

# Contents

<b>1 Results and Analysis</b>	<b>2</b>
1.1 Range of Distortion Values . . . . .	2
1.2 Model Performance . . . . .	3
1.2.1 Parallel Coordinate Plot . . . . .	3
1.2.2 Loss Curve Analysis . . . . .	4
1.2.3 Performance Metrics . . . . .	4
1.2.4 Confusion Matrices . . . . .	5
1.3 Model Predictions . . . . .	6
1.3.1 Visualizations for Synthetic Distorted Images . . . . .	6
1.3.2 Visualizations for Authentic Images . . . . .	6
1.4 Assessing Training and Testing Images Quality . . . . .	9
1.4.1 Training Images Quality . . . . .	9
1.4.2 Test Images Quality . . . . .	11
1.5 Comparison with ARNIQA Predictions . . . . .	12
<b>2 Discussion</b>	<b>13</b>
2.1 Interpretation of Results . . . . .	13
2.1.1 Analysis of Individual Distortion Criteria . . . . .	13
2.1.2 Overall Model Performance on Test Sets . . . . .	14
2.1.3 Comparison with ARNIQA Predictions . . . . .	15
2.2 Key Model Assumptions and Their Implications . . . . .	16
2.3 Reviewing the Objectives of the Thesis . . . . .	17
2.4 Comparison with Related Work . . . . .	17
2.5 Reflection . . . . .	17
2.6 AI Tools Used . . . . .	17
<b>3 Conclusion and Future Work</b>	<b>19</b>
<b>4 Chapter</b>	<b>20</b>
4.1 Section . . . . .	20

# Chapter 1

## 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.

### 1.1 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.

## 1.2 Model Performance

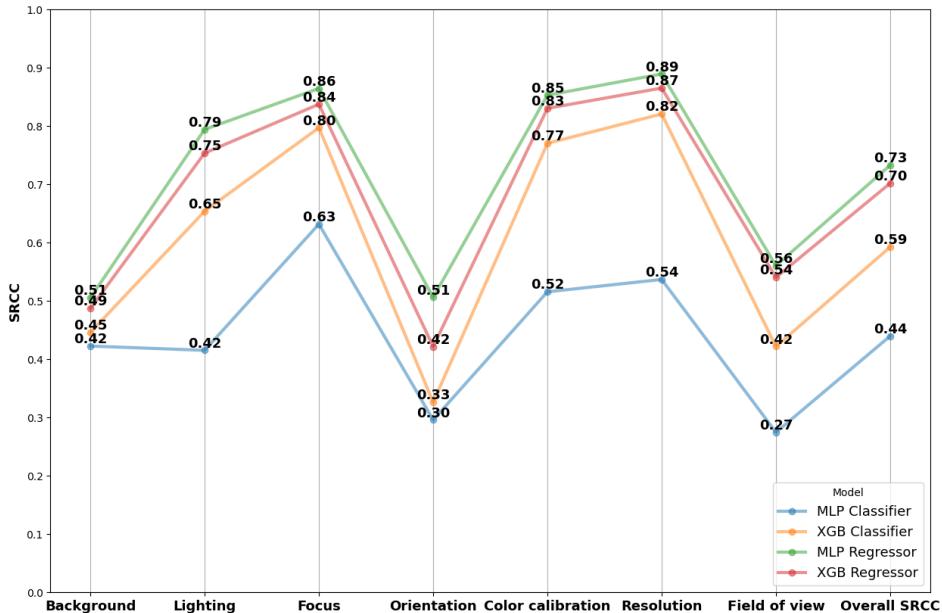
The performance of the four different models was evaluated through cross-dataset evaluation. This involved assessing the models on both the SCIN and Fitzpatrick (F17K) datasets after synthetic distortion, as summarized in Table 1.1. This table highlights how well the models generalize across different datasets.

