Contents

Chapter 1

Introduction

1.1 Background and Problem Statement

In recent years, the way people seek dermatological advice has changed significantly, mainly due to the COVID-19 pandemic. Teledermatology, a branch of telemedicine, has become a popular way to diagnose and manage skin conditions remotely. Telemedicine uses telecommunications technology to provide healthcare services from a distance, allowing patients to consult with healthcare providers without needing to be physically present.

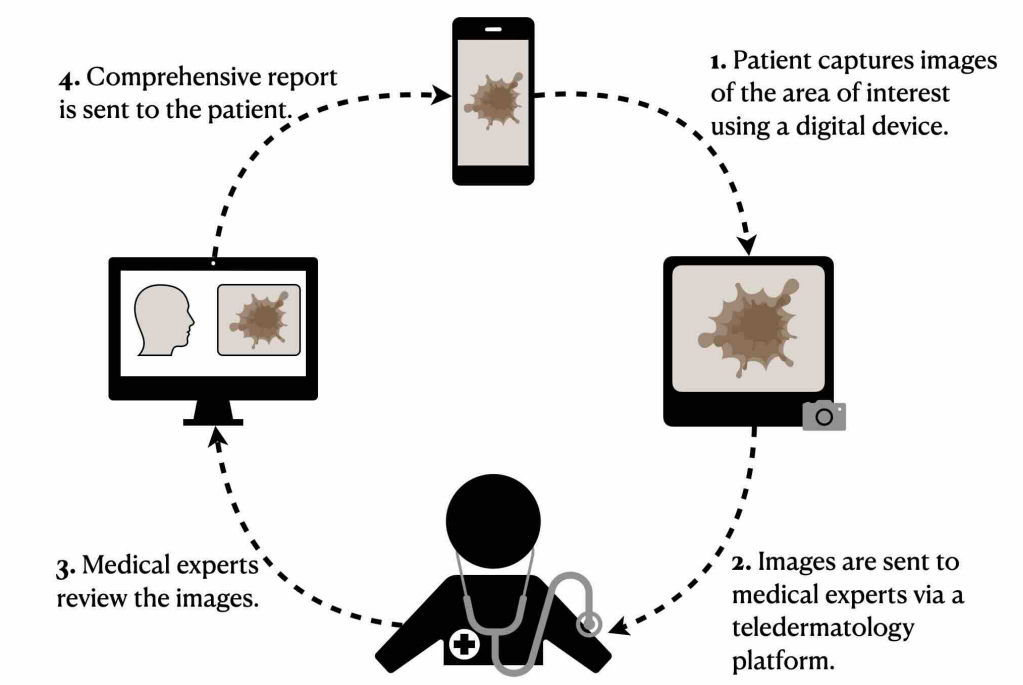


Figure 1.1: This diagram shows the simplified process of a teledermatology consultation, starting from the patient capturing images of their skin condition to receiving a detailed medical report.

In teledermatology, patients use mobile applications to take pictures of their skin conditions with everyday devices like smartphones and tablets. These images are then sent to dermatologists

for analysis, eliminating the need for face-to-face appointments. Figure 1.1 shows each step in the teledermatology process, highlighting the essential role of image quality in ensuring accurate diagnosis and effective patient care.

However, the success of teledermatology relies heavily on the quality of the images patients capture. Despite the convenience of modern technology, many images sent by patients do not meet the necessary standards. Issues such as poor lighting, blurred images, and inadequate representation of skin conditions can greatly limit a dermatologist's ability to make accurate diagnoses. These challenges with image quality reduce the effectiveness of teledermatology.

This common problem highlights the critical need to improve the quality of images taken through mobile applications. This thesis aims to address this problem by developing and implementing automated image quality assessment techniques to enhance the reliability and effectiveness of teledermatology.

1.2 Objectives of the Thesis

The primary goal of this thesis is to develop and evaluate automated methods for assessing image quality within the context of teledermatology. The objectives are varied, starting with a comprehensive literature review of image quality assessment methods from the general imaging domain to determine their suitability for teledermatology applications. This thesis also aims to select appropriate quality metrics, apply these methods to relevant dermatological datasets, and create a reproducible repository for future research.

The specific objectives of this thesis are detailed as follows:

- An extensive review of the literature on image quality assessment (IQA) methods, focusing on their application in teledermatology.
- Identifying and selecting image quality metrics that are 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.

Achieving these objectives will greatly improve the efficiency and accuracy of teledermatology services by creating a way to assess image quality. This improvement will streamline workflows, save time, and reduce frustration in teledermatology. By providing effective tools and methods for evaluating the quality of patient images remotely, this research will ultimately lead to better diagnostic accuracy and overall patient care in remote dermatological consultations.

1.3 Organisation of this Thesis

This thesis is structured into six 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 image quality assessment (IQA) and teledermatology. Chapter 3 details the methodologies, including those used in the literature review and those specific to IQA and teledermatology. In Chapter 4, the experiments conducted are described, showing

the approaches taken, along with the metrics used. Chapter 5 presents the results of these investigations. Finally, Chapter 6 concludes the thesis, summarizing the findings and suggesting directions for future research.

All figures and tables in this thesis are created by the author unless otherwise referenced. If any code is referenced, the path or module is provided in the footnotes.

Chapter 2

Literature Review

2.1 Image Evaluation

There are three ways to evaluate an image: assessing its quality, aesthetics, or fidelity. 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 pertains 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 various types of image degradation. The following subsections will discuss common distortions, datasets that contain these distortions, and the state-of-the-art (SOTA) methods in IQA. But first, the two ways to assess quality are mentioned.

2.1.1 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 a 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 image's quality based on their subjective experience.
- Paired Comparison: In this method, human observers are presented with two images: a 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. 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, the scorer's 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 the human response to an image is the ultimate measure of its quality.

2.1.2 Objective Quality Assessment

Objective quality assessment relies on mathematical algorithms rather than human judgment to evaluate image quality. This approach uses our understanding of human vision system attributes to develop mathematical equations that measure quality, even though not all methods rely on these attributes. Essentially, it involves comparing data points or features extracted from images to determine quality. This assessment is mainly categorized into three methods based on the reference data used: Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA), and No-Reference IQA (NR-IQA).

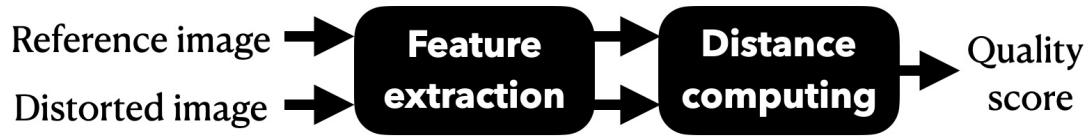


Figure 2.1: General framework of FR-IQA algorithms. Features are extracted from both images, and then the feature distance is calculated.

Full-Reference IQA (FR-IQA) involves a comprehensive 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.

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 thorough comparison of FR-IQA and the independence of NR-IQA, which will be mentioned later, reducing computational demands while still providing meaningful quality assessments based on partial reference data.

Both FR-IQA and RR-IQA utilize two methods to analyze quality:

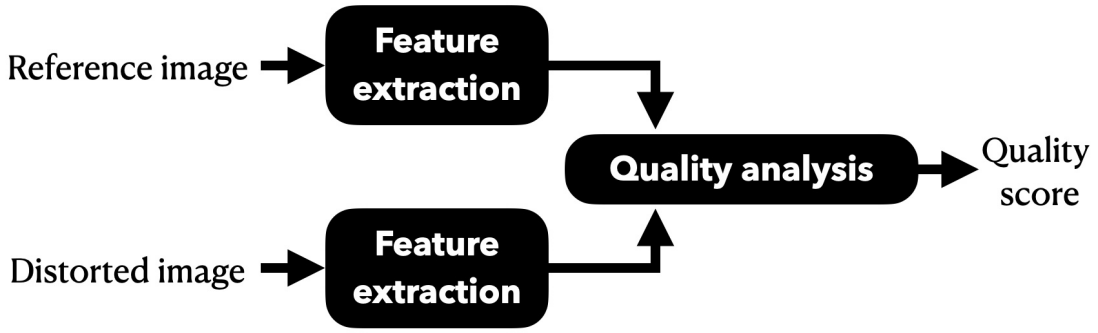


Figure 2.2: General framework of RR-IQA algorithms. Features of the reference and distorted images are extracted and used collectively to compute the quality.

- Spatial-Based Analysis: This method compares images pixel by pixel or region by region, offering straightforward interpretation and efficient computation. However, it may not fully align with how humans process images, lacking 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 robust, it is complex and computationally intensive.

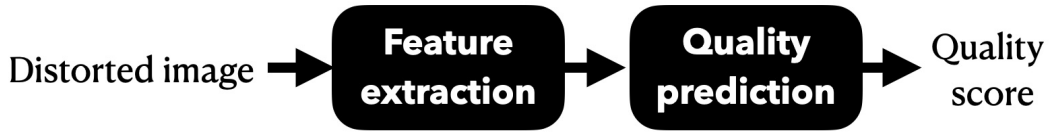


Figure 2.3: General framework of NR-IQA algorithms.

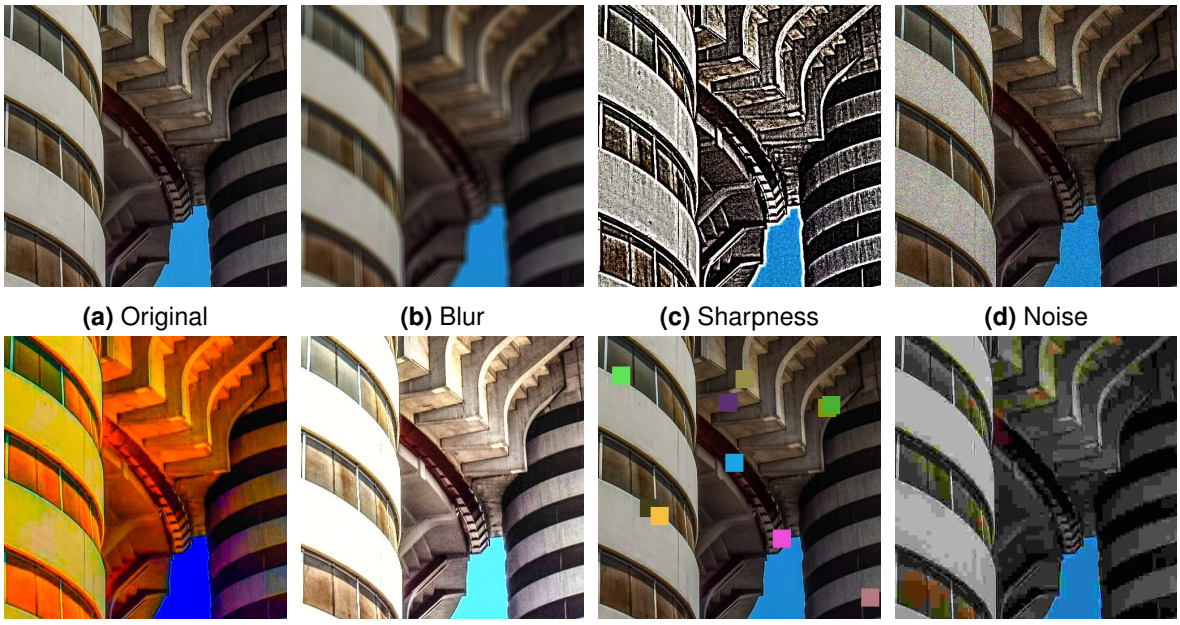
No-Reference IQA (NR-IQA) does not rely on any reference image. Instead, it analyzes the distorted image alone by extracting features indicative of quality (see Figure 2.3). This method is particularly useful when no reference images are available, such as in many practical applications of teledermatology. NR-IQA can be tailored to address specific types of distortions or designed for general-purpose quality assessment, providing versatility across various domains.

