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.

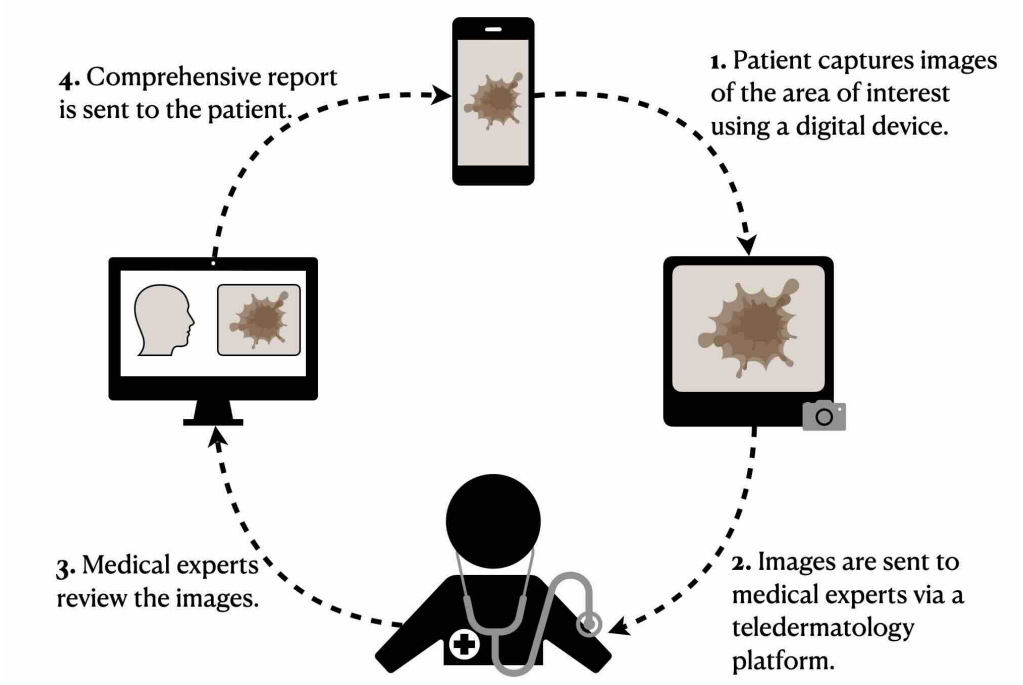


Figure 1.1: This diagram shows the simplified process of a teledermatology consultation, starting from the patient capturing images of their skin condition to receiving a detailed medical report.

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. **??** covers the literature review, discussing previous and related works on image quality assessment (IQA) and teledermatology. Chapter 2 details the methodologies, including those used in the literature review and those specific to IQA and teledermatology. In Chapter 3, the experiments conducted are described, showing the approaches

taken, along with the metrics used. Chapter 4 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

Methodology

Based on the insights from **??**, *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.

2.1 Explorative Approach

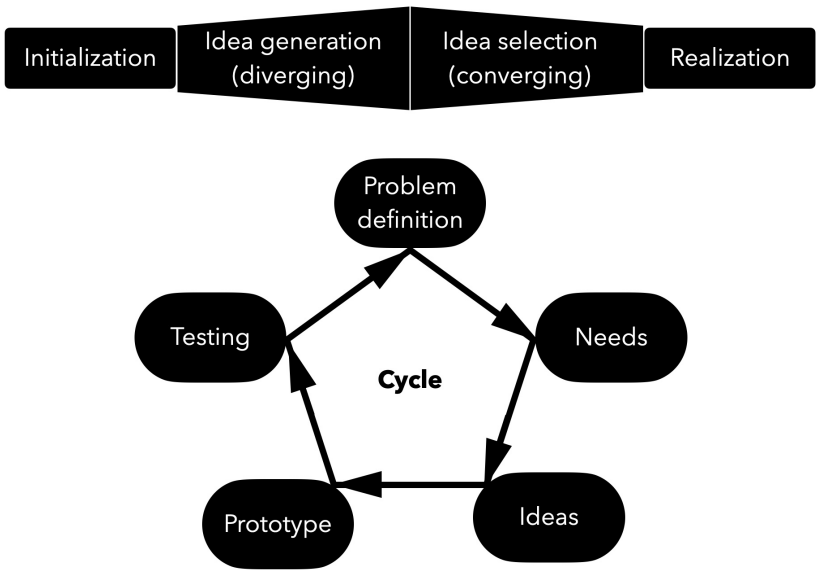


Figure 2.1: Visualization of the explorative approach, including the stages of initialization, idea generation, idea selection, and implementation. The lower part of the figure shows the decision cycles adapted from (cite) (Hoffmann et al., n.d.).

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 2.1, which shows the stages of starting, generating ideas, selecting ideas, and implementation, along with the decision cycles .

2.2 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.

2.3 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:

2.3.1 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.

2.3.2 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 **??**. 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 **??**. 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., n.d.). 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., n.d.), 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

2.3.3 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.

2.3.4 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.

2.3.5 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.

2.3.6 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:

$$SRCC = 1 - \frac{6 \sum_{i=1}^{n} (d_i^2)}{n(n^2 - 1)} \tag{2.1}$$

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:

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.2}$$

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.

2.3.7 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 3

Implementation

This chapter explains the detailed implementation of the methods described in Chapter 2. 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.

3.1 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.

3.1.1 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

3.1.2 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 **??**. 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 3.1. 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.

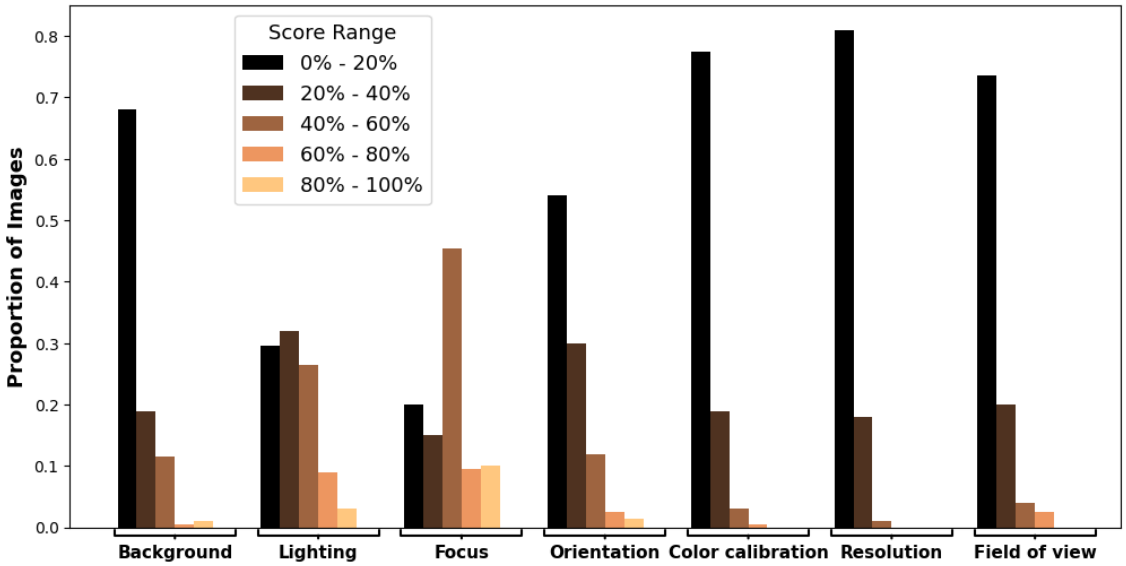


Figure 3.1: Histograms showing the distribution of distortion scores for each quality criterion.

In Figure 3.1 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.

3.2 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 **??**.

