ISC family for their unwavering support along with my university journey and for making the past years very exciting.

List of Figures

List of Tables

List of Abbreviations

GLCM: Gray-Level-Co-Occurrence Matrix

ML: Machine Learning

DL: Deep Learning

NN: Neural Networks

OOD: Out Of Distribution

AE: AutoEncoder

IF: Isolation Forest

OneSVM: One-Class Support Vector Machine

ODIN: Out-of-Distribution detector for Neural networks.

ID: In Distribution

IDD: In Distribution Dataset

FST: Fitzpatrick Skin Tone

FST I-IV: Fitzpatrick light skin tones

FST V-VI: Fitzpatrick brown and dark skin tones

Contents

General Introduction

Skin diseases remain a global health challenge, with skin cancer being the most common cancer worldwide. Following the recent success of Deep Learning (DL) in various computer vision problems, Convolutional Neural Networks (CNNs) have been employed for skin disease classification with improved performance. However, DL models have been shown to be prone to and exacerbate existing societal biases [1, 2]. Thus, as we observe increasing interest in DL for dermatology [3, 4, 5], it is imperative to address the transparency, robustness, and fairness of these solutions to make them adopted clinically for positive societal impact [6, 7, 8, 9, 10]. In dermatology, bias in representations of skin tones in academic materials and clinical care is becoming a primary concern [11, 9]. Recent studies report major disparities in dermatology when treating skin of color as common conditions often manifest differently on dark skin, and physicians are trained mostly to diagnose them on light skin [11, 12]. The growing practice of using machine learning algorithms to aid the diagnosis of skin diseases will further deepen the divide in patient care because these algorithms are trained with such imbalanced datasets [8], with an overwhelming majority of samples with light skin tones. Particularly, when we look at robustness, we are interested in the ability of the models to identify Out-of-Distribution (OOD) samples that differ from the training distribution. For example, OOD samples may come from new skin conditions, different collection protocols [13], or heterogeneous patient sub-populations. However, the fairness of these OOD detection methods has not been explored in the existing literature. Ensembles of multiple models are utilized aiming at maximising performance with limited consideration of shifts in the input data. The absence of fairness in these methods might result in incorrectly classifying new samples with high confidence because these samples might be from previously unknown classes. On the other hand, rare in-distribution samples might still be picked as OOD samples. Thus, it is necessary to detect out-of-distribution (OOD) samples before making decisions to achieve principled transfer of knowledge from in-distribution (ID) training samples to OOD test samples, thereby extending the usability of the models to pre-

viously unseen scenarios. In the dermatological clinical setting, Lam et al. [14] showed that reduced survival rates for ethnic minority patients compared to Caucasian patients may be attributed to several contributing factors. First, a longer time to diagnosis and, as more advanced stages, can lead to a poorer prognosis [15, 16, 17]. Dick et al. [18] found that black patients were significantly more likely to present with advanced-stage disease, even after adjusting for tumor characteristics and demographic factors. Second, socioeconomic status differences may lead to increased barriers that limit access to medical care in minority populations. Communities of lower socioeconomic status tend to have a decreased density of dermatologists, further increasing the disparity of access to care [19]. Therefore, OOD detectors need to guarantee equivalent detection capability across different sub-populations.

In this work, we aim to work towards quantifying and evaluating the detection disparity across skin tones in OOD detectors in different clinical scenarios, study the texture of each skin tone category, and perform a fairness assessment on these OOD detectors. We are interested in answering questions such as: *Are there differences in skin texture among different skin tones? how much does the skin tone representation of the In-Distribution Dataset (IDD) impact the OOD overall performance*? *Do we observe changes in performance for different skin types? Is the average performance of an OOD method a fair measurement across skin tones? How do skin tones differ in terms of texture? How fair are the OOD detectors across different skin tones?*

This report consists of three chapters. The first chapter provides an overview of skin tone diversity and highlights the primary skin conditions and the challenges associated with skin tone representation in dermatology. The second chapter details our methodology for out-of-distribution (OOD) characterization and detection. Finally, the third chapter presents our experiments, discusses the main results, and highlights the key findings of the study.

Chapter I

Skin Tones Representation Overview

I.1 Introduction

Skin is the largest organ in the human body and reflects a remarkable diversity of genetic, environmental, and individual factors. A huge variety of skin is represented due to the different parameters of skin representation that classify the skin into various types and colors leading to the identification of new types of skin diseases and unusual symptoms. This chapter introduces the skin type representation parameters, skin disease identification, and skin representation problems.

I.2 Skin organ and pigmentation

The skin organ is a dynamic anatomical feature, and a responsive entity that plays a pivotal role in human life, consisting of multiple layers intricately designed to protect, regulate, and communicate, the skin serves as a sensory interface between the body and its surroundings. The skin is a sentinel to the overall health, from maintaining temperature balance to acting as a resilient barrier against external threats. Its remarkable adaptability allows it to respond to genetic, environmental, and individual factors, reflecting individual uniqueness. As we delve into the complexities of the skin organ, we must acknowledge its vulnerability to various diseases [20]. Figure I.1 illustrates the key components of skin anatomy and its physiological functions.

The skin is composed of three layers: **the epidermis**, **dermis**, and **hypodermis**. The epidermis forms the outermost layer and acts as a protective barrier against external factors such as pathogens and UV rays. It consists of five sub-layers: Basal layer (basal layer), spiny layer (spiny layer), granular layer (granular layer), clear layer (clear layer), and outer horny layer (horny layer). The major cells within the epidermis include keratinocytes, melanocytes, Langerhans cells, and Merkel cells. The dermis lies beneath the epidermis and consists of the papillary and reticular layers that include structures such as sweat glands, hair follicles, and sensory neurons. The deepest layer, the subcutaneous layer, contains adipose tissue and acts as a link between the skin and the underlying tissues. These layers collectively form an extensive network that regulates temperature, protects against injury, and plays a crucial role in sensory perception.

Skin pigmentation refers to the amount of melanin generated in the body and the color of the skin. The two main types of melanin, eumelanin, and pheomelanin are produced by

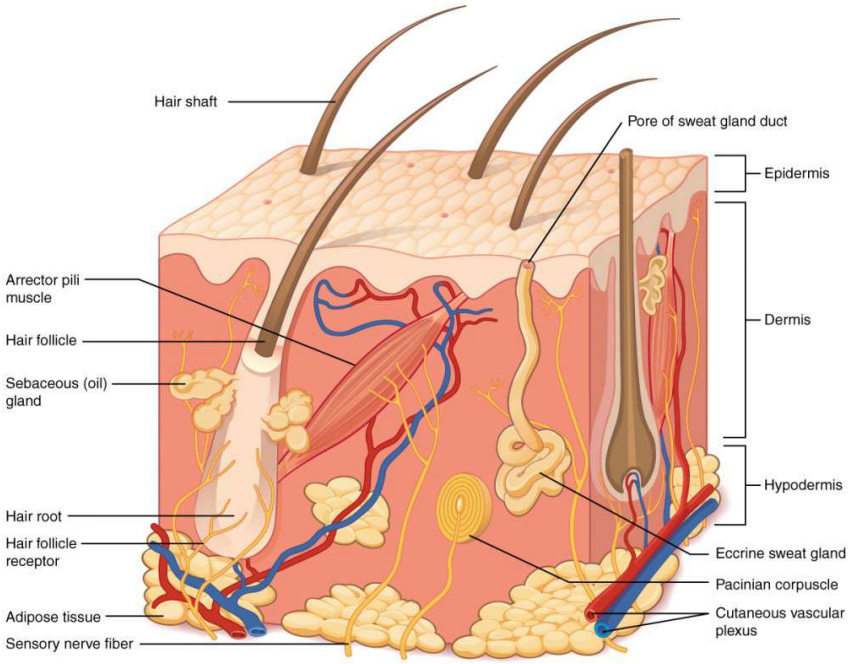


Figure I.1: Schematic representation of the anatomy of human skin [20].

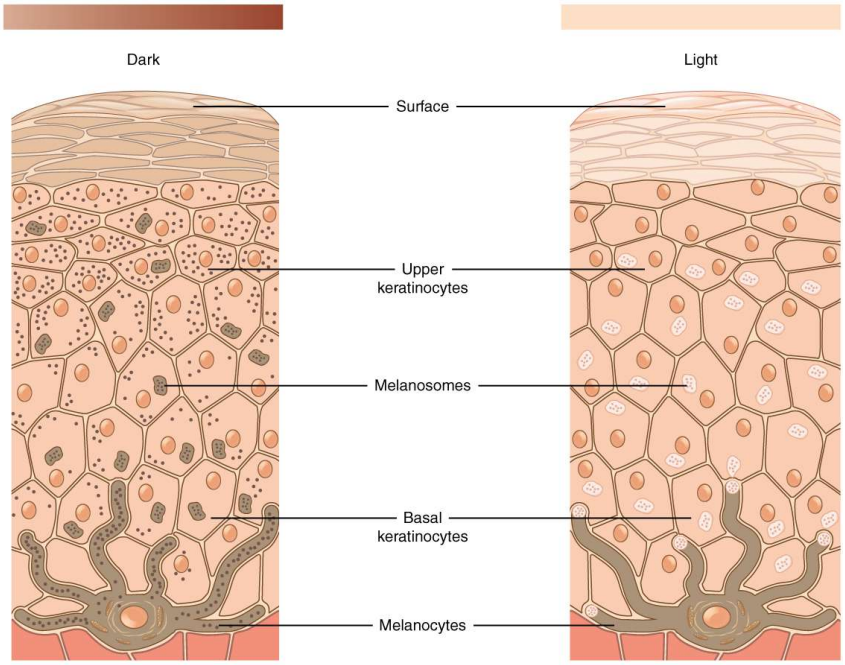


Figure I.2: Skin layers and pigmentation, the amount of melanin generated by melanocytes in the melanocyte basal layer and absorbed by keratinocytes determines the skin's relative coloration [20].

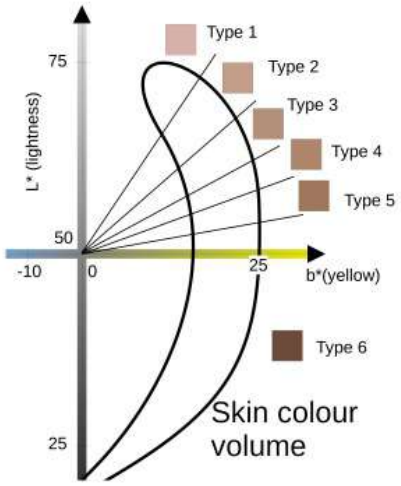
melanocytes in the epidermal layer of the skin. Melanin synthesis and skin pigmentation are mostly impacted by genetics, UV exposure, and some drugs [21]. Many pigments such as hemoglobin, beta-carotene, and melanin affect skin color [20].

I.3 Skin types definitions and diversity

The accurate determination of skin type in relation to understanding sunburn risk, skin cancer, and clinical treatments. Due to the increasing use of laser applications in both cosmetic and medical industries, skin type determination is important. Accurate skin typing is necessary to understand an individual's personal sunburn risk, which is directly related to the risk of developing skin cancer. The different skin types are defined based on two main measurements which are the Individual Typology Angle (ITA) and the Melanin Index (MI) [10].

1. **Individual Typology Angle (ITA)**: a measure of skin pigmentation, used to classify skin types based on spectrophotometric measurements. It categorizes skin types into physiologically relevant groups, ranging from very light to dark skin tones. Lower ITA values represent darker-pigmented skin.
2. **Melanin Index (MI)**: a measure of melanin content, with higher values representing darker pigmented skin.

   Studies have indicated a robust association between Individual Typology Angle (ITA) values and Melanin Index (MI), implying that both methods are viable for appraising skin pigmentation. Skin pigmentation is measured subjectively using several color scales, which are influenced by an individual's bias in how they interpret stimuli. Hence, the Fitzpatrick scale has been widely used for the categorical assessment of skin sensitivity and reactivity to sun exposure and therefore for skin color assessment.



Skin Type	ITA range
Type 1	$\text{ITA} > 55°$
Type 2	$55° \geq \text{ITA} > 41°$
Type 3	$41° \geq \text{ITA} > 28°$
Type 4	$28° \geq \text{ITA} > 19°$
Type 5	$19° \geq \text{ITA} > 10°$
Type 6	$10° \geq \text{ITA}$

Figure I.3: (a) Skin color volume in the L* - b* plane of CIELab color space with ITA thresholds (b) Skin types from ITA thresholds [10].

The **Fitzpatrick** skin type classification categorizes skin into six types based on an individ-

ual's response to ultraviolet (UV) light exposure, particularly focusing on sunburn and tanning ability. This scale ranges from Type I (very fair skin that always burns and never tans) to Type VI (deeply pigmented dark brown or black skin that never burns). The Fitzpatrick scale is widely used in dermatology to assess skin cancer risk, plan phototherapy treatments, and guide cosmetic procedures involving lasers and other light-based technologies [22].

I.4 Skin diseases overview

Skin diseases encompass a wide range of conditions that affect the skin. These conditions can manifest in various forms, ranging from cancerous to non-cancerous infections. The spectrum includes malignant growths, as well as benign conditions, and bacterial or viral infections. Each type of skin disease presents unique symptoms across different skin tones.

I.4.1 Malignant skin conditions

Malignant skin conditions refer to cancerous tumors that grow in the skin tissue. These conditions are characterized by abnormal cell growth that can invade surrounding tissues and potentially spread to other parts of the body if left untreated. Thus, early detection and treatment are crucial in dealing with malignant skin conditions to prevent further complications [23]. There are different types of malignant skin conditions:

1. **Malignant Dermal:** skin conditions that refer to cancerous growths originating in the dermal layer of the skin, which is the deeper layer beneath the epidermis, and developing in the skin tissues. These cancers can be aggressive and have the potential to metastasize to other parts of the body if not detected and treated early. Dermatofibrosarcoma protuberans (DFSP) is a rare, slow-growing soft tissue tumor that arises from cells in the deepest layer of the skin (dermis). It typically appears as a flat or slightly raised skin patch that feels rubbery or hard, often violet, reddish-brown, or skin-colored. As it progresses, it may form lumps near the skin's surface. DFSP rarely metastasizes beyond the skin. On the other hand, Angiosarcoma is a highly invasive and highly malignant tumor composed of tumor endothelial cells. It can occur in individuals of any age with no known gender predominance. Characterized by slow growth with central necrosis, Angiosarcoma results in skin changes overlying it. It appears as multiple purple nodules in the affected muscular skin and is associated with Lymphedema.

2. **Malignant Epidermal:** Skin conditions are considered cancers that develop in the epidermal layer of the skin, the outermost layer responsible for protecting the body from external factors. These cancers often result from the abnormal growth and division of skin cells. The malignant epidermal types include **Basal cell carcinoma** and **Squamous cell carcinoma**. Basal cell carcinoma is the most common form of skin cancer caused by sun damage. It typically presents as small bumps or open sores on the skin and is slow-growing. If not removed, it can spread into local underlying tissues. Squamous cell carcinoma is another common type of skin cancer that often starts in the outermost layer of the skin. It can appear as scaly, reddish patches that may be crusted and can be mistaken for other skin conditions like rashes or eczema. It is often found in sun-exposed areas such as the ears, face, scalp, neck, and hands due to prolonged exposure to UV rays. If not treated, it can grow inward and spread to the interiors of the body.
3. **Malignant Melanoma:** One of the most common Melanoma skin diseases is a malignant tumor that arises from the uncontrolled proliferation of melanocytes —pigment-producing cells. The most common form of melanoma is cutaneous, it can also occur in mucosal surfaces, the uveal tract, and leptomeninges. For decades, melanoma incidence has progressively risen and is projected to continue to rise across the world, while it still represents less than %5 of all cutaneous malignancies, melanoma accounts for the majority of skin cancer deaths. It demonstrates greater variation in incidence rates across different ethnic groups and geographical locations. Melanoma is a malignant tumor that is aggressive and usually spreads beyond its original location which makes it more difficult to treat. However, if Melanoma is detected in its early stages, resection of the lesion can contribute to good survival rates. This demonstrates the necessity of the early detection of this skin disease. Epidemiological studies were done on different populations to identify and characterize the most greatly affected groups, the reasons for their disease, and the treatment management. Its capacity to rapidly metastasize and affect younger patients makes melanoma a significant health and economic burden on society.

Figure I.4 demonstrates the different types of malignant skin conditions:

Figure I.4: Malignant skin conditions [24],[25].

I.4.2 Benign skin tumors

Benign skin tumors cover a variety of non-cancerous growths on the skin, characterized by their lack of both invasiveness and malignant potential. Unlike their malignant counterparts, benign skin tumors cannot metastasize or spread to other body parts. These tumors can originate from different skin types of cells and are often differentiated by their clinical and histological characteristics. Benign skin tumors encompass a spectrum of entities, each with unique characteristics. Common types include dermatofibromas, seborrheic keratosis, lipomas, and nevi (moles) [26, 27].

1. **Benign dermal:** non-cancerous skin growths originating in the dermis, the middle layer of the skin. They can be classified into various types, such as **Dermatofibromas**, **Epidermoid cysts**, and **Lipomas**.

   Dermatofibromas are small, firm, red, or brown bumps caused by an accumulation of fibroblasts under the skin. They are often found on the legs and may itch, being more common in women. Epidermoid cysts are follicular nodules with a central punctum, that appear skin-colored to off-white, dome-shaped, and containing cheesy or yellowish keratin. Lipomas are benign tumors containing fat cells (adipocytes) that are typically slow-growing and non-cancerous. These tumors are usually round, soft, and rubbery lumps located beneath the skin. They are often painless and can be moved with gentle pushing. Most lipomas are commonly found on the upper back, shoulders, arms, buttocks, and upper thighs. Lipomas can occur at any age but are more common between 40 and 60 years old. They do not typically change after formation and have minimal potential for becoming cancerous.
2. **Benign epidermal:** non-cancerous growths that occur in the epidermis, the outermost layer of the skin. These include dermoid cysts, freckles, and seborrheic keratoses. Dermoid cysts are composed of hairs, sweat glands, sebaceous glands, and sometimes cartilage, bone fragments, and teeth. Freckles are darkened, flat spots that appear on sun-exposed areas, com-

mon in individuals with blond or red hair. Seborrheic keratoses are variable warty plaques with a dull, verrucous, or waxy surface, often appearing stuck-on.

3. **Benign Melanocyte:** non-cancerous growths originating from melanocytes, the pigment-producing cells in the skin. These include **Moles** (nevi), **Atypical moles** (dysplastic nevi), and **Pyogenic granulomas**.

   Moles (nevi) are small skin marks caused by pigment-producing cells, varying in color, size, and appearance. Atypical moles (dysplastic nevi) are larger than normal moles, not always round, and can range from tan to dark brown on a pink background. They may occur anywhere on the body. Pyogenic granulomas are red, brown, or bluish-black raised marks due to excessive capillary growth, often forming after skin injury and bleeding easily.

Figure I.5 demonstrates the various types of benign skin conditions:

Figure I.5: Benign skin conditions [24, 25].

I.4.3 Inflammatory skin conditions

Inflammatory skin conditions, are characterized by rashes and skin eruptions and can resolve on their own in a few weeks or turns into a chronic condition that lasts for years. At the tissue level of the skin, there are varying degrees of normal white blood cell accumulation, which are responsible for immunity against infection. This procedure can change the appearance of the skin causing other changes in the skin color and texture. Furthermore, there could be microscopic blood vessel dilatation in the skin, resulting in redness (erythema) [28]. The most commonly known inflammatory skin conditions are: **Eczema**, **Psoriasis**, **Acne**, **Folliculitis** and **Allergic urticaria**.

I.4.4 Genetic skin disorders

Hereditary skin disorders are genetic conditions that predominantly affect the skin and are caused by genetic mutations or abnormalities. They can manifest in different ways, result-

ing in unique changes and symptoms on the skin. Some common genetic skin disorders like **Hereditary skin disorders** are genetic conditions caused by mutations or abnormalities that primarily affect the skin. These disorders exhibit unique changes and symptoms on the skin. Another common genetic skin disorder is albinism, which results in pale skin, light hair, and vision issues due to the lack of melanin synthesis in the skin, hair, and eyes. **Ectodermal dysplasias** are another group of genetic diseases that impact the growth of ectodermal structures such as hair, teeth, nails, and sweat glands. **Ehlers-Danlos Syndrome** (classic type) is an inherited connective tissue disease characterized by hypermobility in the joints and easy bruising. **Ichthyoses** are hereditary skin conditions that cause dry, flaky skin resembling fish scales due to abnormal skin cell turnover and shedding. Tuberous sclerosis, on the other hand, is a genetic condition that can lead to non-cancerous tumors in the skin, brain, kidneys, heart, and other organs. Figure I.6 demonstrates the genetic and inflammatory skin conditions [29].

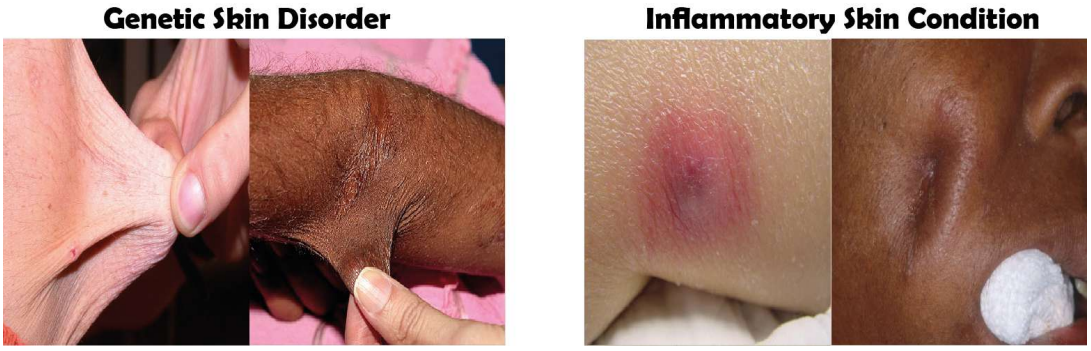


Figure I.6: Genetic and Inflammatory skin conditions [24, 25].