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.

2.1.3 Common Distortions in Image Quality Assessment

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.

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 camera's 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 various applications, including teledermatology.

2.1.4 Benchmark Datasets for IQA

Benchmark datasets play an important role in advancing Image Quality Assessment (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.

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 Appendix B. 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 reproducible.

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

The current state-of-the-art in Image Quality Assessment (IQA) is ARNIQA (Agnolucci et al., 2023), with version 2 released on November 4, 2023. ARNIQA (leArning distoRtion maNifold for Image Quality Assessment) represents a major advancement in NR-IQA. This technology aims to measure image quality based on human perception, even without a reference image.

Overview: ARNIQA is developed using a self-supervised learning approach. It learns a comprehensive model of all possible image distortions, focusing on the types and quality of distortions rather than the content of the images themselves. This makes it highly adaptable across various domains where image content can differ significantly.

Key Features:

1. Image Degradation Model: ARNIQA can synthetically degrade images through up to 1.9 billion distinct degradation patterns. This model can apply up to seven different types of distortions simultaneously, covering a wide range of real-world scenarios. Training with such diverse distortions ensures that ARNIQA can accurately assess image quality across various conditions and avoid the need for large labeled datasets.

2. SimCLR Framework: At the core of ARNIQA is the SimCLR (Simple Framework for Contrastive Learning) framework. This framework learns meaningful representations of image quality by comparing different versions of the same image and focusing on their similarities and differences. SimCLR constructs positive pairs by applying the same distortion settings to two different images, ensuring that the model concentrates on the distortions rather than the content. To further enhance the model's ability to distinguish between different types of distortions, SimCLR introduces subtle variations by downsampling images before cropping

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 first, 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

and applying distortions, creating hard negative examples. These examples help the model differentiate between similar-looking images with different types of degradation. By using this approach, the SimCLR framework ensures that ARNIQA effectively learns to recognize and assess various distortions, improving its ability to provide accurate image quality assessments (see Figure 2.5).

3. Linear Regressor: A Linear Regressor maps the features extracted from the backbone to a qualtiy score ranging from 0 to 1. This score reflects the relative quality of the image based

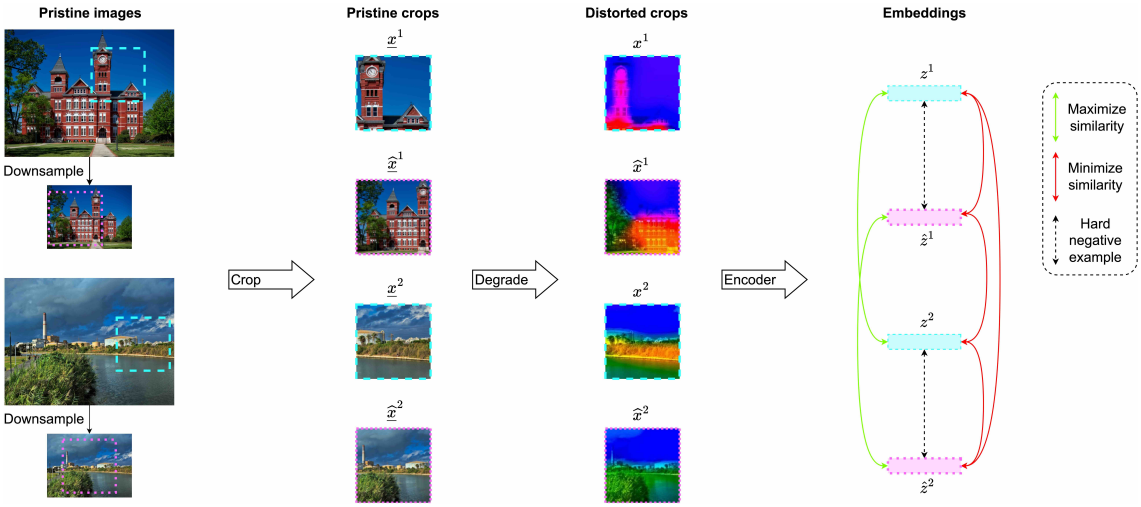


Figure 2.5: Overview of the training strategy for ARNIQA. Two pristine images are cropped and equally degraded. The model maximizes the similarity of their embeddings while minimizing the similarity to embeddings from degraded crops of half-scale versions of the original images. This process creates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

on the distortions present.

Training Strategy: ARNIQA's training strategy involves two main phases:

1. Encoder Pre-training: The model is first trained on a large set of unlabeled images that are synthetically degraded. This helps the encoder learn features related to different types and levels of image degradation.

2. Regressor Training:In the second phase, a specific regressor is trained using the Mean Opinion Scores (MOS) of images. This step translates the learned features into actual quality scores.

Advantages: ARNIQA achieves high performance with only up to 0.5% of the data needed for training compared to other SOTA methods because of its focus on distortion patterns rather than image content. It provides reliable and consistent quality assessments across a wide range of distortions and severities, demonstrating its robustness. Additionally, ARNIQA is particularly suitable for teledermatology as it can handle varying image quality resulting from different lighting conditions, camera quality, and patient handling.

2.1.6 Challenges and Opportunities in Image Quality Assessment

maybe combine it with the other challenges and opportunities section

One major challenge in IQA is that assessing image quality can be very subjective. Different people can have different opinions on what looks good or bad, making it hard to create standard measures. This is especially important in teledermatology, where the quality of images directly affects medical diagnoses. Another challenge is that images can have many types of problems, like blurring, noise, compression artifacts, and color issues. Each problem affects the image in a different way, and it's tough to develop IQA methods that can handle all these issues accurately. Additionally, in many real-world applications, including teledermatology, we often don't have high-quality reference images to compare against. This makes it difficult to evaluate the quality of

distorted images. Therefore, developing methods that don't need reference images (NR-IQA) is essential.

A big opportunity in IQA is the advancement of self-supervised learning techniques. These methods, like those used in ARNIQA, allow models to learn from large amounts of data without needing a lot of labeled examples. This approach saves time and money because it reduces the need for manually labeled data. It also makes it possible to develop high-quality IQA models that can work well even without reference images.

By addressing these challenges and leveraging the opportunities, we can significantly improve how we assess image quality.

2.2 Teledermatology

This section covers teledermatology and highlights the importance of image quality in remote skin evaluations. It starts by explaining what teledermatology is and then discusses why having high-quality images is crucial for accurate diagnoses and treatment.

Next, the quality standards needed for teledermatology images are reviewed, along with public datasets available for research. Different methods used to assess image quality in teledermatology, based on previous studies, are briefly examined. Finally, 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 treatment using telecommunications technology. This is especially helpful for patients in remote areas, ensuring they get timely and effective skin care. There are two main methods used in teledermatology: real-time (synchronous) and store-and-forward (asynchronous).

Real-time teledermatology involves live video consultations between the dermatologist and the patient. This 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 method offers more flexibility since it doesn't require the patient and dermatologist to be available simultaneously (Jiang et al., 2022).

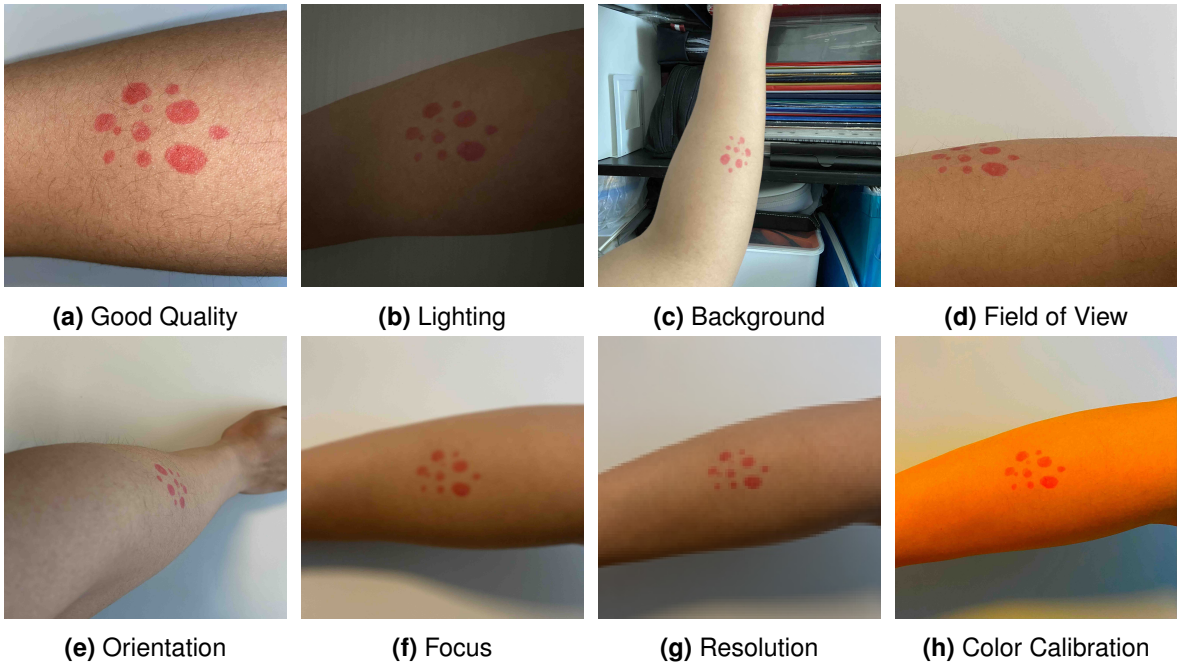
Since store-and-forward teledermatology is the focus, it is important to note that high-quality images are critical in this method. High-quality images are vital in teledermatology because they directly impact the accuracy of remote diagnoses. Poor image quality can lead to incorrect diagnoses or delayed treatment, reducing the benefits of teledermatology. Therefore, ensuring that images meet specific quality standards is crucial for successful teledermatology services.

