Automated Image Quality Assessment in Teledermatology

Choekyel Nyungmartsang

Lucerne University of Applied Sciences and Arts
6343 Rotkreuz, Switzerland

Bachelor Thesis
Bachelor of Science in Artificial Intelligence & Machine Learning

Friday 7th June, 2024

Advisor: Dr. Amruthalingam Ludovic
Lucerne University of Applied Sciences and Arts, 6343 Rotkreuz, Switzerland
ludovic.amruthalingam@hslu.ch

Expert: Andrin Bürli
Centre Suisse dElectronique et de Microtechnique (CSEM), 6055 Alpnach, Switzerland
andrin@bluemail.ch

Bachelor Thesis at Lucerne University of Applied Sciences and Arts School of Computer Science and Information Technology

Title Automated Image Quality Assessment in Teledermatology

Student: Choekyel Nyungmartsang

Degree Program: B.Sc. Artificial Intelligence and Machine Learning

Year of Graduation: 2024

Main Advisor: Dr. Amruthalingam Ludovic

External Expert: Andrin Bürli

Industry partner/provider: Algorithmic Business Research Lab, University Hospital of Basel and derma2go

Code / Thesis Classification:

- ☒ Public (Standard)
- ☐ Private

Declaration

I hereby declare that I have completed this thesis alone and without any unauthorized or external help. I further declare that all the sources, references, literature and any other associated resources have been correctly and appropriately cited and referenced. The confidentiality of the project provider (industry partner) as well as the intellectual property rights of the Lucerne University of Applied Sciences and Arts have been fully and entirely respected in completion of this thesis.

Rotkreuz, Friday 7th June, 2024 _____

Submission of the Thesis to the Portfolio Database:

Confirmation by the student

I hereby confirm that this bachelor thesis has been correctly uploaded to the Portfolio Database in line with the code of practice of the University. I rescind all responsibility and authorization after upload so that no changes or amendments to the document may be undertaken.

Rotkreuz, Friday 7th June, 2024 _____

Expression of Thanks and Gratitude

I would like to express my heartfelt thanks to Dr. Amruthalingam Ludovic for the interesting discussions and professional supervision throughout the entire duration of this research. Further thanks go to my family for their support, and especially to my sister Lhakyi Nyungmartsang for helping to filter the images and to Tenzin Lhundup Tsarma for proofreading.

Abstract

This research focuses on developing and evaluating automated methods to assess image quality in the context of teledermatology. Teledermatology, a growing field within telemedicine, allows patients to receive dermatological consultations remotely by sending photos of their skin conditions to dermatologists. However, the success of these remote consultations largely depends on the quality of the images provided. Poor-quality images can lead to misdiagnosis or require patients to resend images, causing delays and frustration. This research examines different image quality assessment (IQA) techniques for teledermatology. The goal is to make sure that only good quality images are sent to dermatologists, thereby improving their ability to make accurate medical diagnoses. By implementing these IQA techniques, the aim is to simplify the process, reduce back-and-forth communication, and make teledermatology more efficient and reliable.

The methodology involved a detailed review of existing IQA methods, followed by the development of a synthetic distortion pipeline to create a wide range of training datasets from good quality dermatological images. These images were gathered from the Fitzpatrick17k and SCIN datasets. For feature extraction, the state-of-the-art approach called ARNIQA was used, and different machine learning models, including XGBRegressor, XGBClassifier, MLP Regressor, and MLP Classifier, were trained to assess image quality based on seven dermatology quality criteria: lighting, background, field of view, orientation, focus, resolution, and color calibration. The models were trained on synthetic distortions and validated on both synthetically distorted and real-world dermatology images. Performance metrics such as Mean Absolute Error (MAE), R-squared (R^2), Spearmans Rank Order Correlation Coefficient (SRCC), and Cohens Kappa were used for evaluation.

The results showed that the automated IQA methods can assess image quality in the context of teledermatology, closely matching human evaluations and therefore providing reliable feedback on image quality. The final model achieved good performance across multiple dermatology quality criteria, improving the reliability and effectiveness of teledermatology services. This research highlights the potential of automated IQA systems to improve the accuracy of diagnoses and patient care in remote dermatological consultations.

Contents

Chapter 1

Introduction

1.1 Background and Problem Statement

In recent years, teledermatology, a branch of telemedicine, has become increasingly popular, especially due to the COVID-19 pandemic. Telemedicine uses telecommunications technology to provide healthcare services remotely, allowing patients to consult with healthcare providers without needing in-person appointments. Teledermatology takes advantage of this technology to diagnose and manage skin conditions remotely. Patients use their mobile devices to take pictures of their skin conditions and send these images to dermatologists for analysis. This process removes the need for face-to-face appointments, making healthcare more accessible and convenient.

While teledermatology has many benefits, it heavily depends on the quality of the images that patients submit. Many images are not up to standard due to issues like poor lighting, blurriness, or not clearly showing the skin condition (Vodrahalli et al., 2020). These poor-quality images make it difficult for dermatologists to make accurate diagnoses, leading to a back-and-forth exchange of informations between patient and dermatologist. This process can be time-consuming and frustrating, reducing the overall accuracy of teledermatology.

To address this problem, it is very important to improve the quality of images taken by patients. This thesis aims to develop an automated image quality assessment (IQA) techniques that can evaluate the quality of images before they are sent to dermatologists. By making sure that only good-quality images are submitted, the reliability and effectiveness of teledermatology can be greatly improved.

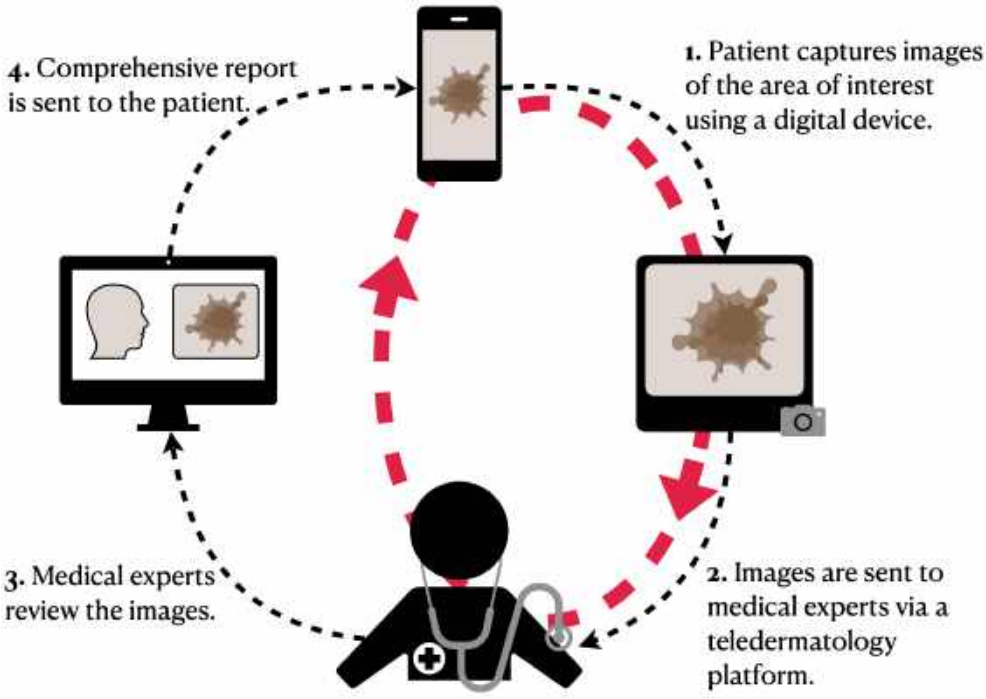


Figure 1.1: Simplified Teledermatology Consultation Process. The red arrows highlight the back-and-forth exchange due to poor-quality images, which can delay diagnosis and treatment.

1.2 Objectives of the Thesis

The primary goal of this thesis is to develop and evaluate automated methods for assessing image quality in teledermatology. The objectives cover several key areas, starting with a detailed review of existing image quality assessment methods. This review will help determine which methods are appropriate for use in the context of teledermatology. The thesis also aims to identify the right dermatology quality criteria, apply selected methods to relevant dermatological images, and create a reproducible repository for future research.

The specific objectives of this thesis are as follows:

- Conduct a detailed review of existing image quality assessment (IQA) methods, focusing on their use in teledermatology.
- Identify and choose the most appropriate metrics for assessing the quality of dermatological images.
- Evaluate the performance of selected image quality metrics on dermatological datasets to determine their effectiveness in assessing image quality.
- Create a reproducible repository of tools and methods for assessing image quality in teledermatology.

Achieving these objectives will greatly improve teledermatology services by providing a reliable way to assess image quality.

1.3 Organisation of this Thesis

This thesis is structured into seven chapters to provide a clear and systematic exploration of image quality assessment in teledermatology. Chapter 2 covers the literature review, discussing previous and related works on IQA and teledermatology. Chapter 3 details the methodologies used for the literature review and specific methods for IQA in teledermatology. Chapter 4 describes the experiments conducted, showing the approaches taken and the metrics used. Chapter 5 presents the results of the research. Chapter 6 interprets the results, discusses key assumptions and reflects on the findings. Finally, Chapter 7 summarizes the findings and suggests directions for future research.

All figures and tables in this thesis are created by the author unless otherwise referenced. Documents and code relevant to this thesis can be downloaded from the following link: https://github.com/Schoggi-Mimi/bachelor-thesis. Any code referenced within the thesis is from this repository, with specific paths or modules provided in the footnotes. For simplicity, image quality assessment will be referred to as IQA throughout the document.

Chapter 2

Literature Review

2.1 Image Evaluation

There are three ways to evaluate an image: assessing its quality, aesthetics, or fidelity (Zhou & Alan, 2007). Each method focuses on different aspects of image evaluation and has unique applications.

Image Quality Assessment (IQA) measures the degradation of an image. This involves comparing an original, undistorted image with a processed version that has undergone changes such as compression, noise addition, or artifact introduction. The goal is to quantify how much the image quality has declined due to these changes.

Image Aesthetics Assessment focuses on the visual appeal of an image. It evaluates how pleasing an image is to the human eye, considering factors like composition, color, and overall aesthetic impact. While related to IQA, since both involve human judgment, this area is not the focus of this thesis because it deals more with subjective perceptions of beauty rather than measurable quality degradations.

Image Fidelity Assessment deals with how accurately an image represents the original scene or view. This is especially relevant in applications involving multiple views or stereo cameras, assessing the correctness of image reconstruction. However, this thesis will also not cover image fidelity assessment, as it relates more to the accuracy of recreating an image rather than evaluating its quality after processing.

The primary focus of this thesis is on image quality assessment, specifically looking at different types of image degradation.

2.1.1 Methods of Image Quality Assessment

There are two main approaches to assess image quality: subjective and objective methods (Zhou & Alan, 2007).

Subjective Quality Assessment

Subjective quality assessment involves human observers evaluating the quality of images based on their visual perception. This method is essential for understanding how humans perceive image quality in real-world situations, especially when technical measurements might not fully capture what people actually see and experience. There are two primary methods used in subjective quality assessment:

- Absolute Categorical Rating: In this approach, human observers are presented with an unlabeled image and asked to rate its quality based on predefined categories. Each observer evaluates the image independently, without comparing it to any reference image. This method allows evaluators to provide a direct judgment on the images quality based on their subjective experience.
- Paired Comparison: In this method, human observers are presented with two images: an unlabeled image and a reference image. Observers then assess the quality of the unlabeled image by comparing it directly to the reference image, assigning a score based on the perceived differences in quality.

Subjective quality assessment is highly valued for its ability to accurately reflect human perception of image quality (Zhou & Alan, 2007). However, this method is also resource-intensive, requiring significant time and effort from human evaluators. Additionally, subjective assessments can be influenced by variability and biases introduced by individual scorers. For example, differences in monitor color calibration, domain knowledge, and personal preferences can affect the consistency and reliability of the evaluations. Despite these challenges, subjective quality assessment remains a critical component of comprehensive image quality evaluation, particularly in applications where how a human perceives the image is the final measure of its quality.

Objective Quality Assessment

Objective quality assessment uses mathematical algorithms to evaluate image quality instead of relying on human judgment. These algorithms are often based on our understanding of how the human vision system works. However, not all methods directly simulate human vision. Some methods use different techniques to measure quality by comparing specific features or data points extracted from images. This type of assessment is mainly categorized into three methods based on the amount of reference data used: Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA), and No-Reference IQA (NR-IQA) (Zhou & Alan, 2007).

Full-Reference IQA (FR-IQA) involves a detailed comparison between a distorted image and a reference image (see Figure 2.1). Features are extracted from both images, and their differences are quantitatively analyzed to compute a quality score. While FR-IQA offers detailed assessments, it requires a reference image for every distorted image evaluated, which can limit its practicality.

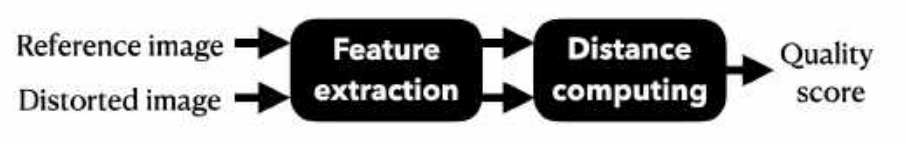


Figure 2.1: General framework of FR-IQA algorithms. Features are extracted from both images, and then the feature distance is calculated.

Reduced-Reference IQA (RR-IQA) operates similarly to FR-IQA but does not need the complete reference image. Instead, it uses a reduced set of features extracted from both the distorted and reference images (see Figure 2.2). This method balances the detailed comparison of FR-IQA with the independence of NR-IQA (which will be discussed later), reducing computational requirements while still providing meaningful quality assessments based on partial reference data.

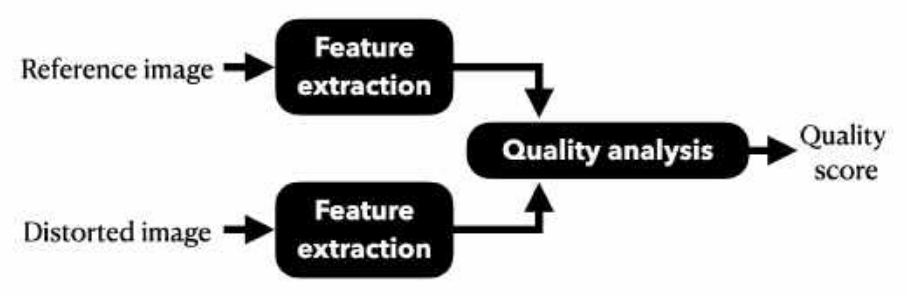


Figure 2.2: General framework of RR-IQA algorithms. Features of the reference and distorted images are extracted and used collectively to compute the quality.

Both FR-IQA and RR-IQA use two methods to analyze image quality (Zhou & Alan, 2007):

- Spatial-Based Analysis: This method compares images pixel by pixel or region by region. It offers straightforward interpretation and efficient computation. However, it may not fully align with how humans process images and can lack robustness in some scenarios.
- Transform-Based Analysis: This approach transforms images into a different domain (such as the frequency domain) that more closely mimics how humans process images. While this method is robust, it is complex and computationally intensive.

No-Reference IQA (NR-IQA) does not rely on any reference image. Instead, it analyzes the distorted image alone by extracting features indicative of quality (see Figure 2.3). This method is particularly useful when no reference images are available, such as in many practical applications of teledermatology. NR-IQA can be customized to target specific types of distortions or created for general-purpose quality assessment, making it adaptable for different fields.

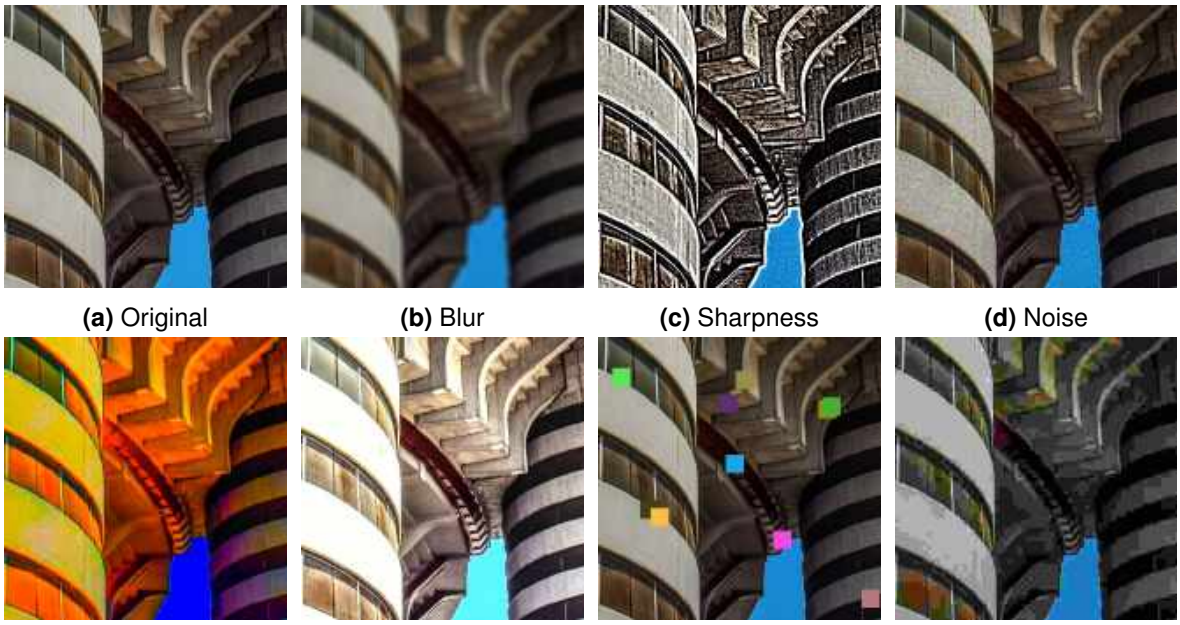
Figure 2.3: General framework of NR-IQA algorithms.

For this thesis, the focus will be on no-reference image quality assessment because it is especially relevant for evaluating teledermatology images where reference images are usually not available. Since IQA measures distortions and NR-IQA can handle various types, it is important to identify the most common distortions encountered. The next subsection will discuss these distortions in detail.

2.1.2 Common Distortions in Image Quality Assessment

IQA must address various distortions that can significantly affect the perceived quality of images. Understanding these common distortions is crucial for developing effective IQA algorithms, particularly in contexts like teledermatology, where accurate image assessment is critical. Figure 2.4 shows the common distortions typically considered in IQA, with a reference image shown first for better comparison (Agnolucci et al., 2023).

The common distortions are:



(a) Original **(b)** Blur **(c)** Sharpness **(d)** Noise

(e) Color Accuracy **(f)** Brightness & Contrast **(g)** Artifacts **(h)** Compression

Figure 2.4: Examples of Common Distortions in Images. (adapted from (Agnolucci et al., 2023))

1. **Blur**: Blurred images lack sharpness and clarity, often resulting from motion during capturing, incorrect focus settings, or imperfections in the camera lens. (See Figure 2.4b for an example of a blurred image.)

2. **Sharpness**: Sharpness refers to how well-defined the edges and fine details in an image appear. High sharpness indicates clear, crisp images, while low sharpness makes an image look soft and unclear. (See Figure 2.4c for an example of a sharpened image.)

3. **Noise**: Noise appears as random variations in brightness or color and is often due to the limitations of the cameras sensor, particularly under low light conditions or at high ISO settings. (See Figure 2.4d for an example of a noisy image.)

4. **Color Accuracy**: Color accuracy refers to how faithfully colors are reproduced in an image. Distortions in color accuracy can lead to inaccurate or unrealistic color representation. (See Figure 2.4e for an example of a color distorted image.)

5. **Brightness & Contrast**: Brightness is the overall light level of an image, while contrast refers to the range between its darkest and lightest areas. Proper balance of both is crucial for maintaining image visibility and detail. Excessive or insufficient brightness and contrast can make an image unusable for detailed analysis. (See Figure 2.4f for an example of an image with altered brightness.)

6. **Artifacts**: Artifacts are unwanted visual anomalies introduced during image acquisition or processing, such as halos, or jagged edges. (See Figure 2.4g for an example of an image with artifacts.)

7. **Compression**: When images are compressed to reduce file size, this often results in lost detail and visible quality degradation. (See Figure 2.4h for an example of a compressed image.)

Each type of distortion affects the visual quality and perceived accuracy of images, influencing the effectiveness of IQA methodologies in assessing image quality. Understanding these distortions is essential for developing robust quality assessment algorithms and improving image clarity in different applications, including teledermatology.

2.1.3 Benchmark Datasets for Image Quality Assessment

Benchmark datasets play an important role in advancing IQA. They provide standardized and diverse image sets with known distortions and corresponding quality annotations, which researchers use to evaluate and improve IQA algorithms. These annotations, often in the form of Mean Opinion Score (MOS) and Differential Mean Opinion Score (DMOS), are used to assess image quality (Zhou & Alan, 2007).

Mean Opinion Score (MOS) is calculated by averaging ratings from human observers who judge the quality of images on a predefined scale. This score reflects the overall perceptual quality as seen by typical viewers and is widely used to compare the performance of different IQA methods against human visual judgment.

Differential Mean Opinion Score (DMOS), on the other hand, is derived from MOS and measures the perceived difference in quality between a reference image and a distorted version. This score is particularly useful for understanding the impact of specific distortions on image quality.

An overview of IQA databases is provided in Table 2.1, and more detailed descriptions can be found in Section A.1. These datasets enable researchers to thoroughly test the robustness, accuracy, and generalization capabilities of different IQA methods. They also help in developing new algorithms by providing reliable quality scores, which are essential for ensuring reproducibility and consistency in research.

2.1.4 State-of-the-Art in Image Quality Assessment

The current state-of-the-art in IQA is ARNIQA (Agnolucci et al., 2023), with version 2 released in late 2023. ARNIQA, which stands for leArning distoRtion maNifold for Image Quality Assessment, represents a significant advancement in No-Reference Image Quality Assessment (NR-IQA). This technology aims to measure image quality based on human perception, even without a reference image.

