

Do Deep Learning Models Generalize Facial Emotion Recognition in Different Age Groups?

Modelos de Aprendizado Profundo Generalizam o Reconhecimento de Emoções Faciais em Diferentes Faixas Etárias?

Neusa Liberato Evangelista^{1,2,*}, Thiago Lopes Trugillo da Silveira^{2,3}

Abstract: The prevalence of images of adults in facial emotion recognition (FER) databases is an evident bias in emotion classification. Although there are sets of FER images exclusively of children or adults aged 60 and over, the capture scenario is generally controlled and has staged emotions. Many studies in this area focus on increasing task performance by using a single data collection for training and testing; however, due to the diversity of individuals portrayed and photographs collected, this strategy generally fails when applied to new images or in real-world circumstances. In this paper, we assess the generalization potential of deep learning models that have been pre-trained in general-purpose databases such as ImageNet and refined using the databases CK+, DEFSS, FACES (60+), MUG, and NIMH-ChEFS with cross-databases in tests. Through this study, which assesses six different datasets involving three distinct age groups through cross-database evaluation using a clear experimental protocol, we were able to verify that models fine-tuned with data sets composed exclusively of children's images have accuracy above 80%, reaching 88% when used to predict images of children; models fine-tuned with images of young adults are more accurate in classifying young adults than children or adults aged 60 and over, and the dataset of adults aged 60 and over in FACES dataset demonstrated accuracy above 80% in predicting images from other datasets and 96% in predicting images within its own test set.

Keywords: FER — CNN — Pre-trained models fine-tuning — Age bias in FER datasets

Resumo: A prevalência de imagens de adultos em bases de dados de reconhecimento de emoções faciais (FER) é um viés evidente na classificação de emoções. Embora existam conjuntos de imagens FER exclusivamente de crianças ou adultos com 60 anos ou mais, o cenário de captura é geralmente controlado e tem emoções encenadas. Muitos estudos nesta área se concentram em aumentar o desempenho da tarefa usando uma única coleta de dados para treinamento e teste. No entanto, devido à diversidade de indivíduos retratados e fotografias coletadas, essa estratégia geralmente falha quando aplicada a novas imagens ou em circunstâncias do mundo real. Neste artigo, avaliamos o potencial de generalização de modelos de aprendizado profundo que foram pré-treinados em bases de dados de uso geral, como ImageNet, e refinados usando as bases de dados CK+, DEFSS, FACES (60+), MUG e NIMH-ChEFS com testes com dados cruzados. Por meio deste estudo, que avalia seis conjuntos de dados diferentes envolvendo três faixas etárias distintas por meio de avaliação entre bases de dados usando um protocolo experimental claro, pudemos verificar que modelos ajustados com conjuntos de dados compostos exclusivamente por imagens de crianças têm precisão acima de 80%, chegando a 88% quando usados para prever imagens de crianças; modelos ajustados com fotos de jovens adultos são mais precisos na classificação de jovens adultos do que crianças ou adultos com 60 anos ou mais, e o conjunto de dados de adultos com 60 anos ou mais no conjunto de dados FACES demonstrou precisão acima de 80% na previsão de imagens de outros conjuntos de dados e 96% na previsão de imagens dentro de seu próprio conjunto de dados de teste.

Palavras-Chave: FER — CNN — Ajuste fino de modelos pré-treinados — Viés de idade em conjuntos de dados para FER

¹ Serviço Federal de Processamento de Dados (Serpro), Fortaleza, Brasil

² Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Brasil

³ Departamento de Computação Aplicada, Universidade Federal de Santa Maria (UFSM), Brasil

*Neusa Liberato Evangelista: neusale@gmail.com

*Thiago Lopes Trugillo da Silveira: drtitsilveira@gmail.com

DOI: <http://dx.doi.org/10.22456/2175-2745.144479> • Received: 06/12/2024 • Accepted: 09/10/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Emotional recognition through facial expressions has long been a focus of computational vision and machine learning research, with several applications in psychology, neuroscience, and learning and cognition [1, 2]. The rising demand for applications for automatic analysis of human behavior, new technologies for human-computer interaction, and multimedia information recovery has sparked interest in Facial Expression Recognition (FER) through images and videos [3].

The automatic analysis by FER systems comprises three stages [4]: facial image acquisition, facial data extraction and representation, and facial expression recognition. Expression recognition systems use a three-stage training procedure [5]: feature learning, feature selection, and classifier construction. The feature learning stage is responsible for extracting all the features related to facial expression variation. Feature selection chooses the best features to represent the facial expression. At the end of the process, one or more classifiers are used to infer the facial expression, given the selected features.

According to Ramis et al. [6], the prevalence of adult photos in FER databases creates a bias in emotion assessment. Images featuring facial expressions of youngsters and adults aged 60 years or older are not distinguished as such, as in the case of FER2013 dataset [7]. Although there are sets of photos for FER entirely of children or adults aged 60 years or older, such as in LIRIS [8] and “a database for emotional interactions of the elderly” [9], respectively, the capture setting is often controlled and with faked emotions.

FER provides cutting-edge results using Convolutional Neural Networks (CNNs) [10]. Pre-trained CNNs are extensively utilized due to their ease of deployment and excellent performance [11]. Pre-training allows dealing with small datasets in common tasks such as classification [12]. To classify emotions through facial expressions in images, some studies opt for training the entire deep network [13, 14] and others apply transfer learning [15].

Many works in this FER area [16, 17, 18] focus on improving task performance using splits from a single database for training and testing. However, this approach generally loses effectiveness when applied to new images or real context due to the diversity of people portrayed and captured images. It is the so-called domain shift problem [19]. By combining various data sets for training and evaluating FER algorithms, we can better understand the area’s problems and opportunities.

This paper investigates the age bias in the FER datasets available, which predominantly contain adult photos. More precisely, we evaluate the potential for generalization of pre-trained deep machine learning models in general-purpose databases such as ImageNet [20] and refined using the Cohn-Kanade+ (CK+) [21], DEFSS [22], FACES [23], MUG [24], and the NIMH Child Emotional Faces Picture Set (NIMH-ChEFS) [25] databases. The DEFSS set, containing photos of children and young adults classified by age, was separated into *DEFSS Child* and *DEFSS Adult*. Images of adults aged 60 or older were chosen from the FACES collection. The

NIMH-ChEFS dataset includes photos of child actors aged 10–17, with an average age of 13.6 years old. CK+ and MUG have images of young adults only.

We hypothesize that the age range of images used to customize the pre-trained model affects classification task performance on samples of the same age group. Based on the experiments, this study aims to answer three key questions:

- Question 1 “*Are emotions in images of children better classified by refined models with images of children?*”
- Question 2 “*Are emotions of young adults better predicted by models refined with images of adults in this age group?*”
- Question 3 “*Are emotions of adults aged 60 and over better predicted by models refined with images of adults in this age group?*”

We thoroughly analyze different datasets and common backbones (pre-trained and fine-tuned) for the FER task. We performed cross-dataset experiments to help answer Questions 1, 2, and 3; and we were able to evaluate and try to answer the first two. Because we only had one dataset of people aged 60 or over, we could not assess the validity of [Question 3](#).

The rest of this paper is organized as follows. [Section 2](#) reviews related works. The proposed methodology is described throughout [Section 3](#). The experimental setup and results are presented in [Section 4](#). [Section 5](#) draws the conclusions and sheds some light on future work.

