Investigating Cross-cultural Generalizability of Facial Emotion Recognition with Multi-dataset Training

YULIN CHEN, University of Twente, The Netherlands

Facial emotion recognition (FER) is among computer vision's most complex fields and has several practical uses in human-computer interaction (HCI) and psychology. Currently, existing FER models are trained on datasets dominated by a singular ethnicity, for instance, AffectNet [19]. AffectNet is a widely used dataset for FER tasks with over 400,000 manually annotated samples, and has 64.4% of its training data represented by White subjects [3]. As a result, the accuracy is often limited when the model is deployed in the real world, where the population is much more culturally diverse. This research will investigate the impact of augmenting existing FER datasets with **EiLA (Emotions in LatAm Dataset)**, a newly curated emotion recognition in-the-wild dataset consisting of video recordings of Latin American populations and their facial expressions, on the accuracy and performance of well-known FER models. The study begins with dataset preparation, where different-sized portions of the EiLA dataset are integrated with two existing FER datasets, to form larger datasets. Next, the integrated datasets are used to train the chosen neural network, followed by testing and metrics evaluation using ground truth labels from the EiLA dataset. Finally, the analysis of results will be interpreted to determine whether augmenting the cultural diversity of datasets positively impacts the efficacy of FER models.

Additional Key Words and Phrases: Facial emotion recognition (FER), Deep convolutional neural networks (DCNN), computer vision, image analysis, data processing

1 INTRODUCTION

Facial expressions and the showcase of emotions through body language are significant in human interaction and contain 55% of the emotional information communicated [18]. Facial expressions especially, carry a lot of information such as the emotions a person is trying to convey [5]. Therefore, teaching machines to perceive and interpret human body language effectively, has become a pivotal task in human-computer interaction (HCI).

The interpretation of facial emotions is significantly affected by the cultural background of an individual [8]. A facial emotion recognition (FER) model, trained on a dataset with most subjects from a singular ethnicity, for instance, White-Caucasian males of the United States, may struggle to differentiate the cultural nuances of facial expressions of people from Southeast Asia. As a result, models trained using such datasets are vulnerable to bias. Racial bias is a prominent problem in state-of-the-art FER methods. Past experiments had shown that FER models trained using datasets of a singular ethnicity performed significantly better when the test subjects were from the same ethnicity than when they were not [15]. At present, many widely used datasets for FER suffer from this problem, such as MMI (MMI Face Database) [21] and CK+ (Extended Cohn-Kanade dataset) [15], which consist of primarily Caucasian subjects. Models trained on such datasets, performed poorly when evaluated on datasets curated for a different region, such as the JAFFE (Japanese Female Facial Expression) dataset [16, 17].

FER models use an array of emotional cues, namely action units (AU) and micro/macro facial expressions to classify the emotion detected. As such, a model trained on the emotional expressions of a certain ethnicity, would not be able to generalize well with other ones, as there are subtle differences in the ways different races and cultures express their emotions, which are not included in the training data. As suggested by Lukáč et al., an ensemble of FER models trained on culturally diverse datasets can allow the model to pick up those differences more easily, which can reduce inter-culture bias in recognizing emotions. Consequently, a FER model that can generalize well can be expected to achieve better accuracy [16].

For this research, instead of devising an ensemble of models and neural network architectures to improve cross-cultural generalizability and model performance, a different approach was taken, with the focus being on how enhancing the training data can achieve better results. The research seeks to answer the following questions:

RQ1: *How does augmenting existing FER datasets with samples from the EiLA dataset affect the accuracy of existing FER methods in recognizing the emotions of individuals?*

RQ2: *To what extent does the integration of EiLA with other FER datasets reduce racial bias in the form of performance discrepancies, for existing FER methods?*

2 SCIENTIFIC BACKGROUND

Several existing research have laid the groundwork for the augmentation of popular Facial Emotion Recognition (FER) datasets. Notably, research by Mollahosseini et al. [20] has suggested the lack of cultural diversity in FER datasets and the need for more inclusive approaches. Specifically, the study discovered the limitations of existing datasets, in capturing facial expressions of different ethnicities and cultural backgrounds equally. Furthermore, the importance of dataset augmentation to address biases and improve the generalization capability of FER models was highlighted.

Additionally, Javadi and Lim [9] worked on the creation of a new FER dataset for Persians, which they considered to be an under-represented ethnic group in the FER world. The research highlighted that people express and interpret emotions differently, across different cultures. Currently, there are no publicly available FER datasets focused on the Latin American population, which makes EiLA a first of its kind.

TScIT 41, July 5, 2024, Enschede, The Netherlands

Furthermore, recent work by Fan et al. [4] attempted to address the problem of racial bias in FER, by sub-sampling training data with different racial distributions into multiple sets and measuring the performance of FER models trained on each set. The study found that in smaller datasets, the racial balance of the training set had a positive correlation with the fairness and performance of FER models. Using a sub-sample of the CAFE (Child Affective Facial Expression) dataset, the F1-score increased by 27.2%, and the demographic parity (similarity in prediction accuracy across different ethnicities) by 15.7% [14].

3 METHODOLOGY & EXPERIMENTAL SETUP

The research will follow the methodology and approach as depicted in Figure 1. The preparation process begins with the collection of data, followed by image pre-processing and preparation of integrated datasets. Then, the experiment phase begins with training the selected model (Figure 6) using the datasets in Table 3. The final step would be to test the model using ground truths from the EiLA test set. The performance and generalizability are to be measured through accuracy and fairness metrics.