I.5 Diversity of skin tones and skin conditions diagnosis

Dermatologists employ various techniques and considerations to diagnose skin conditions in individuals with different skin tones.

Visual examination is an important initial step in diagnosing the different skin conditions. Dermatologists rely on their expertise to carefully inspect the affected area, looking for specific symptoms such as rashes, discoloration, bumps, or lesions. However, certain skin conditions may appear differently depending on the individual's skin color. These variations across skin tone presentation is known by the dermatologists based on their expertise and experience with the nuances of different ethnicities. Additionally, dermatoscopy is a common, non-invasive technique that enables dermatologists to examine the skin more closely. Dermatoscopy involves the use of a specialized magnifying instrument called a dermatoscope, which enhances the

view of the skin's surface and structures. Dermatologists can use dermatoscopy to distinguish between benign skin tumors and potentially malignant tumors, especially in pigmented skin lesions. In complex cases or instances where a diagnosis is challenging, dermatologists may seek input from colleagues or specialists with expertise in specific skin conditions or different skin tones.

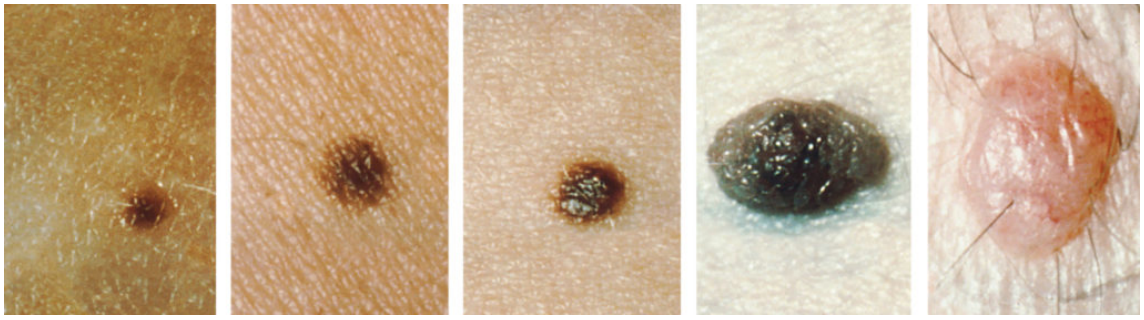


Figure I.7: Moles range from benign accumulations of melanocytes to melanomas across skin tones[20].

I.6 Hardware and illumination settings effects on skin datasets acquisition

Dermatologists are working diligently to capture various images of skin lesions. **Clinical photography** is widely used to provide illustrations and monochrome photographs. However, the most popular technique nowadays is color photography, which uses a camera equipped with a ring-type electric flash. This method removes blur and produces crisp, color-balanced images through an electric light burst lasting milliseconds. Moreover, placing the light source in front of the lens tube eliminates blurring caused by the lens tube in close-up shots. In addition to traditional photography, **dermoscopic imaging** has emerged as a crucial tool in dermatology. Dermoscopy, or the use of a dermoscope, allows for the detailed examination of skin lesions by magnifying the surface structure and providing polarized light to minimize surface reflections. This technique enhances the visualization of features such as pigment networks and vascular patterns that are not easily seen with standard photography. However, traditional color photography has several disadvantages. It does not always reproduce the image as perceived by the naked eye, such as the absence of redness in the peripheral area of erythema multiforme lesions or the difficulty in identifying xerosis and roughness due to the lack of minute scaling records in atopic skin. Additionally, varying illumination conditions and clinical settings of

conventional photographic systems introduce uncertainty in the produced image information, affecting texture, contrast, and light intensity, which impacts color accuracy.

(a) Clinical hardware

(b) Dermoscopic hardware

Figure I.8: Dermoscopic and clinical hardware [30].

I.7 Lack of skin type representations in dermatology datasets

Skin datasets frequently have challenges and restrictions that need to be addressed in various applications, including computer vision, dermatology research, and machine learning methods. The under-representation of skin tones, particularly darker skin tones, and the absence of diversity in skin categories are major problems. Lighter skin tones' over-representation in most skin datasets is one of the main challenges in dermatology research. This imbalance might provide erroneous conclusions and skewed outcomes when investigating skin problems or creating algorithms based on these data. It may lead to a lack of awareness and comprehension of the distinctive traits and variances in skin problems among various ethnic groups causing the lack of understanding of the specific needs and responsiveness of different skin types to treatments.

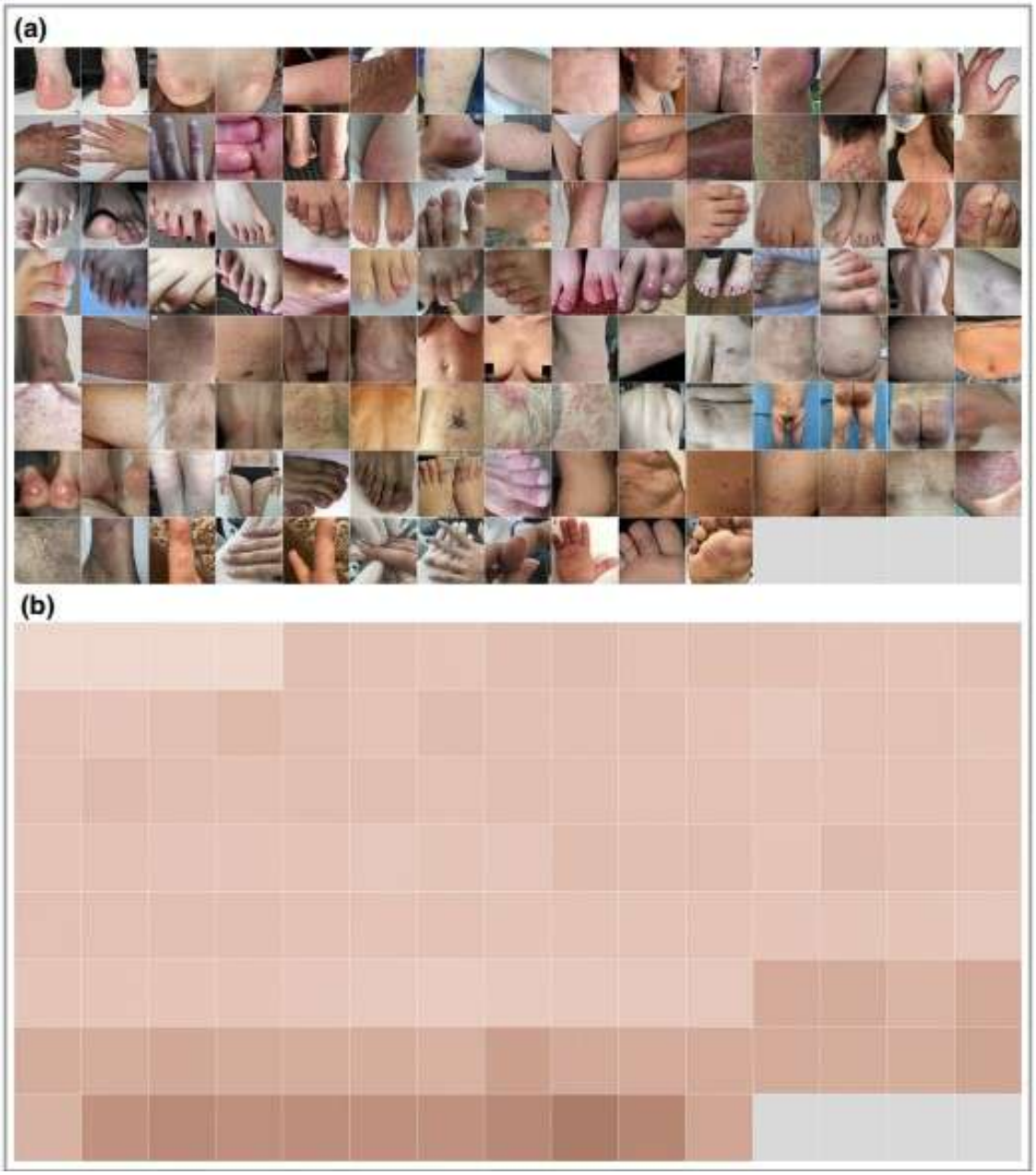


Figure I.9: (a) Images of cutaneous manifestations associated with COVID-19, (b) Approximation of the skin color of the patients [31].

I.8 Summary

In this chapter, we have covered an overview of the various skin tones and skin diseases we will deal with in the experiments. In addition to the diversity challenges, we are addressing in this work.

Chapter II

Out-Of-Distribution Characterization Methods

II.1 Introduction

Anomaly detection is one of the most essential aspects of machine learning and deep learning applications. As machine learning and deep learning models may encounter unusual or unseen data that differ from their initial training data, anomaly detection is extremely significant for overcoming model failures and biases. Identifying anomalies is essential to maintaining the performance and trustworthiness of the models in the face of irregular or unexpected data points. Datasets, on the other hand, are an essential part of the training process as they provide a special distribution of the models' inputs. Identifying out-of-distributions is becoming a significant aspect of ensuring fair and trustworthy AI solutions in different applications. In this chapter, we delve into several essential components for understanding and improving Out-of-Distribution (OOD) detection and analysis in Dermatology. These components include texture analysis methodology, various OOD detection techniques, and AI fairness evaluation.

II.2 Texture Analysis using Gray-Level Co-Occurrence Matrix (GLCM) in Dermatology

Texture analysis is a process that involves extracting and quantifying the textural features of an image. These features are used to classify, segment, and synthesize images. The primary methods used for texture analysis are statistical, structural, model-based, and transform-based approaches. Among these, the gray-level co-occurrence matrix (GLCM) is a widely used statistical method that characterizes the textures of an image by calculating the joint probability of pixel pairs with specific values and spatial relationships. GLCM or Gray-level Co-Occurrence matrix [32] is a statistical method of analyzing texture that takes into consideration the spatial information of pixels, commonly referred to as the gray-level spatial dependence matrix. The GLCM functions analyze the spatial relationship between two pixels with certain intensities in an image and determine how often these pixel pairs occur in the image which generates a GLCM, then build statistics metrics from this matrix. Many statistical properties can be generated from GLCMs which are useful for getting textural details of an image. GLCM texture analysis has been implemented in various skin disease diagnoses and human skin texture analysis. Li et al. [33] demonstrates the application of GLCM-based texture features for the automated detection of various skin diseases. The authors show the effectiveness of these tex-

ture descriptors in discriminating between different skin conditions, highlighting the potential of GLCM analysis for skin disease diagnosis. Hazani et al. [34] used the Gray-Level Co-Occurrence Matrix (GLCM) features and Decision Tree classifier for analyzing human skin texture. Nikita O. et al. [35] used GLCM texture analysis for feature extraction for skin disease recognition. Most of the texture analysis methods in dermatology were addressing skin condition identification making the skin tone representation rarely addressed using GLCM texture analysis.

We will use the GLCM co-occurrence matrix in our study to analyze the different textural features of skin images in dermatology datasets based on diverse skin tone categories [36, 37, 38, 39, 40] . The GLCM formula is presented in Equation II.1:

$$\text{GLCM}(i,j,d,\theta) = \sum_{(x,y)\in I} \begin{cases} 1, & \text{if } I(x,y) = i \ \wedge \ I[x + d\cos\theta, y + d\sin\theta] = j, \\ 0, & \text{otherwise.} \end{cases} \tag{II.1}$$

Where $I(x,y)$ is the pixel intensity at position (x,y) in the image, d is the distance, and θ is the angle.

To conduct a comprehensive texture analysis across diverse skin tones, we use the following statistical texture parameters provided by the `scikit-image` package:

- **Dissimilarity:**

$$\text{Dissimilarity} = \sum_{i,j} |i-j| \cdot p(i,j) \tag{II.2}$$

- **Correlation:**

$$\text{Correlation} = \frac{\sum_{i,j}(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \tag{II.3}$$

- **Homogeneity:**

$$\text{Homogeneity} = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \tag{II.4}$$

- **Energy:**

$$\text{Energy} = \sum_{i,j} p(i,j)^2 \tag{II.5}$$

- **Contrast:**

$$\text{Contrast} = \sum_{i,j} (i-j)^2 \cdot p(i,j) \tag{II.6}$$

- **Skewness:**

$$\text{Skewness} = \frac{\sum_{i,j}(i-\mu)^3 \cdot p(i,j)}{\sigma^3} \tag{II.7}$$

- **Kurtosis:**

$$\text{Kurtosis} = \frac{\sum_{i,j}(i-\mu)^4 \cdot p(i,j)}{\sigma^4} \tag{II.8}$$

- **Mean:**

$$\text{Mean} = \mu = \sum_{i,j} i \cdot p(i,j) \tag{II.9}$$

- **Variance:**

$$\text{Variance} = \sigma^2 = \sum_{i,j}(i-\mu)^2 \cdot p(i,j) \tag{II.10}$$

Where i and j are the row and column indices of the GLCM, respectively, $p(i,j)$ is the (i,j)-th entry of the normalized GLCM, μ_i and μ_j are the means of the i-th and j-th rows and columns of the GLCM, and σ_i and σ_j are the standard deviations of the i-th and j-th rows and columns of the GLCM, respectively.

These parameters allow for the quantification of texture based on the spatial relationships of pixel intensities within an image.

II.3 Out-of-Distribution Detection: How it works?

Out-of-distribution detection (OOD detection) is an important task in machine learning, where the goal is to detect and identify outliers, and data samples that do not belong to the in-distribution dataset (IDD) for which the classifier model has been exposed. OOD data is often denoted as "unseen" data, as the model has not encountered it during training. OOD detection is typically accomplished by training a model to distinguish between data within the in-distribution (IDD), which the model observed during training, and OOD data that was hidden in the training phase. This can be done using a variety of techniques such as training a separate detector for OOD data or modifying the model structure or loss function to make it more sensitive to OOD data. Figure II.1 illustrates the overall concept of OOD detection which is an essential element for the safe and reliable application of machine learning algorithms in biomedical imaging in dermatology. The main key advantages of applying OOD detection in the various machine learning applications are:

- **Improved Safety and Reliability:** The implementation of OOD detection helps identify

samples that differ significantly from the training datasets, which improves the performance of machine learning models by avoiding the generation of incorrect predictions on those samples. This is necessary for deploying models in safety-critical applications such as medical imaging.

- **Robustness to Distribution Shift:** When the distribution of test data deviates from the distribution of training data, as is often the case in practical biomedical imaging applications, OOD methods can enhance the model performance. Their predictions become more accurate when OOD samples are ignored.
- **Adoption and Flexibility:** OOD detection techniques can be easily integrated into production situations, as they can be applied to existing models without changing the architecture or training process. This is essential, as retraining or changing medical imaging models can be costly.

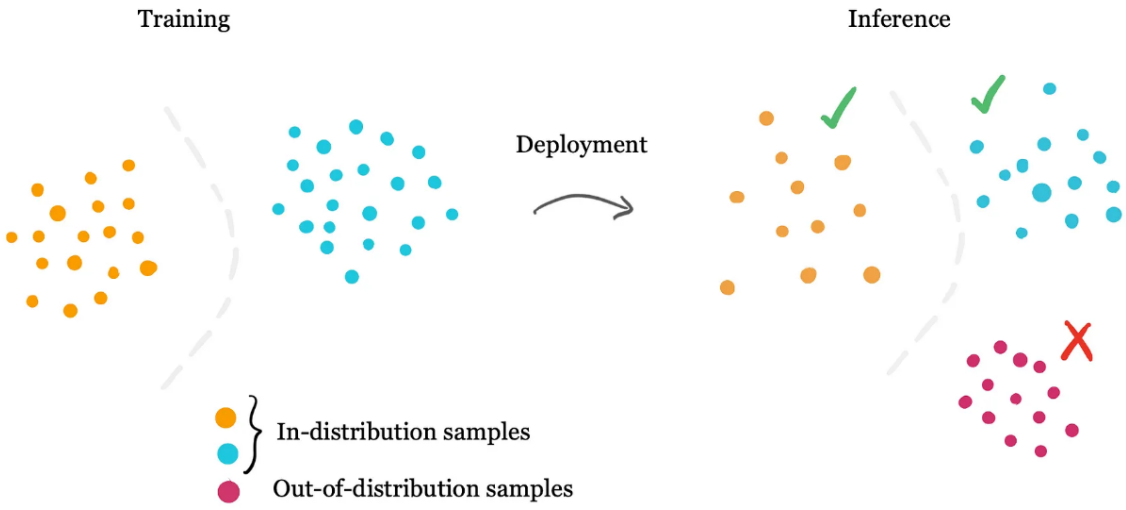


Figure II.1: OOD detection principal [41].

II.3.1 Out-of-Distribution detection methods

Existing OOD detection methods could be grouped into *ensemble methods* such as Isolation Forest (IF) [42], OneClassSVM [43] and *deep learning approaches* [13, 44, 45] based on the type of models employed. Xuan Li et al. [46] used the IF approach on the features computed by a pre-trained CNN to detect OOD images of skin lesions, which is called DeepIF. ODIN [44] and NN Softmax methods [45] utilized CNNs trained for classification to build robust OOD detectors. Hendrycks et al. [45] used temperature scaling in the last layer, and Liang et al. [44] extended this approach, adding small perturbations to the input to separate the softmax score

distributions between in- and out-of-distribution images, allowing for more effective detection. Lastly, as Autoencoders (AE) can model training data distribution, these neural networks are a common option for OOD detection. The majority of the methods discussed in the literature require the training data to consist of in-distribution examples only [47, 48, 49].

We will use different OOD detection methods in our study and systematically compare them to determine the most effective approach for OOD detection across different skin categories.

II.3.1.1 One Class Support-Vector Machines (One-SVM)

A type of SVM or support vector machine that can be used as an OOD detector [43], designed to identify outlier, anomaly, or novelty detection which are samples that do not belong to the in-distribution (ID) dataset. As support vector machines (SVMs) are one of the most robust statistical algorithms for classification problems, due to their broader generalizability of unseen data.

One-Class SVM [43] is an unsupervised learning technique to learn the ability to differentiate the test samples of a particular class from other classes. It works based on the basic idea of minimizing the hypersphere of the single class of examples in training data and considers all the other samples outside the hypersphere to be outliers or out of training data distribution. The One-Class SVM is significantly more effective at modeling the complex shape of the data since it does not make any assumptions about the parametric form of the data distribution.

The mathematical formula to minimize the hypersphere for the One-Class SVM is given by:

$$\min_{w,R,\xi} \quad \frac{1}{2}\|w\|^2 + \frac{1}{vN}\sum_{i=1}^{N}\xi_i \tag{II.11}$$

subject to:

$$w^\top \phi(x_i) \geq R - \xi_i,$$
$$\xi_i \geq 0$$

Where:

- w is the vector of weights that defines the decision boundary (the hypersphere) in the feature space.
- R is the radius of the hypersphere that the One-Class SVM tries to minimize.

- ξ_i are the slack variables that allow some data points to lie outside the hypersphere, effectively treating them as outliers.
- $\phi(x)$ is the mapping function that transforms the input data x_i into a higher-dimensional feature space, making it easier to separate the normal data from outliers.
- v is a hyperparameter that controls the trade-off between minimizing the radius R and the number of training errors (i.e., the number of points lying outside the hypersphere).
- N is the number of training samples.

Figure II.2 illustrates the geometry of the sphere formulation of One-Class SVM.

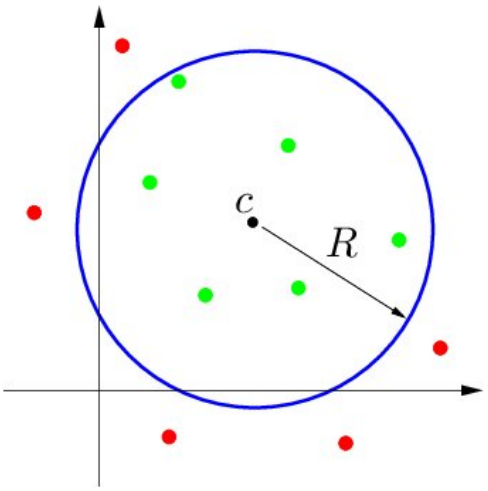


Figure II.2: The geometry of the sphere formulation of one class SVM where R is the radius of the hypersphere and C is a regularization parameter that controls the trade-off between the volume of the hypersphere [50].

We will implement the one SVM by importing the pre-built class from the `libsvm` library of SK-learn named "OneClassSVM". This class implements internally the mathematical model to minimize the hypersphere through training the distribution data samples as one class. the "OneClassSVM" class is implemented using two essential parameters gamma (γ) which controls the width of the Gaussian kernel and nu (ν) which controls the trade-off between the volume of the hypersphere and the number of outliers in the OneClassSVM model.

II.3.1.2 Isolation Forest (IF)

Isolation Forest [42], is a Machine learning approach to identify outliers and detect anomalies using binary trees. Isolation Forest is an ensemble of isolation trees "iTrees" which are binary decision trees that isolate observations by recursive random partitioning represented by the tree structure.

Given the training dataset, random sub-samples are assigned to a binary tree to start the branching by selecting a random feature from the list of features and a random threshold for splitting the node into left and right branches. The data sample is evaluated and assigned to its corresponding branch according to the threshold's value. If the data sample's value is less than the selected threshold values it goes to the left branch and if the data value is higher it goes to the right branch. This process is continued recursively until all data points are isolated or the pre-identified max depth is reached. The IF scoring is based on the aggregation of the depth obtained from each tree known as the anomaly score which is calculated by averaging the path lengths across all the isolation trees in the forest. The anomaly score indicates the outlier sample when it equals -1 and indicates the inlier sample when it equals 1. Figure II.3 illustrates the IF method for detecting anomalies.

