Face Image Quality - A Comparative Study of Motion Blur Detection Algorithms

Muhamad Nadali[1], Wassim Kabbani[2], and Christoph Busch[2]

[1] Technical University of Denmark (DTU), Lyngby, Denmark
mn@nadali.dk
[2] Norwegian Biometrics Laboratory (NBL), Norwegian University of Science and Technology (NTNU), Gjøvik, Norway
{wassim.h.kabbani, christoph.busch}@ntnu.no

Abstract. Motion blur degrades face image quality and impairs recognition accuracy. This paper evaluates five face image quality assessment (FIQA) algorithms for motion blur detection, focusing on accuracy and demographic fairness. Experiments on the EDAMB and MST-E datasets employed Kullback–Leibler (KL) divergence to compare algorithm scores against expert consensus, partial area under the curve (pAUC) from error-versus-discard curves to report prediction of recognition performance, and the Gini coefficient to assess fairness. Densenet169 had the lowest KL divergence, while CNN-R showed the best predicting performance, achieving the lowest pAUC. Fusing CNN-R with Densenet161 further reduced the pAUC by 1.3%. The fairness analysis found that the Fourier Transform and CNN-R methods were the most fair, whereas Laplace was the least fair.

Keywords: Face Image Quality Assessment · Face Recognition Systems · Motion Blur Detection · Demographic Fairness · Human Expert Consensus.

1 Introduction

Face Recognition Systems (FRS) are widely used in smartphone authentication, surveillance, and passport control [20,1]. Despite recent advances, motion blur remains a significant challenge for FRS performance, as it degrades image quality and recognition accuracy [13,19]. Motion blur occurs when the subject or the camera moves during image capture. An object's motion is then projected onto the camera, leading to a blurred effect that traces the object's path. In low light conditions with long exposure times, the object's path is extended in the image, leading to an increased blurring [10]. Improving motion blur detection in FRS is crucial for reducing false non-match rates (FNMR) and enhancing robustness [17,7]. This work evaluates state-of-the-art face image quality assessment (FIQA) algorithms for assessing motion blur intensity and compares their scores to a human expert consensus on the Essen Darmstadt Motion Blur (EDAMB) dataset

[18,21]. This dataset contains 1,341 manually labeled images scored from 10 to 90, where 10 represents full blur and 90 represents no blur. While EDAMB offers valuable insights, its lack of demographic diversity limits fairness evaluation. To address this, the Monk Skin Tone Examples (MST-E) dataset examines algorithm performance across diverse skin tones. The top-performing algorithms are fused and benchmarked to determine whether this fusion improves accuracy. By comparing the algorithms' scores with the human expert labels and evaluating results on EDAMB and MST-E, this study identifies the strengths and limitations of current motion blur detection techniques, including considerations of demographic fairness. a metric to best predict recognition performance. Human Expert Consensus is fusing scores through a weighting method. Additionally, the MST-E dataset will be employed alongside a state-of-the-art fairness metric to assess the fairness of FIQA algorithms across three skin tones. These benchmarks will be applied to determine whether an FIQA algorithm is suitable as an accurate Motion Blur Detector for an FRS system.

2 Related Work

In this work, five FIQA algorithms are compared for motion blur detection, including both classical methods and deep learning approaches. Labels from five human experts in the EDAMB dataset were fused with weighting based on Z-scores to create a ground truth while reducing individual bias. The evaluation is conducted on two datasets and encompasses accuracy, fairness measured using the Gini coefficient, and prediction of recognition performance through error-versus-discard curves. Recent studies have explored score-level fusion in biometric systems. Schlett et al. [23] used Z-score normalization to align quality scores across FIQA models, reducing variation in expert opinions. Min-max normalization is used to rescale scores to a common range, ensuring consistency across expert opinions. Schlett et al. [23] demonstrated the effectiveness of this normalization method for preparing quality scores. The Face Image Quality Assessment Toolkit (fiqat)[3], developed by Schlett et al. [24], supports face recognition and quality assessment experiments. It includes face detection, main face selection, cropping, alignment, trait extraction, and EDC curve generation. Schlett et al. [22] used EDC curves to evaluate FIQA algorithms, introducing partial Area Under Curve (pAUC) as a metric to best predict recognition performance. pAUC focuses on the 0–20% discard range, relevant for operational settings [24]. The Kullback-Leibler (KL) divergence measures the difference between two probability distributions. Introduced by Kullback and Leibler [14], it is widely used in information theory and for evaluating consistency in expert and subjective assessments. Kabbani et al. [11] examined age, gender, and skin tone biases in FIQA algorithms, showing that even high-performing systems can be biased, particularly against individuals with darker skin tones. Dörsch et al. [4] introduced statistical measures to evaluate fairness, noting that higher sample rejection rates in certain demographic groups can lead to bias in systems such