Table 2.1: An overview of IQA databases

Category	Database	Year	#Ref.	#Dist.	#Dist. Type	#Dist. Level	Resolution Type	Ground-truth
General	LIVE	2004	30	779	JPEG, JP2K, WN, GB, FF	5 or 4	768 × 512	DMOS
	TID2008	2008	25	1700	17 [a]	4	512 × 384	MOS
	TID2013	2013	25	3000	24 [b]	5	512 × 384	MOS
	CSIQ	2009	30	866	JPEG, JP2K, WN, GB, APGN, GCD	5 or 4	512 × 512	DMOS
	A57	2007	3	54	DWT, AGWN, JPEG, JP2K, JP2K-DCQ, GB	3	512 × 512	MOS
	WED	2017	4744	94880	JPEG, JP2K, GB, WN	5	-	-
	KADID-10k	2019	81	10125	25 [c]	5	512 × 384	DMOS
	KADIS-700k	2020	140000	700000	25 [d]	5	512 × 384	DMOS
Multiple Dist.	LIVEMD	2012	15	405	GB followed by JPEG, GB followed by WN	-	1280 × 720	DMOS
	MDID2013	2013	12	324	corrupted successively by GB, WN, and JPEG	-	768 × 512 or 1280 × 720	DMOS
	MDID2016	2016	20	1600	GB or CC rst, JPEG or JP2K second and WN last	-	512 × 384	MOS
Screen content	SIQAD	2014	20	980	WN, GB, CC, JPEG, JP2K, MB, LSBC	7	700 × 700	DMOS
	SCIQ	2017	40	1800	WN, GB, MB, CC, JPEG, JP2K, CSC, CQD	5	1280 × 720	MOS
	CCT	2017	72	1320	HEVC and HEVC-SCC coding	11	1280 × 720 to 1920 × 1080	MOS
	HSNID	2019	20	600	WN, GB, MB, CC, JPEG, JP2K	5	-	MOS
Authentic Dist.	LIVE Wild	2016	0	1162	-	-	500 × 500	MOS
	CID2013	2015	0	480	-	-	1600 × 1200	MOS

Note: #Ref.: Total number of pristine images. #Dist.: Total number of distorted images. AGWN: Additive Gaussian white noise. WN: White noise. APGN: Additive pink Gaussian noise. CC: Contrast change. CSC: Color saturation change. CQD: Color quantization with dithering. DWT: Quantization of the LH subbands of a 5-level DWT. FF: Simulated fast fading Rayleigh channel. GB: Gaussian blur. MB: Motion blur. GCD: Global contrast decrements. HEVC-SCC: Screen content coding extension of high efficiency video coding. JPEG: JPEG compression. JP2K: JPEG2000 compression. JP2K-DCQ: JPEG-2000 compression with DCQ. LSBC: Layer segmentation based compression.

[a] See detailed types on database page: https://www.ponomarenko.info/tid2008.htm
[b] See detailed types on database page: https://www.ponomarenko.info/tid2013.htm
[c] See detailed types on database page: https://database.mmsp-kn.de/kadid-10k-database.html
[d] See detailed types on database page: https://database.mmsp-kn.de/kadid-10k-database.html

ARNIQA is developed using a self-supervised learning approach, allowing it to learn a comprehensive model of different image distortions. Instead of focusing on the content of the images, ARNIQA focuses on understanding the types and qualities of distortions. This characteristic makes it highly adaptable across different domains where image content can vary significantly.

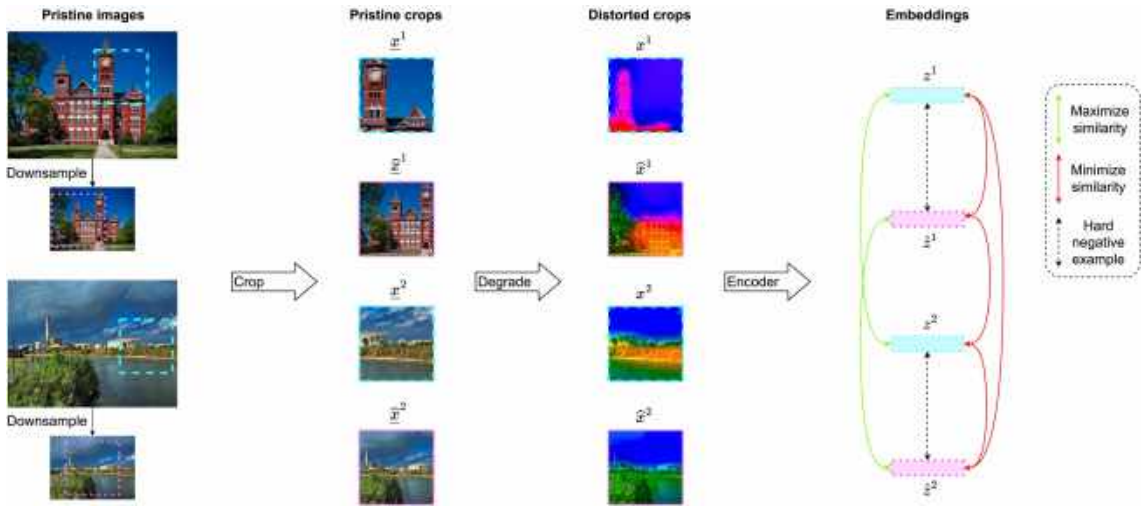


Figure 2.5: Overview of the training strategy for ARNIQA. Two pristine images are cropped and equally degraded. The model maximizes the similarity of their embeddings while minimizing the similarity to embeddings from degraded crops of half-scale versions of the original images. This process creates hard negative examples by introducing downsample distortion, demonstrating how original and half-scale degraded crops differ despite identical degradation. (Agnolucci et al., 2023).

One of the key strengths of ARNIQA is its ability to synthetically degrade images through up to 1.9 billion distinct degradation patterns. It can apply up to seven different types of distortions simultaneously, covering a wide range of real-world scenarios. This extensive training with a wide variety of distortions ensures that ARNIQA can accurately assess image quality in different conditions and avoid the need for large labeled datasets.

At the core of ARNIQA is the SimCLR (Simple Framework for Contrastive Learning) framework. This framework helps ARNIQA learn meaningful representations of image quality by comparing different versions of the same image and focusing on their similarities and differences. By constructing positive pairs through the application of the same distortion settings to different images, SimCLR makes sure the model focuses on the distortions rather than the content. To improve the models ability to distinguish between different types of distortions, SimCLR introduces slight variations by downsampling images before cropping and applying distortions, creating hard negative examples. These examples help the model differentiate between similar-looking images with different types of degradation, thereby improving its ability to provide accurate image quality assessments (see Figure 2.5).

A linear regressor is then used to map the features extracted from the backbone to a quality score ranging from 0 to 1. This score reflects the relative quality of the image based on the distortions.

ARNIQA achieves high performance with only up to 0.5% of the data needed for training compared to other state-of-the-art methods because it focuses on distortion patterns rather than image content. It provides reliable and consistent quality assessments across a wide range of distortions and severities, showing its robustness. This makes ARNIQA particularly suitable for teledermatology, as it can handle different image quality resulting from different lighting conditions, camera quality, and patient handling (Agnolucci et al., 2023).

By leveraging the feature extraction backbone from ARNIQA, this thesis aims to use its advanced capabilities to enhance the assessment of image quality in teledermatology.

2.2 Teledermatology

This section explains what teledermatology is and discusses why having high-quality images is crucial for accurate diagnoses and treatment. It reviews the quality standards needed for teledermatology images, along with public datasets available for research. Additionally, it examines different methods used to assess image quality in teledermatology based on previous studies. In the final section, the challenges and opportunities in the field are explored, focusing on how to improve image quality assessment. This approach helps in understanding the current state of teledermatology and finding ways to enhance it.

2.2.1 Introduction to Teledermatology

Teledermatology is a branch of telemedicine that allows dermatologists to provide remote consultations and treatments using telecommunications technology. This is especially beneficial for patients in remote areas, making sure they receive timely and effective skin care. There are two main methods used in teledermatology: real-time (synchronous) and store-and-forward (asynchronous) (Jiang et al., 2022).

Real-time teledermatology involves live video consultations between the dermatologist and the patient. It allows for immediate interaction and feedback, making it useful for urgent cases. However, it requires both the patient and the dermatologist to be available at the same time, which can be a limitation.

Store-and-forward teledermatology involves sending medical information, including images and patient history, to dermatologists who review it later. This approach offers more flexibility since it doesnt require the patient and dermatologist to be available simultaneously (Jiang et al., 2022).

Given that store-and-forward teledermatology is the focus, it is important to note that high-quality images are essential in this method. The accuracy of remote diagnoses depends directly on the quality of the images provided. Poor image quality can lead to incorrect diagnoses or delayed treatment, thereby reducing the effectiveness of teledermatology. Therefore, ensuring that images meet specific quality standards is crucial for the success of teledermatology services.

2.2.2 Quality Criteria for Teledermatology Images

The International Skin Imaging Collaboration (ISIC) has set guidelines to standardize images based on lighting, background color, field of view for dermoscopic images, image orientation, focus and depth of field, resolution, scale and measurement, color calibration, and image storage. This thesis will focus on seven key criteria from the original nine guidelines that directly impact image quality. The other two criteria, scale and measurement and image storage, are excluded as they do not directly affect image quality. Scale and Measurement are less important here as it involves providing a reference for size within the image, which is not crucial for quality assessment. Image Storage deals more with regulations than with image quality (Finnane et al., 2017).

Throughout this thesis, the term distortion refers to any changes or issues with the seven dermatology quality criteria. This includes problems with lighting, background, field of view, orientation, focus, resolution, and color calibration. These seven key criteria for teledermatology images, along with recommendations on how to meet each criterion, are as follows:

(a) Good Quality **(b)** Lighting **(c)** Background **(d)** Field of View

(e) Orientation **(f)** Focus **(g)** Resolution **(h)** Color Calibration

Figure 2.6: Examples of teledermatology images showing good quality, poor lighting, distracting background, improper field of view, incorrect orientation, lack of focus, low resolution, and poor color calibration.

1. **Lighting**: Good lighting is essential. It should be even and not too harsh, avoiding shadows or bright spots that can hide details. *Use natural light or soft artificial light to clearly show the skin lesion.*

2. **Background**: The background should be plain and uncluttered to keep the focus on the skin issue. A simple, non-reflective background like white or gray works best. *Use a plain, non-reflective background to minimize distractions and keep the focus on the skin lesion.*

3. **Field of View**: The image should include the entire lesion and some surrounding skin. This helps provide context for a more accurate diagnosis. *Make sure the lesion is centered and fully visible in the frame.*

4. **Orientation**: The image should be taken from the correct angle to match standard anatomical positions. This helps the dermatologist understand where the lesion is on the body. *Keep the camera straight and aligned with the lesion.*

5. **Focus & Depth of Field**: The image should be in sharp focus, with the entire lesion clear and detailed. Adjust the camera settings to ensure the lesion is not blurry. *Make sure the camera is in focus and adjust the aperture to achieve sufficient depth of field.*

6. **Resolution**: High resolution is important to show fine details. Use a camera with good resolution settings to capture clear and detailed images of the skin. *Adjust the camera settings to the highest resolution possible to ensure clarity and precision in the image.*

7. **Color Calibration**: Accurate colors are necessary to assess the skin lesion correctly. Make sure the colors in the image match real life. *Use a color reference chart or adjust the white balance settings on the camera to ensure accurate color reproduction.*

By following these guidelines, patients can make sure their images are of high quality, leading to better diagnoses and patient care.

2.2.3 Teledermatology Datasets

In teledermatology, having high-quality image datasets is crucial for developing and testing methods to assess image quality. While many datasets exist for dermatology, they are not always designed specifically for teledermatology. Dermatology datasets often include professional images taken in clinical settings, such as close-up dermoscopic images that provide detailed views of the skin. In contrast, teledermatology images might be taken by patients using their mobile devices, resulting in more inconsistent quality. However, several datasets can still be useful for teledermatology, depending on the specific use case. Here are four public datasets that can be used for teledermatology:

- **DDI** (Daneshjou et al., 2022): This dataset provides 656 high-quality images curated by dermatologists for detailed skin tone evaluation and diagnostic accuracy.
- **Fitzpatrick17k** (Groh et al., 2021): This dataset includes 16'577 images annotated for Fitzpatrick skin type (Thomas, 1998) across 114 different skin conditions. It is valuable for studying a wide range of skin conditions and their presentations across different skin types.
- **Monkeypox Dataset 2022** (Ahsan et al., 2022): This dataset contains approximately 1'905 images focused on monkeypox, useful for developing diagnostic tools.
- **SCIN** (Ward et al., 2024): The Skin Condition Image Network emerges from a crowdsourcing initiative. This dataset contains 10'408 images capturing a broad spectrum of dermatological conditions. Unlike many dermatology datasets that mainly focus on skin cancer diagnostics by classifying malignant and benign tumors, the SCIN dataset covers a broader range of common dermatological conditions, including allergic, inflammatory, and infectious diseases. These conditions are frequently encountered in everyday clinical practice but are underrepresented in existing datasets. The SCIN dataset is particularly valuable because it captures images of early-stage conditions. Over half of the images were taken less than a week from the onset of symptoms, with 30% captured less than a day after symptoms appeared. This dataset includes conditions that patients are likely to consult about via teledermatology platforms before visiting traditional healthcare settings.

For this thesis, the Fitzpatrick17k and SCIN datasets will be particularly important. These datasets provide extensive and diverse image collections that will be used to develop and test image quality assessment methods in teledermatology. The Fitzpatrick17k dataset contains more clinical setting images, which provide good quality but do not represent the variability seen in typical teledermatology images (Groh et al., 2021). Therefore, the Fitzpatrick17k dataset will be used primarily for training purposes to complement the SCIN dataset, which better represents the real-world scenarios encountered in teledermatology.

2.2.4 Related Work on Image Quality Assessment in Teledermatology

In teledermatology, two key methods for detecting image quality have been highlighted in previous studies: TrueImage (Vodrahalli et al., 2020) and ImageQX (Jalaboi et al., 2023). Both methods work closely with dermatologists to ensure their models understand what is needed for accurate diagnoses.

TrueImage (A Machine Learning Algorithm to Improve the Quality of Telehealth Photos), introduced in 2020, uses an automated machine learning system to detect poor-quality dermatology images and help patients take better images. TrueImage uses a semantic segmentation algorithm to identify skin areas, then generates features and classifies the quality. It focuses on common issues like blur, poor lighting, and zoom problems. TrueImage is efficient enough to run on older smartphones and is easy to understand, making it reliable across different skin tones. It was trained on a diverse dataset, including images from Google Images and Stanford Health Care. However, it has limitations: it cannot handle cases where only the background is blurry or poorly lit, it cannot detect framing issues (problems with how the image is arranged, such as when the skin area is not centered or properly aligned), and it cannot discard images that do not contain skin (Vodrahalli et al., 2020).

ImageQX (Explainable Image Quality Assessments in Teledermatological Photography), released in January 2023, is a convolutional neural network that automatically assesses the quality of dermatology images. It focuses on issues like bad framing, poor lighting, blur, low resolution, and distance problems. ImageQX was trained on 26'635 images and validated on 9'c874 images, each annotated by up to 12 board-certified dermatologists. Its main innovation is providing explanations for poor quality and guiding patients on how to take better images. ImageQX is also lightweight, only 15 MB, and can be easily used on mobile devices. It achieves a macro F1-score of 0.73, showing its effectiveness in real-world applications. However, it has limitations in handling certain quality issues, like explaining blurry images, and relies heavily on dermatologist-annotated images, highlighting the need for a diverse and high-quality training dataset (Jalaboi et al., 2023).

Both ImageQX and TrueImage make significant contributions to automated image quality assessment in teledermatology. TrueImage primarily focuses on issues like focus, lighting, and field of view, while ImageQX addresses additional factors such as resolution and distance (e.g., if the skin is too far away in the image). Both methods perform well on detecting blur and lighting issues but have room for improvement in handling field of view, resolution, and distance. Specifically, ImageQX has F1-scores of 0.37 for field of view, 0.61 for lighting, 0.70 for focus, 0.52 for resolution, and 0.42 for distance (Jalaboi et al., 2023). From these methods, it is clear that lightweight models, providing actionable feedback to users, and using a diverse training dataset to ensure robustness are important.

2.3 Challenges and Opportunities in Image Quality Assessment for Teledermatology

2.3.1 Challenges

One major challenge in IQA is that evaluating image quality can be very subjective. This means that different people might have different opinions on what makes an image look good or bad. This subjectivity makes it hard to create standard measures that everyone agrees on. In teledermatology, where image quality directly affects medical diagnoses, this becomes a significant issue (Zhou & Alan, 2007).

Another challenge is the variety of problems that can affect image quality. In teledermatology specifically, patients use different devices under different conditions, leading to inconsistencies in image quality. This variability adds another layer of complexity to developing effective IQA methods. Additionally, in many real-world applications, we often do not have high-quality reference images to compare against. This absence makes it difficult to evaluate the quality of images taken by patients. Therefore, developing methods that do not need reference images (No-Reference IQA or NR-IQA) is essential.

2.3.2 Opportunities

Despite these challenges, there are significant opportunities to improve IQA in teledermatology. One promising area is the advancement of self-supervised learning techniques. These methods, like those used in ARNIQA (Agnolucci et al., 2023), allow models to learn from large amounts of data without needing extensive labeled examples. This approach saves time and money because it reduces the need for manually labeled data. It also enables the development of high-quality IQA models that can work well even without reference images.

Another opportunity lies in collaborating closely with dermatologists. Methods like ImageQX (Jalaboi et al., 2023) and TrueImage (Vodrahalli et al., 2020) have shown the benefits of such collaboration, ensuring that IQA models meet the specific needs of medical professionals. These models can provide real-time, useful guidance to patients on how to take better images, thereby improving the quality of images submitted for remote consultations.

Chapter 3

Methodology

This chapter builds on the findings from the *Literature Review* in Chapter 2, outlining the key ideas and concepts needed to achieve the research objectives. The specifics of implementing these methodologies will be discussed in detail in the next chapter.

3.1 Explorative Approach

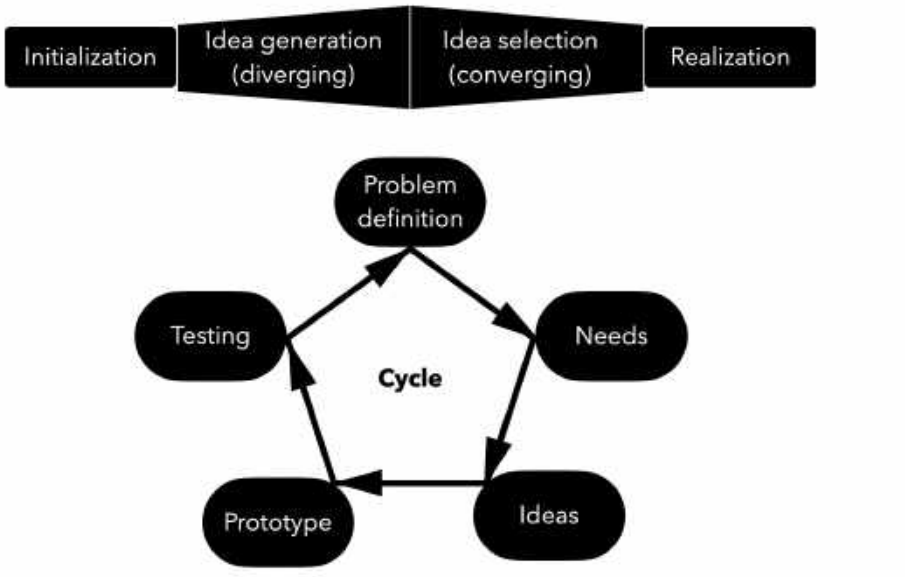


Figure 3.1: Visualization of the explorative approach, including the stages of initialization, idea generation, idea selection, and implementation. The lower part of the figure shows the decision cycles adapted from (Hoffmann et al., 2016).

IQA, particularly in the context of teledermatology, presents many opportunities for innovation because of the diverse types of image distortions and the different methods available to address them. To navigate this complexity, this research uses an exploratory approach. This approach is flexible and allows for changes as new information becomes available. This is different from more traditional methods like the waterfall model, which follows a strict, step-by-step process.

At the beginning, the problem was defined in broad terms to allow flexibility and adaptability. As the research progressed, the approach involved creatively solving problems and refining methods through multiple cycles of learning and improvement. The project was structured into two main phases:

- **Idea Generation Phase**: During this phase, the scope of the research questions was broadened to continuously generate new ideas. This expansion was driven by insights gathered from the ongoing literature review.
- **Idea Refinement Phase**: This phase focused on combining and refining these ideas into clear findings and conclusions, aiming to develop a comprehensive understanding of the initial problem.

This exploratory model is shown in Figure 3.1, which shows the stages of starting the project, generating ideas, selecting the best ideas, and implementing them. It also shows the decision cycles that guide the project (Hoffmann et al., 2016). These cycles help in adjusting the research direction based on new findings and insights, making sure that the methodology stays responsive to new data and trends.

3.2 Project Control

Even with an exploratory approach, it is important to have a rough timeline to guide the research tasks. A workflow was set up before starting, and a detailed Gantt chart is attached to this thesis to show the timeline.

For the first half of the research, three key milestones were set, each playing a critical role for its success:

Understanding Teledermatology: Gaining a thorough understanding of teledermatology is necessary. This makes sure that all subsequent actions are relevant and well-informed.

State of the Art in IQA: Identifying the latest developments in IQA is very important. This helps makes sure that the methods used are up-to-date and effective.

Availability of Teledermatology Images: Securing access to appropriate datasets is necessary for conducting meaningful IQA. Having the right images is important for testing and validating the research methods.

These milestones are important because each phase of the project depends on the successful completion of the previous one. Missing any of these milestones could significantly impact the project and might require a fundamental reassessment of the objectives outlined in Section 1.2.

3.3 Research Steps

As mentioned, this research was exploratory, so it was not possible to follow a strict, step-by-step process. However, a systematic approach was taken for the key steps to stay organized and ensure each step was done in the right order.

3.3.1 Literature Review

The first step was to gain a comprehensive understanding of the research field. Teledermatology and dermatology can be complex areas, so it was essential to build a strong foundation of knowledge. This was done by extensively researching and reading relevant literature to develop a solid understanding of these fields. Additionally, the same approach was taken for understanding IQA. The goal was to learn about the main topics related to the research objectives.

To find relevant literature, several databases were selected, including PubMed[1], Google Scholar[2], IEEE Xplore[3], Connected Papers[4], and Papers with Code[5]. Search filters were applied to narrow down the results, such as limiting the search to articles published after 2020. Special attention was given to state-of-the-art methods, especially those that had published their code and model details. This helped in accessing practical resources and the latest research.

This systematic approach provided a thorough literature review, focusing on the most relevant and up-to-date studies.

3.3.2 Data Collection and Preparation

In the general image domain, IQA commonly uses labels like Mean Opinion Score (MOS) or Differential Mean Opinion Score (DMOS) to train models. However, in the medical field, labeling images with these scores is resource-intensive, so dermatology datasets typically do not have these score labels.