**Table 1.1:** 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

### 1.2.1 Parallel Coordinate Plot

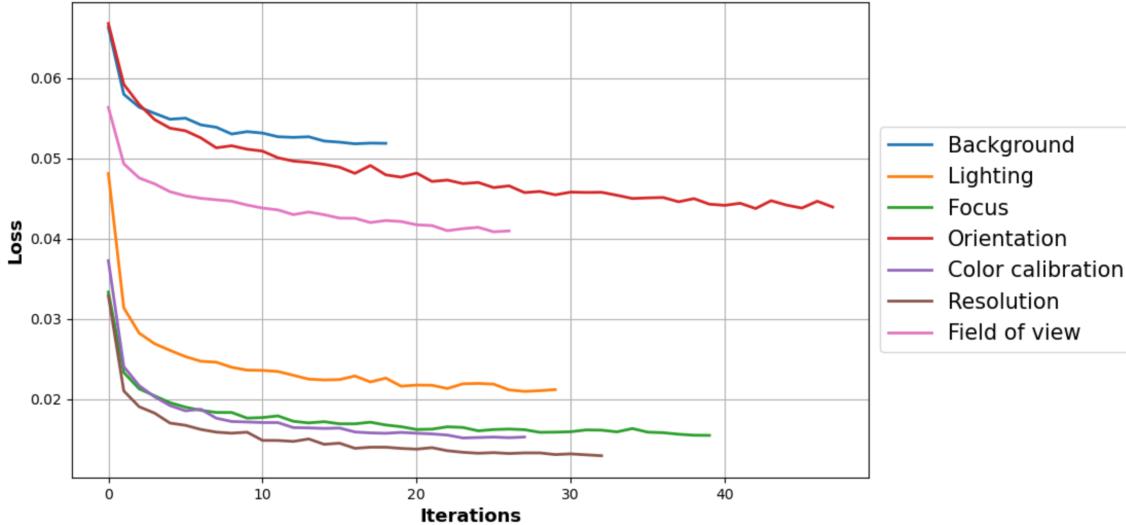
The parallel coordinate plot in Figure 1.1 compares the best-performing models across seven criteria, including the overall SRCC. This visualization highlights the performance of the MLP Regressor, showing that it consistently outperforms the other models.



**Figure 1.1:** 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.

### 1.2.2 Loss Curve Analysis

The loss curve in Figure 1.2 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 1.2:** 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.

### 1.2.3 Performance Metrics

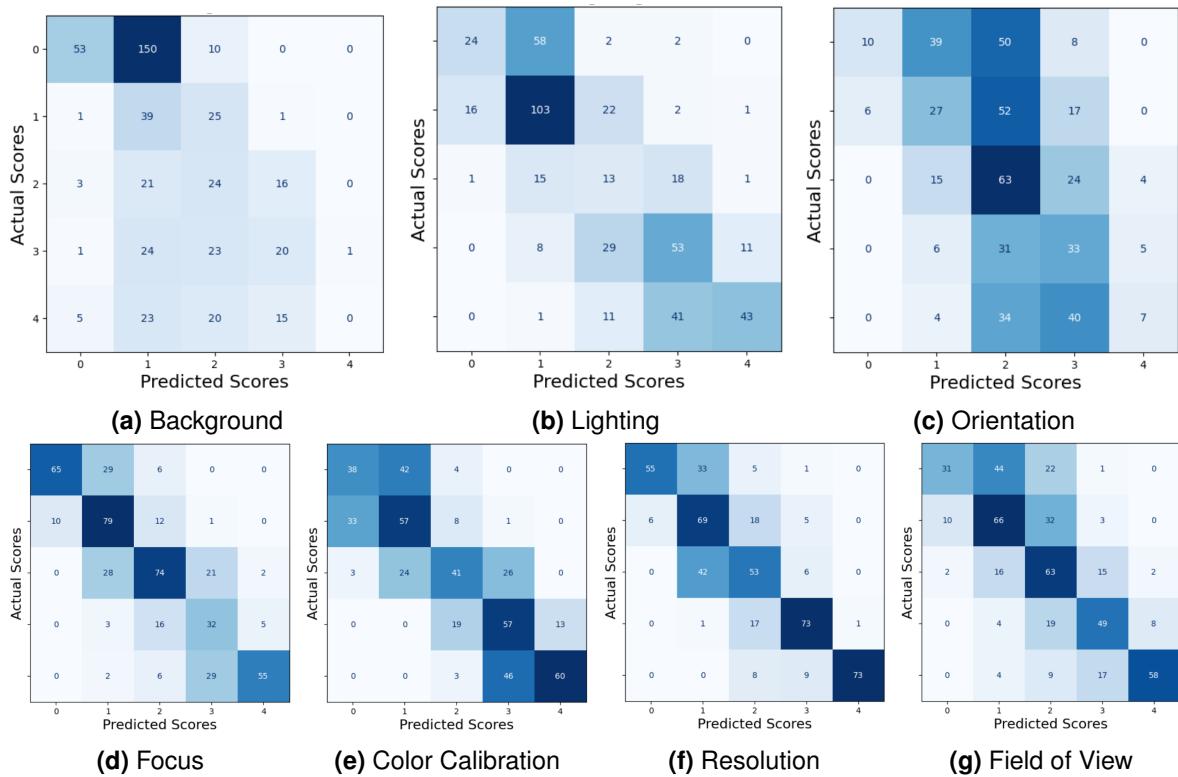
The performance of the final MLP regressor model on individual criteria is shown in Table 1.2. This table presents the performance metrics for the final MLP regressor model on 475 good quality Fitzpatrick images that were synthetically distorted. These metrics give a detailed view of the model's strengths and weaknesses.