2. Related Work

Refining pre-trained models, i.e., fine-tuning, is widely investigated for transfer learning in deep networks and is typically used to compensate for insufficient information for training on the target task [26]. The procedure begins with a pre-trained model on one domain, then refined for a target domain. Fine-tuning a model on a target dataset can significantly improve model performance while reducing annotated data requirements in the target domain [27] instead of training from scratch. Studies such as [13, 14] employ deep network training to categorize emotions using facial expressions in photos, while others such as [15] apply transfer learning.

Utami et al. [13], motivated by the use of FER systems for teacher emotion recognition in the classroom, trained five variations of the EfficientNet [28] exclusively on the FER2013 dataset [7]. To provide variability to the data, transformations such random rotation, horizontal rotation, random noise, and zoom were applied for data augmentation. The EfficientNet network layers were completed with a Global Average Pooling (GAP) layer and a Fully Connected Layer (FCL), activated by softmax. Among the examined alternatives, Utami and colleagues found EfficientNetB3 as the best with an accuracy of 94.44%. The most easily detected facial expression was surprise. The authors claim that EfficientNet can distinguish both

positive and negative emotions. However, they were unable to apply the FER2013 dataset to the educational setting.

Lopes et al. [14] presents a FER solution that combines a simple CNN network with certain pre-processing techniques. Images from the CK+ [21], JAFFE [29], and BU-3DFE [30] datasets were pre-processed for facial normalization (spatial and intensity) before being fed into the convolutional network layers to eliminate the need for controlled conditions and address data shortages. The authors found that combining normalization procedures improved accuracy and yielded competitive results (96.76% on CK+ dataset) compared to other methods using the same facial expression databases and assessment methodology.

Sajjanhar et al. [15] compared a simple CNN model to pre-trained deep networks for facial expression identification. CK+, JAFFE, and FACES datasets were used to retrain the final layer of the VGG16 [31], Inception [32], and VGG-Face [33] networks. The images used were obtained by cutting a region of interest (ROI). The authors concluded that, despite being trained for face recognition, VGG-Faces performed slightly better than the VGG16 and VGG19 networks, which had higher accuracy than Inception V3. Cross-validation was employed as the training, testing, and validation approach.

None of the works cited employed data sets exclusively of children or adults aged 60 years or older, and none of the experiments in those papers used as many models and datasets as those used in this study. To the best of our knowledge, there is no analysis of the impact of age groups in FER using deep learning models.

3. Methodology

This study evaluates pre-trained models for image classification on ImageNet refined for emotion classification through FER, crossing six databases (cross-database tests), i.e., training the model using one set of data and evaluating its accuracy using another set of data.

We decided on the proposed assessment methodology based on key parts of the related works, discussed next. Instead of increasing the number of images available for training with transformations, our study chose to use only augmented images for training, contrasting to [14, 13]. Similarly to Sajjanhar et al. [15], we obtain ROIs from the images during the pre-processing step. Our study refined the final layers of pre-trained models from the DenseNet [34], MobileNet [35], ResNet [36], and VGG [31] architectures, and performed experiments with two configurations for the newly added final layers. Similarly to Utami et al. [13] and Lopes et al. [14], we divided data sets into training, validation, and testing (hold-out [4]), ensuring no data leaking.

The purpose of this study is not to propose a new methodology or architecture but rather to integrate current well-known backbones with a chosen configuration to tackle the FER task. By answering [Question 1](#) and [Question 2](#), we intend to contribute to the FER area and help researchers better select datasets and models for actual FER-based applications.

The following subsections explain the backbones and datasets used and the proposed execution flow for assessment.

3.1 Backbones and additional layers

The basis models were selected from a list of pre-trained models accessible in Keras. The baseline models and complements – final layers added to the pre-trained models – were chosen among those that produced the best results in a preliminary evaluation. In such prior tests, the methods considered the CK+ dataset for training and testing. The models evaluated in this study are: DenseNet121 and DenseNet201 [34]; MobileNetV2 [35]; ResNet50V2, ResNet101V2, and ResNet152V2 [36]; and VGG16 [31].

DenseNet (Dense Convolutional Network) [34] connects each layer to all the others in the feed-forward direction from the input to the output layer. While traditional convolutional networks with L layers have L connections — one between each layer and its subsequent layers — DenseNet has $L(L + 1)/2$ direct connections. Each layer uses the feature maps of all previous layers as inputs and its own feature maps as inputs for all subsequent layers. The aim of DenseNet networks is to alleviate the vanishing gradient problem, strengthen feature propagation, and lead to feature reuse, while substantially reducing the number of parameters. DenseNet networks differ in the number of convolutional layers. In this study, we used DenseNet-121 with 121 layers and DenseNet-201 with 201 layers.

MobileNetV2 [35] introduces a new layer module called the inverted residual block. This module takes a compressed, low-dimensional representation as input, which is first expanded to a high dimension and filtered using a light convolution in depth. The features are then projected back to a low-dimensional representation using linear convolution. This convolutional module is particularly well-suited for embedded projects on mobile devices because it significantly reduces memory consumption by never fully materialising large intermediate tensors.

ResNet (Residual Network) [37] is a network architecture that transforms a simple network into a residual version by inserting shortcut connections, which is a way of dealing with the problem of degradation. ResNetV2 [36] establishes direct paths for information propagation across the entire network, enabling signals to propagate directly from one unit to any other unit in both directions. The number of convolutional layers in ResNetV2 architectures varies, with options including 18, 34, 50, 101 and 152 layers. In this study, we use ResNetV2 50, 101 and 152.

In VGG16 [31], the image passes through a stack of convolutional layers containing 3×3 filters with the smallest possible receptive field (enough to capture the notion of left/right, top/bottom and centre). Spatial grouping is then carried out by five max pooling layers that follow some of the convolutional layers. This is followed by a stack of 16 convolutional layers, then three fully connected layers. The first two of these have

Table 1. Images per emotion in the selected sets

<i>Dataset</i>	<i>Happiness</i>	<i>Fear</i>	<i>Anger</i>	<i>Sadness</i>
CK+	69	45	28	25
DEFSS Child	73	49	44	37
DEFSS Adult	39	22	21	17
FACES 60+	114	114	113	113
MUG	87	47	57	48
NIMH	52	52	52	51

4,096 channels each, while the third has 1,000 channels (one for each ImageNet target class). This third layer is activated by Softmax. All the hidden layers are activated by *ReLU*.

The types of layers appended to the pre-trained models (not concurrently) were:

- Flattening to convert the output of the convolutional layer into a one-dimensional vector - input required for fully connected layers [38].
- GAP (Global Average Pooling) which calculates the average value for each arrangement in the feature map to create a summarised version in a clustered feature map. This capacity added by pooling is called model invariance and helps make the representation almost invariant to small variations in the input [39].
- Batch Normalisation which standardises the inputs and rescales them to solve vanishing/exploding gradients problem and can also act to regularise model [38].
- Random deactivation or Dropout in which, at each stage of training, each neuron (including the input neurons, but always excluding the output neurons) has a probability p of being ignored during this stage and which is also a regularisation method.
- Fully Connected or Dense Layers in which all the neurons in a layer are connected to each neuron in the previous layer, combined with *ReLU* for regularisation, or with *Softmax* for final classification.

3.2 Datasets

Our study uses static images, or images that portray the emotion at its height, and includes the four emotions seen in all sets: happiness, fear, anger, and sadness. We uniformized the number of classes, deleting classes not common to all sets, to maximize classification model performance and reduce problems unrelated to the study's core purpose.

The datasets used for model refinement and cross-dataset testing include CK+ [21], DEFSS [22], FACES [23], MUG [24], and NIMH-ChEFS [25]. The DEFSS set [22], containing photos of children and young adults classified by age, was separated into *DEFSS Child* and *DEFSS Adult*. Only images of adults aged 60 or older were selected from the FACES collection, which we call hereafter FACES 60+. Table 1 lists the number of images per target emotion for each data set.