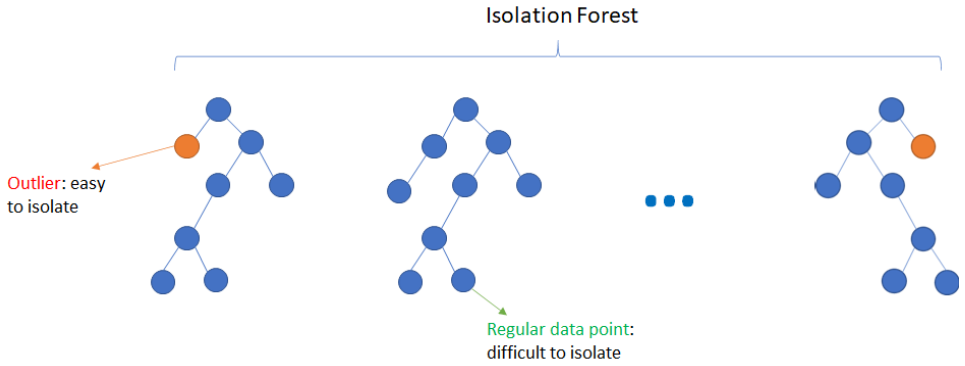


Figure II.3: Isolation Forest for anomaly detection.

The key mathematical model equation for the Isolation Forest is the expected path length formula given by:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{II.12}$$

Where:

$$H(i) = \log(i) + 0.5772156649 \tag{II.13}$$

- c(n) is the expected path length.
- H(i) is the harmonic number.
- n is the number of observations.

We will implement the Isolation Forest (IF) which is available in Scikit-learn [51] by using

the built-in class Isolation Forest imported from `sklearn` ensemble module. The isolation mechanism class is used to discover all the anomalies; this is by choosing a feature at random from all the features of the dataset and a split value between the maximum and minimum values of the chosen feature. The class Isolation Forest is implemented with the following arguments: The parameters are n_estimators which describe the number of base estimators (trees) in the model and max_samples which define the number of samples to draw out of the training set for training the base estimator(s). Moreover, for the decision tree model, the contamination parameter is used to define the percentage of outliers in the respective data set, whereas max_features is used to determine the number of features when finding the best split in the uses of the tree model.

II.3.1.3 AutoEncoder

Autoencoder (AE) is a type of deep learning algorithm designed based on a neural network architecture to compress or encode the input data to its essential features and reconstruct or decode the original input based on its compressed representation [52].

AutoEncoders are trained based on unsupervised machine learning to identify the latent variables of the input data that are hidden or random variables that inform the input data distribution. The AutoEncoder learns to distinguish which latent variables can be used to reconstruct the decoded input data samples. This set of variables constitutes the latent space that represents only the most essential features within the input data to be accurately reconstructed. Hence this deep learning technique is used for Out of Distribution detection because of its ability to extract and reconstruct the main features of the data samples allowing a more accurate OOD detection from the in-distribution features.

The AutoEncoder architecture as shown in Figure II.4 is a subset of the Encoder-decoder architectures that are trained with unsupervised machine learning as they don't rely on labeled datasets and reconstruct their own input data. This architecture is designed to explore the hidden features of unlabelled datasets, rather than to predict known patterns demonstrated in labeled datasets during their training. The Autoencoder's ground truths are the input data samples that are used to measure the reconstructed samples which makes this architecture a self-supervised learning technique. Hence, The Autoencoder is a specific type of Encoder-Decoder architecture characterized by the presence of a bottleneck layer between the encoder and the decoder. This bottleneck captures the latent variables, enabling the encoder to compress the input data

and the decoder to reconstruct the input accurately by focusing on the most relevant features extracted during encoding. This structure helps the Autoencoder learn efficient data representations and improves the reconstruction process. Which makes the architecture consists of three main components:

(a) **Encoder:** The encoder is made up of layers that aim at reducing the dimension of the input layer resulting in a compressed layer representation. Generally, Autoencoder's hidden layers contain fewer nodes than the input layer though it has two sub-layers, an encoder layer and a decoder layer. During passage through these layers, data gets compressed, and this process is often described as data being put through a 'squeeze.' This type of signal data compression involves retaining parts of the input that are beneficial and eliminating information that may be deemed inconsequential.

(b) **Bottleneck:** The bottleneck, or "code," is the layer where the input data is most compressed; as noted earlier, this is the last layer of the stacking process. At the same time, it acts as both the output of the encoder and the input of the decoder. Actually, the main motivation for training an autoencoder and designing it is finding how few salient features are required to capture most of the information contained in the input data. This makes the compressed form important for data reconstruction bearing the name of the latent space or code.

(c) **Decoder:** The decoder is made up of layers that perform the process of decoding by gradually decompressing the information into a larger format. From the bottleneck layer, the decoder gradually adds more nodes to the layers of the model and brings the data out. This output is then measured against the original input to determine the performance of the autoencoder, called the ground truth. The discrepancy between the actual output and the original data is known as the reconstruction error which shows the efficiency with which an autoencoder has encoded the data.

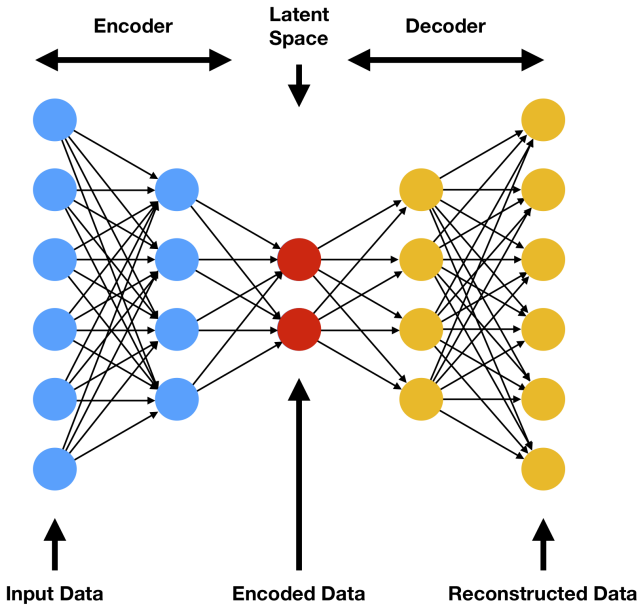


Figure II.4: Autoencoder Architecture[53].

The Autoencoder hyperparameters play a crucial role in the model's performance and efficiency. The code size, which determines the amount of compression, helps balance overfitting and underfitting by regulating the data retained and discarded during compression. The number of layers affects the model's complexity and learning capability, with deeper structures enabling the discovery of complex relations but requiring more processing time. The number of nodes per layer typically decreases from the input layer to the bottleneck layer and increases in the decoder, although this pattern may vary in certain types of autoencoders. Finally, the loss function, which calculates the mean squared error between the input and output during training, is essential for optimizing the model to reconstruct inputs effectively, with its choice depending on the specific problem being addressed.

II.3.1.4 DenseNet

DenseNet [54] is a Densely Connected Convolutional Networks (CNN), designed and characterized by its dense connectivity patterns.

The feature differentiating DenseNet is the dense connection between the layers, where each layer is directly connected to every other layer in a feed-forward manner. This design ensures that problematic issues such as vanishing gradient, weak feature propagation, and feature redundancy are counteracted, while at the same time minimizing the size of the parameters. DenseNet layers are fed with feature maps from all the preceding layers and this is thought to

be filled with information required to enhance learning density and result.

The DenseNet architecture consists of four dense blocks, each followed by a transition layer. The dense blocks contain a series of convolutional layers, where each layer takes input from the feature maps of all preceding layers. This promotes feature reuse and the efficient flow of information through the network. Transition layers between dense blocks perform convolution and pooling operations to progressively reduce spatial dimensions while adjusting the number of feature maps. The final global averaging layer and a Softmax layer produce the output classification. This architecture aims to solve the vanishing gradient problem and improve network parameter efficiency by directly connecting all layers, enabling better feature propagation and reuse. Figure II.5 illustrates the DenseNet121 Architecture which will be used in this study.

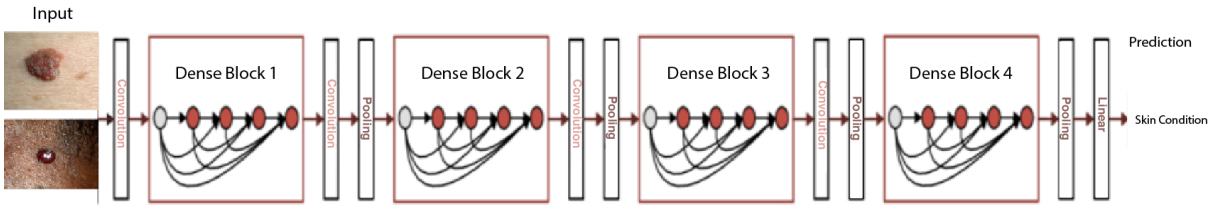


Figure II.5: DenseNet121 architecture.

II.3.1.5 ODIN

ODIN (Out-of-Distribution detector for Neural Networks) is a method for detecting OOD in neural networks. It is designed to improve the reliability of OOD image detection in neural networks by using temperature scaling and adding small perturbations to the input. This method does not require any modification of the pre-trained neural network and is compatible with a variety of network architectures like the DenseNet121 architecture. This OOD detector is built on two main components:

- **Temperature Scaling:** is applied on the neural network $f = (f_1, \ldots, f_N)$ that is trained to classify N classes. For each input x, the neural network assigns a label $\hat{y}(x) = \arg\max_i S_i(x; T)$ by computing the softmax output for each class. As shown in the following equation:

$$S_i(x; T) = \frac{\exp\left(\frac{f_i(x)}{T}\right)}{\sum_{j=1}^{N} \exp\left(\frac{f_j(x)}{T}\right)} \tag{II.14}$$

  Where $T \in R^+$ is the temperature scaling parameter, and it is set to 1 during the training to preserve the standard softmax distribution, ensuring the model focuses on minimizing

loss without altering the output probability distribution. For a given input x, the maximum softmax probability is called the softmax score, i.e., $S_{\hat{y}(x);T} = \max_i S_i(x;T)$ [55].

By using temperature scaling ($S_{\hat{y}(x);T}$ and $S(x;T)$ notations), we can separate the softmax scores between in- and out-of-distribution images, making OOD samples more detectable and their detection more effective.

- **Input Preprocessing:** The input is preprocessed in addition to temperature scaling by adding small perturbations:

$$\tilde{x} = x - \epsilon \operatorname{sgn}(-\nabla_x \log S_{\hat{y}(x);T}), \tag{II.15}$$

where the parameter $\epsilon \ll 1$ is the perturbation magnitude.

The softmax score of any given input can be increased without requiring a class label. This perturbation has a stronger effect on in-distribution images compared to OOD images, as it makes them more separable. These perturbations can be easily computed by back-propagating the gradient of the cross-entropy loss.

The ODIN detector combines the two concepts described above. For each image x. At first the preprocessed image $\tilde{x}$ is calculated. Next, the preprocessed image $\tilde{x}$ is fed into the neural network, calculating its calibrated softmax score $S(\tilde{x};T)$, and comparing the score to the threshold δ. The input image x is then classified as in-distribution if the softmax score is greater than the threshold, and vice versa. Mathematically, the OOD detector can be described as follows:

$$g(x;\delta,T,\epsilon) = \begin{cases} 1 & \text{if } \max_i p(\tilde{x};T) \leq \delta, \\ 0 & \text{if } \max_i p(\tilde{x};T) > \delta. \end{cases} \tag{II.16}$$

The parameters T, ϵ, and δ are chosen such that the true positive rate (i.e., the fraction of in-distribution images correctly classified as in-distribution images) is 95% [55].

II.3.1.6 NN Softmax

NN softmax OOD detector works with the softmax prediction probability as a baseline to compare the in-distribution and the out-of-distribution. It retrieves the maximum predicted class probability from a softmax distribution to identify the OOD sample by computing the maximum softmax probability of the predicted class.

This approach distinguishes between correctly and incorrectly classified examples in the test set and calculates the area under the precision-recall (PR) and receiver operating characteristic (ROC) curves. These areas summarize the performance of a binary classifier using the maximum softmax probabilities as scores for different thresholds. Correctly classified examples are treated as the positive class, labeled "Success", while incorrectly classified examples are treated as the positive class in the "Error" (Err) category, using the negatives of their softmax probabilities as scores. For in-distribution detection ("In"), correctly classified examples in the test set are treated as positive, and their softmax probabilities are used as scores. For OOD ("Out") detection, OOD examples are considered positive and the negative values of their softmax probabilities are used. In addition, the average class prediction probability of misclassified and OOD examples is shown to highlight the potentially misleading confidence of softmax prediction probabilities when considered in isolation [56]. By identifying the threshold, the input image is then classified as in-distribution if the softmax score is greater than the threshold, and vice versa.

II.4 Fairness Analysis

Algorithmic fairness and AI robustness are key concepts to ensure that AI systems do not unfairly discriminate against or favor certain individuals or groups while maintaining sustainable performance in the face of changes in the operating environment or task. To achieve trustworthiness, accountability, fairness, and safety, including the development and AI deployment in an ethical and unbiased manner. There are two main types of fairness [57]:

- **Individual Fairness:** A key concept in AI fairness requires that similar individuals, based on relevant features, should be treated similarly by the AI system. Focusing on ensuring that comparable individuals receive similar outcomes, regardless of their group membership or characteristics. Individual fairness is less restrictive as it does not require the explicit identification of sensitive attributes.
- **Group Fairness:** A key concept in AI fairness that requires fairness at a collective level, which is the idea that individuals who are similar in their features should receive similar modeled predictions. Group fairness is more restrictive compared to individual fairness as it focuses on the collective fairness of different protected groups based on different definitions. The most obvious source of unfairness is unwanted bias, specifically social bias in the

measurement process. Bias in measurement and sampling are considered the most obvious sources of unfairness in machine learning [57].

- **Privileged and Unprivileged Groups:** In the context of AI fairness, privileged and unprivileged groups refer to the characterization of individuals based on their protected characteristics or attributes. A privileged group usually consists of individuals who have historically received advantages or preferential treatment, while the unprivileged group includes only those who have faced discrimination or disadvantage. Observable bias and its direction can be strongly influenced by the characteristics and thresholds used to divide the groups, thus the goal of bias mitigation techniques is to ensure fair treatment and outcomes for both privileged and unprivileged groups.

Several studies propose different ways to analyze skin tones; multiple approaches used individual typology angle (ITA) computed from pixel intensity values [8, 24]. The ITA values were then mapped to Fitzpatrick Skin Types (FST) [58]. This information is key to stratifying further studies regarding the algorithm fairness of classifiers. Rezk et al., [59] proposed data augmentation techniques to improve the diversity of skin tones at the training time of DL models. Moreover, the proximity of skin tones is found to play a significant impact on the classification performance as Groh et al. [25, 24] reported that skin condition classifiers trained on data from only two FST skin tone categories are most accurate on holdout images of the closest FST skin tone categories to the training data. These relationships between the type of training data and holdout accuracy across skin types are consistent with what has been known by dermatologists: skin conditions appear differently across skin types [12].

Although bias and fairness assessment in skin lesion classification has been an active research area [25, 1, 60, 61, 59, 8, 10]. [8] implemented an approach to measure approximate skin tone distributions in public dermatology image datasets using ITA as an estimator, and evaluated the performance of dermatology classification models with respect to the resultant ITA values.

To ensure fairness in our OOD detection methods across diverse skin tone categories, we will incorporate group fairness approaches to address and correct any unwanted biases. Figure II.6 illustrates an example of group fairness evaluation.

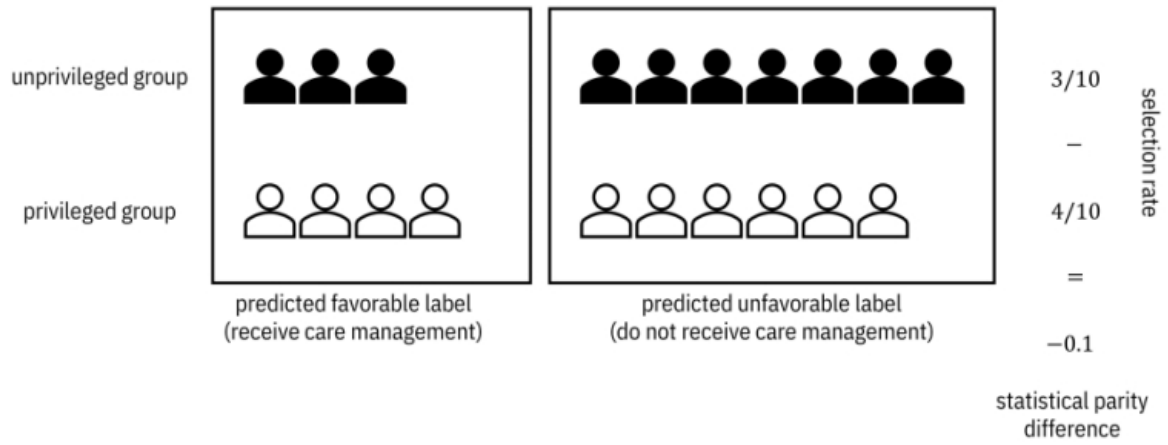


Figure II.6: An example calculation of statistical parity difference for group fairness evaluation [57].

II.5 Proposed Approach

We propose an evaluation framework to assess the fairness of OOD detectors with respect to skin tone categories. Our approach begins with a texture analysis using the Gray Level Co-Occurrence Matrix (GLCM) to identify key differences in the textural features of skin conditions across various skin tones. We categorize both the In-Distribution and Out-of-Distribution datasets into two skin tone groups: FST I-IV (lighter tones) and FST V-VI (darker tones). Subsequently, we will train several OOD detectors on the In-Distribution dataset and test their performance on the Out-of-Distribution dataset. Their effectiveness will be evaluated using standard performance metrics, and fairness will be assessed using group fairness metrics. Figure II.7 illustrates this proposed methodology.

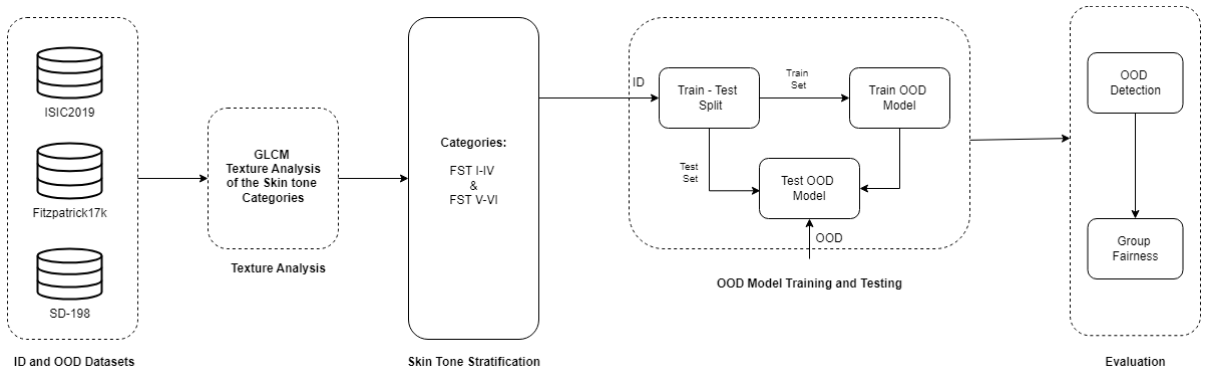


Figure II.7: Proposed Approach.

II.6 Summary

In this chapter, we reviewed the most commonly used OOD detection methods and techniques, detailing the specific methods we will employ in our experiments. Additionally, we discussed statistical parameters for texture analysis and introduced the concept of AI fairness and the need for group fairness evaluation. In the next chapter, we will implement each method, analyze its fairness, and conduct a texture analysis of the dataset used.

Chapter III

Experiments & Results

III.1 Introduction

In this chapter, we investigate the fairness of OOD detectors and compare their performance in identifying anomalies across diverse skin tone representations. We propose an evaluation framework to assess the impact of skin type representation on the performance of OOD detectors using both dermoscopic and clinical datasets. We implement five OOD detection methods, including Isolation Forest and One-Class SVM as baseline models, and three state-of-the-art OOD techniques: AutoEncoder, Neural Network Softmax, and ODIN. These models are trained using in-distribution datasets and tested on OOD datasets to evaluate their performance. To further understand texture differences across skin types, we conduct a texture analysis using the Gray-Level Co-Occurrence Matrix (GLCM). This analysis examines textural features between FST I-IV (lighter skin tones) and FST V-VI (darker skin tones) across various skin conditions. Finally, we conduct a fairness analysis using IBM's AIF360 toolkit to detect any potential biases in our OOD detectors. The goal of this chapter is to explore whether these models show biases towards certain skin tone categories and to propose strategies to address fairness in OOD detection.

III.2 Tools & Libraries

- **Python3:** a popular high-level language acquired for its general-purpose use, easy-to-learn syntax, and flexibility, suitable for use in several contexts including but not limited to data analysis, artificial intelligence, and for creating web pages.
- **Google Colaboratory:** also known as Google Colab is a literate programming web-based cloud service for machine learning in Python and also includes free GPUs and TPUs. Suited best for applications like machine learning and mathematics-based data analysis.
- **TensorFlow:** An open-source machine learning software pioneered by Google that offers tools to architect and train neural networks via computational graphs preferably in deep learning uses.
- **Numpy:** An essential resource for numeric computation in the context of Python that consists of both large, multi-dimensional arrays and matrices, and a set of functions for manipulation of these arrays.
- **Pandas:** An open-source data manipulation and analysis tool for the Python language, providing implemented data structures such as data frames and tools for reading and writing

data between the computer's memory and different formats.