**Table 1.2:** Performance Metrics for Each Distortion Criteria

Criteria	MAE	R <sup>2</sup>	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
<b>Overall</b>	<b>0.6195</b>	<b>0.5646</b>	<b>0.7507</b>	<b>0.7396</b>

### 1.2.4 Confusion Matrices

In addition to numerical metrics, confusion matrices<sup>1</sup> were created for each criterion, as shown in Figure 1.3. 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.



**Figure 1.3:** 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.

<sup>1</sup>from utils.visualization import plot\_all\_confusion\_matrices

## 1.3 Model Predictions

To better understand the model's performance on the two test sets (70 synthetic distorted images and 200 authentic images), radar charts<sup>2</sup> 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.

### 1.3.1 Visualizations for Synthetic Distorted Images

These visualizations, as shown in Figure 1.4, 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.

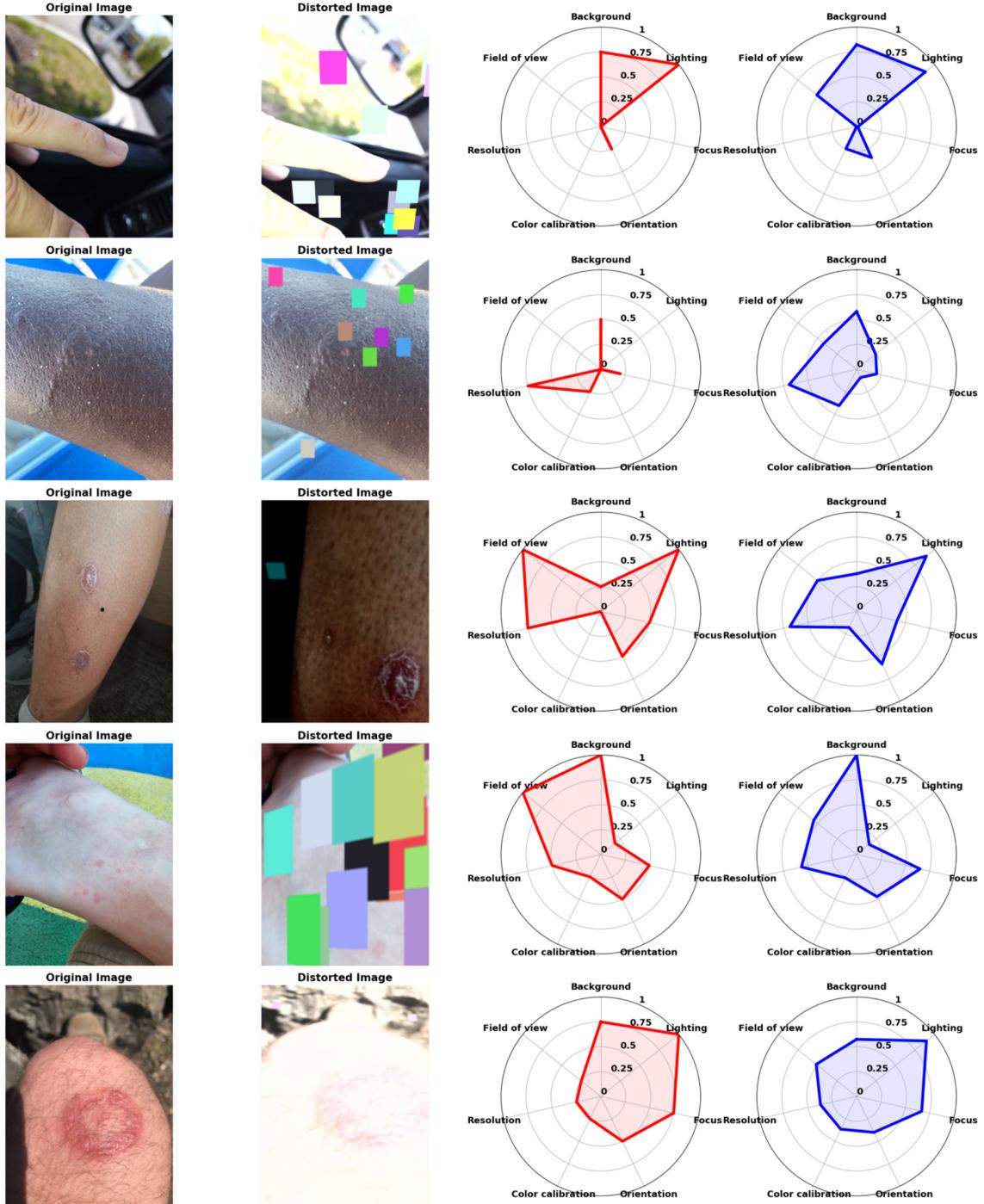
### 1.3.2 Visualizations for Authentic Images