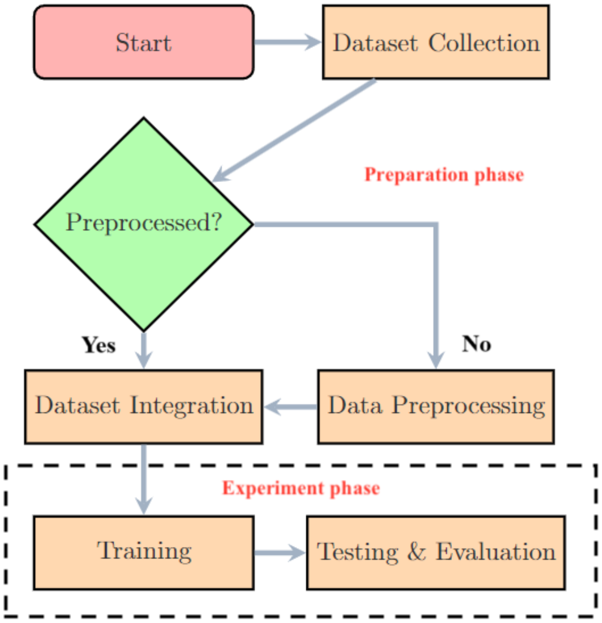


Fig. 1. Research Methodology Pipeline

3.1 Preparation

Data Collection: The three datasets listed in Table 1 were collected for this research. The ***EiLA (Emotions in LatAm)*** dataset, as shown in Figure 3, was provided by The University of Twente and consists of video recordings from TV shows aired in Latin America. ***FER2013*** [6] and ***SFEW (Static Facial Expressions in the Wild)*** [2] were also chosen. FER2013 (Figure 2) is a popular dataset with many pre-processed samples for each emotion class. SFEW (Figure 4) on the other hand, is a smaller subset of static frames of the dataset *AFEW (Acted Facial Expressions In The Wild)*, which originated from popular TV shows and movies. The nature of the images is very similar to that of EiLA, which makes it an ideal candidate for this study.

Table 1. Datasets overview

Dataset	Sample format	Size	Pre-processed
FER2013	48x48 grayscale image	35,887	Yes
SFEW	48x48 grayscale image	1,322	Yes
EiLA	1280x720 color video	8,088	No

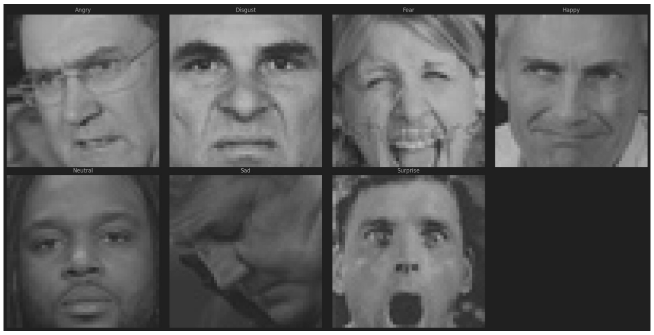


Fig. 2. Samples from FER Dataset

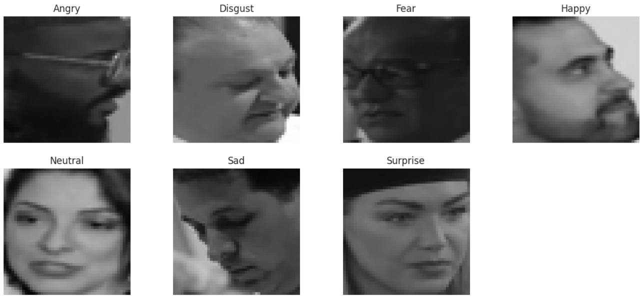


Fig. 3. Samples from EiLA dataset (processed)

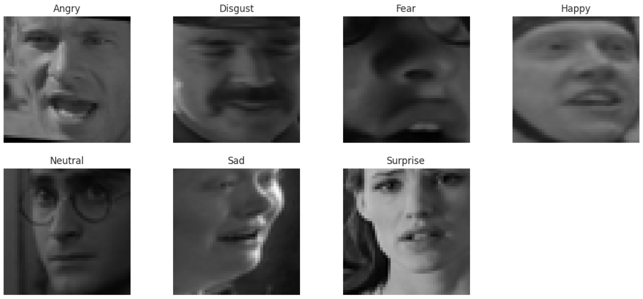


Fig. 4. Samples from SFEW dataset

Table 2. Sample EiLA dataset annotations

Video Tag	Cid	Labels	Frame Number	X	Y	Width	Height	Pid
...	...	...	...	...	...	...	...	...
ngITkMvWuq8	3	[['Happy'], 'No annotation', ['Neutral']]	22471	0.05208333333	6.78219417	73.68036263	93.12521324	9
ngITkMvWuq8	3	[['Happy'], 'No annotation', ['Neutral']]	22477	0.05208333333	6.78219417	72.97850563	93.12521324	9
ngITkMvWuq8	3	[['Happy'], 'No annotation', ['Neutral']]	22483	0.05208333333	6.78219417	72.27664863	93.12521324	9
ngITkMvWuq8	3	[['Happy'], 'No annotation', ['Neutral']]	22489	3.314621716	6.78219417	69.15301945	93.12521324	9
ngITkMvWuq8	3	[['Happy'], 'No annotation', ['Happy']]	22495	22.41555783	0.1421585352	70.11258956	99.71568293	0