[3] `https://hda10196.h-da.io/face-image-quality-toolkit/source/readme.html`

as border control. Merkle et al. [16] highlighted that motion blur and imaging defects can impact performance differently across skin tones, and they advocated for quality metrics that account for demographic variations.

3 Experimental Setup

Experiments were conducted using two datasets: **EDAMB** (real motion blur, 1,341 expert-annotated images) [18,21] and **MST-E** (synthetic motion blur, diverse skin tones) [25]. Images from both datasets were preprocessed consistently using SCRFD from the `fiqat` toolkit, followed by cropping, padding, and resizing to 500×500 pixels.

Five FIQA algorithms, which are described in Section 3.1, were evaluated. Algorithm performance was evaluated using three metrics: KL divergence (to measure alignment with human expert consensus), error-versus-discard characteristic (EDC) curves with partial Area Under the Curve (pAUC) to quantify prediction of recognition accuracy, and the Gini coefficient to evaluate fairness across demographic groups.

The human expert consensus scores were obtained by normalizing the expert labels and fusing the scores with a weighted averaging method. Algorithm fusion was applied using a Z-score-weighted averaging approach to assess potential improvements in accuracy.

The Essen Darmstadt Motion Blur Dataset (EDAMB) consists of 35 subjects captured with varying levels of motion blur, primarily linear but including some complex rotational movements, reflecting real-world conditions. Images were collected in Essen (9 subjects) and Hochschule Darmstadt (26 subjects) using three cameras: Canon EOS50D, Canon PowerShot SC200IS, and Kodak EasyShare DX6490. Five experts labeled each image on a scale from 0 (fully blurred) to 100 (no blur), though some experts used a 10 to 90 scale. After preprocessing with the `fiqat` toolkit, 1,154 images were usable for analysis.

(A) High blur **(B)** Medium blur **(C)** Low blur

Fig. 1. Examples of real motion blur from EDAMB dataset at different blur levels after cropping, preprocessing the images.

Monk Skin Tone Examples Dataset (MST-E), introduced by Google, includes images and videos of 19 individuals photographed under various conditions, including accessories such as masks and glasses. While MST-E enables evaluation across diverse skin tones, it lacks inherent motion blur. Synthetic motion blur was applied using ImageMagick's `-motion-blur` option, varying intensity from 0 to 100, following methods from the NIST FATE Quality SIDD Report [28]. The dataset includes metadata for each image, which was used to filter for frontal, facing-camera images while excluding samples with accessories. After preprocessing, 409 suitable images were retained and categorized into three MST groups: light tones (MST 1–3) with 133 images, medium tones (MST 4–6) with 150 images, and dark tones (MST 7–10) with 125 images. The distribution across these groups ensured balanced representation for comprehensive evaluation.

3.1 FIQA Algorithms for Motion Blur Estimation

Five FIQA algorithms were evaluated, each providing an automatic quality score for a face image (higher score = better quality/less blur).

- Laplace [21]: The algorithm evaluates image quality using Laplacian variance for sharpness, Sobel operator for edge strength, and Fourier Transform analysis for blurriness [26,2,15]. Each method produces a score reflecting image clarity. These are combined to quantify overall quality based on sharpness, edge visibility, and blur.
- FourierTransform [18]: The algorithm converts the image to luminance and applies Fourier and cepstral analysis to assess brightness variations [8]. It then uses the Radon Transform to detect structural lines [3], followed by peak analysis to identify significant features. These steps are combined to produce a final quality score.
- CNN-R [21]: Each image is divided into patches to capture localized quality features [6]. A pretrained CNN (typically with 50x50 patches) then predicts quality scores based on learned features [12]. These scores are aggregated to produce a final quality score per image.
- Densenet161 [18]: A FIQA algorithm based on a Densenet161 deep neural network architecture. Densenets are deep CNNs with densely connected layers. This model, pre-trained and/or fine-tuned for face image quality, outputs a blur-related quality score [27].
- Densenet169 [18]: Similar to Densenet161, but using a slightly different DenseNet (169-layer version). It provides another deep learned scoring mechanism for image quality for blur [5,9].

