A Data-Centric Approach for Detecting a Neutral Facial Expression using Deep Learning

Lúcia Sousa *, Daniel Canedo *, Miguel Drummond *, João Ferreira *, António Neves *

Posted Date: 23 October 2023

doi: 10.20944/preprints202310.1433.v1

Keywords: deep learning; emotion classification; machine learning; neutral emotion classification

Article

A Data-Centric Approach for Detecting a Neutral Facial Expression using Deep Learning

Lúcia Sousa [1,*], Daniel Canedo [1,*], Miguel Drummond [2,*], João Ferreira [3,*] and António Neves [1,*]

[1] IEETA/DETI, University of Aveiro, 810-193 Aveiro, Portugal
[2] Instituto de Telecomunicações (IT), University of Aveiro, 810-193 Aveiro, Portugal
[3] Vision-Box, Rua Casal do Canas 2, 2790-204 Carnaxide, Portugal
* Correspondence: luciamsousa00@ua.pt, danielduartecanedo@ua.pt, mvd@av.it.pt, joao.ferreira@vision-box.pt, an@ua.pt

Abstract: Neutral facial expression recognition is of great importance in various domains and applications. This study introduces a data-centric approach for neutral facial expression recognition, presenting a comprehensive study that explores different methodologies, techniques, and challenges in the field to foster a deeper understanding. The results show that data augmentation plays a crucial role in improving dataset performance. Additionally, the study investigates different model architectures and training techniques to identify the most effective approach, with the InceptionV3 model achieving the highest accuracy of 72%. Furthermore, the research examines the influence of preprocessing methods on the performance of both InceptionV3 and a simplified CNN model. Interestingly, the results indicate that preprocessing techniques positively affect the performance of the simpler CNN model but negatively impact the InceptionV3 model. The implemented system, used to evaluate the findings, demonstrates promising results, correctly classifying 77% of neutral expressions. However, there are still areas for improvement. Creating a specialized dataset that includes both neutral and non-neutral expressions would greatly enhance the accuracy of the system. By addressing limitations and implementing suggested improvements, neutral facial expression recognition can be significantly enhanced, leading to more effective and accurate results.

Keywords: deep learning; emotion classification; machine learning; neutral emotion classification

1. Introduction

The neutral facial expression, also called the poker face, is a facial expression in which there is little or no visible emotion being expressed [1]. A relaxed or impassive appearance, with no detectable changes in the facial muscles or features such as the eyebrows, eyes, mouth, or cheeks, characterizes it. Whereas facial expressions are the opposite, they indicate a demonstration of emotion and, in turn, the contraction of muscles, as shown in Figure 1.

(a) Neutral **(b)** Not Neutral

Figure 1. Example of neutral and not neutral expressions from CK+ dataset [2].

Facial Expression Recognition (FER) is the process of identifying and interpreting human emotions from facial cues [3]. It involves analyzing the movements and positions of facial features such as the eyes, eyebrows, mouth, and cheeks to infer emotions such as happiness, sadness, anger, fear, surprise, and disgust.

 doi:10.20944/preprints202310.1433.v1

FER can be performed autonomously using computer algorithms and machine learning techniques. It involves image capture, preprocessing, feature extraction, classification, and result interpretation. By analyzing key facial features, machine learning algorithms accurately classify expressions, providing a label or probability score as the output [4].

There are several reasons why the ability to detect neutral facial expressions can be important. In nonverbal communication, neutral expressions can serve as a reference point against which other expressions are compared, and accurately detecting them can provide insights into a person's emotional state or intentions [5]. In psychology research, neutral expressions can be used as a control for the influence of facial expressions on the results of experiments [6]. In marketing, detecting neutral expressions can help to understand consumer reactions to different products or advertisements [7]. In the field of human-computer interaction, detecting neutral expressions can be useful for developing more natural and efficient ways for people to communicate with computers. In security contexts, such as border control or law enforcement, accurately detecting neutral expressions can be important for identifying individuals who may be attempting to conceal their emotions or intentions [8].

However, recognizing neutral facial expressions comes with challenges and limitations. Neutral expressions can be difficult to distinguish from other expressions, such as slight smiles or frowns, which can affect the accuracy of detection [9]. Neutral expressions can vary depending on the individual, culture, and context, which can make it challenging to develop a universal standard for detecting them [9]. The context in which a neutral expression is displayed can affect its interpretation and detection. For example, a neutral expression in a social context may be different from a neutral expression in a security context [10]. Individual differences in perception, attention, and cognitive processing can affect the accuracy of detecting neutral expressions [11,12]. The accuracy of detecting neutral expressions can be affected by the technology used, such as the quality of cameras or algorithms used for detection [9]. Despite these challenges and limitations, researchers have developed various methods and databases for detecting neutral expressions, such as face alignment methods and diffusion model analyses [9,12,13]. These methods can help improve the accuracy of detecting neutral expressions in different contexts.

The aim of this study is to develop an advanced facial expression detection and recognition system with a focus on neutrality, aiming to significantly expedite the check-in procedures at airports. To improve efficiency at airports by automating the process of enrolling and identifying individuals. A neutral facial expression recognition system has the potential to enhance airport efficiency by streamlining the check-in process. The system minimizes or eliminates the need for manual verification of identity documents through automated identity verification based on neutral facial expressions. This automation accelerates the check-in process, reduces the necessity for manual intervention, and improves overall accuracy. Passengers have their neutral expressions captured by a camera compared to their pre-registered data, resulting in a quicker and more efficient check-in experience. This system reduces human error, improves identification accuracy, and helps to expedite the overall check-in process.

The present research not only contributes to the field of facial expression detection but also addresses several technical gaps associated with the current problem. These gaps represent areas where advancements and improvements are needed to achieve more accurate and applicable facial expression detection using machine learning. The contributions of this research are the following:

- Extension of existing literature by providing insights into areas for enhancement in facial expression detection through machine learning techniques.
- Despite achieving results that were not as impressive as those of previous studies, this research sheds light on the importance of increasing real-world representativeness in training models.
- By exploring the limitations of training models on datasets that are not person independent, this research highlights the drawbacks and potential biases that can arise when models struggle to generalize to unseen individuals. This contributes to a better understanding of the importance of person-independent datasets in facial expression detection models.

 doi:10.20944/preprints202310.1433.v1

- Identifying imbalanced datasets performing well in classifying non-neutral expressions highlights the need for alternative methods for detecting neutral expressions. This research identifies this as crucial to improve accuracy and prevent biases in real-world applications.
- Also, this research identifies several promising areas for future studies in neutral facial expression detection. These include developing more precise and reliable machine learning models, collecting and annotating more extensive and diverse datasets focusing on neutral expressions and investigating the impact of various factors on neutral expression identification.
- The research emphasizes the need for dedicated datasets that explicitly classify neutral and non-neutral expressions, as existing datasets primarily focus on identifying various emotions without specific categorization.

This paper is divided into the following sections: Section 2 provides a review of the existing literature on the topic of neutral facial expression recognition; Section 3 presents the study conducted on multiple datasets, along with the applied data augmentation techniques; Section 4 presents the conducted study across several models, highlighting their impact; Section 5 provides an analysis of the impact of preprocessing techniques; Section 6 introduces the results of a real-time system implemented to evaluate the effectiveness of the developed work; Section 7 presents the discussion of the results, summarizes the main findings and contributions of the study; Section 8 presents the conclusion and future work.

2. Literature Review

2.1. Facial Expression Detection

The detection of facial expressions can be accomplished using various approaches, with two of these methods being the detection of action units (AUs) and the detection of facial landmarks.

Detection of facial expressions through action units involves identifying specific muscle movements or combinations of muscle movements that correspond to specific emotions or expressions. This method is based on the Facial Action Coding System (FACS), which defines specific AU that correspond to different facial movements [14].

On the other hand, the detection of facial expressions through landmarks involves identifying specific points on the face, such as the corners of the mouth or the inner brow, and tracking their movements over time to infer the expression being made. This method is based on the idea that different facial expressions are characterized by specific patterns of movement in facial landmarks [15].

A typical [3] system comprises several key steps, including image acquisition, pre-processing, feature extraction, and classification or regression [16].

FER researchers use various datasets to build systems that can produce results that can be compared to related works. Most datasets are built on 2D static images or 2D video sequences, but some contain 3D images. 2D systems have limitations in handling different poses, as most 2D datasets only contain frontal faces. 3D systems have the potential to handle the pose variation problem [16].

Most FER datasets are labelled with the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and the neutral expression. Some datasets are built in controlled environments with controlled lighting conditions, while others are built in uncontrolled or wild environments. Some datasets have subjects pose certain emotions towards a reference, while others try to stimulate spontaneous and genuine facial expressions [16].

Pre-processing is a crucial step in FER and other computer vision systems. It involves cleaning raw data and extracting valuable information. Popular approaches include face detection, geometric modifications, and data normalization.

Once the pre-processing phase is completed, relevant features can be extracted. In a conventional FER system, these features are facial features. The quality of these features plays a significant role in determining the system's accuracy. Therefore, several techniques have been developed in computer vision to extract these features. The most popular feature extraction techniques used in reviewed works

 doi:10.20944/preprints202310.1433.v1

include Local Binary Pattern (LBP), Optical Flow (OF), Active Appearance Model (AAM), Action Unite (AU), and Facial Animation Parameter (FAP) and Gabor Filter (GF) [16].

A classification model predicts a specific label given an input image or attributes, whereas a regression model determines the connection between a dependent variable and independent variables. Both strategies were utilized in the publications studied, with classification being the most common.

The most commonly used classification and regression algorithms include Convolutional Neural Network (CNN), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF) and Decision Tree (DT). Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the specific task and dataset [16].

2.2. Neutral Facial Expression Detection

Three significant studies conducted in recent years have focused on detecting neutral facial expressions. Each study presents unique approaches and utilizes diverse datasets to address the challenges in this domain. The first study, "Automatic Neutral Face Detection Using Location and Shape Features", addresses the challenge of automatically detecting neutral faces. The researchers used the CK+ (Cohn-Kanade) [2] and Yale Face datasets. In the second study, "Neutral Face Classification Using Personalized Appearance Models for Fast and Robust Emotion Detection", the datasets used were CK+ and ISL ((Indian Spontaneous Expression) [17]. In addition, the authors have developed two novel datasets, SRID1 and SRID2 [18]. In the third study, "Deep Peak-Neutral Difference Feature for Facial Expression Recognition", the researchers propose a feature called the deep peak-neutral difference and utilize three datasets: CK+, BU4DFE [19], and the newly created CCNUFE dataset [20]. By analyzing these three studies, valuable insights are gained into the recent advancements in neutral facial expression detection.