Many other datasets for FER exist, such as FER2013 [7] and FER+ [40]. Although these datasets are commonly used in training models for facial expression recognition, as in Utami et al. [13], we chose not to use them in our study because they are images obtained through web scrapping, presenting duplicates [6] and adult and children with no separation, making impossible to analyze the impact of age group bias on performance of the models.

3.3 Execution flow

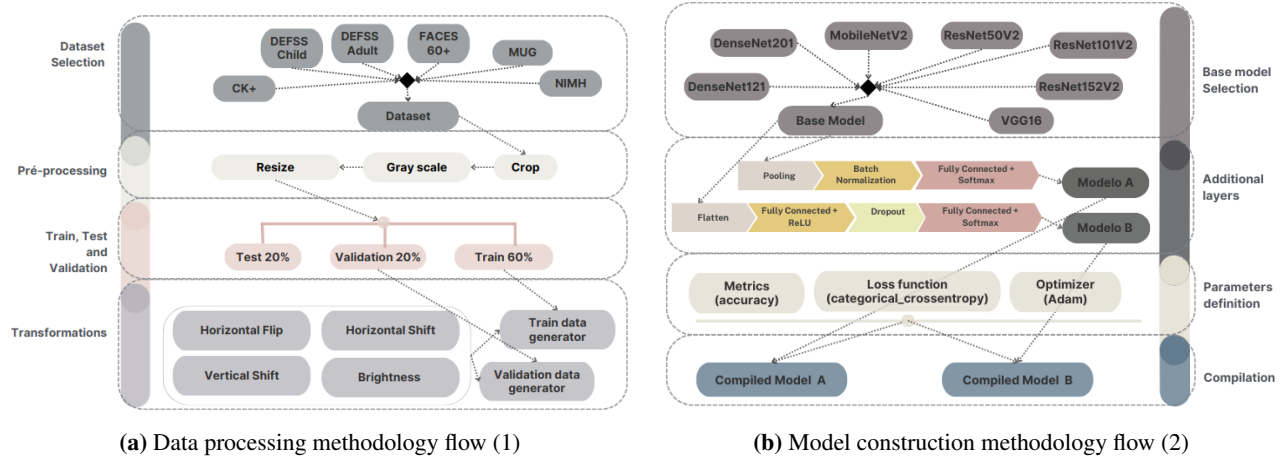
The methodology used to conduct the experiments defines a flow divided into three macro-sections: (1) steps for data preparation and pre-processing; (2) steps to create the models that will be used to classify the images; and (3) steps required to refine the constructed models resulting from flow (2) using the data prepared in flow (1).

Different capture conditions, such as image quality, color intensity, or resolution, depend on the capture process and environment and can affect classification accuracy, particularly in cross-dataset evaluations [6]. For this reason and to ensure consistency between the datasets, some image transformations are performed. For example, we opted to convert all photos to grayscale as made in [6, 14, 15] because CK+ includes both colored and grayscale images.

Fig. 1a shows the steps for processing the obtained data sets. Pre-processing entails cropping the images to highlight the face in the image; afterward converting the colors of each image to grayscale; and finally resizing them to the format expected by the models to be refined – in this work, they receive input images 224 pixels high by 224 pixels wide with three color channels (replicated grayscale). Fig. 2 depicts images in their original form and after the pre-processing steps applied.

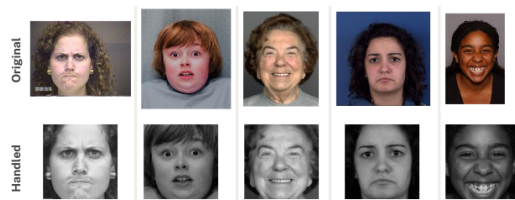
The dataset is then randomly shuffled into training (60%), validation (20%), and testing (20%). The training and validation subsets are fed into the following stage, which employs an image augment generator whose goal is to produce new picture variations at each training epoch. This work used data augmentation techniques such as 20% horizontal and vertical displacement, horizontal mirroring, and brightness variation ranging from 50% to 150%. The test samples are used later in the workflow for prediction by each refined model.

Fig. 1b illustrates building and compiling models. To create the models, at each iteration, one of the base models, that is, one of the pre-trained networks that will serve as the base for the FER model, is chosen. In this work, some final layers were fine-tuned using training data from the target set, while the remaining layers were left frozen. The choice of which and how many layers to adjust was primarily based on refinement examples available in the Keras reference documentation, as well as preliminary and specific experimentation with select models and minimal variations in trainable layers. We adhere to the recommendations of several experiments conducted and published on Kaggle [41], such as [42], which propose augmenting pre-trained models with additional layers to adapt



(a) Data processing methodology flow (1)

(b) Model construction methodology flow (2)

Figure 1. Data processing and model construction flows.

Figure 2. Original and pre-processed images. Each image was taken from a different set.

to the new classification challenge. As a result, new layers are added to the existing layers to tailor the model to its objective of categorizing facial emotions.

A model that includes Clustering, Batch Normalization, and Fully Connected Layers with Softmax activation received the name Model A (model with complement A). Model B is a model created by combining Flattening layers, Fully Connected Layers with ReLU, Dropout, and Fully Connected Layers activated with Softmax. Table 2 shows all the base models and the complements (Comp), highlighting the number of parameters added by the additional layers of the complement, as well as the number of trainable and non-trainable parameters.

The models created during flow (2) are trained using the training and validation data generated by flow (1). This approach begins by defining callbacks to stop training when there is no progress, evaluated on the validation set for a specific number of epochs, and adjust the learning rate to prevent training from getting stuck on a plateau with no improvement in the loss function and accuracy.

The next step is to define the number of epochs (30) and batch size (20) hyperparameters, which were the same in all training sessions. All the models use the Categorical Cross Entropy loss function. We set per-class weights (inversely proportional to the class frequency) to deal with class imbalance. And finally, the parameters specified above are passed to the models for training. At the end of the process, Model A and Model B of each base model are trained with each data set, resulting in 84 trained models (7 base models $\times$ 2

Table 2. Model parameters with the final layer complement.

Model	Comp	Added	Trainable	Frozen
DenseNet121	A	0,01M	0,33M	6,71M
DenseNet121	B	12,85M	13,17M	6,71M
DenseNet201	A	0,01M	0,56M	17,76M
DenseNet201	B	24,08M	24,64M	17,76M
MobileNetV2	A	0,01M	1,04M	1,22M
MobileNetV2	B	16,06M	17,09M	1,22M
ResNet50V2	A	0,01M	5,52M	18,05M
ResNet50V2	B	25,69M	31,20M	18,05M
ResNet101V2	A	0,01M	5,52M	37,11M
ResNet101V2	B	25,69M	31,20M	37,11M
ResNet152V2	A	0,01M	5,52M	52,82M
ResNet152V2	B	25,69M	31,20M	52,81M
VGG16	A	0,01M	4,72M	10,00M
VGG16	B	25,69M	30,41M	9,99M

complement types $\times$ 6 sets of data) for predicting the stored test data.

The final stage is to predict and collect the outcomes from the testing set from each dataset using each trained model, which equals 504 results.

4. Results and Discussion

The experiments were conducted on an Intel Core i5 11445 G7 2.6GHz machine using a 64-bit Ubuntu 22.04 operating system equipped with 16 Gbytes RAM. The Python 3 programming language was used. The libraries TensorFlow 2.13.0 were used to build and train the models and NumPy 1.23 for data manipulation. The scikit-learn 1.3 library was used to split the data into training, test, and validation and calculate the target classes' weights. The implementation is publicly available on GitHub¹.