Data Preprocessing: The EiLA dataset is the only one of the three chosen datasets that need to be pre-processed, as the other two datasets both contain samples of 48x48 grayscale images, whereas the EiLA dataset consists of high-resolution video recordings. To integrate EiLA with FER2013 & SFEW, samples from the EiLA dataset must be converted to static 48x48 grayscale images. As illustrated in Figure 5, the pre-processing begins with the extraction of static frames, using existing annotations as shown in Table 2. The annotations also contain the XY coordinates of the bounding box of the person in interest and their emotion in the frame. Then, the bounding box is cut from the frame, and face detection is used to isolate the face from the rest of the framework using YOLOv8, a state-of-the-art object detection tool that can also be applied to faces [22]. Frames where the tool cannot identify a face were discarded. Next, the remaining images were cropped and resized to match the standard dimensions of the integrating dataset. The resulting set of EiLA samples was separated into one of the three sets below, based on the number of valid annotations for the emotions perceived by a maximum of three human annotators:

Train: 3 out of 3 annotations
Validation: 2 out of 3 annotations
Test: 1 out of 3 annotations

Furthermore, the EiLA dataset provides demographic data for every person represented in the frame. This also includes each individual's race, which is represented as Black, White or Mixed. After all images have been pre-processed, the demographic information was extracted and added to the test set, stored in a *numpy* dataframe.

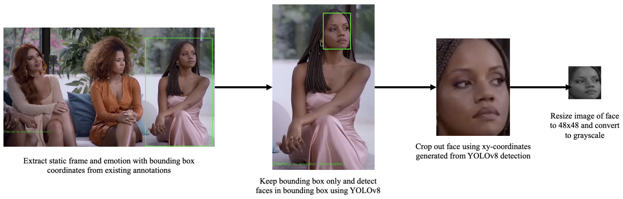

Extract static frame and emotion with bounding box coordinates from existing annotations
Keep bounding box only and detect faces in bounding box using YOLOv8
Crop out face using xy-coordinates generated from YOLOv8 detection
Resize image of face to 48x48 and convert to grayscale

Fig. 5. Pre-processing of EiLA dataset

Table 3. Created Experiment Configurations

Experiment Set	Training Set Size	Validation Set Size
EiLA only	**4,852**	**1362**
FER2013 only	**28,709**	**7,178**
SFEW only	**891**	**431**
SFEW + 10% EiLA	1373	564
SFEW + 30% EiLA	2,342	837
SFEW + 50% EiLA	3,316	1,111
SFEW + 100% EiLA	5,743	1,793
FER2013 + 10% EiLA	29,191	7,311
FER2013 + 30% EiLA	30,160	7,584
FER2013 + 50% EiLA	31,134	7,858
FER2013 + 100% EiLA	33,561	8,540

Table 4. Racial Composition of EiLA test set

Race	Count	Portion
Black	163	20.00%
White	588	72.15%
Mixed	64	7.85%
Total	815	100%

Dataset Integration: The next step involves integrating various proportions of the pre-processed EiLA dataset as listed in Figure 3, selected randomly using Python's *random* module, with each of the existing datasets chosen for this study. Eight augmented datasets were created, each consisting of the full FER2013 or SFEW dataset, plus a portion of EiLA. For example, if the EiLA training and validation sets contained 100 and 20 images respectively, an augmented FER2013 dataset containing 50% of data from EiLA will include all the training and validation samples from FER2013, as well as 50 training samples and 10 validation samples from EiLA.

3.2 Model Selection

In this research, a deep learning approach is proposed for the classification of emotions. The approach consisted of using ResNet50V2 as the base model and pre-trained ImageNet [1] weights for transfer learning. This approach initializes the model with meaningful pre-trained weights, which saves a considerable amount of time needed for training.

ResNet50V2 Architecture

Figure 6 presents the basic architecture of ResNet50v2, a deep convolutional neural network (DCNN) specifically built for the task

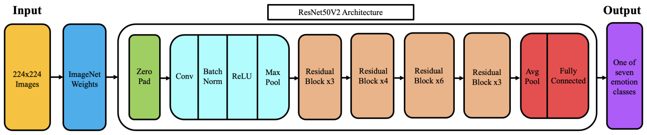


Fig. 6. ResNet50V2 Architecture

of image recognition. It is a 50-layer variant of the Deep residual neural network *ResNet*, which had previously achieved state-of-the-art performances on benchmarks such as ImageNet [1]. The network specifically makes use of residual blocks with shortcut connections (Figure 7) to mitigate common issues such as degradation, vanishing/exploding gradients and slow Convergence [7]. The model is initialized with pre-trained weights and receives an input of 224x224x3 images. First, the images are padded with zeros and pass through a convolution, followed by batch normalization, the ReLU activation function in Equation (1) and max pooling. Then, the image is passed through five stages with 3 residual blocks each, where each block consists of three convolutions (1x1, 3x3, 1x1). Next, a global average pooling layer is applied, which reduces each 7x7 feature map from the image to a single value, by averaging the values in every feature map. Finally, a fully connected dense layer is used for the final classification. As shown in Equation 2, the Softmax activation function is applied to compute a vector of probabilities of each class, and the class which receives the highest probability becomes the final output. [24]