The first study on automatic neutral face detection using location and shape features was conducted in 2001 by Ying-li Tian and Ruud M. Bolle [21]. The authors aimed to develop a system that can identify neutral expressions on a person's face by analyzing specific features' locations and shapes.

The proposed system consists of three main steps. First, the authors employed the Senior method for face detection, which involves using filters and machine learning algorithms to detect faces in images. This method applies a skin-tone filter to classify pixels as skin-tone or non-skin-tone and rejects regions without a sufficient percentage of skin-tone pixels. It uses a linear discriminant and a Distance From Face Space measure to filter candidates and rank overlapping candidates, ultimately selecting the highest-scoring candidate. The Senior method has been proven effective in detecting faces under various lighting conditions and poses.

After capturing a face image, the authors proceeded to locate specific features or landmarks on the face using a Fisher discriminant and distance from feature space measurements. Six stable and discriminative landmarks were selected, including the centre of the pupils, inner endpoints of the eyebrows, and corners of the mouth.

These landmarks were used to calculate distances between them, which served as location features.

Additionally, the authors considered the opening of the eyes by calculating the distance between the upper and lower eyelids. This was achieved by applying a binary image and fitting an ellipse around the eyes.

To extract the shape of the mouth, an edge detector was applied to the image, and the face was divided into a 3x3 grid. Histograms were calculated for the three squares in the mouth region, providing a shape description for each zone. The mouth shape was represented as a feature vector consisting of 12 components, comprising three histograms, each with four components.

To detect a neutral face, the authors utilized a three-layer neural network with one hidden layer. Nineteen features, including seven location features and twelve shape histogram features of the mouth, were used as input to classify the face as neutral or non-neutral. Through experiments, the authors determined that using four hidden units resulted in the best performance.

 doi:10.20944/preprints202310.1433.v1

The authors evaluated their system using two datasets. The first dataset was the CK+ dataset, divided into a training set and a testing set. The system achieved an average detection rate of 97.2% on this dataset. The second dataset was the Yale Face dataset, and the system achieved an average detection rate of 95.3%. However, the system occasionally made errors in recognizing neutral faces as non-neutral faces, primarily due to difficulties in calculating shape features accurately for individuals with facial hair. To address this, the authors suggested incorporating a feature for detecting facial hair in future work.

In the second study by Pojala Chiranjeevi, Viswanath Gopalakrishnan, and Pratibha Moogi, a personalized appearance model was developed for fast and robust emotion detection, specifically focusing on the classification of neutral facial expressions in real-time video footage. The authors aimed to reduce the computational time of neutral versus emotion classification by using tailored appearance models.

The research introduced a preprocessing algorithm to determine if a frame is neutral or not. The personalized model learned the neutral appearance of each user by using a set of reference neutral frames. To construct the model, users were prompted to display various emotions before generating the reference frames, assuming that they would start the application with a neutral expression and maintain it for a significant portion of the time.

The algorithm involved several key steps, including Procrustes analysis, selection and tracking of key emotion points, patch representation, generation of affine noise-based statistical models, multi-neighbour comparison based on structural similarity, a fusion of distances of key emotion point pairs, and using textural change inferences to improve emotion classification accuracy.

Procrustes analysis was used to align facial features and extract relevant shape information for analysis. The Constrained Local Model (CLM) approach was utilized to track key feature points on the face, compensating for affine variations. Additional key emotion points were generated using offsets from the CLM points and accurately tracked over time.

The effectiveness of the proposed algorithm was validated using datasets such as CK+, ISL, and two collected datasets, SRID1 and SRID2. The comparison with a previous study showed that Tian and Bolle's approach outperformed the proposed approach in terms of True Positive Rate (TPR), but the textural change inferences improved the f-measure for each emotion, particularly on the SRID1 and SRID2 datasets.

The study demonstrated that the proposed approach consumed less time compared to other studies. However, it has limitations, such as not handling talking faces and potential difficulties with large and abrupt pose variations.

The third study by Jingying Chen, Ruyi Xu, and Leyuan Liu proposes the use of a Deep Peak-Neutral Difference (DPND) feature for facial expression recognition. The authors aim to address the overfitting problem in facial expression recognition by utilizing handcrafted features and fine-tuning a pre-trained neural network. They also propose a semi-supervised method for selecting key neutral and peak frames without manual labelling.

The DPND feature extraction process involves three steps.

Firstly, deep representations are extracted from the fully connected layer of a VGG-16 network fine-tuned on the BU4DFE and CK+ datasets. The authors find that the FC6 layer provides the best performance for facial expression classification. Secondly, peak and neutral frames are detected and classified using a clustering algorithm (K-means++) and a classification algorithm. Finally, the DPND feature is calculated as the difference between the deep representation of the key peak frame and the key neutral frame. This feature is then used as input for a multiclass Support Vector Machine (SVM) classifier to recognize facial expressions.

To evaluate their approach, the authors used the CK+, BU4DFE, and CCNUFE datasets. They trained the VGG-16 network on the VGG face dataset and fine-tuned it on a subset of the BU4DFE dataset. The authors achieved high accuracy in classifying neutral and peak frames using their

 doi:10.20944/preprints202310.1433.v1

semi-supervised method. The DPND feature outperformed the DPR alone, showing improved accuracy on the CK+ and BU4DFE datasets.

The proposed system offers a novel approach to accurately detect neutral facial expressions using personalized appearance models. Integrating deep learning and handcrafted features, along with the semi-supervised method for frame classification, contributes to improved performance in facial expression recognition. The DPND feature extraction enhances the discrimination of differences between neutral and peak expressions. The comparative evaluations demonstrate the superiority of the DPND method over existing convolutional neural network models on publicly available datasets.

The present research extends the existing literature by providing insights into areas that can be enhanced in facial expression detection using machine learning. Despite achieving results that were not as impressive as those of previous studies, this research sheds light on the importance of increasing real-world representativeness in training models. By exploring the limitations of training models on datasets that are not person-independent, this research highlights the drawbacks and potential biases that can arise when models struggle to generalize to unseen individuals. This contributes to a better understanding of the importance of person-independent datasets in facial expression detection models. Identifying imbalanced datasets performing well in classifying non-neutral expressions highlights the need for alternative methods for detecting neutral expressions. This research identifies this as crucial to improve accuracy and prevent biases in real-world applications. Also, this research identifies several promising areas for future studies in neutral facial expression detection. These include developing more precise and reliable machine learning models, collecting and annotating more extensive and diverse datasets focusing on neutral expressions and investigating the impact of various factors on neutral expression identification. The research emphasizes the need for dedicated datasets that explicitly classify neutral and non-neutral expressions, as existing datasets primarily focus on identifying various emotions without specific categorization.

3. Datasets

The investigation focused on three datasets: FER2013, CK+, and JAFFE. The FER2013 dataset was selected for its large number of images and labels, while the CK+ dataset was chosen for its diverse age range. The JAFFE dataset, despite its limited size, was included for additional analysis.

For neutral facial expression detection, the FER2013, CK+, and JAFFE datasets have been used; nevertheless, they have drawbacks. These datasets lack diversity in terms of age, gender, and cultural representation, which restricts their applicability to broader demographic and cultural contexts. The absence of contextual information and limited variation in environmental factors in these datasets undermines the model's capacity to understand neutral expressions in real-world situations. Moreover, the small sample sizes of CK+ and JAFFE datasets pose a risk of overfitting and hinder the development of robust models. Therefore, while these datasets offer insights, their constraints must be recognized when using them for neutral facial expression detection.

The FER2013 dataset contains grayscale images of faces 48x48 pixels in size. The faces have been automatically registered so that the face is more or less in the centre of each image and occupies roughly the same amount of space. The facial expressions are classified as follows: (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral) [22]. Considering the purpose was to classify a facial emotion as Neutral or Not Neutral, the categories Angry, Disgust, Fear, Happy, Sad, and Surprise were combined as Not Neutral, with a total of 29688 photos (83%) belonging to the Neutral category and 6198 (17%) belonging to the Not Neutral category.

The CK+ dataset is formed through the collection of facial expression images of subjects ranging from 18 to 50 years old [2]. These images display various facial expressions, anger, contempt, disgust, fear, happiness, sadness and surprise. The dataset contains 920 image sequences from 123 subjects, 327 (36%) images belonging to the Not Neutral category and 593 (64%) images belonging to the Neutral category.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

The JAFFE dataset consists of 213 images of different facial expressions from 10 different Japanese female subjects. Each subject was asked to do 7 facial expressions, Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral [23–25]. Resulting in a total of 183 (86%) Not Neutral images and 30 (14%) Neutral images.

The distribution of Neutral and Not Neutral expressions for each of the datasets is presented in Table 1.

Table 1. Distribution of Neutral and Not Neutral expressions in different datasets.

	Neutral	Not Neutral	Total
FER2013	6198 (17%)	29688 (83%)	35886
CK+	593 (64%)	327 (36%)	830
JAFFE	30 (14%)	183 (86%)	213

In this study, a sequential deep CNN architecture was adopted, shown in Figure 2, which is a popular neural network design for image classification applications. The selection of this model as a starting point was a deliberate choice based on the promising performance in previous studies. However, this model was chosen as a starting point. It is not considered a final solution. It serves as a foundation to study the impact of the datasets.

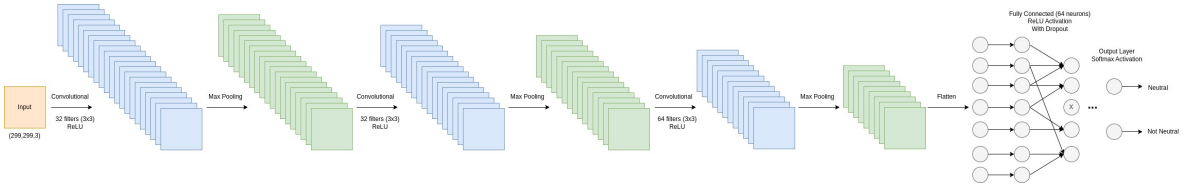


Figure 2. CNN architecture.

To extract high-level information from the input photos, the model used many convolutional layers followed by pooling layers. The ReLU (Rectified Linear Unit) activation function was utilized in the convolutional layers, which is extensively used in deep learning models.

The first layer in the model is a convolutional layer, which applies a set of learnable filters to the input image. In this case, the layer has 32 filters of size 3x3. The activation function used in this layer is ReLU, which introduces non-linearity into the model.

Following the convolutional layer, a max pooling layer is added to minimize the spatial dimensions of the convolutional layer's output by taking the maximum value of each non-overlapping sub-region. This reduces the number of parameters in the model and makes it less susceptible to minor picture translations.

This method of adding a convolutional layer followed by a max pooling layer is repeated twice more, raising the number of filters to 32 and then 64.