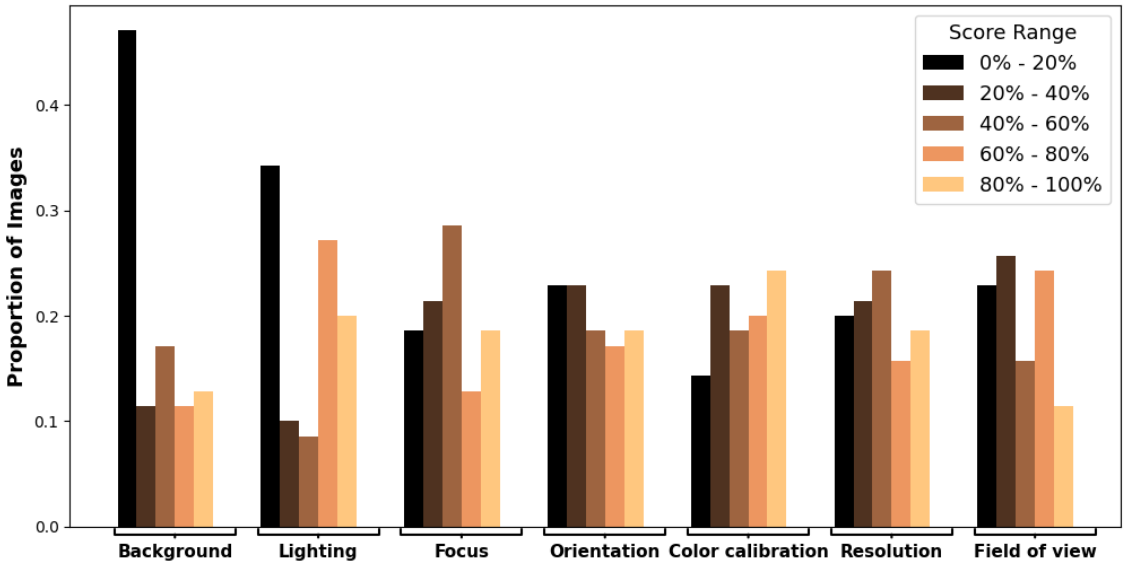


Figure 3.2: Histograms showing the distribution of distortion scores for each quality criterion.

3.2.1 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., n.d.) image degradation model, which was inspired by the KADID (Lin et al., n.d.) 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.

3.3 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 3.3(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

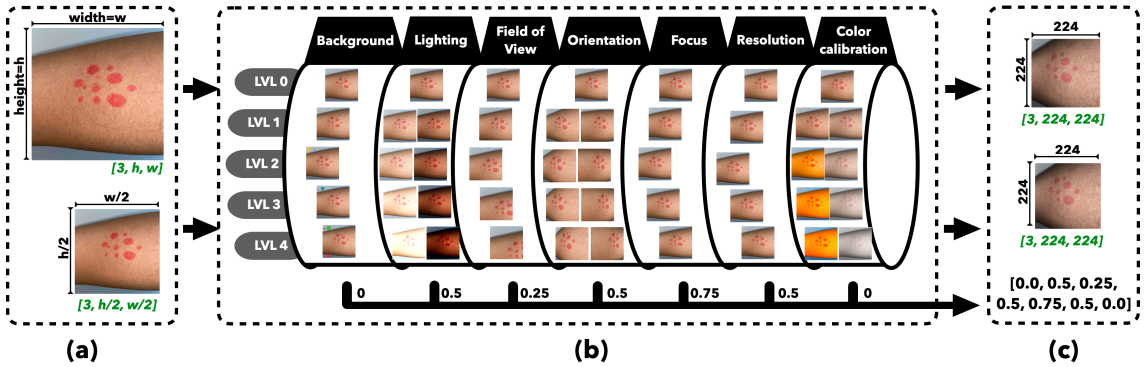


Figure 3.3: Distortion pipeline for generating training images with different levels of distortion. (a) shows the original image and its downscaled version. (b) illustrates the distortion pipeline where a type of distortion and a random level for each criterion are selected, with the corresponding mapped 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., n.d.). 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.

3.4 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-

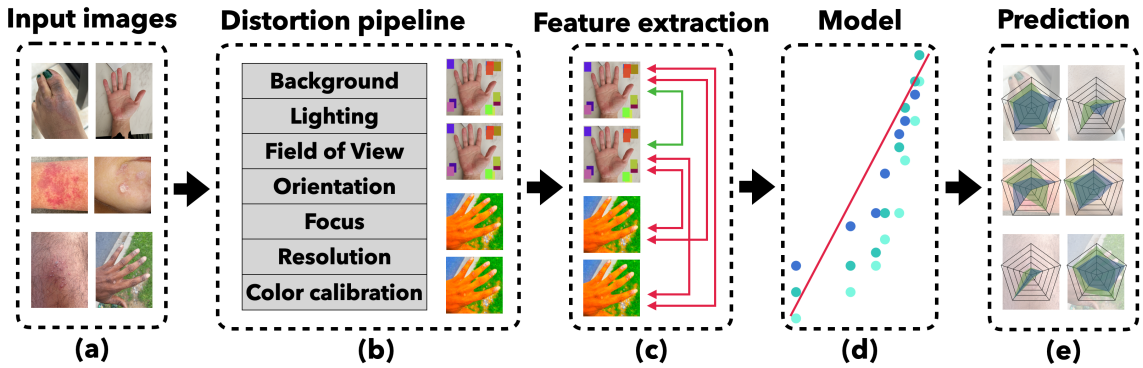


Figure 3.4: Overview of the entire process for training and evaluating the image quality assessment model. (a) shows the album of input images. (b) depicts the distortion pipeline that randomly distorts images across seven criteria at five severity levels. (c) illustrates the feature extraction using 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.

3.5 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 3.5 and Figure 3.6, 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.

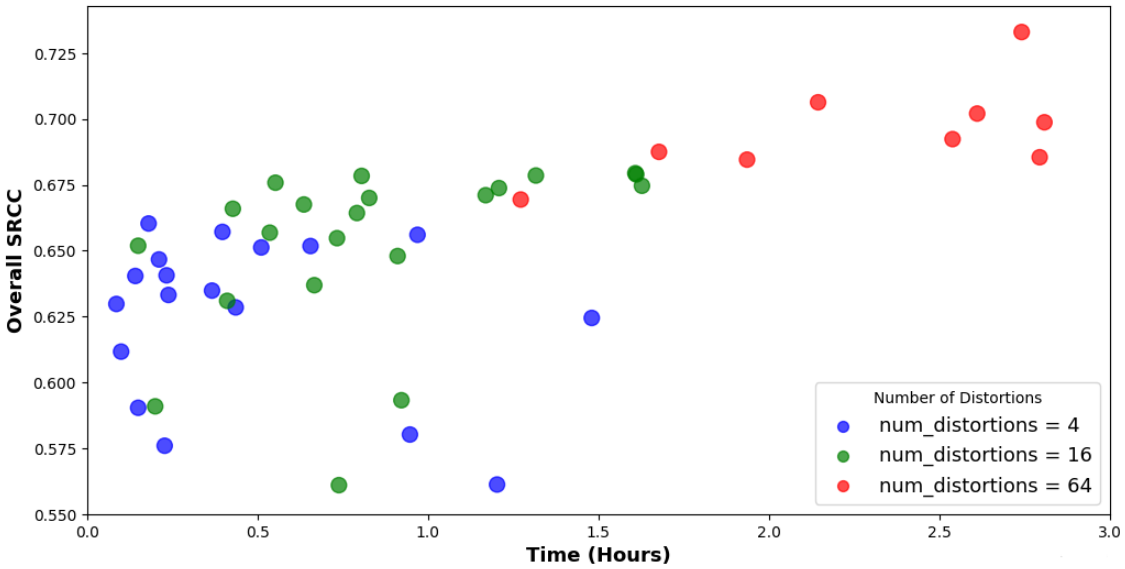


Figure 3.5: Overall SRCC for XGB Regressor and MLP Regressor with different numbers of distortions. The x-axis shows the time it took to train, and the y-axis shows the SRCC values, demonstrating better performance with larger datasets.