¹<https://github.com/neusale/facial-emotion-age-generalization.git>

4.1 Impact of Age Groups in FER

Our main goal is to investigate the age bias in the datasets available for FER, which predominantly consist of adult photographs. We hypothesize that the age range of the portrait people used to fine-tune pre-trained models affects classification task. Based on the experiments, our study can address two central questions from Section 1. Given that we only have one dataset of adults aged 60 and over, the Question 3 results in the intuition that a refined model with images from a set accurately identifies images from the same set. Nonetheless, we still perform such experiment.

To answer Question 1, we use images of children and teenagers from two datasets: DEFSS Child and NIMH-ChEFS. Table 3 shows that, in predicting emotions from the DEFSS Child set, the models refined with training images from the DEFSS Child database itself and with training images from the NIMH-ChEFS database lead to the best accuracy (87.80%). By brevity, this and the other tables will show only results with accuracy above 80%.

Table 3. Emotion classification accuracy in the DEFSS Child Test set (above 80%)

<i>Model</i>	<i>Training set</i>	<i>Accuracy (%)</i>
ResNet152V2 A	DEFSS Child	87.80
DenseNet201 B	NIMH-ChEFS	87.80
ResNet50V2 A	DEFSS Child	85.37
VGG16 B	DEFSS Child	85.37
VGG16 B	DEFSS Adult	85.37
ResNet152V2 B	MUG	85.37
DenseNet121 B	MUG	85.37
ResNet152V2 A	NIMH-ChEFS	82.93
VGG16 A	CK+	80.49
ResNet50V2 B	DEFSS Child	80.49
ResNet152V2 B	DEFSS Child	80.49
DenseNet201 B	DEFSS Child	80.49
ResNet152V2 B	DEFSS Adult	80.49
ResNet101V2 B	DEFSS Adult	80.49
DenseNet201 A	FACES 60+	80.49
ResNet101V2 B	MUG	80.49
ResNet152V2 B	NIMH-ChEFS	80.49

Table 4. Emotion classification accuracy in the NIMH-ChEFS Test set (above 80%)

<i>Model</i>	<i>Training set</i>	<i>Accuracy (%)</i>
DenseNet201 B	NIMH-ChEFS	85.71
DenseNet121 B	NIMH-ChEFS	83.33
ResNet152V2 B	DEFSS Child	80.95
DenseNet201 A	FACES 60+	80.95
ResNet50V2 B	NIMH-ChEFS	80.95
MobileNetV2 A	NIMH-ChEFS	80.95

For the NIMH-ChEFS test, models trained with the NIMH-ChEFS demonstrated superior accuracy (85.71%), followed by models trained with FACES 60+ and DEFSS Child, which

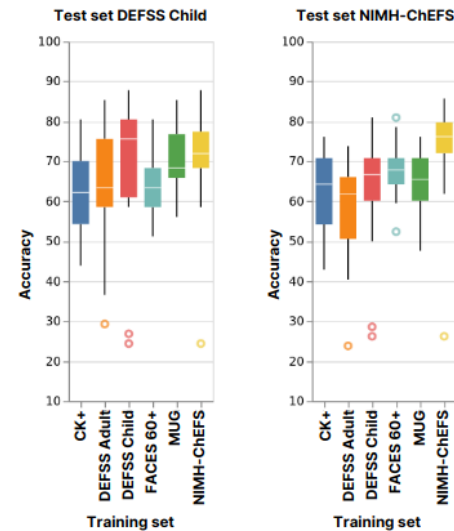


Figure 3. Accuracy distribution of the DEFSS Child and NIMH-ChEFS tests.

reached close results. The distributions of accuracy had low dispersion for the three configurations, as shown in Table 4.

Fig. 3 shows box plots of model accuracy in the DEFSS Child test and the NIMH-ChEFS test per training base, without distinguishing the model employed. Table 5 displays the accuracy statistics for the DEFSS Child test, whereas Table 6 displays the statistics for the NIMH-ChEFS test.

Table 5 shows that, of all the models used to classify emotions from the DEFSS Child set, the highest accuracies (87.80%) and the highest medians (75.61% and 71.95%) were obtained using models refined with the child datasets; and that the models trained with the MUG, DEFSS Adult, FACES 60+ and CK+ adult sets showed lower maximum and median accuracies, albeit above 80%. The statistics displayed in Table 6 show that, in the NIMH-ChEFS test set prediction experiments, the maximum accuracy was obtained by the model refined with the set itself (85.71%), followed by models refined with DEFSS Child and FACES 60+ (80.95%). Models refined with adult images (MUG, CK+, and DEFSS Adult) showed maximum accuracy below 80% when classifying NIMH-ChEFS child images.

Table 5. Statistics on DEFSS Child Test accuracy (%)

<i>Train set</i>	<i>Max</i>	<i>Q3</i>	<i>M</i>	<i>Q1</i>	<i>Min</i>
CK+	80.49	70.12	62.19	54.27	43.90
DEFSS Adult	85.37	75.61	63.41	58.54	29.27
DEFSS Child	87.80	80.49	75.61	60.98	24.39
FACES 60+	80.49	68.29	63.41	58.54	51.22
MUG	85.37	76.83	68.29	65.85	56.10
NIMH-ChEFS	87.80	77.44	71.95	68.29	24.39

After obtaining encouraging results regardless Question 1 – as showed the higher accuracy in classification of emotions in images of children were reached by models refined with images of children –, we can add the following question:

Table 6. Statistics on NIMH-ChEFS Test accuracy (%)

<i>Train set</i>	<i>Max</i>	<i>Q3</i>	<i>M</i>	<i>Q1</i>	<i>Min</i>
CK+	76.19	70.83	64.28	54.16	42.86
DEFSS Adult	73.81	66.07	61.90	50.59	23.81
DEFSS Child	80.95	70.83	66.67	60.11	26.19
FACES 60+	80.95	70.83	67.86	64.29	52.38
MUG	76.19	70.83	65.48	60.11	47.62
NIMH-ChEFS	85.71	79.76	76.19	72.02	26.19

Table 7. Emotion classification accuracy using NIMH-ChEFS for fine-tuning (above 80%)

<i>Model</i>	<i>Test set</i>	<i>Accuracy (%)</i>
DenseNet201 B	DEFSS Child	87.80
DenseNet201 B	NIMH-ChEFS	85.71
VGG16 B	FACES 60+	85.71
DenseNet121 B	NIMH-ChEFS	83.33
ResNet152V2 A	DEFSS Child	82.93
ResNet50V2 B	NIMH-ChEFS	80.95
MobileNetV2 A	NIMH-ChEFS	80.95
ResNet152V2 B	DEFSS Child	80.49

“The models trained with the DEFSS Child and NIMH-ChEFS image sets present the best accuracy in general, regardless of the range age of the test set?”.

To answer this question, we highlight the accuracies above 80% in the models refined with child images. In Table 7, models refined with NIMH-ChEFS, we see that DenseNet201 B obtained 87.80%, ResNet152V2 A obtained 82.93% and ResNet152V2 B obtained 80.49% in predicting emotions from DEFSS Child test images. In this same table, we can also see that DenseNet201 B, DenseNet121 B, ResNet50V2 B, and MobileNetV2 A obtained 85.71%, 83.33%, 80.95%, and 80.95%, respectively, in predicting NIMH-ChEFS test images. The only test set with adult images with an accuracy above 80% is FACES 60+.