All five algorithms take a face image (cropped and aligned) as input and output a numeric quality score. For consistency, we normalized the algorithms' output scales if needed, so that their scores are in the 0–100 range.

4 Establishing Human Expert Consensus

Establishing a reliable human expert consensus serves as the foundation for objectively evaluating algorithm performance. By comparing scores by algorithms to this human expert consensus, we assess how well each model aligns with human perception of motion blur. To evaluate algorithm performance, the consensus value was established from five human experts' opinions of the EDAMB dataset. Since experts used different scoring scales (e.g., 10–90 or 0–100), all scores were first normalized to a common 0–100 range using min-max normalization:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \times 100 \tag{1}$$

This ensured comparability across all expert scores.

To fuse the scores, a Z-score-based weighted average was used, giving higher weight to scores closer to the group mean:

$$\text{Weight}_Z = \frac{1}{1 + |Z|} \tag{2}$$

$$\text{Weighted Average}_{\text{Zscore}} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i} \tag{3}$$

Where:

- w_i is the weight assigned to expert i,
- x_i is the score provided by expert i,
- n is the total number of experts.

This approach ensures that outlier-like scores have reduced influence on the final consensus value, promoting a consensus-based, statistically grounded fusion.

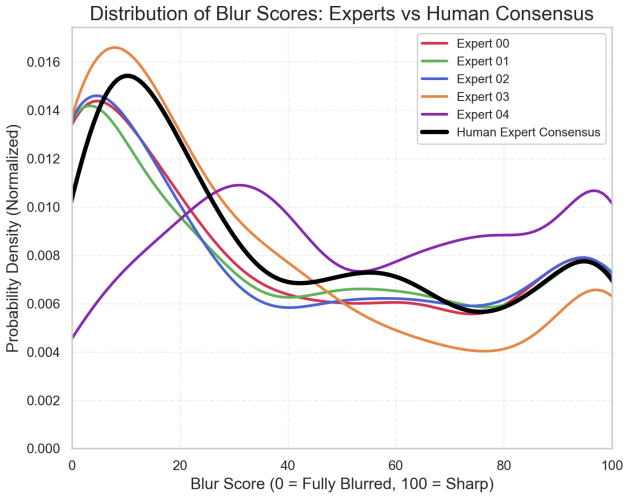

Fig. 2. Kernel Density Estimate plot showing the distribution of blur scores given by individual experts and the human expert consensus (in black).

Figure 2 is a KDE plot of the labels from five expert and their consensus. The KDE plot shows the variations of each expert's labels and the result of human expert consensus based on these five experts. The KDE plot also shows Expert 04 curve is different from the rest of the experts, suggesting a greater tolerance for blur or a differing interpretation of motion blur.

5 Evaluating Motion Blur FIQA Algorithms

The KL divergence analysis quantifies the alignment between each algorithm's score distribution and the consensus value, with values ranging from 0.059 to 2.443. Lower values indicate a stronger alignment, where a KL divergence below 0.1 suggests high similarity, while values above 1.0 reflect substantial distribution differences.

Comparison	KL Divergence ↓
CNN-R	1.535
Laplace	0.586
Densenet161	0.093
Densenet169	**0.059**
Fourier Transform	2.443

Table 1. KL Divergence results for various models compared to the consensus value.

As shown in Table 1, Densenet169 achieved the most substantial alignment with consensus value (0.0588), followed closely by Densenet161 (0.0927). Laplace showed moderate divergence (0.5862), while CNN-R and Fourier Transform demonstrated the highest deviation with values above 1.5.

Figure 3 illustrates the EDC curves for the FIQA algorithms, which assess the quality of face recognition images based on motion blur. This figure effectively illustrates the variation in FNMR as the proportion of discarded images increases, specifically focusing on the discard range of 0% to 20%.

