I. J. Engineering and Manufacturing, **2025, 1, 1-10**
Published Online on February 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2025.01.01

Racial Bias in Facial Expression Recognition Datasets: Evaluating the Impact on Model Performance

Ridwan O. Bello*
Department of Computer Science, The University of Alabama, Tuscaloosa, Tuscaloosa, AL 35487, USA
E-mail: robello2@crimson.ua.edu
ORCID iD: https://orcid.org/0000-0002-7994-6522
*Corresponding Author

Joseph D. Akinyemi
Department of Computer Science, University of York, YO10 5DD, Heslington, United Kingdom
E-mail: joseph.akinyemi@york.ac.uk
ORCID iD: https://orcid.org/0000-0003-3121-4231

Khadijat T. Ladoja
Department of Computer Science, University of Ibadan, Ibadan, Nigeria
E-mail: kt.bamigbade@ui.edu.ng
ORCID iD: https://orcid.org/0000-0001-9479-5825

Oladeji P. Akomolafe
Massload Technologies, SK Canada
E-mail: akomspatrick@gmail.com
ORCID iD: https://orcid.org/0000-0002-0892-2603

Received: 27 July, 2024; Revised: 21 August, 2024; Accepted: 25 September, 2024; Published: 08 February, 2025

Abstract: Despite extensive research efforts in Facial Expression Recognition (FER), achieving consistent performance across diverse datasets remains challenging. This challenge stems from variations in imaging conditions such as head pose, illumination, and background, as well as demographic factors like age, gender, and ethnicity. This paper introduces NIFER, a novel facial expression database designed to address this issue by enhancing racial diversity in existing datasets. NIFER comprises 3,481 images primarily featuring individuals with dark skin tones, collected in real-world settings. These images underwent preprocessing through face detection and histogram equalization before being categorized into five basic facial expressions using a deep learning model. Experiments conducted on both NIFER and FER-2013 datasets revealed a decrease in performance in multiracial FER compared to single-race FER, underscoring the importance of incorporating diverse racial representations in FER datasets to ensure accurate recognition across various ethnicities.

Index Terms: Bias, Ethnicity, Deep learning, Facial expression recognition, Computer vision.

1. Introduction

Expressions and emotions play a significant role in human-to-human interactions and non-verbal communication as they make people express themselves without using words. Facial expressions are important in facilitating human communications and interactions. Human emotion recognition from images or video streams is a widely researched area in Computer Vision, with implications for facial expression recognition (FER) [1]. FER is an essential visual recognition technology used to detect emotions from facial images, with applications in affective computing, human-computer interaction, education, healthcare, driver monitoring for autonomous driving, and psychological treatments [2]. Facial expression recognition deals with visually analyzing and recognizing different facial motions. It is a combination

of face detection and recognition techniques as well as expression recognition in analyzing static images [3] or dynamic image sequences [4].

Significant advancements have been made in the development of facial expression recognition systems, particularly with the adoption of Deep Neural Networks [5]. However, these systems still struggle with generalization across diverse datasets, rendering them unreliable for real-world applications. Human facial expressions encompass a wide range of variations that are often underrepresented in the training datasets for these models. Moreover, a major limitation of deep learning models is their need for extensive, varied, and large datasets to achieve optimal performance. Although some methods attempt to address this challenge, many do not undergo thorough evaluations across varied and well-diversified databases. The major contributions of this work are as follows:

- A new dataset named Nigerian Facial Expression Recognition (NIFER), predominantly featuring black/ Nigerian faces, was created to enhance racial diversity and generalization in facial expression recognition.
- The Contrast Limited Adaptive Histogram Equalization (CLAHE) method was applied to preprocess face images before training and validation, with the aim of enhancing image quality and features.
- Extensive experiments were conducted to investigate the generalization of the VGGFace2 architecture in recognizing facial expressions across multiple ethnicities.
- The findings indicate the necessity for diversity in FER datasets to achieve robust multi-ethnic facial expression recognition.

The paper proceeds with a literature review in Section 2, followed by the research methodology in Section 3, which details the new dataset, existing datasets, and the experimental setup. Section 4 presents and discusses the experimental results, while the paper concludes in Section 5.