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

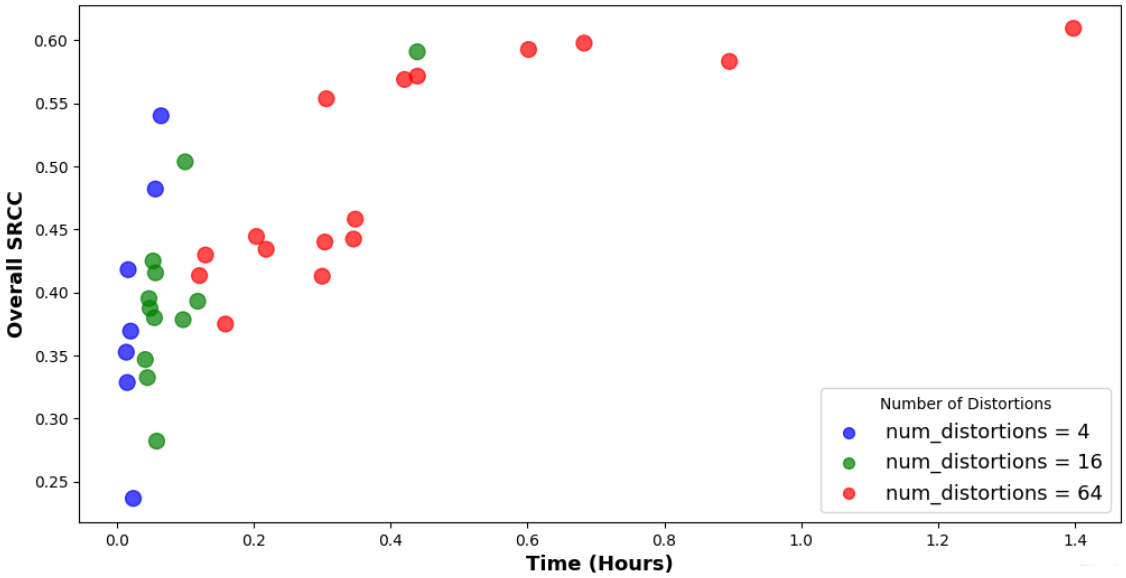


Figure 3.6: Overall SRCC for XGB Classifier and MLP Classifier with different numbers of distortions. Similar to Figure 3.5, this figure shows improved performance with larger datasets.

3.5.1 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 3.1: 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 3.2: 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)

3.6 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 4

Results and Analysis

In this chapter, the main objective is to present the findings from the experiments and analysis conducted in the previous chapters. The section is designed to show the performance of the trained models, especially focusing on the MLP regressor model, which was trained on the combined SCIN and Fitzpatrick datasets. This chapter will include various tables, figures, and visualizations that display the results without going into detailed explanations. That will be covered in the next chapter.

4.1 Range of Distortion Values

In this section, the goal is to show the chosen range of values for each distortion type used in the study. Each distortion type was visualized individually to ensure they reflect realistic scenarios for teledermatology applications. **??** in the supplementary material includes images showing each criterion with different distortion types and five severity levels. It is important to note that images should not be normalized before viewing because normalization can make them appear overly colorful and unrealistic. However, normalization is necessary during training and testing because the feature extraction backbone from ARNIQA(Agnolucci et al., n.d.) was trained on ResNet50 with ImageNet images.

4.2 Model Performance

The performance of the four different models was evaluated through cross-dataset evaluation. This involved assessing the models on both the SCIN and Fitzpatrick (F17K) datasets after synthetic distortion, as summarized in Table 4.1. This table highlights how well the models generalize across different datasets.

Table 4.1: Spearman's Rank Correlation Coefficient (SRCC) of Different Models on SCIN and F17K Datasets. Note that the Fitzpatrick dataset is referred to as F17K for simplicity.

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

4.2.1 Parallel Coordinate Plot

The parallel coordinate plot in Figure 4.1 compares the best-performing models across seven criteria, including the overall SRCC. This visualization highlights the performance of the MLP Regressor, showing that it consistently outperforms the other models.

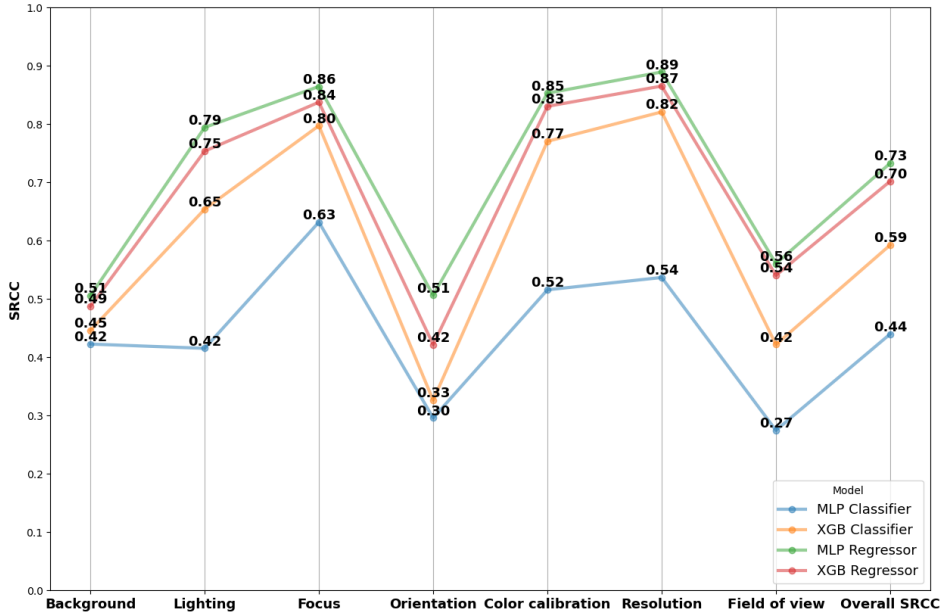


Figure 4.1: Parallel coordinate plot showing the best SRCC values for the four different models across the seven criteria and the overall SRCC. This plot highlights the performance of the MLP Regressor.

4.2.2 Loss Curve Analysis

The loss curve in Figure 4.2 shows the model's loss decreases over time for each distortion criterion during training. This visualization highlights the model's performance improvement over time with each iteration. The maximum number of iterations was set to 500, and early stopping was enabled to prevent overfitting.

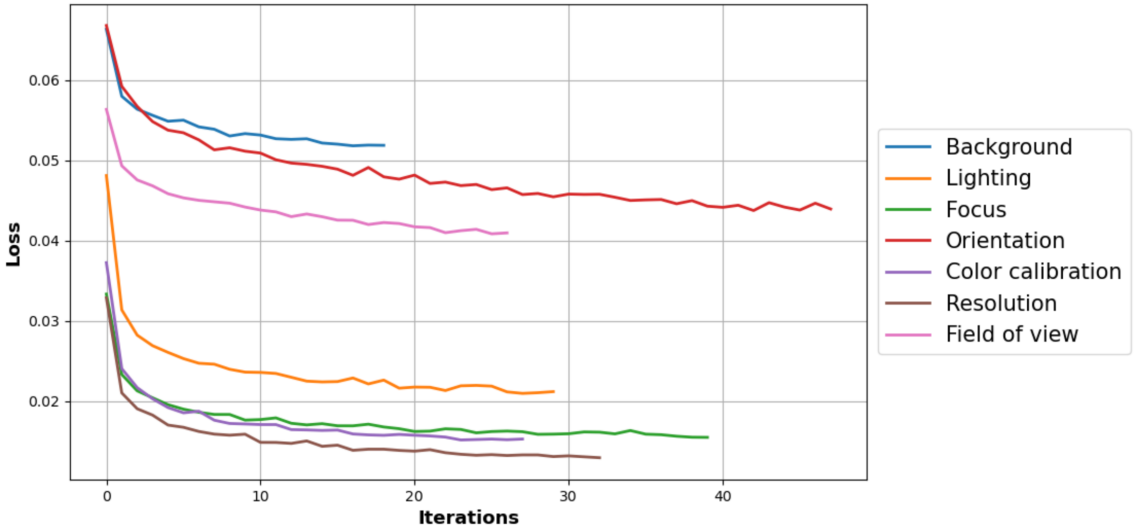


Figure 4.2: Loss curve showing the reduction in loss for each distortion criterion during the training process. Each line represents a different criterion, showing how the model's performance improves over time with each iteration.