2.2.2 Quality Criteria for Teledermatology Images

The International Skin Imaging Collaboration (ISIC) has set guidelines for standardizing images in terms of 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. Out of these nine recommended criteria, this thesis will focus on seven key criteria that directly impact image quality. The other two, "Scale and Measurement" and "Image Storage," are not relevant because they are not directly related to image quality. "Scale and Measurement" is less important in this context because it focuses on 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).

The seven key criteria for teledermatology images, shown in Figure 2.6, along with recommendations on how to meet each criterion, are as follows:

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*



(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.

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. *Ensure 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, teledermatology practitioners can ensure 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. The main difference is that dermatology datasets often include more professional images taken in clinical settings, including close-up dermoscopic images which provide detailed views of the skin. In teledermatology, images might be taken by patients using their mobile devices, resulting in more varied quality. However, there are several datasets that can still be useful for teledermatology depending on the specific use case. Here are seven public datasets that can be used for teledermatology:

- **ACNE04**: This dataset focuses on acne severity and lesion counting, containing 1,457 images with detailed annotations for training and testing purposes (Wu et al., 2019).
- **DDI**: Provides 656 high-quality images curated by dermatologists for detailed skin tone evaluation and diagnostic accuracy (Daneshjou et al., 2022).
- **Derm7pt**: Utilizes 1,011 lesion cases to train a neural network for classifying skin lesions and melanoma using the 7-point checklist (Kawahara et al., 2019).
- **Fitzpatrick17k**: Includes 16,577 images annotated for Fitzpatrick skin type across 114 different skin conditions (Groh et al., 2021).
- **Monkeypox Dataset 2022**: Contains approximately 1,905 images focused on monkeypox, useful for developing diagnostic tools (Ahsan et al., 2022).
- **PAD-UFES-20**: Comprises 2,298 clinical images from smartphones, enriched with clinical metadata for comprehensive research (Pacheco & Krohling, 2020).
- **SCIN**: Emerged from a crowdsourcing initiative, this dataset contains 10,408 images capturing a broad spectrum of dermatological conditions (Ward et al., 2024).

These datasets provide valuable images and annotations that help develop and test image quality assessment methods for 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 2021, uses an automated machine learning system to detect poor-quality dermatology images and help patients take better images.This method was developed in response to the disruption caused by many low-quality images submitted by patients in clinical workflows. 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 composed, 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). Another limitation is that it only considers three common distortions.

Released in January 2023, **ImageQX** (Explainable Image Quality Assessments in Teledermatological Photography) 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 photos and validated on 9,874 photos, each annotated by up to 12 board-certified dermatologists. Its main innovation is providing explanations for poor quality, guiding patients on how to take better images. ImageQX is 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. They both address common issues like blur and poor lighting. ImageQX excels in providing detailed feedback on how to improve image quality, while TrueImage focuses on being computationally efficient and interpretable, making it suitable for older smartphones. 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. However, there is still room for improvement in handling complex lighting conditions and ensuring accurate zoom detection.

2.2.5 Challenges and Opportunities in Image Quality Assessment for Teledermatology

Teledermatology faces several challenges similar to those in Subsection 2.1.6. A major issue is the subjectivity of image quality assessment. Different dermatologists might have different opinions on what makes a good image, making standardization difficult. This variability can affect the accuracy of medical diagnoses since the quality of images is crucial. Common problems in teledermatology images include blurring, poor lighting, compression artifacts, and color issues. Each problem affects the image differently, making it hard to develop methods that handle all these issues well. Additionally, high-quality reference images are often unavailable, making it tough to evaluate the quality of patient-taken images accurately. Patients also use various devices and capture images under different conditions, adding to the complexity.

Despite these challenges, there are significant opportunities to improve teledermatology. Collaboration with dermatologists, as seen in methods like ImageQX and TrueImage, can improve IQA models by ensuring they meet the needs of medical professionals. These models can provide real-time, actionable feedback to patients on how to take better images, improving the quality of images submitted for remote consultations.

Chapter 3

Methodology

Based on the insights from Chapter 2, *Literature Review*, this chapter provides an overview of the key ideas and concepts needed to achieve the research objectives. The following sections will explore important concepts related to image quality assessment in teledermatology and explain the reasoning behind this work. Detailed implementation of these steps will be covered in the next chapter.

3.1 Explorative Approach

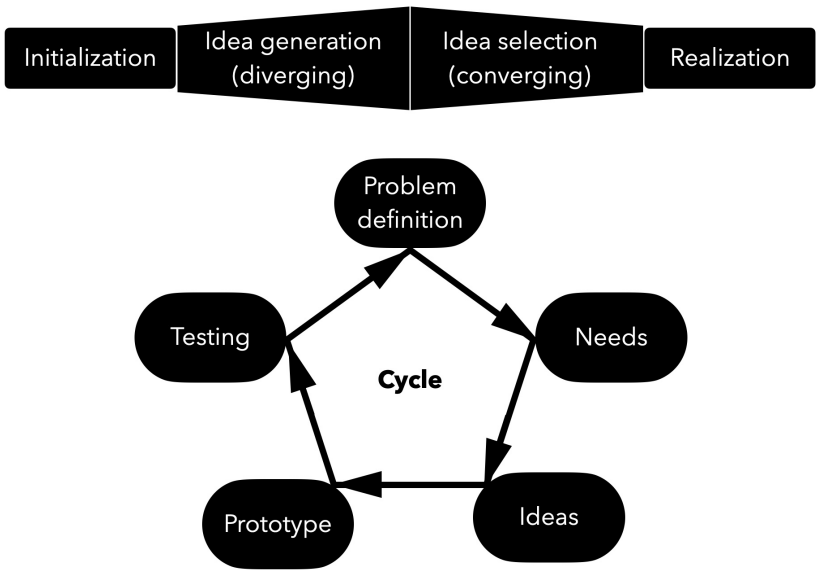


Figure 3.1: Visualization of the explorative approach, including the stages of initialization, idea generation, idea selection, and implementation. The lower part of the figure shows the decision cycles adapted from (Hoffmann et al., 2016).

Teledermatology, especially when focused on Image Quality Assessment (IQA), offers many opportunities for innovation due to the different types of image distortions and ways to address them. To handle this complexity, an exploratory approach was used in this research. This approach is flexible and innovative, allowing for adjustments as new information is discovered, unlike

traditional methods like the waterfall approach.

At the beginning, the problem was broadly defined, allowing for a flexible and adaptive approach. The research progressed through creative problem-solving with multiple learning cycles, refining ideas and methods iteratively. The project was divided into two main phases. In the *Diverging Phase*, the research question's scope was expanded to generate new ideas continuously, based on insights from the ongoing literature review. In the *Converging Phase*, the focus on combining these ideas into clear findings and conclusions, aiming to create a unified understanding of the initial problem. This exploratory model is shown in Figure 3.1, which shows the stages of starting, generating ideas, selecting ideas, and implementation, along with the decision cycles (Hoffmann et al., 2016).

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 established before starting, detailed in the Gantt chart attached to this thesis.

There are three key milestones identified in the first half of the project, each crucial for its success:

Understanding Teledermatology: Gaining a thorough understanding of the field to ensure all subsequent actions are relevant and well-informed.

State of the Art in IQA: Identifying the latest developments in IQA to ensure the methods used are up-to-date and effective.

Availability of Teledermatology Data: Securing access to appropriate datasets for conducting meaningful IQA.

These milestones are essential because each phase of the project relies 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 study was exploratory, so it was not possible to follow a strict, step-by-step process. However, for the individual key steps, I took a systematic approach to stay organized and ensure that each step was done in the right order:

3.3.1 Literature Review

First, I began by getting an overview of my research field. As I was new to the domain of teledermatology and dermatology, this initial step was crucial. By researching and reading relevant literature, I gradually built a solid understanding of the field. Next, I identified the core topics related to my research objectives. Once I had these key topics, I carefully selected the databases to search, focusing on those most relevant to my field: PubMed[1], Google Scholar[2], IEEE Xplore[3],

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

Connected Papers[4], and Papers with Code[5]. Using these databases, I applied search filters to narrow down the results, such as limiting the search to articles published after 2020.

I reviewed the titles of the search results and opened the ones that seemed interesting. After that, I read the abstracts to determine their relevance. Depending on the relevance of the abstract and some of the figures, I decided whether to read the full paper. Additionally, for state-of-the-art methods, I focused on finding and reading papers that had published their code and model weights if models were trained. This systematic approach ensured that my literature review was thorough and focused on the most relevant and up-to-date research.

3.3.2 Data Collection and Preparation

revisit this section

In searching for a suitable dataset to evaluate image quality in teledermatology, a major challenge was the lack of Mean Opinion Score (MOS) or Differential Mean Opinion Score (DMOS) in teledermatology datasets, as commonly found in traditional IQA datasets mentioned in Subsection 2.1.4. This scarcity is due to the resource-intensive nature of labeling images in the medical field.

To address this gap, I created a distortion pipeline that synthetically distorts images based on the seven criteria defined in Subsection 2.2.2. Each type of distortion has five levels of severity, with the 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 according to the severity and type of distortion applied, creating a dataset that not only includes the distorted images but also features precise annotations regarding their quality. This allowed me to artificially create labels for my images. For this, I needed good quality images to start with. I chose two datasets: the SCIN dataset for its relevance and uniqueness, and the Fitzpatrick17k dataset to complement the SCIN dataset.

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 (Ward et al., 2024). I chose this dataset because it includes conditions that patients are likely to consult about via teledermatology platforms before visiting traditional healthcare settings. 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), so I used the Fitzpatrick17k dataset only for training purposes to complement the SCIN dataset.

In total, I selected 475 high-quality images from the Fitzpatrick17k dataset and another 475 high-quality images from the SCIN dataset for training and evaluation. Additionally, I randomly chose 200 test images from the SCIN dataset and 70 independent high-quality images from SCIN for testing. The 70 high-quality images previously selected from the SCIN dataset are fed through the distortion pipeline to introduce distortions, allowing for a consistent basis to test the model against the same types of distortions. I also labeled 200 test images, scoring each one on the seven quality criteria to ensure the model's performance can be compared to human evaluation.

[4] https://www.connectedpapers.com
[5] https://paperswithcode.com

3.3.3 Feature Extraction

Feature extraction is the next important step where the backbone from ARNIQA is used to identify key features from the distorted images. These features capture the patterns of distortions that affect image quality. The extracted features and the generated labels are then used to train different models, including Extreme Gradient Boosting (XGBoost) regressor, XGBoost classifier, and Multi-Layer Perceptron (MLP) regressor and MLP classifier, to see which one works best for assessing image quality.

3.3.4 Training and Validation

The training of the models is based on the prepared training images. Since I am generating labels and distorted images, I am not restricted by the original amount of images. I can run the images through the distortion pipeline multiple times, creating various versions of distortions from the original images. The models are then trained with these images to develop their ability to assess image quality. Validation is done in parallel with training by using a portion of the data as a validation set. This helps evaluate and monitor the performance of the models.