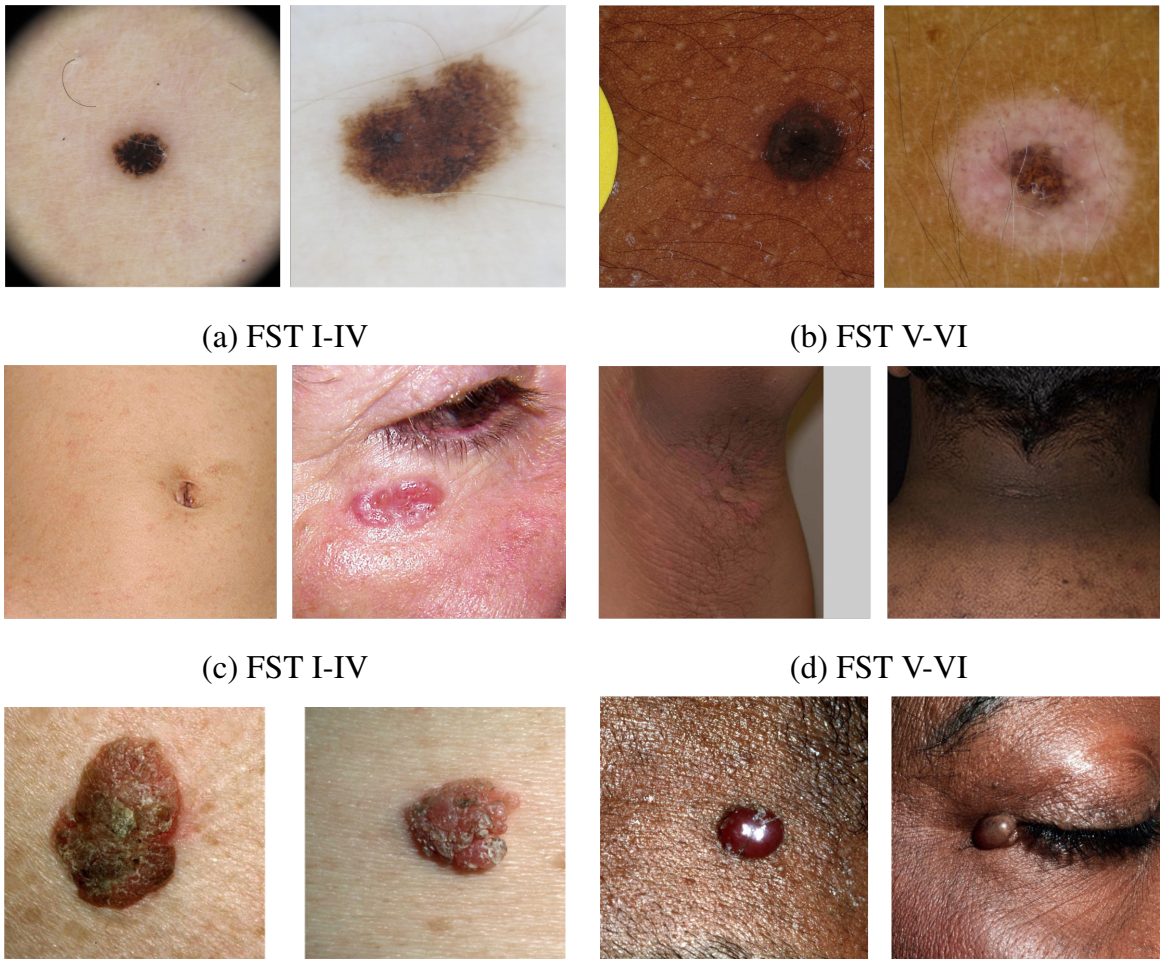
- **Seaborn:** Python library built on top of Matplotlib that offers utilities for creating visually pleasing and clear statistical figures that help with EDA.
- **Matplotlib:** A comprehensive plotting library for Python that produces high-quality static and interactive visualizations. It offers fine-grained control over figures, axes, and plot elements.
- **Scikit-learn:** A simple and efficient machine learning library for Python, providing a wide range of tools for classical machine learning tasks such as classification, regression, clustering, and model selection.
- **Keras:** An open-source deep learning library written in Python that provides a user-friendly API to build and train neural networks. It can run on top of TensorFlow.
- **PyTorch:** An open-source machine learning framework primarily developed by Facebook's AI Research lab. It supports dynamic computational graphs and is popular for its ease of use and flexibility in building and training neural networks.
- **AI Fairness 360 (AIF360):** is an open-source toolkit developed by IBM that provides a comprehensive set of algorithms, metrics, and bias mitigation techniques to detect and mitigate biases in machine learning models and datasets.

III.3 Datasets

We use in our study three different datasets with different clinical and dermoscopic settings. ISIC 2019 [62], SD-198 [63] are used for clinical samples, and Fitzpatrick 17k [64], [25] for dermoscopic samples from different collection protocols. We stratify the samples from both datasets based on skin tones (FST I-IV for light skin tones and FST V-VI for brown and dark skin tones) [64]. Figure III.1 shows reference examples for each skin tone across all the datasets, and Figure III.2 shows the sample count of the different datasets that we used across the two main skin tone categories FST I-IV and FST V-VI according to the Fitzpatrick labeling system.

1. **ISIC2019 Dataset:** consists of 25331 dermoscopic images among eight diagnostic categories [62]. Non-dermatologists trained on previous examples labeled skin images as FST I-IV and FST V-VI. We manually annotated the skin tone for this dataset, and after carefully curating the labels, we have 25327 samples categorized as FST I-IV and 4 as FST V-VI.

2. **FitzPatrick17k Dataset:** contains 16577 clinical images with skin type labels based on the Fitzpatrick scoring system. The images are sourced from two online open-source dermatology atlases and are annotated with Fitzpatrick skin type labels by a team of human annotators from Scale AI. The Fitzpatrick labeling system is a six-point scale originally developed for classifying sun reactivity of skin and adjusting clinical medicine according to skin phenotype. We grouped the six labels provided in the dataset into two classes, 13844 as FST I-IV and 2168 as FST V-VI.
3. **SD-198 Dataset:** consists of 6473 clinical images among diagnostic categories. Non-dermatologists trained on previous examples labeled skin images as FST I-IV and FST V-VI. We manually annotated the skin tone for this dataset, as this information was missing, the labels are available at the repository. After carefully curating the labels, we have 6214 samples categorized as FST I-IV and 210 as FST V-VI.

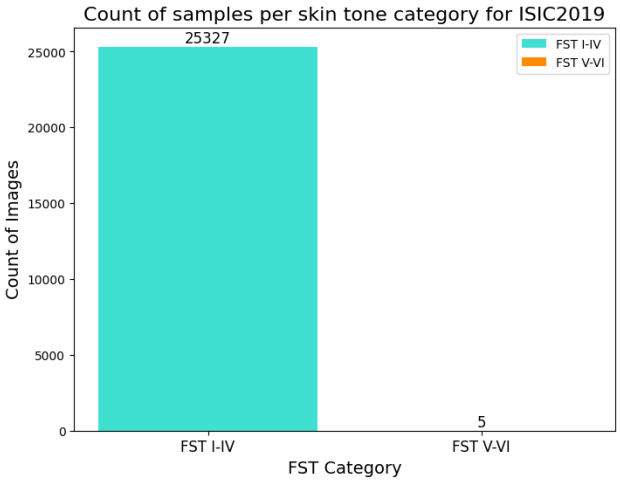


(a) FST I-IV (b) FST V-VI

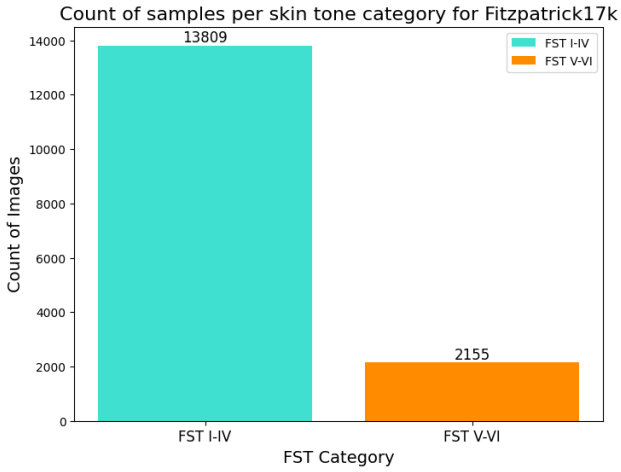
(c) FST I-IV (d) FST V-VI

(e) FST I-IV (f) FST V-VI

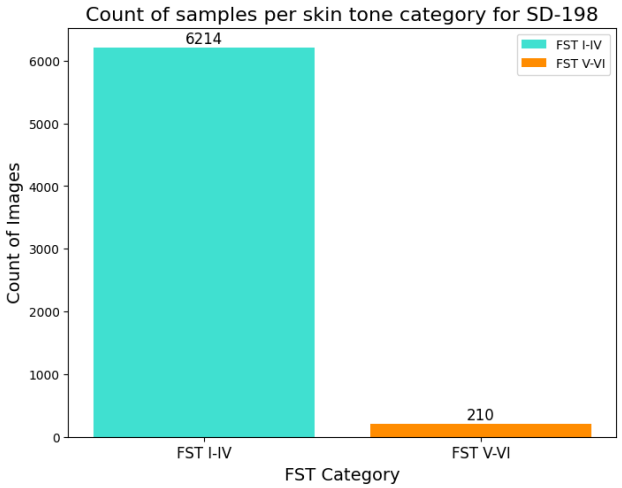
Figure III.1: Examples for FST I-IV and FST V-VI skin types in all datasets. Samples in (a) and (b) belong to the ISIC 2019 dataset, (c) and (d) to the Fitz17k dataset, (e) and (f) belong to the SD-198 dataset.



(a) ISIC2019 Dataset



(b) Fitzpatrick17k Dataset



(c) SD-198 Dataset

Figure III.2: Datasets samples count across skin tone categories.

III.4 Performance & Fairness Evaluation

Our evaluation of OOD detectors considers both their performance in OOD detection and their fairness across various skin tone categories.

III.4.1 Performance metrics

To effectively evaluate the performance of our OOD detection methods, we apply a set of standard evaluation metrics that are widely used in OOD detection. In addition to the average evaluation metrics we introduce a novel metric to measure OOD detection performance across different skin tone categories:

- **Mean Squared Error (MSE) loss:** is a common loss function used for machine learning and statistical modeling. It measures the average squared difference between the predicted values and the actual target values. We use MSE in OOD detection for setting thresholds and assigning OOD labels based on reconstruction errors. The formula for MSE is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{III.1}$$

  Where n is the number of data points, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

- **Area Under the Receiver Operating Characteristic Curve ($AUROC$):** is a performance metric used for evaluating the ability of the OOD detection model to distinguish between positive (in-distribution) and negative (out-of-distribution). The AUROC measures the entire two-dimensional area underneath the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. A model with a higher AUROC value indicates better performance in distinguishing between the classes, with a value of 1 representing perfect discrimination and a value of 0.5 indicating no discriminative power, equivalent to random guessing [65]. Using numerical integration (specifically, the trapezoidal rule), the AUROC can be approximated as shown in Equation III.2:

$$\text{AUROC} \approx \sum_{i=1}^{N}\frac{x_{i+1} - x_i}{2} \times (y_i + y_{i+1}) \tag{III.2}$$

  where x_i and y_i are the values of the false positive rate (FPR) and true positive rate (TPR) at different threshold levels.

- **F1 score (F_1):** a performance metric used to identify the overall performance of an anomaly detection method, by combining the precision and recall using the harmonic mean [66]. Its formula is shown in Equation III.3:

$$F_1 = \frac{2 \times (P \times R)}{(P + R)} \tag{III.3}$$

Where:

$$R = \frac{TP}{TP + FN} \quad \text{(Recall)} \tag{III.4}$$

$$P = \frac{TP}{TP + FP} \quad \text{(Precision)} \tag{III.5}$$

Here:

TP True Positives

FN False Negatives

FP False Positives

- **Representation Gap (RG):** is a metric used to measure the difference in the performance of the OOD detector under different skin types compared to overall performance as shown in Equation III.6.

$$\text{RG} = |\text{F1}_{\text{FST I-IV}} - \text{F1}_{\text{FST V-VI}}| \tag{III.6}$$

Where:

– $\text{F1}_{\text{FST I-IV}}$ denotes the F1 score for light skin tones.

– $\text{F1}_{\text{FST V-VI}}$ denotes the F1 score for dark skin tones.

By taking the absolute difference, the RG metric ensures that the result is always non-negative, reflecting the magnitude of the performance disparity without concern for direction. A smaller RG indicates more equitable performance across different skin tones, while a larger RG suggests a disparity in OOD detection performance between these categories.

III.4.2 Group Fairness Metrics

To measure the fairness of our OOD detectors across the different skin categories, we employ known group fairness metrics using the AIF360 toolkit [67]:

1. **Statistical Parity Difference (SPD):** a group fairness metric that measures the difference in the proportion of positive outcomes between the privileged and unprivileged groups [68]. It evaluates how equally the model's predictions are distributed across different groups. The value (SPD = 0) means an equal rate, a negative value means the unprivileged group is at a

disadvantage, and a positive value means the privileged group is at a disadvantage.

2. **Disparate Impact Ratio (DI):** a group fairness metric evaluates the ratio of the positive prediction rates between the privileged and unprivileged groups [69]. It aims to ensure that the model's predictions do not have a disproportionately negative impact on the disadvantaged group. A value of DI close to 1 indicates fairness, where the rate of receiving the favorable outcome is the same for both the unprivileged and privileged groups. A value less than 1 indicates unfairness, where the unprivileged group is at a disadvantage, and a value greater than 1 indicates unfairness, where the privileged group is at a disadvantage.

III.5 Experiments & Results

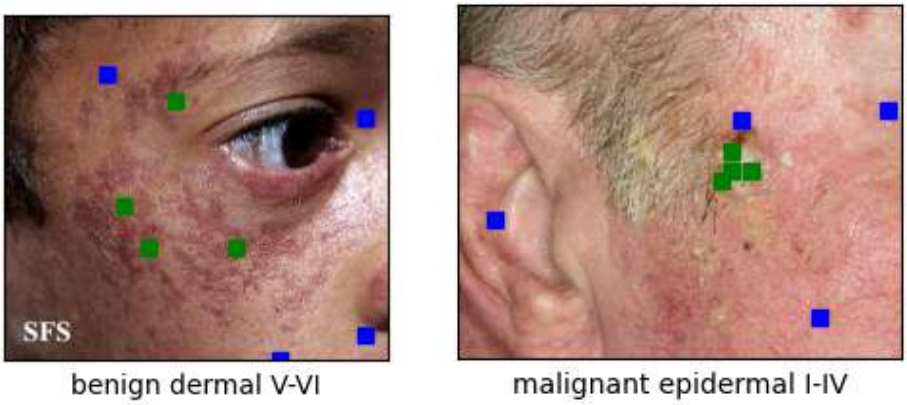
III.5.1 Experiment I: Texture analysis using GLCM-Grey Level Co-Occurrence Matrix

The purpose of this experiment is to study the texture analysis of the skin tone representation across different skin conditions. We aim to use the GLCM (Grey level co-occurrence matrix) on the FitzPatrick17k dataset which contains clinical images from different collection protocols and across various skin types. Our goal is to extract the main features that differentiate the skin texture based on the skin tones across each skin condition from the Fitzpatick17k dataset that contains more FST V-VI (darker) samples. This approach allows us to understand the main features of each skin category when analyzing the different skin conditions. In this analysis, we use the most common GLCM statistical parameters to identify the main features of each skin tone and spot the main differences of each category (FST V-VI and FST I-IV). We start our analysis by scaling the given images of multiple skin conditions into gray level and identifying different patches of each area of the image. There are two main area types in the image:

(a) Scratch (sky) area: corresponds to the healthy or the normal part of the skin.

(b) Cell area: corresponds to the unhealthy or the affected part of the skin.

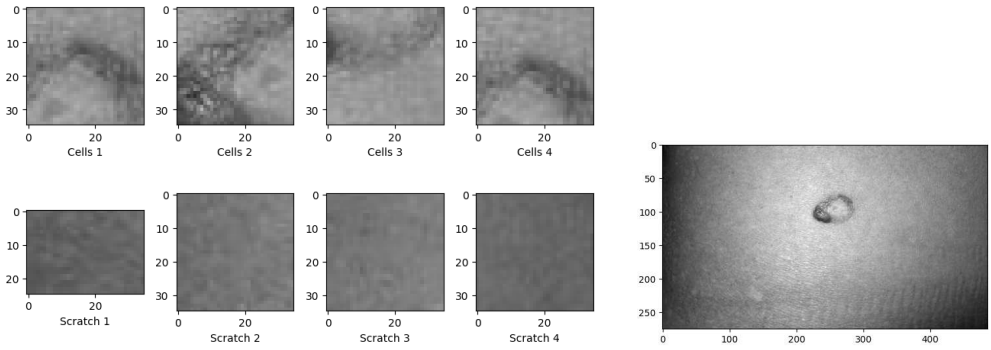
We take four patches from each area to extract the main features of the image by calculating the statistical parameters across each skin condition stratified by skin tones and comparing the results as illustrated in Figure III.3.



(a) (b)

Figure III.3: Cell and scratch area patches representation across different skin conditions and skin tones.

Each patch from both the cell and scratch areas represents a specific feature that corresponds to its corresponding area. After defining the grey co-matrix and turning the image to the grey level we extract the patches as a list of cell and scratch patches and calculate the needed statistical parameters as shown in Figure III.4.



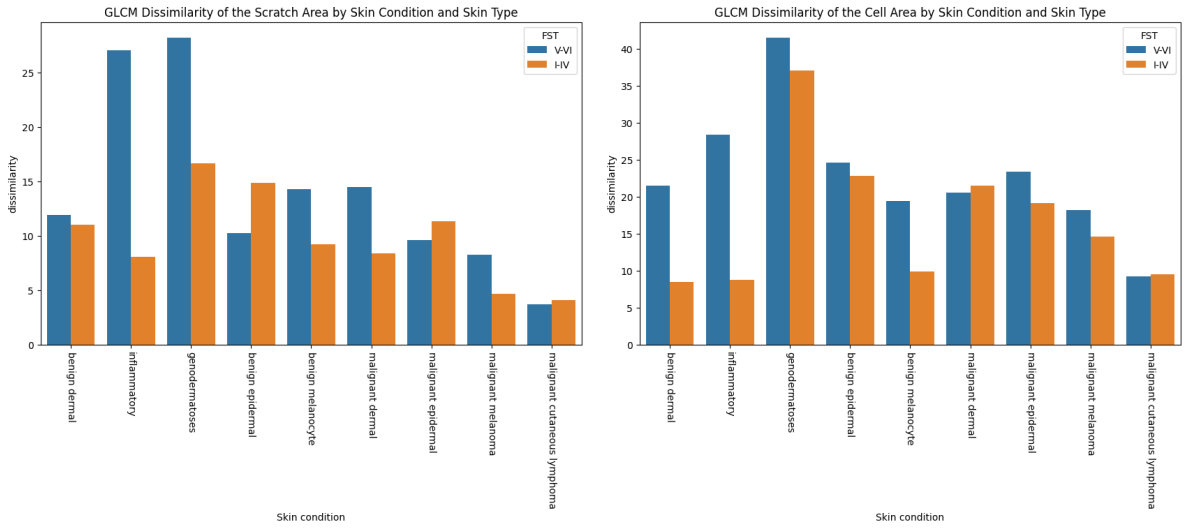
(a) (b)

Figure III.4: Grey level co-occurrence matrix features extraction from patches.

The GLCM parameters are extracted for each area type of the captured images for analysis. We plot the statistical features of each skin tone across the skin conditions available in the dataset.

The following plots show the variations of textural statistical parameters across 9 skin conditions stratified by two skin tone categories FST V-VI and FST I-IV.

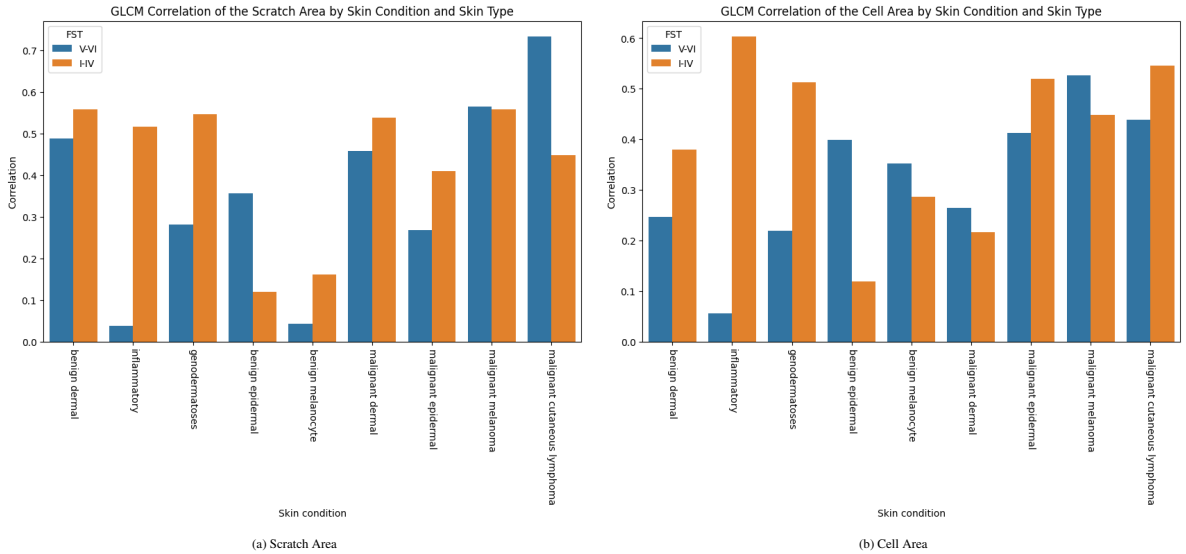
1. **Dissimilarity:** The bar plot in Figure III.5 indicates that the dissimilarity is greater for the FST V-VI skin category across most skin conditions, in both cell and scratch areas.