The visualizations, as shown in Figure 1.5, compare the model's predictions with human-labeled scores for authentic images. This highlights the model's performance in real-world scenarios.

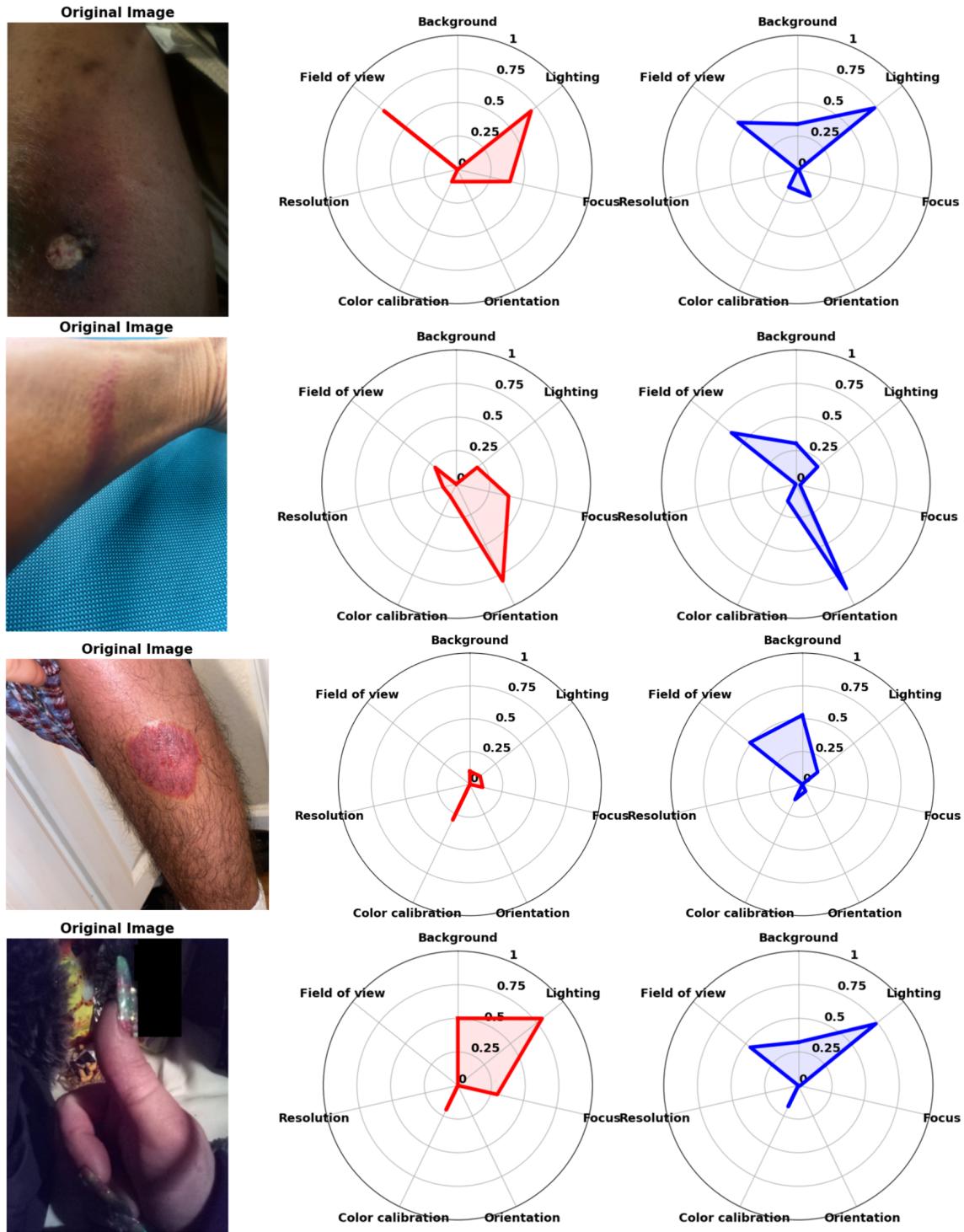
The first column shows the image, the second column displays the human-labeled scores, and the third column presents the model's predictions. This comparison helps show how well the model's predictions align with the human evaluations.