4.2.3 Performance Metrics

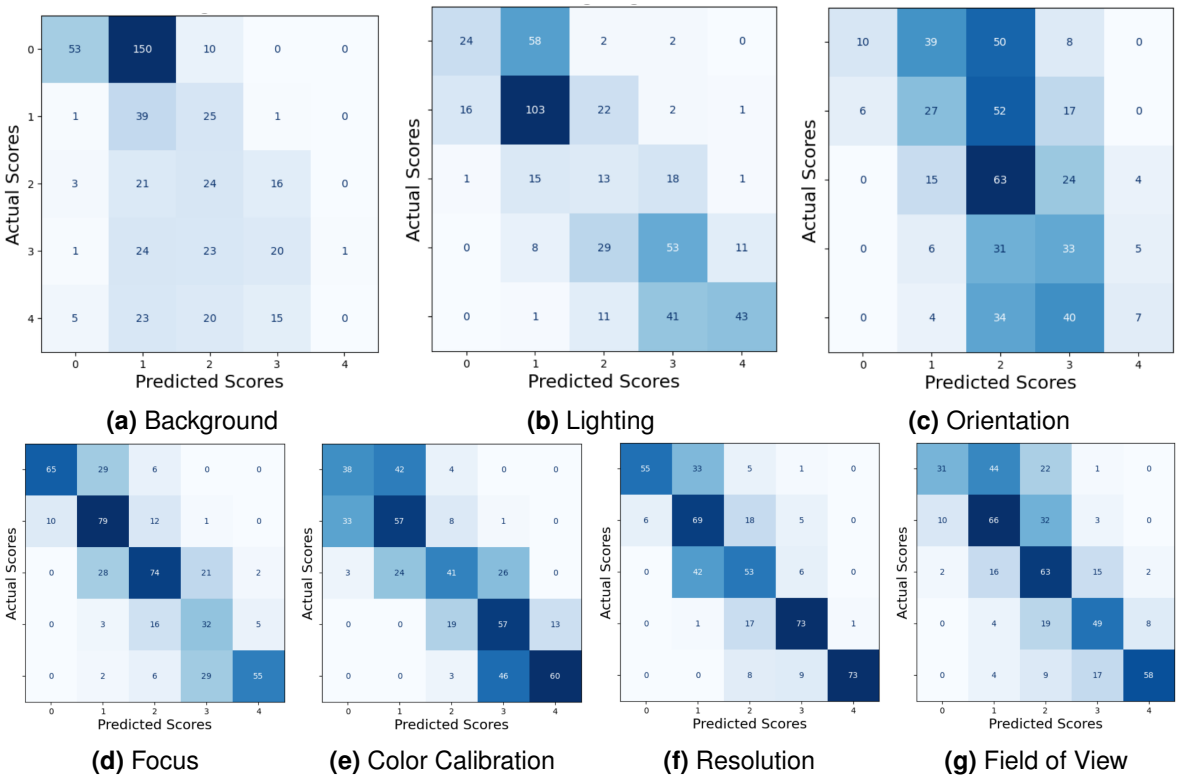
The performance of the final MLP regressor model on individual criteria is shown in Table 4.2. This table presents the performance metrics for the final MLP regressor model on 475 good quality Fitzpatrick images that were synthetically distorted. These metrics give a detailed view of the model's strengths and weaknesses.

Table 4.2: 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**

4.2.4 Confusion Matrices

In addition to numerical metrics, confusion matrices[1] were created for each criterion, as shown in Figure 4.3. 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. Furthermore, the confusion matrices also reveal any biases the model might have toward certain severity ranges, indicating whether it tends to predict only low or high severity levels, or if its predictions are skewed in some way.



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

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

Figure 4.3: 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.

[1] from utils.visualization import plot_all_confusion_matrices

4.3 Model Predictions

To better understand the model's performance on the two test sets (70 synthetic distorted images and 200 authentic images), radar charts[2] were created. These charts show the criteria on the outside, with severity ranges going from the center (0) to the outer edge (1), indicating high distortion for each criterion. These visualizations provide a clear and simple view of the model's performance, showing its strengths and areas for improvement.

4.3.1 Visualizations for Synthetic Distorted Images

These visualizations, as shown in Figure 4.4, help to compare the model's predictions with actual distortions for synthetic test images. This also helps to demonstrate the model's accuracy in predicting various types of distortions.

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

4.3.2 Visualizations for Authentic Images

The visualizations, as shown in Figure 4.5, compare the model's predictions with human-labeled scores for authentic images. This highlights the model's performance in real-world scenarios.

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.

[2] from utils.visualization import plot_results

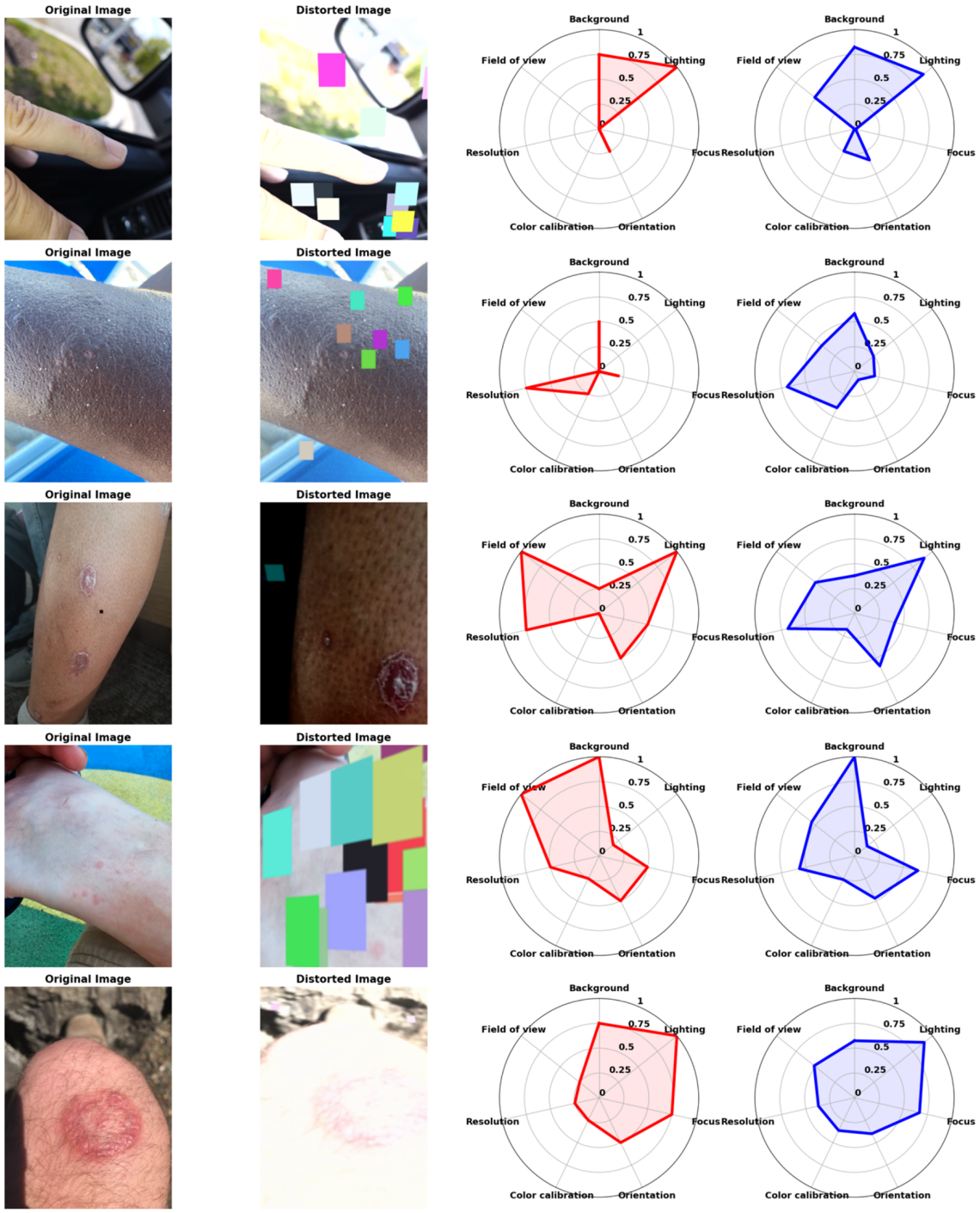


Figure 4.4: Visualizations for the MLP Regressor model on 70 synthetic distorted images. The four-column layout shows the original image, the distorted image, the actual labels, and the model's predictions.

