Beyond Accuracy: Fairness, Scalability, and Uncertainty Considerations in Facial Emotion Recognition

Fromberg, Laurits; Nielsen, Troels; Frumosu, Flavia Dalia; Clemmensen, Line Katrine Harder

Published in:
Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL)

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Beyond Accuracy: Fairness, Scalability, and Uncertainty Considerations in Facial Emotion Recognition

Laurits Fromberg[1], Troels Nielsen[2], Flavia Dalia Frumosu[1], and Line Katrine Harder Clemmensen*[1,2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Tetatet.AI, Copenhagen, Denmark

Abstract

Facial emotion recognition (FER) from images or videos is an emerging subfield of emotion recognition that in recent years has achieved increased traction resulting in a wide range of models, datasets, and applications. Benchmarking computer vision methods often provide accuracy rates above 90% in controlled settings. However, little focus has been given to aspects of fairness, uncertainty, and scalability within facial emotion recognition systems. The increasing applicability of FER models within assisted psychiatry and similar domains underlines the importance of fair and computational resource compliant decision-making. The primary objective of this paper is to propose methods for assessment of existing open source FER models to establish a thorough understanding of their current fairness, scalability, and robustness.

1 Introduction

Facial emotion recognition (FER) is the task of classifying the emotional state of an individual based on the facial expressions of said individual. FER models have gone from hand-crafted features to various deep convolutional neural network architectures which achieve state-of-the-art performance on benchmark datasets [1, 2]. Transfer learning is also widely used with numerous applications of pre-trained models (e.g., GoogleNet [3], ResNet [4], etc.) trained on large datasets such as VGGFace2 [5]. In lab conditions, FER models generally provide high classification accuracies (above 90%), whereas accuracies closer to 50% are observed for in-the-wild datasets [2].

The transition from controlled to in-the-wild datasets such as [6, 7] using images collected from the internet has increased in popularity for training of large neural network models [1, 8]. Challenges and problems associated with occlusion have also been addressed with occlusion frequently arising in the setting of in-the-wild facial images [9].

FER has successfully been applied in AI assisted psychiatry for diagnosing schizophrenia [10] and depression [11], or in automatic behavioral coding to support research and quality assurance [12, 13]. Given the nature of the applications, it is vital to prioritize fair decision-making in the development phase of FER models to avoid any form of discriminatory behavior. Considering the prevalence of video-streams in many areas of applications, real-time operation is a fundamental requirement [14], while at the same time balancing the need for scalability with fairness and robustness. However, these aspects of facial emotion recognition are underrepresented in the literature.

A range of new databases focus on individuals of diverse ethnic backgrounds [2, 15, 16]. Additionally, other studies have concentrated on illumination conditions [17], and different viewpoints [17, 18]. Recently, the use of generative adversarial networks (GANs) have been proposed to address challenges originating from the vast diversity in facial images such as level of occlusion, differences in head-poses, or light conditions by generating synthetic data resembling the original images e.g. without glasses (in the case of occlusion) or with a different head-pose [19, 20].

To the best of our knowledge, only a few studies directly address issues of fairness in FER. One such study [21], evaluates gender-related fairness and its influence on facial expressions. They find that the Inception FER model they trained with gender balanced data is more fair, but also that when training with male faces and testing on female faces, results are more fair than when the training bias is reversed. In another work, gender bias has been analyzed using different deep learning architectures for face emotion recognition on the SASE-FE dataset [22]. The study finds similar results in terms of worse performance when training on female faces, and they find the largest concern to be classification of surprise, where the models generally perform worse on female faces. The authors encourage more explorations of a large range of datasets. Wang et al. [23] proposed an effective visual recognition benchmark for studying bias mitigation. A study by Xu et al. [24] focuses on sensitive attributes of gender, age and/or race. Here, three approaches, namely a baseline, an attribute-aware, and a disentangled approach have been performed with and without

*Corresponding Author: lkhc@dtu.dk

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

augmentation. Among these, it is found that the disentangled approach is the most effective at reducing demographic bias tested on the RAF-DB and CelebA-DB datasets.