---

<sup>2</sup>from utils.visualization import plot\_results



**Figure 1.4:** 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.



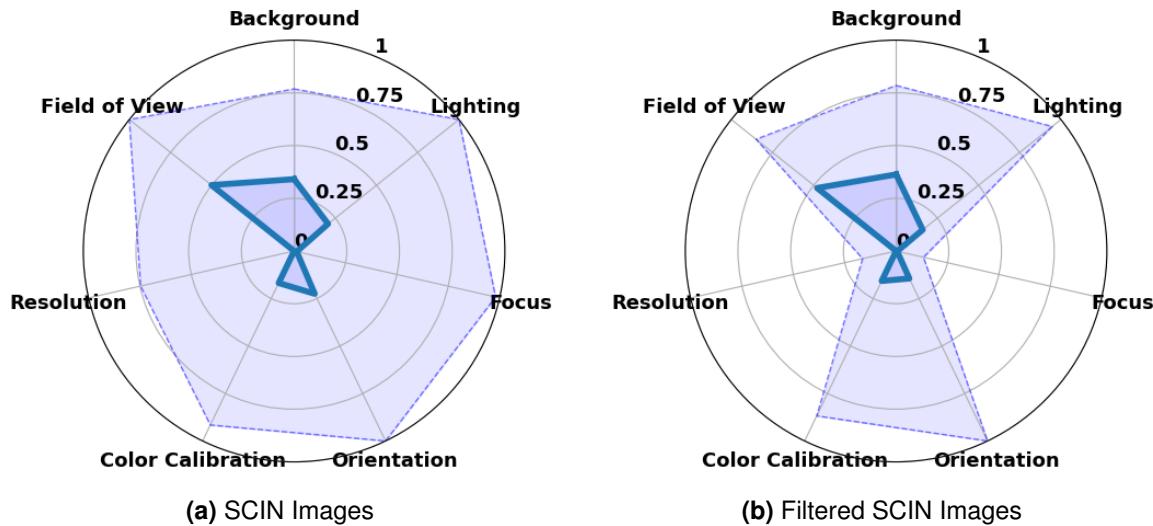
**Figure 1.5:** Visualizations for the MLP Regressor model on 200 authentic images. The three-column layout shows the image, the human-labeled scores, and the model's predictions.

## 1.4 Assessing Training and Testing Images Quality

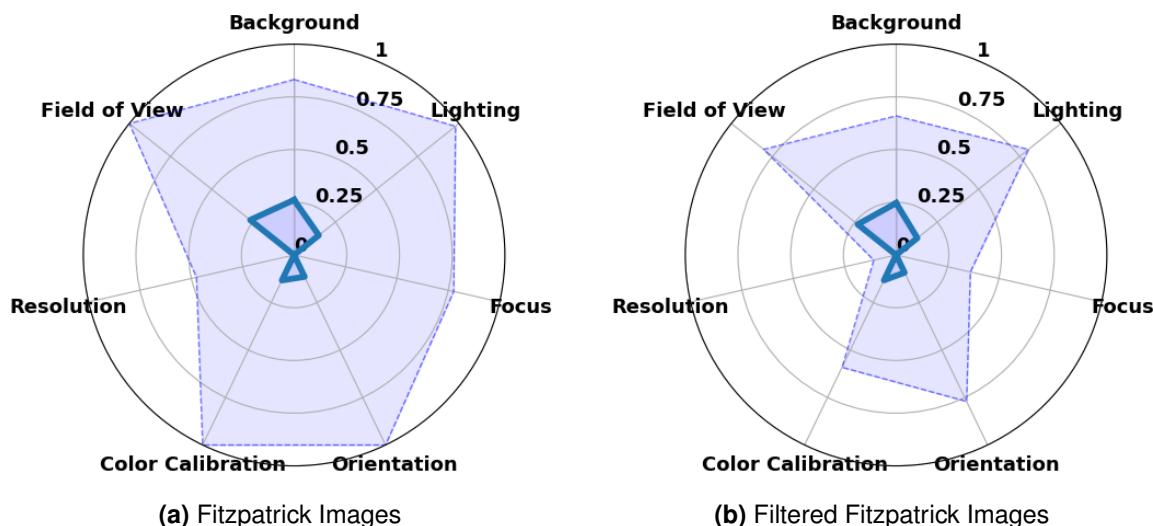
To verify the quality of the images used for training and see how they change after synthetic distortion, radar charts were created. These charts show the quality of the original training images and how they are affected by the distortions. Additionally, the quality of both the synthetic and authentic 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.