To address this gap, a distortion pipeline was created to synthetically distort images based on the seven dermatology quality criteria defined in Subsection 2.2.2. Each type of distortion has five levels of severity, indicating how poor the image quality is. These distortions are carefully selected to simulate real-world imperfections commonly encountered in teledermatology. Each image is then labeled with seven values corresponding to the severity and type of distortion applied, creating a dataset that includes both the distorted images and precise annotations regarding their quality.

To start with, good quality images were needed. Two datasets were chosen for this purpose: the SCIN (Ward et al., 2024) dataset for its relevance and uniqueness, and the Fitzpatrick17k (Groh et al., 2021) dataset to complement the SCIN dataset. Filtered good quality images from these datasets were passed through the distortion pipeline, creating different versions of distortions for each image. This approach means that the original number of images can be expanded significantly through synthetic distortion, allowing for a more robust training and evaluation process.

[1] https://pubmed.ncbi.nlm.nih.gov
[2] https://scholar.google.com
[3] https://ieeexplore.ieee.org/Xplore/home.jsp
[4] https://www.connectedpapers.com
[5] https://paperswithcode.com

In total, 475 good quality images were filtered from the Fitzpatrick17k (Groh et al., 2021) dataset and another 475 good quality images from the SCIN (Ward et al., 2024) dataset for training and evaluation. Additionally, 200 test images were selected from the SCIN dataset and 70 independent good quality images from SCIN dataset for testing. The 70 good quality images were also fed through the distortion pipeline to introduce synthetic distortions, providing a consistent basis to test the model against the same types of distortions. Furthermore, the 200 test images were labeled, scoring each one on the seven dermatology quality criteria to allow the models performance to be also compared to human evaluation.

Table 3.1: Summary of the datasets used in the research. Note that the Fitzpatrick17k dataset is referred to as F17K for simplicity.

Dataset	Description
$SCIN_{good}$	475 good quality images filtered from the SCIN dataset.
$F17K_{good}$	475 good quality images filtered from the Fitzpatrick17k dataset.
$SCIN_{distorted}$	Synthetic distortions applied to $SCIN_{good}$.
$F17K_{distorted}$	Synthetic distortions applied to $F17K_{good}$.
$COMB_{distorted}$	Combined synthetic distortions.
$SCIN_{authentic}$	200 test images labeled with human evaluation scores.
$SCIN_{synthetic}$	70 good quality test images, synthetically distorted.

3.3.3 Feature Extraction

Feature extraction is the next important step where the state-of-the-art approach from ARNIQA (Agnolucci et al., 2023) is used to identify key features from the synthetically distorted images. ARNIQA is chosen because it has been trained on many different types of image distortions. This training allows ARNIQA to recognize and understand different distortion patterns effectively. By using ARNIQA, this knowledge can be applied to teledermatology images, improving the ability to assess image quality.

The features extracted by ARNIQA capture the patterns of distortions that affect image quality. These features, along with the generated labels from the synthetic distortion pipeline, are then used to train different models, including Extreme Gradient Boosting (XGBoost) regressor, XGBoost classifier, and Multi-Layer Perceptron (MLP) regressor and MLP classifier. By training and comparing these models, the most effective approach for assessing image quality in teledermatology can be identified.

3.3.4 Training and Validation

The training of the models is based on the prepared training images. Because labels and distorted images are generated, there is no restriction on the original number of images. The images can be passed through the distortion pipeline multiple times, creating different versions of distortions from the original images. This allows for a larger and more diverse training set.

The models are then trained using these distorted images to develop their ability to assess image quality. Validation is done alongside training by setting aside a portion of the data as a validation set. This validation set helps to evaluate and monitor the performance of the models, ensuring they are learning correctly and adjusting as needed.

3.3.5 Evaluation Metrics

The evaluation of the models is done using defined metrics such as Mean Absolute Error (MAE), R-squared (R^2), Spearmans Rank Order Correlation Coefficient (SRCC), and Cohens Kappa. These metrics help understand the strengths and weaknesses of the models and guide further improvements or adjustments.

Understanding the Metrics and Their Importance

MAE measures the average difference between the predicted image quality scores and the actual scores. It helps in understanding how accurate the models predictions are on average. A lower MAE indicates better model performance, meaning the predictions are closer to the actual values.

R^2 indicates how well the predicted scores match the actual data. It tells us how much of the variance in the actual scores is explained by the models predictions. A higher R^2 means better model performance, showing that the models predictions fit the actual data well. Using MAE and R^2 together provides a clear picture of the models accuracy and how well it fits the actual data.

SRCC measures the strength and direction of the association between two ranked variables. In simpler terms, it evaluates how well the predicted rankings of image quality match the actual rankings. For example, if the model predicts the severity of distortions in the same order as the actual severity, it will have a high SRCC. SRCC is calculated as:

$$SRCC = 1 - \frac{6 \sum_{i=1}^{n} (d_i^2)}{n(n^2 - 1)} \tag{3.1}$$

where,
n : Number of images
d_i : Difference in ranks between predicted and actual scores for image i

An SRCC of 1 means perfect rank correlation, and -1 means perfect negative correlation. This metric is crucial because, in many cases, getting the rank order correct is more important than predicting the exact value. If images are ranked correctly in terms of severity, even if the predicted values are not exact, the model can still be useful in prioritizing cases for further review.

Cohens Kappa measures how well the models predictions agree with the actual labels. Unlike SRCC, which focuses on ranking, Cohens Kappa evaluates the exact agreement between predictions and actual labels. Unlike simple accuracy, which only looks at the proportion of correct predictions, Cohens Kappa accounts for the possibility that some agreement might occur by chance. It is calculated as:

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.2}$$

where,
p_o : Observed agreement
p_e : Expected agreement

Cohens Kappa ranges from -1 to 1, with 1 indicating perfect agreement and 0 indicating agreement by chance.

3.3.6 Testing and Evaluation

After completing the training, the models are evaluated using independent test images. There are two test sets used in this evaluation. The first test set consists of images that have been synthetically distorted to simulate common types of distortions. The purpose of this test set is to assess the performance and reliability of the models when faced with consistent types of distortions. By using this controlled environment, its possible to see how well the models handle different distortion levels.

The second test set includes real-world images with authentic distortions. The performance of the models on this test set is compared to human evaluations. This comparison helps to understand how well the models assessments align with human judgment, providing a more realistic measure of their effectiveness. The results from these test sets are then compared using radar charts to visualize the models strengths and weaknesses across different quality criteria.

3.3.7 Model Comparison

To compare the performance of the image quality assessment model for teledermatology, a baseline comparison was done against two methods: Structural Similarity Index Measure (SSIM) and ARNIQA (Agnolucci et al., 2023). This helps to show how effective the proposed approach is and where it can be improved.

SSIM is a widely-used method that compares the similarity between two images by measuring changes in structure, brightness, and contrast. It can only be used when both the original image and its distorted version are available. ARNIQA, on the other hand, is a pre-trained No-Reference Image Quality Assessment (NR-IQA) method that gives quality scores without needing the original image.

The comparison involves plotting the single quality scores from the proposed model, SSIM, and ARNIQA, as well as calculating the Spearmans Rank Correlation Coefficient (SRCC) to see how well the predicted scores match the true quality scores. This comparison also helps to understand where the proposed model does better or needs improvement compared to traditional and state-of-the-art methods.

In addition to these comparisons, an out-of-distribution test was done. Images that are different from the training data, taken from the KADID10K (Lin et al., 2019) dataset, were used for this test. These images include nature scenes, vehicles, animals, and everyday objects. The models predictions on these unrelated images were checked to see if the model produces reasonable scores or shows large differences, which could mean the model is too specialized to the teledermatology domain.

3.3.8 Discussion and Further Development

In conclusion, the results of the project are analyzed and discussed. This discussion includes an evaluation of the achieved goals, an analysis of the challenges and limitations of the project, and a look at possible further developments.

Chapter 4

Implementation

This chapter explains the detailed implementation of the methods described in Chapter 3. It covers the specific processes, experiments, and analyses done during the research. This includes the practical steps taken to prepare images, apply distortions, extract features, and train the models to assess image quality in teledermatology.

4.1 Image Selection and Labeling Process

This section describes the first stages of the implementation, focusing on the selection and preparation of the image datasets used in the research. The images considered for this research were in JPG and PNG formats.

4.1.1 Image Filtering and Selection

The first step in preparing the images involves carefully selecting good quality images from the SCIN (Ward et al., 2024) and Fitzpatrick17k datasets (Groh et al., 2021). This manual selection process makes sure that no mistakes are made and only clear images are included in the training set. The focus during selection is on images that are well-framed and free from distortions that could affect their usefulness for making accurate medical diagnoses.

Key criteria for selecting good quality images include clarity, proper lighting, and accurate color representation. Clear images are important for accurate diagnosis, as they allow for detailed examination of the skin condition. A good indication of clarity is the visibility of fine details like hair follicles, which shows that the image is not blurry. Proper lighting is also necessary, as it helps in accurately displaying the skins condition. Images should neither be too bright nor too dark, as this can hide important details. Balanced contrast, where the light and dark areas of the image are evenly distributed, is important for a clear view of the skins texture and color. Accurate color representation is another critical factor. The images must realistically represent skin tones and colors, as this is needed for correct diagnosis. Any change in color can lead to misinterpretation of the skin condition.

By carefully filtering and selecting images based on these criteria, the dataset is prepared with good quality images that are appropriate for further analysis and training in the context of teledermatology.

4.1.2 Labeling of the Test Set

The labeling process involves manually scoring 200 images from the SCIN (Ward et al., 2024) dataset. Among these 200 images, around 50 are of good quality and are also included in the test set. Each image is scored on a scale from 0 to 1 for each dermatology quality criterion, where 0 indicates no distortion and 1 indicates extreme distortion. A custom Python script[1] is used for this labeling process. The script displays each image and prompts the user to enter scores for each quality criterion. The collected scores are organized in a structured format and stored in a JSON file for later analysis. This method ensures that each image is consistently and thoroughly evaluated.

The labeling is done using an absolute categorical rating method. This method is time-consuming and requires significant effort from the evaluator. Each of the 200 images is scored on 7 different criteria, resulting in a total of 1400 labels. To maintain accuracy and avoid fatigue, the labeling process is spread out over multiple sessions. This careful and methodical approach helps to ensure that the scores are reliable and that each image is evaluated fairly.

4.2 Distortion Types

Here, each dermatology quality criteria type is briefly described, highlighting how they simulate different aspects of image degradation:

1. Lighting:
    - *Brighten*: This operation increases the brightness of an image by applying color space transformations and adjustments, enhancing the overall visual intensity.
    - *Darken*: Similar to the brighten operation but reduces the visual intensity, making the image darker.
2. Focus:
    - *Gaussian blur*: Applies a Gaussian kernel to create a blurred effect, which softens the image by averaging the pixel values.
    - *Lens blur*: Uses a circular kernel to simulate the effect of a camera lens blur, causing a more uniform blur across the image.
    - *Motion blur*: Simulates the effect of motion, either from the camera or the subject, by applying a linear blur in a specified direction.
3. Orientation:
    - *Top perspective*: Alters the image to appear as if viewed from a higher angle, distorting the top part of the image.
    - *Bottom perspective*: Alters the image to appear as if viewed from a lower angle, distorting the bottom part of the image.
    - *Left perspective*: Alters the image to appear as if viewed from the left side, distorting the left part of the image.
    - *Right perspective*: Alters the image to appear as if viewed from the right side, distorting the right part of the image.

[1] playground/create_labels.ipynb

4. Color calibration:
    - *Color saturation 1*: Adjusts the saturation in the HSV color space, either increasing or decreasing the vividness of the colors.
    - *Color saturation 2*: Modifies the color channels in the LAB color space to change the saturation levels, affecting the color intensity.
5. Background:
    - *Color Block*: Uses skin segmentation to apply color block artifacts in the background, simulating background distortions and maintaining focus on the skin area.
6. Resolution:
    - *Change Resolution*: Alters the image resolution to simulate low-quality images by downsampling and then upsampling the image.
7. Field of view:
    - *Crop Image*: Crops the image to simulate different levels of field of view, reducing the visible area of the image.

The distortions for Lighting, Focus, and Color Calibration were adapted from the ARNIQA (Agnolucci et al., 2023) image degradation model, which was inspired by the KADID (Lin et al., 2019) dataset. These distortions originally provided an extensive range of severity levels. The severity levels were modified to better fit real-world distortions commonly encountered in teledermatology. The rest of the distortions were designed based on observations of real-world image quality issues in teledermatology.

For the orientation distortion, the perspective of the image is changed to simulate different viewing angles. By tilting, the image appears as if viewed from a higher, lower, left, or right angle, giving the effect that the camera is not perpendicular to the skin. The resolution distortion is done by first downsampling the image to a lower resolution and then upsampling it back to its original size. This process introduces pixelation and a loss of detail, simulating the effect of low-quality images when enlarged. The field of view distortion involves cropping the image from the left corner to reduce the visible area. Normally, in good quality images, the skin lesion is centered. By cropping the left corner, the lesion moves to the bottom right, simulating poor framing or incomplete capture of the lesion area. Lastly, the background distortion involves segmenting the skin from the background and adding color blocks to create a noisy and cluttered background. This simulates real-world situations where the background is not clean, causing issues in image quality.

4.3 Distortion Implementation Process

The distortion implementation process for each dermatology quality criteria involves several key steps to create a range of distorted images, which helps train and evaluate the IQA model.

For each image, the RGB version is used, and a downsampled version of the image at half the resolution is created. This involves resizing the image to half its original dimensions to simulate lower resolution, following the approach used by ARNIQAs (Agnolucci et al., 2023) feature extraction backbone. Distortions are then applied in a specific sequence, as shown in Figure 4.1(b). The order matters only for the background distortion, which has to be applied first because it depends on identifying the skin area in the undistorted image. For all other distortions, the order does not matter. If the images have less than 10% background in proportion

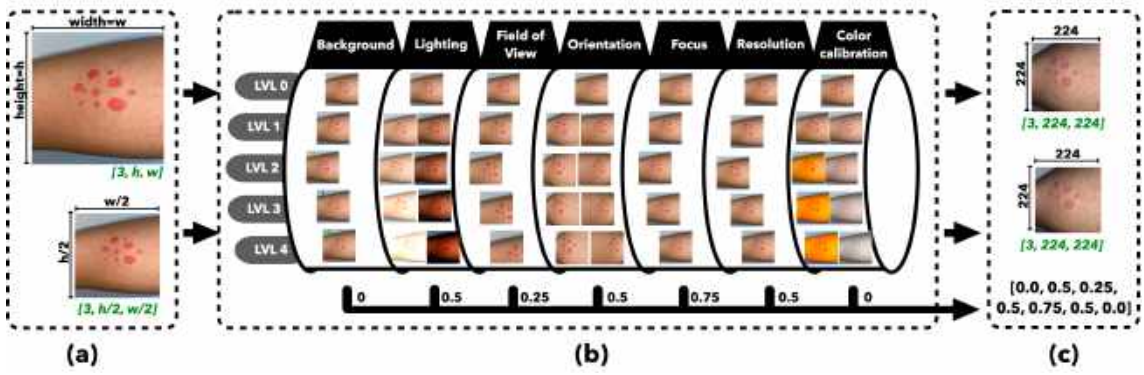


Figure 4.1: Distortion pipeline for generating training images with different levels of distortion. (a) shows the original image and its downscaled version. (b) shows the distortion pipeline where a type of distortion and a random level for each criterion are selected, with the corresponding mapped values shown at the bottom. (c) shows the output where the distorted original image and the distorted downscaled image are resized to 224x224 pixels, along with the seven distortion values for the image.

to skin, no color blocks are added in the background, and a range value of 0 is used. After that, other distortions are applied based on randomly chosen severity levels. This creates a range of distortion levels across the dataset.

Once the distortions are applied, both the original and downsampled images are resized to 224x224 pixels to match the requirements for the backbone. Following resizing, both images are normalized using the mean and standard deviation values of the ImageNet dataset (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]). This normalization ensures the images are processed consistently, as the feature extraction backbone expects the images to have these properties.

The severity of each applied distortion is mapped to a value between 0 and 1. This is done because the ranges are, for some distortions, integers and for others, floats in different ranges. Standardizing them, by taking the minimum and maximum possible values of the distortion and scaling the actual distortion value within this range, ensures consistency and maintains the correct ratio of severity across different types of distortions.

This process can generate multiple combinations of distorted images because of the random selection of distortion types and severity levels. This significantly increases the number of training images by factors of 4, 8, 16, 32, or 64, though one could experiment with even larger factors. By following this detailed and structured approach, the distortion pipeline effectively simulates a wide range of real-world image quality issues in teledermatology, providing a comprehensive dataset for training and evaluating the image quality assessment model.

To make it easier to repeat the experiments, the features and scores from the multiplied images were saved in *.npy* files. These files were named according to the number of distortions, such as *features_num_distortions.npy* and *scores_num_distortions.npy*. Saving these files means that the exact same sets of distorted images and their related scores can be used again for training and testing different models, allowing for consistent comparisons and evaluations across different tests. By saving these files, there is no need to regenerate the distortions each time, making the experimental process simpler and keeping the evaluation consistent and fair.

4.4 Feature Extraction with the ARNIQA Backbone

After creating the distorted images and their half-scaled versions with mapped labels, the next step is to use the pretrained backbone from ARNIQA (Agnolucci et al., 2023), which is loaded via *torch.hub*. This pretrained model has already learned useful features from a large dataset, and these features are transferred to this specific task, a process known as transfer learning. This approach saves time and computational resources.

The backbone from ARNIQA generates feature vectors that represent the distortion patterns in the images. By using both the original and downscaled images, the model effectively learns to distinguish between different levels of distortion. This dual-input method provides a comprehensive understanding of image quality variations.

The extracted features, which have a shape of (*num_images*, 4096), represent the learned characteristics of the images. The target labels, indicating distortion severity, have a shape of (*num_images*, 7) corresponding to the seven distortion criteria. These features and labels are then used to train the final image quality assessment model.

4.5 Model Training

After expanding the dataset with distorted images, the images were divided into training and validation sets. The training set consisted of 75% of the images, while the validation set contained the remaining 25%. This split allowed the models to train on the majority of the data while still having a separate set for evaluating their performance.

Four different multi-output models were experimented with: XGBRegressor, XGBClassifier, MLP Regressor, and MLP Classifier. These models were selected because they can handle complex relationships and predict multiple outputs simultaneously, which is essential for assessing various aspects of image quality. These models were imported from well-established libraries: *xgboost* for XGBRegressor and XGBClassifier, and *sklearn.neural_network* for MLP Regressor and MLP Classifier. Using these pre-built models has several benefits, including access to optimized and tested implementations, extensive documentation, and community support, which saves time and ensures reliability compared to writing custom models from scratch.

Training was conducted using an NVIDIA A16 GPU, equipped with 16GB of vRAM, 1280 CUDA Cores, 40 Tensor Cores, and 512 GB of RAM. This setup made efficient use of resources and sped up the training process.

The training involved using the $\text{SCIN}_{\text{distorted}}$, $\text{F17K}_{\text{distorted}}$, and $\text{COMB}_{\text{distorted}}$ datasets. Each model was trained individually on these datasets. Additionally, cross-dataset evaluations were conducted to test how well the models generalized across different datasets. For instance, a model trained on $\text{SCIN}_{\text{distorted}}$ was evaluated on $\text{F17K}_{\text{distorted}}$ to assess its robustness and ability to generalize. Combining both datasets and evaluating the models on each dataset individually provided further insights into their performance with a more diverse set of images, ensuring the models did not overfit to a particular dataset.

During training, the loss curve was evaluated with early stopping to monitor the models performance and prevent overfitting. Early stopping stops the training process if the models performance on the validation set does not improve after a certain number of iterations, ensuring the model does not learn noise from the training data.

4.5.1 Handling Continuous Scores and Discretization

An important aspect of training the regressors is handling continuous scores. Directly comparing continuous scores to fixed numbers from the distortion pipeline can lead to minor errors. To minimize these errors and accurately calculate metrics like rank correlation and Cohens Kappa, the regressor predictions were clipped to the range of 0 to 1 and then categorized into severity levels using a discretization function[2] that converts continuous scores to discrete categories based on defined thresholds. This process effectively categorizes the severity levels and reduces errors in score comparison.

The function ensures that the continuous scores are mapped correctly to discrete categories, providing a standardized way to handle various severity levels. Additionally, the discretization function was used for the classifier models to convert continuous scores to categorical ones, as these models require categorical labels as input. This approach ensured that the models could effectively predict and categorize the severity of distortions in the images.

4.5.2 Hyperparameter Configuration

To find the best hyperparameters, a hyperparameter sweep was performed using Weights and Biases[3]. This process involved randomly searching for the best hyperparameters to maximize the overall Spearmans Rank Order Correlation Coefficient (SRCC). The Table 4.1 and Table 4.2 shows the configurations used in the hyperparameter sweep.

The hyperparameter sweep aimed to prevent overfitting and help the models perform well on new data. Key parameters were chosen for their ability to control model complexity and improve generalization.

The batch size was set to 10 due to hardware limitations. L2 regularization (reg_lambda) and subsampling (subsample) were used to help the model generalize better. L2 regularization prevents the model from having large coefficients, and subsampling exposes the model to different subsets of data, reducing the chance of overfitting.

Table 4.1: Hyperparameter Configurations for MLP Models

MLP Parameter	**Sweep Values**
model_type	[mlp_reg, mlp_cls]
num_distortions	[4, 16, 64]
hidden_layer_sizes	[(512,), (1024,), (512, 256), (1024, 512), (512, 512)]
alpha	{"min": 0.0001, "max": 0.01}
learning_rate_init	{"min": 0.0001, "max": 0.1}
max_iter	[200, 300, 500]
	Fixed Values
batch_size	10
activation	relu
solver	adam
early_stopping	True

For the XGB models, the min child weight (min_child_weight) parameter was adjusted to make sure each tree leaf had a minimum number of instances, preventing the trees from becoming too complex. The gamma parameter added a constraint on tree growth, which helped to further avoid overfitting by regulating how trees expand during training.

[2] from utils.utils_data import discretization
[3] https://wandb.ai/site

Table 4.2: Hyperparameter Configurations for XGB Models

XGB Parameter	**Sweep Values**
model_type	[xgb_reg, xgb_cls]
num_distortions	[4, 16, 64]
n_estimators	[50, 100, 200, 300]
learning_rate	{"min": 0.0001, "max": 0.1}
min_child_weight	{"min": 1, "max": 150}
early_stopping_rounds	[10, 20, 30, 40]
subsample	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
max_depth	[3, 5, 7, 9]
gamma	{"min": 0.001, "max": 0.5}
multi_strategy	[one_output_per_tree, multi_output_tree]
	Fixed Values
batch_size	10
reg_alpha	0.0
reg_lambda	1.0
tree_method	hist
objective	reg:pseudohubererror (specific to XGBRegressor), multi:softprob (specific to XGBClassifier)
n_jobs	16 (specific to XGBRegressor)
booster	gbtree (specific to XGBClassifier)
eval_metric	['mlogloss', 'merror', 'auc'] (specific to XGBClassifier)

By carefully tuning these parameters, the models were designed to balance complexity and generalization, resulting in robust performance on both the training and validation datasets.