After the final max pooling layer, the output is flattened into a 1D array and passed to a fully connected Dense layer, which has 64 neurons and uses the ReLU activation function. To prevent overfitting, a dropout layer is added, which randomly drops out 50% of the neurons during training.

Lastly, the model's output layer is a Dense layer with the same number of neurons as the number of classes in the dataset. This layer's activation function is softmax, which computes the probability distribution over the classes.

Overall, this model is a relatively simple CNN architecture that can be used as a baseline for image classification tasks. It consists of several convolutional layers to learn spatial features, followed by fully connected layers to make a prediction on the output. The dropout layer helps to prevent overfitting, and the softmax activation function in the output layer produces a probability distribution over the classes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

It is important to note that the only difference between the training of each dataset was the input size, with the CK+ dataset having an input size of 96x96 and the FER2013 and JAFFE datasets having an input size of 48x48. This adjustment was made by resizing the images to ensure that the model was able to learn the features and patterns specific to each dataset effectively.

Initially, the performance of the model was evaluated using the original datasets. The FER2013 dataset exhibited higher accuracy for the Not Neutral class, while the CK+ dataset performed better on the Neutral class. These results indicated variations in model performance due to the dataset characteristics.

Due to the limited number of images in the JAFFE dataset, even though it achieved high accuracy on train, test and validation sets, the results were not reliable for the neutral classification task as the model misclassified more neutral images as not neutral compared to the number of neutral images classified as not neutral. The confusion matrix showed that 7 neutral images were classified as not neutral, and 6 neutral images were classified as neutral, as shown in Table 2.

Table 2. Confusion matrix for JAFFE dataset.

		Predicted Class	
		Neutral	Not Neutral
Actual Class	Neutral	6	7
	Not Neutral	0	30

To address the imbalances within the datasets, the FER2013 and CK+ datasets were balanced by randomly removing samples. This balancing led to improved performance in predicting the Neutral class for the FER2013 dataset, while the CK+ dataset's overall performance decreased. It's worth noting that the CK+ dataset initially contained fewer images compared to the FER2013 dataset. The process of deleting additional images to balance the CK+ dataset likely contributed to the observed decrease in performance.

Data augmentation techniques were applied to enhance the training process on the balanced FER2013 and CK+ datasets. Brightness and contrast modifications were performed on the FER2013 dataset, while no images were removed from the CK+ dataset due to the smaller imbalance. Resulting in 18594 (50%) Not Neutral images and 18594 (50%) Neutral images for the FER2013 dataset, and 981 (45%) Not Neutral images and 1186 (55%) Neutral images for the CK+ dataset. The data augmentation on CK+ was performed without removing any images because the difference was not as significant as the FER2013 dataset. The models trained on these augmented datasets exhibited improved accuracy and F1-scores compared to their balanced counterparts.

Comparing the results of the FER2013 dataset with data augmentation to the balanced FER2013 dataset, the former demonstrated better accuracy and F1-scores. Similarly, the CK+ dataset with data augmentation improved performance compared to the balanced CK+ dataset.

To improve the performance of the model and increase the dataset size, a decision was made to merge the CK+, JAFFE, and FER2013 datasets. By combining these three datasets, it is expected to result in better model performance. There are two ways to generate a training set and a testing set for a facial expression recognition system, person-dependent or person-independent. In the person-dependent case, the individuals included in the testing images also show up in the training images. It means that the classifier has seen the individuals included in the testing images. However, in the person-independent case, the individuals included in the testing images never appear in the training images [26]. Having a person-independent dataset brings several advantages to face recognition and expression analysis tasks. It helps evaluate the generalization capability of a model, promotes the development of robust models, ensures that the model's performance is not limited to specific individuals but can be generalized to a wide range of users, helps mitigate the risk of overfitting and reduces the risk of bias and fairness issues. However, there are also potential disadvantages to consider.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

Collecting and curating a person-independent dataset can be challenging. It requires careful selection from diverse sources and may contain inherent variations and inconsistencies due to differences in lighting conditions, poses, and image quality, which can introduce additional complexities for model training and evaluation.

In the context of person-independent datasets, it is important to note that the two studies yielded less favourable results compared to person-dependent datasets. Studies [26,27] reported inferior performance when employing person-independent datasets. These findings suggest that the utilization of person-independent datasets may pose challenges and introduce complexities in achieving optimal accuracy and generalization.

The dataset consists of two classes, Neutral and Non-Neutral, is composed of the FER2013, CK+ and JAFFE datasets, and is person-independent. The percentage of samples in each class is shown in Table 3 for the test, train, and validation sets. In the test set, there are 1367 samples of the Neutral class and 6110 samples of the Not Neutral class, making a total of 7477 samples (20% of the total samples). The train dataset has 4909 samples of the Neutral class and 21690 samples of the Not Neutral class, making a total of 26599 samples (72% of the total samples). The validation dataset has 545 samples of the Neutral class and 2409 samples of the Not Neutral class, making a total of 2954 samples (8% of the total samples). The Total row shows the overall distribution of samples in each class and the total number of samples.

Table 3. Distribution of samples in train, validation, and test sets by class for the merged dataset.

	Neutral	**Not Neutral**	**Total**
Test	1367	6110	7477 (20%)
Train	4909	21690	26599 (72%)
Validation	545	2409	2954 (8%)
Total	6821 (18%)	30209 (82%)	**37030**

To increase the number of neutral images, it was applied a 0.6 contrast and brightness and a 1.5 contrast and brightness only in the train set. Here is an example in Figure 3. The decision to use specific values (0.6 and 1.5) instead of random values for the contrast and brightness adjustments was based on experimental considerations. It was observed that the images from the three datasets exhibited significant variations in contrast and lighting conditions. Applying random values could result in some images becoming overly bright or too dark, hindering the augmentation process. The `convertScaleAbs` function in OpenCV [28] was used to convert an image to a different data type and scale the pixel values. It performs a linear transformation on the input image, multiplying each pixel value by a scale factor and adding an optional offset. The resulting pixel values are then converted to absolute values and clipped to the range of the output data type. This function is used to perform contrast adjustments on images. By adjusting the "alpha" and "beta" parameters, the scaling and shifting of the pixel values are controlled, enhancing or reducing the contrast of the image.

(a) Original image. **(b)** Modified (0.6). **(c)** Modified (1.5).

Figure 3. Changes applied in the original image for the data augmentation.

 doi:10.20944/preprints202310.1433.v1

During the training of the model, early stopping was implemented based on the validation accuracy and loss. The decision to use early stopping was driven by the observation that the validation loss was increasing. To address this issue and prevent overfitting, early stopping was employed as a solution. This approach helps select the best model that achieves optimal performance while avoiding further deterioration of the validation loss.

As the validation accuracy and loss did not show significant improvement, another strategy was employed. The validation set was expanded by transferring a portion of the images from the training set to the validation set. This expansion aimed to provide the model with a larger and more diverse set of validation data, facilitating better evaluation and monitoring of the model's performance. Importantly, this adjustment was made while maintaining the person-independent nature of the dataset.

To ensure that the data augmentation techniques remained effective, they were exclusively applied to the training set. By keeping the data augmentation limited to the training data, the model could still benefit from increased variability and robustness during training while allowing for a fair evaluation of the validation set.

By combining early stopping, expanding the validation set, and applying data augmentation selectively, efforts were made to improve the model's training process, mitigate overfitting, and enhance overall performance.

In an effort to further increase the data, it was used online data augmentation techniques while training the model, disregarding the use of class weights. By incorporating data augmentation into the training process and not relying on class weights, the aim was to increase the diversity and variability of the dataset, thereby improving the model's ability to generalize and make accurate predictions. This approach allowed for a more comprehensive exploration of the data space and offered potential insights into the impact of data augmentation without considering class weights on model performance.

The first technique utilized is rotation. This involves randomly turning images within a specified degree range for improved robustness while recognizing objects from varying angles in real-world scenarios effectively. Then comes horizontal and vertical shifts introduced by randomly shifting image components based on specific fractions of their total dimensions (width/height). Zooming was another augmentation technique used. Random zooming is applied to the images, altering their scale and helping the model detect objects at different levels of detail. The model learns to recognize objects, regardless of their left-right or up-down orientation, by flipping images. To ensure proper data scaling, a rescaling operation is performed. This involves dividing each pixel value by a specified constant, effectively scaling the pixel values to a range between 0 and 1. Rescaling helps the model converge faster during training. And random adjustments to brightness were made. By modifying the brightness of images within a given range, the model becomes more robust to variations in lighting conditions. These data augmentation techniques collectively enhance the diversity and quantity of the training data, allowing the model to generalize better and improve its performance when presented with real-world scenarios.

In order to address class imbalance within the dataset, class weights were calculated and incorporated during the online data augmentation process. The weights assigned to each class were, for the Neutral class, 2.71 and, for the Not Neutral class, 0.61. By considering these class weights, the data augmentation techniques applied online aimed to provide a balanced representation of both classes during the training process. The higher weight for the Neutral class indicates its relative importance and ensures that the model gives appropriate attention to this class, despite its imbalance in the dataset. Similarly, the lower weight for the Not Neutral class indicates its relatively larger representation in the dataset and helps mitigate any bias that may arise due to class imbalance. Incorporating class weights in data augmentation is an effective strategy to improve the model's performance by accounting for the uneven distribution of classes and ensuring fair treatment during training.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

Overall, the investigation highlighted the impact of dataset balancing and data augmentation techniques on model performance. While balancing the datasets addressed the class imbalance, it also introduced some biases and affected the overall performance. Data augmentation, on the other hand, enhanced the model's ability to generalize and perform well on unseen data.

By addressing the class imbalance and utilizing online data augmentation techniques with class weights, the model was able to improve its ability to accurately classify neutral expressions while maintaining strong performance in identifying non-neutral expressions.

The best results were obtained for the merged dataset when employing offline data augmentation techniques for the Neutral class and when employing online data augmentation for both classes using class weights. However, the decision was to use the merged dataset with online data augmentation for both classes using class weights because it increases data diversity, balanced augmentation, adaptability to real-world scenarios, and helps avoid biases towards the majority class. These factors can contribute to improved model performance, better generalization, and more equitable treatment of all classes in the training process.

4. Models

The selection of an appropriate model is crucial in the field of machine learning to obtain accurate and reliable results. This section explores various models utilized in this research, including InceptionV3, VGG16, ResNet50, and MobileNet, highlighting the modifications and strategies implemented to enhance their performance. These models were chosen based on their effectiveness in emotion recognition tests and their widespread adoption in the scientific community.

After the dataset preparation phase described in the previous section, the dataset used for studying the four models was created by merging images from the FER2013, CK+, and JAFFE datasets. This merged dataset is augmented using online data augmentation techniques and incorporates class weights. The dataset consists of images classified into two distinct categories: neutral and not-neutral. The dataset is divided into three subsets to facilitate model training and evaluation: the training, test, and validation sets.