By emphasizing the significance of further exploration, our work aims to delve into the aspect of fairness and the influence of scalability and uncertainty, for pre-trained models. Stoychev and Gunes [25] show that compression and quantization techniques effectively reduce model size while maintaining high overall accuracy in experimental results carried out over selected datasets. To this end, we propose a metric for assessing the balance between accuracy and model size.

Based on recently published reviews of FER, which give a comprehensive overview of the state of the art models and existing benchmark datasets [2, 20, 26], we choose to focus on a range of models of convolutional neural networks of different complexity and datasets with exhaustive labelling (subpopulations, head-pose, occlusion type, etc.) as well as class imbalances within datasets.

The contributions of our study are the following:

Scalability: We propose a novel measure, *perceived effectiveness*, for comparing the trade-off between the accuracy and the number of parameters of a model. We assess the perceived effectiveness of five FER models of different sizes. In addition, we test the possibility for achieving scalability through downsampling of the images.

Fairness: We assess the fairness of the model with the highest perceived effectiveness with respect to sex, age, and ethnicity.

Uncertainty: We propose an algorithm to add an additional class for observations with less confidence to make alternative decisions in cases of uncertainty. We evaluate the change in performance for our proposed algorithm.

2 Methods

2.1 FER models

We consider models of varying size, complexity, and architecture, to compare simple models, approaches designed to be operable on mobile devices, and more complex models and explore the limits of what is achievable in regards to performance. We chose the following pre-trained FER models, all available in `Python`: `HSE MobileNet (Mobile)` [5], `HSE EfficientNet-B0 (B0)` [5], `HSE EfficientNet-B2 (B2)` [5], `EMO-AffectNetModel (EMO)` [8], `Paz mini-Xception (mini-X)` [27], `Ad-Corre (Ad)` [1], and `DeepFace (DF)` [28]. We are interested in finding a model small enough to run in real-time, while still producing satisfactory results. The model complexities in terms of parameters can be seen in Figure 1.

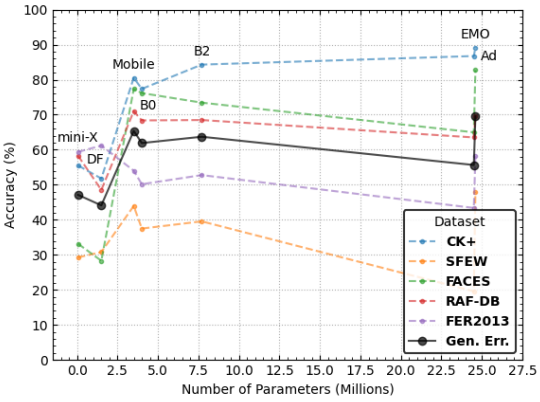

Figure 1. Model-size vs. Accuracy. Presented for the five datasets and for the average across the five sets.

2.2 Fairness

To assess the FER models for fairness across gender, age, and ethnicity, a suitable and simple method is to ensure demographic parity i.e., fairness across groups according to a specified statistic, often the prediction accuracy. We use demographic parity as we are aware of specific historical biases (gender, ethnicity) that may affect the model. To formalize this condition: A binary classifier h is said to satisfy demographic parity under a distribution over (X, A), where X is a feature vector and A is a sensitive feature, if the prediction $h(X)$ is statistically independent of the sensitive feature A [29]. This means that the difference in predicting a positive outcome should be zero (or as small as possible) across the sensitive feature for the model to be considered fair. We express this mathematically as:

$$\mathbb{E}[h(X)|A = a] = \mathbb{E}[h(X)], \quad \forall a, \tag{1}$$

where h expresses the prediction accuracy of a given emotion. In practice, we compute this using the `Python` library *Fairlearn* [29].