4.5.3 Interpreting Model Performance with Plots

To understand the models performance, several metrics and visual tools were used. Including Mean Absolute Error (MAE), R-squared (R^2), Spearmans Rank Order Correlation Coefficient (SRCC), and Cohens Kappa, for each of the seven dermatology quality criteria. These metrics provided detailed insights into where the model performs well and captures distortions effectively and where it does not. An overall score was also calculated to provide a overview of the models performance.

In addition to numerical metrics, Confusion matrices[4] were used for each criterion to show where the model makes correct predictions and where it makes mistakes. These matrices help visualize the accuracy for each type of distortion. They also reveal any biases the model might have, such as predicting only low or high severity levels, or if its predictions are skewed in some way. This detailed view helps identify areas where the model needs improvement and provides insights into its decision-making process.

[4] from utils.visualization import plot_all_confusion_matrices

4.6 Model Testing

The best model was tested on two specific sets of images to evaluate its performance. The first set, referred to as $SCIN_{synthetic}$, included 70 good quality images that were synthetically distorted using a pipeline to introduce consistent types of distortions. This helped assess how well the model handled controlled distortions. The second set, referred to as $SCIN_{authentic}$, consisted of 200 images with authentic distortions, allowing for a comparison of the models performance with human labels.

For the $SCIN_{authentic}$ images, they were first downscaled to half their size, resized to 224x224 pixels, and normalized. These preprocessed images were then passed through the ARNIQA backbone to extract features, and their actual scores were retrieved from a JSON[5] file where the labels were stored. For the $SCIN_{synthetic}$ images, the features and scores were saved after extraction from the backbone, and these were stored in a *.npy*[6] file to allow reproducibility and easier comparison across different tests.

To visualize the predictions and test scores, radar charts[7] were used to observe the models strengths and weaknesses across different dermatology quality criteria. For $SCIN_{synthetic}$, a four-column layout was used to present the original image, the distorted image, the actual labels, and the models predictions. For $SCIN_{authentic}$, a three-column layout displayed the image, the human-labeled scores, and the models predictions.

4.7 Baseline Comparison

To evaluate the performance of the proposed image quality assessment model, a baseline comparison was done using both traditional and state-of-the-art methods on the two test sets. The synthetic test set included comparisons with SSIM, ARNIQA, the models predictions, and actual scores. The authentic test set included comparisons with ARNIQA, the models predictions, and human-labeled scores. This comparison shows how well the proposed model compares to known image quality assessment techniques.

Before diving into the details of the individual methods, it is important to understand how the predictions and actual scores were converted into a single quality score. This process makes sure of a consistent and fair comparison.

To do this, a weighted average[8] method was used to give more importance to higher distortion values, which have a bigger effect on overall image quality. The function calculates the average of the seven distortion scores by first creating weights based on how severe each distortion is. This means that higher distortion scores will have a bigger impact on the final quality score. The scores are squared to highlight higher distortion values, and these squared scores are then normalized by dividing each by the sum of all the squared scores. This step makes sure that the weights add up to 1, keeping the balance between the different scores.

The function then uses these normalized weights to compute the weighted average of the distortion scores. This results in a single score ranging from 0 to 1, where 0 indicates good quality (no distortion) and 1 indicates poor quality (high distortion). This method makes sure that the final quality score accurately shows the severity of the distortions in the image.

[5] src/test_200/scores.json
[6] src/test_70/embeddings/features.npy
[7] from utils.visualization import plot_results
[8] playground/create_labels.ipynb

4.7.1 Traditional Image Quality Assessment (SSIM)

For the SSIM (Structural Similarity Index) implementation, the $SCIN_{synthetic}$ and $SCIN_{synthetic}$,before distortion, were first converted to RGB format and then resized to 224x224 pixels to match the size used in the model. The SSIM[9] was calculated for each pair of images, and the scores were saved in a CSV file for later comparison.

It is important to note that SSIM values range from -1 to 1, where 1 indicates perfect similarity, 0 indicates no similarity, and -1 indicates perfect anti-correlation. However, since the quality assessment scores in this research work inversely, with higher values indicating worse quality, the SSIM scores were inverted. This adjustment ensures that the SSIM values can be compared directly to the predicted and actual scores.

4.7.2 No-Reference Image Quality Assessment Model (ARNIQA)

For the ARNIQA(Agnolucci et al., 2023) implementation, the *single_image_inference.py* script from the ARNIQA codebase was referenced and adapted[10]. The preprocessing steps were done similarly to the model training, involving downscaling, resizing, and normalizing the images. A key difference in this implementation was the need to specify a regressor. ARNIQA provides six different regressors for image evaluation, and for this comparison, the default *KADID10K* regressor was used. Additionally, the *"scale_score=True"* parameter was set for the model.

Similar to the SSIM adjustment, the ARNIQA scores had to be inverted. This is because ARNIQA predicts quality on a scale where 1 indicates good quality and 0 indicates bad quality, which is opposite to the scoring system used in this research. This inversion allows for a direct comparison of ARNIQA scores with the predicted and actual scores from the model.

Finally, a line plot was created to compare the different scores from the proposed model, SSIM, and ARNIQA to the actual scores. Additionally, the SRCC values were also calculated to see how well the scores matched the actual quality scores.

4.7.3 Out of Distribution Testing

In addition to these comparisons, out-of-distribution testing was done. The script[11] used the same preprocessing and inference steps as in testing the model. This script processes an image and generates a two-column layout, with the image on the left side and a radar chart showing the seven quality criteria for that image on the right side.

Different images were chosen from the KADID10K (Lin et al., 2019) dataset, which are different from teledermatology images. For example, images with known distortions such as brightened images were included to check if the model recognizes the brightening, along with color-calibrated images and some good quality images for comparison. This method shows how the model assesses image quality outside its usual teledermatology context.

This approach allows for comparing the models predictions to see if it identifies distortions correctly. If not, it helps identify limitations, such as difficulties in assessing distortions in general image domains.

[9] src/ssim_inference.py
[10] src/ARNIQA_test.py
[11] src/single_image_inference.py

Chapter 5

Results and Analysis

In this chapter, the main objective is to present the findings from the experiments and analysis conducted in the previous chapters. The section is designed to show the performance of the trained models, especially focusing on the MLP regressor, which was trained on the $COMB_{synthetic}$ images. This chapter will include several tables, figures, and visualizations that show the results without going into detailed explanations. That will be covered in the next chapter.

5.1 Label Distribution

The histograms in Figure 5.1 for $SCIN_{authentic}$ and Figure 5.2 for $SCIN_{synthetic}$ show the spread and severity of distortions in the test datasets. Each histogram has five bins for each dermatology quality criterion. The first bin shows no distortion, and the other bins show increasing levels of distortion severity.

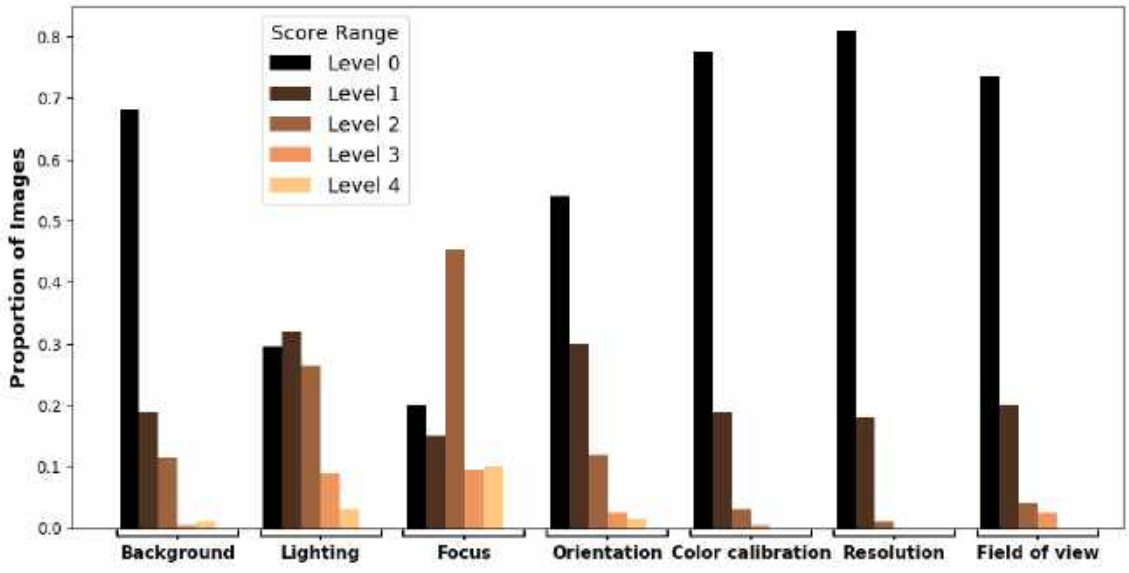


Figure 5.1: Distribution of distortion scores for each dermatology quality criteria in the $SCIN_{authentic}$ test set. The histograms show the proportion of images at different levels of distortion severity, ranging from Level 0 (no distortion) to Level 4 (high distortion).

Figure 5.1 shows the distribution of distortion scores for each dermatology quality criterion in the $SCIN_{authentic}$ test set. Most criteria are right-skewed, meaning higher levels of distortion are less common. However, focus and lighting have a more balanced spread of distortion levels.

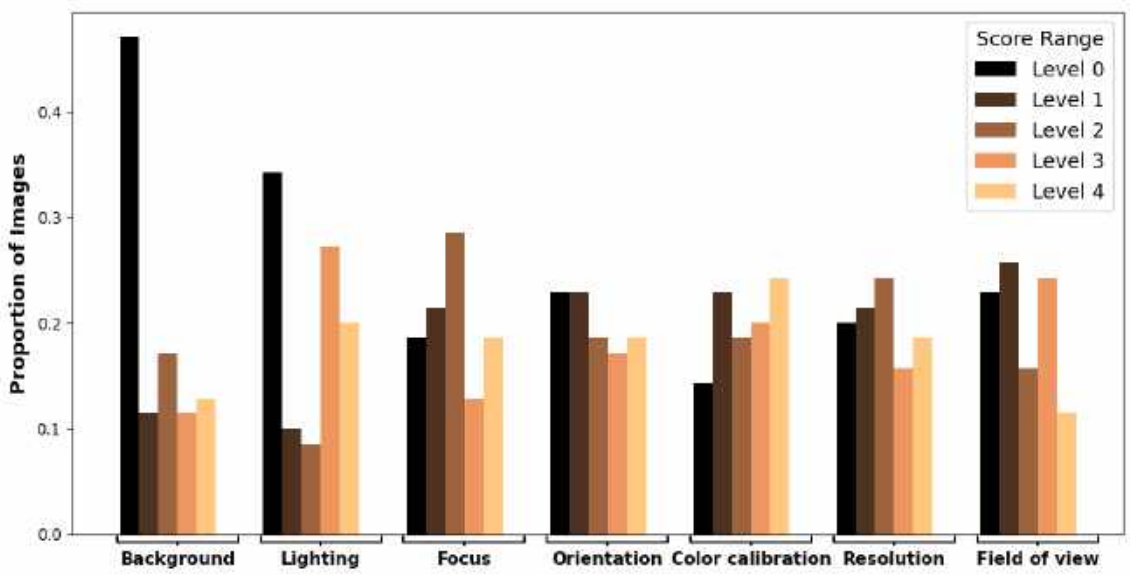


Figure 5.2: Distribution of distortion scores for each dermatology quality criteria in the $SCIN_{synthetic}$ test set. Unlike the $SCIN_{authentic}$ test set, these distributions are more balanced, showing that synthetic distortions were applied evenly across all severity levels.

Figure 5.2 shows the distribution of distortion scores for each quality criterion in the $SCIN_{synthetic}$ test set. These distributions are more balanced, showing that the synthetic distortions were applied evenly across all severity levels.

5.2 Visual Examples of Distortions

The distortion pipeline is key to creating realistic image quality issues in the context of teledermatology. Each dermatology quality criterion includes multiple types of distortions, each with five levels of intensity. The distortions range from zero, meaning no distortion, to higher values that show increasing levels of the specified distortion. Visual examples of these distortions at different levels for each quality criterion are included in Section A.2. These examples help to understand what each type of distortion looks like and how it impacts the images.

5.3 Effect of Distortion Quantity on Performance

The scatter plots in Figure 5.3 and Figure 5.4 show how the overall Spearman Rank Order Correlation Coefficient (SRCC) changes with training time for different numbers of distortions. The points are colored based on the number of distortions in the dataset, helping to see how the size of the dataset affects model performance.

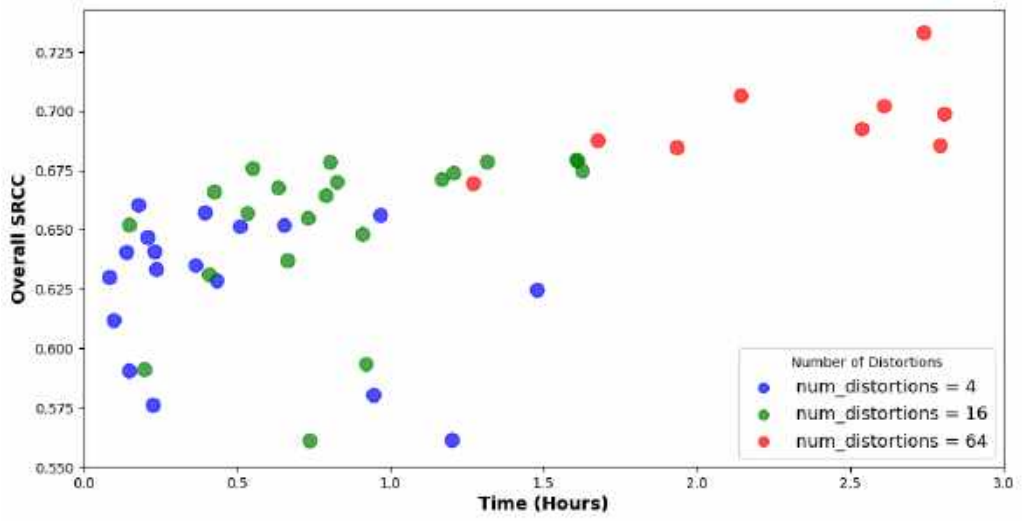


Figure 5.3: Overall SRCC for XGB Regressor and MLP Regressor with different numbers of distortions. The x-axis shows the training time, and the y-axis shows the SRCC values. Larger datasets generally lead to better performance but take more time to train.

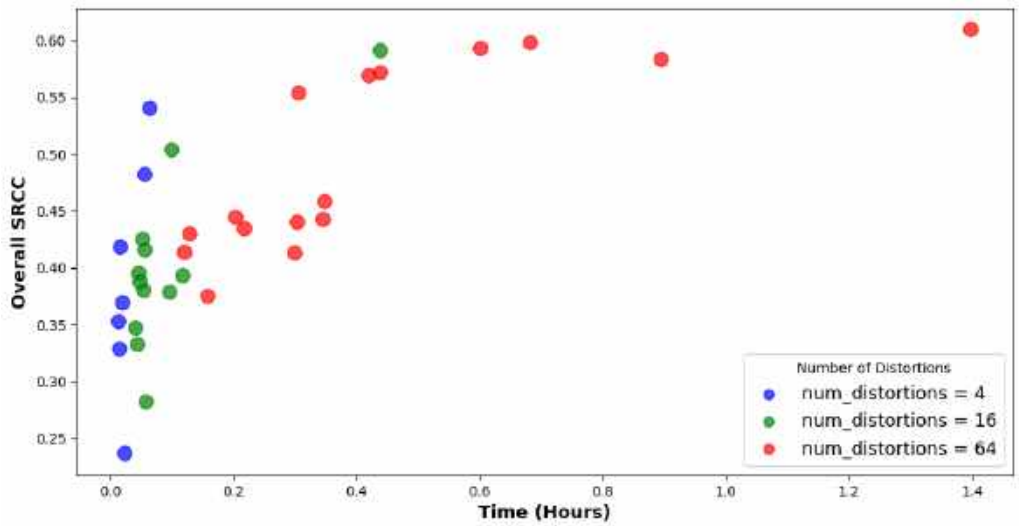


Figure 5.4: Overall SRCC for XGB Classifier and MLP Classifier with different numbers of distortions. Similar to Figure 5.3, this figure shows better performance with larger datasets but longer training times.

5.4 Cross-Dataset Evaluation

The performance of the four different models, trained on $SCIN_{distorted}$, $F17K_{distorted}$, and $COMB_{distorted}$ datasets, was evaluated through cross-dataset testing. This means the models were tested on both the $SCIN_{distorted}$ and $F17K_{distorted}$ individually, as shown in Table 5.1. This table helps to understand how well the models perform on different datasets and shows how well each model can adapt to different datasets, demonstrating their ability to generalize.

Table 5.1: Spearmans Rank Correlation Coefficient (SRCC) of Different Models on $SCIN_{distorted}$ and $F17K_{distorted}$ Datasets.

Model	**$SCIN_{distorted}$**	**$F17K_{distorted}$**
$COMB_{distorted}$ MLP Regressor	<u>0.66</u>	<u>0.75</u>
$COMB_{distorted}$ XGB Regressor	0.65	0.73
$COMB_{distorted}$ XGB Classifier	0.58	0.61
$COMB_{distorted}$ MLP Classifier	0.43	0.46
$F17K_{distorted}$ MLP Regressor	0.54	0.69
$SCIN_{distorted}$ MLP Regressor	0.62	0.49
$F17K_{distorted}$ XGB Regressor	0.53	0.67
$SCIN_{distorted}$ XGB Regressor	0.61	0.48
$SCIN_{distorted}$ MLP Classifier	0.53	0.45
$F17K_{distorted}$ MLP Classifier	0.47	0.58
$SCIN_{distorted}$ XGB Classifier	0.54	0.43
$F17K_{distorted}$ XGB Classifier	0.46	0.59

5.5 Loss Curves

The loss curve in Figure 5.5 shows how the models loss decreases over time for each dermatology quality criterion during training. This graph helps to identify which criteria have higher or lower loss, indicating areas where the model performs well or needs improvement.

This plot shows that as training progresses, the loss for most criteria steadily decreases, indicating that the model is learning and improving. Some criteria such as background, orientation, and field of view have higher loss values, meaning the model finds these criteria more challenging, and there should be more focus on improving performance for these specific distortions. On the other hand, criteria with lower loss, like color calibration and resolution, indicate that the model performs relatively well in these areas. The training process was set to run for a maximum of 500 iterations, with early stopping enabled to avoid overfitting, which stops the training when the models performance does not to improve.

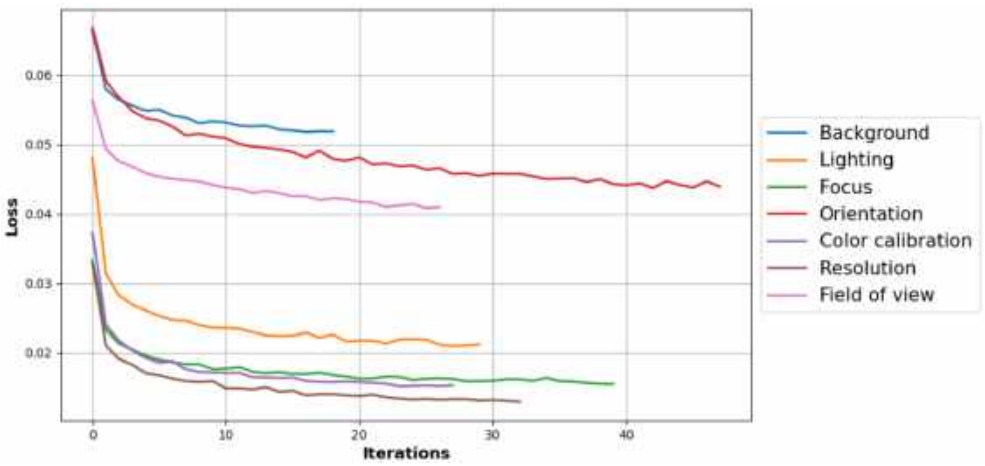


Figure 5.5: Loss curve showing the reduction in loss for each distortion criterion during the training process. Each line represents a different criterion, showing how the models performance improves with each iteration.

5.6 Parallel Coordinate Plot

The parallel coordinate plot in Figure 5.6 compares the best-performing models across seven dermatology quality criteria, including the overall SRCC. This plot helps visualize variations in model performance for each quality criterion and the overall score, making it easier to see which models perform well in different areas.