### 1.4.1 Training Images Quality

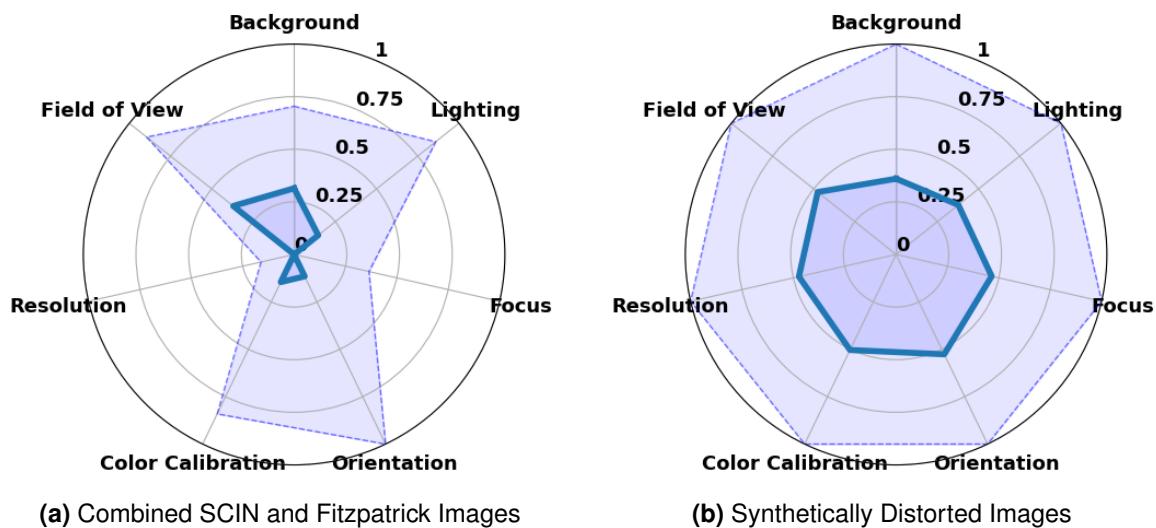
Figure 1.6 and Figure 1.7 show the quality of the original SCIN and Fitzpatrick images, and the filtered good quality images, respectively. Figure 1.8 shows the quality of the combined SCIN and Fitzpatrick images and the synthetically distorted images used for training the model.



**Figure 1.6:** Radar charts for the SCIN dataset. (a) Original images from the SCIN dataset (10'379 images). (b) Filtered good quality images (475 images).



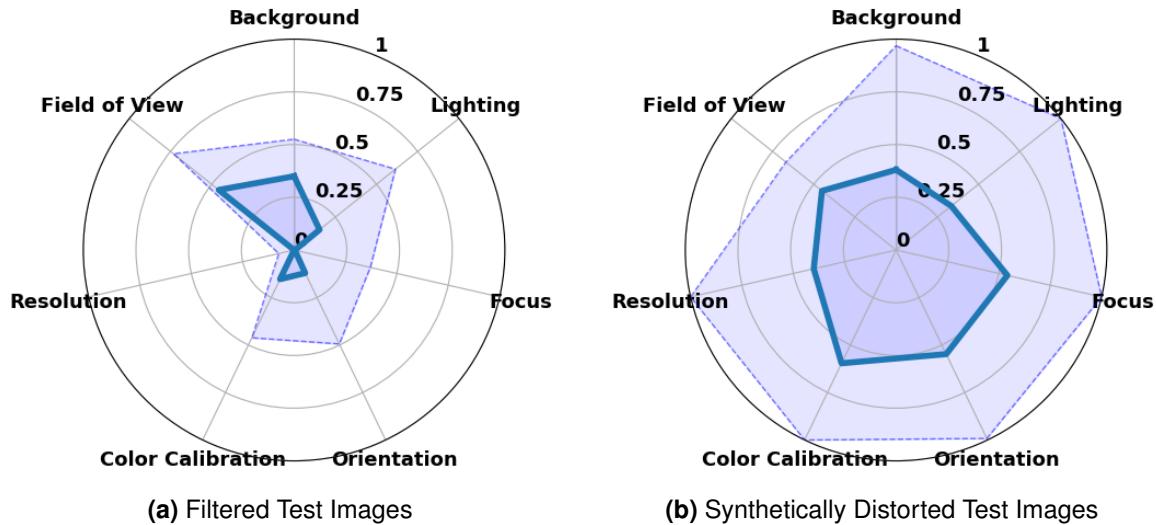
**Figure 1.7:** Radar charts for the Fitzpatrick dataset. (a) Original images from the Fitzpatrick dataset (16'577 images). (b) Filtered good quality images (475 images).