2.3 Scalability

We want FER models that scale computationally and preferably run real-time, while still having satisfactory performance. To assess this trade-off, we propose a novel measure for comparing the accuracy of neural networks with reference to the number of parameters. We call this measure the *perceived efficiency (PE)*, and define it as:

$$PE_\gamma = A - 2 \cdot \gamma \cdot P^{3/4}, \quad \gamma \in (0, 1], \tag{2}$$

where A is the accuracy in percentage, P is the number of parameters in millions, and γ is a regularisation parameter where a value close to 1 puts more emphasis on a small model, while a value close to 0 puts more weight on the accuracy of the model. For a fixed value of γ and a fixed accuracy, the perceived efficiency satisfies $PE_\gamma \to -\infty$ for $P \to \infty$ and $PE_\gamma \to A$ for $P \to 0$. Hence, an *optimal* model has a perceived efficiency of 100%. Figure 2 illustrates the perceived efficiency as a function of the number of parameters for different values of γ.

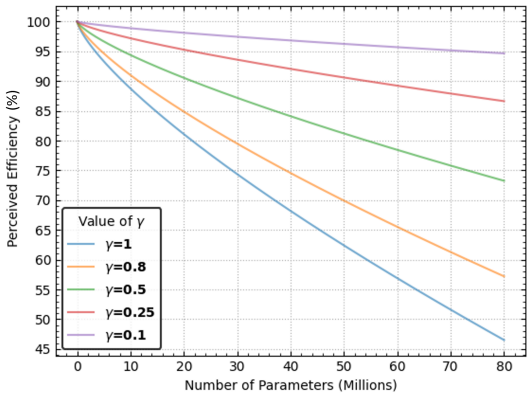

Figure 2. Perceived Efficiency. The perceived efficiency for a fixed accuracy of 100%, while varying the number of parameters for different values of the regularisation parameter γ.

2.4 Uncertainty

All the considered models return a probability for each expression class, which provides a natural measure of uncertainty of the predictions. We will exploit this information to create an additional class: "uncertain". When a prediction probability is lower than a pre-specified threshold C, the observation is considered uncertain. In practice, it is possible to make alternative decisions for observations predicted to the "uncertain" class. The pseudocode of this procedure is presented in Algorithm 1.

2.5 Data

An overview of the datasets used in this paper is presented in Table 1. The datasets were chosen to cover a wide variety of images and settings. For reference, the FER2013 dataset only yields a human accuracy of $68 \pm 5\%$ [32].

Algorithm 1 Uncertainty cut-off algorithm

Initialize threshold value, C
Create class "uncertain", $u = \{\}$
Number of samples, n
Classes, $c_i = \{\}$
while $n \neq 0$ **do**
  $n \leftarrow n - 1$
  Compute class probabilities : $\boldsymbol{p}_n \leftarrow f(\boldsymbol{x}_n)$
  Find class: $i \leftarrow \arg\max_i p_{i,n}$
  if $p_{i,n} < C$ **then**
    $u \leftarrow u \cup \{(p_{i,n}, i, n)\}$
  else
    $c_i \leftarrow c_i \cup \{(p_{i,n}, i, n)\}$
  end if
end while
Make predictions based on u and c_i =0

3 Results

3.1 Baseline and Scalability

We evaluated the performance of the pre-trained FER models on the five benchmark dataset, see Table 2. The accuracies are based on all available data, i.e. training and test sets were combined in cases where data were pre-defined in several splits. We also assessed the perceived efficiency (PE) of the models and here HSE MobileNet was best.

The average drop in accuracy from the best model (EMO-AffectNetModel) to HSE MobileNet is around 5% and it is the second best performing model in terms of accuracy. Therefore, we propose the use of HSE MobileNet and continue illustrating the assessment of fairness and robustness with this model only. Figure 1 illustrates the model-size vs. accuracy for the five datasets. Here it is also clear how HSE MobileNet outperforms the models of similar size, while EfficientNet-B2 only yields a slight increase in accuracy despite being almost double the size. Likewise, EMO-AffectNetModel performs better than Ad-Corre, albeit being similar in size.