The first model studied was the pre-trained InceptionV3 model for emotion classification, shown in Figure 4. InceptionV3 is a deep convolutional neural network architecture trained on ImageNet, a large-scale picture categorization challenge [29]. This pre-training procedure allows the model to learn rich and general features from various images.

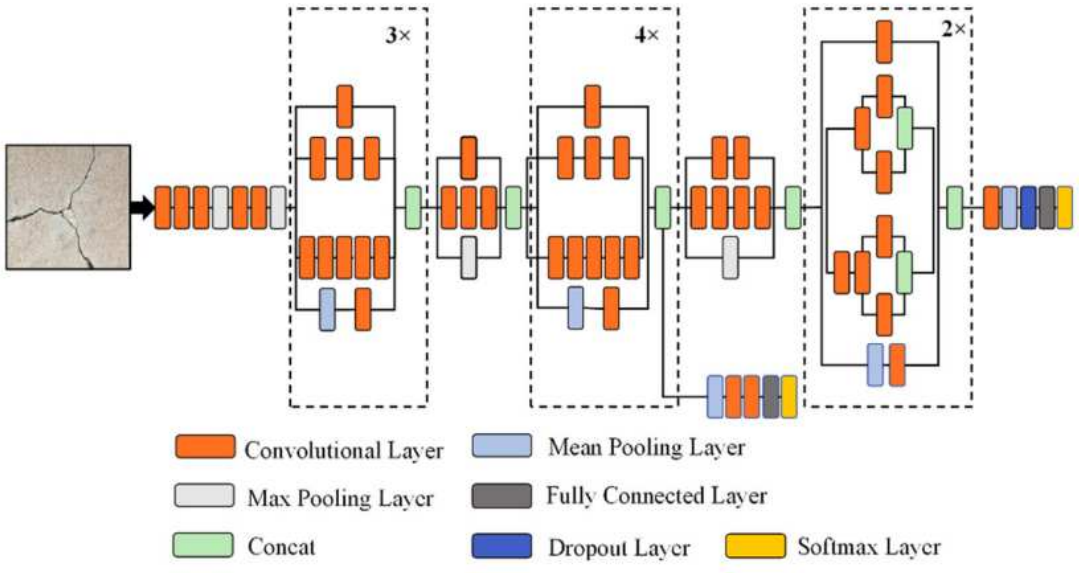


Figure 4. InceptionV3 architecture [30].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

The InceptionV3 model was loaded, but the top layer, which consists of fully connected layers responsible for the final classification, was removed. In order to adapt the model for a particular goal, in this instance, emotion categorization, the top layer must be removed. After removing the top layer, the remaining layers of the InceptionV3 model are set to non-trainable. This means that the weights of these layers will remain the same during the training process. The knowledge obtained from the ImageNet dataset is kept by freezing the weights of the pre-trained layers, preventing it from being overwritten or disrupted during training on the new emotion classification task.

The model can be customized for the particular characteristics and patterns of emotions available in the dataset for emotion classification by altering the InceptionV3 model and making only the top layers trainable. This method uses pre-trained features while fine-tuning the model to improve its performance on the emotion categorization job.

The architecture for emotion classification is enhanced by incorporating several top layers into the pre-trained InceptionV3 model. These layers include a Global Average Pooling Layer, which performs pooling over the spatial dimensions of the convolutional feature maps. By averaging the values in each feature map, this layer reduces the spatial dimensions and captures a condensed representation of the learned features. It effectively extracts crucial global information from the convolutional features [31].

Next, a Dense Layer is introduced after the global average pooling layer. This fully connected layer consists of multiple neurons that facilitate complex computations and feature transformations. Taking the output from the previous layer, the dense layer applies a Rectified Linear Unit (ReLU) activation function. The inclusion of ReLU introduces non-linearity into the model, enabling it to learn and represent intricate patterns within the data, improving the performance of the network [32].

To prevent overfitting and improve the model's generalization ability, a Dropout Layer is included. During training, this layer randomly sets a fraction of the input units to zero. By doing so, it reduces the reliance on specific features, encouraging the model to learn more robust and generalized representations. The provided code snippet specifies a dropout rate of 0.5, indicating that 50% of the input units are randomly dropped during each training iteration.

The architecture is further enhanced with a Softmax Output Layer, which is the final layer added. This layer consists of two neurons corresponding to the two emotion classes for classification (neutral or not neutral). The softmax activation function is applied to the output of this layer, generating probabilities for each class. These probabilities represent the model's confidence in assigning a given input to each emotion category.

By combining the pre-trained InceptionV3 model with the global average pooling layer, dense layer, dropout layer, and softmax output layer, the architecture is customized to effectively classify emotions based on the learned features from the InceptionV3 backbone.

The model is compiled by specifying the optimizer as Adam with a learning rate of 0.001. The optimizer updates the model weights during training based on the calculated gradients. In addition, the loss function is set to categorical cross-entropy, which assesses the difference between predicted and actual class labels. The accuracy metric is used to monitor the model's performance during training. Accuracy is calculated by dividing the number of correctly identified samples by the total number of samples. Analyzing how well the model performs on the classification task is easy when accuracy is tracked as a statistic. Then, the training procedure is carried out by iteratively changing the model's parameters based on the training data. To prevent overfitting, early stopping is implemented with a patience of 3. This means the training procedure will be terminated early if the validation loss does not improve for three consecutive epochs.

After training, the model's performance is tested on both the validation and training sets. The evaluation method is used on these datasets to compute and acquire the model's accuracy and loss values. The accuracy metric represents the fraction of correctly classified samples, whereas the loss value quantifies the difference between predicted and actual class labels.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

Once the model has been evaluated, the code saves the trained model as an h5 file. This file format allows for easy storage and future utilization of the trained model. Saving the model enables its reuse without having to retrain it from scratch.

The code predicts the labels of the test set using the trained model to analyze the model's performance further. The classification report and confusion matrix are then generated using the expected labels. In order to provide insight into the model's performance on several emotion categories, the classification report includes data such as precision, recall, and F1-score for each class. Furthermore, the code plots the accuracy and loss curves during training, allowing for visual monitoring of the model's learning progress and potential overfitting.

The second model studied was the pre-trained VGG16 model for emotion classification, shown in Figure 5. VGG16 is a deep convolutional neural network architecture that has been trained on ImageNet. The VGG16 followed the same procedure as the InceptionV3 model.

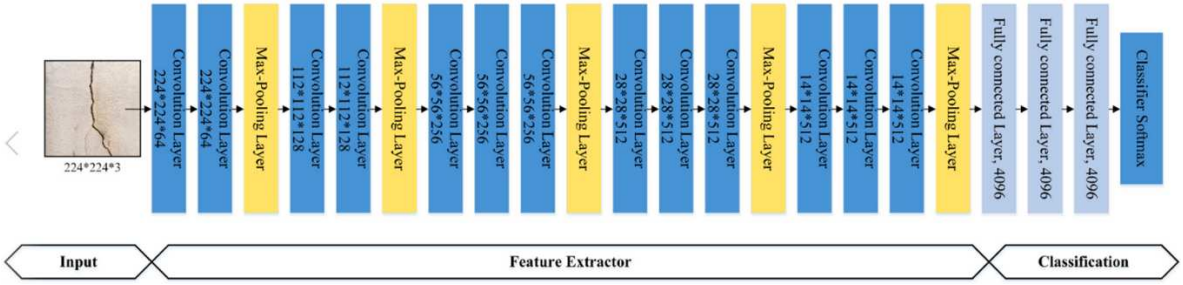


Figure 5. VGG16 architecture [30].

The third model studied was the pre-trained ResNet50 model for emotion classification, shown in Figure 6. ResNet50 is a deep convolutional neural network architecture that has been trained on ImageNet. The ResNet followed the same procedure as the InceptionV3 model.

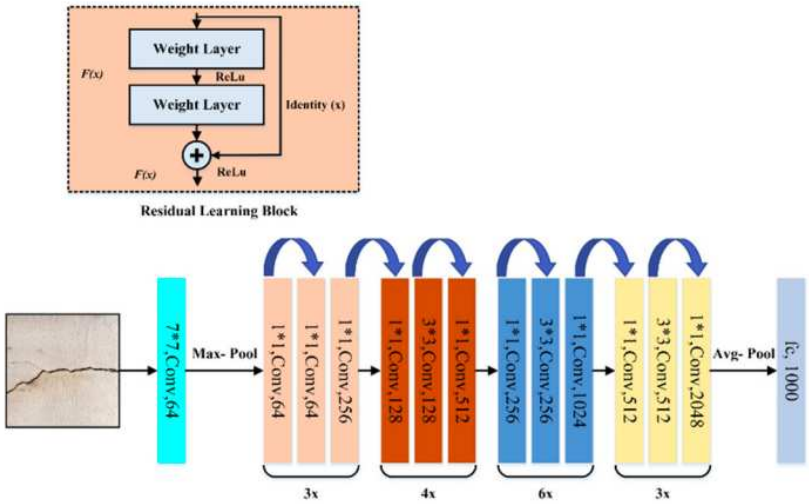


Figure 6. ResNet50 architecture [30].

The fourth model studied was the pre-trained MobileNet model for emotion classification, shown in Figure 7. MobileNet is a deep convolutional neural network architecture that has been trained on ImageNet. The MobileNet followed the same procedure as the InceptionV3 model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

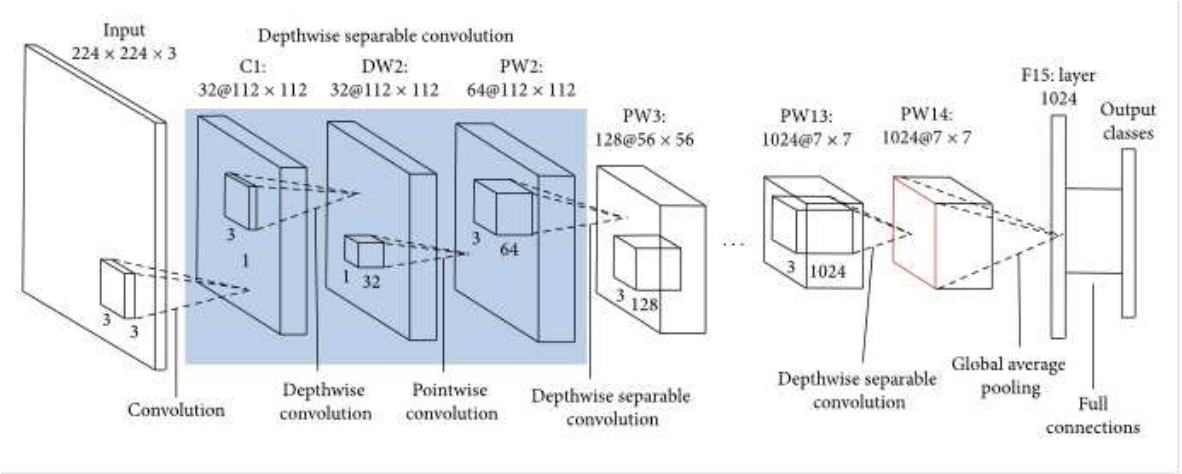


Figure 7. MobileNet architecture [33].