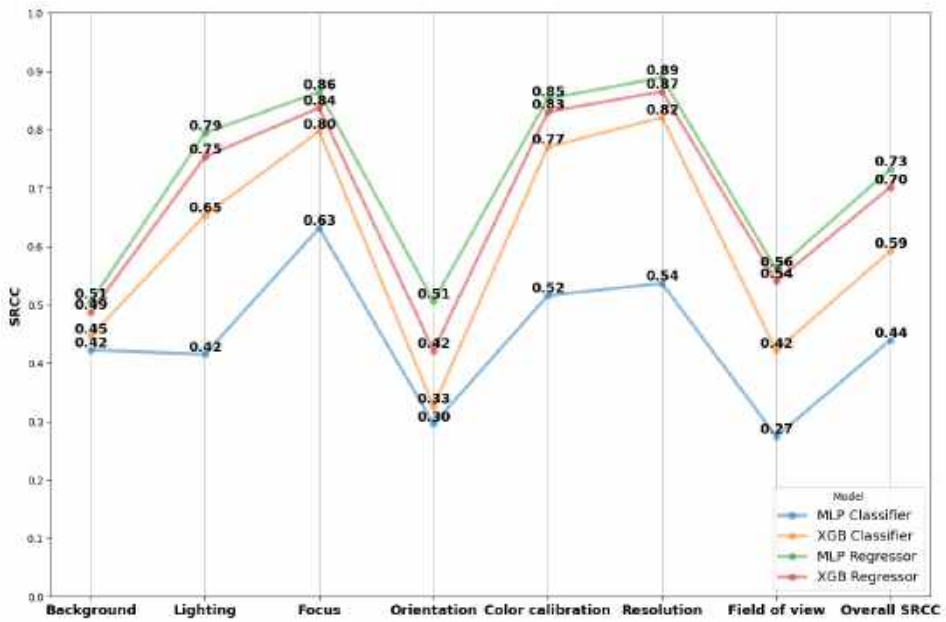
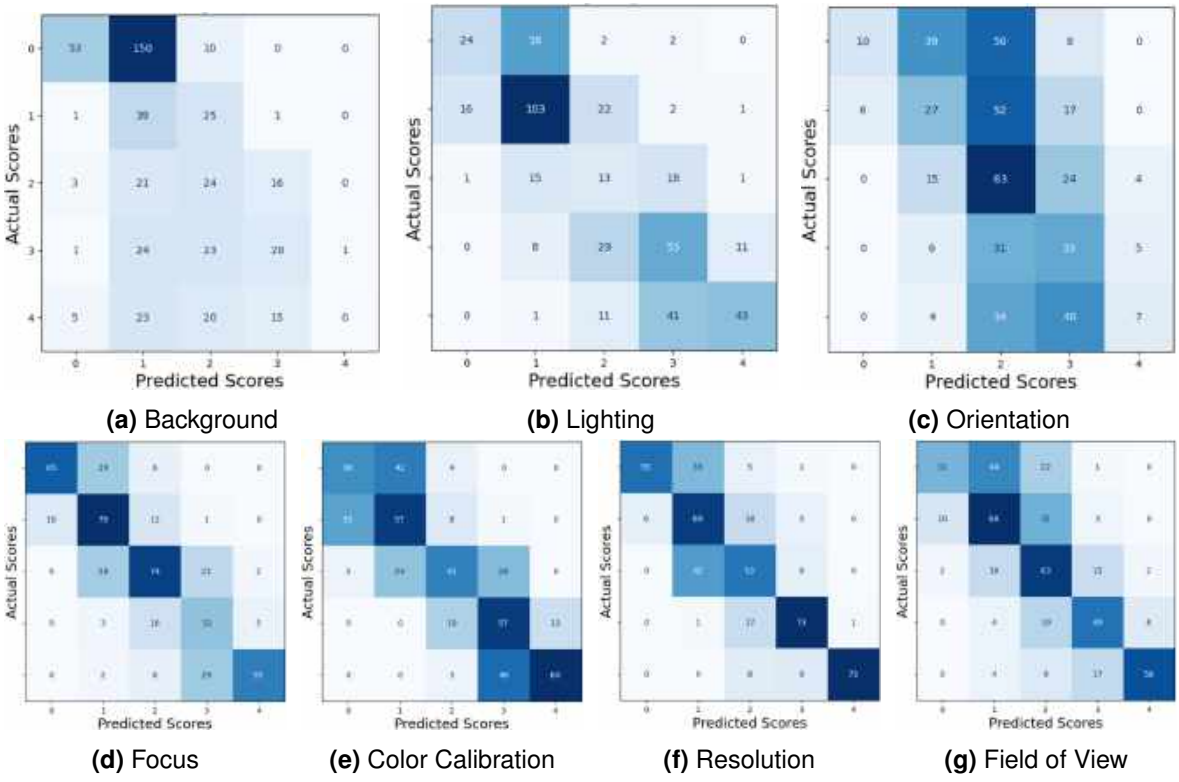


Figure 5.6: Parallel coordinate plot showing the best SRCC values for the four different models across the seven criteria and the overall SRCC. This plot highlights how each model performs in different areas and shows that the MLP Regressor generally performs the best.

5.7 Confusion Matrices

The confusion matrices in Figure 5.7 show how well the MLP Regressor model performs on the $F17K_{distorted}$ dataset for each distortion criterion. These matrices help visualize where the model makes correct predictions and where it makes errors, providing a detailed view of its accuracy. Additionally, they reveal any biases the model might have towards certain severity levels, showing if it tends to predict only low or high severity levels or if its predictions are skewed in some way.



(a) Background **(b)** Lighting **(c)** Orientation

(d) Focus **(e)** Color Calibration **(f)** Resolution **(g)** Field of View

Figure 5.7: Confusion matrices for the MLP Regressor model evaluated on the $F17K_{distorted}$ dataset. Each matrix corresponds to a specific distortion criterion and shows the actual scores on the y-axis and the predicted scores on the x-axis. Darker shades indicate higher counts, highlighting where the model's predictions match the actual values and where discrepancies occur.

For the background criterion (Figure 5.7a), the model mostly predicts lower severity levels, showing a habit of underestimating the severity of background distortions. In the lighting criterion (Figure 5.7b), the models predictions are quite accurate, though there are some differences across the severity levels. When it comes to orientation (Figure 5.7c), the model shows a tendency to predict middle severity levels for these distortions.

The focus criterion (Figure 5.7d) shows that the model performs well, with accurate predictions across different severity levels. For color calibration (Figure 5.7e, the model shows good performance with some differences but no significant bias. In the resolution criterion (Figure 5.7f), the models predictions are quite accurate across different severity levels. Finally, for the field of view criterion (Figure 5.7g), the model performs well, with predictions spread across severity levels without significant bias.

5.8 Performance Metrics

The performance metrics of the final MLP regressor model on individual dermatology quality criteria are shown in Table 5.2. This table provides a detailed view of how well the model performs across different criteria and overall. It includes four key metrics: Mean Absolute Error (MAE), R^2, Spearman Rank Correlation Coefficient (SRCC), and Cohens Kappa.

This table shows the models strengths and weaknesses. For example, the SRCC values for background, orientation, and field of view are the lowest among the seven criteria, indicating that the model struggles more with these distortions. On the other hand, higher SRCC values for criteria like lighting and focus suggest better model performance in those areas.

Table 5.2: Performance Metrics for Each Distortion Criteria

Criteria	MAE	R^2	SRCC	Cohen's Kappa
Background	0.9684	0.2595	0.5422	0.4399
Lighting	0.5726	0.6440	0.8028	0.7913
Focus	0.4042	0.7385	0.8622	0.8568
Orientation	0.9895	0.1824	0.4735	0.4102
Color calibration	0.4905	0.7334	0.8622	0.8583
Resolution	0.3642	0.7656	0.8722	0.8726
Field of view	0.5474	0.5976	0.7710	0.7660
Overall	**0.6195**	**0.5646**	**0.7507**	**0.7396**

5.9 Model Predictions

To understand how well the model performs on the two test sets ($SCIN_{synthetic}$ and $SCIN_{authentic}$), radar charts were used. These charts show the different criteria around the outside, with the severity levels ranging from the center (0) to the outer edge (1), indicating high distortion for each criterion. These visualizations give a clear and simple view of the models performance, showing both its strengths and areas where it can improve.

5.9.1 Visualizations for Synthetic Distorted Images

Figure 5.8 shows how the model predicts distortions for synthetic test images. These visualizations compare the models predictions with the actual distortions. This layout also helps to see how accurately the model can predict different types of distortions.

The first column shows the original image, the second shows the distorted image, the third contains the actual labels, and the fourth presents the models predictions.

5.9.2 Visualizations for Authentic Images

Figure 5.9 shows the models performance on authentic images, comparing its predictions with human-labeled scores. This helps to understand how well the model works in real-world situations and how closely the models predictions match with human evaluations.

The first column shows the image, the second column displays the human-labeled scores, and the third column presents the models predictions.

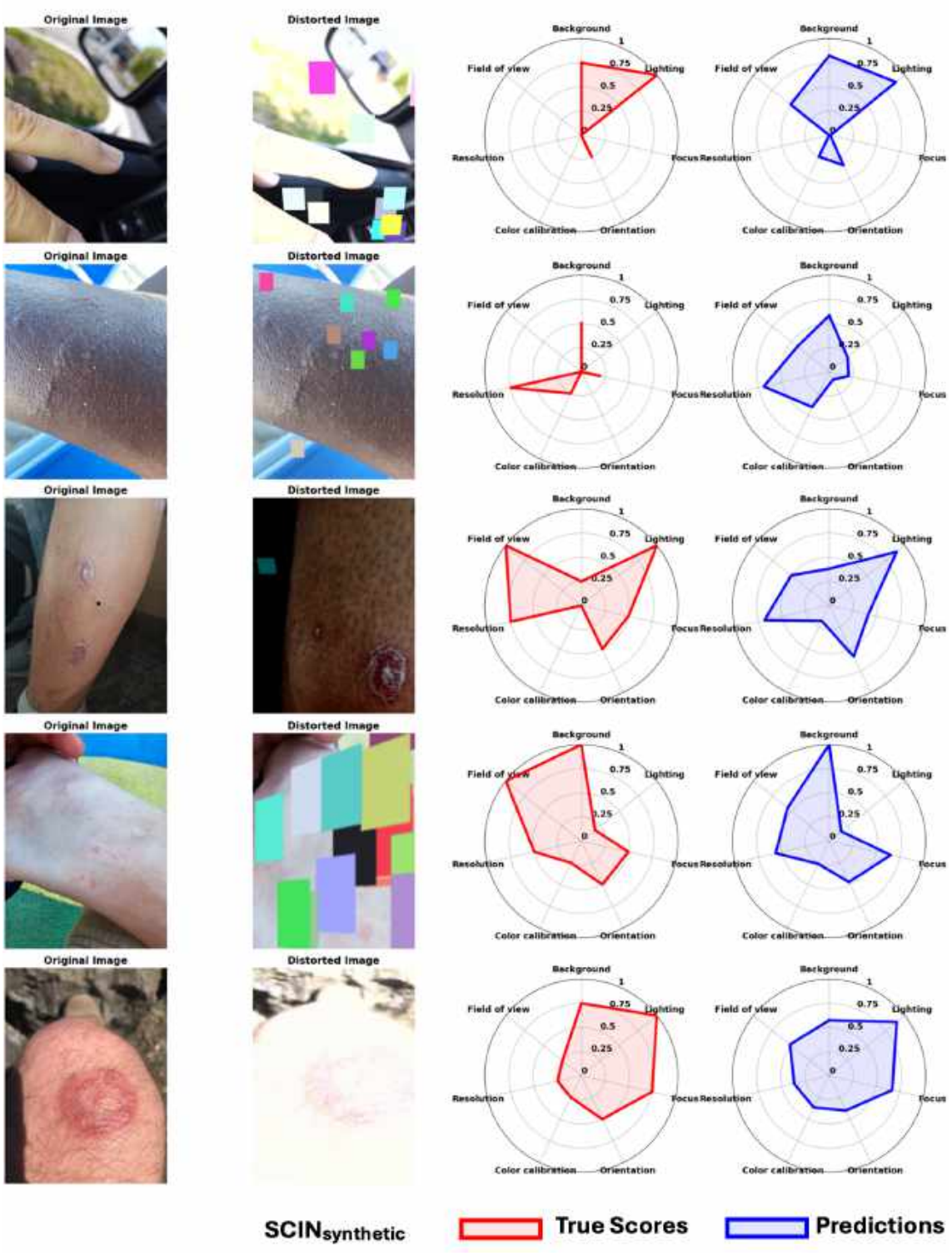


Figure 5.8: Comparison of model predictions with synthetic distorted test images using a four-column layout. The four-column layout shows the original image, the distorted image, the actual labels, and the model's predictions.

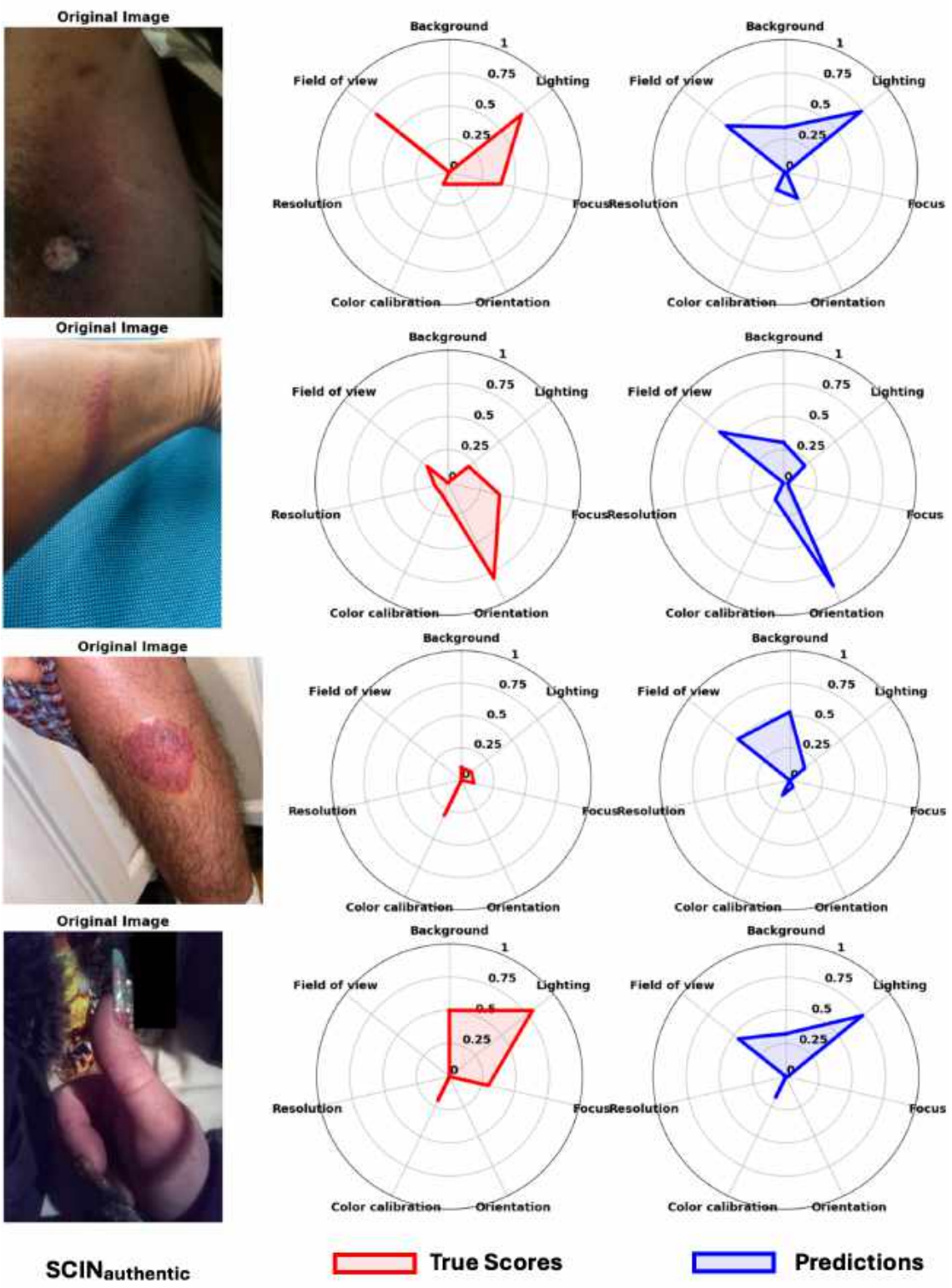


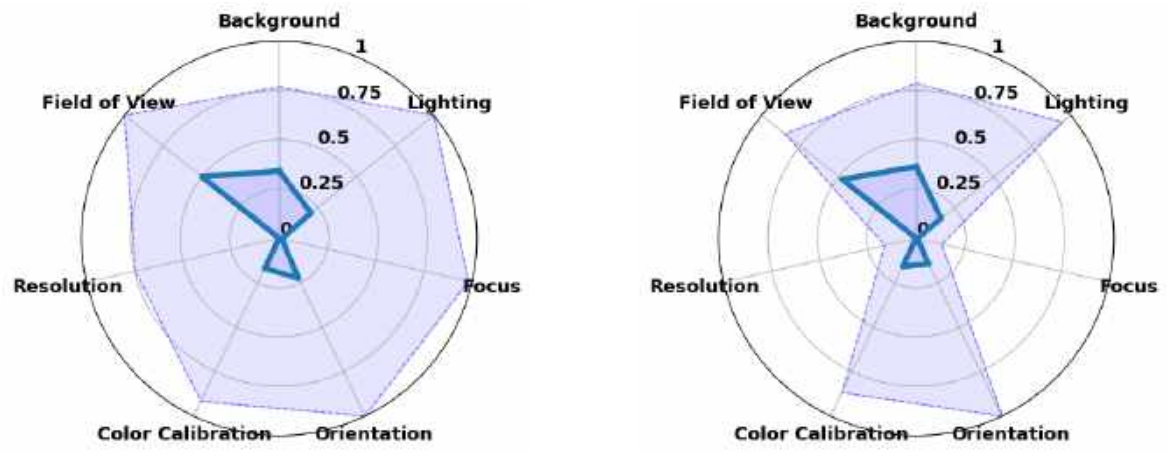
Figure 5.9: Comparison of model predictions with human-labeled scores for authentic images using a three-column layout. The three-column layout shows the image, the human-labeled scores, and the model's predictions.

5.10 Assessing Training and Testing Images Quality

To verify the quality of the images used for training and see how they change after synthetic distortion, radar charts were created. These charts show the quality of the original training images and how they are affected by the distortions. Additionally, the quality of both the synthetic and authentic test images is assessed using the same method. These radar charts show a simple visual representation of the quality and the level of distortion across the seven criteria. In these radar charts, the median value is shown as the blue line, while the dotted line represents the maximum value.

5.10.1 Training Images Quality

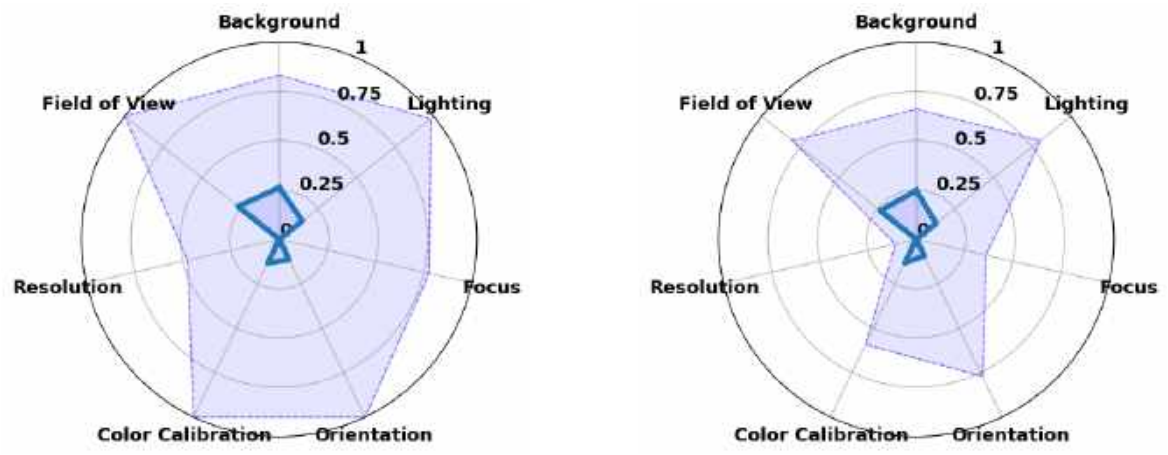
Figure 5.10 and Figure 5.11 show the quality of the original SCIN(Ward et al., 2024) and Fitzpatrick17k(Groh et al., 2021) images, and the filtered good quality images, respectively. Figure 5.12 shows the quality of the combined $SCIN_{good}$ and $F17K_{good}$ images and the synthetically distorted $COMB_{distorted}$ images used for training the model. These charts show the differences in quality before and after filtering and distortion.



(a) SCIN Images

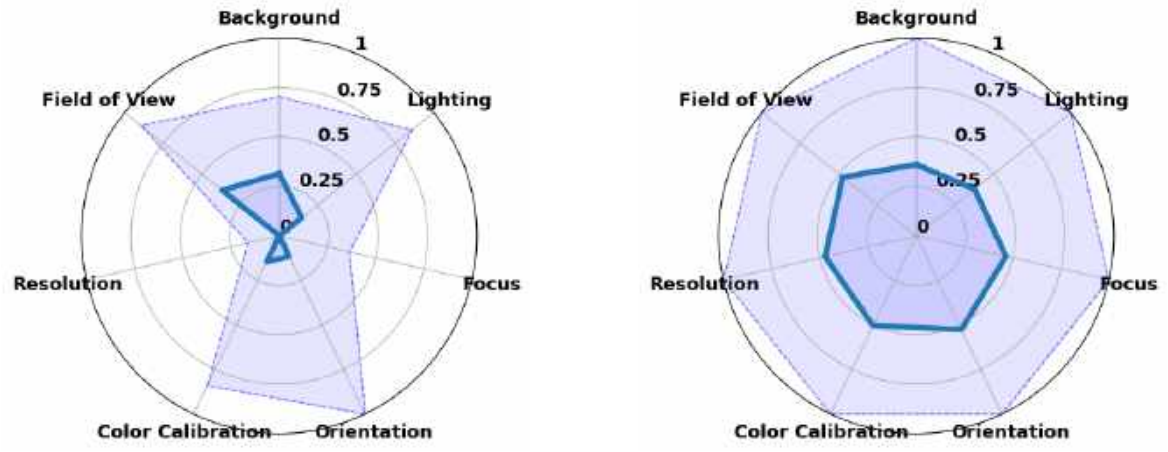
(b) Filtered SCIN Images

Figure 5.10: Radar charts for the SCIN dataset. (a) Original images from the SCIN dataset (10'379 images). (b) Filtered good quality images ($SCIN_{good}$). Note: The blue line represents the median, and the dotted line represents the maximum.



(a) Fitzpatrick Images **(b)** Filtered Fitzpatrick Images

Figure 5.11: Radar charts for the Fitzpatrick dataset. (a) Original images from the Fitzpatrick dataset (16'577 images). (b) Filtered good quality images ($F17K_{good}$). Note: The blue line represents the median, and the dotted line represents the maximum.

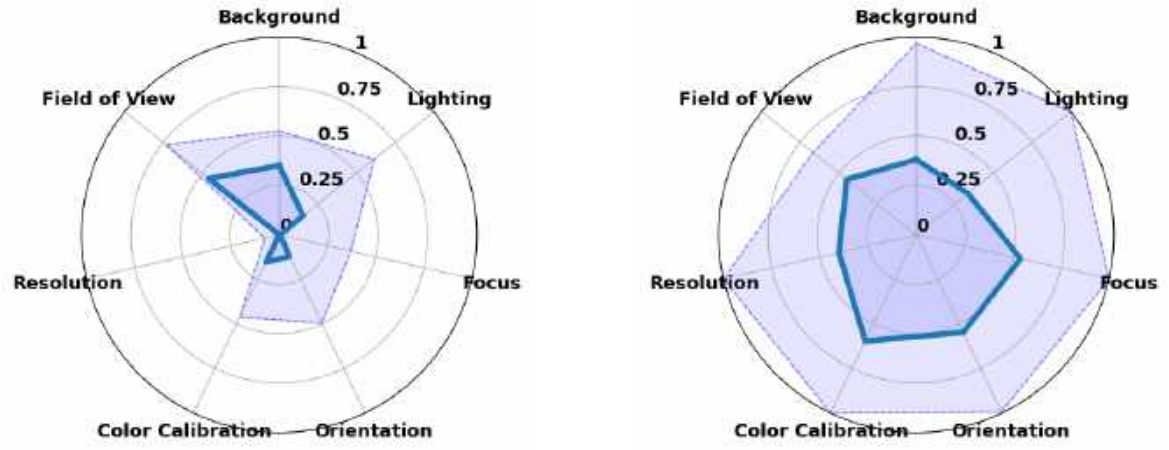


(a) Combined SCIN and Fitzpatrick Images **(b)** Synthetically Distorted Images

Figure 5.12: Combined dataset analysis. (a) Combined $SCIN_{good}$ and $F17K_{good}$ images. (b) Synthetically distorted images ($COMB_{distorted}$). Note: The blue line represents the median, and the dotted line represents the maximum.