2. Related Works

2.1 Previous Approaches in Facial Expression Recognition

Numerous methods in the literature aim to address challenges in facial expression recognition. However, many of these approaches rely on limited databases with similar attributes, such as race, during both training and evaluation phases, making it difficult to demonstrate their effectiveness across diverse conditions. This section reviews related works that tackle the generalization problem. Li, Jin, Akram, Han, and Chen [6] addressed data limitations and efficient feature extraction in facial recognition by introducing innovative face cropping and rotation techniques along with a simplified CNN architecture. Their approach aimed to improve dataset quality and feature extraction accuracy. Evaluation on the CK+ [7]and JAFFE [8] databases yielded high recognition accuracies of 97.38% and 97.18% for 7-class experiments, respectively. The study also analyzed the impact of each method and the CNN simplification. Their approach demonstrated competitiveness in terms of training time, testing time, and recognition accuracy compared to existing methods. Schoneveld and Othmani [9] developed a generalized deep feature extractor (DeepFEVER) to tackle the generalizability problem in most FER systems. Using self-distillation, they trained two labeled sets and one unlabeled set, employing teacher and student model phases. The teacher model was trained using the FaceNet pre-trained network on both Affect-Net [10] and Google Facial Expression Comparison (FEC) databases [11]. The resulting model was then used to predict the output of the unlabeled dataset in the student model. When compared to other models, the DeepFEVER model achieved significantly higher validation accuracies, achieving 65.4% on AffectNet, 86.5% on FEC, and 87.4% on the Real-World Affective Faces database.

Farzaneh and Qi in [12] proposed a flexible method called Deep Attentive Center Loss (DACL) for Facial Expression Recognition (FER) under improbable circumstances. DACL combines a sparse re-formulation of center loss with Deep Metric Learning to adaptively control deep feature contributions. They trained ResNet-18 on RAF-DB for 60 epochs and AffectNet for 20 epochs, using a learning rate decay strategy. DACL incorporates an attention mechanism parameterized by a customizable neural network to estimate contribution probabilities. Experimental results demonstrated DACL's superiority over baseline methods and state-of-the-art approaches on RAF-DB and AffectNet. Minaee Minaei, and Abdolrashidi [13] proposed an attentional convolutional network for facial expression recognition (FER), integrating a spatial transformer module to focus on critical facial regions. Their visualization technique highlights the key image areas influencing classifier outcomes. This approach, which uses a convolutional network with fewer than 10 layers and incorporates attention, consistently high accuracies of 70% on FER-2013, 99.3% on FERG, 92.8% on JAFFE, and 98.0% on CK+.

Stoychev and Gunes [14] explored fairness and accuracy in model compression on facial expression recognition systems. The compressed model was developed to make FER systems easier to train and deployable on resource-constrained devices like mobile phones and robots. Their results showed that model compression has a significant effect on increasing bias on sensitive attributes like age, gender, and race. This is most evident in the CK+ database which consists of highly dispersed data distribution.

Pourramezan, Mahoor and Fard in [3] proposed an Adaptive Correlation based Loss (Ad-Corre) model to address intra-class variations and inter-class similarities in facial expression tasks. The Ad-Corre model comprises three

components: Feature Discriminator, Mean Discriminator, and Embedding Discriminator. These components guide the network to generate highly correlated features for the same category and lower correlations for different categories. Using ResNet-50 as the backbone model and the Ad-Corre loss metric, an accuracy of 71.48%, 85.93% and 67.78% were recorded for FER-13 [15], RAF-DB [16] and AffectNet [10] datasets respectively. The Facial Expression database FER-2013 [15] was introduced during the ICML 2013 Workshop on Challenges in Representation Learning, utilized 184 emotion-related keywords combined with various demographic descriptors to diversify search results. This dataset comprises 35,887 images representing seven emotions sourced via the Google Image Search API. Following approval, images were converted to 48 × 48 grayscale format. The dataset includes 4953 images for “Anger”, 547 for “Disgust”, 5121 for ”Fear”, 8989 for ”Happiness”, 6077 for ”Sadness”, 4002 for ”Surprise”, and 6198 for ”Neutral” expressions.

Researchers have made significant progress in improving facial expression recognition (FER) using advanced algorithms such as Attentive Center Loss [13] and generative adversarial models [16]. Moreover, efforts in data augmentation, cross-database analysis, and data preprocessing have contributed to performance enhancements. However, existing datasets often lack adequate representation of African ethnicity, presenting challenges in image classification and recognition. This study aims to enhance FER by creating a new dataset primarily comprising African faces and comparing pre-trained models across this new dataset and existing datasets for a more robust FER model.