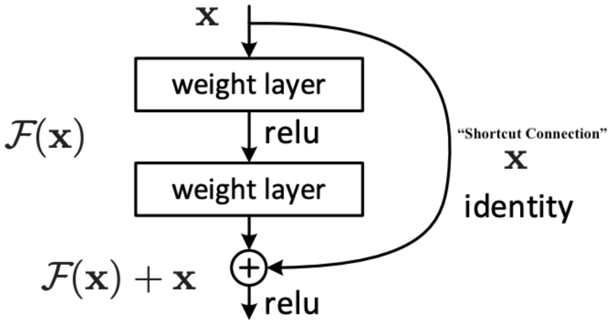


Fig. 7. Example of a Residual Block

$$\mathrm{ReLU}(z) = \max(0, z) \tag{1}$$

$$\mathrm{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{2}$$

where the input is a vector $z = [z_1, z_2, \ldots, z_n]$, and z_i is the i-th element.

Loss Function & Optimizer

For the training of our model, the Adamax optimizer [11] is used to adjust the weights and biases of the model to minimize the loss function. Loss is a measure of the differences between the model's predictions and the ground truth. A lower loss value indicates better model performance. For this research, categorical cross-entropy is used to calculate loss, as shown in Equation 3.

Categorical Cross-Entropy Loss

The cross-entropy loss function measures the difference between the predicted probabilities of a given model, and the true label, and is defined as:

$$Loss(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

where y is the true label and $\hat{y}$ is the predicted probability.

Adamax

Adamax is an extension of the Adam optimizer [12], which uses the infinity norm (max norm) to scale the gradient updates. The following section describes how the Adamax optimizer operates in a neural network. First, the following parameters are initialized:

θ_0: Initial parameter vector
m_0: 1st moment vector
u_0: Exponentially weighted infinity norm
η: Learning rate (small value constant)
β_1 and β_2: Exponential decay rates (constants between 0-1)

Then, while θ_t has not converged, at every time step t, given the parameters θ_t, learning rate η, and decay rates β_1 and β_2:

1. Compute the gradient of the loss function:

$$g_t = \nabla_\theta J(\theta_{t-1})$$

2. Compute biased first-moment estimate

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

3. Compute exponentially weighted infinity norm:

$$u_t = \max(\beta_2 u_{t-1}, |g_t|)$$

4. Compute bias-corrected first-moment estimate:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

5. Update parameters:

$$\theta_{t+1} = \theta_t - \frac{\eta}{u_t}\hat{m}_t$$

where:
- g_t is the gradient of the loss function at time step t
- m_t is the first-moment vector (moving average of the gradients)
- u_t is the exponentially weighted infinity norm
- $\hat{m}_t$ is the bias-corrected first moment estimate

Callbacks

Callbacks are useful functions that are executed in between epochs during the training of our model. They provide the ability to automate adjustments to the model midst training, which allows for greater control and flexibility over the process. For this research, two callback functions, Early Stopping and ReduceLROnPlateau were used.

Early Stopping is a regularization method used to prevent overfitting. It is triggered when the validation accuracy stops improving during training. Another benefit of this method is the saved computational time by terminating early.

ReduceLROnPlateau is another regularization method used to monitor the model's performance and reduce the learning rate when there is no improvement for several epochs. The learning rate reduction on plateau strategy is defined as:

$$\eta_{\text{new}} = \eta_{\text{old}} \times \text{factor}$$

where η is the learning rate of the model, and $factor$ is a number between $0-1$ that determines how much the learning rate is reduced by each method call.

3.3 Experimentation

Training: The experimentation phase started with loading the datasets for training. For that, *ImageDataGenerator* from the library *keras* was used to fetch and process the datasets from the directories where they were stored. The tool allowed for convenient batching, normalization and resizing of the images to match the standard input size of 224x224x3 for ResNet50V2. The next step in the experiment was to choose the loss function, optimizer and training hyper-parameters. For this research, categorical cross-entropy was used to evaluate model performance throughout training. The initial training consisted of 15 epochs with an initial learning rate of 1×10^{-4} using the Adamax optimizer. Afterwards, the last 10 layers of the base model were unfreezed for fine-tuning of an additional 35 epochs, with a reduced initial learning rate of 1×10^{-5}. Furthermore, Dropout with a rate of 10%, Early Stopping with a patience of 15 epochs and ReduceLROnPlateau with a factor of 0.2, patience of 3 epochs, a minimum learning rate of 1×10^{-7} were added to prevent over-fitting.

Testing & Evaluation: For this research, the EiLA test set was used as the testing benchmark, for all models trained using the datasets configurations in Table 3. To evaluate whether augmenting FER datasets with samples from EiLA affects the accuracy of existing FER methods, the weighted f1 score (Equation 4 & 5), accuracy (Equation 6) and categorical cross-entropy loss (Equation 3) were computed for each experiment. The performance metrics were compared with the baseline, which are the experiments highlighted in bold in Table 3. Additionally, to measure the extent to which dataset integration affects racial bias in FER models, the model's accuracy per racial group (Equation 7), and the standard deviation in percentages, were calculated using the ground truths and demographic information from the test set.