5.10.2 Test Images Quality

Figure 5.13 shows the quality of the $SCIN_{good}$ test images and the synthetically distorted $SCIN_{synthetic}$ test images. Figure 5.14 shows the quality of the $SCIN_{authentic}$ test images from the SCIN dataset. These radar charts show a simple visual representation of the quality of the test images and the level of distortion across the seven criteria.



(a) Filtered Test Images **(b)** Synthetically Distorted Test Images

Figure 5.13: Synthetic test set analysis. (a) Filtered good quality test images (70 images, independent of training set). (b) Synthetically distorted test images ($SCIN_{synthetic}$) Note: The blue line represents the median, and the dotted line represents the maximum..

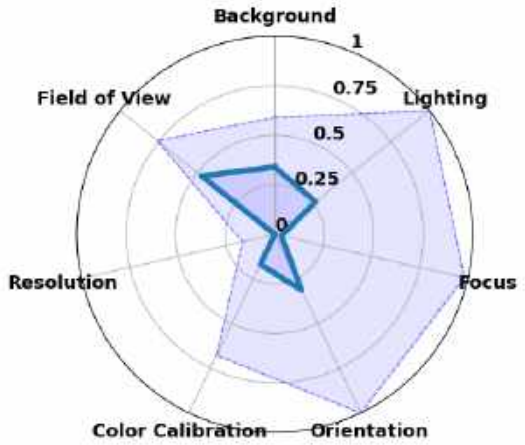


Figure 5.14: Authentic test set from the SCIN dataset, independent of the training images, showing real-world distortions ($SCIN_{authentic}$). Note: The blue line represents the median, and the dotted line represents the maximum.

5.11 Baseline Comparison on Synthetic and Authentic Distortions

Figure 5.15 compares single quality scores for synthetic distortions across different methods and Figure 5.16 compares single quality scores for authentic distortions. The x-axis shows the number of images, while the y-axis shows the quality score for each image. The SRCC values in the legend tell us how well each method matches the actual scores and the main focus is on which smoothed line best matches the trend of the actual scores.

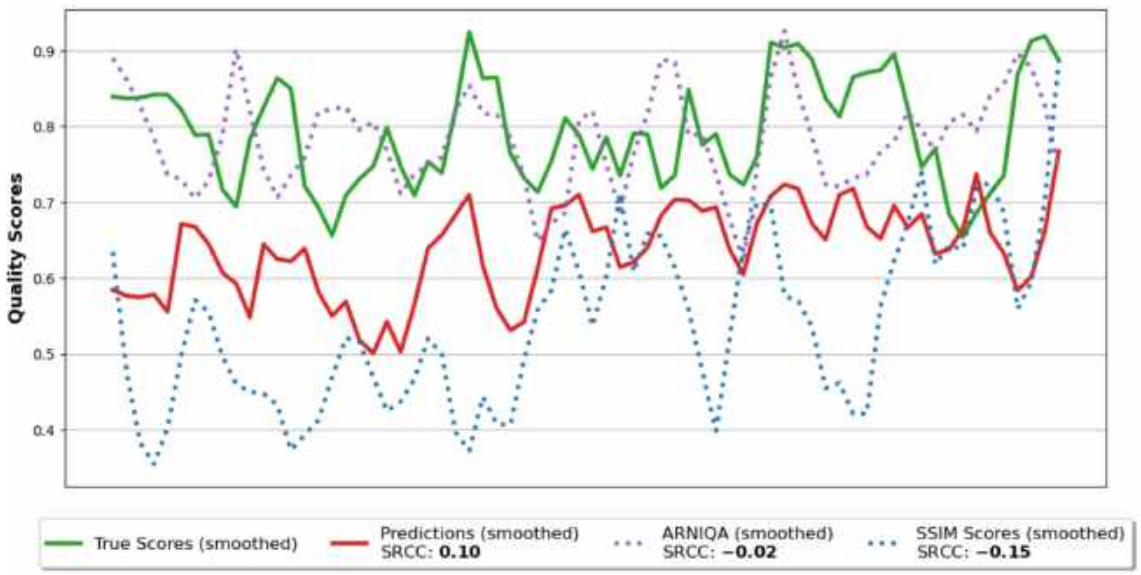


Figure 5.15: Comparing single quality scores for synthetic distortions using the proposed model, SSIM, and ARNIQA.

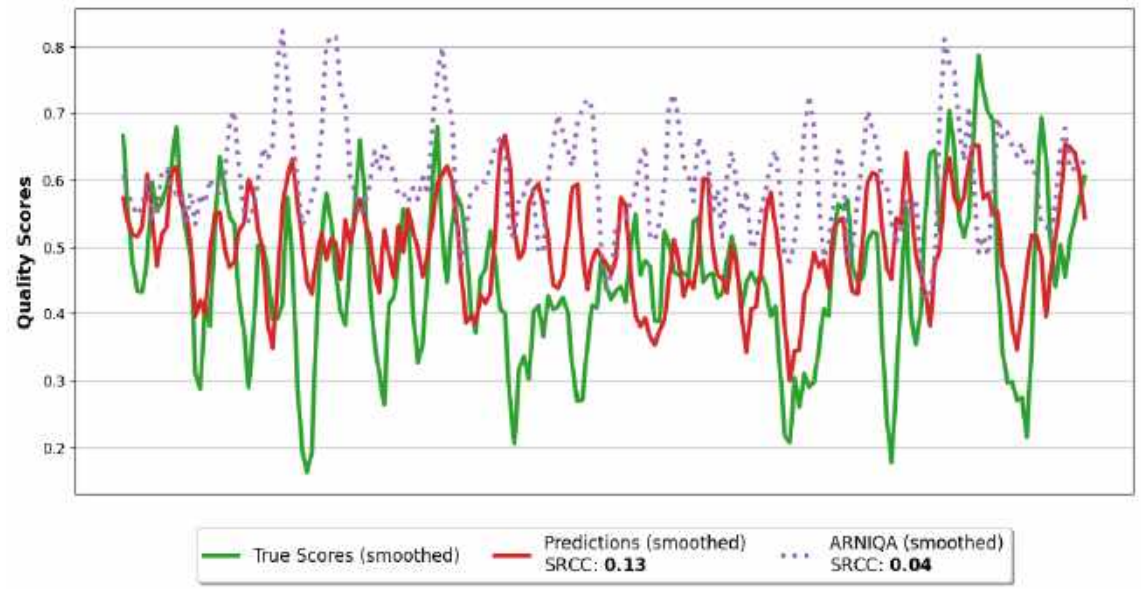
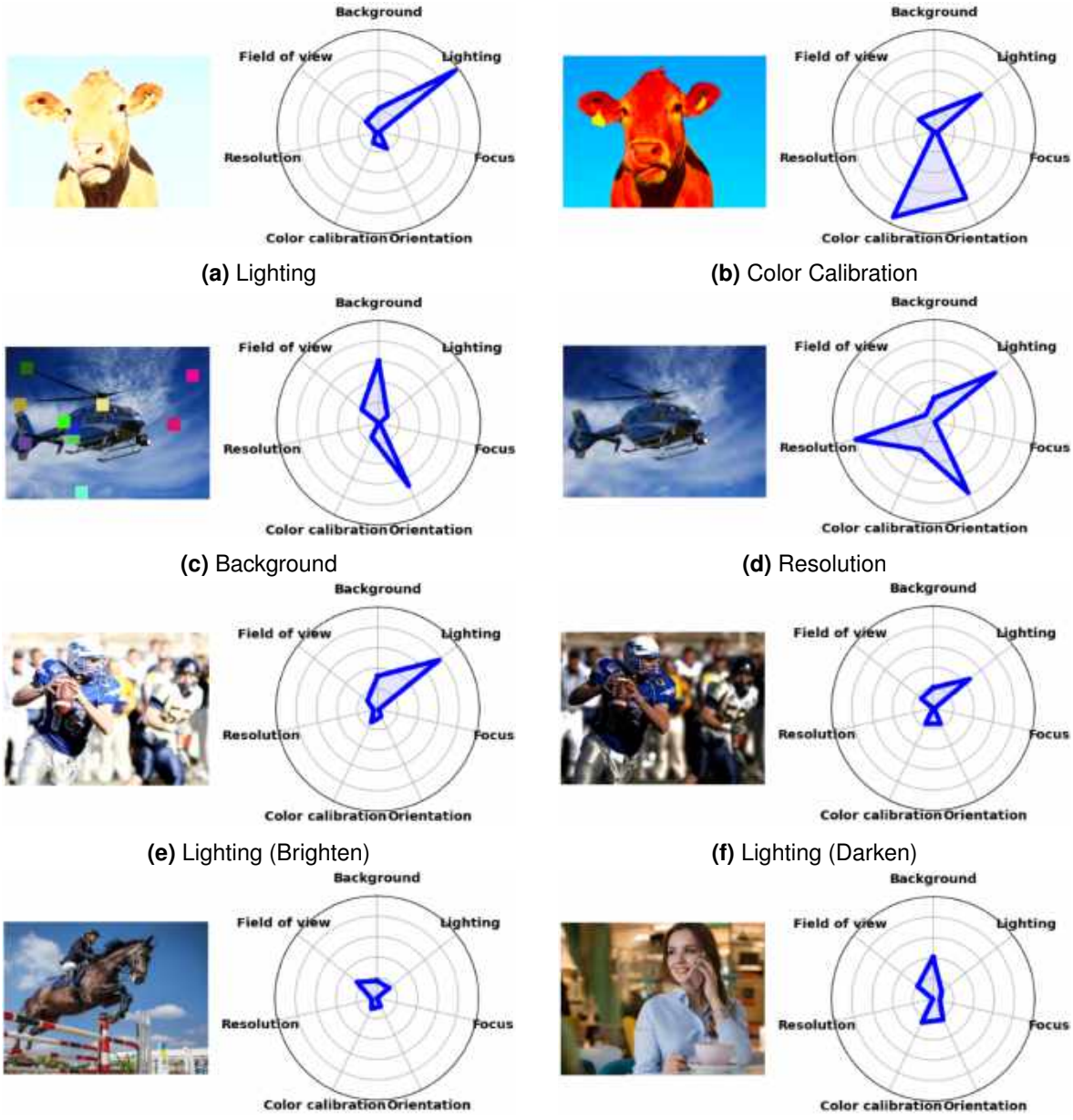


Figure 5.16: Comparing single quality scores for authentic distortions using the proposed model and ARNIQA.

5.11.1 Out of Distribution Testing

Figure 5.17 shows images which are different form teledermatology images. The left side shows the images, and the right side shows radar charts that indicate the models assessment of the seven quality criteria for each image.



(a) Lighting

(b) Color Calibration

(c) Background

(d) Resolution

(e) Lighting (Brighten)

(f) Lighting (Darken)

(g) Good Quality

(h) Good Quality

Figure 5.17: Out of distribution testing with images from the KADID10K dataset.

Chapter 6

Discussion

The chapter covers the entire work and, in particular, the results will be interpreted, the objectives of the thesis will be reviewed, and recommendations for future research in teledermatology image quality assessment will be provided.

6.1 Interpretation of Results

This section will discuss the results presented in Chapter 5. The performance of the final model will be highlighted, compared with other models, and explained in detail using different evaluation metrics. This analysis will cover model performance, loss curves, confusion matrices, and predictions on test sets. Additionally, the model will be compared with baseline methods like ARNIQA and SSIM to understand its strengths and weaknesses. The goal is to provide a clear understanding of how well the model performs and pinpoint areas that need improvement.

6.1.1 Final Model Selection and Cross-Dataset Evaluation

The final model chosen for this research is the MLP Regressor. This model was selected based on its best performance in terms of the SRCC metric. The cross-dataset evaluation results, summarized in Table 5.1, show that the MLP Regressor trained on $COMB_{distorted}$ dataset achieved the highest SRCC scores of 0.66 on $SCIN_{distorted}$ and 0.75 on $F17K_{distorted}$ images. This indicates that the model performs well across different datasets.

From the evaluation, it was clear that regressors performed better than classifiers, and models trained on the combined dataset outperformed those trained on individual datasets. The combined dataset approach likely works better because it includes a wider variety of images. For example, the $SCIN_{distorted}$ dataset has more images with background distortions, leading to better performance in predicting background problems, while the $F17K_{distorted}$ dataset has more diverse field of view distortions. This balanced approach improves overall performance. The following Table 6.1 shows the performance metrics for field of view distortion using an MLP Regressor on $SCIN_{distorted}$ and $F17K_{distorted}$ images individually. The results suggest that field of view accuracy depends on the type of images in the dataset. $F17K_{good}$, which has fewer images with backgrounds, performs better for field of view distortion. In contrast, $SCIN_{good}$ has more images with background elements, making it harder to accurately handle field of view distortions.

Table 6.1: Performance metrics for field of view distortion using an MLP regressor on synthetically distorted SCIN and F17K images.

Dataset	MAE	R^2	SRCC	Cohen's Kappa
$SCIN_{distorted}$	1.20	0.05	0.23	0.08
$F17K_{distorted}$	0.63	0.50	0.72	0.71

Also, multiplying the number of distortions was tested, as shown in Figure 5.3 and Figure 5.4. The results show that increasing the number of distortions in the dataset generally leads to better model performance. However, this improvement comes with longer training times. This balance is important when deciding how many distortions to include in the training data. Both XGB and MLP models, whether regressors or classifiers, perform better with larger, more diverse training data, which helps them assess image quality more accurately.

6.1.2 Analysis of Parallel Coordinate Plot and Loss Curves

The parallel coordinate plot in Figure 5.6 shows how the MLP Regressor compares to other models (XGB Regressor, XGB Classifier, and MLP Classifier) when trained on the $COMB_{distorted}$ dataset. It is clear that the MLP Regressor performs the best overall across all dermatology quality criteria based on SRCC. All four models show similar patterns, where they perform well on focus, color calibration, resolution, and lighting but have trouble with background, orientation, and field of view.

This pattern is also seen in the loss curves shown in Figure 5.5. The loss curves show that background, orientation, and field of view have higher losses compared to the other criteria. This means that the model finds it harder to learn these three criteria effectively.

One possible reason for this trend is the use of the ARNIQA (Agnolucci et al., 2023) backbone for feature extraction. The ARNIQA model was originally trained on a general imaging domain, focusing also on criteria like focus, color calibration, resolution, and lighting. This pre-training likely results in better performance on these criteria, as the model has already learned to handle them well. However, this also means that the backbone is less effective for domain specific criteria like background, orientation, and field of view, which are more relevant to teledermatology.

6.1.3 Performance Metrics and Confusion Matrices

When looking at the performance metrics in Table 5.2 and the confusion matrices in Figure 5.7, it is clear that the model does well in predicting focus, color calibration, and resolution. Although there are some differences between predicted and actual severities, these are usually small, varying by just one severity level. This small difference shows that the models predictions are quite accurate. This accuracy may also be due to the ARNIQA backbone used for feature extraction (this point was discussed before in Subsection 6.1.3).

However, the models performance on the lighting criterion is only moderate, even though ARNIQA also focused on this distortion. This could be because the lighting criterion includes two opposite types of distortions: brightening and darkening. If the training set mostly contains brightened images while the validation set includes darkened images, this mismatch could negatively affect the models performance. As a result, the model has trouble accurately predicting lighting severity due to these opposite distortion types.

For the background criterion, the confusion matrix shows that higher severity levels are rarely predicted. This is because, in the distortion process, no color blocks are added if the background is less than 10% of the image, resulting in a score of 0 for background distortion. Many images, therefore, have a 0 value for background distortion. The radar chart for the combined synthetically distorted images in Figure 5.12b also shows that the median value for background distortion is lower than for other dermatology quality criteria, meaning there are fewer strong severity values.

The confusion matrix for the orientation criterion show that predictions are generally unclear, often grouping around middle severity levels. This could be due to the different perspective changes (top, bottom, right, left) applied during training. As a result, the model detects some level of perspective distortion but cannot precisely figure out the direction or severity, leading to middle-level predictions.

Finally, field of view distortion is the most challenging criterion. In teledermatology, it is very important to center the lesion or area of interest in the image. However, in general photography, subjects are often placed off-center for artistic reasons (for example golden ratio). This difference might explain why the ARNIQA model, trained for general image quality assessment, has difficulty with field of view distortions specific to teledermatology.

6.1.4 Model Predictions

Synthetic Distorted Images

The models performance on the $SCIN_{synthetic}$ in Table 6.2 matches well with the validation split and cross-dataset evaluation.

Table 6.2: Model Performance Metrics for $SCIN_{synthetic}$ Test Images

Dataset	MAE	R^2	SRCC	Cohen's Kappa
$SCIN_{synthetic}$	0.71	0.48	0.69	0.65

Figure 5.8 shows randomly selected images in a four-column layout: the original image, the distorted image, the actual labels, and the models predictions.

For background distortion, the actual severity closely matches the predictions for most images. To further improve the models performance, an experiment could involve applying color blocks randomly on the whole image without any skin segmentation. This less typical approach for teledermatology images could test the models ability to handle random artifacts. The second and fourth images in the figure suggest this method could be effective.

One noticeable issue is the field of view distortion, where the predictions are not accurate. This was expected from earlier evaluations. Although background, lighting, and orientation distortions are not well-represented in the confusion plot, they appear quite accurate in the synthetic test set images. This suggests that while the confusion plot highlights general trends, individual image predictions can differ.

An important factor to consider is the distribution of the different distortion severity levels. The distortion pipeline selects ranges randomly, and the figures in Section 5.1 show the differences. Synthetic distorted images are evenly distributed, but authentic distortions are more right-skewed, showing higher distortions are less common. To improve this, an experiment was started to choose distortion ranges according to a Gaussian distribution centered at 0 severity with a standard deviation of 2.5. This approach might include more distortions with lower severity, which are more common in teledermatology images. For example, heavily brightened images, as seen in the last synthetic distorted image, may not happen often in real-world scenarios. Training the model on more common distortions could improve its performance. However, this experiment was not fully evaluated due to time constraints.

Authentic Distorted Images

For the $SCIN_{authentic}$ test images, random samples are shown with the labels provided by manual annotation in Figure 5.9. At first glance, the predictions do not match the manual labels well, as also reflected in the following Table 6.3.

Table 6.3: Model Performance Metrics for $SCIN_{authentic}$ Test Images

Dataset	MAE	R^2	SRCC	Cohen's Kappa
$SCIN_{authentic}$	0.97	-0.96	0.12	0.16

The SRCC is low, but the MAE suggests the model does not have large errors. Several factors could explain this difference. Firstly, the manual labeling focused primarily on the skin lesion, often overlooking the background, which might not align with the models overall assessment. Labeling 1,400 instances (200 images, 7 criteria each) likely introduced some human error.

Additionally, images with multiple distortions make it harder to assess individual criteria. For example, a heavily darkened image might hide other distortions like focus, resolution, and color calibration, making it difficult to evaluate accurately. Labeling image quality on images with lesions or marks on darker skin tones was challenging, which might have affected labeling accuracy. This highlights the wider issue of skin tone diversity in medical imaging, an important consideration for future research.

Even with these challenges, if the field of view distortion, which is the worst-performing criterion, is removed, then the models predictions are fairly accurate. This distortion significantly affects the radar charts. To keep the evaluation fair and accurate, the images were not re-labeled to avoid introducing any bias from the test set into the training process.

6.1.5 Comparison with Baselines

Figure 5.15 true scores show the actual distortion levels of the images, which were used as the ground truth. The SSIM scores have a negative correlation with the true scores (SRCC = -0.15), meaning SSIM is not good at capturing the true image quality. The fluctuations and differences between the quality scores suggest that SSIM is not a reliable measure for this specific use case. The ARNIQA quality scores, although somewhat close to the actual scores, have a slight negative correlation with the true scores (SRCC = -0.02). This means that ARNIQA also does not effectively assess the quality of these synthetically distorted images in teledermatology. The models predictions, however, show a slightly positive correlation with the true scores (SRCC = 0.10), which is better than SSIM and ARNIQA. This indicates that the proposed synthetic distortions pipeline used for preprocessing the images improves the models performance.

For the authentic distortions in Figure 5.16, both ARNIQA and the models predictions show a slight positive correlation with the true scores. ARNIQA achieves an SRCC of 0.04, while the proposed model achieves an SRCC of 0.13. Although the quality scores in synthetic and authentic are minimal, they can be improved by fine-tuning the weights assigned to the quality criteria.

Field of view, orientation, and background currently show less accuracy and negatively affect the overall quality score compared to the other four dermatology quality criteria. Adjusting the weights to reduce the influence of these three less accurate distortions improves the predictions, as shown in Figure 6.1. Here, the weights for field of view, orientation, and background are multiplied by 0.5, while the weights for focus, color calibration, and resolution are multiplied by 1.5. The same squaring of the weights is used and the average is taken. The final quality score improves, resulting in an SRCC of 0.16 for the synthetic distortions compared to SRCC of 0.10 from before. However, this adjusted weighting was not used for the final model, as the goal should be to achieve accuracy in all criteria, including field of view, orientation, and background, which play an important role in image quality assessment in teledermatology.