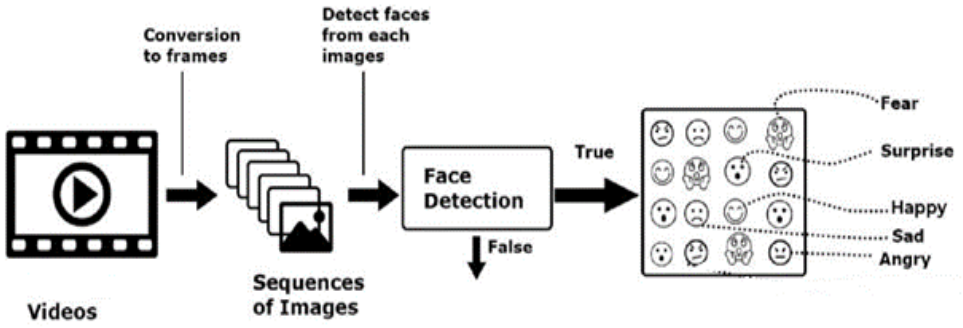


Fig. 1. Dataset collection process from videos.

3. Data and Models

3.1 Experimental Setup

The newly created Nigerian Facial Expression Recognition (NIFER) dataset consists of 3,481 images before preprocessing and 2,846 images after preprocessing. The dataset images were collected in two categories: posed and non-posed facial expressions. The posed refers to images obtained when the subjects intentionally posed a facial expression while the non-posed refers to images obtained from online video streams where the facial expressions were mostly real. Generally, the images comprise a wide range of illumination, poses, ethnicities, resolution, and brightness as depicted in most benchmarked Computer Vision datasets. For posed facial expressions, there are 1400 facial images (before preprocessing) from 350 subjects. These images were taken using mobile phone cameras with participants ages between 15 and 40 years old. The participants were students at the University of Ibadan, Ibadan, Nigeria. Consent to collect data was gotten from the Department of Computer Science of the institution. Additionally, the selected students agreed and consented to participate in the study. The collected images were then labelled based on the subject’s facial expression into 5 basic emotion categories: Happy, Sad, Angry, Fear and Surprised.

The non-posed facial expressions are made up of 2,081 images (before preprocessing) extracted from downloaded video streams. The videos were downloaded from popular social media websites such as Instagram and TikTok and converted into frames. The search keywords used on these sites include “Top SAD Nigeria movie scenes”, “Top HAPPY Nigeria movie scenes”, “Top ANGRY Nigeria movie scenes”, “Top FEAR Nigeria movie scenes” and “Top SURPRISE” Nigeria movie scenes”. The Multi-task Cascaded Neural Network (MTCNN) [17] face detection algorithm was used to detect faces in each frame and cropped before downloading. After pre-processing and annotation, the resulting images amount to a total of 2,846 posed and non-posed images. The flowchart of the data collection process for the non-posed images is shown in Figure 1. A similar process is used for the posed images but without the need to convert to frames since still photographs are obtained as the input.

The resulting 2846 images were labelled according to their facial expressions with the 5 basic emotions: Happy, Sad, Angry, Fear and Surprised (see table 1). Using these labels as ground truth can be defective as facial expressions often vary from the stereotypical definition. Therefore, the labelling process was done in two steps. The first step involves the comparison between each facial image and the combination of action units defined by [18] for the five basic emotions. If the blend of these action units matches an image, it's labelled accordingly; otherwise, it proceeds to the second step for perceptual judgment. However, the second step isn't entirely independent, as expressions meeting emotional attributes are likely categorised as such. Table 1 shows a nearly balanced distribution of images across the 5 expression classes in the NIFER dataset.

Table 1. Number of images (post-preprocessing) in each expression category within the NIFER dataset.

Category	Number of Images
Happy	571
Sad	593
Angry	582
Fear	544
Surprise	556
Total	**2846**

3.2 Data Pre-processing

In the pre-processing stage, images were passed through MTCNN for face detection and landmark localization and images with undetected faces were removed from the dataset. This was followed by face extraction, adaptive histogram equalization and resizing of images. Contrast-Limited Adaptive Histogram Equalization (CLAHE) [19] was applied to the cropped face image for intensity normalization and contrast enhancement. Adaptive Histogram Equalization (AHE) works on small sections of an image often referred to as tiles, instead of the entire image in ordinary Histogram Equalization. CLAHE is a modified version of AHE which solves the problem of over-amplification of the contrast experienced in AHE. It distributes the part of the histogram which is greater than the clip limit evenly across all histograms. It does histogram equalization in small patches or small tiles with high accuracy and contrast limiting. Figures 2 and 3 show a comparison between a sample image and its histogram with and without CLAHE.