(a) Scratch Area (b) Cell Area

Figure III.5: Dissimilarity results of FST I-IV and FST V-VI skin types across nine skin conditions for scratch and cell areas

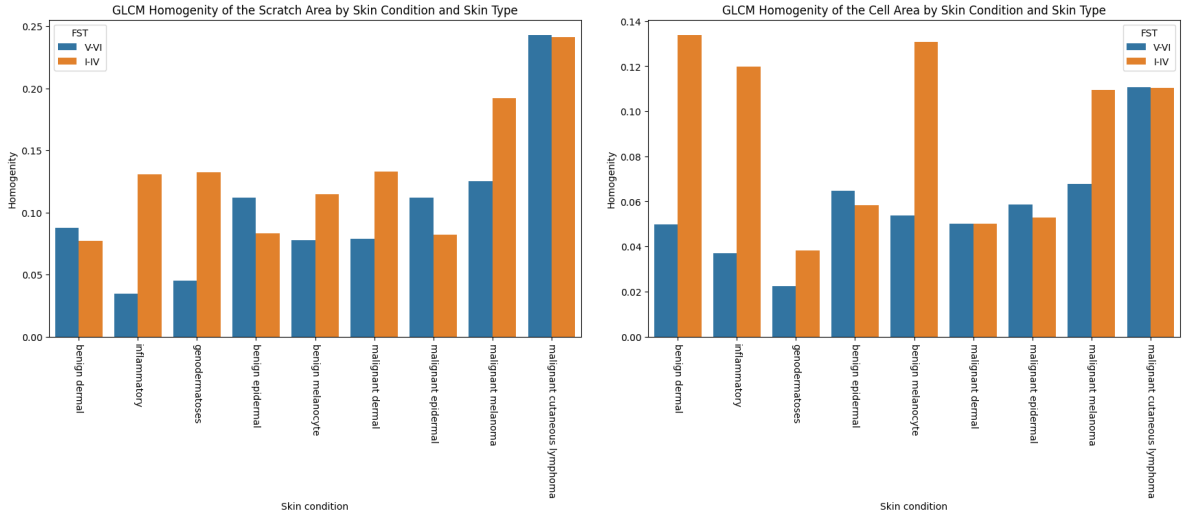
2. **Correlation:** The bar plot in the Figure III.6 shows that the correlation rate is higher for the FST I-IV skin category across most skin conditions in both scratch and cell areas.



(a) Scratch Area (b) Cell Area

Figure III.6: Correlation results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas.

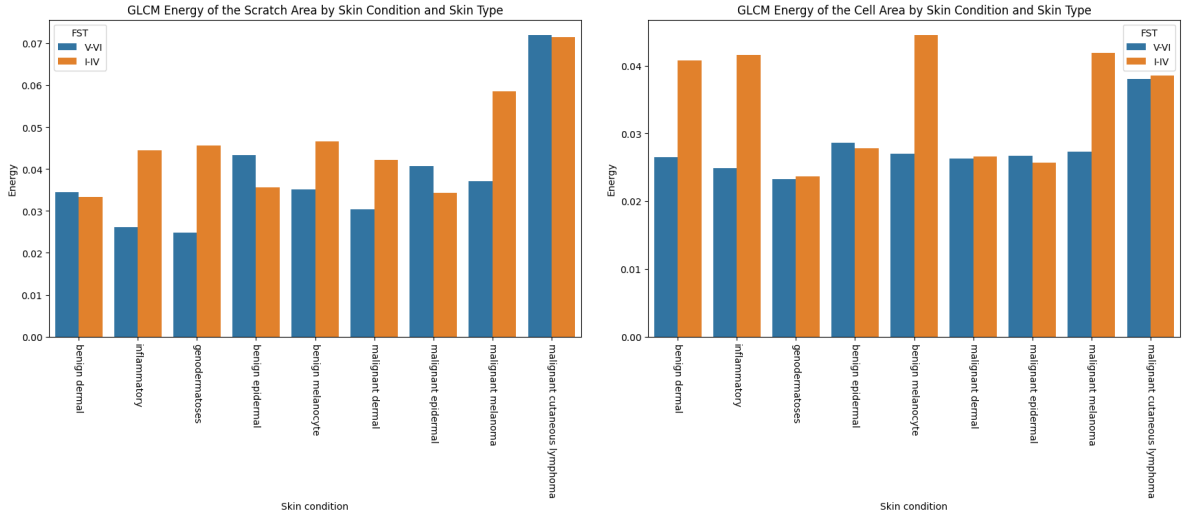
3. **Homogenity:** According to the FigureIII.7, the homogeneity parameter is generally higher in the FST I-IV skin category.



(a) Scratch Area (b) Cell Area

Figure III.7: Homogenity results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

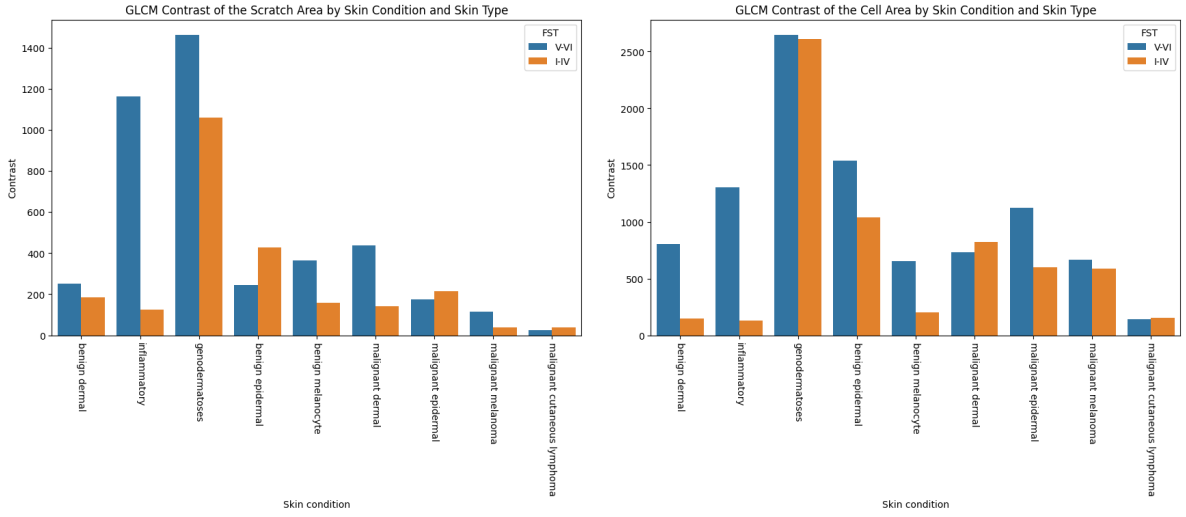
4. **Energy:** According to Figure III.8, the energy parameter appears higher for the FST I-IV skin tone category across most skin conditions in both scratch and cell areas.



(a) Scratch Area (b) Cell Area

Figure III.8: Energy results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

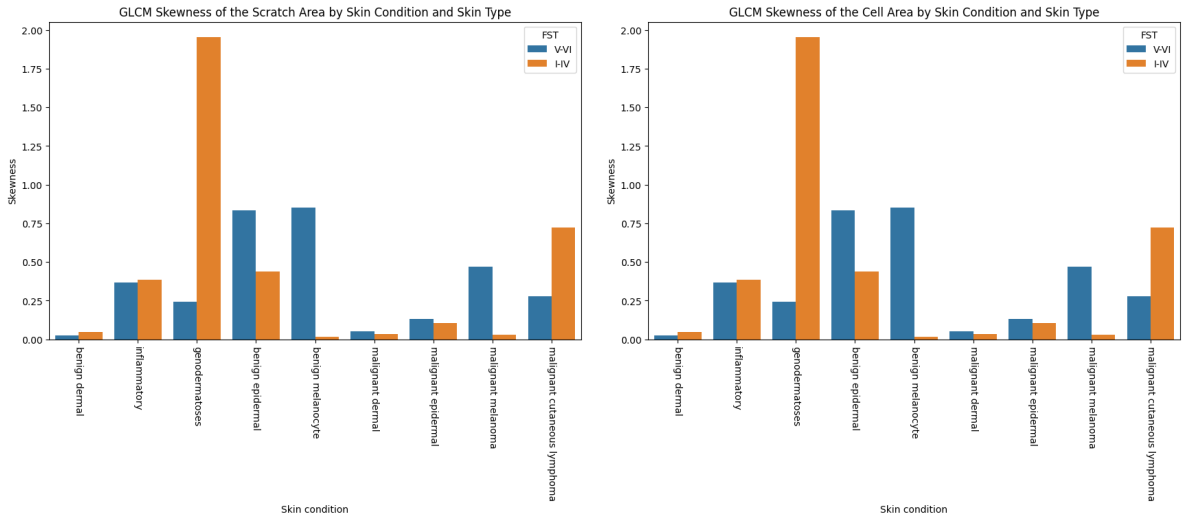
5. **Contrast:** From the contrast results shown in Figure III.9, the contrast appears to be higher for the FST V-VI skin tone category across most skin conditions in both cell and scratch areas of the images.



(a) Scratch Area (b) Cell Area

Figure III.9: Contrast results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

6. **Skewness:** From Figure III.10, we observe a range of positive and negative skewness values, indicating varying degrees of asymmetry in the cell area texture distributions. The FST I-IV group shows higher positive skewness values, whereas the FST V-VI group exhibits lower skewness values, indicating a more symmetric or left-skewed distribution of the cell area textures.

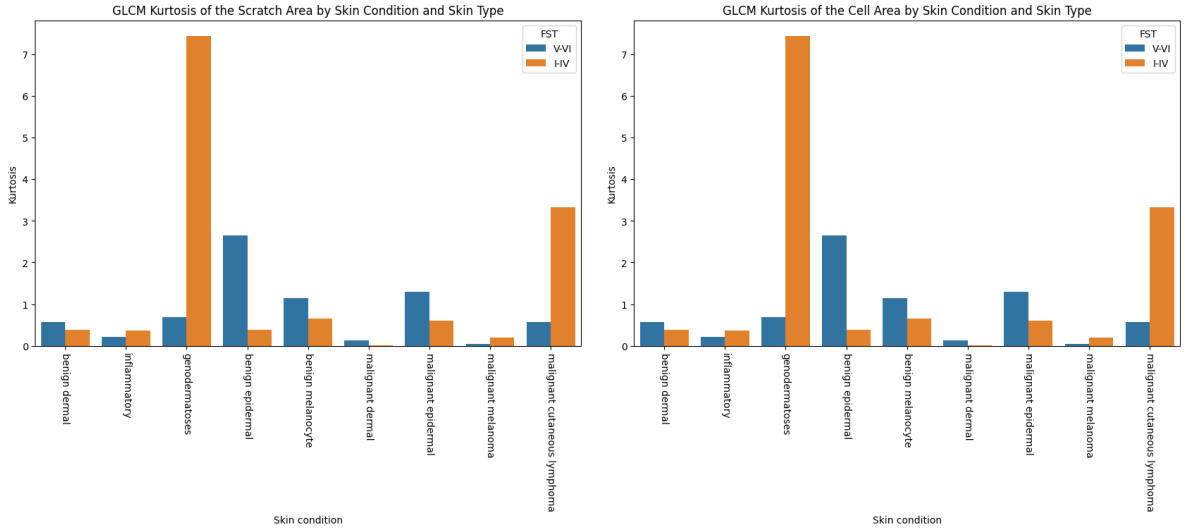


(a) Scratch Area (b) Cell Area

Figure III.10: Skewness results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

7. **Kurtosis:** As shown in Figure III.11, the FST I-IV skin category exhibits higher kurtosis, indicating that their cell area texture distributions are more peaked and have heavier tails. In

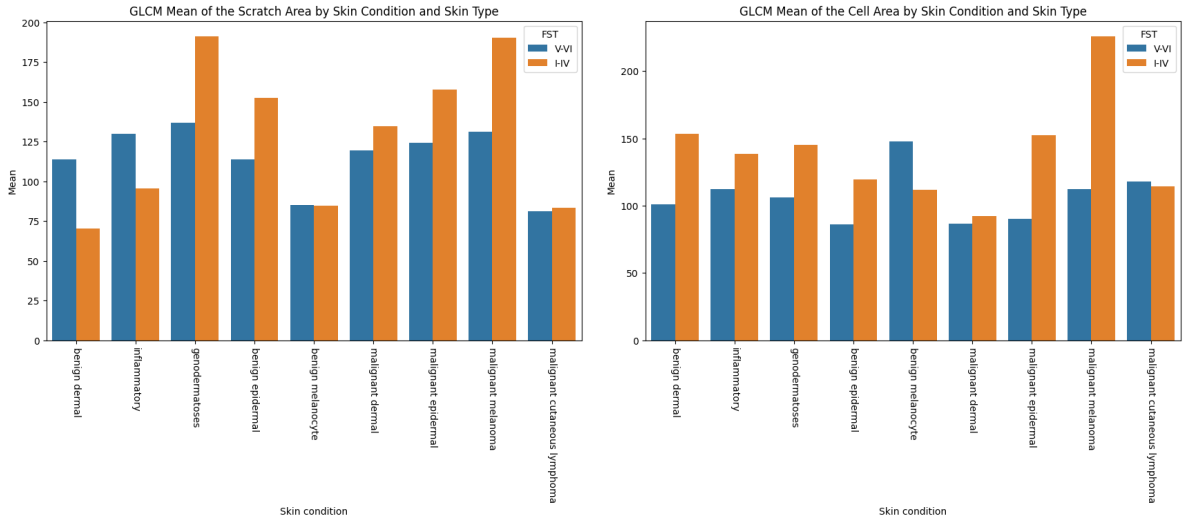
contrast, the FST V-VI skin category tends to have lower kurtosis, suggesting their texture distributions have lighter tails compared to the light skin tone.



(a) Scratch Area (b) Cell Area

Figure III.11: Kurtosis results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

8. **Mean:** From the results shown in Figure III.12, the mean values appear to be higher for the FST I-IV skin category compared to the FST V-VI skin category.

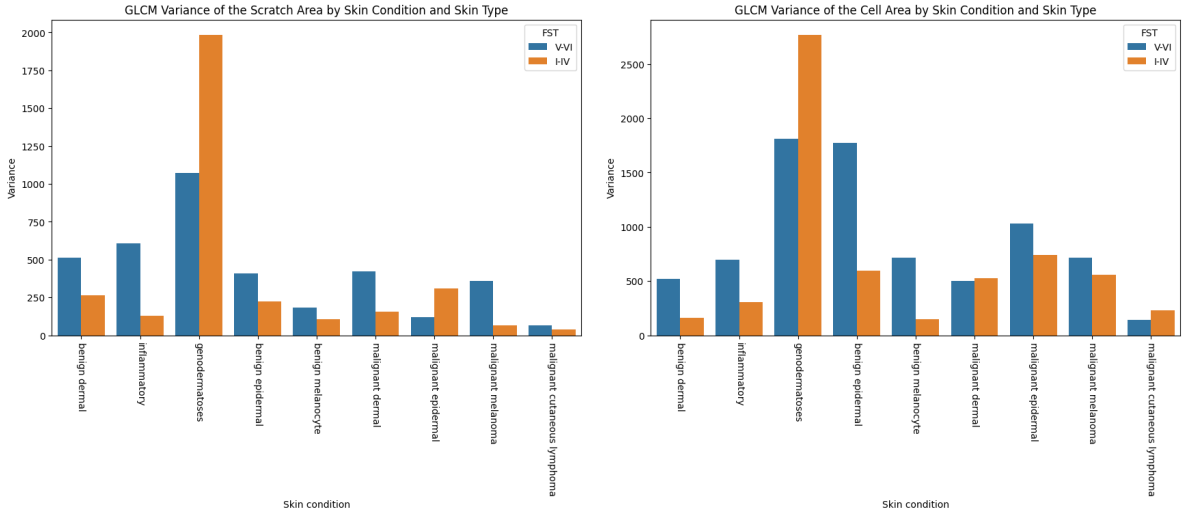


(a) Scratch Area (b) Cell Area

Figure III.12: Mean results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

9. **Variance:** From Figure III.13, we observe that for most skin conditions, variance remains higher for the FST V-VI skin category compared to the FST I-IV skin category in both cell and scratch areas. This indicates that the texture properties are more variable in the darker

skin tone compared to the lighter skin tone.



(a) Scratch Area

(b) Cell Area

Figure III.13: Variance results of FST I-IV and FST V-VI skin types across all skin conditions for scratch and cell areas

III.5.2 Summary table of GLCM texture analysis results

Table III.1 summarises all the statistical features of the texture analysis for some Fitzpatrick17k dataset samples across the different skin diseases available in the dataset stratified by the skin tones categories FST I-IV and FST V-VI.

Table III.1: FitzPatrick17K Dataset skin conditions samples textural features stratified by skin tones FST I-IV and FST V-VI, All the samples belong to the Fitz17k.

Skin condition	FST	Dissimilarity	Correlation	Homogenity	Energy	Contrast	Skewness	Kurtosis	Mean	Variance
Benign Dermal	V-VI	11.9307	0.4878	0.0877	0.0344	250.4408	0.0244	0.5610	113.7421	510.9848
Benign Dermal	I-IV	10.9879	0.5579	0.0771	0.0334	184.4512	0.0478	0.3830	70.3931	263.7079
Inflammatory	V-VI	27.0156	0.0375	0.0343	0.0262	1161.2953	0.3658	0.2184	129.8025	609.0989
Inflammatory	I-IV	8.0738	0.5160	0.1304	0.0445	126.3005	0.3866	0.3689	95.3314	129.4696
Genodermatoses	V-VI	28.2279	0.2807	0.0450	0.0248	1463.4360	0.2448	0.6940	136.9131	1070.0848
Genodermatoses	I-IV	16.6462	0.5461	0.1324	0.0456	1060.0576	1.9563	7.4469	191.2769	1985.7399
Benign Epidermal	V-VI	10.2676	0.3566	0.1118	0.0434	244.3543	0.8331	2.6566	113.6373	407.6705
Benign Epidermal	I-IV	14.8898	0.1188	0.0834	0.0357	426.9783	0.4385	0.3851	152.3800	225.7625
Benign Melanocyte	V-VI	14.2739	0.0428	0.0775	0.0351	365.1156	0.8510	1.1428	84.9363	183.1117
Benign Melanocyte	I-IV	9.2006	0.1607	0.1148	0.0466	158.3584	0.0165	0.6467	84.7889	106.7244
Malignant Dermal	V-VI	14.4810	0.4578	0.0786	0.0304	437.5686	0.0532	0.1248	119.4553	420.2027
Malignant Dermal	I-IV	8.3514	0.5389	0.1331	0.0421	143.1767	0.0324	0.0079	134.7110	156.4214
Malignant Epidermal	V-VI	9.5993	0.2688	0.1119	0.0408	174.6802	0.1332	1.3014	124.2565	121.6026
Malignant Epidermal	I-IV	11.2999	0.4107	0.0821	0.0343	216.2976	0.1042	0.6001	157.4645	308.5526
Malignant Melanoma	V-VI	8.2246	0.5650	0.1251	0.0371	115.3283	0.4692	0.0463	130.9070	360.3003
Malignant Melanoma	I-IV	4.6807	0.5579	0.1923	0.0585	38.8674	0.0309	0.1911	190.4288	65.4713
Malignant Cutaneous-lymphoma	V-VI	3.6662	0.7344	0.2431	0.0721	26.1429	0.2792	0.5640	81.0149	64.5359
Malignant Cutaneous-lymphoma	I-IV	4.0710	0.4488	0.2412	0.0715	38.7643	0.7212	3.3311	83.2720	40.1855

III.5.2.1 Discussion

Analyzing the GLCM (Gray-Level Co-occurrence Matrix) features across the different skin conditions and skin tone categories (FST I-IV and FST V-VI) provides valuable insights into the underlying textural characteristics of the cell and scratch areas. The key findings from the GLCM analysis are as follows: Correlation, Homogeneity, Energy, Kurtosis, Skewness, and Mean were generally higher for the lighter skin tone category (FST I-IV) compared to the darker skin tone category (FST V-VI) across most of the skin conditions, suggesting that the cell and scratch area textures in the lighter skin samples exhibit greater correlation, more uniform gray-level distributions, higher energy, increased kurtosis and skewness, and higher overall mean in the GLCM values, indicating more consistent and less varied textural properties compared to the darker skin samples. On the other hand, the GLCM features of Dissimilarity, Contrast, and Variance were higher for the darker skin tone category (FST V-VI) compared to the lighter skin tone category (FST I-IV) across most of the skin conditions, implying that the cell and scratch area textures in the darker skin samples exhibit greater dissimilarity between neighboring pixel pairs, higher contrast, and a more variance of GLCM values, suggesting more heterogeneous and varied textural characteristics. The observed differences in these GLCM-based textural features between the light and dark skin tone categories across the various skin conditions indicate that the underlying cell and scratch area textures may have distinct properties that differentiate them. Further exploration of these GLCM feature patterns could lead to a deeper understanding of the relationship between skin tone, skin condition, and the underlying cellular and structural properties of the skin. The texture analysis findings highlight the critical importance of considering skin tone representation in the classification of skin condition samples to ensure fairness across diverse skin tone categories.

III.5.3 Experiment II: OOD Detection using ISIC2019 as ID and Fitzpatrick17k as OOD

The purpose of this experiment is to evaluate different OOD detection methods and compare their performance across the two main skin categories: FST V-VI for darker skin tones and FST I-IV for lighter skin tones. We use the ISIC2019 dataset as the in-distribution (ID) and the Fitzpatrick17k dataset as the OOD dataset. We adopt Isolation Forest [42], One-Class SVM [43] as baselines, and Autoencoder [49], Neural Network Softmax [45], and ODIN [44]

as state-of-the-art OOD methods.