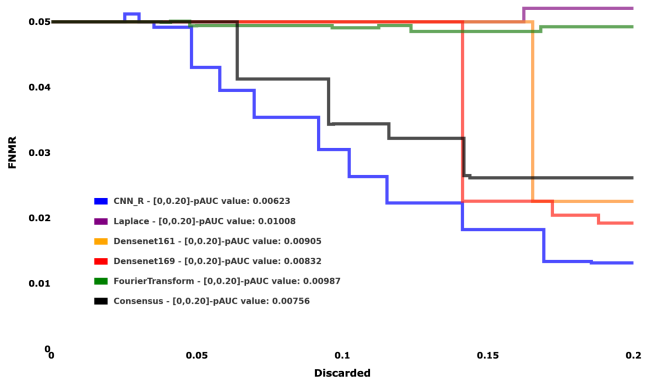

Fig. 3. The EDC curve represents data from the five algorithms, the consensus value for comparison. The x-axis is limited to 0.2, as values beyond 20% are discarded

The performance of these algorithms is quantified using the pAUC values noted beside each curve label, where lower values signify more accurate recognition of face images.

The CNN-R curve shows a lower FNMR across the discard spectrum, demonstrating higher consistency and reliability in retaining high-quality images. The other algorithms, such as Laplace and Densenet169, show a steeper rise in FNMR as more images are discarded, suggesting they may mistakenly discard images that, despite perceived lower quality, are crucial for accurate biometric verification.

The fusion methodology combined the outputs of multiple algorithms using a Z-score weighted averaging approach. This method assigned higher weights to algorithms whose outputs aligned more closely with the overall distribution, ensuring that outlier predictions had less influence on the final score. A total of 26 possible algorithm combinations were evaluated to determine the most effective fusion. The fused results were evaluated using the same performance metrics as those used for the individual algorithms. The analysis demonstrated that the Z-score weighted fusion consistently outperformed the single-algorithm approaches across all evaluation metrics.

Combination	pAUC $[0, 0.20] \downarrow$
CNN-R + Densenet161	**0.00615**
CNN-R + Densenet169	0.00618
CNN-R + Densenet161 + Densenet169	0.00629

Table 2. Top 3 performing methods based on lowest pAUC scores within the [0, 0.20] discard fraction range. Lower values indicate better error reduction.

Among the tested combinations, the fusion of CNN-R and Densenet161 yielded the most effective results. Specifically, the Z-score-weighted fusion of these two models achieved the lowest pAUC value of 0.00615, representing an improvement of approximately 1.3% compared to the best individual algorithm.

6 Fairness Evaluation of FIQA algorithm

We group the results by skin tone to assess fairness, using the MST-E dataset's skin tone labels (based on the Monk Skin Tone scale). We define an algorithm to be fair if its performance remains consistent across all skin tone categories (e.g., error rates, discard behavior) is consistent across these groups. No additional training or adaptation was done for different demographics; We applied each algorithm to the entire MST-E dataset and then analyzed the results for each subgroup. We use the traditional Gini Coefficient (GC) as a fairness metric to evaluate potential demographic biases in motion blur quality assessment, following the framework introduced by Dörsch et al. [4]. The GC quantifies inequality in quality score distributions across demographic groups and produces a normalized score between 0 and 1, where 1 indicates perfect fairness. As recommended by the authors, mean quality scores are used to capture subtle demographic differences better. This measure assesses whether the NBL algorithms perform consistently across demographic groups in the MST-E dataset.

In table 3, the FourierTransform algorithm demonstrated superior fairness with a score of 0.94, significantly outperforming other approaches. In contrast, the Laplace algorithm showed the lowest fairness score of 0.12, showing significant differences in quality assessments across groups. The CNN-R algorithm achieved the second-best performance with a score of 0.82, while Densenet161 and Densenet169 showed moderate fairness levels of 0.71 and 0.64, respectively.

Algorithm	Fairness Score (GC) ↑
Densenet161	0.71
Densenet169	0.64
CNN-R	0.82
Fourier Transform	**0.94**
Laplace	0.12

Table 3. Fairness scores calculated using the Gini coefficient, where higher scores indicate greater fairness. Fourier Transform is the fairest algorithm, while Laplace is the least fair.