Table 8 shows models refined with DEFSS Child with accuracies above 80%. In this table, we see that in the predictions of DEFSS Child itself, the ResNet152V2, ResNet50V2, VGG16, ResNet50V2, ResNet152V2, and DenseNet201 models achieved accuracies of 87.80%, 85.37%, 85.37%, 80.49%, 80.49%, and 80.49%, respectively. The ResNet152V2 model achieved an accuracy of 80.95% when predicting NIMH-ChEFS images, and VGG16 B achieved accuracies of 90% and 83.33% when predicting DEFSS Adult and MUG images, respectively.

Thus, to sum it up and answer the last question, the models refined with images of children do not present high accuracies indistinctly, i.e., the highest accuracies of these models, according to the experiments, are achieved when predicting images of children.

Fig. 4 depicts the confusion matrix and model errors for the two top DEFSS Child predictors. In the confusion matrix of the ResNet152V2 A model refined with DEFSS Child (Fig. 4a), which was the model that achieved the highest accu-

Table 8. Emotion classification accuracy using DEFSS Child for fine-tuning (above 80%)

<i>Model</i>	<i>Test set</i>	<i>Accuracy (%)</i>
VGG16 B	DEFSS Adult	90.00
ResNet152V2 A	DEFSS Child	87.80
ResNet50V2 A	DEFSS Child	85.37
VGG16 B	DEFSS Child	85.37
VGG16 B	MUG	83.33
ResNet152V2 B	NIMH-ChEFS	80.95
ResNet50V2 B	DEFSS Child	80.49
ResNet152V2 B	DEFSS Child	80.49
DenseNet201 B	DEFSS Child	80.49

racy in predicting the DEFSS Child test images, we see that the test set contained images classified as anger (9), fear (10), happiness (15) and sadness (7). All those classified as fear were correctly predicted by the model, although one image of happiness was also classified as fear by the model (false positive). The emotion sadness had the highest proportion of errors (28.5%), where 2 (two) of the 5 (five) images were classified as anger.

The confusion matrix of the DenseNet201 B model re-trained with NIMH-ChEFS images, the model with the second highest accuracy in predicting the DEFSS Child test images, shows the same distribution as the test set - anger (9), fear (10), happiness (15), and sadness (7). As seen in the matrix of Fig. 4b, all the images classified as happiness were correctly predicted, and one fear image was also classified as happiness by the model (false positive). Roughly 33% of the fear images were classified as sadness, i.e., three out of nine images. We hypothesize that some of the characteristics seen in the picture, such as more open or closed eyes and frowning eyebrows, are possible reasons for the classification errors.

Related to Question 2 and the performance of refined models with images of young adults (CK+, MUG and DEFSS Adult), the tested deep learning models perform best when used to predict emotions in their own databases as well as in other adult populations.

DEFSS is one unique dataset which, in this study, was split into two subsets, and share some characteristics that can influence the prediction task, such as the conditions under which the images were captured, the quality of the capture, and any pre-processing that was done, for example.

We compare the performance of models refined with CK+ and MUG when used to classify images of adults and children from the DEFSS set because accuracy results show that CK+ and MUG perform better when categorizing DEFSS Adult images than when classifying DEFSS Child images.

Table 9 shows the statistics when predicting each test set with models refined with young adult's images from each data set. The highest accuracy (94.12%) was obtained by a model refined with CK+ when used to predict the class images from the same database (CK+ test set); in this scenario of training and testing with CK+, 50% of the predictions had

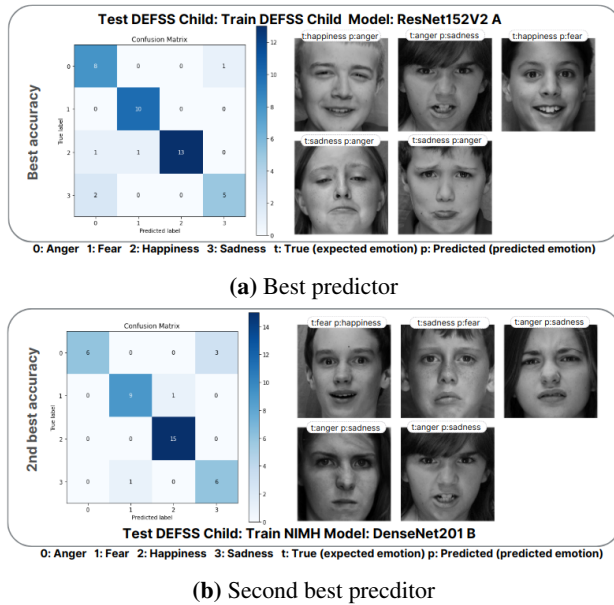


Figure 4. Confusion matrix and errors of the best predictors for the DEFSS Child test set (t of *true* to denote the real class of the image and p of *predict* to denote the class predicted by the model).

an accuracy greater than 80.88%. The second-best accuracy was achieved when testing images from DEFSS Adult (85%), and the lowest results (maximum accuracy of 74.73%, 3rd quartile of 59.74%, and median of 52.75%) were obtained when testing on images from FACES 60+.

The models refined with MUG showed similar results to the models refined with CK+, with 93.75% maximum accuracy and 50% of the accuracies above 85.42% when testing images from the MUG set itself. The second best accuracy was achieved when testing images from DEFSS Adult (90%) and the lowest results (maximum accuracy, 3rd quartile and median, 75.82%, 66.48% and 59.89% respectively) were also those obtained when testing images from FACES 60+.

On the other hand, the models refined with DEFSS Adult did not have as high accuracies as the others, with maximum accuracies of 85.37% in the classification of the set itself and 80% in the classification of DEFSS Child images.

The maximum accuracy achieved when using the CK+ dataset for training and testing is close to that achieved by studies that use CNN such as [43] whose proposed method showed an average accuracy of 96.76%.

Related to **Question 3**, FACES was the only study database that allowed us to separate photos of persons over 60. The models that employed the FACES 60+ set in refinement had very high accuracy in categorizing the FACES 60+ test sample, as well as good performance on the other test sets, as can be seen from the results presented in Table 10.

Although these results might be related to a dataset dependency, we do not discard the influence of aging marks as a significant dataset characteristic for domain shift accuracy drop. More elderly datasets should be considered to confirm

Table 9. Statistics on model accuracy (%) - fine-tuned with CK+, MUG e DEFSS Adult per test set

Test set	Max	Q3	M	Q1	Min
CK+ training set					
CK+	94.12	87.50	80.88	69.12	58.82
DEFSS Adult	85.00	73.75	70.00	65.00	50.00
DEFSS Child	80.49	70.12	62.20	54.27	43.90
FACES 60+	74.73	59.34	52.75	48.90	41.76
MUG	77.08	65.63	53.13	45.83	37.50
NIMH-ChEFS	76.19	70.84	64.29	54.17	42.86
DEFSS Adult training set					
CK+	70.59	69.12	63.24	52.94	17.65
DEFSS Adult	85.00	75.00	70.00	65.00	25.00
DEFSS Child	85.37	75.61	63.41	58.54	29.27
FACES 60+	74.73	59.89	51.65	43.41	29.67
MUG	70.83	61.98	60.42	41.15	22.92
NIMH-ChEFS	73.81	66.08	61.91	50.60	23.81
MUG training set					
CK+	76.47	70.59	67.65	62.50	55.88
DEFSS Adult	90.00	83.75	77.50	71.25	65.00
DEFSS Child	85.37	76.83	68.29	65.85	56.10
FACES 60+	75.82	66.48	59.89	54.13	42.86
MUG	93.75	90.63	85.42	77.08	54.17
NIMH-ChEFS	76.19	70.84	65.48	60.12	47.62

Table 10. Statistics on model accuracy (%) - fine-tuned with FACES 60+ per test set

Test set	Max	Q3	M	Q1	Min
CK+	82.25	67.65	64.71	59.55	52.94
DEFSS Adult	80.00	75.00	70.00	70.00	60.00
DEFSS Child	80.49	68.29	63.41	58.54	51.22
FACES 60+	95.60	92.31	91.21	87.36	64.84
MUG	75.00	70.31	67.71	59.89	35.42
NIMH-ChEFS	80.95	70.83	67.86	64.29	52.38

or refute this hypothesis.