3.3.5 Testing and Experiments

After completing the training, the models are evaluated using independent test data. There are two test sets used in this evaluation. The first test set consists of 70 images that were previously selected from the SCIN dataset and fed through the distortion pipeline to introduce similar distortions. This set is used to assess the actual performance and reliability of the model against consistent types of distortions. The second test set includes 200 images from the SCIN dataset, which I labeled myself, scoring each one on the seven quality criteria to ensure the model's performance can be compared to human evaluation.

3.3.6 Evaluation Metrics

The evaluation is conducted using defined metrics such as Mean Absolute Error (MAE), R-squared (R^2), Spearman's Rank Order Correlation Coefficient (SRCC), and Cohen's 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 model's predictions are on average. A lower MAE indicates better model performance.

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 model's predictions. A higher R^2 means better model performance. Using MAE and R^2 together gives a clear picture of the model's 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. For example, 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.

Cohen's Kappa measures how well the model's predictions agree with the actual labels. Unlike SRCC, which focuses on ranking, Cohen's Kappa evaluates the exact agreement between predictions and actual labels. Also unlike simple accuracy, which only looks at the proportion of correct predictions, Cohen's 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.

3.3.7 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. Additionally, potential applications of the developed image quality assessment models in teledermatology are considered, along with the opportunities and challenges that arise from these applications.

Chapter 4

Implementation

This chapter explains the detailed implementation of the methods described in Chapter 3. It covers the specific processes, experiments, and analyses conducted. 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 initial stages of the implementation, focusing on the selection and preparation of the image datasets used in the study.

4.1.1 Image Filtering and Selection

The first step in preparing the images involves carefully choosing good quality pictures from the SCIN and Fitzpatrick17k datasets. This selection is done manually to ensure that each image is clear and useful for clinical purposes. The primary focus during selection is on images that are well-framed and free of any distortions that might affect their usefulness in diagnosis.

Each selected image is checked to ensure it is not blurred, as clear images are crucial for accurate diagnosis. Additionally, it is important that the images have proper lighting and true contrast, meaning they should not be too bright or too dark. Proper lighting and contrast help in accurately showing the skin's condition. Lastly, the images must represent realistic skin tones and colors because accurate color representation is critical for correct diagnoses. Some pictures from the dataset are included in the appendix for reference (see Appendix).

maybe include some pictures in appendix

4.1.2 Labeling of the Test Set

The labeling process involves manually scoring 200 images from the SCIN dataset. Of these 200, around 50 are good quality images, which I wanted to represent in the test set as well. Each image is scored on a scale from 0 to 1 for each criterion, where 0 indicates no distortion and 1 indicates extreme distortion. This manual labeling is done using a custom Python script [1] , which displays each image and prompts the user to enter scores for each distortion criterion. The scores are collected in a structured format and stored in a JSON file for later analysis.

correct this later

[1] src/create_labels.ipynb

This structured approach ensures consistent and thorough evaluation of each image. I did the labeling myself, using an absolute categorical rating method as described in Subsection 2.1.1. This method is very time consuming and requires significant effort from the evaluator. My labeling process involved scoring 200 images on 7 criteria each, resulting in 1400 labels. To ensure accuracy and avoid rushing, I deliberately spread out the labeling over multiple sessions.

Visualization of Label Distribution for the Test Set

To understand the distribution of labels and how often distortions occur across different criteria, see Figure 4.1. These histograms are useful for visualizing the prevalence and severity of distortions in the dataset. The histograms are plotted with 5 bins for each criterion, where the first bin indicates no distortion, and the remaining bins represent increasing levels of distortion severity for that type.

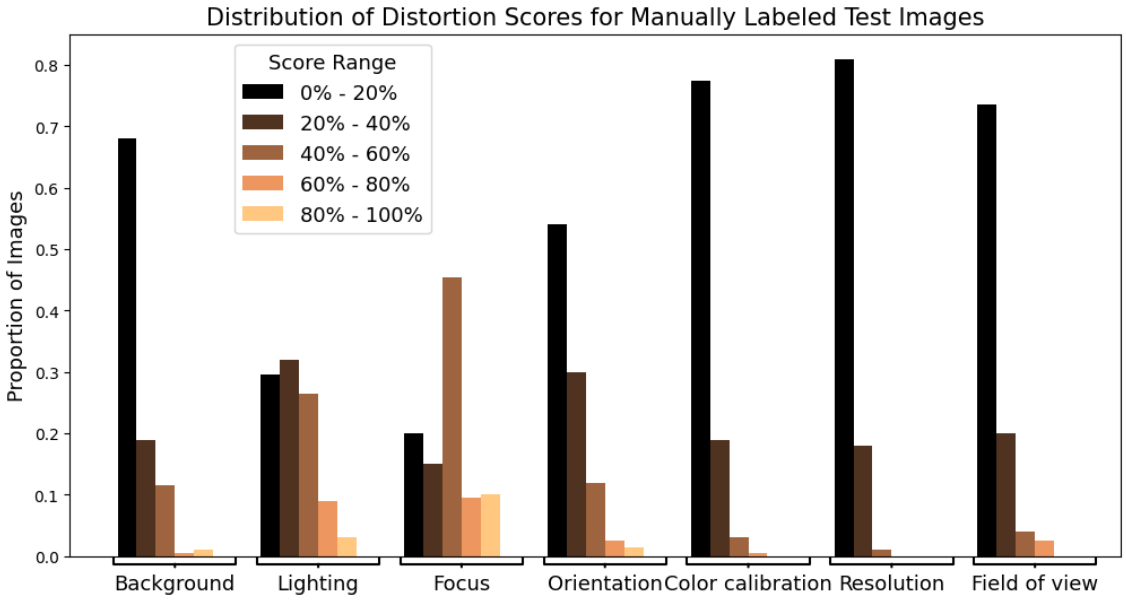


Figure 4.1: Histograms showing the distribution of distortion scores for each quality criterion.

In Figure 4.1 most criteria show a right-skewed distribution, indicating that higher levels of distortions are less common. This is evident in the criteria such as orientation, color calibration, background, resolution, and field of view. In contrast, the distributions for lighting and focus are more symmetrical, suggesting a more even spread of distortion severity levels. Since an image can have multiple distortions at once, it was difficult to separate them individually. For example, when an image is dark due to lighting issues, it becomes hard to judge other factors like focus, resolution, background, or color accuracy. These findings highlight the need to handle multiple distortions together during model training. They also point out the challenges in accurately labeling and assessing images that have several overlapping distortions.

4.2 Distortion Pipeline

The distortion pipeline is central to simulating realistic image quality issues in teledermatology. Each quality criterion has multiple types of distortions, each having five levels of intensity, increasing in severity. All distortion types begin at zero, indicating no distortion applied, and progress to higher values that represent increasing levels of the specified distortion. Visual representations of the types of degradations at different ranges for each quality criterion are provided in Appendix A.

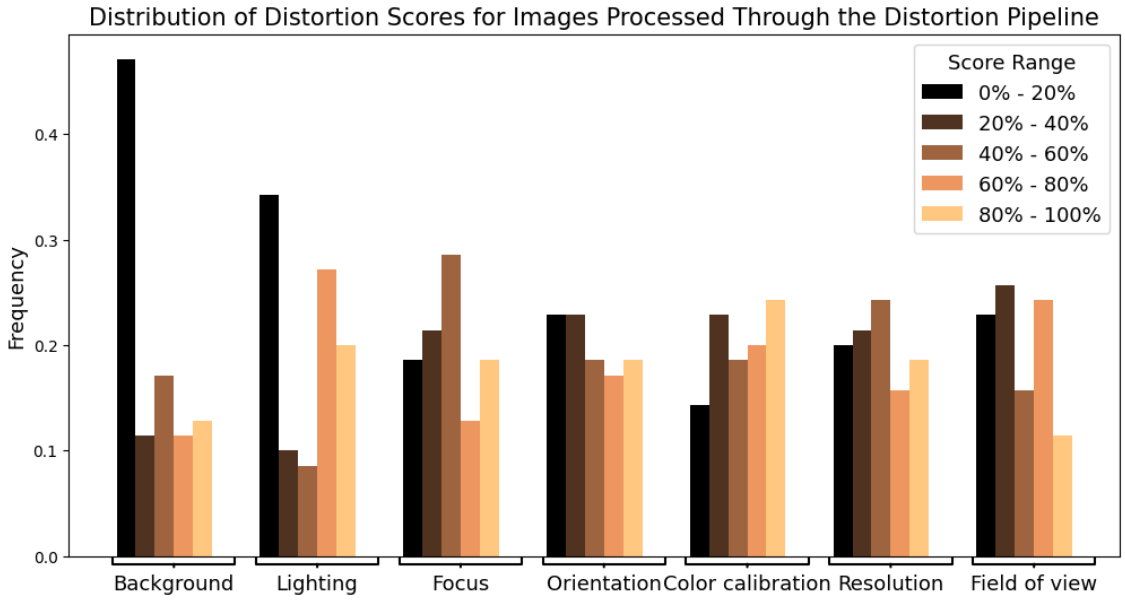


Figure 4.2: Histograms showing the distribution of distortion scores for each quality criterion.

4.2.1 Distortion Types

Here, each distortion 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.
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 my own 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. This gives the effect that the camera is not perpendicular to the skin, as if the camera was not held straight. For the resolution distortion, it was done by first downsampling the image to a lower resolution and then upsampling it back to its original size. This process simulates the effect of low-quality images by introducing pixelation and a loss of detail, similar to what happens when a low-resolution image is enlarged. For the field of view distortion, the image is cropped from the left corner to reduce the visible area. Normally, in good quality images, the skin lesion is centered. By cropping the corner, the lesion moves to the bottom right, simulating poor framing or incomplete capture of the lesion area. Lastly, the background distortion involved segmenting the skin from the background and depending on the amount of background present, color blocks are added to create a noisy background. This makes the background look noisy and cluttered, which can distract the model from focusing on the skin. 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 involves several key steps to create many realistic set of distorted images, which helps train and test the image quality assessment model.