Another approach to obtain scalability is image resizing. Table 3 summarizes accuracies from rescaling the images between 16×16 to 128×128 pixels. As expected, accuracy generally decreases as the image size is decreased. However, the drop is minimal until sizes of 48×48 or 32×32 pixels.

3.2 Fairness

We examined the fairness of HSE MobileNet across the protected attributes sex, age, and ethnicity in the RAF-DB dataset; see Figure 3.

HSE MobileNet is generally fair although with smaller differences across the attributes. Of concern is the emotion "surprise", where the model yields a discrepancy for the ethnic subpopulations of asian

Table 1. Datasets. Description of FER datasets used in this paper. P = Posed emotions (controlled setting), I = In-the-wild, b = basic categorical emotion labels (happy, sad, angry, surprise, fear, disgust), n = neutral emotion labels. The FACES dataset is missing the surprise emotion label.

Dataset	Type	#Subject	#Samples	Expr.	Attributes	Notes
CK+ [30]	P	123	327	6 b + n	Age: 18-50	Videos
RAF-DB [31]	I	∼ 29672	29672	6 b + n	Age: 0-70 52% Females (5% unsure)	Static
FER-2013 [32]	P + I	∼ 35887	35887	6 b + n	N/A	Static
SFEW[33]	I	95	700	6 b + n	Age: 1-70	Static
FACES [34]	P	171	2052	5 b + n	Young: 19-31 Middle-aged: 39-55 Older: 69-80 50% Females Caucasian	Static

If there is no specification regarding ethnicity, then either no information has been provided or there are multiple ethnicities.

Table 2. Pre-trained Models. Performance of pre-trained FER models using benchmark datasets.

Dataset	HSE MobileNet	HSE EfficientNet-B0	HSE EfficientNet-B2	EMO-AffectNetModel	Paz mini-Xception	Ad-Corre	DeepFace
CK+	80.53%	77.37%	84.30%	89.19%	55.25%	86.75%	51.78%
SFEW	43.95%	38.50%	39.56%	47.96%	29.27%	19.59%	30.79%
FACES	77.33%	76.17%	73.44%	82.78%	33.17%	64.98%	28.31%
RAF-DB	70.98%	68.37%	68.48%	69.57%	58.31%	63.51%	48.52%
FER2013	54.02%	50.16%	52.76%	58.20%	59.37%	43.39%	61.54%
Gen. Err.	65.36%	62.11%	63.71%	**69.54%**	47.07%	55.64%	44.19%
$\overline{\text{PE}}_{\{\gamma=0.7\}}$	**61.78**	57.95	57.24	54.08	46.96	40.23	42.22

vs. black/white. Fear and disgust also display some differences across age and ethnicity.

3.3 Uncertainty

As expected, there is a positive linear trend between the cut-off value and the accuracy of the emotion predictions, see Figure 4. Employing a threshold approach yields a significant increase in accuracy, when compared to the baseline. However, the proportion of samples in the "uncertain" class quickly tends towards being the most dominating class, which for many applications may be undesirable. It is often more appropriate to implement a cut-off value of 40-60% to allow for an increase in performance while controlling the proportion of samples in the "uncertain" class. We recommend fine-tuning the cut-off value to achieve the desired level of accuracy to the proportion of samples in the "uncertain" class.