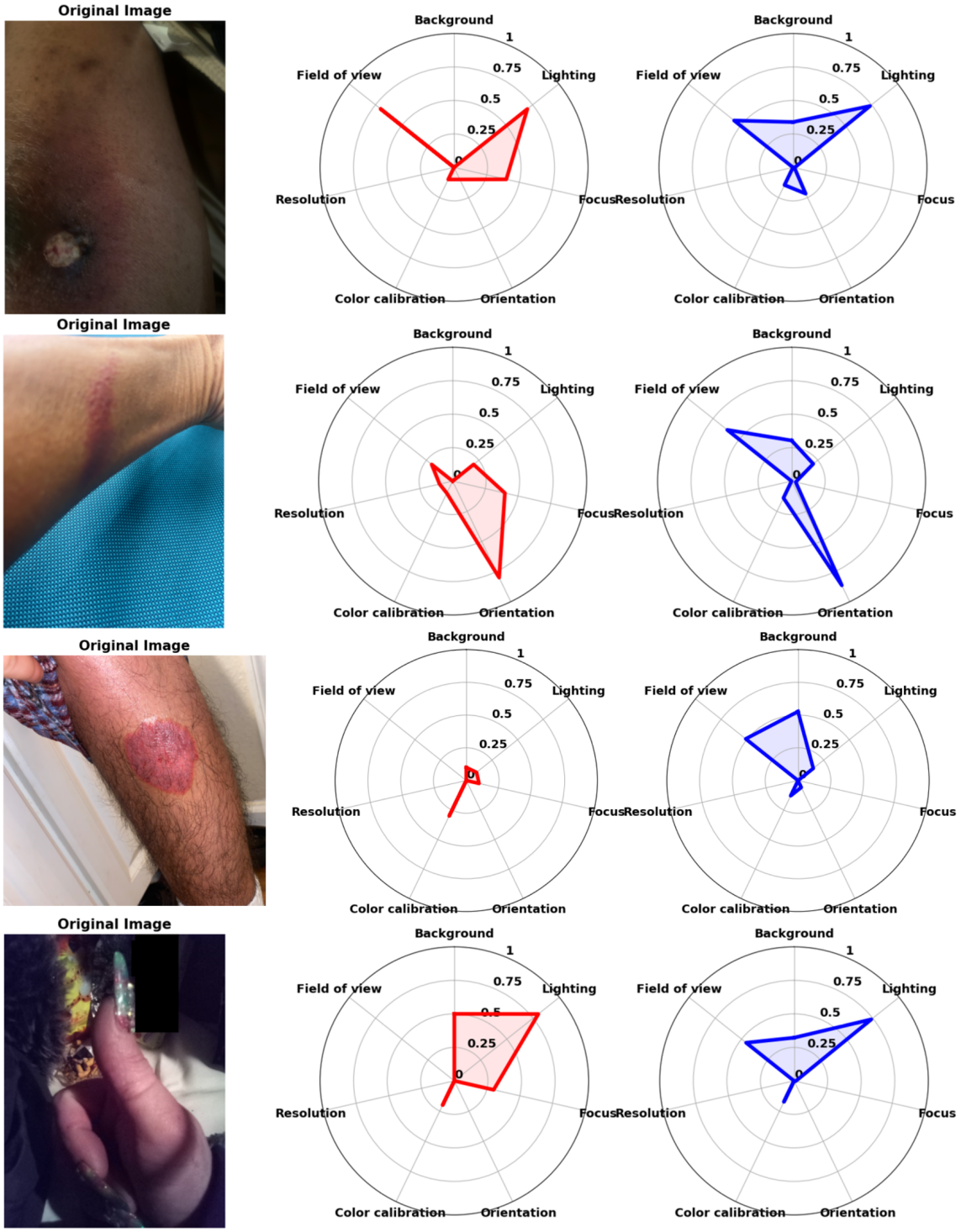


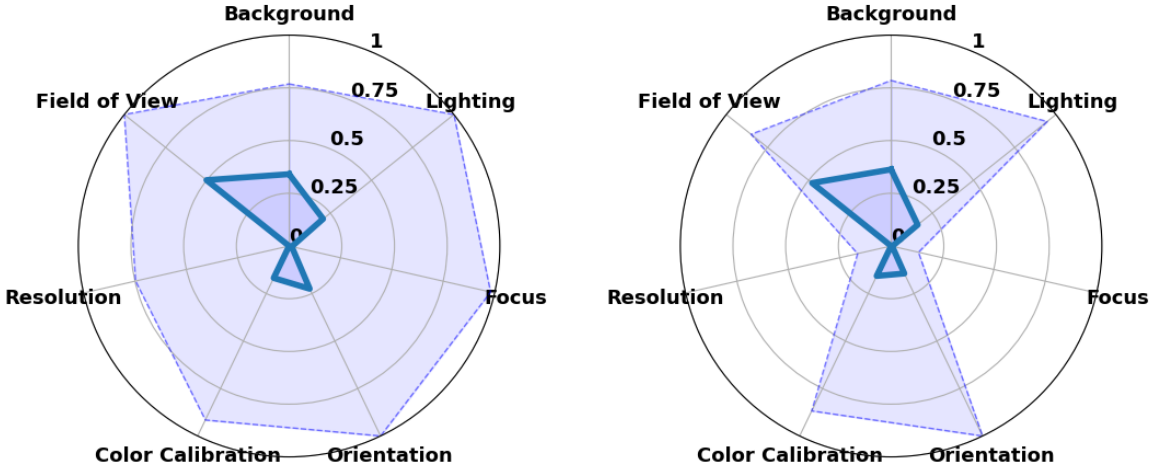
Figure 4.5: Visualizations for the MLP Regressor model on 200 authentic images. The three-column layout shows the image, the human-labeled scores, and the model's predictions.

4.4 Assessing Training and Testing Images Quality

To verify the quality of the images used for training and see how they change after synthetic distortion, radar charts were created. These charts show the quality of the original training images and how they are affected by the distortions. Additionally, the quality of both the synthetic and authentic test images is assessed using the same method. These radar charts show a simple visual representation of the quality and the level of distortion across the seven criteria.

4.4.1 Training Images Quality

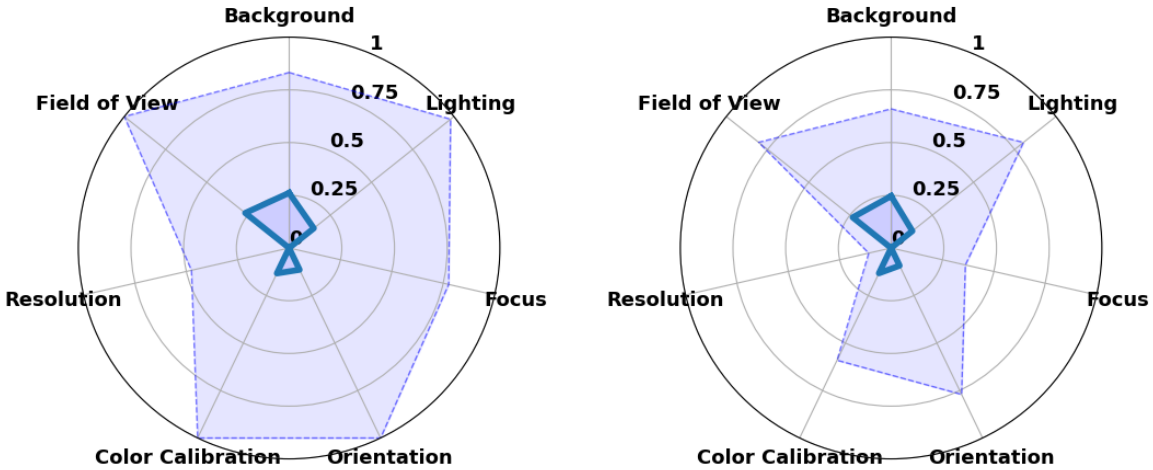
Figure 4.6 and Figure 4.7 show the quality of the original SCIN and Fitzpatrick images, and the filtered good quality images, respectively. Figure 4.8 shows the quality of the combined SCIN and Fitzpatrick images and the synthetically distorted images used for training the model.