For each image, the RGB version is taken 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. Distortions are then applied in a specific sequence (see Figure 4.3(b)) to ensure realistic simulation. The background distortion is applied first because it depends on identifying the skin area in the undistorted image. If the images have less than 10% background in proportion to skin, no color blocks are added in the background. Therefore, the range value of 0 is used as

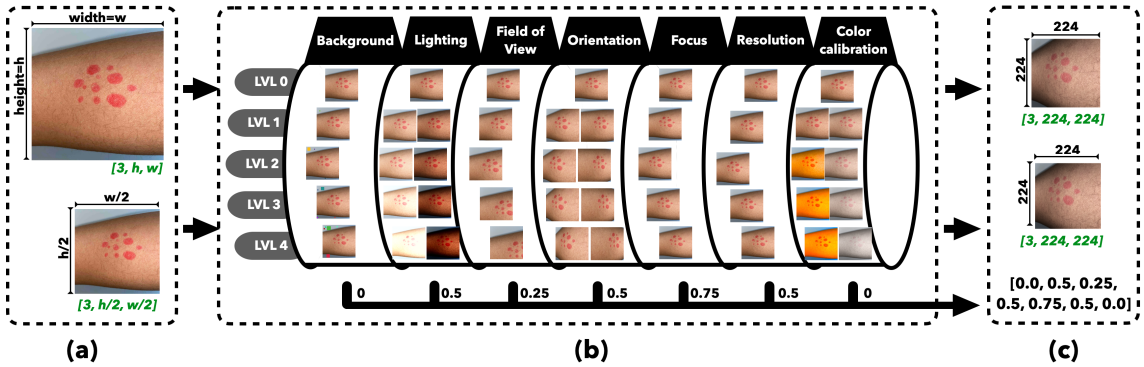


Figure 4.3: Distortion pipeline for generating training images with different levels of distortion. (a) shows the original image and its downscaled version. (b) illustrates the distortion pipeline where a type of distortion and a random level for each criterion are selected, with the corresponding mapped 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 7 distortion values for the image.

value. After that, other distortions are applied based on randomly chosen severity ranges. This ensures a variety 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 of ARNIQA (Agnolucci et al., 2023). 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 makes sure the images are processed consistently, as the model expects the images to have these properties, even though it might make the images look a bit different.

The severity of each applied distortion is mapped to a value between 0 and 1. This is done by taking the minimum and maximum possible values of the distortion and scaling the actual distortion value within this range. This standardized representation allows for consistent training and evaluation of the model later on.

This process can generate 3'750'000 possible combinations of distorted images because of the random selection of distortion types and severity levels. This highlights the robustness and adaptability of the pipeline. 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.

4.4 Feature Extraction with the ARNIQA Backbone

After creating the distortions and their half-scaled versions with mapped labels, the next step is to use the pretrained backbone from ARNIQA, which is loaded via *torch.hub*. This pretrained model has already learned useful features from a large dataset, and these features are transferred to our specific task, a process known as transfer learning. This approach saves time and computational resources while improving the performance of the image quality assessment model.

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 ensures a comprehen-

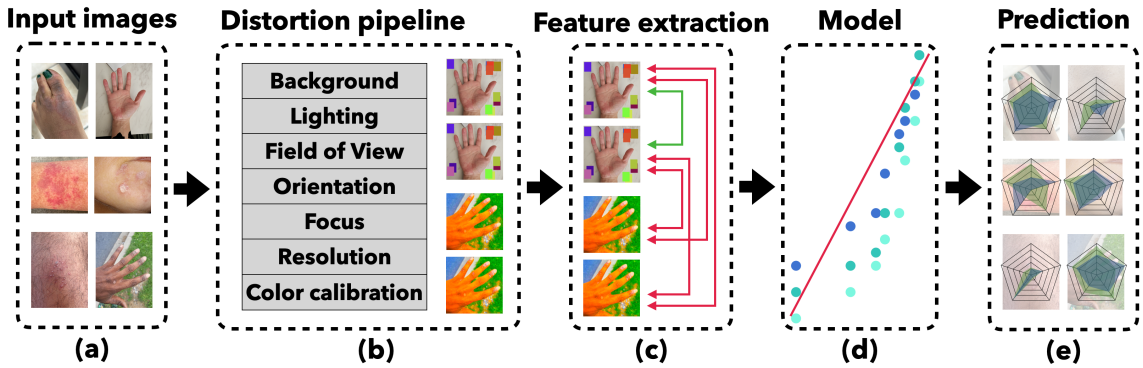


Figure 4.4: Overview of the entire process for training and evaluating the image quality assessment model. (a) shows the album of input images. (b) depicts the distortion pipeline that randomly distorts images across seven criteria at five severity levels. (c) illustrates the feature extraction using the backbone from ARNIQA, which uses the SimCLR framework. (d) shows a scatter plot representing the model training process, with a diagonal red line indicating the fit. (e) presents the prediction results, comparing the model's output to the actual labels.

sive understanding of image quality variations.

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

4.5 Model Selection and Training

Hardware and Resources
Training was done using an NVIDIA A16 GPU, which has 16GB vRAM, 1280 CUDA Cores, 40 Tensor Cores, and 512 GB RAM. This setup ensured efficient use of resources and sped up the training process. This information is important because I was limited to using a batch size of 10 for extracting features from the ARNIQA backbone. Larger batch sizes could potentially improve feature extraction quality because the SimCLR framework in ARNIQA benefits from larger batches, but this was not tested due to resource limitations. Nonetheless, I could work with this setup effectively.

Data Preparation and Splitting
The dataset was expanded by multiplying the original images by factors of 4, 8, 16, 32, or 64 to examine how increasing the dataset size affected performance. This approach tested the hypothesis that larger datasets would lead to better performance. Indeed, as shown in Figure 4.5 and Figure 4.6, where the overall SRCC for XGBRegressor and MLP Regressor, as well as XGBClassifier and MLP Classifier, improved with an increasing number of distortions.

However, it's important to note that larger datasets also require more training time. Even with smaller datasets, it is possible to achieve good performance by finding optimal parameters. After expanding the dataset, the images were split into training and validation sets, with the training set containing 75% of the images and the validation set containing the remaining 25%. This split allowed the model to train on most of the data while still having a separate set for evaluating its performance.

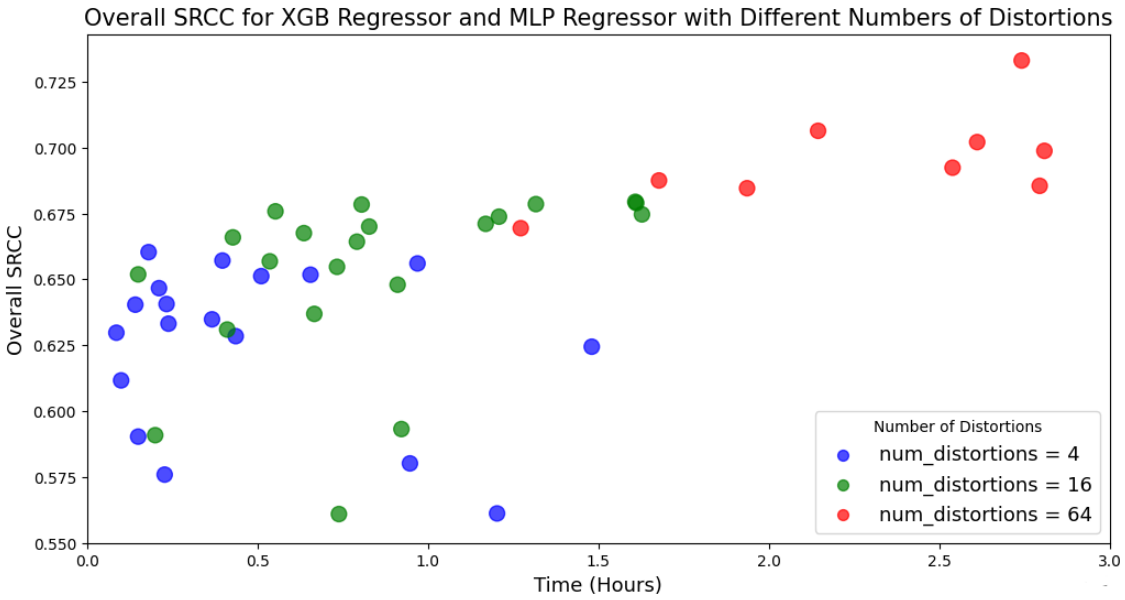


Figure 4.5: Overall SRCC for XGB Regressor and MLP Regressor with different numbers of distortions. The x-axis shows the time it took to train, and the y-axis shows the SRCC values, demonstrating better performance with larger datasets.

Modle Selection

Four different multi-output models were experimented with: XGBRegressor, XGBClassifier, and both MLP Regressor and MLP Classifier. These models were chosen for their strengths in handling complex relationships and their ability to predict multiple outputs at once. Multi-output models can predict multiple quality criteria simultaneously, making them particularly suitable for this task, where assessing multiple aspects of image quality is important.

The models were trained individually on the SCIN and Fitzpatrick datasets and also on a combination of both datasets to assess performance. This approach, known as cross-dataset evaluation, helps to understand how well the models generalize across different datasets. By training on one dataset and evaluating on another, such as training on SCIN and evaluating on Fitzpatrick, the robustness and generalizability of the models can be tested. Combining both datasets and evaluating on each individually further helps in understanding how the models perform with a more diverse set of images, ensuring that the models are not overfitting to a particular dataset.

One important aspect of regressor training is handling continuous scores. If continuous scores were compared directly to fixed numbers from the distortion pipeline, there would always be some minor errors. To minimize these errors and accurately calculate metrics like rank correlation and Cohen's Kappa, the regressor predictions were clipped to the range of 0 to 1. The scores were then categorized into severity levels using a function[2] that converts continuous scores to discrete categories based on defined thresholds. This process, known as discretization, helps in effectively categorizing the severity levels and reducing errors in score comparison. Additionally, the discretization function was also used for the classifier models to convert continuous scores to categorical ones because these models require categorical labels as input.

[2] from utils.utils_data import binarize_scores

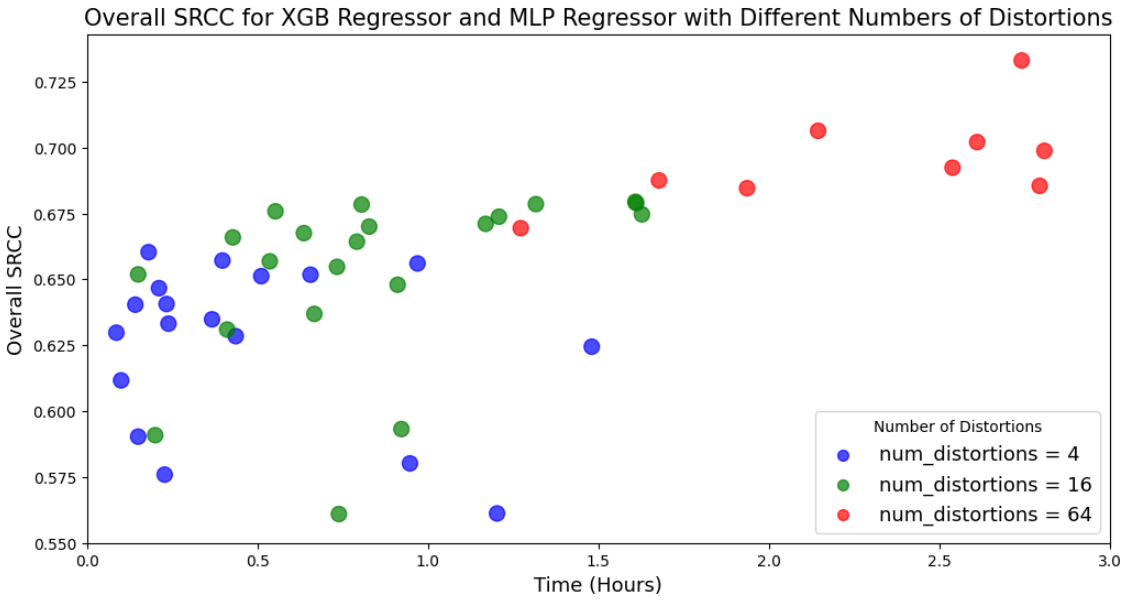


Figure 4.6: Overall SRCC for XGB Classifier and MLP Classifier with different numbers of distortions. Similar to Figure 4.5, this figure shows improved performance with larger datasets.