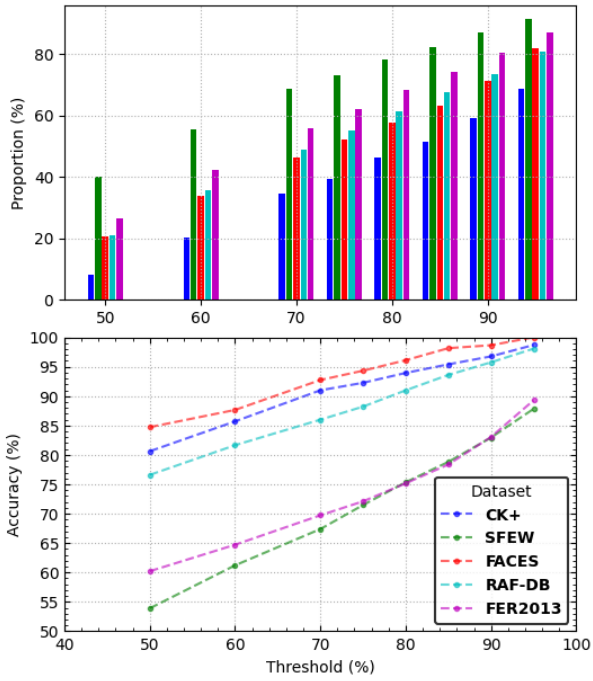

Figure 4. Proportion/Accuracy vs. Threshold. The proportion of samples in the "uncertain" class and the accuracy vs. the threshold for HSE MobileNet.

Table 3. Resolution. The accuracy of the datasets at different resolutions using HSE MobileNet. The column [**Acc.**] refers to the baseline i.e. the accuracy at the original resolution. All resolutions are symmetric so, a resolution of 16 implies 16×16 pixels and so forth. The resizing is done using bilinear interpolation with the `Python` library *OpenCV*.

Dataset	16	32	48	64	88	128	Acc.
CK+	38.94%	76.76%	80.53%	79.61%	79.92%	80.53%	80.53%
SFEW	24.81%	40.39%	43.95%	44.48%	44.86%	44.10%	43.95%
FACES	16.28%	41.61%	62.03%	72.47%	75.11%	75.86%	77.33%
RAF-DB	42.28%	62.22%	69.07%	70.13%	70.64%	70.88%	70.98%