4.2 Other relevant findings

This section discusses other relevant findings that are not central to this research but might help other researchers to set up target datasets, models, and age groups for real applications.

4.2.1 Dataset dependency

The CK+, FACES 60+, NIMH-ChEFS and MUG datasets validated the intuition that the best image predictors are models refined using the dataset itself, as expected.

Table 11 shows the top three accuracies above 80% per adult test set. Observe that models trained and tested with images from the same dataset obtained the highest accuracies in predicting the CK+, FACES 60+, and MUG test sets. In predicting the CK+ test set, one of the models achieved 94.12% accuracy (ResNet50V2 A), three achieved 88.24% (ResNet101V2 B, DenseNet121 B, and VGG16 B), and two other achieved 85.29% (DenseNet201 B and MobileNetV2 B).

VGG16 A model achieved 95.60%, the ResNet50V2 A and ResNet152V2 B models achieved 93.41%, and VGG16 B and ResNet101V2 B achieved 92.31% accuracy when predicting the FACES 60+ test set. DenseNet201 A achieved 93.75%, three other models achieved 91.67% (VGG16 A, ResNet50V2 B, and VGG16 B), and MobileNetV2 A and ResNet101V2 B achieved 87.50% accuracy when predicting MUG images.

DEFSS Adult set was the only one among the adult datasets that did not show this behavior in the tests since models that obtained accuracies of 90%, 85%, and 80% were each refined with a training set from a different base. Three out of the 20 results above 80% relate to models trained with images from DEFSS Adult itself – ResNet101V2 A with an accuracy of 85% and ResNet152V2 B and VGG16 B with 80% accuracy.

Table 12 shows the models achieving the top three accuracies (above 80%) per child test set. For the predictions of the images in the DEFSS Child set, models trained with NIMH-ChEFS, DEFSS Child, MUG, and DEFSS Adult achieved accuracies between 82.93% and 87.80%. The DenseNet201 B model trained with NIMH-ChEFS showed the highest accuracy in predicting NIMH-ChEFS itself (87.80%) and predicting the other set of child images, DEFSS Child (85.71%).

As for the predictions with images from the child sets, for the test images from the NIMH-ChEFS set, the highest accuracies were obtained by models trained with images from the dataset itself (DenseNet201 B with 85.71% and DenseNet121 B with 83.33%), and in third place with an accuracy of 80.95% are models trained with NIMH-ChEFS (MobileNetV2 A and ResNet50V2 B), with FACES 60+ (DenseNet201 A) and with DEFSS Child (ResNet152V2 B).

4.2.2 Overall model performance

Although no model was the best when classifying all test sets (measurements specified in Table 13), examining basic models connected with complements A or B, regardless of the data set used for refinement, shows that the VGG16 Model B presented:

- Maximum accuracy of 92.31%;
- The highest minimum value for accuracy, that is, no test set prediction obtained an accuracy lower than 59.52%;
- The smallest amplitude (32.79%), that is, with the accuracy more concentrated around the median with low dispersion.
- Absence of outliers.

ResNet101V2 Model B achieved similar results as VGG16 Model B, with a maximum accuracy of 92.31%, a median of 70.59%, and a minimum accuracy of 55%.

4.3 Models performance in terms of emotions

To analyze the performance of models in terms of emotions, we chose the models with accuracy over 75% (the pre-trained models' average prediction accuracy on the ImageNet validation dataset [44]). We calculated the F1-Score metric per

Table 11. Top three accuracy above 80% in emotions classification per adult test set

<i>Model</i>	<i>Training set</i>	<i>Accuracy</i>
Test set: CK+		
ResNet50V2 A	CK+	94.12
ResNet101V2 B	CK+	88.24
DenseNet121 B	CK+	88.24
VGG16 B	CK+	88.24
DenseNet201 B	CK+	85.29
MobileNetV2 B	CK+	85.29
Test set: DEFSS Adult		
VGG16 B	DEFSS Child	90.00
ResNet152V2 B	MUG	90.00
VGG16 B	MUG	90.00
ResNet152V2 A	MUG	90.00
ResNet101V2 B	MUG	85.00
VGG16 B	CK+	85.00
ResNet101V2 A	DEFSS Adult	85.00
VGG16 A	CK+	80.00
ResNet152V2 B	DEFSS Adult	80.00
VGG16 B	DEFSS Adult	80.00
ResNet152V2 B	DEFSS Child	80.00
DenseNet201 B	DEFSS Child	80.00
DenseNet201 A	FACES 60+	80.00
DenseNet201 B	FACES 60+	80.00
VGG16 A	MUG	80.00
DenseNet201 B	MUG	80.00
ResNet50V2 B	MUG	80.00
ResNet50V2 A	NIMH-ChEFS	80.00
ResNet152V2 A	NIMH-ChEFS	80.00
ResNet152V2 B	NIMH-ChEFS	80.00
Test set: FACES 60+		
VGG16 A	FACES 60+	95.60
ResNet50V2 A	FACES 60+	93.41
ResNet152V2 B	FACES 60+	93.41
ResNet101V2 B	FACES 60+	92.31
VGG16 B	FACES 60+	92.31
Test set: MUG		
DenseNet201 A	MUG	93.75
VGG16 A	MUG	91.67
ResNet50V2 B	MUG	91.67
VGG16 B	MUG	91.67
MobileNetV2 A	MUG	87.50
ResNet101V2 B	MUG	87.50

emotion and produced the distribution map shown in Figure 5. Table 14 shows the F1-Score statistical indicators for the different emotions.

For happiness, the F1-Score is consistently high, with a maximum of 100% and a minimum of 83.33%, and the median and the 3rd quartile close to 94%, showing that the model is good at classifying happiness in most models. Fear detection oscillates with 75% of the models achieving an F1-Score above 76.92% (1st quartile). Figure 5 shows outliers

Table 12. Top three accuracy above 80% in emotions classification per child test set

Model	Training set	Accuracy
Test set: DEFSS Child		
DenseNet201 B	NIMH-ChEFS	87.80
ResNet152V2 A	DEFSS Child	87.80
DenseNet121 B	MUG	85.37
ResNet152V2 B	MUG	85.37
VGG16 B	DEFSS Adult	85.37
VGG16 B	DEFSS Child	85.37
ResNet50V2 A	DEFSS Child	85.37
ResNet152V2 A	NIMH-ChEFS	82.93
Test set: NIMH-ChEFS		
DenseNet201 B	NIMH-ChEFS	85.71
DenseNet121 B	NIMH-ChEFS	83.33
MobileNetV2 A	NIMH-ChEFS	80.95
ResNet50V2 B	NIMH-ChEFS	80.95
DenseNet201 A	FACES 60+	80.95
ResNet152V2 B	DEFSS Child	80.95