4.5.1 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 Spearman's Rank Order Correlation Coefficient (SRCC).

The following table shows the configurations used in the hyperparameter sweep:

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

L2 regularization and subsampling were used to improve the generalization of the model and prevent it from memorizing the training data. L2 regularization helps to avoid large coefficients, and subsampling trains the model on different subsets of data to reduce variance.

[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)
objective	multi:softprob (specific to XGBClassifier)
n_jobs	16 (specific to XGBRegressor)
booster	gbtree (specific to XGBClassifier)
eval_metric	’mlogloss’, ’merror’, ’auc’] (specific to XGBClassifier)

4.6 Model Testing

The best model was tested on two specific sets of images to evaluate its performance. The first set 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 consisted of 200 images with authentic distortions, allowing for a comparison of the model’s performance with my manual evaluations.

For the 200 authentic images, they were half-scaled, resized to 224x224 pixels, and normalized. These preprocessed images were then passed through the ARNIQA backbone to extract features, and their scores were taken from a JSON[4] file where my labels were stored. For the 70 synthetically distorted images, features were extracted from the backbone, and the scores and features were saved in a *.npy*[5] file to ensure reproducibility and easier comparison across different tests.

In addition to testing the model, I also validated the effectiveness of my approach by using ARNIQA itself to score both sets of images. ARNIQA provided quality scores ranging from 0 to 1, where higher scores meant better image quality. This comparison helped verify whether adding synthetic distortions improved the model’s performance.

[4] src/test_200/scores.json
[5] src/test_70/embeddings

Chapter 5

Results and Analysis

In this chapter, the performance of the trained models is analyzed and discussed. The main focus is on the final MLP regressor model, which was found to be the best performing model across multiple criteria. The selection of the MLP regressor is supported by the results shown in Figure 5.1, which presents a parallel coordinate plot comparing the best-performing models across all seven criteria and the overall SRCC.

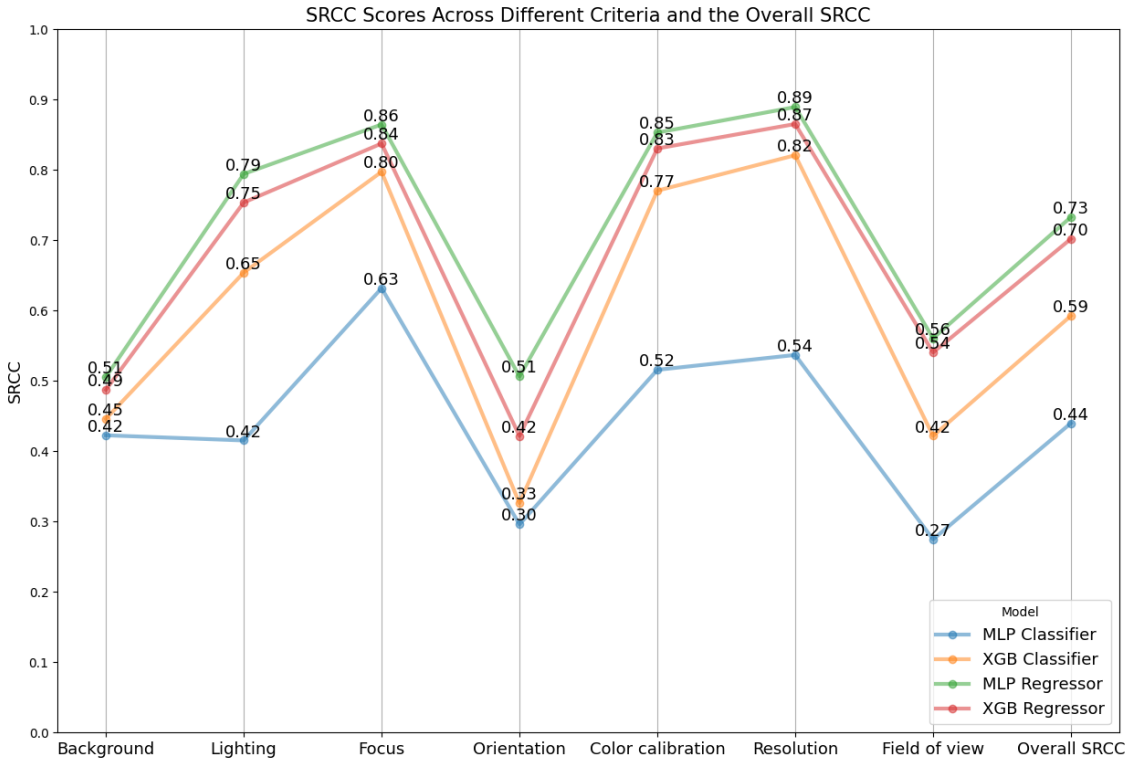


Figure 5.1: Parallel coordinate plot showing the best SRCC values for the four different models across the seven criteria and the overall SRCC. This plot highlights the performance of the MLP Regressor.

The parallel coordinate plot shows that the MLP regressor consistently performs better than the other models across all criteria. Although all models perform similarly for different criteria, no model stands out in a specific criterion. This could be because the same features were used for

every criterion, with only the target labels differing.

In addition to the parallel coordinate plot, Table 5.1 summarizes the cross-dataset evaluation results, showing the generalizability of the models. The table lists the models in the left column and their evaluation results on the SCIN and Fitzpatrick (F17K) datasets in the right columns. This table shows how well the models, trained on one dataset, perform when evaluated on another, providing insights into their robustness and adaptability. All data were synthetically distorted through the pipeline to ensure consistent evaluation conditions.

Table 5.1: Spearman's Rank Correlation Coefficient (SRCC) of Different Models on SCIN and F17K Datasets. F17K refers to the Fitzpatrick17k dataset.

Model	SCIN	F17K
Combined MLP Regressor	0.66	0.75
Combined XGB Regressor	0.65	0.73
Combined XGB Classifier	0.58	0.61
Combined MLP Classifier	0.43	0.46
F17K MLP Regressor	0.54	0.69
SCIN MLP Regressor	0.62	0.49
F17K XGB Regressor	0.53	0.67
SCIN XGB Regressor	0.61	0.48
SCIN MLP Classifier	0.53	0.45
F17K MLP Classifier	0.47	0.58
SCIN XGB Classifier	0.54	0.43
F17K XGB Classifier	0.46	0.59

Given these findings, the MLP regressor was chosen as the final model for further testing. The following sections will detail the performance of the MLP regressor on the test images, providing a clear analysis of its strengths and weaknesses in assessing image quality in teledermatology.

5.1 Model Performance

To fully understand the model's performance, both overall metrics and individual criteria performance were analyzed[1]. Table 5.2 shows the results for the final MLP regressor model evaluated on the 475 good quality Fitzpatrick images, which were synthetically distorted using the distortion pipeline. The results indicate that the model performs very well on focus, color calibration, and resolution, with low MAE, high R^2, SRCC, and Cohen's Kappa values. However, the most problematic criteria are background and orientation. These metrics provide a comprehensive view of the model's strengths and weaknesses, highlighting areas that may need improvement.

In addition to numerical metrics, visual tools were used to gain a clearer understanding of the model's performance. For each criterion, a confusion matrix[2] was created. As shown in Figure 5.2 these matrices display where the model makes correct predictions and where it makes mistakes, showing a detailed view of its accuracy for each type of distortion. The confusion matrix also shows the comparison between the actual scores and the predicted scores.

Examining the confusion plots shows that the bottom row (Figure 5.2d to 5.2g) shows good performance, where the diagonal indicate correct predictions with only minor fluctuations. In contrast, the top row (Figure 5.2a to 5.2c) has more noticeable issues. For instance, Figure 5.2a

[1] from utils.visualization import print_metrics
[2] from utils.visualization import plot_all_confusion_matrices

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**

shows that there are rarely predictions on the higher severity for background distortion. This is because, in the distortion pipeline, if the background proportion is less than 10%, no color blocks are added, resulting in a 0 value for background distortion. This indicates many images were given a 0 for background distortion. Improving the training dataset to include more images with background could address this issue.

Orientation predictions, as shown in Figure 5.2c, is generally unsure and tend to cluster in the middle. This might be due to the various perspective changes (top, bottom, right, left) applied, making the model predict around the middle as it detects perspective distortions but not precisely which way and how strong.

Lighting predictions, as shown in Figure 5.2b, are reasonably accurate, but errors may occur because the criteria include two opposite types of distortion: brightening and darkening the image. This could lead to mispredictions as the inherent distortions look opposite but have the same values for the criteria.

These observations highlight the strengths of using a combined dataset, making the model more robust. Testing with datasets containing more background, such as the SCIN dataset, shows higher background scores, validating this hypothesis. Conversely, the Fitzpatrick dataset, with images taken in controlled settings or with dermatoscopes, shows better field of view predictions due to less background, supporting the combined dataset's robustness. This detailed analysis helps to understand where the model performs well and where improvements are needed.

5.2 Visualizing Model Predictions

To better understand the model's performance, I created visualizations[3] for the test images. These visualizations provided a clear and detailed view of the model's performance, highlighting its strengths and areas for improvement. They were particularly useful for identifying specific cases where the model performed well and where it struggled.

5.2.1 Visualizations for Synthetic Distorted Images

To better understand the model's performance on synthetically distorted images, visualizations were created for 70 test images. These visualizations, as shown in the four-column layout, help to compare the model's predictions with the actual distortions introduced by the pipeline. This approach clearly demonstrates the model's ability to handle various types of distortions.