To enhance the performance of the emotion classification model, the InceptionV3 architecture was selected due to its superior results in preliminary evaluations.

Data augmentation techniques are applied to improve the model's ability to generalize and handle variations in the input data. This involves randomly applying rotation, shift, zoom, and flip transformations to the training images. These augmented images are then used to train the model, allowing it to learn robust features.

Since the dataset may suffer from class imbalance, it is essential to account for this during training. Class weights are calculated and are helpful to give more importance to the minority class, which is the Neutral class, during the training process, ensuring that the model is not biased towards the majority class.

The pre-trained InceptionV3 model is loaded, without the top layer, as it is specific to the original classification task and needs to be replaced for neutral and not neutral classification.

To adapt the InceptionV3 model for emotion classification, as opposed to what was explained above, the first step was to freeze most of the layers in the InceptionV3 model to prevent them from being modified during training and unfreeze the last five layers of the model. Also, custom top layers are added on top of the pre-trained base layers. These top layers consist of global average pooling, dense layers with activation functions, and dropout regularization. The model is then compiled with an appropriate optimizer and loss function.

The model is then trained on the augmented training dataset, and the training is monitored using early stopping, which stops the training if the validation loss does not improve for three consecutive epochs. A learning rate scheduler is employed to dynamically adjust the learning rate during training. The Adam optimizer is employed for model training, with an initial learning rate of 0.001. Additionally, a learning rate scheduler, specifically `ReduceLROnPlateau`, is incorporated to dynamically adjust the learning rate based on validation loss, with a reduction factor of 0.1 after 3 epochs of no improvement and a minimum learning rate of 0.0001.

To determine the impact of not using the early stopping, another test was performed on the InceptionV3 model, continuing with the five layers unfrozen but this time without the early stopping and training for around 20 epochs.

The graph (Figure 8) provides a visual comparison of the loss and accuracy results for different models: InceptionV3 (M1), VGG16 (M2), ResNet50 (M3), MobileNet (M4), InceptionV3 Unfreezed (M5), and InceptionV3 No Early Stopping (M6). The bars represent each model's test, train, and validation loss and accuracy. Among the models, ResNet50 shows the highest test and validation loss, indicating that it may struggle to generalize well to unseen data. InceptionV3 and MobileNet demonstrate lower test and validation loss values, suggesting better generalization capabilities than other models. InceptionV3 No Early Stopping exhibits the lowest test, train, and validation loss, indicating that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

the model may have achieved the best overall performance in minimizing the loss. Regarding test accuracy, InceptionV3 and VGG16 demonstrate the highest values, suggesting better performance in correctly classifying unseen data. ResNet50 and MobileNet show relatively lower test accuracy, indicating potential challenges in predicting classes for new samples accurately. InceptionV3 No Early Stopping exhibits the highest train and validation accuracy, indicating that the model achieved the best overall performance in correctly classifying the training and validation data. Based on these observations, InceptionV3 No Early Stopping outperforms the other models in terms of both loss and accuracy.

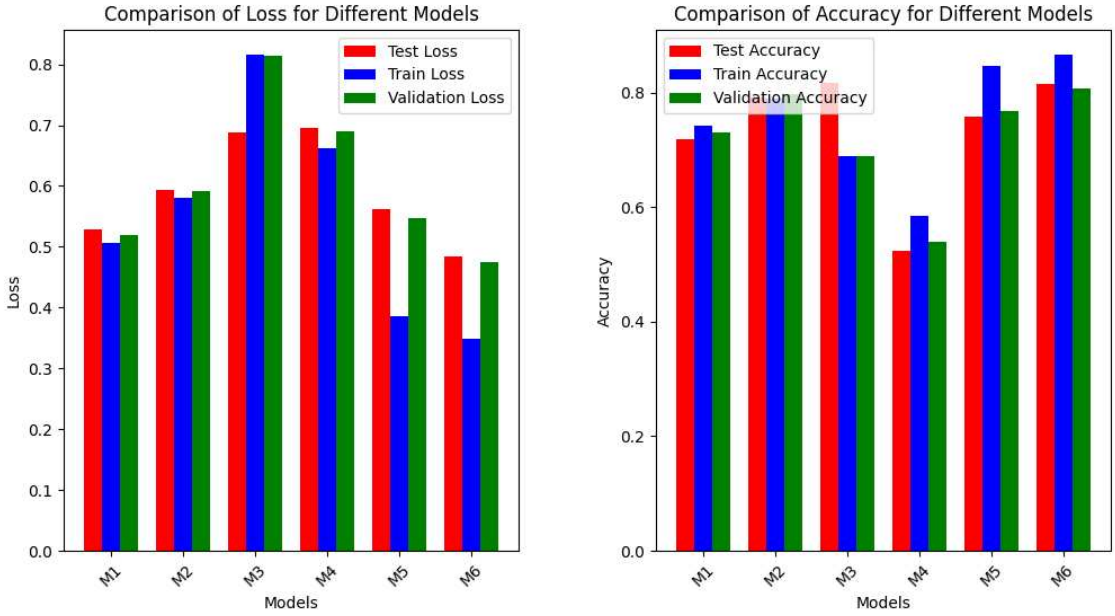


Figure 8. Loss and accuracy comparison for different models.

Table 4 includes the precision, recall, F1-score (weighted average) and the number of epochs for each model. The InceptionV3 model demonstrates competitive performance across all metrics, including precision, recall, and F1-score.

Table 4. Comparison of classification reports for models.

Model	Precision (Weighted Avg)	Recall (Weighted Avg)	F1-Score (Weighted Avg)	Epochs
InceptionV3	0.82	0.72	0.75	6
VGG16	0.77	0.79	0.78	4
ResNet50	0.67	0.82	0.73	7
MobileNet	0.74	0.76	0.75	4
InceptionV3 Unfreezed	0.78	0.81	0.78	4
InceptionV3 No Early	0.78	0.81	0.78	20

The InceptionV3 model is considered a better choice because it achieved a higher precision (0.82) compared to other models, indicating that it can better classify both the Neutral and Not Neutral classes accurately. While InceptionV3 had a recall of 0.72 for both classes, it still performed competitively compared to other models. Recall measures the model's ability to capture positive instances, and InceptionV3 demonstrated a good balance in this aspect. InceptionV3 achieved an F1-score of 0.75, indicating a good balance between precision and recall. This suggests that the model can accurately classify instances while effectively capturing positive ones. The model also exhibited reliable and favourable outcomes across multiple metrics, including precision, recall, and F1-score. This consistency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

in performance indicates that the model is robust and dependable in differentiating between the classes. Considering the higher precision, competitive recall, balanced F1-score, and consistent performance, the InceptionV3 model is a better choice among the evaluated models.

5. Preprocessing

Preprocessing is necessary for data preparation for analysis and machine learning activities. The preprocessing phases were used on this study's training and validation sets. It was divided into five phases, as shown in Figure 9, each of which aimed to address particular data characteristics and improve the quality of the data for later analysis.

Figure 9. Preprocessing phases.

The first phase, Phase 0, refers to the raw set. The raw set initially includes the raw, unprocessed data gathered throughout the data-collecting procedure. This is the initial step of the preprocessing pipeline and serves as the basis for subsequent phases; however, because it was the results of the raw data, it will consider the findings from the InceptionV3 model from the previous chapter.

Phase 1 refers to the rotation correction set. In the next phase, the rotation correction set focuses on correcting any misalignments or variances in the orientation of the data during this step. For example, image rotation and alignment techniques are used to normalize the orientation of data samples, assuring uniformity across the collection.

Phase 2 refers to the cropped set. The emphasis moves to deleting irrelevant or extraneous information from the data. Cropping techniques are used within each sample to separate and extract the relevant regions of interest. This phase aids with removing background noise and extraneous visual features, improving the clarity and emphasis of the data.

Phase 3 refers to the histogram equalization set. In this phase, histogram equalization is used to increase the data samples' overall contrast and visual quality. This technique helps to equalize the histogram by shifting the intensity values in the photos, resulting in a more balanced depiction of the data. This step is especially effective for increasing the visibility of characteristics and patterns in the data.

Phase 4 refers to the smoothed set. The smoothed set's final phase seeks to reduce noise and artefacts in the data. Several smoothing or noise-cancelling techniques, such as Gaussian smoothing or median filtering, are used to reduce the impact of random changes or inconsistencies in the data. This phase contributes to the overall quality and dependability of the data samples.

Systematically progressing through these preprocessing phases, the data is refined, standardized, and enhanced, setting the phase for subsequent analysis or machine learning tasks. Each phase explicitly addresses different aspects of the data, contributing to the overall data preparation process.

Preprocessing is an important step in facial expression detection that can improve the accuracy of the algorithm by handling variations in images. By applying techniques such as rotation correction, cropping, histogram equalization, and smoothing, the algorithm can better detect and analyze facial features, leading to more accurate results [34,35].

The first step is to analyze the impact of each preprocessing phase, on the CNN Model.

The CNN model was trained using the selected dataset. Figure 10 illustrates the accuracy and loss trends throughout the epochs for each phase, providing a comprehensive visualization of their respective performances.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

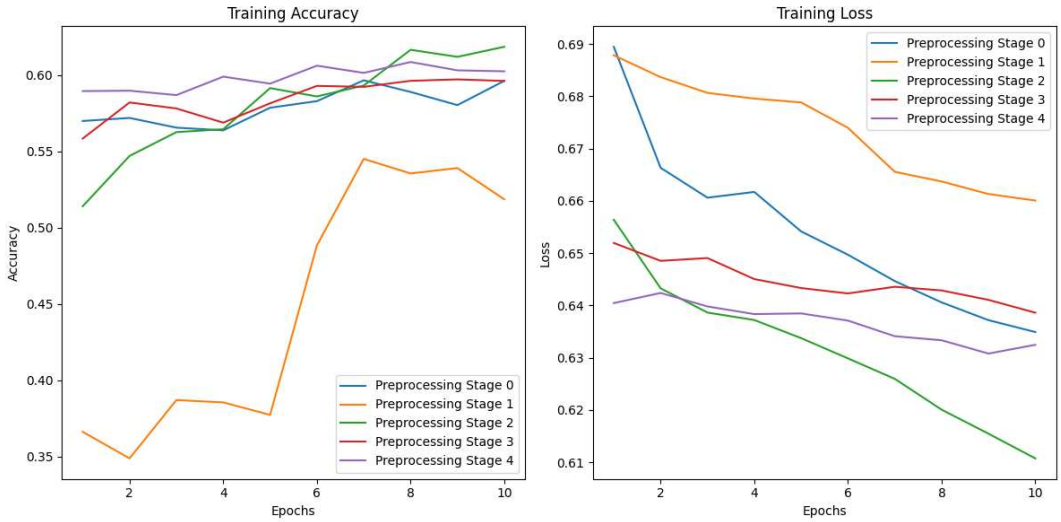


Figure 10. Accuracy and loss graphs for the CNN Model for each of the phases.