Table 13. Statistics on model accuracy (%) - all tests

Model	Max	Q3	M	Q1	Min
DenseNet121 A	77.08	64.40	58.39	43.86	23.81
DenseNet121 B	91.21	74.11	64.92	59.21	35.42
DenseNet201 A	93.75	73.29	64.71	54.67	17.65
DenseNet201 B	87.80	79.56	72.85	68.98	51.65
MobileNetV2 A	89.01	62.50	56.51	49.59	31.87
MobileNetV2 B	90.11	70.00	56.62	45.64	25.00
ResNet101V2 A	91.21	74.11	67.97	62.12	42.86
ResNet101V2 B	92.31	76.65	70.59	63.84	55.00
ResNet152V2 A	91.21	70.00	61.83	49.59	17.65
ResNet152V2 B	93.41	80.00	73.23	64.93	37.50
ResNet50V2 A	94.12	72.03	64.50	61.57	31.25
ResNet50V2 B	91.67	73.26	66.26	61.43	50.55
VGG16 A	95.60	75.15	66.26	52.31	14.71
VGG16 B	92.31	83.75	75.00	67.86	59.52

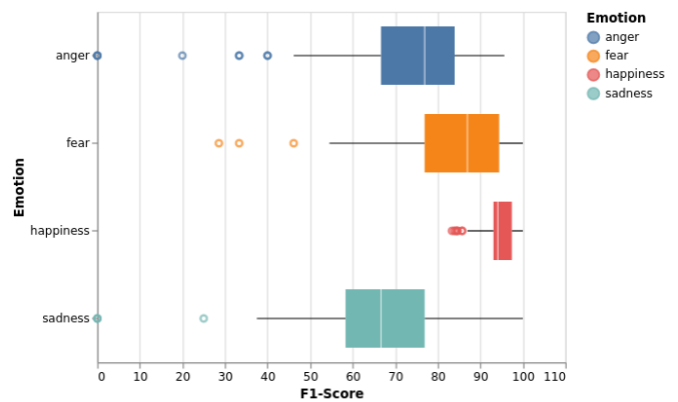
with F1-Scores between 28.57% and 50%, indicating difficulty in predicting this emotion by some models.

The models had a higher amplitude in the categorization of anger and sadness, 96.65% and 100%, respectively, since some models, despite having an accuracy of more than 75%, missed all predictions for one of these two emotions. Table 15 shows models with a minimum F1-Score of zero. To illustrate this situation, Figure 6 shows the confusion matrix of the ResNet101V2 A model refined with MUG and used to predict CK+, displaying hits and misses for each emotion class, and Figure 7 shows the images improperly categorized by this model. In this case, sadness was labeled as fear or anger, whereas fear was classified as happiness.

Happiness was the emotion most discriminable by the pre-trained and refined models used in cross-testing.

4.4 Discussion on ethical concerns

Ethical considerations must be taken into account when dealing with facial datasets. Researchers should be aware that

Figure 5. F1-Score of models per emotion**Table 14.** F1-Score per emotions

Emotion	Max	Q3	M	Q1	Min
happiness	100.00	97.40	94.12	93.09	83.33
fear	100.00	94.43	86.96	76.92	28.57
anger	96.65	83.96	76.92	66.67	0.00
sadness	100.00	76.92	66.67	58.33	0.00

Table 15. Models with Zero F1-Score

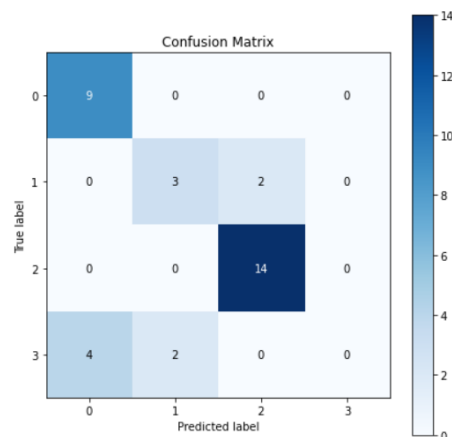
Model	Train	Test
Anger		
VGG16 B	DEFSS Adult	DEFSS Adult
DenseNet201 B	FACES 60+	DEFSS Adult
DenseNet201 A	FACES 60+	DEFSS Adult
Sadness		
ResNet101V2 A	MUG	CK+
ResNet101V2 A	DEFSS Adulto	DEFSS Adult
ResNet101V2 A	MUG	NIMH-ChEFS

many such datasets contain images of individuals who have not explicitly consented to using their images, raising privacy concerns that must be addressed [45]. The datasets used in this study are all publicly available resources intended to enrich FER research. However, many large-scale datasets contain pictures that have been scraped from the internet or captured in public spaces, and those portrayed may not even be aware of this.

Datasets of images portraying people often reflect societal biases, which can be inadvertently incorporated into the models trained on them. It is crucial to promote fairness and equality when identifying and addressing bias in computer vision models. This involves examining datasets for potential biases, implementing strategies to address identified issues, and evaluating model fairness before deployment [46].

McStay [47] coined the term *Emotional Artificial Intelligence (EAI)* to draw attention to the growing number of systems that track, infer, and measure signals of human emotional expression. Andrew McStay and Pamila Pavliscak [48] have developed guidelines for the ethical use of EAI tech-

Figure 6. Confusion Matrix of ResNet101V2 A model trained with MUG used to predict CK+ test images



nologies, designed to enable companies to innovate ethically and legally regarding emotion and AI. The three main reflections that the authors encourage researchers to consider are that (i) the physical display of emotion is only one aspect of emotion; (ii) past expressions do not always predict future feelings; and, finally, about our study, (iii) we should consider the stereotypes and assumptions about emotions that can have a significant impact on individuals or groups.

5. Conclusion

This study focuses on transfer learning by modifying pre-trained deep networks to assess the age generalization of those particular pre-trained models for FER. The models were refined by training their final layers with the CK+, DEFSS, FACES 60+, MUG, and NIMH-ChEFS databases. The performance evaluation also used database crossover in the tests. The pre-processing of images was limited to resizing for the pre-trained model's resolution, converting to grayscale, and cropping to generate the ROI image. The selected pre-models, DenseNet121 and DenseNet201; MobileNet V2; ResNet50 V2, ResNet101 V2 and ResNet152 V2; and VGG16, were extended with a few additional layers to be adjusted for the task of emotion classification by recognizing facial expressions in images from the datasets above.

Refined models using only images of children (DEFSS Child and NIMH-ChEFS) achieved accuracy above 80%, reaching 88%, when used to predict images of children, including diverse sets. Refined models incorporating photographs of young adults (CK+, DEFSS Adult, and MUG) are more accurate at classifying young adults than images of children or adults over 60. The FACES 60+ dataset, wherein only photographs of persons aged 60 years and above were picked, showed generally good performance in the predictions of all sets (above 80%) and notably high accuracy (96%) in predicting tests from the same data collection. Happiness was the most discriminative emotion in the pre-trained and refined

Figure 7. F1-score of ResNet101V2 A model trained with MUG used to predict CK+ test images



model cross-dataset tests.

Although it is not possible to point out one of the models as the overall best performing, the model VGG16 Model B, in the analysis of the distribution of accuracy obtained across different tests, was the one in which the highest minimum value for accuracy was found (no prediction obtained an accuracy lower than 60%), maximum accuracy of 92% (the highest accuracy obtained among all was 96%), the lowest range (32%), and accuracy is clustered around the median with low dispersion and no outliers.

Future work will include tests with all photos available in the FACES dataset, including young adults and adults aged 60 and up, as FACES 60+ has proven to be a promising set for FER. Furthermore, we can combine the bases used here to refine the models with the best-observed performance and evaluate them using broader bases such as FER+, enabling statistical significance tests. Another line of research involves using class activation maps (CAMs) to investigate which regions in the input contributed to the decision of the predicted class and whether these regions differ for models refined with images of different age groups.