[3] from utils.visualization import plot_results

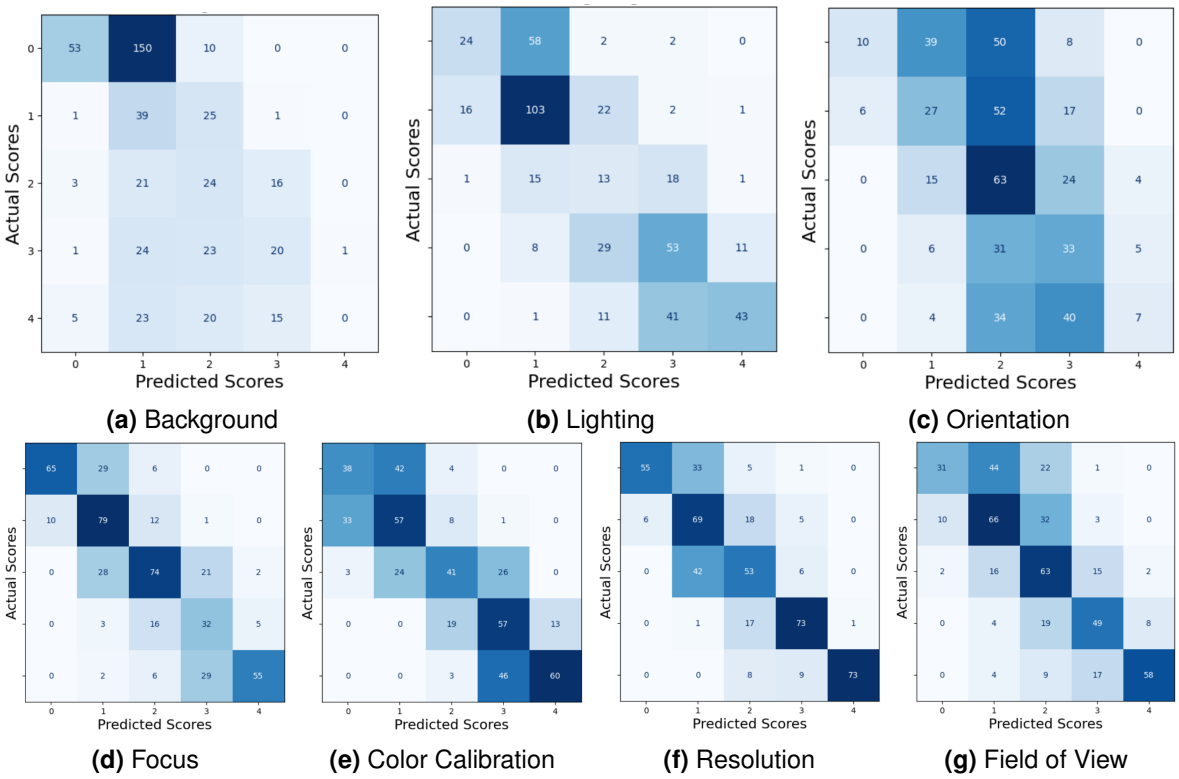


Figure 5.2: Confusion matrices for the MLP Regressor model evaluated on the 475 images from the Fitzpatrick dataset. Each matrix corresponds to a specific distortion criterion and shows the actual scores on the y-axis and the predicted scores on the x-axis. Darker shades indicate higher counts, highlighting where the model's predictions match the actual values and where discrepancies occur.

The first column shows the original image, the second displays the distorted image, the third contains the actual labels, and the fourth presents the model's predictions. This setup makes it easy to compare the model's predictions with the actual distortions.

5.2.2 Visualizations for Authentic Images

The visualizations for the 200 images with authentic distortions use a three-column layout to compare the model's predictions with human-labeled scores. This method highlights the model's performance in real-world scenarios, showing its strengths and areas for improvement.

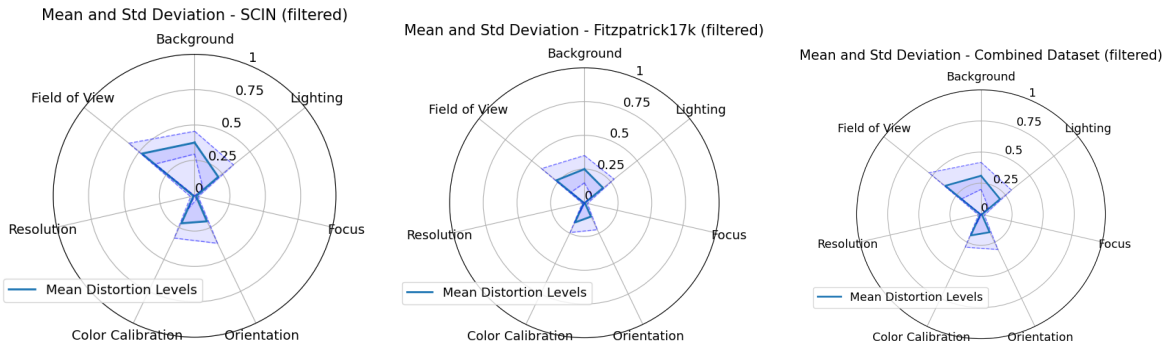
The first column shows the image, the second column displays the human-labeled scores, and the third column presents the model's predictions. This comparison helps show how well the model's predictions align with the human evaluations.

5.3 Testing on Filtered Images

Furthermore, the final model was tested on the original training images filtered for good quality. The radar charts in Figure 4.9 provide a visual representation of distortion levels across seven quality criteria. These charts confirm that the SCIN images exhibit more distortion compared to the Fitzpatrick17k images, which is expected due to the controlled environment in which the Fitzpatrick17k images were taken. The absence of distortion in resolution and focus across both datasets confirms the effectiveness of the filtering process.

Furthermore, the final model was tested on the original training images that were filtered as good quality images. This test was done to confirm that the images are indeed of good quality. Figure 5.3 shows radar charts for the mean distortion levels and standard deviations across seven quality criteria for the 475 good quality SCIN, Fitzpatrick17k, and combined images. These charts provide a visual representation of the distortion levels across the seven criteria, with values ranging from 0 (center) to 1 (outer edge). The standard deviations indicate the variability in distortion levels for each criterion.

The radar charts reveal that the SCIN images has more distortion compared to the Fitzpatrick17k images, with the combined images falling in between. This suggests that the SCIN images have more distortions than the Fitzpatrick17k images, which is expected since the Fitzpatrick17k images were taken in a controlled environment. Additionally, both the SCIN and Fitzpatrick17k images show no distortion in resolution and focus, confirming that the filtering process was effective in selecting good quality images.



(a) Filtered SCIN Images **(b)** Filtered Fitzpatrick Images **(c)** Combined Images

Figure 5.3: Radar charts on the mean distortion levels and standard deviations across seven quality criteria for the 475 good quality SCIN, Fitzpatrick17k, and combined images.

Chapter 6

Discussion and Conclusion

text

Bibliography

Agnolucci, L., Galteri, L., Bertini, M., & Del Bimbo, A. (2023, November 4). *ARNIQA: Learning Distortion Manifold for Image Quality Assessment*. arXiv: 2310.14918 [cs]. 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*. arXiv: 2206.01774 [cs, eess]. 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*. arXiv: 2104.09957 [cs]. 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*. arXiv: 2209.04699 [cs]. 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

Kawahara, J., Daneshvar, S., Argenziano, G., & Hamarneh, G. (2019). Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 538–546. https://doi.org/10.1109/JBHI.2018.2824327

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

Pacheco, A. G., & Krohling, R. A. (2020). The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, *116*, 103545. https://doi.org/10.1016/j.compbiomed.2019.103545

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

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*. arXiv: 2010.02086 [cs, eess]. 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*. arXiv: 2402.18545 [cs]. Retrieved April 28, 2024, from http://arxiv.org/abs/2402.18545

Wu, X., Wen, N., Liang, J., Lai, Y.-K., She, D., Cheng, M.-M., & Yang, J. (2019). Joint Acne Image Grading and Counting via Label Distribution Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10641–10650. https://doi.org/10.1109/ICCV.2019.01074

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

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

Documents and code relevant to the thesis can be downloaded from the following link: **https://github.com/Schoggi-Mimi/bachelor-thesis**.

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

include risk management and project planning pdfs

Appendix B

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

search database and maybe add KonIQ-10k and other from ARNIQA!

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.

Appendix C

Degradation Types

As mentioned in Subsection 2.2.2, the dataset used in this thesis is augmented with synthetic degradations. The following figures Figure C.1, Figure C.2, Figure C.3, Figure C.4, Figure C.5, Figure C.6, Figure C.7 show the different levels of intensity for the degradations of each distortion group.

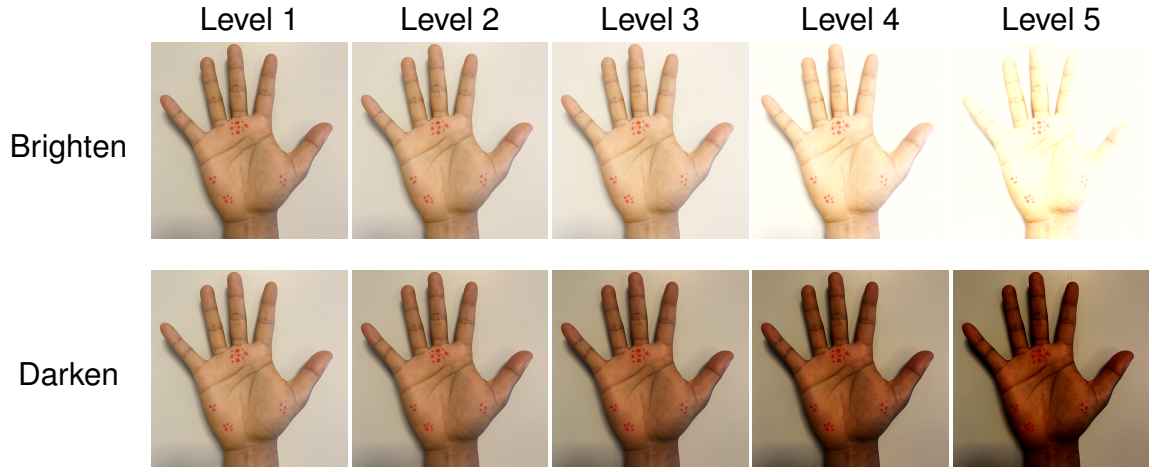


Figure C.1: Visualization of the degradation types belonging to the *Brightness change* group for increasing levels of intensity.

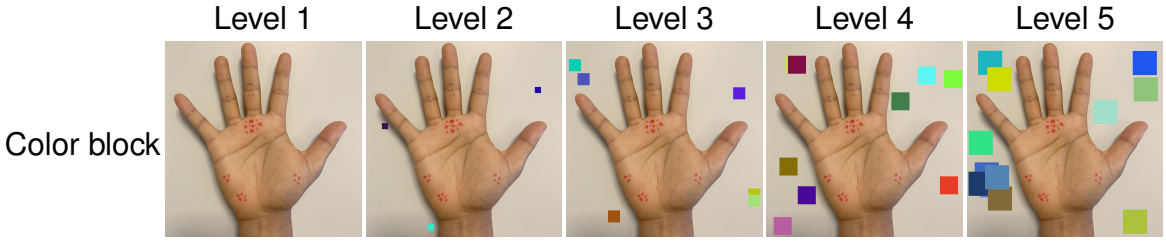


Figure C.2: Visualization of the degradation types belonging to the *Background color* group for increasing levels of intensity.