To calculate the F1 score for class i:

$$F1Score_i = 2 \times \left(\frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \right) \tag{4}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

where for some class i:

TP_i (True Positives) is the number of correct class i predictions.

FP_i (False Positives) is the number of incorrect class i predictions.

FN_i (False Negatives) is the number of predictions made to another class, even though the ground truth is class i.

Once the f1 score has been computed for each class in the dataset. The weighted f1-score can be calculated using the formula below:

$$\text{Weighted F1 Score} = \frac{\sum_{i=1}^{n} \text{Support}_i \times F1Score_i}{\sum_{i=1}^{n} \text{Support}_i} \tag{5}$$

where n is the number of classes, and $Support_i$ is the number of samples labelled as class i) in the ground truth.

To calculate overall model accuracy:

$$\%\text{Accuracy} = \left(\frac{\text{Total \# of correct predictions}}{\text{Total \# of samples}} \right) \times 100 \tag{6}$$

To calculate model accuracy for racial group G:

$$\%\text{Accuracy(G)} = \left(\frac{\text{\# correct predictions in G}}{\text{\# of samples in G}} \right) \times 100 \tag{7}$$

4 RESULTS & DISCUSSION

Table 5 presents the performance metrics recorded for all experimental setups listed in Table 3. This includes the F1 score, accuracy, and categorical cross-entropy loss. For the baseline, the model trained only using the EiLA training data achieved an F1 score of 0.267, accuracy of 31.9% and loss of 2.733 when evaluated against the EiLA test set. The model trained using the SFEW dataset reported an F1 score of 0.254, accuracy of 28.3% and loss of 1.838. As for the model trained using the entirety of the FER2013 training set, it achieved the best baseline performance of the three with an F1 score of 0.342, accuracy of 40.2% and loss of 2.274.

As for integrated models, integrating 10% of training data from EiLA with SFEW helped improve the F1 score from 0.254 to 0.278, and accuracy from 28.3% to 34.0%, a 5.7% increase. On the contrary, cross-entropy loss increased slightly from 1.838 to 2.030. Other models with more data integration from EiLA also yielded improvements in comparison to the baseline. However, there was no correlation between higher integration percentages and consistent performance

Table 5. Measured Performance For Each Model

Experiment	Weighted F1 score	%Accuracy	Loss
EiLA only	**0.267**	**31.9**	**2.733**
SFEW only	**0.254**	**28.3**	**1.838**
SFEW + 10% EiLA	0.278	34.0	2.030
SFEW + 30% EiLA	0.255	30.3	2.285
SFEW + 50% EiLA	0.283	32.8	2.311
SFEW + 100% EiLA	0.256	30.9	2.675
FER2013 only	**0.342**	**40.2**	**2.274**
FER2013 + 10% EiLA	0.332	38.9	2.420
FER2013 + 30% EiLA	0.334	38.5	2.520
FER2013 + 50% EiLA	0.341	39.8	2.492
FER2013 + 100% EiLA	0.327	36.9	2.608

Table 6. Measured Performance For Each Model on Represented Race

Experiment	%Acc(Black)	%Acc(White)	%Acc(Mixed)	%Std
EiLA only	**36.8**	**31.1**	**26.6**	**5.14**
SFEW only	**38.7**	**25.9**	**25.0**	**7.65**
SFEW + 10% EiLA	45.4	31.3	29.7	8.65
SFEW + 30% EiLA	41.1	27.6	28.1	7.66
SFEW + 50% EiLA	42.9	30.3	29.7	7.49
SFEW + 100% EiLA	43.6	26.2	32.8	8.77
FER2013 only	**58.9**	**34.5**	**45.3**	**12.2**
FER2013 + 10% EiLA	46.0	37.6	32.8	6.68
FER2013 + 30% EiLA	43.6	37.9	31.3	6.16
FER2013 + 50% EiLA	50.9	37.4	32.8	9.41
FER2013 + 100% EiLA	42.3	35.5	35.9	3.81

gains, as the F1 score and accuracy fluctuated as more EiLA samples were integrated with SFEW. An additional observation was that the loss value continued to increase, as more samples were being integrated. The model trained using SFEW and 100% of the EiLA training set, recorded the highest loss of 2.675, much higher in comparison to the baseline value of 1.838. On the other hand, integrating the EiLA dataset with FER2013 resulted in minor reductions in performance across all experiments. The best-performing model of them all, trained with FER2013 and 50% of samples from EiLA, recorded an F1 score of 0.341, accuracy of 39.8% and loss of 2.492, which are slightly worse than the baseline model.