The second step is to analyze the impact of each preprocessing phase, on the InceptionV3 Model.

The InceptionV3 model was trained using the selected dataset, and Figure 11 shows the accuracy and loss tendencies throughout the epochs for each phase, providing a visualization of their individual performances. A distinct pattern appears after analyzing the accuracy and loss graphs. During Phase 0, the precision is at its maximum, while the loss is at its lowest. Surprisingly, despite using numerous preprocessing steps, they all resulted in no improvement in the findings. The accuracy stays stable, and the loss does not drop beyond the initial values obtained in Phase 0.

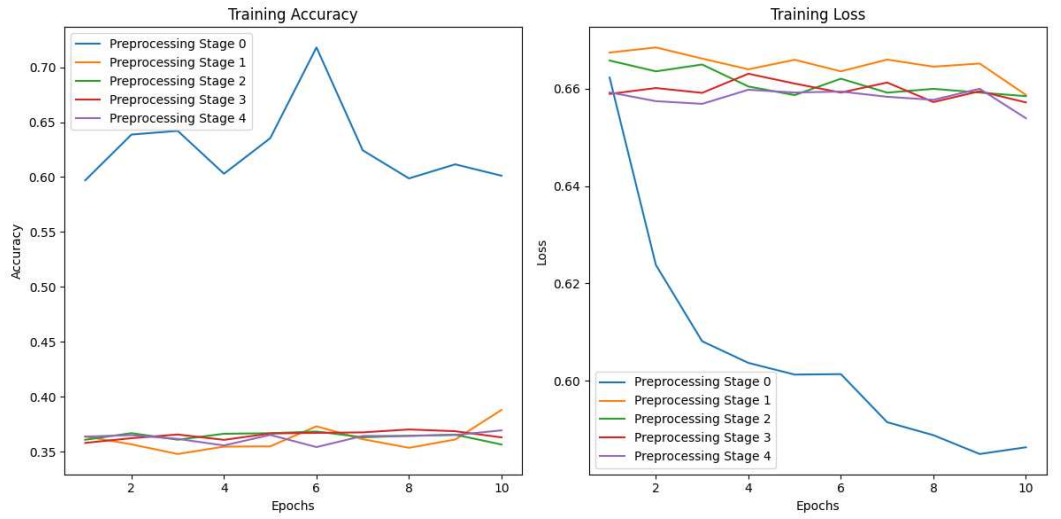


Figure 11. Accuracy and loss graphs for the InceptionV3 Model for each of the phases.

The difference in the impact of preprocessing between the simple CNN model and the InceptionV3 model could be attributed to the models' complexity.

The preprocessing techniques for the simple CNN model enhanced the model's ability to extract significant features from the dataset. This could have resulted in improved accuracy and reduced losses. The simplicity of the model allows for easier manipulation and adaptation to the preprocessing steps, leading to noticeable improvements in performance.

On the other hand, the InceptionV3 model has a more complex architecture that incorporates sophisticated convolutional operations and intricate feature extraction mechanisms. The preprocessing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

techniques may not have significantly impacted the model's performance due to the model's inherent ability to handle a wide range of input variations and extract complex features.

6. Results

The following protocol was developed to test a real-time system for detecting a neutral facial expression using a camera and to create a dataset consisting of frames specifically focused on neutral expressions:

1. Privacy consent: Before proceeding with the data collection process, it is obtained the participant's informed consent regarding the capturing and usage of their facial expressions.
2. Set up the camera: Determine that the camera is correctly connected and positioned to capture the subject's facial expressions clearly. As needed, change camera settings such as resolution and frame rate.
3. Position the subject: Instruct the participant in the research to stand in front of the camera, facing it. Make sure there is enough light to accurately capture face characteristics.
4. Start the real-time system and collect data: Launch the real-time system for neutral facial expression detection. This system incorporates the model for the classification of neutral facial expressions, as well as the result if the expression is indeed neutral and its confidence value. Begin the data capturing process, and the camera starts recording.
5. Instruct the participant to start by executing a neutral facial expression: Request that the subject relax their face and maintain a neutral expression.
6. Capture frames of the neutral expression: Request the subject to keep the neutral expression for a set amount of time, such as 5 seconds.
7. Instruct the participant to perform other facial expressions: Instruct the subject to execute various facial expressions, such as happy, sad, surprised or angry.
8. End the data capture: Stop the data collection procedure on the camera after the participant has performed the neutral expression and other expressions. Ensure that each participant's neutral frame with the highest confidence level is correctly saved to the chosen storage place.
9. Validate the captured frames: Conduct a visual inspection of the captured frames to ensure the quality and clarity of the neutral facial expressions. Remove any frames that do not correspond to a neutral facial expression.

The technology that has been deployed focuses on real-time facial expression recognition. It starts by gathering video input from a webcam, presuming the participant is facing the camera. The system recognizes faces inside each frame of the video using the dlib face detector. During face detection, the system extracts facial landmarks. Figure 12 shows the system phases.

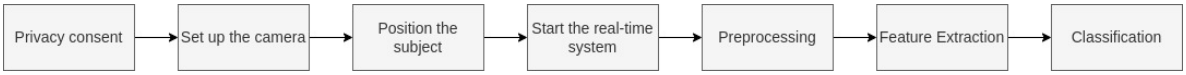


Figure 12. Data collection steps.

In order to attain correct alignment, a rotation correction based on face landmarks is conducted. The landmarks are then used to crop the face, and different preprocessing procedures are used to improve the facial characteristics. Histogram equalization is used to increase contrast, Gaussian blurring is used to minimize noise, and picture scaling is used to match the input size required by the pre-trained model.

The pre-trained model predicts the participant's emotion as neutral or not neutral. For each frame, the model assigns a prediction label and a confidence score. The frames with emotion labels are presented in real time and written to an output video file. If a frame with the highest confidence for the neutral emotion is discovered, it is saved as a separate image file. This frame is an example of a neutral representation that can be used for additional analysis. The whole video recording, which includes all frames that have been analyzed, is then saved by the system.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

A total of 32 participants took part in this test, consisting of 23 women and 9 men. Table 5 presents the results obtained from analyzing the videos of the participants. The table provides key insights into the expressions captured and saved in the dataset. The table's column "Expression Detected" indicates what expression (neutral or not neutral) was detected when the participant was asked at the beginning of the video to perform a neutral expression. "Expression Saved" indicates if the image saved on the dataset was a neutral expression captured or if it saved a not neutral detected as neutral. "Confidence Value" indicates the confidence value for the image saved on the dataset. This indicates the system's level of certainty in detecting and categorizing the expressions. The lower confidence values suggest a lower level of confidence in the accuracy of the detected expressions, while the higher values indicate a higher level of confidence.

Table 5. Analysis of Image Expressions with Detection of Neutral Expression.

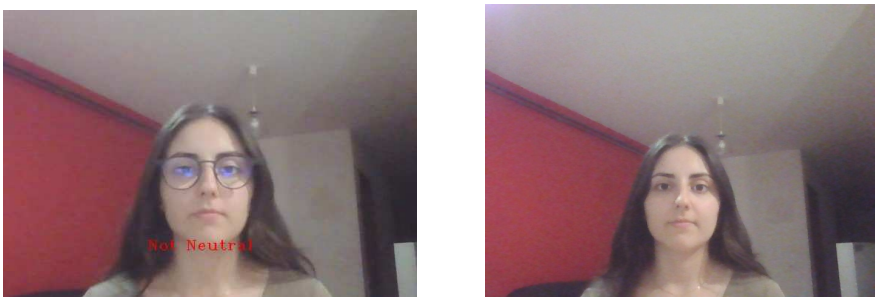
Person	Expression Detected	Expression Saved	Confidence
1	Neutral	Neutral	60%
2	Neutral	Neutral	58%
3	Neutral	Neutral	55%
4	Neutral	Neutral	51%
5	Not Neutral	-	-
6	Neutral	Neutral	71%
7	Neutral	Neutral	70%
8	Neutral	Neutral	53%
9	Neutral	Not Neutral	59%
10	Neutral	Neutral	63%
11	Neutral	Neutral	57%
12	Neutral	Neutral	63%
13	Neutral	Neutral	69%
14	Not Neutral	-	-
15	Neutral	Neutral	58%
16	Neutral	Neutral	62%
17	Neutral	Neutral	69%
18	Neutral	Not Neutral	61%
19	Neutral	Neutral	58%
20	Neutral	Neutral	58%
21	Not Neutral	-	-
22	Not Neutral	-	-
23	Neutral	Neutral	79%
24	Neutral	Neutral	77%
25	Neutral	Neutral	78%
26	Neutral	Neutral	52%
27	Neutral	Neutral	70%
28	Not Neutral	-	-
29	Neutral	Neutral	63%
30	Neutral	Not Neutral	58%
31	Neutral	Neutral	73%
32	Neutral	Neutral	70%

In five participants (5, 14, 21, 22 and 28), despite starting with a neutral expression and subsequently performing other expressions such as happiness or surprise, the neutral expression

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

was never detected. In all frames captured for these participants, the expressions were classified as Not Neutral.

Figure 13 shows an example of a participant that was wearing glasses, performing a neutral expression, but was never detected. Removing the glasses, the neutral expression was correctly detected.

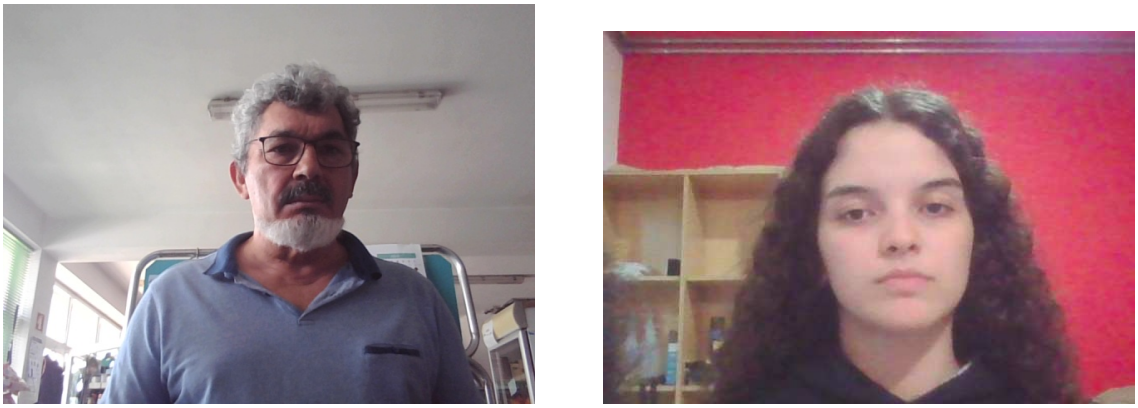


(a) Neutral expression not detected. **(b)** Neutral expression detected.