III.5.3.1 One SVM & Isolation Forest

1. **Loading and pre-processing the datasets:** Each image from the datasets is opened, converted to RGB format, and resized to 32x32 pixels The resized images are then flattened into 1D arrays.
2. **Data labeling :** The data labeling is done by categorizing our data samples based on whether they are ID or OOD. Additionally, we stratify the OOD samples by skin tone categories, specifically FST V-VI and FST I-IV.
3. **PCA analysis:** We use Principal Component Analysis (PCA) visualization to examine the correlation between our inliers and OOD data samples based on their skin type categories. This involves plotting the data points and identifying relationships between them to better understand the distribution and clustering of samples with respect to their skin types as shown in Figure III.14.

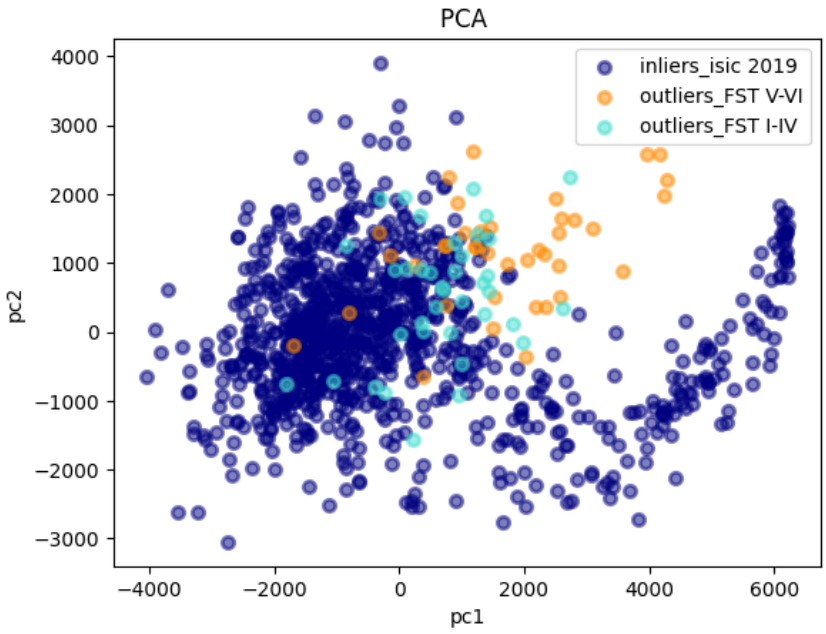


Figure III.14: PCA Visualization taking ISIC2019 as IDD and Fitzpatrick17k as OOD dataset.

4. **Grid search and models training:** We perform a grid search analysis for our One-Class SVM (OneSVM) and Isolation Forest (IF) models to determine the optimal parameters for training. As a result of the grid search, the OneSVM is configured with $\nu = 0.01$ and $\gamma = 0.0001$, and the IF is configured with 300 estimators and a contamination rate of 0.1. Additionally, we split our dataset into 60% for training and 40% for testing, ensuring the

split is stratified by the target variable to maintain the distribution of classes. The models are then trained on the training set using the optimized parameters.

5. **Models evaluation:** Both One SVM and IF models are evaluated using two primary performance metrics for OOD detection: the $F1 - score$ and $AUROC$. Detailed results of the model's performance can be found in Table III.2.

   The following histograms shown in Figure III.15 showcase the Abnormal scores distributions for the IF as an OOD method, stratified by skin tone for the IF model. We can observe that the IF assigned higher abnormal scores (≥ 0.5) to $\approx 16\%$ of true outliers from FST I-IV skin type and $\approx 36\%$ of outliers in FST V-VI.

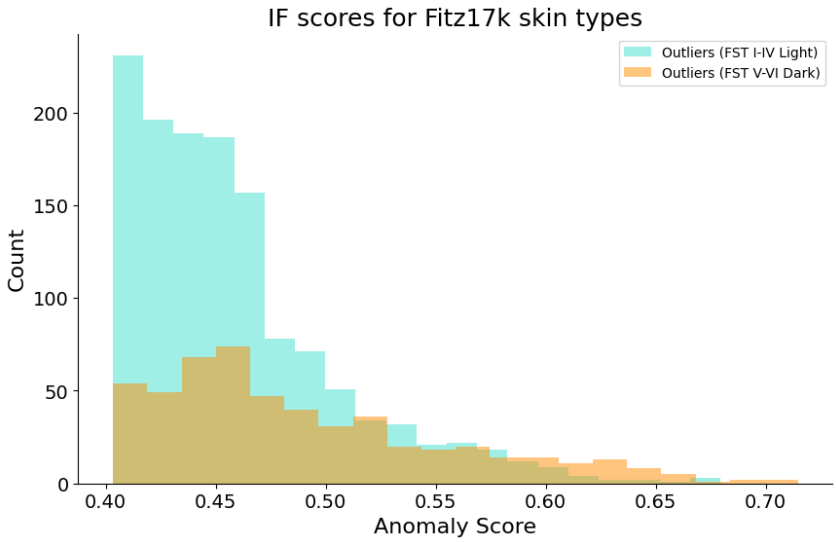
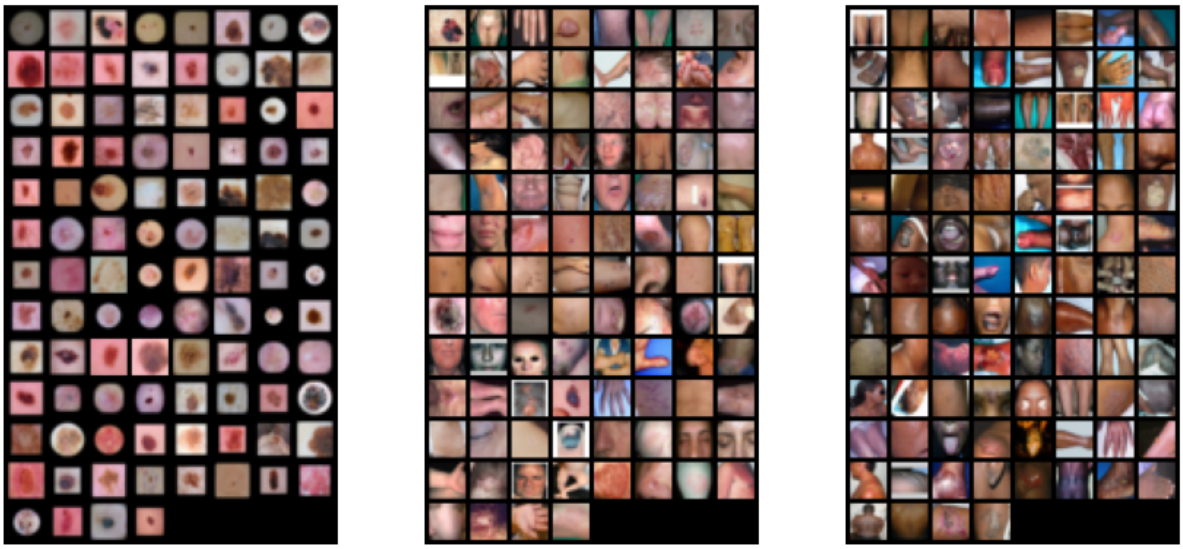


Figure III.15: Abnormal scores distributions for the Isolation Forest.

III.5.3.2 Autoencoder

We train the Autoencoder on the ISIC2019 dataset as IDD and evaluate its performance on the Fitzpatrick17k dataset as OOD.

1. **Loading and preprocessing the datasets:** Both the train and test datasets are preprocessed and loaded via Torch data loader. Each image is loaded after being resized and transformed into tensors. We split the train images into train and validation data loaders with 80% for the train and 20% for the validation while the test data loaders are stratified by skin categories, to get the FST I-IV test data loader, FST V-VI data loader, and all the test images data loader.



(a) Train dataloader (b) Test data loader FST I-IV (c) Test data loader FST V-VI

Figure III.16: Image dataloaders.

2. **Creating the Autoencoder architecture:** The Autoencoder architecture is designed to compress and reconstruct 32x32 pixel images with three color channels, as illustrated in Figure II.4. The encoder consists of three convolutional layers, each followed by a ReLU activation function. The first convolutional layer reduces the input image size from 3 x 32 x 32 to 12 x 16 x 16. The second layer further reduces it to 24 x 8 x 8, and the third layer compresses it to 48 x 4 x 4. The decoder mirrors this structure with three transposed convolutional layers, also followed by ReLU activations, which progressively upsample the encoded representation back to the original image size. The first transposed convolutional layer increases the size from 48 x 4 x 4 to 24 x 8 x 8, the second to 12 x 16 x 16, and the final layer reconstructs the output to the original size of 3 x 32 x 32 using a sigmoid activation function to ensure the output values are between 0 and 1. This architecture includes 47,355 trainable parameters, allowing the network to effectively compress and decompress input images while minimizing reconstruction error. The Autoencoder is trained on the in-distribution dataset.

```
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [-1, 12, 16, 16]             588
              ReLU-2          [-1, 12, 16, 16]               0
            Conv2d-3            [-1, 24, 8, 8]           4,632
              ReLU-4            [-1, 24, 8, 8]               0
            Conv2d-5            [-1, 48, 4, 4]          18,480
              ReLU-6            [-1, 48, 4, 4]               0
   ConvTranspose2d-7            [-1, 24, 8, 8]          18,456
              ReLU-8            [-1, 24, 8, 8]               0
   ConvTranspose2d-9          [-1, 12, 16, 16]           4,620
             ReLU-10          [-1, 12, 16, 16]               0
  ConvTranspose2d-11           [-1, 3, 32, 32]             579
          Sigmoid-12           [-1, 3, 32, 32]               0
================================================================
Total params: 47,355
Trainable params: 47,355
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 0.01
Forward/backward pass size (MB): 0.20
Params size (MB): 0.18
Estimated Total Size (MB): 0.39
----------------------------------------------------------------
```

Figure III.17: Simple Autoencoder Architecture

3. **Training the Autoencoder:** We trained the Autoencoder on the ISIC2019 dataset as the in-distribution data, applying early stopping with the patience of 5 epochs to obtain the best-performing model and prevent overfitting. The training process stopped after 22 epochs, with the training and validation losses depicted in Figure III.18. Our trained Autoencoder successfully extracts key features from the input images and reconstructs them by passing the encoded features through the decoder. Figure III.19 displays the original images alongside their reconstructed counterparts, demonstrating the Autoencoder's ability to capture and replicate essential characteristics of the input data.

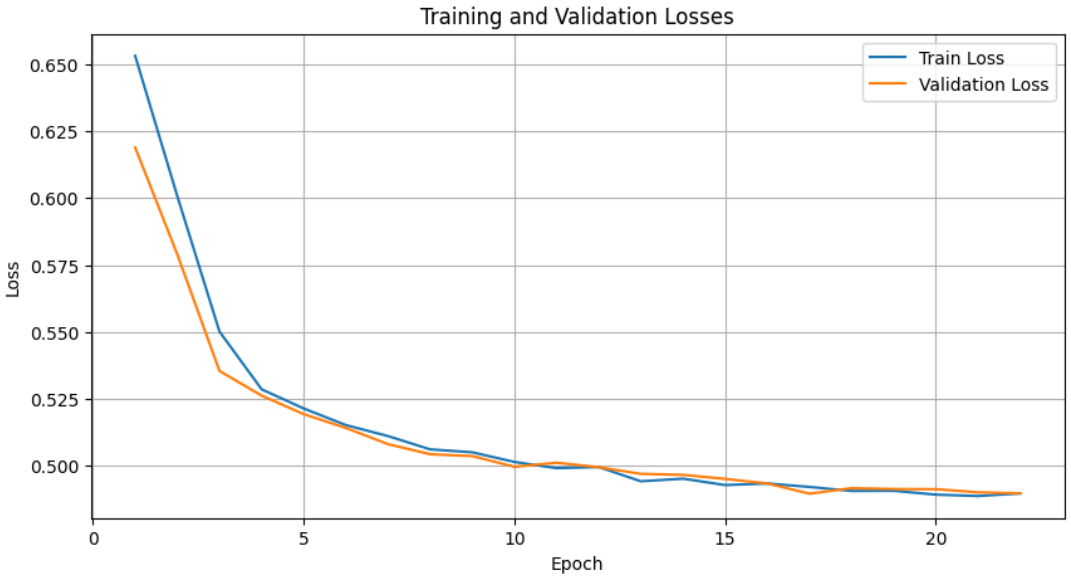


Figure III.18: Train and validation losses of the Autoencoder trained on ISIC2019.

(a) Real ISIC2019 (b) Real Fitz17k (c) Real Fitz17k

(d) Reconstructed (e) Reconstructed (f) Reconstructed

Figure III.19: AE real and reconstructed images.

4. **Threshold Calculation:** We first apply the MSE loss function to our predictions to find the reconstruction error of our reconstructed samples from both the in-distribution dataset and the OOD dataset stratified by skin tone categories (FST V-VI and FST I-IV). The threshold is then calculated using Brent's method [70], to find the root of the reconstruction error distributions for ID and OOD samples. Figure III.25 illustrates the calculated threshold.

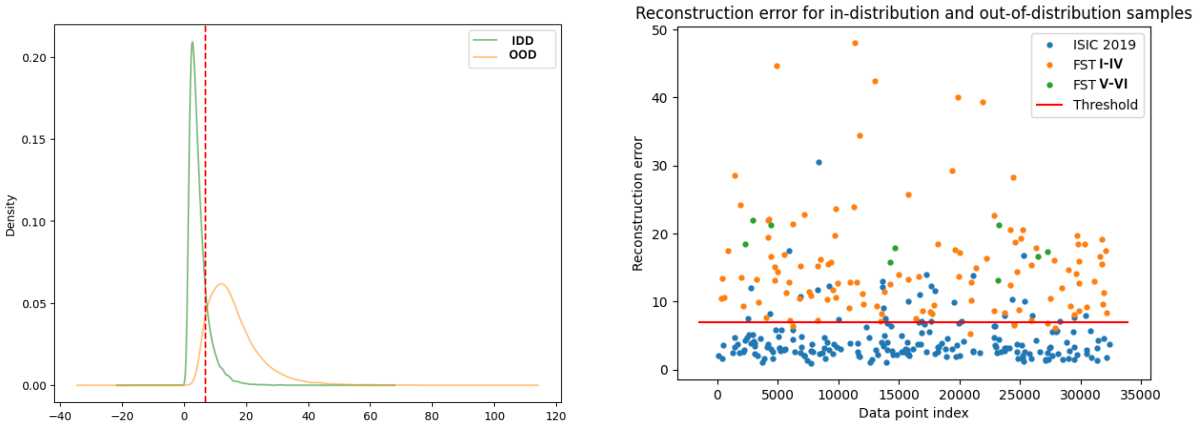


Figure III.20: AE reconstruction error thresholds.

5. **Model evaluation:** The Autoencoder assigned $\approx 91\%$ above threshold ($t_0 = 7$ and $t_1 = 8$) for FST I-IV and $\approx 98\%$ for FST V-VI.

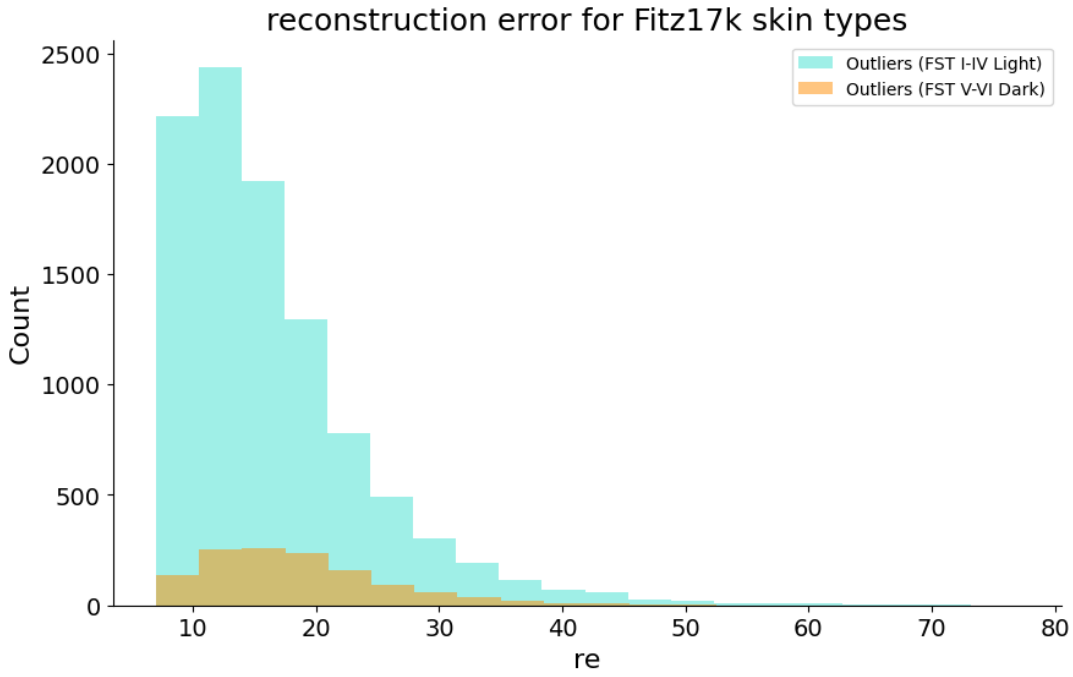


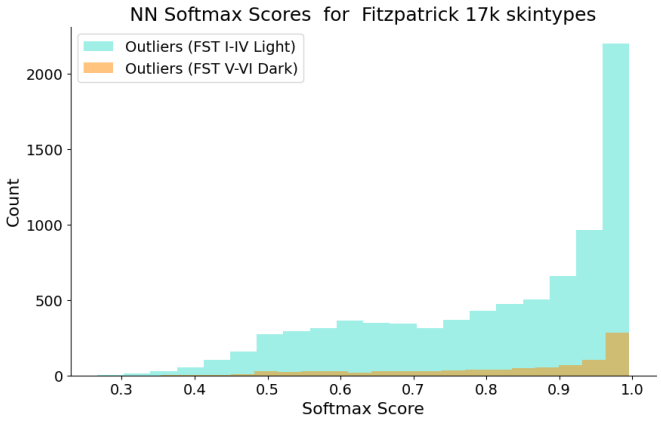
Figure III.21: Autoencoder histograms.

III.5.3.3 NN Softmax & ODIN

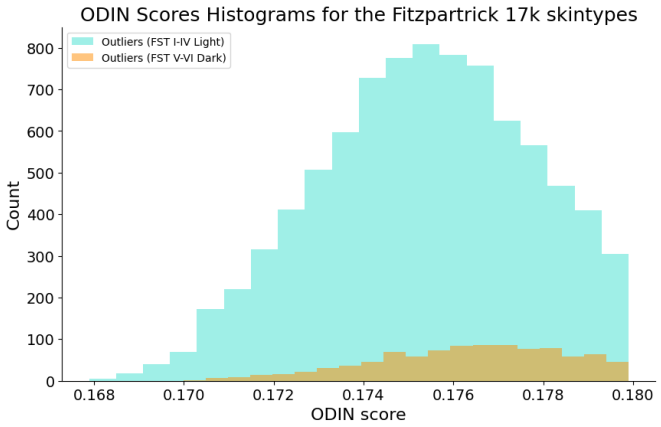
We employ NN Softmax and ODIN as our OOD detection methods. Following the approach outlined in [13], we use a pre-trained DenseNet model for diagnosis classification on the ISIC2019 dataset. We assess the performance of our OOD detectors by using the ISIC2019 dataset as the in-distribution data and the Fitzpatrick17k dataset as the OOD data.

1. **Loading and preprocessing datasets:** We preprocess and load images from the ISIC2019 and Fitzpatrick17k datasets using data generators for evaluating the ODIN and NN Softmax OOD detectors. During training, data augmentation techniques, such as resizing and random cropping, are applied to enhance model robustness. For validation, images are standardized to a consistent size and normalized. Data shuffling is performed at the end of each epoch to prevent order dependencies.
2. **Base line scores calculation:** We obtain initial confidence values for both in-distribution and OOD datasets using the model. The process involves passing images through a data generator, where each image is resized and normalized before model prediction. The maximum softmax score for each image is extracted to represent the model's confidence, facilitating the evaluation of prediction accuracy.
3. **ODIN scores calculation:** ODIN scores are computed by enhancing the separation between ID and OOD data through perturbations and temperature scaling. First, temperature scaling adjusts the sharpness of the softmax distribution. A grid search over temperature and perturbation parameters is conducted to optimize performance. Perturbations are generated by scaling the gradient of the loss with respect to each image and adding this to the original image. The perturbed image is then passed through the model to obtain softmax scores from the scaled logits. The highest softmax score for each image is recorded as the ODIN score. This approach improves OOD detection by lowering the confidence scores for OOD samples and increasing them for ID samples.
4. **Optimal thresholds:** To get the optimal parameters for both OOD models, we employ a grid search by testing a variation of parameters consisting of temperature scaling ($\tau = 200$) and magnitudes of perturbation ($\epsilon = 0.0002$) for ODIN. The threshold in these models is called optimal delta ($\delta = 0.996$ for NN Softmax and $\delta = 0.179$ for ODIN. Scores below δ are considered OOD samples, while scores above the threshold are considered in-distribution samples [13, 44, 45].
5. **Model evaluation:** According to the NN softmax and ODIN histograms shown in Fig-

ure III.22, NN Softmax assigned below threshold (optimal delta 0.996) $\approx 70\%$ of FST V-VI samples of true outliers and $\approx 75\%$ for FST I-IV samples, reducing even further the gap between skin types. While ODIN assigned bellow threshold (optimal $\delta = 0.179$) $\approx 28\%$ of FST V-VI samples of true outliers and $\approx 84\%$ for FST I-IV samples.