Acknowledgements

We express our gratitude to Serviço Federal de Processamento de Dados, Brazil, for partially funding this research.

Author contributions

Neusa Liberato contributed to the development of the code and the writing of the article. Thiago contributed to the implementation adjustments and text refinement.

References

- [1] BHATTI, Y. et al. Facial expression recognition of instructor using deep features and extreme learning machine. *Computational Intelligence and Neuroscience*, v. 2021, p. 1–17, 2021.
- [2] MIT MEDIA LABORATORY. *Multimodal Affect Recognition in Learning Environments*. 677–682 p.
- [3] EDWARDS, C. et al. Human-machine communication: What does/could communication science contribute to hri? In: *ACM/IEEE International Conference on Human-Robot Interaction*. [S.l.: s.n.], 2019. p. 673–674.
- [4] SENIOR, A. W.; PANKANTI, S. *Handbook of Face Recognition*. 2. ed. London: Springer London, 2011. 699 p.
- [5] LIU, P. et al. Facial expression recognition via a boosted deep belief network. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1805–1812.
- [6] RAMIS, S. et al. A novel approach to cross dataset studies in facial expression recognition. *Multimedia Tools and Applications*, v. 81, p. 39507–39544, 2022.
- [7] GOODFELLOW, I. J. *Papers with Code - FER2013 Dataset*. 2013.
- [8] KHAN, R. A. et al. A novel database of childrens spontaneous facial expressions (LIRIS-CSE). *Image and Vision Computing - Elsevier*, v. 83–84, p. 61–69, 2019.
- [9] WANG, K. et al. A database for emotional interactions of the elderly. In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. [S.l.: s.n.], 2016. p. 1–6.
- [10] CANAL, F. Z. et al. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, v. 582, p. 593–617, 2022.
- [11] LOPES, A. T.; AGUIAR, E. de; OLIVEIRA-SANTOS, T. A facial expression recognition system using convolutional networks. In: *Conference on Graphics, Patterns and Images*. [S.l.: s.n.], 2015. p. 273–280.
- [12] TAMAYO-MONSALVE, M. A. et al. Coffee maturity classification using convolutional neural networks and transfer learning. *IEEE Access*, v. 10, p. 42971–42982, 2022.
- [13] UTAMI, P.; HARTANTO, R.; SOESANTI, I. The efficientnet performance for facial expressions recognition. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2022. p. 756–762. ISBN 9781665455121.
- [14] LOPES, A. T. et al. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, v. 61, p. 610–628, 2017.
- [15] SAJJANHAR, A.; WU, Z.; WEN, Q. Deep learning models for facial expression recognition. *Digital Image Computing: Techniques and Applications*, p. 1–6, 2018.
- [16] KHAIREDDIN, Y.; CHEN, Z. *Facial Emotion Recognition: State of the Art Performance on FER2013*. 2021.
- [17] MOLLAHOSSEINI, A.; CHAN, D.; MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. *IEEE Winter Conference on Applications of Computer Vision*, p. 1–10, 2016.
- [18] AGRAWAL, A.; MITTAL, N. Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *The Visual Computer*, v. 36, p. 405–412, 2020.
- [19] WILSON, G.; COOK, D. J. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, v. 11, 2020.
- [20] RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, v. 115, p. 211–252, 2015.
- [21] LUCEY, P. et al. 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*. [S.l.: s.n.], 2010. p. 94–101.
- [22] MEUWISSEN, A. S.; ANDERSON, J. E.; ZELAZO, P. D. The creation and validation of the developmental emotional faces stimulus set. *Behav Res Methods*, v. 49, p. 960–966, 2017.
- [23] EBNER, N. C.; RIEDIGER, M.; LINDENBERGER, U. Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, v. 42, p. 351–362, 2010.
- [24] AIFANTI, N.; PAPACHRISTOU, C.; DELOPOULOS, A. The mug facial expression database. In: *International Workshop on Image Analysis for Multimedia Interactive Services*. [S.l.: s.n.], 2010. p. 1–4.
- [25] EGGER, H. et al. The nimh child emotional faces picture set (nimh-chefs): a new set of children’s facial emotion stimuli. *International journal of methods in psychiatric research*, v. 20, p. 145–56, 2011.
- [26] GUO, Y. et al. Spottune: Transfer learning through adaptive fine-tuning. In: . [S.l.: s.n.], 2018.

- [27] NGUYEN, D. et al. Meta-transfer learning for emotion recognition. *Neural Computing and Applications*, v. 35, p. 10535–10549, 2023.
- [28] TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In: . [S.l.: s.n.], 2019.
- [29] LYONS, M. J.; KAMACHI, M.; GYOBA, J. *Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)*. [S.l.: Zenodo, 2020.
- [30] YIN, L. et al. A 3d facial expression database for facial behavior research. In: *International Conference on Automatic Face and Gesture Recognition*. [S.l.: s.n.], 2006. p. 211–216.
- [31] SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014.
- [32] SZEGEDY, C. et al. *Rethinking the Inception Architecture for Computer Vision*. 2015.
- [33] CHEN, H.; HAOYU, C. Face recognition algorithm based on vgg network model and svm. *Journal of Physics: Conference Series*, v. 1229, p. 012015, 2019.
- [34] HUANG, G. et al. Densely connected convolutional networks. In: . [S.l.: s.n.], 2016.
- [35] SANDLER, M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: . [S.l.: s.n.], 2018.
- [36] HE, K. et al. Identity mappings in deep residual networks. In: . [S.l.: s.n.], 2016.
- [37] HE, K. et al. *Deep Residual Learning for Image Recognition*. 2015.
- [38] GERON, A. *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017. ISBN 978-1491962299.
- [39] BROWNLEE, J. *A Gentle Introduction to Pooling Layers for Convolutional Neural Networks*. 2019. (<https://machinelearningmastery.com/pooling-layers-for-convolutional-neural-networks/>).
- [40] BARSOUM, E. et al. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: . [S.l.: s.n.], 2016.
- [41] KAGGLE. 2024. Disponível em: (<https://www.kaggle.com/>).
- [42] NELSON, S. G. *Emotion Recognition using Facial Expressions*. 2023. Disponível em: (<https://www.kaggle.com/code/sandragracenelson/emotion-recognition-using-facial-expressions>).
- [43] KHRYASHCHEV, V.; IVANOVSKY, L.; PRIOROV, A. Deep learning for real-time robust facial expression analysis. In: *Proceedings of the International Conference on Machine Vision and Applications*. Association for Computing Machinery, 2018. p. 66–70. Disponível em: (<https://doi.org/10.1145/3220511.3220518>).
- [44] KERAS APPLICATIONS. *Keras Applications*. 2023. (<https://keras.io/api/applications/>).
- [45] BIRHANE, A.; PRABHU, V. U. Large image datasets: A pyrrhic win for computer vision? . In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2021. p. 1536–1546. Disponível em: (<https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00158>).
- [46] BELLAMY, R. et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, PP, p. 1–1, 09 2019.
- [47] MCSTAY, A. *Emotional AI: The Rise of Empathic Media*. [S.l.: s.n.], 2018. ISBN 9781473971103.
- [48] MCSTAY, A.; PAVLISCAK, P. *Emotional Artificial Intelligence: Guidelines For Ethical Use*. 2019. (https://drive.google.com/file/d/1frAGcvCY_v25V8ylqgPF2brTK9UVj_5Z/view).