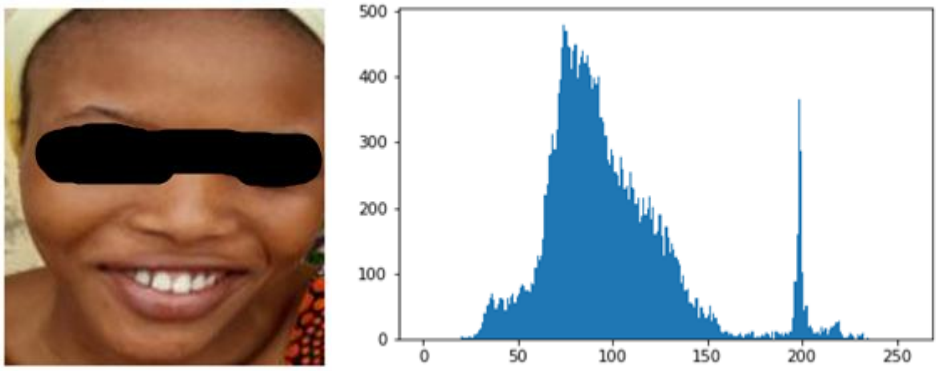


Fig. 2. Sample image and its histogram before CLAHE.

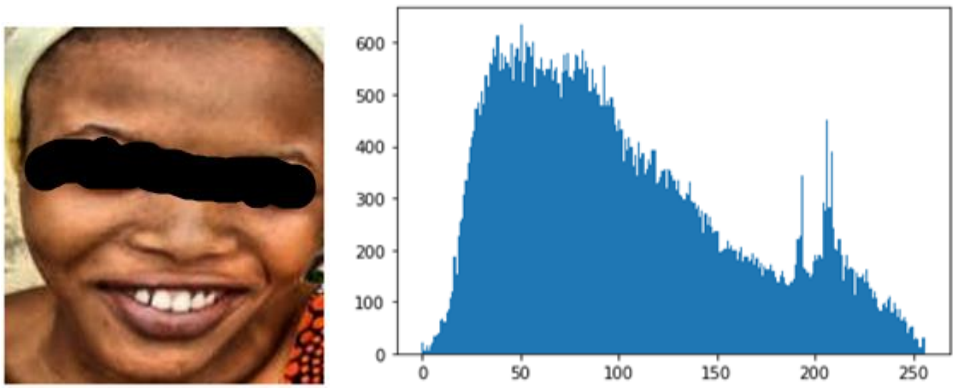


Fig 3. Sample image and its histogram after CLAHE

3.3. Model Architecture and Training

The VGGFace2 architecture [20] with ResNet pre-trained weights was selected due to its superior performance, especially since it was pre-trained on face images rather than generic objects. The loss function used was categorical cross-entropy, and the Adam optimizer was adopted with a learning rate of 0.0001. For each dataset (FER-2013 and NIFER), VGGFace2 was trained for 40 epochs with dropout and early stopping parameters set to prevent overfitting. The accuracy of the VGGFace2 model was evaluated by alternately using these two datasets as training and testing sets. The stated model and parameter values were applied consistently throughout the experiments.

3.4. Evaluation Metrics

In this work, the performances of the face recognition models are measured by common classification metrics such as accuracy, precision, recall and F1-score. The overall F1 score (also referred to as the macro-averaged accuracy) is the arithmetic mean of the F1 scores across all expression classes. F1 score is an important metric for assessing generalization performance across the two imbalanced datasets.

4. Results and Discussion

Experiments were performed on the two datasets, FER-2013 and NIFER by interchangeably using them as training and test sets and the results are presented in this section. Since the NIFER dataset only contains five classes, we excluded the "Disgust" and "Neutral" classes in the FER2013 dataset to match with our class labels for fair comparison. Two sets of experiments were performed. The first set of experiments was intended to investigate the individual impact of each dataset for facial expression recognition while the second was intended for a fair comparison between the two datasets for facial expression recognition.

In the first set of experiments, out of the 36,300 images in the FER-2013 dataset, we selected 5,476 images for training, and 4,394 images for testing while for NIFER, 2,273 images were used for training and 572 images for testing. In the second experiments, we selected the same number of images from each class in both datasets for training and testing. For training, we selected 435 images from each class making a total of 2,175 training images across 5 classes. For testing, we selected 110 images per class making a total of 550 test images across 5 classes.

The results presented in Table 4 underscore the critical importance of including diverse ethnic and racial groups in facial expression recognition datasets. Specifically, as shown in Table 2, training on the FER-2013 dataset yielded an accuracy of 56% on its own test set but only 34% on the NIFER test set, highlighting the model's limited ability to generalize across datasets representing different ethnic groups. Similarly, training on the NIFER dataset resulted in an accuracy of 82% on its test set, but only 32% on the FER-2013 test set, as demonstrated in Table 3. However, when we combined the two datasets, we observed a significant improvement in performance, with accuracies of 54% on the FER-2013 test set and 83% on the NIFER test set. This performance boost clearly indicates that including multiple ethnic groups in the training data enables the model to better capture the variability in facial expressions across diverse populations. The results suggest that ethnic inclusivity in benchmark datasets is essential for developing robust, generalizable facial expression recognition systems that perform well across different demographic groups. In Figure 4, ten samples of images showing five expression classes for both males and females in the NIFER dataset are shown. The images express variations of illumination, pose and background mimicking the real world.