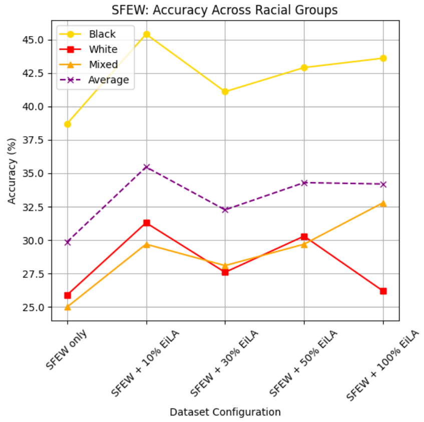


Fig. 8. Accuracy of models trained on SFEW+EiLA across racial groups

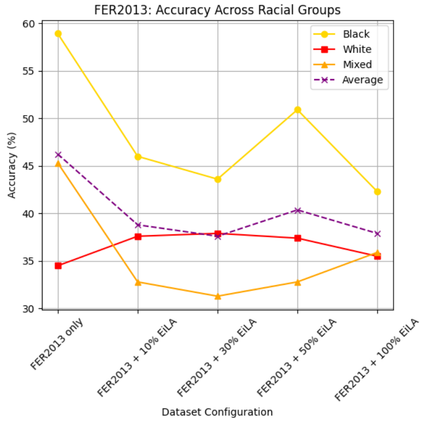


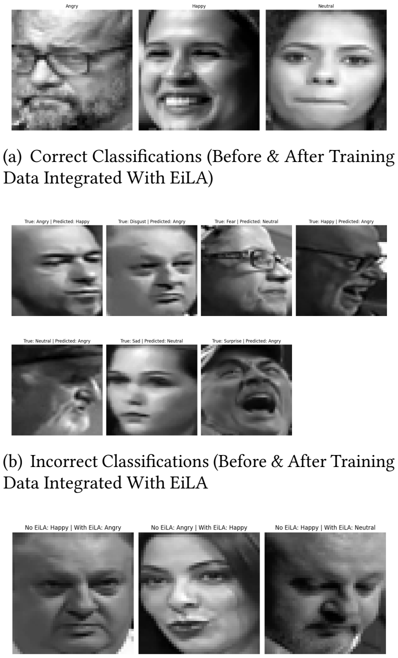
Fig. 9. Accuracy of models trained on FER2013+EiLA across racial groups

Table 6 details the accuracy of different models at predicting emotions for each of the racial groups represented in the EiLA test set (Black, White, Mixed). The baseline model trained solely using EiLA's training data, showed the highest accuracy for Black individuals (36.8%), followed by White (31.1%) and Mixed (26.6%). Similarly, the baseline model trained only using SFEW was the most accurate for Black individuals (38.7%), and less accurate at predicting subjects with White (25.9%) and Mixed (25.0%) skin tones. The model trained with FER2013 data showed the highest baseline accuracy for all racial groups at 58.9% for black individuals, 34.5% and 45.3% respectively for White and Mixed groups. As demonstrated in Figure 8, adding data samples of EiLA to SFEW improved the accuracy for all racial groups, with the most significant improvement (6.7%, 5.4% & 4.7% for Black, White and Mixed groups respectively) observed in the model with 10% of EiLA integrated with SFEW. Despite yielding worse performances compared to the 10% model, further integration still performed better than the baseline. Conversely, as depicted in Figure 9, integrating EiLA data with FER2013 only resulted in a 3.4% improvement at most for white individuals. On the other hand, significant reductions in accuracy (-16.6% & -14.0% at most) were observed for Black and Mixed groups.

Discussion

How does augmenting existing FER datasets with samples from the EiLA dataset affects the accuracy of existing FER methods in recognizing the emotions of individuals?

The results from the above experiments suggest that augmenting a small dataset such as SFEW with a small portion of samples from EiLA can improve the overall model performance, as indicated by the accuracy metrics. However, higher levels of integration yielded diminishing performance gains. This could be because EiLA is several times larger than SFEW, and the model might overfit when EiLA make up the majority of the integrated dataset. As a result, the model may be unable to learn the characteristics of the SFEW samples and become unable to generalize between datasets. As for the much larger FER2013 dataset, integration with EiLA resulted in

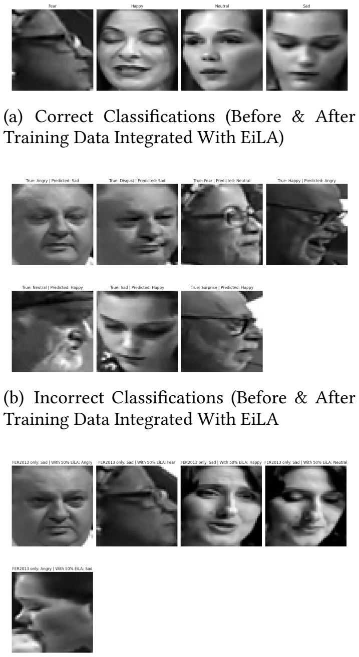

Angry
Happy
Neutral
True: Angry | Predicted: Happy
True: Disgust | Predicted: Angry
True: Fear | Predicted: Neutral
True: Happy | Predicted: Angry
True: Neutral | Predicted: Angry
True: Sad | Predicted: Neutral
True: Surprise | Predicted: Angry
No EiLA: Happy | With EiLA: Angry
No EiLA: Angry | With EiLA: Happy
No EiLA: Happy | With EiLA: Neutral

(a) Correct Classifications (Before & After Training Data Integrated With EiLA)

(b) Incorrect Classifications (Before & After Training Data Integrated With EiLA

(c) Corrected Classifications After Training Data Integrated With EiLA

Fig. 10. Classification Examples Between Model Trained Using Only SFEW versus SFEW+10%EiLA (Best-performing)