(a) NN Softmax



(b) ODIN

Figure III.22: NN softmax & ODIN Histograms

III.5.3.4 Performance metrics calculation

We evaluate our OOD detectors' performance across the skin tone categories FST I-IV and FST V-VI using two metrics that are widely used for OOD detection which are F1 score and AUROC. We perform K-fold cross-validation to ensure robust evaluation of each metric and to mitigate potential biases from the dataset split. We also calculate the RG score to have a better overview and quantify the disparity in model performance across different skin tone groups.

Table III.2: OOD detection performance for samples from two skin tone categories (FST I-IV and FST V-VI) with ISIC2019 as ID and Fitzpatrick17k as OOD.

	Datasets		AUROC ↑			F_1 ↑			$\mathcal{RG}$ ↓
Methods	IDD	OOD	FST I-IV	FST V-VI	All	FST I-IV	FST V-VI	All	
OneSVM [43]	ISIC 2019	Fitz17k	0.52 ± 0.011	0.53 ± 0.033	0.51 ± 0.011	0.67 ± 0.014	0.70 ± 0.015	0.64 ± 0.008	0.03
IF [42]	ISIC 2019	Fitz17k	0.53 ± 0.004	0.47 ± 0.013	0.52 ± 0.012	0.80 ± 0.009	0.89 ± 0.004	0.84 ± 0.014	0.09
AE [49]	ISIC 2019	Fitz17k	$\mathbf{0.97 \pm 0.005}$	$\mathbf{0.98 \pm 0.007}$	$\mathbf{0.97 \pm 0.005}$	$\mathbf{0.92 \pm 0.010}$	$\mathbf{0.93 \pm 0.013}$	$\mathbf{0.91 \pm 0.012}$	**0.02**
ODIN [44]	ISIC 2019	Fitz17k	0.67 ± 0.006	0.55 ± 0.006	0.64 ± 0.003	0.84 ± 0.001	0.50 ± 0.009	0.84 ± 0.001	0.34
NN Softmax [45]	ISIC 2019	Fitz17k	0.88 ± 0.002	0.84 ± 0.005	0.87 ± 0.001	0.85 ± 0.004	0.75 ± 0.014	0.84 ± 0.003	0.1

III.5.3.5 Discussion

Table III.2 shows the OOD detection performance for samples of different skin tone categories (FST I-IV and FST V-VI) across traditional and deep learning-based OOD methods. We can observe that in poorly general performance models such as IF and ODIN (AUROC 0.52, and 0.64), the representation gap measured between skin types is wider (0.09 and 0.3 respectively). Compared to AE (AUROC 0.97), skin type performance only differs for 0.02. Similar behavior can be seen in histograms shown in Figures III.15, III.21, III.22b, and III.22a. When we observe the different anomalous scores assigned by each method to both skin categories. We can observe that IF assigned higher abnormal scores (≥ 0.5) to $\approx 16\%$ of true outliers from FST I-IV skin type, while $\approx 36\%$ of outliers in FST V-VI were assigned. Similar behavior can be seen in the scores generated by ODIN. In comparison, the

AE ($t_0 = 7$ and $t_1 = 8$) assigned $\approx 91\%$ above the threshold for FST I-IV and $\approx 98\%$ for FST V-VI, while we see a reduction in the gap, from 20% to 6.9% between scores. Additionally, in Figure III.21, a more concentrated set of scores is shown within a small range for all samples across skin types compared to the rest of the score distributions.

III.5.4 Experiment III: OOD Detection using FitzPatick17k as ID and ISIC2019 as OOD

The purpose of this experiment is to evaluate the different OOD detection methods and compare their performance across the skin categories (FST V-VI and FST I-IV) using the Fitzpatrick17k dataset as ID, allowing us to train our OOD detectors on more FST V-VI samples. We use the ISIC2019 dataset as OOD for testing the performance of the OOD detectors. We adopt Isolation Forest and OneClassSVM as baselines, and an Autoencoder as state-of-the-art OOD methods.

III.5.4.1 One SVM & Isolation Forest

We follow the same experimental steps as $ExperimentI$ (data loading, preprocessing, labeling, model training, and evaluation). We conduct a grid search evaluation to get the optimal parameters of the One-class SVM and IF models rained on the Fitzpatrick17k dataset as in distribution. As a result of the grid search, the OneSVM was configured with $\nu = 0.01$ and $\gamma = 0.00001$. While the IF model was configured with 100 estimators and a contamination rate of 0.1 with 160 max samples.

III.5.4.2 Autoencoder

We trained the Autoencoder on the Fitzpatrick17k dataset by following the same steps as Experiment1 for data loading and preprocessing and the architecture, the training halted after 33 epochs. The trained Autoencoder can extract the main features and reconstruct the input images from each category as shown in Figure III.24. The thresholds are calculated for each category using Brent's method. The threshold for FST I-IV is $t_0 = 5$ and for FST V-VI is $t_1 = 8$.

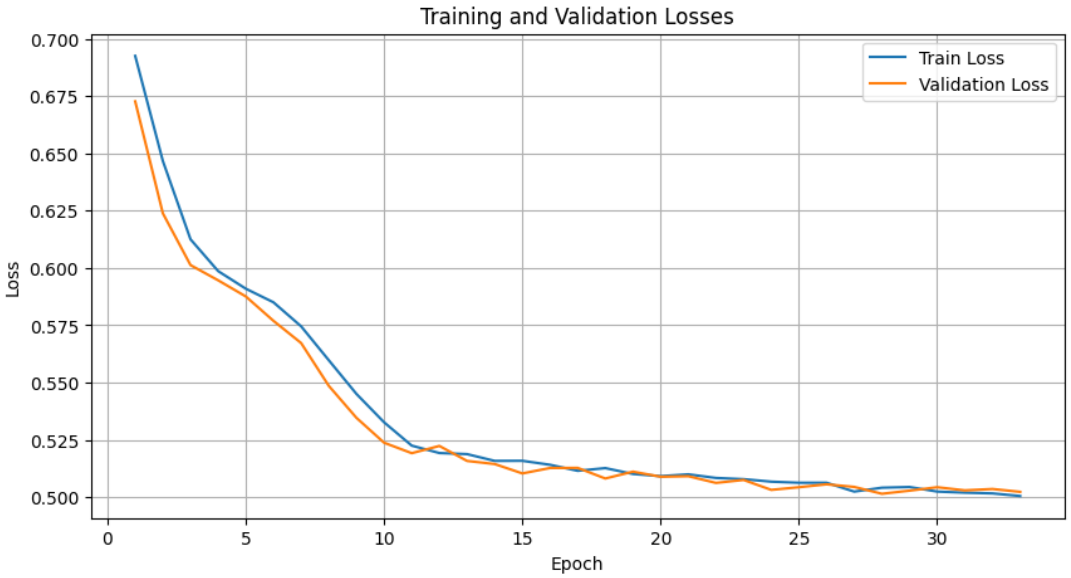
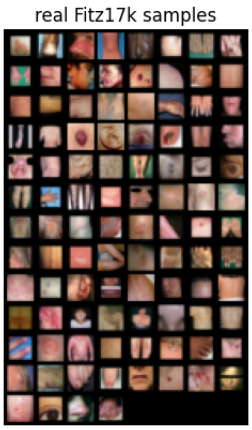
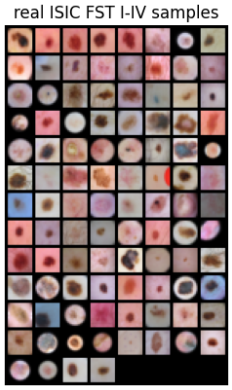


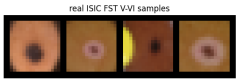
Figure III.23: Training and validation losses of the AE trained on Fitzpatrick17k.



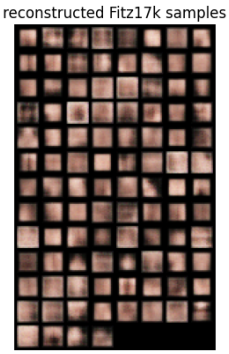
(a) Real Fitz17k



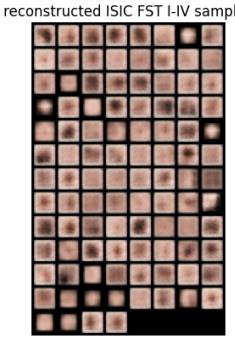
(b) Real ISIC2019



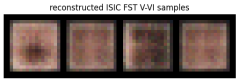
(c) Real ISIC2019



(d) Reconstructed



(e) Reconstructed



(f) Reconstructed

Figure III.24: AE real and reconstructed images.

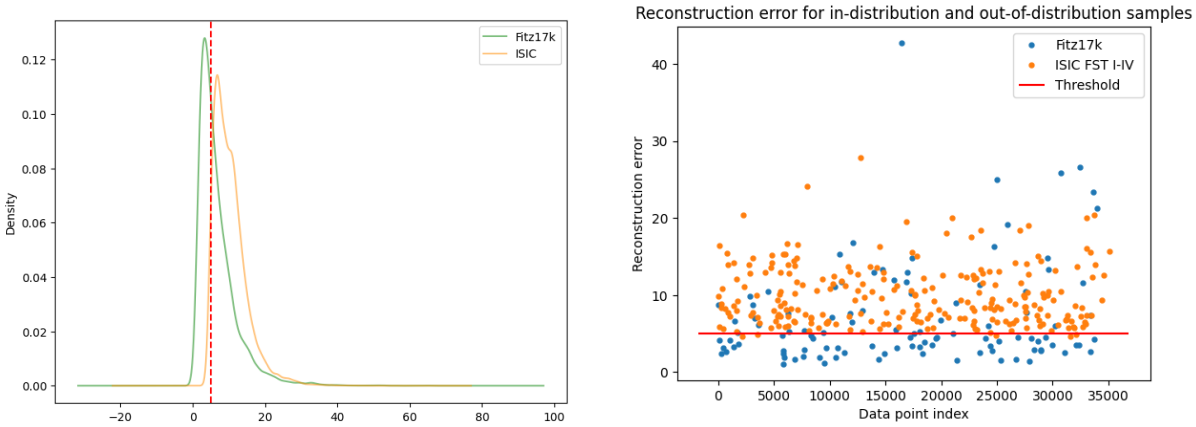


Figure III.25: AE reconstruction error thresholds.

III.5.4.3 Performance metrics calculation

We evaluate our OOD detectors on the ISIC2019 as out of distribution across the different skin tone categories. We perform K folds cross-validation in evaluating the F_1 and $AUROC$ metrics for each category then calculate the RG score. As we observed earlier, the FST V-VI category contains only 4 samples for testing.

Table III.3: OOD detection performance for samples from two skin tone categories(FST I-IV and FST V-VI). Fitzpatrick17k as ID and ISIC2019 as OOD. (*) Results were obtained over the only four samples FST V-VI of ISIC2019. (-): No DenseNet model available trained on Fitz17k.

	Datasets		AUROC ↑			F_1 ↑			$\mathcal{RG}$ ↓
Methods	IDD	OOD	FST I-IV	FST V-VI	All	FST I-IV	FST V-VI	All	
OneSVM [43]	Fitz17k	ISIC 2019	0.51 ± 0.014	0.27 ± 0.004(*)	0.51 ± 0.019	0.66 ± 0.028	0.72 ± 0.016(*)	0.66 ± 0.007	**0.06**
IF [42]	Fitz17k	ISIC 2019	0.57 ± 0.015	0.44 ± 0.000(*)	0.42 ± 0.008	0.85 ± 0.002	0.94 ± 0.004(*)	0.86 ± 0.005	**0.09**
AE [49]	Fitz17k	ISIC 2019	$\mathbf{0.93 \pm 0.002}$	$\mathbf{0.95 \pm 0.003}$(*)	$\mathbf{0.93 \pm 0.001}$	$\mathbf{0.89 \pm 0.002}$	0.84 ± 0.002(*)	$\mathbf{0.89 \pm 0.001}$	**0.05**
ODIN [44]	Fitz17k	ISIC 2019	-	-	-	-	-	-	-
NN Softmax [45]	Fitz17k	ISIC 2019	-	-	-	-	-	-	-

III.5.4.4 Discussion

Table III.3 presents the performance of traditional and deep learning-based OOD detection methods across different skin types, FST V-VI and FST I-IV. In poorly performing models such as IF and OneSVM (AUROC 0.51 and 0.42 respectively), the representation gap (RG) is approximately 0.06 and 0.09. In contrast, the Autoencoder, with a higher AUROC of 0.93, demonstrates a smaller RG of 0.05. The reduction in the RG score can be attributed to the limited number of FST V-VI samples in the OOD dataset.

III.5.5 Experiment IV: OOD Detection FitzPatick17k as ID and SD-198 as OOD

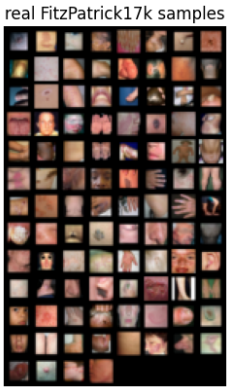
The purpose of this experiment is to evaluate our OOD methods and compare their performance across the skin categories FST V-VI and FST I-IV. We use the Fitzpatrick17k dataset as ID, and the SD-198 as OOD. This approach is chosen to focus on training and testing our OOD detectors with a higher representation of FST V-VI samples.

III.5.5.1 One SVM & Isolation Forest

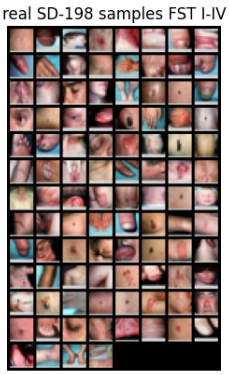
We train the OneSVM and IF models using the same data loading, preprocessing, and labeling steps as in the previous experiments. The models are trained on the Fitzpatrick17k dataset as the in-distribution data, using the same optimal parameters identified in $Experiment III$. The SD-198 dataset is used for testing as OOD.

III.5.5.2 AutoEncoder

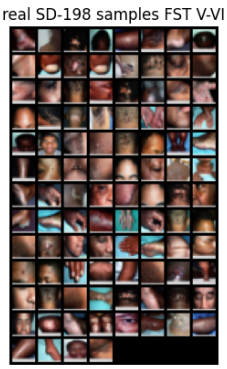
We use the trained AE on the Fitzpatrick17k dataset to evaluate its performance on the SD-198 dataset as OOD data. The model reconstructs the SD-198 samples, effectively identifying the main features of the input data. These results are illustrated in Figure III.26. The thresholds were determined using Brent's method, resulting in $t_0 = 9$ for FST I-IV and $t_1 = 11$ for FST V-VI. Figure III.27 illustrates the calculated threshold values.



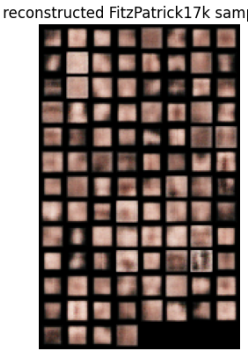
(a) Real Fitz17k



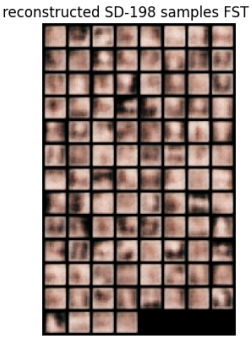
(b) Real SD-198



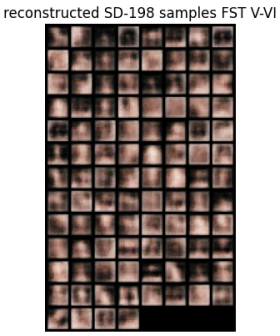
(c) Real SD-198



(d) Reconstructed



(e) Reconstructed



(f) Reconstructed

Figure III.26: AE real and reconstructed images.

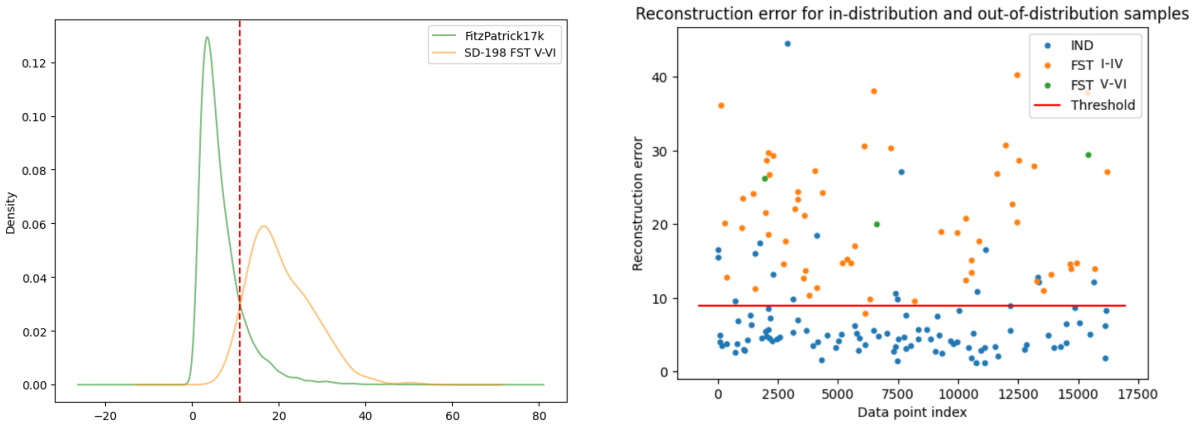


Figure III.27: Reconstruction error thresholds.

III.5.5.3 Performance metrics calculation

We evaluate the OOD detectors on the SD-198 dataset as OOD data across different skin tone categories. The evaluation metrics, including the F_1 score and AUROC, are calculated

using K-fold cross-validation. The representation gap (RG) score is derived from the F_1 metric results for the FST I-IV and FST V-VI categories.

Table III.4: OOD detection performance for samples from two skin tone categories (FST I-IV and FST V-VI) using Fitzpatrick17k as ID and SD-198 as OOD. (-): No DenseNet model available trained on Fitz17k.

	Datasets		AUROC ↑			F_1 ↑			$\mathcal{RG}$ ↓
Methods	IDD	OOD	FST I-IV	FST V-VI	All	FST I-IV	FST V-VI	All	
OneSVM [43]	Fitz17k	SD-198	0.5002 ± 0.00007	0.504 ± 0.01	0.499 ± 0.0006	0.88 ± 0.013	0.991 ± 0.001	0.874 ± 0.009	0.11
IF [42]	Fitz17k	SD-198	0.505 ± 0.01	0.63 ± 0.03	0.52 ± 0.023	0.84 ± 0.012	0.94 ± 0.008	0.842 ± 0.015	0.1
AE [49]	Fitz17k	SD-198	$\mathbf{0.883 \pm 0.0358}$	$\mathbf{0.932 \pm 0.0039}$	$\mathbf{0.944 \pm 0.0034}$	$\mathbf{0.824 \pm 0.0332}$	$\mathbf{0.785 \pm 0.0104}$	$\mathbf{0.868 \pm 0.0044}$	**0.039**
ODIN [44]	Fitz17k	SD-198	-	-	-	-	-	-	-
NN Softmax [45]	Fitz17k	SD-198	-	-	-	-	-	-	-

III.5.5.4 Discussion

According to Table III.4, the performance of our traditional and deep learning-based OOD detectors was evaluated across different skin types, FST V-VI and FST I-IV, using the SD-198 dataset as OOD. The OneClassSVM and Isolation Forest models continue to show poor performance, with AUROC scores of 0.49 and 0.52, respectively, and a higher representation gap (RG) score of 0.1. In contrast, the AutoEncoder demonstrates better performance, with an AUROC of 0.94 and a smaller RG score of 0.03. The improvement in the RG score is attributed to the increased number of FST V-VI samples in the OOD dataset.

III.5.6 Experiment V: Fairness analysis of OOD detectors

In this experiment, we employ mitigation techniques using the AIF360 toolkit to analyze the fairness of our OOD detectors across different skin type categories [67], specifically FST V-VI and FST I-IV. Our objective is to investigate the correlation between group fairness metrics and the Representation Gap (RG) score. We use group fairness to compare the average performance between members of the privileged group (FST I-IV) and the unprivileged group (FST V-VI) within our datasets.

To analyze group fairness, we use the following metrics: Statistical Parity Difference (SPD) and Disparate Impact (DI). Initially, we label our datasets and define the protected attribute as

the FST skin type. The privileged group is identified as FST I-IV, and the unprivileged group as FST V-VI. Favorable labels are defined as those selected as OOD. Figure III.28 illustrates the group fairness evaluation process with the AIF360 toolkit.