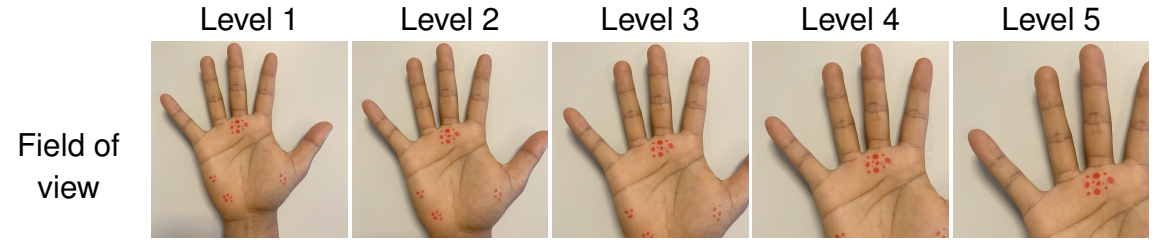


Figure C.3: Visualization of the degradation types belonging to the *Field of View* group for increasing levels of intensity.

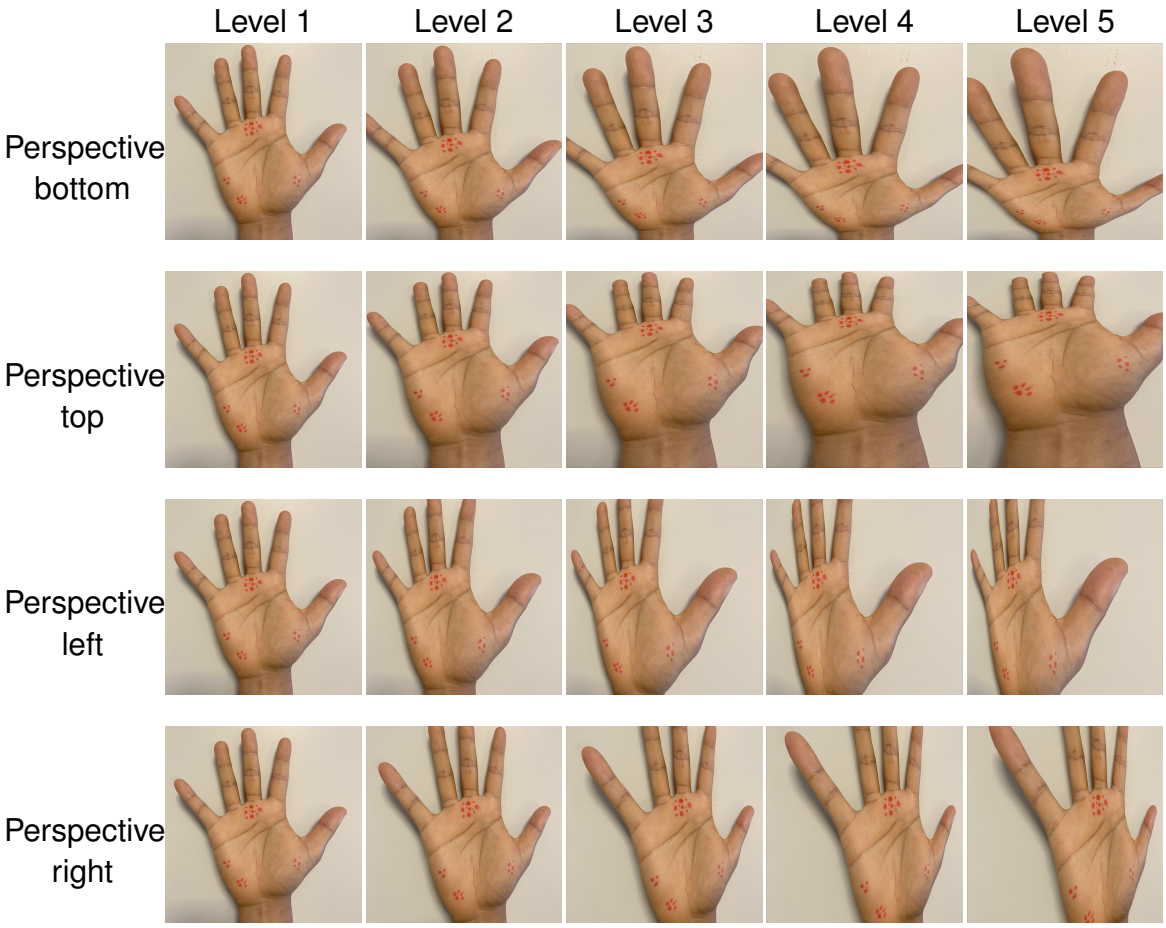


Figure C.4: Visualization of the degradation types belonging to the *Image orientation* group for increasing levels of intensity.

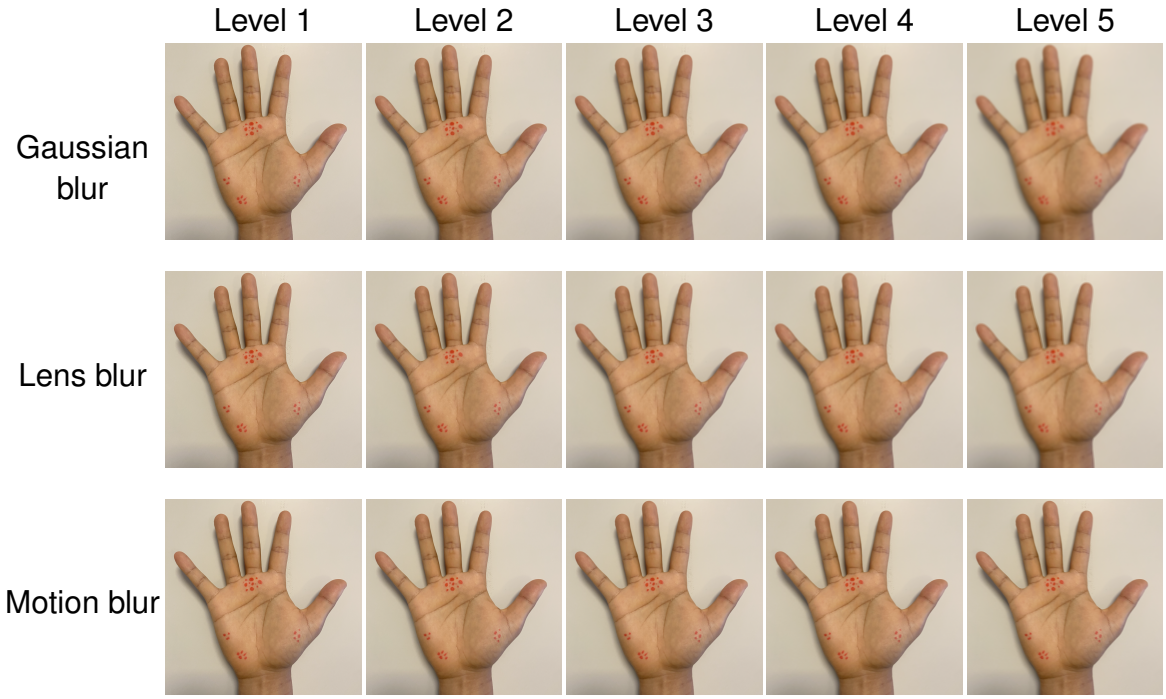


Figure C.5: Visualization of the degradation types belonging to the *Focus* group for increasing levels of intensity.

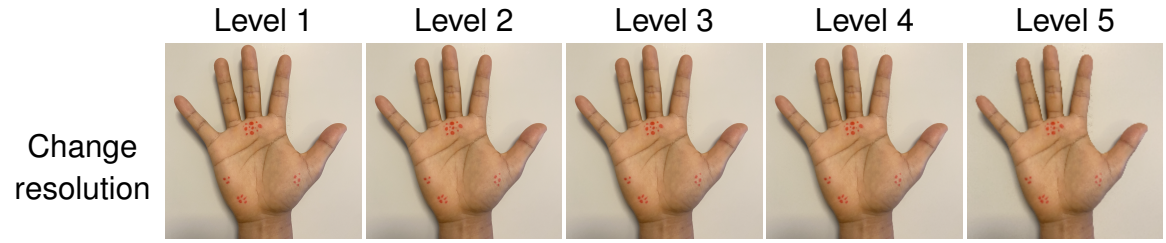


Figure C.6: Visualization of the degradation types belonging to the *Resolution* group for increasing levels of intensity.

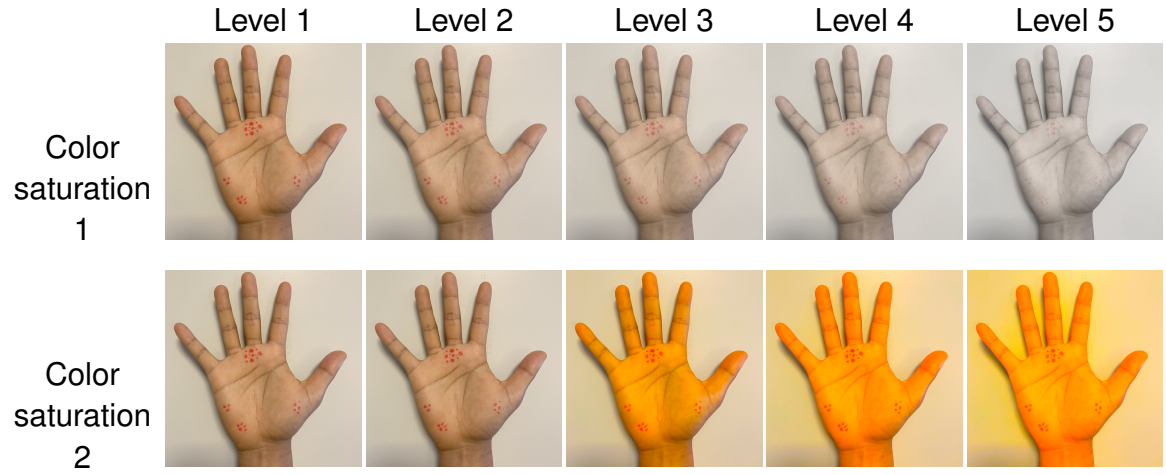


Figure C.7: Visualization of the degradation types belonging to the *Color calibration* group for increasing levels of intensity.

Appendix D

Code

Anhang, Abkürzungs-, Abbildungs-, Tabellen-, Formel-Verzeichnis, Literaturverzeichnis nicht vergessen!

Anhänge

Projektspezifisch können weitere Dokumentationsteile angefügt werden wie:

Aufgabenstellung, Projektmanagement-Plan/Bericht, Testplan/Testbericht, Bedienungsanleitungen, Details zu Umfragen, detaillierte Anforderungslisten, Referenzen auf projektspezifische Daten in externen Entwicklungs- und Datenverwaltungstools etc.

Listing D.1: Caption on PDF

```
import numpy as np
```