Fear
Happy
Neutral
Sad
True: Angry | Predicted: Sad
True: Disgust | Predicted: Sad
True: Fear | Predicted: Neutral
True: Happy | Predicted: Angry
True: Neutral | Predicted: Happy
True: Sad | Predicted: Happy
True: Surprise | Predicted: Happy
FER2013 only: Sad | With 50% EiLA: Angry
FER2013 only: Sad | With 50% EiLA: Fear
FER2013 only: Sad | With 50% EiLA: Happy
FER2013 only: Sad | With 50% EiLA: Neutral
FER2013 only: Angry | With 50% EiLA: Sad

(a) Correct Classifications (Before & After Training Data Integrated With EiLA)

(b) Incorrect Classifications (Before & After Training Data Integrated With EiLA

(c) Corrected Classifications After Training Data Integrated With EiLA

Fig. 11. Classification Examples Between Model Trained Using Only FER2013 versus FER2013+50%EiLA (Best-performing)

worse performances, compared to the baseline. This may suggest that the diversity introduced by EiLA may be insufficient as it is too small to make a significant representation in the integrated dataset.

Another possible reason for the observed behaviour is the differences in characteristics such as angle, lighting and background of the samples, which can confuse the model. The photographic styles of samples in FER2013 and EiLA are quite different. As depicted in Figure 2 and 3, the majority of FER2013 samples are forward-facing and make eye contact with the lens, whereas in the EiLA dataset, many samples show a side profile of the face, or are filmed at an angle where the subject does not make eye contact with the camera. This is a similar trend in SFEW as it was sourced similarly from TV shows and films. The results tend to support this theory, as the accuracy improved when EiLA was merged with SFEW, but not with FER2013.

Furthermore, in Figure 10 & 11, a small batch of prediction results from two of the best performing models (SFEW+10%EiLA & FER2013+50%EiLA) show how the models perform on the different samples in the EiLA test set. Most of the correct predictions are images where the subject faces forward, or at an angle where the facial features are still quite clear. On the other hand, a significant portion of the incorrect predictions come from samples depicting the side profile of subjects and/or containing some occlusion to facial features, such as glasses.

To what extent does the integration of EiLA with other FER datasets reduce racial bias in the form of performance discrepancies, for existing FER methods?

In most cases, combining the EiLA dataset with SFEW increased model accuracy for each ethnic group, although the percentage standard deviation of accuracy remained practically the same as the benchmark. This suggests that, even though the integration with EiLA enhanced the model's ability to detect emotions across all racial groups, it failed to eliminate performance gaps between the racial groups represented. On the contrary, integrating EiLA with FER2013 failed to boost model performance across racial groups in most instances. However, the percentage standard deviation reduced substantially, indicating that the model's performance across racial backgrounds became equal, even though the overall performance was worse than the baseline.

Therefore, it is reasonable to believe that the impact of dataset integration on accuracy and racial bias of FER models is **dataset-dependent** since results are quite clearly contrasting for experiments involving FER2013 and SFEW. Furthermore, there is a noticeable trade-off between a model's overall accuracy and balanced performance across all racial groups, as improvements in one metric result in a setback for the other.

Limitations

Class Imbalance was a significant limitation that impacted the accuracy of the models. As illustrated in Tables 7, 8, 9, classes such as Happy and Neutral are more represented across the datasets compared to Disgust. Such imbalance can lead to training a model biased towards predicting only the majority classes and ignoring the smaller classes altogether. Evidence of this behaviour can be found in Figure 12, in which the models trained on the FER2013 dataset, predicted almost exclusively one of its majority classes.
Domain Mismatch is another potential cause of poor model performance. The facial orientation and placement of the subject in images appear to be different for the datasets, especially FER2013 when compared to EiLA and SFEW. Additionally, accessories such as glasses partially obstruct the face. This distinction can increase the difficulty for models trained with FER2013 to learn the facial features present in the EiLA test set.
Insufficient Sample Size was another issue for this research. In contrast to FER2013 with over 35,000 samples, the EiLA dataset only consisted of roughly 8,000 samples, still more than SFEW, which holds an even lower figure at just over 1,300. Despite efforts to integrate datasets, the combined size is possibly too small to train a robust and generalized model that can identify the nuances of emotional expression between different cultural backgrounds.
Lack of Publicly Available Data was problematic, considering the limited time available for this research. There are very few FER datasets that can be freely accessed from the internet. Many datasets can only be obtained by making a formal request to their creators, which is in most cases a time-consuming process.

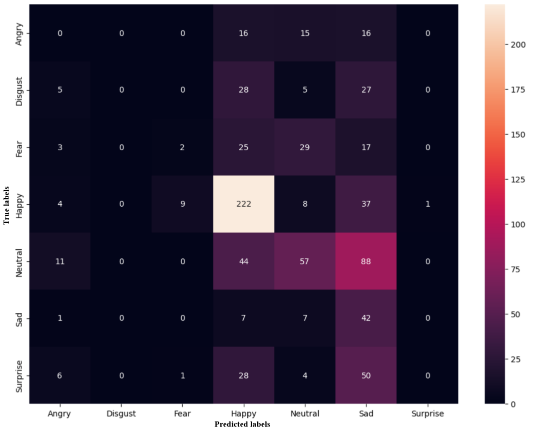


Fig. 12. Confusion matrix of model trained on FER2013 training set

Future work

Future research should begin with addressing the limitations of this study, by increasing the size, diversity and quality of datasets. Some methods to address this include data augmentation and synthetic data generation for smaller classes. Synthetic Minority Over-Sampling Technique (SMOTE) is a suitable over-sampling method, which interpolates between similar samples of a minority class to create new samples of that class [13]. Furthermore, Generative Adversarial Networks (GANs) can be used to generate synthetic data based on existing datasets. The synthetic data generated can closely resemble the original data, which makes this a good approach for populating minority classes [23]. To reduce the effects of face orientation and occlusions on the performance of FER models, existing methods such as CFR-GAN can be used in future research to apply facial re-orientation and de-occlusion to faces [10].