7 Conclusion and Future Work

The evaluation revealed that the CNN-R model consistently delivered the most reliable performance on the EDAMB dataset for motion blur detection. It showed strong agreement with the human expert consensus and achieved the best results when evaluated using the EDC curves. Although the CNN-R model did not achieve the lowest KL divergence, indicating some divergence from the human expert consensus, this difference was minor and did not negatively impact its overall performance. CNN-R demonstrated stable and accurate performance across varying levels of motion blur, suggesting that it may better distinguish between usable and unusable images than human experts. This implies that the five experts may tend to label certain moderately blurred images as too degraded for recognition, whereas CNN-R assigns more appropriate quality scores that preserve biometric utility. Additionally, fairness evaluation showed good fairness scores across skin tone groups, suggesting that the CNN-R is not biased toward specific demographic groups. Algorithm fusion was also explored as a potential performance enhancement technique. For the EDAMB dataset, fusing CNN-R with Densenet161 resulted in a slight increase in performance, with its lower pAUC value. Despite these findings, the dataset we used is limited. The MST-E dataset relied on synthetically blurred images and lacked genuine motion blur, which may not fully capture real-world blur characteristics; additionally, the EDAMB dataset's limited demographic variety constrained the scope of the fairness analysis. Future work will benefit by addressing these limitations by expanding the evaluation to more diverse, real-world blur datasets and by developing fairness-aware FIQA training strategies to improve algorithmic consistency across demographic groups.

Acknowledgements

This research has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101121280 (EINSTEIN). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect the views of the EU/Executive Agency. Neither the EU or the granting authority can be held responsible for them.

References

1. Chaki, N.M., Ashour, M.W.: Automated border control systems: A literature review. Iet Conference Proceedings **2021**(11), 152–157 (2021). https://doi.org/10.1049/icp.2022.0331
2. Chaple, G.N., Daruwala, R.D., Gofane, M.S.: Comparisions of robert, prewitt, sobel operator based edge detection methods for real time uses on fpga. Proceedings - International Conference on Technologies for Sustainable Development, Ictsd 2015 p. 7095920 (2015). https://doi.org/10.1109/ICTSD.2015.7095920
3. Cho, T.S., Paris, S., Horn, B.K.P., Freeman, W.T.: Blur kernel estimation using the radon transform. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 241–248 (2011). https://doi.org/10.1109/CVPR.2011.5995479
4. Dörsch, A., Schlett, T., Munch, P., Rathgeb, C., Busch, C.: Fairness measures for biometric quality assessment. In: Proceedings of the 27th International Conference on Pattern Recognition (ICPR). Springer (2024). https://doi.org/10.1007/978-3-031-87657-8_20, https://link.springer.com/chapter/10.1007/978-3-031-87657-8_20
5. Guo, N., Qingge, L., Huang, Y.C., Roy, K., Li, Y.G., Yang, P.: Blind image quality assessment via multiperspective consistency. International Journal of Intelligent Systems **2023**(1), 4631995 (2023). https://doi.org/10.1155/2023/4631995
6. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-**, 2424–2433 (2016). https://doi.org/10.1109/CVPR.2016.266
7. Huo, L., Xiong, Y., Sun, J., Nie, Y., Su, W.: Unsupervised face image quality assessment based on face recognition. Mechatronics and Automation Technology, J. Xu (Ed.), IOS Press pp. 77–82 (2022). https://doi.org/10.3233/ATDE221152
8. Ibrar-Ul-Haque, M., Qadri, M.T., Siddiqui, N.: Image quality assessment using image details in frequency domain. Mehran University Research Journal of Engineering and Technology **36**(4), 8 pp. (2017). https://doi.org/10.22581/muet1982.1704.04
9. Jiang, T., Hu, X.j., Yao, X.h., Tu, L.p., Huang, J.b., Ma, X.x., Cui, J., Wu, Q.f., Xu, J.t.: Tongue image quality assessment based on a deep convolutional neural network. BMC Medical Informatics and Decision Making **21**(1), 147 (2021). https://doi.org/10.1186/s12911-021-01508-8
10. Joshi, N.: Motion Blur, pp. 815–818. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-63416-2_512
11. Kabbani, W., Raja, K., Ramachandra, R., Busch, C.: Demographic variability in face image quality measures. In: BIOSIG 2024. pp. 1–12. Gesellschaft für Informatik (2024)
12. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 1733–1740 (2014). https://doi.org/10.1109/CVPR.2014.224
13. Knežević, K., Mandić, E., Petrović, R., Stojanović, B.: Blur and motion blur influence on face recognition performance. In: 2018 14th Symposium on Neural Networks and Applications (NEUREL). pp. 1–6. IEEE (2018). https://doi.org/10.1109/NEUREL.2018.8587012

14. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22**(1), 79–86 (1951). https://doi.org/10.1214/aoms/1177729694
15. Mavridaki, E., Mezaris, V.: No-reference blur assessment in natural images using fourier transform and spatial pyramids. 2014 Ieee International Conference on Image Processing, Icip 2014 pp. 566–570 (2014). https://doi.org/10.1109/ICIP.2014.7025113
16. Merkle, J., Rathgeb, C., Tams, B., Lou, D.P., Dörsch, A., Drozdowski, P.: Facial metrics for ees: State of the art of quality assessment of facial images. Tech. rep., Federal Office for Information Security (2022)
17. Nouyed, I., Zhang, N.: Face image quality enhancement study for face recognition. arXiv preprint arXiv:2307.05534 (2023), https://arxiv.org/abs/2307.05534
18. Pedersen, J.B.: Motion Blur Estimation for Face Image Quality Assessment. Master's thesis, Technical University of Denmark, Kongens Lyngby, Denmark (Jul 2024)
19. Punnappurath, A., Rajagopalan, A.N., Taheri, S., Chellappa, R., Seetharaman, G.: Face recognition across non-uniform motion blur, illumination, and pose. IEEE Transactions on Image Processing **24**(7), 2067–2082 (July 2015)
20. Qinjun, L., Tianwei, C., Yan, Z., Yuying, W.: Facial recognition technology: A comprehensive overview. Academic Journal of Computing & Information Science **6**(7), 15–26 (2023). https://doi.org/10.25236/AJCIS.2023.060703
21. Schachner, S.: Analysis of Motion Blur in Biometric Facial Images. Master's thesis, Hochschule Darmstadt, Fachbereich Informatik, Darmstadt, Germany (Aug 2024), master of Science (M.Sc.)
22. Schlett, T., Rathgeb, C., Henniger, O., Galbally, J., Fierrez, J., Busch, C.: Face image quality assessment: A literature survey. ACM Computing Surveys (CSUR) (December 2021)
23. Schlett, T., Rathgeb, C., Tapia, J., Busch, C.: Evaluating face image quality score fusion for modern deep learning models. In: Intl. Conf. of the Biometrics Special Interest Group (BIOSIG). pp. 1–8. LNI, GI (September 2022)
24. Schlett, T., Rathgeb, C., Tapia, J., Busch, C.: Considerations on the evaluation of biometric quality assessment algorithms. Trans. on Biometrics, Behavior, and Identity Science (TBIOM) (December 2023)
25. Schumann, C., Ollanubi, G., Wright, A., Monk, E., Heldreth, C., Ricco, S.: Consensus and subjectivity of skin tone annotation for ml fairness. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS). pp. 30319–30348 (2023)
26. So, C.W., Yuen, E.L.H., Leung, E.H.F., Pun, J.C.S.: Solar image quality assessment: a proof of concept using variance of laplacian method and its application to optical atmospheric condition monitoring (2024). https://doi.org/10.48550/arXiv.2405.11490
27. Tian, H., Zhang, B., Zhang, Z., Xu, Z., Jin, L., Bian, Y., Wu, J.: Densenet model incorporating hybrid attention mechanisms and clinical features for pancreatic cystic tumor classification. Journal of Applied Clinical Medical Physics **25**(7), e14380 (2024). https://doi.org/10.1002/acm2.14380
28. Yang, J., Grother, P., Ngan, M., Hanaoka, K., Hom, A.: Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment. Tech. Rep. NIST IR 8485, National Institute of Standards and Technology (NIST). https://doi.org/10.6028/NIST.IR.8485, https://doi.org/10.6028/NIST.IR.8485