4.1. Experiment 1

Figure 5 shows the training/validation accuracy and loss graphs over 40 epochs on the FER2013 dataset. The model starts training at around 27% accuracy and plateaus around 50% accuracy. The validation accuracy starts from 31% to 54%. The training loss starts from 8.3434 down to 1.23423 while the validation loss also decreases from 1.923 to 1.3422. Although the accuracy obtained on the FER-2013 is not very impressive, the training and validation plots show a steady progression towards convergence that could be achieved with more computational resources, more epochs and more parameter tuning. More so, the validation performance being consistently better than the training performance indicates a "good fit".

Figure 6 shows the training/validation accuracy and loss graphs over 40 epochs on the NIFER dataset. The model starts training at around 32% accuracy and plateaued at around 72% accuracy while the validation accuracy starts from 55% to 78%. The training loss starts from 13.3434 down to 1.23423 while the validation loss also decreases from 2.023 to 1.3422. As in the case of the FER-2013 dataset, a steady progression towards convergence can be observed in both the accuracy and loss curves. However, the model's performance on NIFER is better than on FER-2013 in both accuracy and loss. This is largely due to the very small resolution of the FER-2013 images (48 × 48) compared to the larger resolution of NIFER images (160 × 160) which means most of the visual information in the FER-2013 images is lost when resized to 224 × 224 for use in the VGGFace2 model. Again, the training/validation performance indicates a good fit and that the model was able to generalize well on the NIFER dataset.

(a) Angry-Female.

(b) Fear-Female

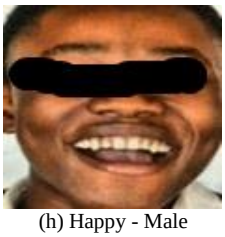
(c) Happy-Female

(d) Sad-Female

(e) Surprise-Female


(f) Angry- Male


(g) Fear - Male


(h) Happy - Male


(i) Sad-Male


(j) Surprise - Male

Fig. 4. Sample images in the NIFER dataset.

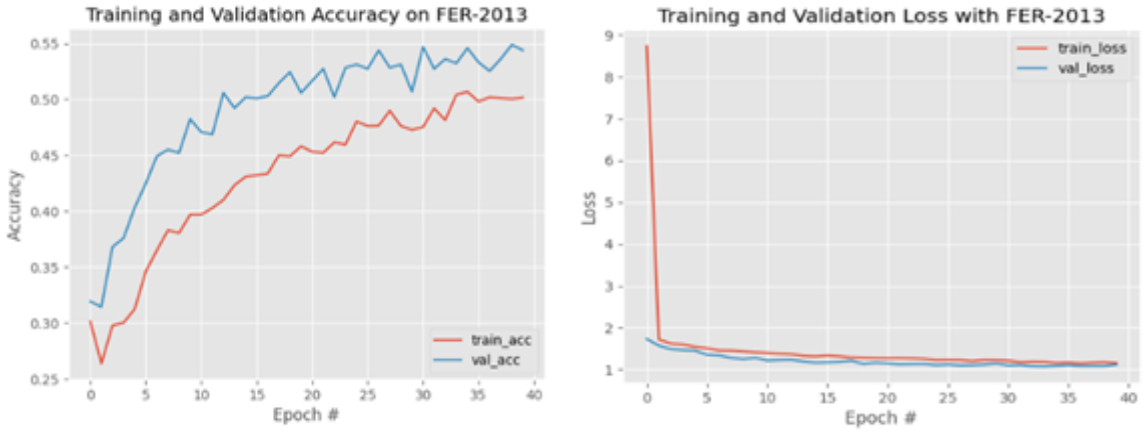


Fig. 5. Training/Validation Accuracy for FER-2013 after 40 epochs.

4.2. Experiment 2

To investigate the effect of multi-ethnicity in the datasets, the second phase of experiments was conducted using both datasets as training and test sets, interchangeably and in combination, with both datasets containing an equal number of images per class in both their training and test sets. As earlier stated, in all experiments here, we used the same model and parameters as in Experiment 1.