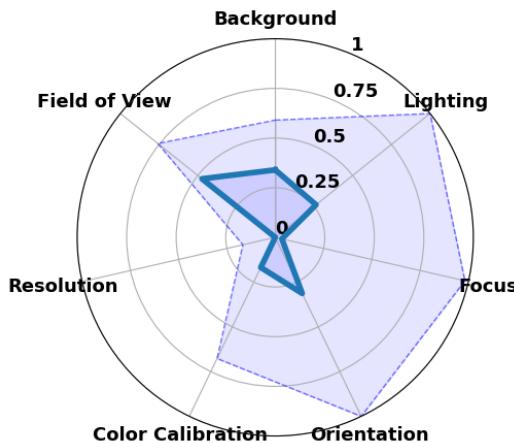
**Figure 1.8:** Combined dataset analysis. (a) Combined SCIN and Fitzpatrick images (950 images). (b) Synthetically distorted images.

### 1.4.2 Test Images Quality

Figure 1.9 shows the quality of the filtered good quality test images and the synthetically distorted test images. Figure 1.10 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.



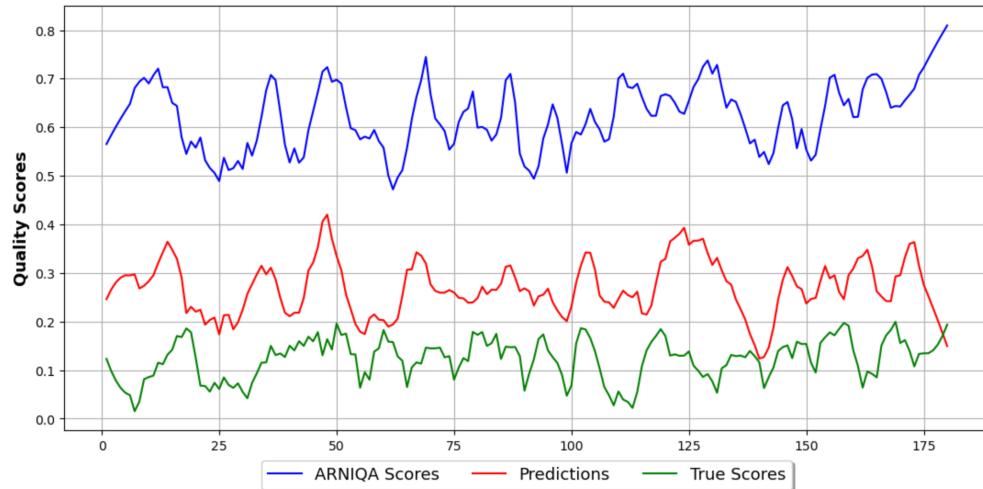
**Figure 1.9:** Synthetic test set analysis. (a) Filtered good quality test images (70 images, independent of training set). (b) Synthetically distorted test images.