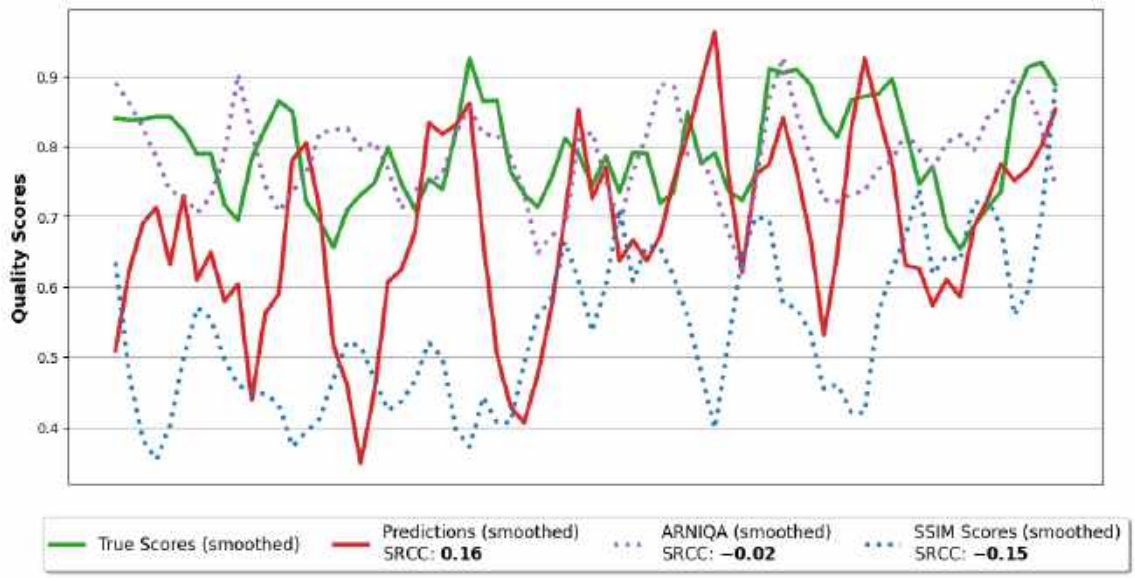


Figure 6.1: Improved quality scores for synthetic distortion predictions with adjusted weights for less accurate quality criteria. The weights for field of view, orientation, and background were reduced, while the weights for focus, color calibration, and resolution were increased, resulting in an improved SRCC score.

6.1.6 Out of Distribution Testing

In Figure 5.17, images of animals, vehicles, and people were used to test the models performance on general images. The radar charts show that the models predictions for these images are quite accurate. This is because the model focuses on the types of distortions rather than the content of the image. For example, Figure 5.17e and Figure 5.17f show the same image with different lighting distortionsone is brightened and the other is darkened. The model correctly spots these differences. Similarly, the good quality images in Figures Figure 5.17g and Figure 5.17h are also accurately predicted.

However, the model struggles with orientation and field of view distortions, which are more specific to teledermatology. This shows that while the model is not overfitted to teledermatology images, it is specially tuned to handle distortions important for this field. This means the model is flexible enough to work well on many types of images for common distortions but still needs improvement for specific teledermatology quality criteria. This balance makes sure the model is useful in many situations without losing its effectiveness for teledermatology.

6.2 Key Model Assumptions and Their Implications

The model assumes that the features extracted by ARNIQAs backbone are detailed enough to identify the main distortions in teledermatology images. This assumption works well for lighting, focus, color calibration, and resolution, which cover four out of the seven dermatology quality criteria. However, the criteria for background, orientation, and field of view need more testing and adjustments. The current performance shows that the model is somewhat effective, and if the assumption that ARNIQAs backbone can capture all key distortions is not entirely correct, it would mean that the model might not perform well in real-world scenarios where these distortions are common. To confirm these assumptions, more experiments with different datasets and real-world images should be conducted. By including a wider range of images in training, especially those with more background details, the model can be better prepared to handle real-world distortions. This is important for making the model stronger and more reliable in different situations.

Overall, the ARNIQA feature extraction backbone shows great potential for teledermatology applications, but continuous improvement and validation are essential to achieve the best possible performance.

6.3 Reviewing the Objectives of the Thesis

At the beginning of this thesis, several specific objectives were outlined:

- Conduct an extensive review of the literature on image quality assessment (IQA) methods, focusing on their application in teledermatology.
- Identify and select image quality metrics most suitable for assessing the quality of dermatological images.
- Evaluate the performance of selected image quality metrics on dermatological datasets to determine their effectiveness in assessing image quality.
- Develop a reproducible repository of image quality assessment tools and methodologies for teledermatology applications.

The first objective was to perform a detailed review of the literature on image quality assessment methods and how they apply to teledermatology. This process took a lot of time but was very important for the rest of the work. Through this review, key concepts such as IQA, teledermatology, and related studies were thoroughly examined and recorded.

The second objective was achieved during the literature review process. During this phase, the seven quality criteria from the International Skin Imaging Collaboration (ISIC) were found and chosen as the best metrics for assessing the quality of dermatological images. Selecting these dermatology quality criteria was very important for the next steps of the research.

For the third objective, the performance of the selected image quality metrics was evaluated on dermatological datasets. These evaluations were very detailed, involving tests on independent images that were not part of the model training. The datasets included both synthetically distorted images and images with real-world distortions, providing a complete assessment of how well the metrics worked.

The final objective was to develop a reproducible repository of image quality assessment tools and methods for teledermatology applications. This was successfully completed, creating a strong framework for future experiments and research. The repository makes sure that the methods and tools developed in this thesis can be used and built upon, helping further advancements in this field.

6.4 Reflection

This research showed that assessing image quality in teledermatology is possible with the proposed method, providing a good starting point for future studies. The research had both difficult and easier phases. Initially, understanding and connecting key concepts in teledermatology required a lot of reading and careful organization of information without losing sight of the main goals. These early challenges were important parts of the research process. Detailed records of these challenges and the steps taken to address them, as discussed in bi-weekly meetings, are included in Appendix A for a complete overview.

Filtering good quality images and labeling the test set was the most time consuming and repetitive part of the research, but it was necessary. Accurate labeling ensured that model performance comparisons were reliable, and good quality images were crucial for effective training. While these task were tedious, their importance in achieving the researchs goals cannot be overstated.

Creating synthetic distortions and using severity ranges to generate artificial values for images was particularly interesting. Developing the distortion pipeline that could generate multiple synthetically distorted images from a single good quality image was one of the most innovative parts of this research. This new approach effectively addressed the lack of annotated images for teledermatology image quality assessment. By generating different distorted versions of the same image, the dataset was significantly expanded, which greatly improved the models performance.

The results showed that the model is robust and generalizes well, even when trained with a relatively small number of original images. The final model performed well, especially for criteria like lighting, focus, color calibration, and resolution. However, it still requires fine-tuning for field of view, background, and orientation distortions. Generally, high expectations are set, and leaving research unfinished is not preferred. However, due to time constraints, the research is complete up to this point. There remains potential for further improvements and more extensive testing to enhance the overall performance and reliability of the image quality assessment in the context of teledermatology.

6.5 AI Tools Used

In this work, several AI tools were used. ChatGPT was used to compress and summarize content. Additionally, it was used to optimize sentences and sections to make them more reader-friendly. Furthermore, GitHub Copilot was used in the development environment. It primarily helped in developing the Python scripts and models. These tools made the work more efficient and helped improve the overall quality of the thesis.

Chapter 7

Conclusion and Future Research

This thesis has shown that assessing image quality in teledermatology is possible using the proposed approach. The main findings highlight the models strength and ability to handle different types of distortions well, especially in areas like lighting, focus, color calibration, and resolution. The innovative distortion pipeline played a key role in creating a comprehensive dataset, which greatly improved the models performance. By generating multiple synthetic distortions, the model was trained on a wide range of image variations, improving its accuracy in assessing image quality in the context of teledermatology.

The MLP regressor stood out as the best-performing model, consistently achieving better SRCC scores than the other models on the combined dataset across all seven criteria. However, the research also found that the model had difficulties with certain criteria, such as background, orientation, and field of view. These areas had higher errors and less accurate predictions, indicating a need for further improvement and specific data collection.

Future research should focus on expanding the dataset with more diverse and representative images of teledermatology, particularly those that address the challenging criteria. Adding more varied background conditions and different perspectives will be important. Improving the labeling process by collaborating with dermatologists to help filter and label images will reduce human error and increase accuracy. Finding new methods for creating synthetic distortions that better reflect real-world scenarios will help refine the model further.

Another promising direction for future research is to make the model more understandable. Techniques like GradCam can be used to show which parts of an image the model focuses on when making predictions. This would help in understanding the models decision-making process and in identifying areas where the model might be making mistakes.

The reproducible repository developed in this research allows for ongoing exploration and development. It provides a strong foundation for future experiments, allowing researchers to build upon and improve the methods and tools developed in this thesis. Improving image quality assessment in the context of teledermatology can greatly improve remote consultations, making sure that dermatologists can rely on good quality images for accurate diagnoses and effective treatment.

List of Figures

List of Tables

Appendix A

Supplementary Material

The following pages contain the supplementary material for this thesis. This section includes documents specific to project planning and management. The documents are attached in this order:

- Project Assignment
- Risk Management
- Project Planning

Additionally, the appendix contains detailed information on the IQA databases from the general image domain, showcases the different dermatology quality criteria and their types across five severity ranges, and provides an overview of the proposed framework for automated IQA in teledermatology.

Aufgabenstellung

Modul:	Dept I BAA FS24

Titel:	Automated Image Quality Assessment in Teledermatology
Ausgangslage und Problemstellung:	ABIZ has been researching artificial intelligence applications in dermatology for the past decade with the objective to develop decision support systems to effectively support clinical practice. In collaboration with the University Hospital of Basel and the Swiss company Derma2go, we are tackling the issue of automatically assessing the quality of patient images for diagnosis, since this factor heavily impacts the effectiveness of teledermatology worflows.
Ziel der Arbeit und erwartete Resultate:	The objective of this work is to conduct an extensive review of state-of-the-art quality assessment methods in the general image domain and evaluate how they can be applied to teledermatology. The project deliverables include: - A comprehensive review of state-of-the-art image quality assessment methods. - A review of image quality criteria for teledermatology diagnosis. - An evaluation of selected quality assessment methods on public dermatology datasets. - A well-written repository enabling to reproduce reported results and assess the quality of new patient images.
Gewünschte Methoden, Vorgehen:	The project will start with a literature review of existing quality assessment methods and patient image quality criteria in dermatology. Together with the supervisor, adapted methods will be selected, which the student will then evaluate on public dermatology datasets. The student will present his work to the supervisor on bi-weekly meetings. One day before the meeting, the student will share a 1-page document describing in bullet points: - What work was performed during the last reporting period. - What work is planned for the next period. - Project status, comparison with planning, reasons for deviations if applicable. - Top three risks incl. planned measures. For the meeting, the student will prepare slides to present these information in more details.
Kreativität, Methoden, Innovation:	This thesis will encourage innovative approaches, including but not limited to proposing new metrics and relevant changes to adapt methods to the teledermatology context. The student will have the opportunity to fine-tune deep learning models on public dermatology datasets and work closely with both clinicians and researchers from ABIZ and the partner institutions.
Sonstige Bemerkungen:	Candidates should have a strong background in computer science. Prior experience with medical imaging or teledermatology is beneficial but not mandatory. The project will require a creative approach to problem-solving and an eagerness to work in interdisciplinary teams.

Projektteam

Student:in 1:	Choekyel Nyungmartsang
Betreuer:in:	Dr. Ludovic Amruthalingam

Auftraggeber

Firma:	Algorithmic Business Research Lab
Ansprechperson:	Dr. Ludovic Amruthalingam
Funktion:	
Strasse:	
PLZ/Ort:	
Telefon:	+41 41 349 30 74
E-Mail:	ludovic.amruthalingam@hslu.ch
Website:	

Version 13.06.2023 / bcl

Challenges and Planned measures

- "Staff taking images for teledermatology need to be good equipped, trained, identifiable and competent" [*4.1.3 TQS*]. What are requirements for patients taking images?
    - [Find updated quality standards, interpolate the camera/photographic specifications and photography protocol]
- Technical and medical terms/vocabulary
    - [Get familiar with the vocabulary]
- "Scale and Measurement Using Digital Imaging Software" [*7 domain ISIC*]. How does that work? I understand the concept, but I would like to "see" it.
    - [Maybe not necessary]
- Evaluation (adapted) methods?
    - [I have a general idea, but I would like to have a better understanding.]

- Many methods or processes written in the papers need long time to understand. [Get just an overview on the methods and remark it. If needed I can come back in the future.]
- Literature review and synthesis of the gathered information. [Sort the papers read into different categories and come back later to it if needed]

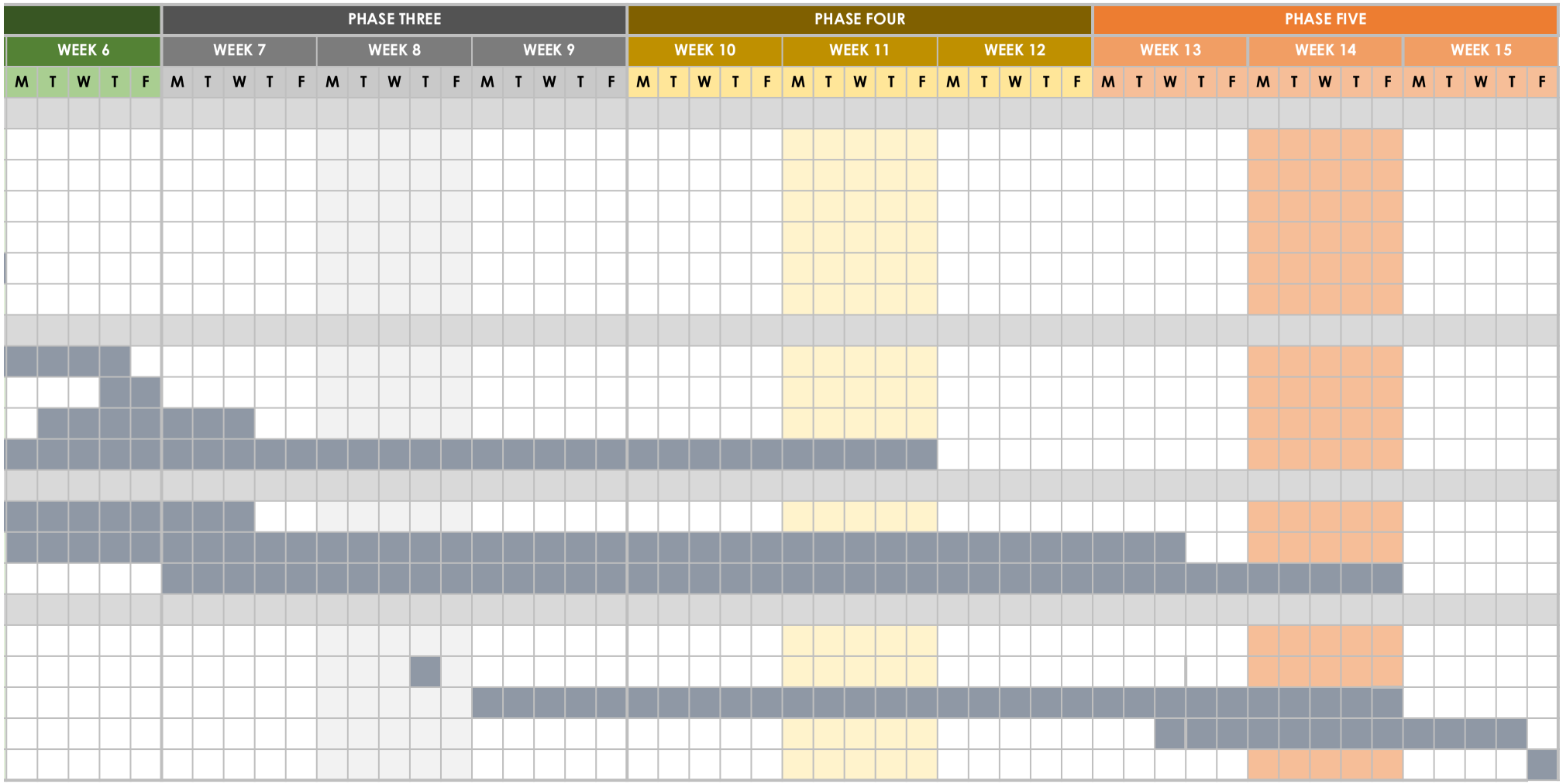
- SOTA IQA methods are mainly focused on distortions and feature extractions. Teledermatology IQA focuses additionally on framing and depth because the orientation o the skin and if the skin is too far away matters. [incorporate skin segmentation preprocess to mitigate depth and for framing... I must find other literature.]

- For finetuning I needed MOS or DMOS scores, where Fitzpatrick did not have. [SCIN dataset has dermatology confidence score, ranging from 1 to 5. I used that as an alternative to MOS. A single image could have multiple conditions so it can also have multiple confidence scores.]
- Getting an even distribution of the confidence score was at first a little challenging. [Since I wanted a single score per image, I took the median of the scores and took the min or max of the scores depending on, if the score is <2 or >2. This was done deliberately so most of the scores were then evenly distributed at the extremes.]
- SCIN dataset has 10'379 images. After preprocessing I am left with 6'503 images. Could be small for finetuning. [Getting more images!]
- The first results were not very satisfying because the model makes mistakes! [look at the features that were extracted from the encoder model with a t-SNE plot or look at if the dermatology confidence score matches the image in SCIN.]

- Background and orientation distortions are difficult to synthetically reproduce.
  [I decided to use "colorblock" for background distortion where randomly blocks are placed in the image to add artifacts. The idea behind this is to distract the model from the skin lesion.]
  [For orientation distortion I decided to change the perspective of the image, so I get different angles from the image. It is like tilting the image.]
- Field of view distortion is challenging to synthetically reproduce for FR IQA.
  [I could crop (upper right corner) from the images and the crops should then have the skin lesion at the down left corner, but I am unsure if this is a valid idea.]

- I tried to work with overleaf to write my thesis but since I have no access to the pro version I encountered some issues with syncing. [I switched back to writing in VSCode.]
- The DDI dataset was not very helpful. When sifting through the images I was not content with the images because it was not representative of teledermatological images. [I will not include it in my thesis.]
- The MAE and MSE values are not good enough. [Try other models that can capture the complexity]

- The metrics such as precision and recall are around 40%. [Hyperparameter tune with grid search or sweeps in wandb. Cross validate the dataset.]


PHASE THREE
PHASE FOUR
PHASE FIVE
WEEK 6
WEEK 7
WEEK 8
WEEK 9
WEEK 10
WEEK 11
WEEK 12
WEEK 13
WEEK 14
WEEK 15
M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F M T W T F

Automated Image Quality Assessment in Teledermatology

Choekyel Nyungmartsang	BAA	07/06/24

TASK ID	TASK TITLE	START DATE	DUE DATE	DURATION IN DAYS	PCT OF TASK COMPLETE
1	Project Initiation and Literature Review			86	
1.1	Kickoff Meeting	19.2.2024	19.2.2024	1	**100%**
1.2	Literature Review Planning	19.2.2024	20.3.2024	31	**100%**
1.2.1	Reviewing SOTA in IQA	19.2.2024	6.3.2024	17	**100%**
1.2.2	Reviewing SOTA in Teledermatology	7.3.2024	20.3.2024	14	**100%**
1.3	Dataset Acquisition Planning	11.3.2024	23.3.2024	13	**100%**
1.4	Define Objectives and Scope	19.2.2024	28.2.2024	10	**100%**
2	Model Development Phase			59	
2.1	Implementing IQA Models	20.3.2024	26.3.2024	7	**100%**
2.2	Data Annotation	28.3.2024	29.3.2024	2	**100%**
2.3	Reviewing First Results	26.3.2024	3.4.2024	9	**100%**
2.4	Iterative Model Refinement	20.3.2024	29.4.2024	41	**100%**
3	Project Launch and Execution			121	
3.1	Experiment for IQA	20.3.2024	3.4.2024	15	**100%**
3.2	Experiment for Teledermatology	25.3.2024	15.5.2024	52	**100%**
3.3	Analyzing Results + Drawing Conlusions	1.4.2024	24.5.2024	54	**100%**
4	Documentation + Presentation			65	
4.1	Introduction + SOTA	5.3.2024	15.3.2024	11	**100%**
4.2	Mid-term Presentation	8.4.2024	8.4.2024	1	**100%**
4.3	Writing Thesis Report	19.4.2024	24.5.2024	36	**100%**
4.4	Revision + Finalization	15.5.2024	30.5.2024	16	**100%**
4.5	Submission	7.6.2024	7.6.2024	1	**100%**

A.1 Dataset

Detailed information on image quality assessment (IQA) databases:

- **LIVE** (Laboratory for Image & Video Engineering) dataset (Sheikh et al., 2006) includes 29 reference images and 779 manually distorted images corrupted by 5 types of distortions: JPEG compression (JPEG), JPEG2000 compression (JP2K), white noise (WN), Gaussian blur (GB), and simulated fast fading Rayleigh channel (FF). Each distortion type contains 5 or 4 distortion levels. Most images are 768 $\times$ 512 pixels in size. Each distorted image in this dataset is associated with a Differential Mean Opinion Score (DMOS), scaled from 0 to 100, where 0 indicates no perceivable distortion.
- **TID2008** (Tampere image database 2008) dataset (Ponomarenko et al., 2009) includes 25 reference images and 1700 distorted images corrupted by 17 types of distortions, with 4 levels for each distortion type. All images have a fixed resolution of 512 $\times$ 384. This dataset provides MOS values and their standard deviations, with MOS ranging from 0 to 9, where 9 signifies a distortion-free image.
- **TID2013** (Tampere image database 2013) dataset (Ponomarenko et al., 2015) is extended from TID2008 (Ponomarenko et al., 2009) by increasing the number of distortion levels to 5, and the number of distortion types to 24. Therefore, 3000 distorted images are generated from 25 pristine images. The subjective testing and data processing steps are similar to that of TID2008. DMOS values for this dataset were derived from over half a million ratings given by nearly a thousand observers, with values ranging from 0 to 9, where higher values denote poorer image quality.
- **CSIQ** (Categorical subjective image quality (CSIQ) database) (D. M. Chandler, 2010) contains 30 pristine images and 866 distorted images corrupted by JPEG, JP2K, WN, GB, additive pink Gaussian noise, and global contrast decrements, with 5 or 4 levels for each distortion type. The resolution is 512 $\times$ 512. Each image in CSIQ is associated with DMOS values obtained from subjective ratings by 25 testers, with DMOS values scaled from 0 to 1, where higher values indicate worse quality.
- **A57** (D. Chandler & Hemami, 2007) includes 3 pristine images and 54 distorted images corrupted by 6 types of distortions, with 3 levels for each distortion type. All images are in gray scale. The resolution is 512 $\times$ 512.
- **WED** (Waterloo exploration database) (Ma et al., 2017) includes 4744 pristine natural images and 94880 distorted images corrupted by JPEG, JP2K, GB, and WN, with 5 levels for each distortion type. The images have various resolutions. No human opinion score is provided, but the authors introduce several alternative test criteria to evaluate the IQA models.