Figure 13. Neutral expression misclassified and correctly classified.

In two other participants (9 and 18) who followed the same procedure, the initial frames were classified as Neutral. However, the saved frame for these participants represented a Not Neutral expression. The saved frame captured a smiling expression instead of the intended neutral expression. The smiling expression obtained a higher confidence value than the neutral expression detected in the beginning.

The confidence values associated with the frames saved vary from 51% to 79%. These scores belong to participants 4 and 23, respectively, shown in Figure 14.



(a) Neutral expression with the lowest confidence. **(b)** Neutral expression with the highest confidence.

Figure 14. Neutral expressions saved with the lowest and highest confidence values.

Of the 32 participants in this analysis, 24 had their neutral expressions correctly captured and saved in the dataset. This accounts for approximately 75% of the participants.

However, during the recording, approximately 16% of the total participants (five individuals) did not have any neutral expressions detected. In every frame captured for these individuals, their expressions were classified as Not Neutral. It is worth noting that two of these participants were wearing glasses, and there was no request made for them to remove their glasses. The presence of glasses could potentially have been a factor contributing to the inability to detect neutral expressions accurately, as seen in the example in Figure 13.

Additionally, three individuals (approximately 9%) had their saved images mistakenly classified as neutral expressions despite displaying different emotions. Specifically, two participants had their saved images representing a smiling expression, while one participant had their saved image representing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2023 doi:10.20944/preprints202310.1433.v1

a sad expression. Despite the initial frames being classified as Neutral, these had lower confidence values than those misclassified Not Neutral frames.

These findings highlight the success and challenges in accurately detecting and capturing neutral expressions, with most participants achieving accurate results. However, the presence of participants without any detected neutral expressions and instances of capturing not neutral expressions demonstrate the complexities involved in accurately assessing and categorizing facial expressions.

The relatively low confidence values associated with the detected neutral expressions indicate room for improvement in the dataset and model training.

Figure 15 shows several examples of expressions that were correctly classified as Not Neutral.

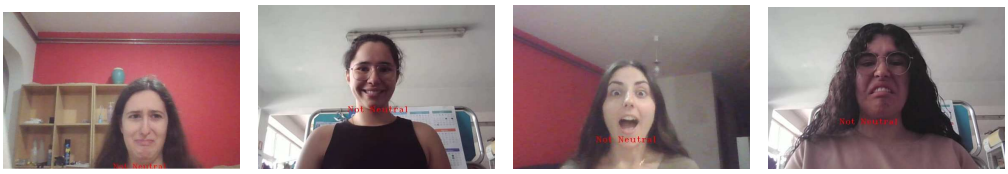


Figure 15. Not Neutral expressions correctly classified.

Figure 16 shows several examples of expressions that were Not Neutral and were misclassified as Neutral.

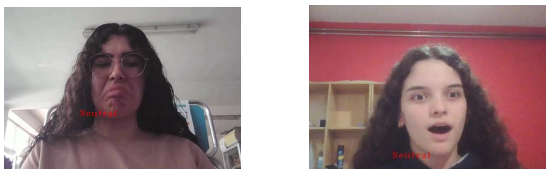


Figure 16. Not Neutral expressions misclassified.

Additional tests should be conducted to improve the accuracy and reliability of the neutral expression detection. This would help gather more data and further refine the model's training. The dataset can include a broader range of facial expressions and variations, ensuring a more comprehensive representation of neutral expressions. By incorporating these improvements, it is expected that the confidence levels of the detected neutral expressions will be significantly enhanced.

Therefore, future efforts should focus on iteratively refining the dataset and model training to achieve higher accuracy and confidence in capturing and categorizing neutral expressions.

7. Discussion

This work consolidates the knowledge gained from exploring datasets, model architectures and examining preprocessing methods. The findings and recommendations presented in this paper align with state of the art in facial expression detection research, particularly concerning the challenges and areas for improvement. The comparison with existing studies highlights the need for continuous improvement in the field and the importance of addressing real-world representativeness, dataset limitations, and model generalizability.

The choice of datasets significantly influences the model's performance. Each dataset, namely FER2013, CK+, and JAFFE, was individually tested, and distinct outcomes were observed for each dataset. The JAFFE dataset, with its limited number of images, presented challenges in accurately classifying neutral expressions.

On the other hand, data augmentation techniques proved successful, particularly in addressing the class imbalance issue present in the FER2013 and CK+ datasets. To overcome these limitations and leverage the strengths of each dataset, all three datasets were merged. This combined dataset offered a more comprehensive and diverse representation of facial expressions, leading to improved performance.

 doi:10.20944/preprints202310.1433.v1

Offline and online data augmentation techniques were applied to enhance the dataset further. The augmentation techniques played a crucial role in increasing the data diversity and addressing the class imbalance, resulting in better model generalization and performance. The best results were achieved using online data augmentation with class weights, promoting data diversity and mitigating biases, obtaining an accuracy of 64%.

Moving on to the choice of models, InceptionV3, with all layers frozen, outperformed other models such as VGG16, MobileNet, and ResNet50. Despite employing other techniques during training, InceptionV3, with all layers frozen and applying early stopping, exhibited superior results compared to the alternative models, achieving an accuracy of 72%.

Regarding preprocessing, the impact can vary depending on the model's complexity and architecture. The preprocessing techniques applied to the simple CNN model enhanced its feature extraction capabilities, resulting in improved accuracy and reduced losses. The model's simplicity allowed for easy manipulation and adaptation to the preprocessing steps, leading to noticeable enhancements in performance. In contrast, the InceptionV3 model's complex architecture incorporates sophisticated convolutional operations and intricate feature extraction mechanisms. As a result, the impact of preprocessing techniques may not have significantly influenced the model's performance. The InceptionV3 model is designed to automatically learn and extract complex features, making it less reliant on specific preprocessing steps.

The findings indicate that choosing datasets significantly impacts model performance, and merging datasets while applying appropriate data augmentation techniques can yield better results. The InceptionV3 model, with its complexity and inherent feature extraction capabilities, demonstrated superior performance even with no preprocessing. The combination of the merged dataset and the InceptionV3 model, along with the applied data augmentation techniques, resulted in a robust and accurate facial expression recognition system, outperforming other models and achieving an overall accuracy of 72%.

The study encountered several challenges that affected the training of models. One significant challenge was the issue of overfitting. Despite efforts to mitigate overfitting through techniques such as early stopping, the risk of overfitting persisted, mainly due to the limited size of the datasets. Another challenge arose from the unbalanced nature of the datasets, particularly regarding the need for more neutral images. This imbalance introduced bias in the training process, leading to difficulties in accurately categorizing neutral expressions.

As for the real-time system implemented, while the overall results were promising, with a significant number of participants correctly capturing their neutral expressions (75%), there were instances where neutral expressions were not detected, or not neutral expressions were incorrectly categorized. One factor that potentially impacted the accurate detection of neutral expressions was the presence of participants wearing glasses. In the recording process, some participants were wearing glasses, and the presence of glasses could have impeded the accurate detection of neutral expressions.

The model failed in specific areas during its evaluation. The frames in which the model failed may vary, but they generally show inaccurate or inadequate results regarding the model's task or objective. Examining the frames in which the model failed makes it possible to identify a pattern in the failures. Depending on the problem's nature and the data's characteristics, there may be different visual explanations for these failures. There may be inconsistencies in the face or facial expression detection, errors in emotion classification, or difficulties in recognizing subtle nuances in expressions. For three participants, it detected the neutral expression correctly. However, it saved the not-neutral expression. The not-neutral expression was, in two cases, a smiling face and, in the other case, a sad face. The sad face was not so evident. Because of that, the system could have trouble identifying that correctly. However, more tests should be done on more participants so a more noticeable pattern could be detected.

These findings emphasize the complexities in accurately detecting facial expressions, necessitating further refinement in dataset construction and model training.

 doi:10.20944/preprints202310.1433.v1

Even though this paper may not have achieved the same level of accuracy and performance as the state-of-the-art, it still provides valuable contributions and insights. It sheds light on the challenges and limitations in neutral facial expression detection, identifies areas for improvement, and proposes recommendations for future research.

8. Conclusion

In conclusion, the study comprehensively investigated neutral facial expression recognition, exploring various datasets, data augmentation techniques, model architectures and preprocessing methods. The results demonstrated the significance of dataset selection, with data augmentation proving an effective technique for addressing class imbalances. The choice of model architecture, such as InceptionV3, also played a crucial role in achieving high performance. The findings highlight the success and challenges in accurately detecting and capturing neutral expressions, emphasizing the need for further refinement in dataset construction and model training. By incorporating suggested improvements and addressing the identified limitations, neutral facial expression recognition accuracy can be improved.

The study's limitations include the impact of limited datasets on model performance, with challenges stemming from datasets like JAFFE having too few images for accurate neutral expression identification, and from data augmentation techniques whose potential drawbacks weren't examined. While InceptionV3 outperformed other models, the comparison lacked a comprehensive range of architectures. Overfitting risks, though mitigated with techniques like early stopping, were mentioned without details on effectiveness. The real-time system, promising yet with limitations, faced difficulties in correctly detecting and categorizing neutral expressions, possibly affected by participants wearing glasses. Failures of the model during evaluation weren't extensively analyzed, and the study's achieved accuracy raised questions about the approach's applicability to broader scenarios. Participant diversity, like glasses-wearing, received limited attention, and broader aspects such as age, gender, and ethnicity effects on model performance weren't discussed.

Despite the limitations encountered, the findings of this study contribute to advancing knowledge in this field and provide directions for future research.

Future efforts should focus on iteratively refining the dataset and model training to achieve higher accuracy and confidence in capturing and categorizing neutral expressions. This can be accomplished by expanding the dataset to include a broader range of facial expressions and variations, ensuring a more comprehensive representation of neutral expressions. Strategies like cross-validation, regularization techniques, and more diverse datasets should be considered to combat overfitting. Efforts should also be directed towards augmenting the dataset with a more significant number of neutral images to ensure a more balanced representation of facial expressions. By addressing these limitations and refining the training process, the accuracy and generalization capabilities of the models can be significantly improved. Including individuals who wear glasses is recommended to improve the dataset's representation.