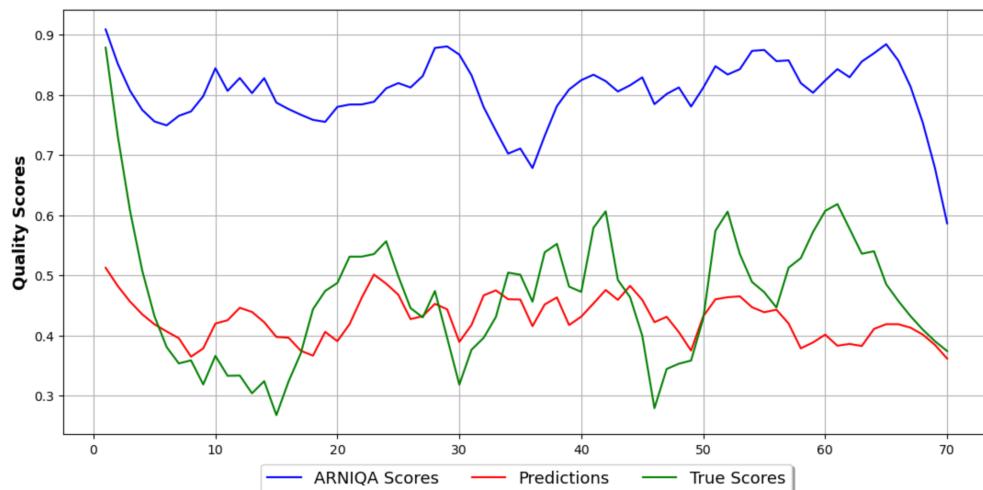
**Figure 1.10:** Authentic test set from the SCIN dataset, independent of the training images, showing real-world distortions.

## 1.5 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 1.11:** 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 1.12:** 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 2

## Discussion

The chapter on Discussion covers the entire work and, in particular, the results achieved in Chapter 1. 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.

### 2.1 Interpretation of Results

The differences in model performance shown in Figure 1.1 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.

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.

check this paragraph

#### 2.1.1 Analysis of Individual Distortion Criteria

When we analyze the metrics in Table 1.2 alongside the confusion plots in Figure 1.3, 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 1.8b, 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 2.1 and compare Figure 1.6 with Figure 1.7).

**Table 2.1:** 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 <sup>2</sup>	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.

### 2.1.2 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 1.4. 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 1.4 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 1.3, 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.

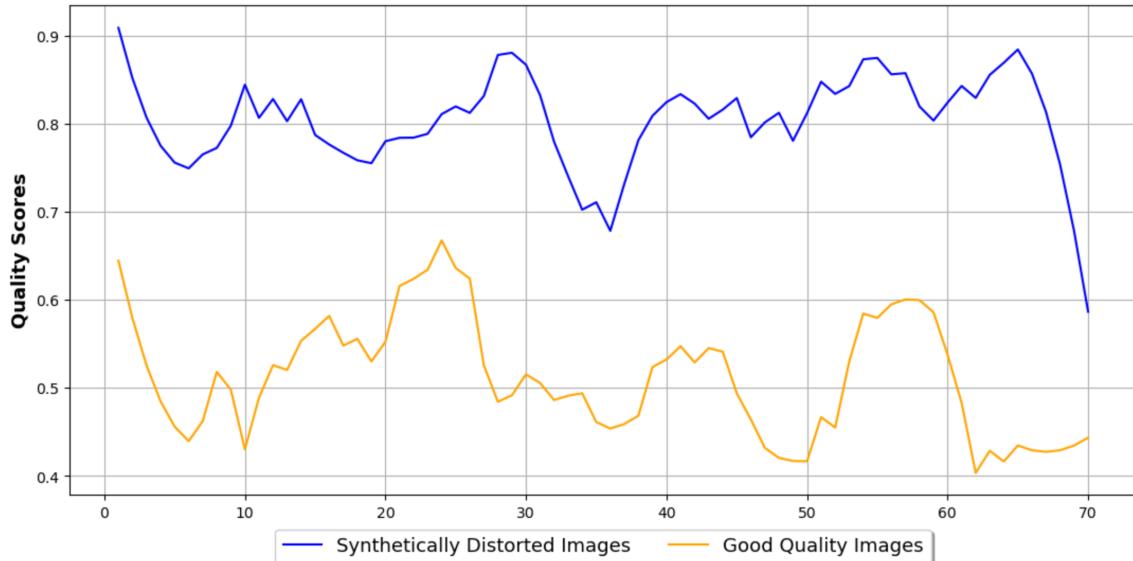
Despite these challenges, if we take out the field of view distortion, the model's predictions are quite accurate. This distortion significantly affects the radar charts. I chose not to label the images again to avoid "leaking information" from the test set into the training process, keeping the evaluation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

#### 2.1.3 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 1.11 and Figure 1.12 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 2.1, 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 2.1:** 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.

## 2.2 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.

## 2.3 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.

## 2.4 Comparison with Related Work

## 2.5 Reflection

## 2.6 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 3**

## **Conclusion and Future Work**

text

# **Chapter 4**

## **Chapter**

Ausblick

Reflexion der eigenen Arbeit, ungelöste Probleme, weitere Ideen.

### **4.1 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>