(a) SCIN Images

(b) Filtered SCIN Images

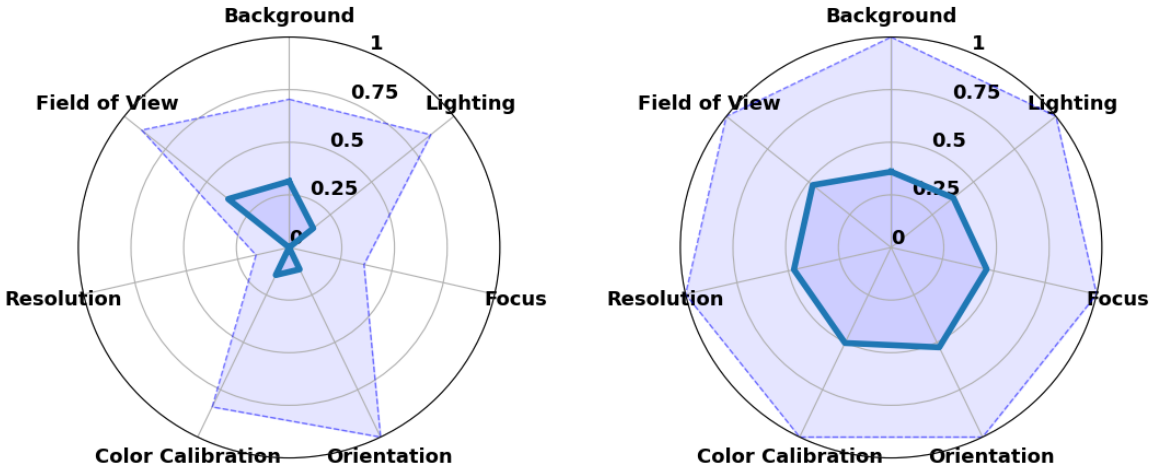
Figure 4.6: Radar charts for the SCIN dataset. (a) Original images from the SCIN dataset (10'379 images). (b) Filtered good quality images (475 images).



(a) Fitzpatrick Images

(b) Filtered Fitzpatrick Images

Figure 4.7: Radar charts for the Fitzpatrick dataset. (a) Original images from the Fitzpatrick dataset (16'577 images). (b) Filtered good quality images (475 images).



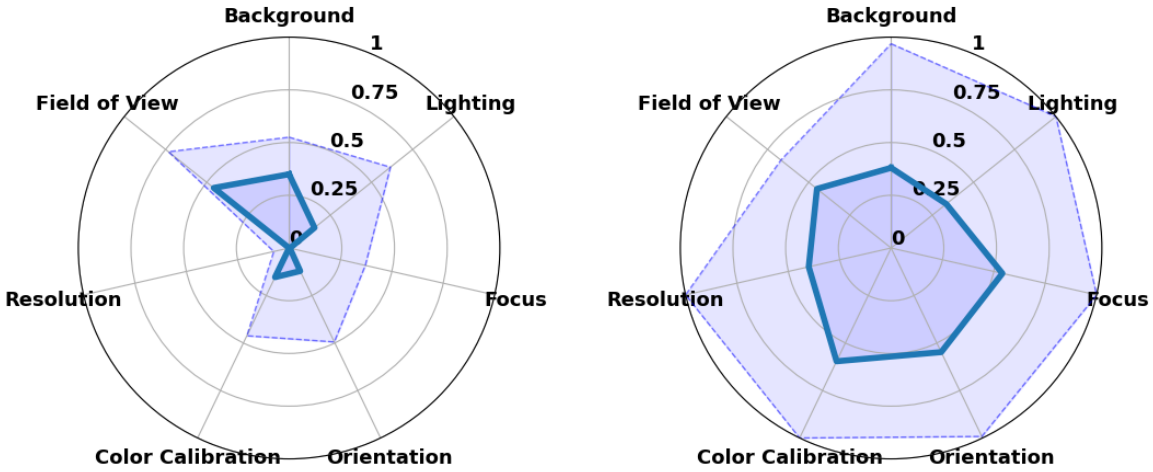
(a) Combined SCIN and Fitzpatrick Images

(b) Synthetically Distorted Images

Figure 4.8: Combined dataset analysis. (a) Combined SCIN and Fitzpatrick images (950 images). (b) Synthetically distorted images.

4.4.2 Test Images Quality

Figure 4.9 shows the quality of the filtered good quality test images and the synthetically distorted test images. Figure 4.10 shows the quality of the authentic test images from the SCIN dataset. These radar charts provide a visual representation of the quality of the test images and the level of distortion across the seven criteria.



(a) Filtered Test Images

(b) Synthetically Distorted Test Images

Figure 4.9: Synthetic test set analysis. (a) Filtered good quality test images (70 images, independent of training set). (b) Synthetically distorted test images.

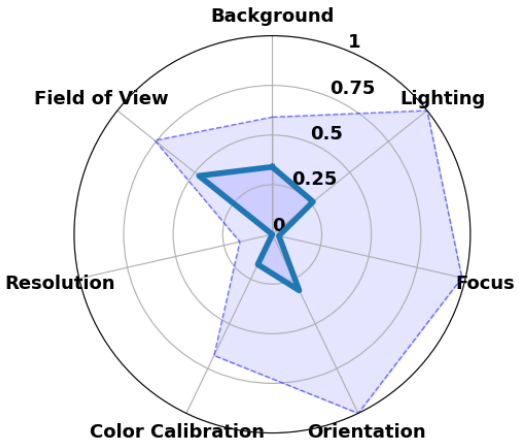


Figure 4.10: Authentic test set from the SCIN dataset, independent of the training images, showing real-world distortions.

4.5 Comparison with ARNIQA Predictions

This section compares the model's predictions with the predictions from ARNIQA for both synthetic and authentic test sets, highlighting how well the model aligns with established quality scores.

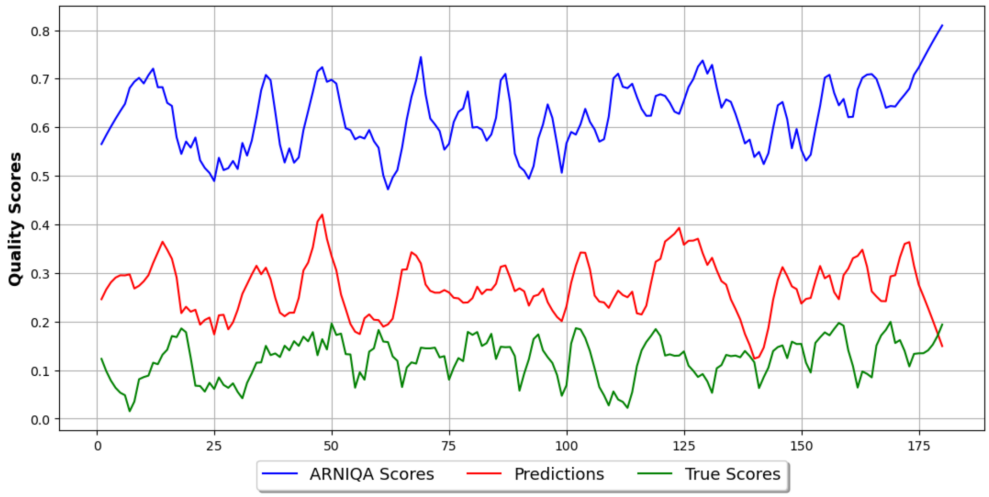


Figure 4.11: Comparison of ARNIQA Scores, Model Predictions, and True Scores for 200 authentic images from the SCIN Dataset. This plot presents the quality scores for each image, showcasing how well the model's predictions align with the ARNIQA scores and the true scores.