Additionally, conducting additional tests and gathering more data can help further refine the model's training and improve the overall performance. Further exploration of preprocessing techniques tailored to the specific model architecture can also enhance the model's capabilities in capturing and categorizing facial expressions.

The study highlights several limitations that provide insight into the complexity of facial expression detection. The effects of limited datasets on model performance, particularly the challenges posed by datasets with a scarcity of images like JAFFE for identifying neutral expressions accurately, underline the need for more diverse and comprehensive data. Additionally, while data augmentation techniques helped, not exploring potential drawbacks leaves room for further investigation into the best practices for enhancing data quality. The comparison of models primarily focusing on InceptionV3's success, with a limited scope of other architectures, suggests a promising avenue for future research exploring a broader range of models. Addressing overfitting concerns and

 doi:10.20944/preprints202310.1433.v1

delving into the effectiveness of methods like early stopping could shed light on strategies to enhance model generalizability. The real-time system's limitations, especially in accurately detecting neutral expressions and potential influences of participant eyewear, emphasize the importance of refining real-world applicability. Furthermore, a deeper analysis of model failures during evaluation would provide a more comprehensive understanding of its shortcomings. The study's achieved accuracy, while not reaching state-of-the-art levels, still offers valuable insights. However, it's essential to consider the study's specific context, dataset sizes, and experimental conditions when evaluating its contributions. The study's focus on participant glasses is a starting point for addressing wider participant diversity aspects, like age, gender, and ethnicity, which could greatly impact real-world deployment. Ultimately, these limitations and insights encourage further refinements in dataset construction, model training, and real-world deployment strategies, making the study a meaningful step towards improving facial expression detection systems.

Author Contributions: Conceptualization, Lúcia Sousa; methodology, Lúcia Sousa; software, Lúcia Sousa; validation, Daniel Canedo, António Neves and Miguel Drummond; formal analysis, Lúcia Sousa; investigation, Lúcia Sousa; resources, Daniel Canedo; data curation, Lúcia Sousa; writing—original draft preparation, Lúcia Sousa; writing—review and editing, Daniel Canedo, António Neves, Miguel Drummond and João Ferreira; visualization, Lúcia Sousa; supervision, Daniel Canedo, António Neves, Miguel Drummond and João Ferreira; project administration, António Neves and João Ferreira. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AAM	Active Appearance Model
AU	Action Unit
AUDN	Action Unit-Inspired Deep Network
BU-3DFE	Binghamton University 3D Facial Expression
BU-4DFE	Binghamton University 4D (3D+time) Facial Expression
CCNU-FE	Central China Normal University Facial Expression
CK+	Extended Cohn–Kanade
CLM	Constrained Local Model
CNN	Convolutional Neural Network
CRF	Conditional Random Forest
DPND	Deep Peak-Neutral Difference
DT	Decision Tree
DPR	Deep Representation Feature
ED	Euclidean Distance
ELM	Extreme Learning Machine
EmotiW	Emotion Recognition in the Wild
ER	Emotion Recognition
FAP	Facial Animation Parameter
FACS	Facial Action Coding System
FC	Fully Connected
FER	Facial Expression Recognition
FDR	False Discovery Rate
HOG	Histograms of Oriented Gradients
HMM	Hidden Markov Model
ISL	Intelligent Systems Lab
JAFFE	Japanese Female Facial Expression
KE	Key Emotion
KDEF	Karolinska Directed Emotional Faces
KNN	K-Nearest Neighbor
LPQ	Local Phase Quantization
LBP	Local Binary Patterns
MTCNN	Multi-task Cascade Convolutional Neural Network

OF	Optical Flow
OpenCV	Open Computer Vision
RaFD	Radboud Faces Database
RF	Random Forest
ROI	Region of Interest
SFEW	Static Facial Expression in the Wild
SVM	Support Vector Machine
TL	Transfer Learning
TPR	True Positive Rate
RGB	Red, Green and Blue
ReLU	Rectified Linear Unit

References

1. Ekman, P. Basic emotions. *Handbook of cognition and emotion* **1999**, *98*, 16.
2. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression **2010**. pp. 94–101. doi:10.1109/CVPRW.2010.5543262.
3. Karnati, M.; Seal, A.; Krejcar, O.; Yazidi, A. FER-net: facial expression recognition using deep neural net. *Neural Computing and Applications* **2021**, *33*. doi:10.1007/s00521-020-05676-y.
4. Samadiani, N.; Huang, G.; Cai, B.; Luo, W.; Chi, C.H.; Xiang, Y.; He, J. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* **2019**, *19*, 1863.
5. Kuenecke, J.; Wilhelm, O.; Sommer, W. Emotion Recognition in Nonverbal Face-to-Face Communication. *Journal of Nonverbal Behavior* **2017**, *41*. doi:10.1007/s10919-017-0255-2.
6. Joormann, J.; Gotlib, I.H. Is this happiness I see? Biases in the identification of emotional facial expressions in depression and social phobia. *Journal of abnormal psychology* **2006**, *115*, 705.
7. Barreto, A. Application of facial expression studies on the field of marketing. *Porto: FEELab Science Books* **2017**.
8. Al-Modwahi, A.A.M.; Sebetela, O.; Batleng, L.N.; Parhizkar, B.; Lashkari, A.H. Facial expression recognition intelligent security system for real time surveillance **2012**.
9. Li, P.; Phung, S.L.; Bouzerdoum, A.; Tivive, F.H.C. Automatic Recognition of Smiling and Neutral Facial Expressions. 2010 International Conference on Digital Image Computing: Techniques and Applications, 2010, pp. 581–586. doi:10.1109/DICTA.2010.103.
10. Saito, A.; Sato, W.; Yoshikawa, S. Rapid detection of neutral faces associated with emotional value. *Cognition and Emotion* **2022**, *36*, 546–559.
11. Uono, S.; Sato, W.; Sawada, R.; Kawakami, S.; Yoshimura, S.; Toichi, M. Schizotypy is associated with difficulties detecting emotional facial expressions. *Royal Society Open Science* **2021**, *8*, 211322.
12. Sawada, R.; Sato, W.; Nakashima, R.; Kumada, T. How are emotional facial expressions detected rapidly and accurately? A diffusion model analysis. *Cognition* **2022**, *229*, 105235. doi:10.1016/j.cognition.2022.105235.
13. Jia, S.; Wang, S.; Hu, C.; Webster, P.J.; Li, X. Detection of genuine and posed facial expressions of emotion: databases and methods. *Frontiers in Psychology* **2021**, *11*, 580287.
14. Shao, Z.; Liu, Z.; Cai, J.; Wu, Y.; Ma, L. Facial Action Unit Detection Using Attention and Relation Learning. *IEEE Transactions on Affective Computing* **2022**, *13*, 1274–1289. doi:10.1109/TAFFC.2019.2948635.
15. Rizwan, S.A.; Jalal, A.; Kim, K. An Accurate Facial Expression Detector using Multi-Landmarks Selection and Local Transform Features **2020**. pp. 1–6. doi:10.1109/ICACS47775.2020.9055954.
16. Canedo, D.; Neves, A.J. Facial expression recognition using computer vision: A systematic review. *Applied Sciences* **2019**, *9*, 4678.
17. Tong, Y.; Liao, W.; Ji, Q. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on pattern analysis and machine intelligence* **2007**, *29*, 1683–1699.
18. Chiranjeevi, P.; Gopalakrishnan, V.; Moogi, P. Neutral Face Classification Using Personalized Appearance Models for Fast and Robust Emotion Detection. *IEEE Transactions on Image Processing* **2015**, *24*, 2701–2711. doi:10.1109/TIP.2015.2421437.
19. Yin, L.; Chen, X.; Sun, Y.; Worm, T.; Reale, M. A high-resolution 3D dynamic facial expression database **2008**. pp. 1–6. doi:10.1109/AFGR.2008.4813324.

20. Chen, J.; Xu, R.; Liu, L. Deep peak-neutral difference feature for facial expression recognition. *Multimedia Tools and Applications* **2018**, *77*. doi:10.1007/s11042-018-5909-5.
21. Li, Y.; Bolle, R.; Bolle, R. Automatic Neutral Face Detection Using Location and Shape. *RC Computer Science* **2001**, *22259*, 111–73.
22. Khaireddin, Y.; Chen, Z. Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588* **2021**.
23. Kamachi, M.; Lyons, M.; Gyoba, J. The japanese female facial expression (jaffe) database. *Availble: http://www. kasrl. org/jaffe. html* **1997**.
24. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with Gabor wavelets **1998**. pp. 200–205. doi:10.1109/AFGR.1998.670949.
25. Lyons, M. "Excavating AI" Re-Excavated: Debunking a Fallacious Account of the Jaffe Dataset. *SSRN Electronic Journal* **2021**. doi:10.2139/ssrn.3900990.
26. Xue, M.; Liu, W.; Li, L. Person-independent facial expression recognition via hierarchical classification. *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing* **2013**, pp. 449–454.
27. Toda, K.; Shinomiya, N. Machine learning-based fall detection system for the elderly using passive RFID sensor tags. *2019 13th International Conference on Sensing Technology (ICST)* **2019**, pp. 1–6.
28. Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* **2000**.
29. Meena, G.; Mohbey, K.; Kumar, S.; Chawda, R.; Gaikwad, S. Image-Based Sentiment Analysis Using InceptionV3 Transfer Learning Approach. *SN Computer Science* **2023**, *4*. doi:10.1007/s42979-023-01695-3.
30. Ali, L.; Alnajjar, F.; Jassmi, H.; Gochoo, M.; Khan, W.; Serhani, M. Performance Evaluation of Deep CNN-Based Crack Detection and Localization Techniques for Concrete Structures. *Sensors* **2021**, *21*, 1688. doi:10.3390/s21051688.
31. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv preprint arXiv:1312.4400* **2013**.
32. Javid, A.M.; Das, S.; Skoglund, M.; Chatterjee, S. A relu dense layer to improve the performance of neural networks **2021**. pp. 2810–2814.
33. Darapaneni, N.; Tanndalam, A.; Gupta, M.; Taneja, N.; Purushothaman, P.; Eswar, S.; Paduri, A.R.; Arichandrapandian, T. Banana Sub-Family Classification and Quality Prediction using Computer Vision. *arXiv preprint arXiv:2204.02581* **2022**.
34. Htay, M.; Sinha, P.G.; Htun, H.; Thwe, P. COMPARISON OF PREPROCESSING METHOD USED IN FACIAL EXPRESSION RECOGNITION. 2019.
35. Revina, I.; Emmanuel, W.S. A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University - Computer and Information Sciences* **2021**, *33*, 619–628. doi:https://doi.org/10.1016/j.jksuci.2018.09.002.