First, we trained the VGGFace2 model on 2175 images of the FER-2013 and tested it on the FER-2013 and NIFER test sets and the results obtained are shown in Table 2. From the results in Table 2, it can be observed that the trained model really struggles to achieve a good performance on any of the test sets. Even though it was trained on the images of the FER-2013 dataset, its overall F1 score on the test images of the same FER-2013 dataset is only 56% and, as expected, it is way lower on the NIFER test set where it achieves an overall F1-score of 34%. This gives two indications, the first is that the training set is not large and diverse enough and the second is that the images in the FER-2013 dataset are nothing like the ones in the NIFER dataset even though they exhibit the same facial expressions. This could signal the impact of ethnic variations even in the way facial expressions are displayed. To further investigate the above observation, we also trained the VGGFace2 model on NIFER and tested it on FER-2013 and NIFER test sets; the results are shown in Table 3.

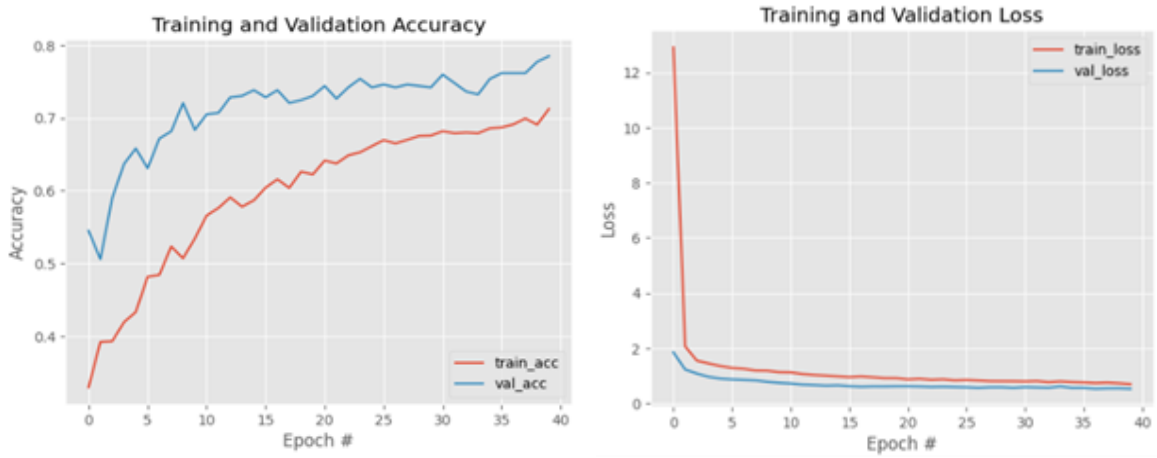


Fig. 6. Training/Validation Accuracy for NIFER after 40 epochs.

From the results in Table 3, it was observed that the model, when trained on the images of the NIFER dataset, correctly recognises expressions from the test images of the NIFER dataset (the second half of Table 3) with an F1-score of 82% which is significantly higher than the performance when trained and tested on images of the FER-2013 dataset (the first half of Table 2). This is most likely due to the better quality of the images in the NIFER dataset, which seems to make up for the ethnic diversity of the dataset. Much like the result in the second half of Table 2, the model does not perform quite well when trained on the images of NIFER and tested on the images of FER-2013 (the first half of Table 3).

Table 2. Results of Training with images of the FER-2013 dataset

	Class	Precision	Recall	F1 Score
FER-2013 Test Set	Anger	0.50	0.59	0.54
	Fear	0.36	0.33	0.34
	Happy	0.81	0.73	0.77
	Sad	0.45	0.47	0.46
	Surprise	0.68	0.66	0.67
	Accuracy			0.56
	Macro avg.	0.56	0.56	0.56
	Weighted avg.	0.56	0.56	0.56
NIFER Test Set	Anger	0.26	0.34	0.29
	Fear	0.22	0.16	0.19
	Happy	0.58	0.65	0.61
	Sad	0.30	0.43	0.35
	Surprise	0.44	0.18	0.26
	Accuracy			0.35
	Macro avg.	0.36	0.35	0.34
	Weighted avg.	0.36	0.35	0.34

Finally, we used a combination of images from both datasets for training and testing. To achieve this, we combined the training sets from both datasets to make a single training set of 4350 images and tested on the 550 test images of each dataset. The results are shown in Table 4, and they depict nearly the same performance trend as seen in Tables 2 and 3. Despite the larger training set size, the performance on the FER-2013 test set does not improve, while that of NIFER improves from 82% to 83%. This demonstrates that the FER-2013 may be more challenging because of its low-quality images but also shows that multi-ethnic facial expression recognition can significantly improve by including in the training set, more faces across multiple ethnicities. This second indication is important as it directly relates to our research question. The fact that the recognition improves for the NIFER test set from Table 3 to Table 4 shows that more ethnic diversity is beneficial for improved multi-ethnic facial expression recognition.