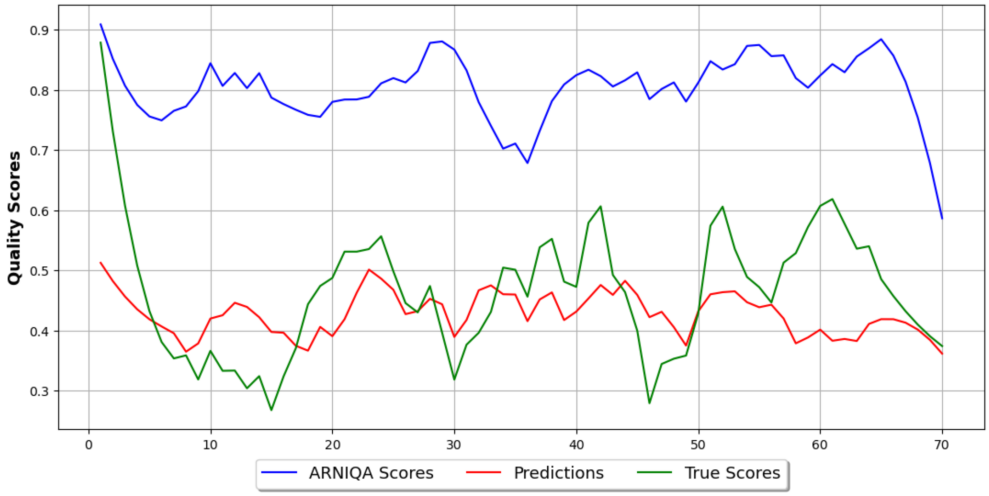


Figure 4.12: Comparison of ARNIQA Scores, Model Predictions, and True Scores for 70 synthetically distorted images from the SCIN dataset. This plot shows the quality scores for each image, highlighting the performance of the model against ARNIQA's predictions and the actual scores.

Chapter 5

Discussion

The chapter on Discussion covers the entire work and, in particular, the results achieved in Chapter 4. In the following sections, the results will be interpreted, the objectives of the thesis will be reviewed, and recommendations for future research in teledermatology image quality assessment will be provided.

5.1 Interpretation of Results

The differences in model performance shown in Figure 4.1 reveal that both classifiers (XGB Classifier and MLP Classifier) were not as effective as the regressors (XGB Regressor and MLP Regressor). This is likely because the task involves predicting continuous severity scores, which are more suited to regression models. Classifiers categorize the severity into fixed levels, which can lead to less precise predictions.

check this paragraph

The experiments demonstrated that the cross-dataset evaluation showed good generalization. The models performed well not only on the datasets they were trained on but also on previously unseen datasets. This indicates that the models, especially the MLP Regressor, can generalize well to different data distributions and are robust.

5.1.1 Analysis of Individual Distortion Criteria

When we analyze the metrics in Table 4.2 alongside the confusion plots in Figure 4.3, it becomes clear that the criteria for focus, color calibration, and resolution are captured very well by the model. Although there are some fluctuations between the predicted severity and the actual severity, these deviations are typically only by one severity level. This minor difference suggests that the model is making reasonably accurate predictions. This high level of accuracy can be explained by the design of the ARNIQA backbone. When training ARNIQA to extract features, their goal was to assess general image quality, and they focused also on distortions related to focus and color calibration. Since I am using their backbone to extract features, the model performs better in these criteria.

On the other hand, lighting was also one of the distortions that ARNIQA focused on, but this criterion performed only moderately well. This can be reasonably explained by the fact that the lighting criterion includes two opposite types of distortions: brightening and darkening. If the majority of the training images were brightened and the validation set included darkened images,

this could negatively impact performance. That is why the model struggles to accurately predict the lighting severity due to these opposing distortion types.

For the background criterion, it is clear from the confusion plot that there are rarely predictions on the higher severity levels. This is because, in the distortion pipeline, if the background proportion is less than 10% relative to the skin, no color blocks are added, resulting in a 0 value for background distortion. This indicates that many images were given a 0 value for background distortion. Additionally, when looking at the radar chart of the combined synthetically distorted images in Figure 4.8b, the median value for background distortion is lower than that of other criteria, indicating fewer strong severity values.

The confusion plot also shows that the orientation criterion is generally uncertain in its predictions, tending to cluster around the middle severity levels. This might be due to the various perspective changes (top, bottom, right, left) applied during training. As a result, the model detects that there is some perspective distortion but cannot precisely determine the direction or severity, leading to predictions that hover around the middle severity levels.

Field of view distortion was the most experimental criterion. In teledermatology, it is crucial to have the lesion or area of interest centered in the image. However, in general photography, the subject is often placed off-center to create a more aesthetically pleasing composition. This is why ARNIQA, trained for general image quality assessment, might have difficulties with field of view distortions specific to teledermatology. This was confirmed by comparing the performance metrics between SCIN and Fitzpatrick images, where Fitzpatrick images performed better in field of view distortion, likely due to having less background (see Table 5.1 and compare Figure 4.6 with Figure 4.7).

Table 5.1: Performance metrics for field of view distortion using an MLP regressor on synthetically distorted SCIN and F17K images. F17K refers to the Fitzpatrick17k images.

Dataset	MAE	R^2	SRCC	Cohen's Kappa
SCIN (synthetically distorted)	1.20	0.05	0.23	0.08
F17K (synthetically distorted)	0.63	0.50	0.72	0.71

These observations highlight the strengths of using a combined dataset. By integrating diverse images from different sources, the model benefits from a wider variety of distortions and scenarios, enhancing its ability to generalize and perform well across different conditions.

5.1.2 Overall Model Performance on Test Sets

Synthetic Distorted Images

The model's performance on the 70 synthetic distorted test images aligns well with the validation split and the cross-dataset evaluation, as shown in Figure 4.4. In this figure, I randomly selected some images, and the four-column layout presents the original image, the distorted image, the actual labels, and the model's predictions.

First, let's discuss background distortion. The actual severity matches closely with the predictions for most images. To further improve the model's performance, an experiment I didn't have time to conduct would involve applying color blocks randomly on the whole image without any skin segmentation. This approach, though unconventional for teledermatology images, could test the model's robustness by introducing random artifacts. The second and fourth images in Figure 4.4 suggest this method could work effectively.

One noticeable issue is the field of view distortion, where the predictions are not accurate. Despite background, lighting, and orientation distortions not being well-represented in the confusion plot in Figure 4.3, they appear to be quite accurate in the synthetic test set images shown. This suggests that while the confusion plot highlights general trends, individual image predictions can vary.

One important factor to consider is the distribution of distortion severity levels. The distortion pipeline selects ranges randomly. To improve this, I experimented with choosing distortion ranges according to a Gaussian distribution centered at 0 severity with a standard deviation of 2.5. This approach might include more distortions with lower severity, which are more common in teledermatology images. For instance, heavily brightened images, as seen in the last synthetic distorted image, may not occur frequently in real-world scenarios. Training the model on more common distortions could enhance its performance. This experiment was initiated, but no evaluation could be conducted due to time constraints.

Authentic Test Images

For the 200 authentic test images, random samples are shown where I labeled the images. At first look, the predictions do not match my labels well. This difference could be due to several factors. Firstly, I labeled the images primarily by focusing on the skin lesion, often ignoring the background, which might not align with the model's overall assessment. Labeling 1,400 instances (200 images, 7 criteria each) likely introduced some human error.