Table 7. FER2013 Training set class distribution

Total	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
28,709	3,995	436	4,097	7,215	4,965	4,830	3,171

Table 8. SFEW Training set class distribution

Total	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
891	178	52	78	184	144	161	94

Table 9. EiLA Training set class distribution

Total	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
4,852	944	96	76	1,053	2,516	122	45

5 CONCLUSIONS

Facial emotion recognition is a challenging task, due to the subtle nuances in the way we as humans express emotions. The complexity is further amplified when cultural differences are introduced to the problem, which highlights the need for FER models to be able to generalize facial expressions of people from different cultural backgrounds. This research explored the effects of integrating multiple datasets, namely *EiLA*, *FER2013* and *SFEW*, on the cross-cultural generalizability and performance of FER models. The experiments revealed that models trained using data consisting of EiLA and SFEW made notable improvements in performance, whereas models trained on FER2013 and EiLA performed slightly worse than the baseline. Furthermore, the study investigated the effect of dataset integration on the accuracy of FER models across different racial groups. The findings noted that while such a technique has the potential to reduce racial bias and improve overall accuracy, the limitations discussed in this paper must be appropriately addressed first in future research.

A APPENDIX

A.1 Declaration use of AI

During the preparation of this work, the author used Grammarly & Overleaf, in order to correct spelling, grammar, sentence structure and used expressions. In addition, ChatGPT had been used to assist with proper formatting of the scientific paper in LaTeX and written definitions of mathematical formulae. After using the above-mentioned tools & services, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[2] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimedia* 19 (09 2012), 34–31.

[3] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. 2022. Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition. arXiv:2205.10049 [cs.CV] https://arxiv.org/abs/2205.10049

[4] Alex Fan, Xingshuo Xiao, and Peter Washington. 2023. Addressing Racial Bias in Facial Emotion Recognition. arXiv:2308.04674 [cs.CV]

[5] Chris Frith. 2009. Role of facial expressions in social interactions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364 (12 2009), 3453–8. https://doi.org/10.1098/rstb.2009.0142

[6] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In *Neural Information Processing*, Minho Lee, Akira Hirose, Zeng-Guang Hou, and Rhee Man Kil (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 117–124.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[8] Rachael E. Jack. 2013. Culture and facial expressions of emotion. *Visual Cognition* 21, 9-10 (2013), 1248–1286. https://doi.org/10.1080/13506285.2013.835367 arXiv:https://doi.org/10.1080/13506285.2013.835367

[9] Roya Javadi and Angelica Lim. 2021. The Many Faces of Anger: A Multicultural Video Dataset of Negative Emotions in the Wild (MFA-Wild). arXiv:2112.05267 [cs.CV]

[10] Yeong-Joon Ju, Gun-Hee Lee, Jung-Ho Hong, and Seong-Whan Lee. 2022. Complete Face Recovery GAN: Unsupervised Joint Face Rotation and De-Occlusion From a Single-View Image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3711–3721.

[11] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[12] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980

[13] Joffrey Leevy, Taghi Khoshgoftaar, Richard Bauder, and Naeem Seliya. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5 (11 2018). https://doi.org/10.1186/s40537-018-0151-6

[14] Vanessa LoBue and Cat Thrasher. 2015. The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults. *Frontiers in Psychology* 5 (2015). https://doi.org/10.3389/fpsyg.2014.01532

[15] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. 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*. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[16] Martin Lukac, Gulnaz Zhambulova, Kamila Abdiyeva, and Michael Lewis. 2023. Study on emotion recognition bias in different regional groups. *Scientific Reports* 13 (05 2023). https://doi.org/10.1038/s41598-023-34932-z

[17] Michael J Lyons, Miyuki Kamachi, and Jiro Gyoba. 2020. Coding Facial Expressions with Gabor Wavelets (IVC Special Issue). (Sept. 2020). https://doi.org/10.5281/zenodo.4029680

[18] A. Mehrabian. 2008. *Communication Theory*. Transaction Publishers; Piscataway, NJ, USA.

[19] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10 (2017), 18–31. https://api.semanticscholar.org/CorpusID:37515850

[20] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[21] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*. 5 pp.–. https://doi.org/10.1109/ICME.2005.1521424

[22] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. 2023. Real-Time Flying Object Detection with YOLOv8. arXiv:2305.09972 [cs.CV]

[23] Rick Sauber-Cole and Taghi M. Khoshgoftaar. 2022. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. *J. Big Data* 9, 1 (2022), 98. https://doi.org/10.1186/S40537-022-00648-6

[24] Liang Xiu-jian and Sun He. 2022. Deep Learning Based Image Forgery Detection Methods. *Journal of Cyber Security* 4 (01 2022), 119–133. https://doi.org/10.32604/jcs.2022.032915