Table 3. Results of Training with images of the NIFER dataset

	Class	Precision	Recall	F1 Score
FER-2013 Test Set	Anger	0.38	0.14	0.20
	Fear	0.23	0.15	0.18
	Happy	0.43	0.74	0.55
	Sad	0.29	0.53	0.38
	Surprise	0.43	0.22	0.29
	Accuracy			0.35
	Macro avg.	0.35	0.35	0.32
	Weighted avg.	0.35	0.35	0.32
NIFER Test Set	Anger	0.88	0.77	0.82
	Fear	0.90	0.86	0.88
	Happy	0.79	0.91	0.84
	Sad	0.71	0.70	0.73
	Surprise	0.86	0.83	0.84
	Accuracy			0.82
	Macro avg.	0.83	0.82	0.82
	Weighted avg.	0.83	0.82	0.82

Table 4. Results of Training with images of the combined dataset (FER-2013 and NIFER)

	Class	Precision	Recall	F1 Score
FER-2013 Test Set	Anger	0.52	0.54	0.53
	Fear	0.37	0.34	0.35
	Happy	0.72	0.75	0.73
	Sad	0.40	0.42	0.41
	Surprise	0.68	0.67	0.68
	Accuracy			0.54
	Macro avg.	0.54	0.54	0.54
	Weighted avg.	0.54	0.54	0.54
NIFER TEST SET	Anger	0.83	0.83	0.83
	Fear	0.80	0.85	0.85
	Happy	0.84	0.88	0.86
	Sad	0.78	0.71	0.74
	Surprise	0.82	0.87	0.85
	Accuracy			0.83
	Macro avg.	0.83	0.83	0.83
	Weighted avg.	0.83	0.83	0.83

5. Conclusion

In this research, the importance of racially diverse training examples to the facial expression recognition problem was demonstrated. A new dataset, the Nigerian Facial Expression Recognition (NIFER) dataset, was introduced, primarily consisting of 3,481 images featuring dark-skinned faces with diverse attributes such as illumination, poses, resolution, and brightness. This dataset aims to increase diversity within facial expression databases, which predominantly feature non-black faces. Images in NIFER were classified into five basic emotions: Anger, Fear, Happiness, Sadness, and Surprise. A deep learning model was trained and tested on both the NIFER dataset and the FER-2013 dataset, with results compared. The findings showed that racially biased datasets cannot produce effective facial recognition models, highlighting the need for subjects from multiple ethnicities. The NIFER dataset represents a first step in this direction, introducing racial diversity into existing facial expression recognition datasets.

Future work will focus on expanding the size and scope of the NIFER dataset and conducting more extensive experiments across facial expression datasets representing other ethnicities. We also plan to further investigate the performance of advanced models like Vision Transformer (ViT) and Deep Attention Network on interracial facial expression datasets, focusing on how well these models generalize across diverse racial groups. By comparing their performance, we aim to identify the model that best captures the nuances of facial expressions in different demographic groups, thereby improving the inclusivity of FER systems. The aim is to develop FER systems that are robust and invariant to racial diversity, thus advancing the field of facial expression recognition.

References

[1] H. Yang, U. Ciftci, and L. Yin, “Facial Expression Recognition by De-expression Residue Learning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168–2177.

[2] S. Umer, R. K. Rout, C. Pero, and M. Nappi, “Facial expression recognition with trade-offs between data augmentation and deep learning features,” Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 2, pp. 721–735, 2021. [Online]. Available: https://doi.org/10.1007/S12652-020-02845-8.

[3] A. Pourramezan, M. H. Mahoor, and A. P. Fard, “Ad-Corre: Adaptive Correlation-Based Loss for Facial Expression Recognition in the Wild,” IEEE Access, vol. 10, pp. 26756-26768, 2022, doi: 10.1109/ACCESS.2022.3156598.

[4] F. Zhang, T. Zhang, Q. Mao, and C. Xu, “Joint Pose and Expression Modeling for Facial Expression Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 3395-3368.

[5] T. Shahzad, K. Iqbal, M. A. Khan, and N. Iqbal, “Role of Zoning in Facial Expression Using Deep Learning,” IEEE Access, vol. 11, pp. 16493-16508, 2023, doi: 10.1109/ACCESS.2021.DOI.

[6] K. Li, Y. Jin, M. W. Akram, R. Han, and J. Chen, “Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy,” Visual Computer, vol. 36, no. 2, pp. 391-404, 2020, doi:10.1007/s00371-019-01627-4.

[7] 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.