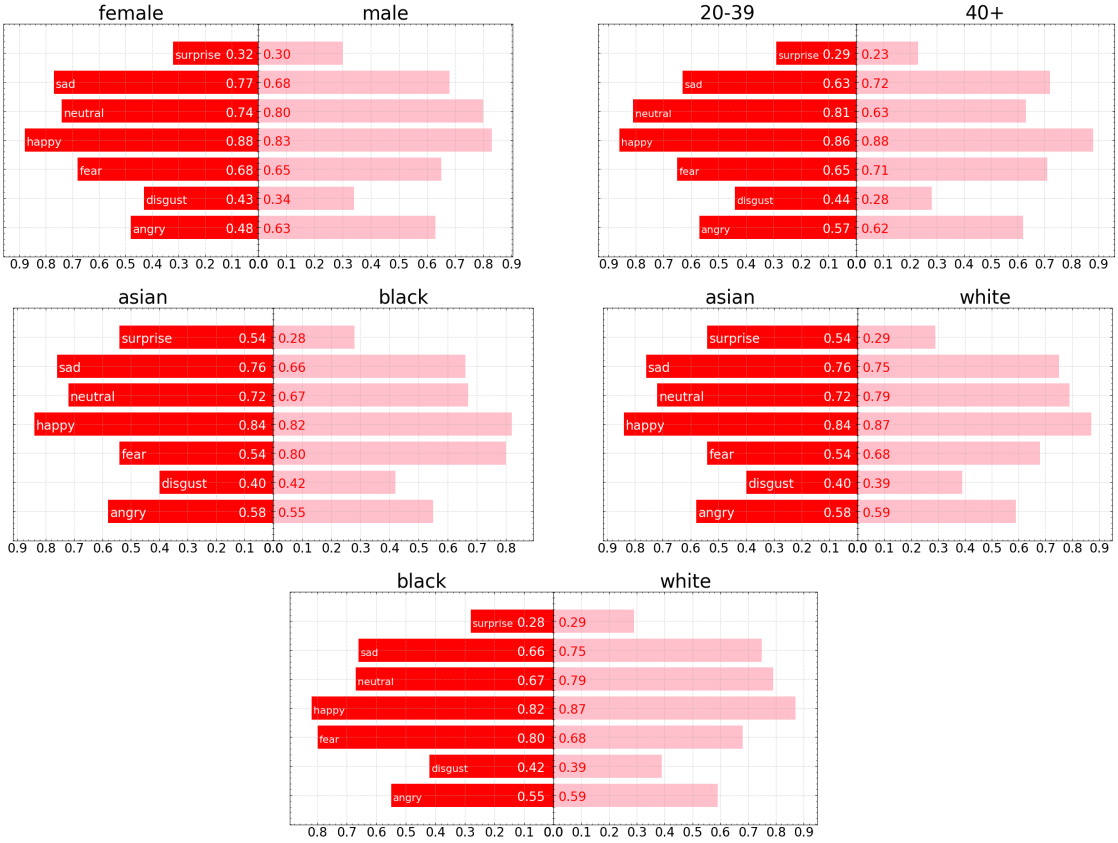

Figure 3. Fairness. The demographic parity for different subpopulations of the RAF-DB dataset using HSE MobileNet. Note that the last two age groups have been combined into a single group due to the last age group containing very few samples, hence skewing the perception of the model's fairness.

Figure 5 shows an example of predicted class probabilities and illustrates how all predictions below a cut-off value would determine the uncertainty class.

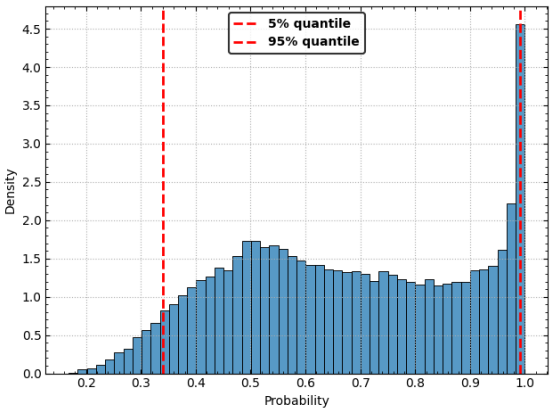

Figure 5. Distribution. Histogram of probabilities for the FER2013 dataset using HSE MobileNet.

4 Discussion

Our study is limited by existing limitations in the available data sources, which we have based our investigations on. In future work, addressing these weaknesses are of importance to reach fair open source FER models.

We have chosen demographic parity to measure fairness, but other metrics like equalized odds or equal opportunity exist, and are relevant. The metric for measuring fairness should be chosen carefully with the objective of the model in mind. As we are examining open source models, which could potentially be used for various purposes, all such metric may be relevant. Therefore, we recommend future studies examine different fairness metrics.

We acknowledge that our proposed perceived effectiveness measure restricts itself to looking at the number of parameters in the models, which is a simplified view. Considering computation time and interpretability would also be relevant.

Future studies should, apart from the development of unbiased data sets, also consider how we address the biases in terms of model developments.

5 Conclusion

Complimentary to existing literature, we have assessed the fairness for a case-specific pre-trained model in relation to multiple benchmarks, where we generally found no results of unfair behaviour except for a few emotions, like "surprise" for asian vs black or white. We made a comprehensive investigation of the scalability of pre-trained models and proposed a measure for comparing neural networks while balancing between accuracy and scalability. We explored the reliance on high-resolution images, discovering that even images with a resolution as low as 32 to 48 pixels produced little to no reduction in the performance. Finally, we demonstrated the importance of considering prediction uncertainty and proposed how to do this through a simple cut-off value on the predicted class probabilities. For general-purpose applications, we suggest adopting a cut-off value in the vicinity of 50%.

Acknowledgments

We would like to thank the anonymous reviewers for their relevant input, in particular towards the discussion of the limitations of the presented work.