Multiple Distortions IQA Databases

- **LIVEMD** (LIVE multiply distorted) (Jayaraman et al., 2012) database consists of 15 reference images and 405 multiply distorted images. The database includes one/double-fold artifacts. Each multiply distorted image is corrupted under two multiple distortion scenarios: Gaussian blur followed by JPEG and Gaussian blur followed by white noise. All images have a resolution of 1280 $\times$ 720. DMOS values for each distorted image range from 0 to 100.
- **Multiply distorted image database 2013 (MDID2013)** (Gu et al., 2014): MDID2013 has a total of 12 pristine images and 324 distorted images. Each pristine image is corrupted successively by Gaussian blur, white noise, and JPEG. The images have resolutions of 768 $\times$ 512 or 1280 $\times$ 720.

- **Multiply distorted image database 2016 (MDID2016)** (Sun et al., 2017): MDID2016 consists of 20 reference images and 1600 distorted images. Five distortion types are introduced, i.e., white noise, Gaussian blur, JPEG, JPEG2000, and contrast change (CC). The order of distortions is as follows: Gaussian blur or CC first, JPEG or JPEG2000 second, and white noise last. All distorted images are with random types and levels of distortions. The image resolution is 512 $\times$ 384.

Screen Content IQA Databases

- **Screen Image Quality Assessment Database (SIQAD)** (Yang et al., 2014): SIQAD includes 20 pristine and 980 distorted screen content images (SCIs). Distortion types include white noise (WN), Gaussian blur (GB), color cast (CC), JPEG, JPEG2000 (JP2K), motion blur (MB), and layer segmentation-based compression, with 7 levels for each type. The images have various resolutions near 700 $\times$ 700.

- **Screen Content Image Quality (SCIQ) Database** (Ni et al., 2017): SCIQ consists of 40 pristine and 1800 distorted SCIs corrupted by 9 types of distortions, including WN, GB, MB, CC, JPEG, JP2K, color saturation change (CSC), color quantization with dithering (CQD), and the screen content coding extension of High Efficiency Video Coding (HEVC-SCC). Five distortion levels are considered. The resolution is fixed at 1280 $\times$ 720.

- **Cross-Content-Type (CCT) Database** (Min et al., 2017): CCT is constructed to conduct cross-content-type IQA research. CCT consists of 72 pristine and 1320 distorted natural scene images (NSIs), computer graphic images (CGIs), and SCIs. Two distortion types are considered, i.e., HEVC and HEVC-SCC coding, with 11 distortion levels for each type. The image resolution is either 1920 $\times$ 1080 or 1280 $\times$ 720.

- **Hybrid Screen Content and Natural Scene Image Database (HSNID)** (Gu et al., 2020): HSNID has 10 pristine NSIs and 10 pristine SCIs, and 600 distorted NSIs and SCIs corrupted by WN, GB, MB, CC, JPEG, and JP2K, with 5 distortion levels for each type.

Authentic Distortions IQA Databases

- **LIVE in the wild image quality challenge database** (Ghadiyaram & Bovik, 2016) includes 1162 authentically distorted images captured using a variety of mobile devices. Complex real distortions, which are not well-modeled by the synthetic distortions are included. All images are cropped to the resolution of 500 $\times$ 500. A novel crowdsourcing system was employed to gather over 350,000 opinion scores from 8100 observers, ensuring the objectivity of the MOS values obtained.

- **Camera image database (CID2013)** (Virtanen et al., 2015): CID2013 is designed to test no-reference IQA algorithms. It includes 480 real images captured from 8 typical scenes using 79 consumer cameras and mobile phones. The images are rated from 5 aspects: the overall quality, sharpness, graininess, lightness, and color saturation scales. The images are scaled to a size of 1600 $\times$ 1200.

A.2 Degradation Types

The dataset used in this thesis is augmented with synthetic degradations. The following figures below show the different levels of intensity for the degradations of each distortion group.

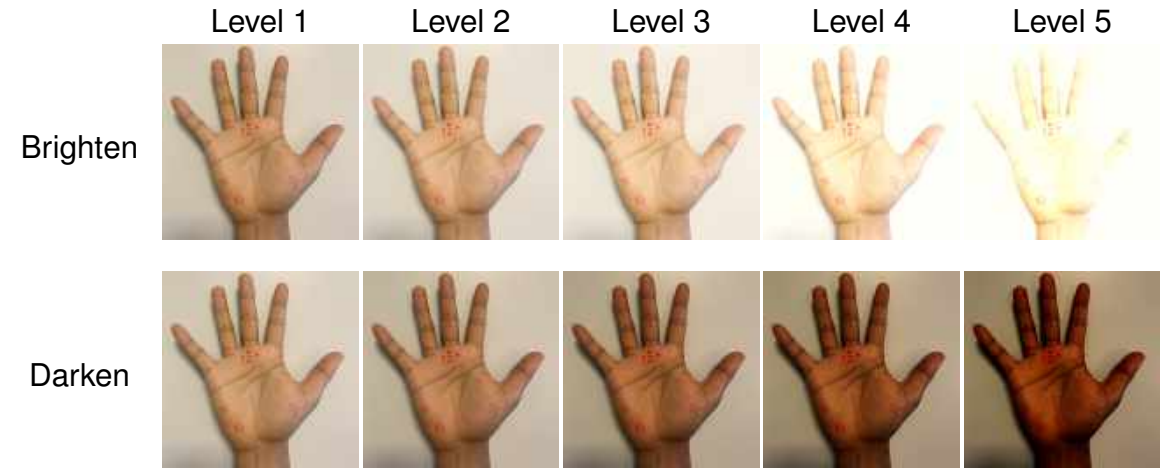


Figure A.1: Visualization of the degradation types belonging to the *Brightness change* group for increasing levels of intensity.

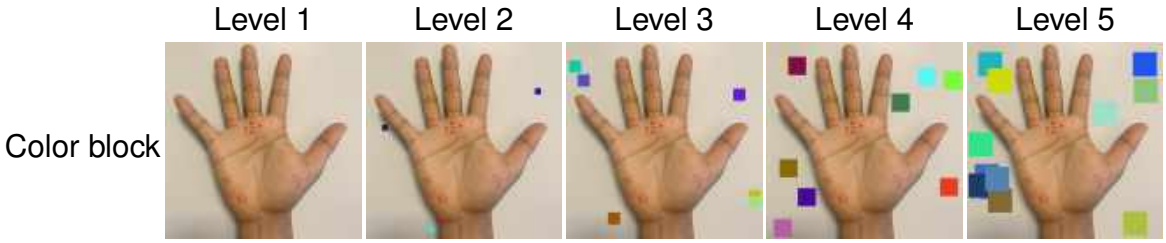


Figure A.2: Visualization of the degradation types belonging to the *Background color* group for increasing levels of intensity.

Figure A.3: Visualization of the degradation types belonging to the *Field of View* group for increasing levels of intensity.

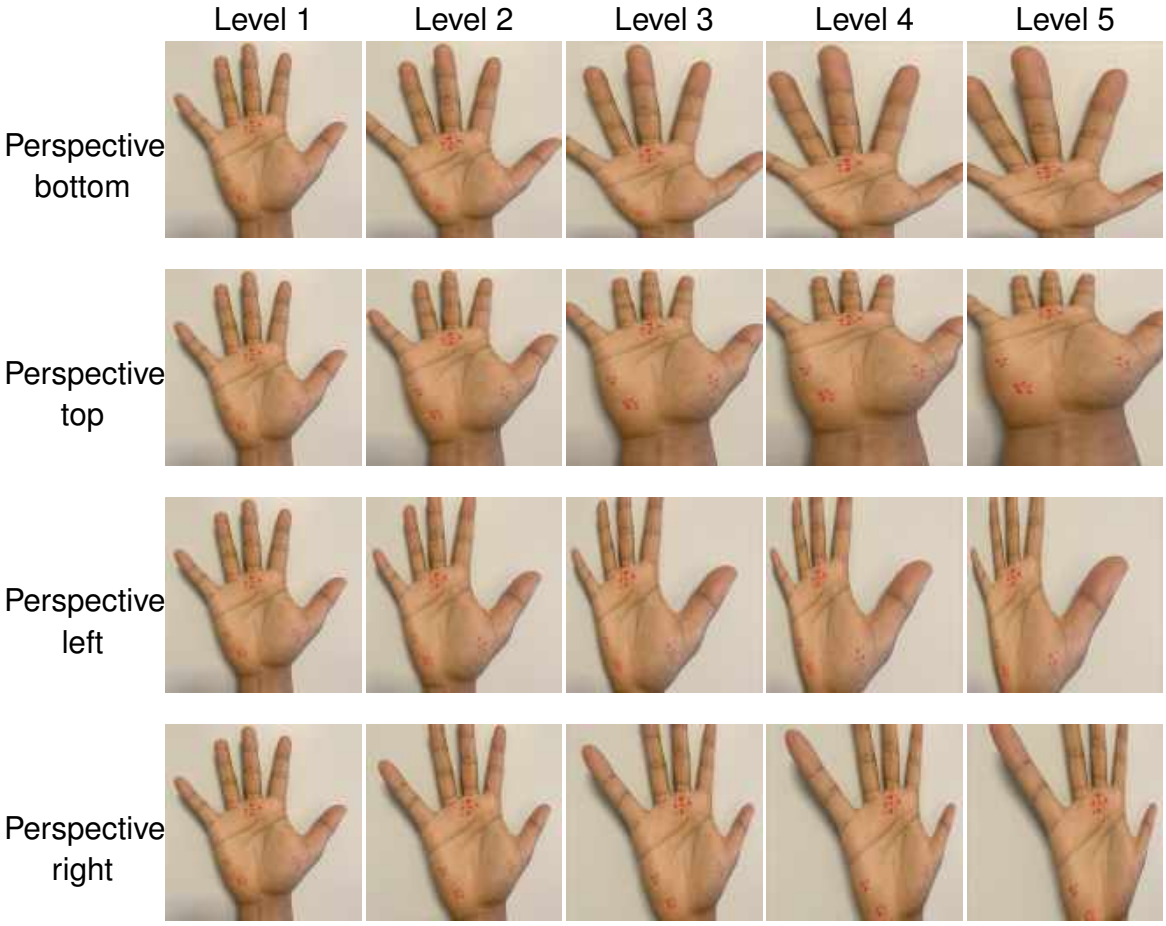


Figure A.4: Visualization of the degradation types belonging to the *Image orientation* group for increasing levels of intensity.

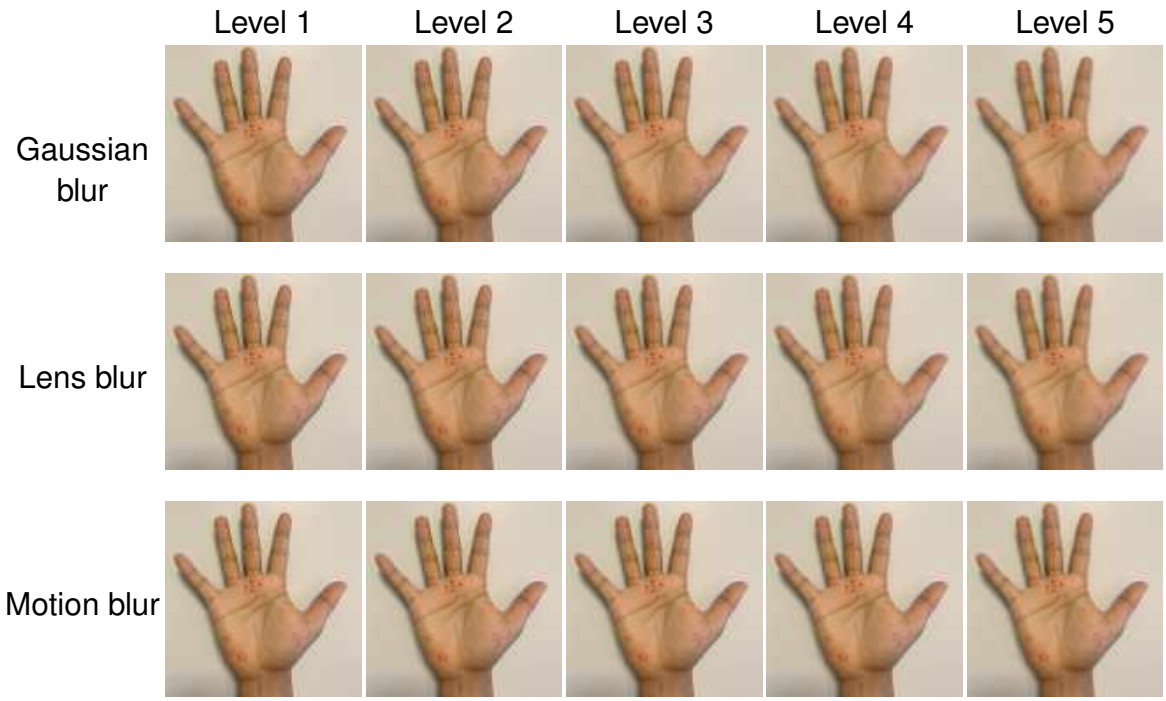


Figure A.5: Visualization of the degradation types belonging to the *Focus* group for increasing levels of intensity.

Figure A.6: Visualization of the degradation types belonging to the *Resolution* group for increasing levels of intensity.

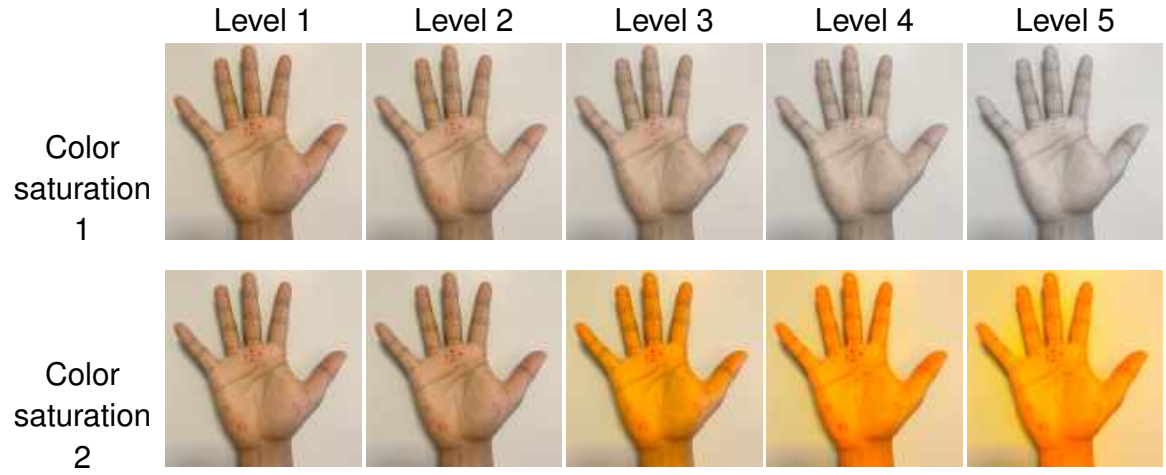


Figure A.7: Visualization of the degradation types belonging to the *Color calibration* group for increasing levels of intensity.

A.3 Proposed Framework for Automated Image Quality Assessment in Teledermatology

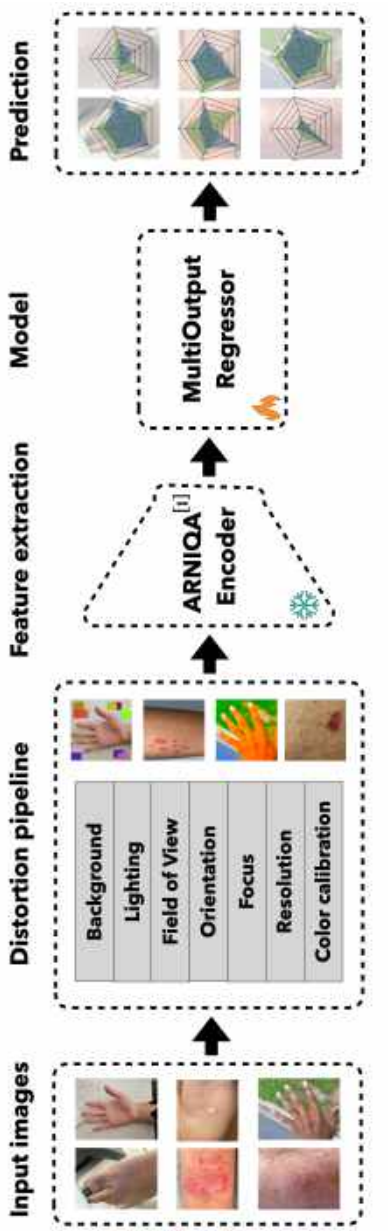


Figure A.8: Overview of the proposed framework for automated image quality assessment in teledermatology. It starts with good quality input images, which go through a distortion pipeline covering seven dermatology quality criteria. The ARNIQA backbone extracts features, and a multi-output regressor predicts the seven quality criteria. The final predictions are visualized with radar charts.

Bibliography

Agnolucci, L., Galteri, L., Bertini, M., & Del Bimbo, A. (2023, November 4). ARNIQA: Learning distortion manifold for image quality assessment. Retrieved April 23, 2024, from http://arxiv.org/abs/2310.14918

Ahsan, M. M., Uddin, M. R., & Luna, S. A. (2022, June 3). Monkeypox image data collection. Retrieved April 28, 2024, from http://arxiv.org/abs/2206.01774

Chandler, D. M. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, *19*(1), 011006. https://doi.org/10.1117/1.3267105

Chandler, D., & Hemami, S. (2007). VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, *16*(9), 2284–2298. https://doi.org/10.1109/TIP.2007.901820

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, *8*(32), eabq6147. https://doi.org/10.1126/sciadv.abq6147

Finnane, A., Curiel-Lewandrowski, C., Wimberley, G., Caffery, L., Katragadda, C., Halpern, A., Marghoob, A. A., Malvehy, J., Kittler, H., Hofmann-Wellenhof, R., Abraham, I., Soyer, H. P., & On behalf of the International Society of Digital Imaging of the Skin (ISDIS) for the International Skin Imaging Collaboration (ISIC). (2017). Proposed technical guidelines for the acquisition of clinical images of skin-related conditions. *JAMA Dermatology*, *153*(5), 453. https://doi.org/10.1001/jamadermatol.2016.6214

Ghadiyaram, D., & Bovik, A. C. (2016). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, *25*(1), 372–387. https://doi.org/10.1109/TIP.2015.2500021

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021, April 20). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. Retrieved April 28, 2024, from http://arxiv.org/abs/2104.09957

Gu, K., Xu, X., Qiao, J., Jiang, Q., Lin, W., & Thalmann, D. (2020). Learning a unified blind image quality metric via on-line and off-line big training instances. *IEEE Transactions on Big Data*, *6*(4), 780–791. https://doi.org/10.1109/TBDATA.2019.2895605

Gu, K., Zhai, G., Yang, X., & Zhang, W. (2014). Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting*, *60*(3), 555–567. https://doi.org/10.1109/TBC.2014.2344471

Hoffmann, C. P., Lennerts, S., Schmitz, C., Stölzle, W., & Uebernickel, F. (Eds.). (2016). *Business innovation: Das st. galler modell*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07167-7

Jalaboi, R., Winther, O., & Galimzianova, A. (2023, January 23). Explainable image quality assessments in teledermatological photography. Retrieved April 23, 2024, from http://arxiv.org/abs/2209.04699

Jayaraman, D., Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). Objective quality assessment of multiply distorted images. *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 1693–1697. https://doi.org/10.1109/ACSSC.2012.6489321

Jiang, S. W., Flynn, M. S., Kwock, J. T., & Nicholas, M. W. (2022). Store-and-forward images in teledermatology: Narrative literature review. *JMIR Dermatology*, *5*(3), e37517. https://doi.org/10.2196/37517

Lin, H., Hosu, V., & Saupe, D. (2019). KADID-10k: A large-scale artificially distorted IQA database. *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. https://doi.org/10.1109/QoMEX.2019.8743252

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., & Zhang, L. (2017). Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, *26*(2), 1004–1016. https://doi.org/10.1109/TIP.2016.2631888

Min, X., Ma, K., Gu, K., Zhai, G., Wang, Z., & Lin, W. (2017). Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Transactions on Image Processing*, *26*(11), 5462–5474. https://doi.org/10.1109/TIP.2017.2735192

Ni, Z., Ma, L., Zeng, H., Chen, J., Cai, C., & Ma, K.-K. (2017). ESIM: Edge similarity for screen content image quality assessment. *IEEE Transactions on Image Processing*, *26*(10), 4818–4831. https://doi.org/10.1109/TIP.2017.2718185

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., & Jay Kuo, C.-C. (2015). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, *30*, 57–77. https://doi.org/10.1016/j.image.2014.10.009

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Astola, J., Carli, M., & Battisti, F. (2009). TID2008 a database for evaluation of full- reference visual quality assessment metrics.

Sheikh, H., Sabir, M., & Bovik, A. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, *15*(11), 3440–3451. https://doi.org/10.1109/TIP.2006.881959

Sun, W., Zhou, F., & Liao, Q. (2017). MDID: A multiply distorted image database for image quality assessment. *Pattern Recognition*, *61*, 153–168. https://doi.org/10.1016/j.patcog.2016.07.033

Thomas, B. F. (1998). The validity and practicality of sun-reactive skin types i through VI. *Archives of Dermatology*, *124*, 869–871. https://doi.org/https://doi.org/10.1001/archderm.124.6.869

Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., & Hakkinen, J. (2015). CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, *24*(1), 390–402. https://doi.org/10.1109/TIP.2014.2378061

Vodrahalli, K., Daneshjou, R., Novoa, R. A., Chiou, A., Ko, J. M., & Zou, J. (2020, October 1). TrueImage: A machine learning algorithm to improve the quality of telehealth photos. Retrieved April 23, 2024, from http://arxiv.org/abs/2010.02086

Ward, A., Li, J., Wang, J., Lakshminarasimhan, S., Carrick, A., Campana, B., Hartford, J., S, P. K., Tiyasirichokchai, T., Virmani, S., Wong, R., Matias, Y., Corrado, G. S., Webster, D. R., Siegel, D., Lin, S., Ko, J., Karthikesalingam, A., Semturs, C., & Rao, P. (2024, February 28). Crowdsourcing dermatology images with google search ads: Creating a real-world skin condition dataset. Retrieved April 28, 2024, from http://arxiv.org/abs/2402.18545

Yang, H., Yuming Fang, Lin, W., & Wang, Z. (2014). Subjective quality assessment of screen content images. *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 257–262. https://doi.org/10.1109/QoMEX.2014.6982328

Zhou, W., & Alan, C. B. (2007). *Modern image quality assessment*. Springer Cham. https://link.springer.com/book/10.1007/978-3-031-02238-8