Also, images with multiple distortions complicate the assessment of individual criteria. For instance, a heavily darkened image might hide other distortions like focus, resolution, and color calibration, making it difficult to evaluate accurately. Additionally, labeling lesions or marks on darker skin tones presented challenges, which might have affected the labeling accuracy. This highlights the broader issue of skin tone diversity in medical imaging, an important factor that was not within the scope of this thesis but is worth mentioning for future research.

Despite these challenges, if we take out the field of view distortion, the model's predictions are quite accurate. This distortion significantly affects the radar charts. I chose not to label the images again to avoid "leaking information" from the test set into the training process, keeping the evaluation fair and accurate.

5.1.3 Comparison with ARNIQA Predictions

The first thing to note is that ARNIQA predicts image quality scores ranging from 0 to 1, where higher values indicate better image quality. In contrast, my model predicts distortion severity scores, also ranging from 0 to 1, but where higher values indicate worse image quality. My model also evaluates seven different criteria, and to make a direct comparison with ARNIQA's single quality score, I took the median of all distortion scores. This provides a single, balanced score that represents the overall quality of the image by aggregating the severity of distortions across all criteria. Additionally, the scores from ARNIQA had to be inverted for comparison. ARNIQA also provides six different regressors for image evaluation, and for this analysis, I used the default regressor, which was trained on the KADID10K dataset.

Figure 4.11 and Figure 4.12 show the differences in predictions. ARNIQA tends to predict that all images have very low quality, while my model's predictions align more closely with both the self-labeled values and the synthetically generated values. This suggests that while ARNIQA is highly sensitive to distortions, it may not be as finely tuned to the specific context of teledermatology as my model.

To further validate ARNIQA's ability to differentiate image quality, I compared its predictions on 70 filtered good quality images with those on synthetically distorted images. As shown in Figure 5.1, ARNIQA can indeed distinguish between good and bad quality images, confirming its general effectiveness. However, my approach of using synthetically distorted images appears to improve performance in teledermatology-specific tasks, as my model's predictions align better with the expected outcomes in this domain.

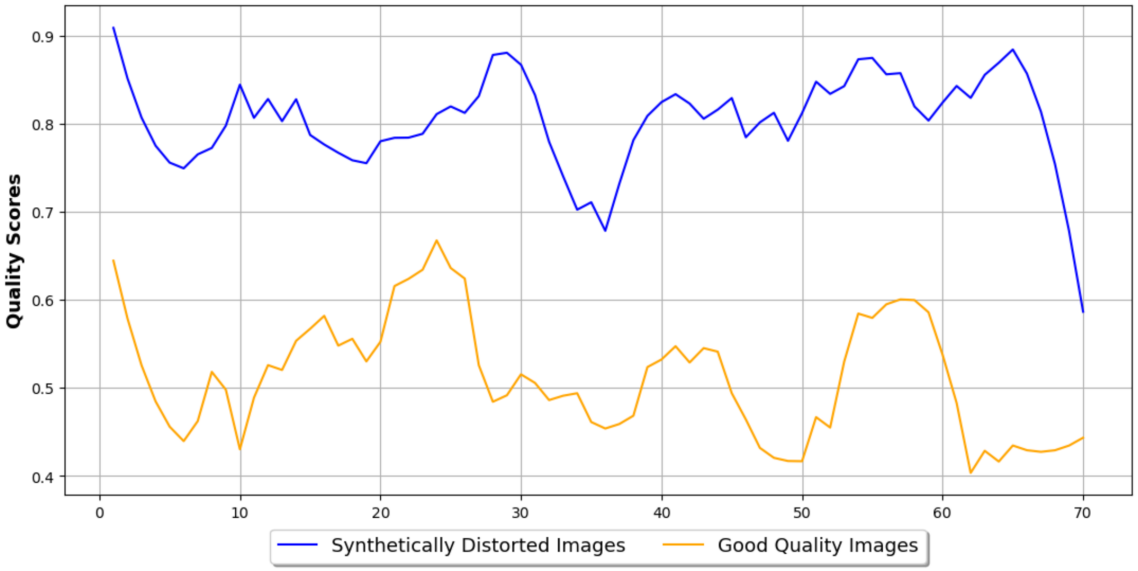


Figure 5.1: Comparison of ARNIQA's quality scores for 70 synthetically distorted images (blue) and 70 good quality images (orange). The scores range from 0 to 1, where higher values indicate better image quality.

These comparisons highlight the strengths of a domain-specific approach and suggest that while general image quality assessment models like ARNIQA are useful, customized models tailored to specific applications can offer significant improvements. My approach of synthetically distorting images and training the model specifically for teledermatology not only aligns better with the expected outcomes but also enhances the model's robustness and accuracy in this specific field.

5.2 Key Model Assumptions and Their Implications

The models assume that the features extracted by ARNIQA's backbone are comprehensive enough to capture the key distortions in teledermatology images. This assumption holds true for lighting, focus, color calibration, and resolution, covering 4 out of the 7 criteria. The remaining criteria: background, orientation, and field of view need further experimentation and fine-tuning. While the current performance indicates some level of effectiveness, more targeted data collection and model adjustments are necessary to fully validate these assumptions.

If the assumption that ARNIQA's backbone can capture all key distortions is not entirely correct, it would mean that the model might not perform well in real-world scenarios where these distortions are prevalent. To ensure these assumptions are valid, additional experiments with varied datasets and real-world images should be conducted. By expanding the variety of images used in training, particularly those with significant background presence, the model can be better equipped to handle real-world distortions. This is crucial for improving the model's robustness and generalizability.

While the backbone has proven effective for certain distortions, the uncertainties with background

and orientation distortions highlight the need for further refinement. Addressing these uncertainties through targeted data collection and further model tuning can enhance the overall performance and reliability of the image quality assessment in teledermatology. Overall, the ARNIQA backbone shows great potential for teledermatology applications, but continuous improvement and validation are essential to achieve the best possible performance.

5.3 Reviewing the Objectives of the Thesis

At the beginning of this thesis, the specific objectives were detailed:

- 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.

The first objective involved carrying out an in-depth review of the literature on image quality assessment methods and their application in teledermatology. This took a lot of time but was very important for the rest of the work. Through this review, key concepts such as IQA, teledermatology, ARNIQA, and related works were explored and documented.

The second objective was achieved during the literature review process. In this phase, the seven quality criteria from ISIC were identified and chosen as the best metrics for assessing the quality of dermatological images. This selection was critical for the next steps.

For the third objective, the performance of the selected image quality metrics was evaluated on dermatological datasets. These evaluations were thorough, involving tests on independent images not included in the model training. The datasets included both synthetically distorted images and images with authentic distortions, ensuring a complete assessment of how well the metrics worked.

The final objective was to develop a reproducible repository of image quality assessment tools and methods for teledermatology applications. This was successfully accomplished, making it possible for further experiments and research to build on this thesis. The repository provides a solid framework for future work in this field, ensuring that the methods and tools developed here can be effectively used and expanded upon.

5.4 Reflection

5.5 AI Tools Used

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

Chapter 6

Conclusion and Future Work

text

Chapter 7

Chapter

Ausblick
Reflexion der eigenen Arbeit, ungelöste Probleme, weitere Ideen.

7.1 Section

text

Bibliography

Agnolucci, L., Galteri, L., Bertini, M., & Del Bimbo, A. (n.d.). *ARNIQA: Learning Distortion Manifold for Image Quality Assessment*. arXiv: 2310.14918 [cs]. http://arxiv.org/abs/2310.14918

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (n.d.). *Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset*. arXiv: 2104.09957 [cs]. http://arxiv.org/abs/2104.09957

Hoffmann, C. P., Lennerts, S., Schmitz, C., Stölzle, W., & Uebernickel, F. (Eds.). (n.d.). *Business Innovation: Das St. Galler Modell*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07167-7

Lin, H., Hosu, V., & Saupe, D. (n.d.). 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

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. (n.d.). *Crowdsourcing Dermatology Images with Google Search Ads: Creating a Real-World Skin Condition Dataset*. arXiv: 2402.18545 [cs]. http://arxiv.org/abs/2402.18545