Laurits Fromberg, Flavia Frumosu, and Line Clemmensen were supported by Novo Nordisk Foundation (grant number: NNF19OC0056795, project Wrist Angel) and Discovery Grant (grant number: DTU112009, project AI Powered Emotional Awareness). Line Clemmensen together with Troels Nielsen are co-founders of Tetatet AI.

References

[1] A. P. Fard and M. H. Mahoor. "Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild". In: *IEEE Access* 10 (2022), pp. 26756–26768. DOI: 10.1109/ACCESS.2022.3156598.

[2] S. Li, L. Guo, and J. Liu. "Towards East Asian Facial Expression Recognition in the Real World: A New Database and Deep Recognition Baseline". In: *Sensors* 22.21 (2022). ISSN: 1424-8220. DOI: 10.3390/s22218089. URL: https://www.mdpi.com/1424-8220/22/21/8089.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[5] A. V. Savchenko. *Frame-level Prediction of Facial Expressions, Valence, Arousal and Action Units for Mobile Devices*. 2022. arXiv: 2203.13436 [cs.CV].

[6] D. Kollias and S. Zafeiriou. "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace". In: *arXiv preprint arXiv:1910.04855* (2019).

[7] A. Mollahosseini, B. Hasani, and M. H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10.1 (Jan. 2019), pp. 18–31. DOI: 10.1109/taffc.2017.2740923. URL: https://doi.org/10.1109%2Ftaffc.2017.2740923.

[8] E. Ryumina, D. Dresvyanskiy, and A. Karpov. "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study". In: *Neurocomputing* 514 (2022), pp. 435–450. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2022.10.013. URL: https://www.sciencedirect.com/science/article/pii/S0925231222012656.

[9] Y. Li, J. Zeng, S. Shan, and X. Chen. "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism". In: *IEEE Transactions on Image Processing* 28.5 (May 2019), p. 2439. DOI: 10.1109/TIP.2018.2886767.

[10] X. Zhang, T. Li, C. Wang, T. Tian, H. Pang, J. Pang, C. Su, X. Shi, J. Li, L. Ren, J. Wang, L. Li, Y. Ma, S. Li, and L. Wang. "Recognizing schizophrenia using facial expressions based on convolutional neural network". In: *Brain and Behavior* 13.5 (2023), e3002. DOI: https://doi.org/10.1002/brb3.3002. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.3002. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/brb3.3002.

[11] Y. S. Lee and W. H. Park. "Diagnosis of Depressive Disorder Model on Facial Expression Based on Fast R-CNN". In: *Diagnostics (Basel, Switzerland)* 12.2 (2022), p. 317. DOI: 10.3390/diagnostics12020317. URL: https://doi.org/10.3390/diagnostics12020317.

[12] F. Frumosu, N. Lønfeldt, A. Mora-Jensen, S. Das, N. Lund, A. Pagsberg, and L. Clemmensen. "Interpretability by design using computer vision for behavioral sensing in child and adolescent psychiatry". English. In: *Proceedings of Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning*. 38th International Conference on Machine Learning, ICML 2021 ; Conference date: 18-07-2021 Through 24-07-2021. 2022. URL: https://icml.cc/Conferences/2021.

[13] N. N. Lønfeldt, S. Das, F. D. Frumosu, A.-R. C. Mora-Jensen, A. K. Pagsberg, and L. Clemmensen. "Scaling-up Behavioral Observation with Computational Behavior Recognition". In: *PsyArxiv* (2023), pp. 1–14. DOI: 10.3758/BRM.42.1.351.

[14] A. V. Savchenko. "HSEmotion: High-speed emotion recognition library". In: *Software Impacts* 14 (2022), p. 100433. ISSN: 2665-9638. DOI: https://doi.org/10.1016/j.simpa.2022.100433. URL: https://www.sciencedirect.com/science/article/pii/S2665963822001178.