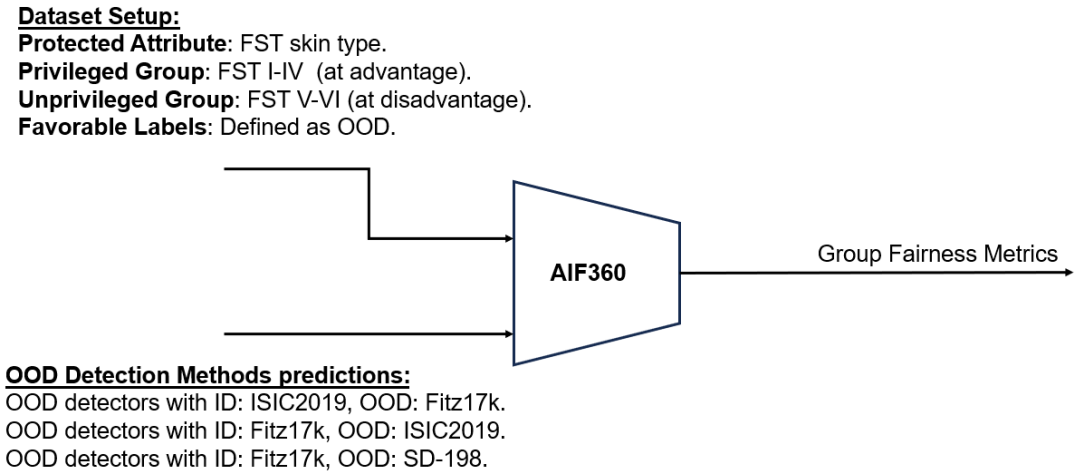
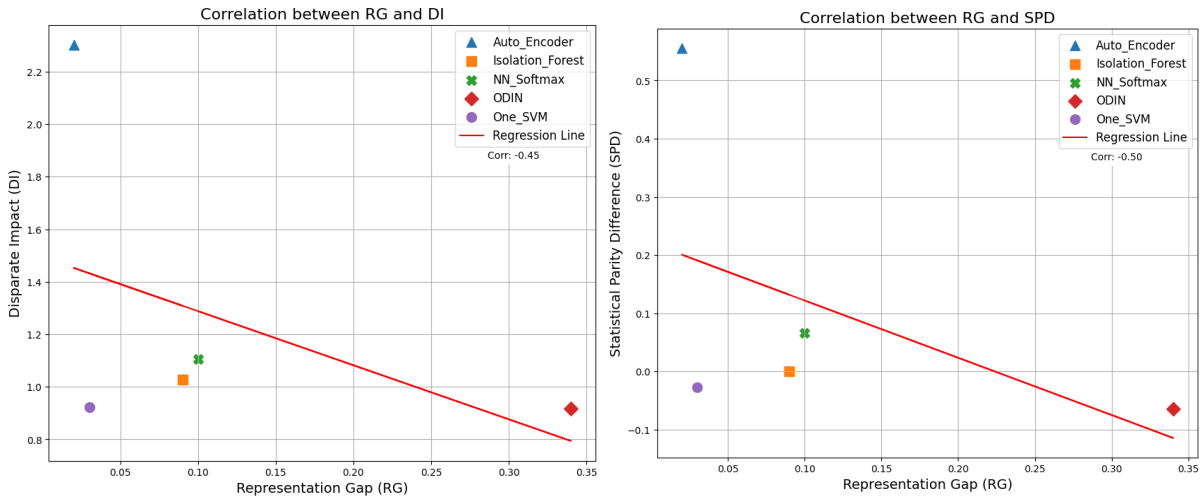


Figure III.28: Group fairness evaluation with AIF360 toolkit.

The group fairness measurements are presented in Table III.5, while the correlation between the group fairness metrics and the Representation Gap (RG) is depicted in the scatter plots shown in Figure III.29.



(a) RG&DI

(b) RG&SPD

Figure III.29: Correlation of Group Fairness metrics and RG score.

Table III.5: Group fairness metrics across the privileged group and unprivileged group.

Methods	IDD	OOD	SPD	DI
One SVM [43]	ISIC 2019	Fitzpatrick17k	-0.027	0.923
IF [42]	ISIC 2019	Fitzpatrick17k	0.0028	1.028
AE [49]	**ISIC 2019**	**Fitzpatrick17k**	**0.556**	**2.303**
ODIN [44]	ISIC 2019	Fitzpatrick17k	-0.065	0.917
NN Softmax [45]	ISIC 2019	Fitzpatrick17k	0.066	1.105
One SVM [43]	Fitzpatrick17k	ISIC2019	-0.0097	0.98
IF [42]	Fitzpatrick17k	ISIC2019	0.396	4.845
AE [49]	**Fitzpatrick17k**	**ISIC2019**	**0.162**	**1.193**
ODIN [44]	Fitzpatrick17k	ISIC2019	-	-
NN Softmax [45]	Fitzpatrick17k	ISIC2019	-	-
One SVM [43]	Fitzpatrick17k	SD-198	-0.069	0.863
IF [42]	Fitzpatrick17k	SD-198	0.267	3.991
AE [49]	**Fitzpatrick17k**	**SD-198**	**0.477**	**1.929**
ODIN [44]	Fitzpatrick17k	SD-198	-	-
NN Softmax [45]	Fitzpatrick17k	SD-198	-	-

III.5.6.1 Discussion

Based on the group fairness metrics results shown in Table III.5 and the correlation scatter plots in Figure III.29 , we can deduce several important findings. Firstly, in most of the poorly performing models, we see that the Statistical Parity Difference (SPD) values are close to 0, and the Disparate Impact (DI) values show that most OOD methods are closer to 1 which

are the default fairness values. While we observe that the best performing OOD detection method, the Autoencoder, displays a slight increase in both SPD and DI values (0.556 and 2.303 respectively), indicating some bias in OOD detection, with the unprivileged group (FST V-VI skin category) being at a disadvantage. We also observe a reduction in both SPD and DI values when training the Autoencoder on more FST V-VI values. From the correlation analysis, we observe a negative correlation between the RG score and group fairness metrics SPD and DI. This implies that the fairness of the OOD models is inversely related to the RG score, meaning that a smaller RG score corresponds to a fairer OOD detector across both the privileged and unprivileged groups, which in our study are the FST V-VI and FST I-IV skin tone categories.

III.5.7 General Discussion

Understanding the textural features of diverse skin tones is essential for our study and conducting a texture analysis on image samples from the Fitzpatrick17k dataset that contains the highest number of FST V-VI samples is essential to understanding the main textural features of skin tones categories (FST I-IV and FST V-VI) across the different skin conditions. The texture analysis results indicate the main key differences between the textural features of the skin categories making the FST V-VI (dark skin tone) texture more varied as it has higher Dissimilarity, Contrast, and Skewness statistical features, and the FST I-IV (light skin tone) texture more consistent with higher Correlation, Homogeneity, Energy, Kurtosis, Mean, and Variance as shown in Table III.1. In addition, understanding the performance of OOD methods beyond average metrics is critical for understanding potential blind spots and developing more fair approaches across diverse skin categories. A clear example of this can be seen in Table III.2, where IF, ODIN, and NN Softmax have similar overall F_1 scores, but when we observe performance by skin type, we see a ≈ 0.2 difference in the representation gap in both methods impacting FST I-IV and FST V-VI differently in each approach. This instability of performance may be partially because the Densenet used for NN softmax and ODIN is trained on a dataset that heavily lacks samples of Dark skin tones (ISIC2019). Training the OOD detectors on more FST V-VI samples leads to improvement in the average performance metrics with a slight increase in the RG scores as seen in Table III.3 and Table III.4. For instance, FST V-VI (brown and dark skin samples) used to train our OOD detectors constitute only 13.5% of Fitz17k and less than 0.01% of ISIC-2019. This could also encourage OOD detectors to classify them to be out of distribution easily. Getting an overall stable rate of the RG score for the AE performance makes it the best-performing OOD detector in all the experiments with 0.02, 0.05, and 0.03 respectively. Finally, the use of group fairness metrics is essential in analyzing the skin tone representation by determining the bias resulting from our OOD detectors. Table III.5 shows the variation of the group fairness across OOD detectors trained on ISIC2019 and Fitz17k datasets. We see that the group fairness metrics are reduced when training the OOD detectors on more FST V-VI samples which makes our models less biased to the FST V-VI samples as OOD. Making it a fairer OOD detector across the two skin categories.

III.6 Summary

In this chapter, we conducted a comparative study of baseline and state-of-the-art OOD detectors across different datasets with varied skin tone representations of clinical and dermoscopic images. In addition, we could understand the main differences in skin tone textures and representation by conducting a texture analysis and fairness analysis of our OOD detectors across diverse skin categories.

General Conclusion

We propose an evaluation framework to assess the impact of skin tone representation on OOD detection. the GLCM texture analysis demonstrated the key differences between skin tone categories across different skin conditions available in the Fitzpatrick17k dataset which proves that the skin tones differ in terms of texture across the different skin conditions. We stratify OOD samples based on skin tone and observe imbalanced detection performance for FST V-VI samples, where the samples from darker skin tones are detected as OOD with higher performance in most cases making the OOD detector biased to this category. We showcase the importance of quantifying the representation gap, as the existing OOD models with similar overall performance diverge differently on skin types. This information should be considered when deciding the OOD method type to implement in the robustness pipeline. Furthermore, we provided labeled samples for ISIC-2019 and SD-198, and we highlighted the need for more diverse dermatoscopic datasets. While clinical datasets, such as the Fitzpatrick17k dataset, yield more representation across both labels. Finally, we assessed the OOD detectors' fairness using the AIF360 toolkit. The group fairness analysis results demonstrated the bias reduction of group fairness metrics across skin tone groups in various scenarios.

Future work aims to further understand the impact of different proportions of skin types and potential interventions that can be done during training to reduce the representation gap we observed in this study and to train the DenseNet architecture on clinical datasets with diverse skin representations and test more DL-based OOD detectors.

List of Publications

The following paper has been published as part of the work conducted in this thesis:

- Benmalek, A., Cintas, C., Tadesse, G.A., Daneshjou, R., Varshney, K., Dalila, C.: ”Evaluating the impact of skin tone representation on out-of-distribution detection performance in dermatology.” In: *IEEE International Symposium on Biomedical Imaging* (2024).

Bibliography

[1] Yawen Wu et al. “FairPrune: Achieving Fairness Through Pruning for Dermatological Disease Diagnosis”. In: *MICCAI*. Springer. Cham, 2022, pp. 743–753. ISBN: 978-3-031-16431-6.

[2] Abeba Birhane et al. “On Hate Scaling Laws For Data-Swamps”. In: *arXiv preprint arXiv:2306.13141* (2023).

[3] Andre Esteva et al. “Dermatologist-level classification of skin cancer with deep neural networks”. In: *Nature* 542.7639 (2017), pp. 115–118.

[4] Arieh Gomolin et al. “Artificial intelligence applications in dermatology: Where do we stand?” In: *Front. Med.* 7 (2020).

[5] Pengyi Zhang, Yunxin Zhong, and Xiaoqiong Li. “MelaNet: A Deep Dense Attention Network for Melanoma Detection in Dermoscopy Images”. In: (2019).

[6] Adewole S Adamson and Avery Smith. “Machine learning and health care disparities in dermatology”. In: *JAMA Derm.* 154.11 (2018), pp. 1247–1248.

[7] Adnan Qayyum et al. *Secure and robust machine learning for healthcare: A survey*. arXiv:2001.08103. 2020.

[8] Newton M Kinyanjui et al. “Fairness of classifiers across skin tones in dermatology”. In: *Proc. Int. Conf. Med. Image Comp. Comp.-Assist. Interv.* 2020, pp. 320–329.

[9] Girmaw Abebe Tadesse et al. “Skin Tone Analysis for Representation in Educational Materials using machine learning”. In: *npj Digital Medicine* 6.1 (2023), p. 151.

[10] Thorsten Kalb et al. “Revisiting Skin Tone Fairness in Dermatological Lesion Classification”. In: *Workshop on Clinical Image-Based Procedures*. Springer. 2023, pp. 246–255.

[11] Usha Lee Mcfarling. *Dermatology faces a reckoning: Lack of darker skin in textbooks and journals harms care for patients of color*. July 2020.

[12] Ademide Adelekun, Ginikanwa Onyekaba, and Jules B Lipoff. "Skin color in dermatology textbooks: an updated evaluation and analysis". In: *Journal of the American Academy of Dermatology* 84.1 (2021), pp. 194–196.

[13] Hannah Kim et al. "Out-of-distribution detection in dermatology using input perturbation and subset scanning". In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–4.

[14] Megan Lam et al. "Racial differences in the prognosis and survival of cutaneous melanoma from 1990 to 2020 in north America: a systematic review and meta-analysis". In: *Journal of Cutaneous Medicine and Surgery* 26.2 (2022), pp. 181–188.

[15] Sean M Dawes et al. "Racial disparities in melanoma survival". In: *Journal of the American Academy of Dermatology* 75.5 (2016), pp. 983–991.

[16] Shasa Hu et al. "Disparity in melanoma: a trend analysis of melanoma incidence and stage at diagnosis among whites, Hispanics, and blacks in Florida". In: *Archives of dermatology* 145.12 (2009), pp. 1369–1374.

[17] Shasa Hu et al. "Comparison of stage at diagnosis of melanoma among Hispanic, black, and white patients in Miami-Dade County, Florida". In: *Archives of Dermatology* 142.6 (2006), pp. 704–708.

[18] Mary Dick, Sarah Aurit, and Peter Silberstein. "The odds of stage IV melanoma diagnoses based on socioeconomic factors". In: *Journal of cutaneous medicine and surgery* 23.4 (2019), pp. 421–427.

[19] Toral Vaidya et al. "Socioeconomic and geographic barriers to dermatology care in urban and rural US populations". In: *Journal of the American Academy of Dermatology* 78.2 (2018), pp. 406–408.

[20] Lumen Learning. *Pigmentation*. 2023. URL: https://courses.lumenlearning.com/wm-biology2/chapter/pigmentation/ (visited on 06/09/2023).

[21] Amin Mahmood Thawabteh et al. "Skin Pigmentation Types, Causes and Treatment—A Review". In: *Molecules* 28.12 (2023), p. 4839. DOI: 10.3390/molecules28124839.

[22] V. Gupta and V. K. Sharma. “Skin typing: Fitzpatrick grading and others”. In: *Clinical Dermatology* 37.5 (Sept. 2019), pp. 430–436. DOI: 10.1016/j.clindermatol.2019.07.010.

[23] V. E. Nambudiri. *Overview of Skin Cancer*. In MSD Manual Professional Edition. Accessed: 2024-09-19. 2023. URL: https://www.msdmanuals.com/professional/dermatologic-disorders/cancers-of-the-skin/overview-of-skin-cancer.

[24] Matthew Groh et al. “Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1820–1828.

[25] Matthew Groh et al. “Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm”. In: *arXiv preprint arXiv:2207.02942* (2022).

[26] DermNet NZ. *Benign Skin Lesions*. Accessed: 2024-09-17. 2024. URL: https://dermnetnz.org/topics/benign-skin-lesions.

[27] Johns Hopkins Medicine. *Other Benign Skin Growths*. Accessed: 2024-09-19. 2024. URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/other-benign-skin-growths.

[28] V. E. Nambudiri. *Overview of Skin Cancer*. In MSD Manual Professional Edition. Accessed: 2024-09-19. 2024. URL: https://www.msdmanuals.com/professional/dermatologic-disorders/cancers-of-the-skin/overview-of-skin-cancer.

[29] *Genetic Skin Disorders*. 3rd ed. In Oxford Monographs on Medical Genetics. Oxford Academic, n.d. URL: https://academic.oup.com/oxford-monographs-on-medical-genetics.

[30] Medpace. *Dermatology Clinical Trials: Best Practices in Digital Photography*. https://www.medpace.com/wp-content/uploads/2023/11/Whitepaper-Dermatology-Clinical-Trials-Best-Practices-in-Digital-Photography.pdf. Accessed: 2024-08-26. 2023.

[31] J.C. Lester et al. “Absence of images of skin of colour in publications of COVID-19 skin manifestations”. In: *British Journal of Dermatology* (2023).

[32] R.M. Haralick. “Statistical and structural approaches to Texture”. In: *Proceedings of the IEEE* 67.5 (1979), pp. 784–804.

[33] Le-Qing Li and Hui Fu. “Automated detection of skin diseases using texture features”. In: *2011 4th International Congress on Image and Signal Processing* 4 (2011), pp. 1512–1516.

[34] Nur Hazwani Hazani and Wan Mahani Hafizah Wan Mahmud. “Analysis of Human Skin Texture by using Machine Learning Approaches”. In: *Evolution in Electrical and Electronic Engineering* 4.1 (2023), pp. 356–362. DOI: https://publisher.uthm.edu.my/periodicals/index.php/eeee/article/view/10808.

[35] Prof. Nikita O. Paunikar et al. “SKIN DISEASE RECOGNITION USING TEXTURE ANALYSIS”. In: *International Research Journal of Modernization in Engineering Technology and Science* 5.6 (2023), pp. 3238–3245. DOI: 10.56726/IRJMETS42405.

[36] Jessica Hayes. *Grey Scale Co-Occurrence Matrix*. Portland State University. Accessed: 2024-09-19. 2007. URL: https://web.pdx.edu/~jduh/courses/Archive/geog481w07/Students/Hayes_GreyScaleCoOccurrenceMatrix.pdf.

[37] *Plot GLCM*. In scikit-image. Accessed: 2024-09-19. URL: https://scikit-image.org/docs/stable/auto_examples/features_detection/plot_glcm.html.

[38] *GLCM Texture Features*. In Echoview Support. Accessed: 2024-09-19. URL: https://support.echoview.com/WebHelp/Windows_And_Dialog_Boxes/Dialog_Boxes/Variable_Properties_Dialog_Box/Operator_Pages/GLCM_Texture_Features.htm.

[39] *The Equations of GLCM Features*. In ResearchGate. Accessed: 2024-09-19. URL: https://www.researchgate.net/figure/The-Equations-of-GLCM-Features_tbl1_332247376.

[40] *The Equations of GLCM Features*. In HAL. Accessed: 2024-09-19. URL: https://hal.science/hal-01376668/file/supplemental_material.pdf.

[41] Karina Zadorozhny and Giovanni Cinà. “Out-Of-Distribution Detection in Medical AI”. In: *Geek Culture* (June 2021). Accessed: 2024-05-27. URL: https://medium.com/geekculture/out-of-distribution-detection-in-medical-ai-b638b385c2a3.

[42] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. “Isolation forest”. In: *2008 eighth ieee international conference on data mining*. IEEE. 2008, pp. 413–422.

[43] Stephan Dreiseitl et al. “Outlier Detection with One-Class SVMs: An Application to Melanoma Prognosis”. In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2010 (Nov. 2010), pp. 172–6.

[44] Shiyu Liang, Yixuan Li, and R. Srikant. *Principled Detection of Out-of-Distribution Examples in Neural Networks*. arXiv:1706.02690. 2017.

[45] Dan Hendrycks and Kevin Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. In: *arXiv preprint arXiv:1610.02136* (2016).

[46] Xuan Li et al. “Out-of-Distribution Detection for Skin Lesion Images with Deep Isolation Forest”. In: *arXiv preprint arXiv:2003.09365* (2020).

[47] Yuchen Lu and Peng Xu. “Anomaly detection for skin disease images using variational autoencoder”. In: *arXiv preprint arXiv:1807.01349* (2018).

[48] Muhammad Zaida et al. “Out of distribution detection for skin and malaria images”. In: *arXiv preprint arXiv:2111.01505* (2021).

[49] Jonathan Masci et al. “Stacked convolutional auto-encoders for hierarchical feature extraction”. In: *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*. Springer. 2011, pp. 52–59.

[50] David M. J. Tax and Robert P. W. Duin. “Intrusion Detection in Unlabeled Data with Quarter-sphere Support Vector Machines”. In: *Machine Learning* 45 (2001), pp. 5–8.

[51] *IsolationForest example — scikit-learn 1.6.dev0 documentation*. `https://scikit-learn.org/dev/auto_examples/ensemble/plot_isolation_forest.html`. 2023.

[52] Dave Bergmann and Cole Stryker. *What Is an Autoencoder? — IBM*. Published: 23 November 2023. Nov. 2023. URL: `https://www.ibm.com/topics/autoencoder`.

[53] Steven Flores. *Variational Autoencoders are Beautiful*. Apr. 2019. URL: `https://www.compthree.com/blog/autoencoder/`.

[54] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[55] Shiyu Liang, R Srikant, and Yixuan Li. “Enhancing the reliability of out-of-distribution image detection in neural networks”. In: *International Conference on Learning Representations*. 2018.

[56] Dan Hendrycks and Kevin Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. In: *arXiv preprint arXiv:1610.02136* (2017).

[57] Kush R. Varshney. *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022. URL: http://www.trustworthymachinelearning.com/trustworthymachinelearning.pdf.

[58] Marcus Wilkes et al. “Fitzpatrick Skin Type, Individual Typology Angle, and Melanin Index in an African Population”. In: *JAMA Dermatol.* 151.8 (Aug. 2015), pp. 902–903.

[59] Eman Rezk, Mohamed Eltorki, Wael El-Dakhakhni, et al. “Improving skin color diversity in cancer detection: deep learning approach”. In: *JMIR Dermatology* 5.3 (2022), e39143.

[60] Peter J. Bevan and Amir Atapour-Abarghouei. “Skin Deep Unlearning: Artefact and Instrument Debiasing in the Context of Melanoma Classification”. In: (2021). DOI: 10.48550/ARXIV.2109.09818.

[61] Neda Alipour, Ted Burke, and Jane Courtney. “Skin Type Diversity: a Case Study in Skin Lesion Datasets”. In: (July 2023). DOI: 10.21203/rs.3.rs-3160120/v1.

[62] Noel C. F. Codella et al. *Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI)*. arXiv:1710.05006. 2017.

[63] Xiaoxiao Sun et al. “A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images”. In: *Proc. Europ. Conf. Comput. Vis.* 2016, pp. 206–222.

[64] Thomas B Fitzpatrick. “The validity and practicality of sun-reactive skin types I through VI”. In: *Archives of dermatology* 124.6 (1988), pp. 869–871.

[65] Andrew P. Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern recognition* 30.7 (1997), pp. 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2.

[66] C. J. Van Rijsbergen. *Information Retrieval*. 2nd ed. Butterworth-Heinemann, 1979.

[67] Rachel KE Bellamy et al. “AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias”. In: *arXiv preprint arXiv:1810.01943* (2018).

[68] Michaela Wick, Golnoosh Mokhtari, and Jean-Baptiste Tristan. “Fairness Metrics: A Comparative Analysis”. In: *arXiv preprint arXiv:2004.04870* (2020).

[69] Michael Feldman et al. “Certifying and removing disparate impact”. In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015), pp. 259–268.

[70] Richard P. Brent. *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall, 1973.