[8] M. J. Lyons, M. Kamachi, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)," ArXiv Preprint ArXiv:2009.05938, 2020.
[9] L. Schoneveld and A. Othmani, "Towards a General Deep Feature Extractor for Facial Expression Recognition," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2021, pp. 2339-2342, doi: 10.1109/ICIP42928.2021.9506025.
[10] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 18-31, Jan.-Mar. 2019, doi:10.1109/TAFFC.2017.2740923.
[11] R. Vemulapalli, G. Ai, and A. Agarwala, "A Compact Embedding for Facial Expression Similarity," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5683-5692.
[12] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in Proc. 2021 IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2021, pp. 2401-2410, doi: 10.1109/WACV48630.2021.00245.
[13] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," Sensors, vol. 21, no. 9, article 3046, 2021, doi: 10.3390/s21093046.
[14] S. Stoychev and H. Gunes, "The effect of model compression on fairness in facial expression recognition," in Proc. Int. Conf. Pattern Recognit., Aug. 2022, pp. 121-138, Cham, Springer Nature Switzerland.
[15] I. J. Goodfellow 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, vol. 20, pp. 117-124, 2013.
[16] J. Wang, "Improved Facial Expression Recognition Method Based on GAN," Scientific Programming, 2021, doi:10.1155/2021/2689029.
[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499-1503, Oct. 2016.
[18] R. Zhi, C. Zhou, T. Li, S. Liu, and Y. Jin, "Action unit analysis enhanced facial expression recognition by deep neural network evolution," Neurocomputing, vol. 425, pp. 135-148, 2021, doi: 10.1016/j.neucom.2020.03.036.
[19] A. M. Reza, "Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement," J. VLSI Signal Process. Syst., vol. 38, no. 1, pp. 35-44, 2004, doi:10.1023/B:VLSI.0000028532.53893.82.
[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in Proc. 2018 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2018), 2018, pp. 67-74.

Authors' Profiles

Ridwan O. Bello is currently pursuing a Ph.D. in Computer Science at The University of Alabama, Tuscaloosa, USA. He received his B.Sc. in Computer Science from Fountain University, Osogbo, Nigeria, in 2017, and his M.Sc. in Computer Science from the University of Ibadan, Nigeria, in 2022. His research interests include Computer Vision, Affective Computing, and Speech Recognition, with a focus on developing efficient AI systems for low-resource communities in Africa.

Joseph D. Akinyemi is currently with the University of York, York, United Kingdom. He received his Bachelor's degree in Computer Science from the University of Ilorin, Ilorin, Nigeria in 2010. He received his Master's degree in Computer Science from the University of Ibadan, Ibadan, Nigeria, in 2014 and a Ph.D. degree in Computer Science from the same institution in 2020. His research spans areas of Computer Vision such as facial and medical image processing as well as aspects of Natural Language Processing such as Sentiment Analysis. He is a 2022 Heidelberg Laureate Forum Fellow in Germany, a recipient of the Google Developers Machine Learning Bootcamp sponsorship for Sub-Saharan Africa and a member of the ACM.

Khadijat Ladoja received her B.Sc. degree in Computer Science from the University of Ilorin, Nigeria, in 2010. She then completed her M.Sc. and Ph.D. degrees in Computer Science at the University of Ibadan, Nigeria, in 2014 and 2021, respectively. Currently, she is a faculty member in the Department of Computer Science at the University of Ibadan, with over five years of teaching and research experience. Her research focuses on Natural Language Processing, specifically targeting language models for low-resource Nigerian languages and computer vision. Dr. Ladoja's dedication to teaching and research has earned her the fellowship of the "Empowering the Teachers" program at MIT, USA, and recognition as one of the Top 200 young researchers by the Heidelberg Laureate Foundation (HLF).

Akomolafe Oladeji Patrick is currently a Full Stack Developer at Massload Technologies, SK Canada. Before joining Massload, he was a Lecturer at the Department of Computer Science, University of Ibadan, Nigeria. He obtained a Bachelor of Technology (B.Tech) in Computer Engineering at Ladoke Akintola University of Technology, Ogbomoso, Nigeria (LAUTECH) in 1999, a Master of Science Degree (M.Sc) in Computer Science at the University of Ibadan, Ibadan, Nigeria in 2004 and a Ph.D. Degree in Computer Science from LAUTECH, Ogbomoso, Nigeria in 2014. He also has a Master degree in Data Science from University of West in England in 2022. His research interests include Pervasive and Mobile Computing, Mobile Agent Technology, Context Aware Computing and Software Engineering.

How to cite this paper: Ridwan O. Bello, Joseph D. Akinyemi, Khadijat T. Ladoja, Oladeji P. Akomolafe, "Racial Bias in Facial Expression Recognition Datasets: Evaluating the Impact on Model Performance", International Journal of Engineering and Manufacturing (IJEM), Vol.15, No.1, pp. 1-10, 2025. DOI:10.5815/ijem.2025.01.01