[15] T. Yang, Z. Yang, G. Xu, and et al. "Tsinghua facial expression database - A database of facial expressions in Chinese young and older women and men: Development and validation". In: *PLoS One* 15.4 (Apr. 2020). Published 2020 Apr 15, e0231304. DOI: 10.1371/journal.pone.0231304.

[16] J. Muhammad, Y. Wang, C. Wang, K. Zhang, and Z. Sun. *CASIA-Face-Africa: A Large-scale African Face Image Database*. 2021. arXiv: 2105.03632 [cs.CV].

[17] N. Aifanti, C. Papachristou, and A. Delopoulos. "The MUG facial expression database". In: *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. 2010, pp. 1–4.

[18] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg. "Presentation and validation of the Radboud Faces Database". In: *Cognition and Emotion* 24.8 (2010), pp. 1377–1388. DOI: 10.1080/02699930903485076. eprint: https://doi.org/10.1080/02699930903485076. URL: https://doi.org/10.1080/02699930903485076.

[19] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee. "Complete Face Recovery GAN: Unsupervised Joint Face Rotation and De-Occlusion from a Single-View Image". In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 1173–1183. DOI: 10.1109/WACV51458.2022.00124.

[20] M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. Alaya Cheikh, M. Hijji, K. Muhammad, and J. J. Rodrigues. "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines". In: *Alexandria Engineering Journal* 68 (2023), pp. 817–840. ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2023.01.017. URL: https://www.sciencedirect.com/science/article/pii/S1110016823000327.

[21] C. Manresa-Yee, S. Ramis Guarinos, and J. M. Buades Rubio. "Facial Expression Recognition: Impact of Gender on Fairness and Expressions". In: *Proceedings of the XXII International Conference on Human Computer Interaction*. Interacción '22. Teruel, Spain: Association for Computing Machinery, 2022. ISBN: 9781450397025. DOI: 10.1145/3549865.3549904. URL: https://doi.org/10.1145/3549865.3549904.

[22] A. Domnich and G. Anbarjafari. "Responsible AI: Gender bias assessment in emotion recognition". In: *arXiv preprint arXiv:2103.11436* (2021).

[23] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. "Towards fairness in visual recognition: Effective strategies for bias mitigation". In: *Proceedings of the IEEE/CVF conference on computer*

vision and pattern recognition. 2020, pp. 8919–8928.

[24] T. Xu, J. White, S. Kalkan, and H. Gunes. "Investigating bias and fairness in facial expression recognition". In: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer. 2020, pp. 506–523.

[25] S. Stoychev and H. Gunes. "The effect of model compression on fairness in facial expression recognition". In: *arXiv preprint arXiv:2201.01709* (2022).

[26] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. "A survey on facial emotion recognition techniques: A state-of-the-art literature review". In: *Information Sciences* 582 (2022), pp. 593–617. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2021.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0020025521010136.

[27] O. Arriaga, M. Valdenegro-Toro, and P. Plöger. *Real-time Convolutional Neural Networks for Emotion and Gender Classification*. 2017. arXiv: 1710.07557 [cs.CV].

[28] S. I. Serengil and A. Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework". In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE. 2021, pp. 1–4. DOI: 10.1109/ICEET53442.2021.9659697. URL: https://doi.org/10.1109/ICEET53442.2021.9659697.

[29] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Tech. rep. MSR-TR-2020-32. Microsoft, May 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

[31] S. Li, W. Deng, and J. Du. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2584–2593. DOI: 10.1109/CVPR.2017.277.

[32] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. "Challenges in representation learning: A report on three machine learning contests". In: *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer. 2013, pp. 117–124.

[33] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark". In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 2011, pp. 2106–2112. DOI: 10.1109/ICCVW.2011.6130508.

[34] N. Ebner, M. Riediger, and U. Lindenberger. "FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation". In: *Behavior research methods* 42 (Feb. 2010), pp. 351–62. DOI: 10.3758/BRM.42.1.351.