**Real-Time Facial Emotion Recognition System**Technical Report

Ricardo de Deijn[[1]](#footnote-2) · Greeshma Kalsham1 · Neha Sure1

**Abstract**

Real-time facial emotion detection is important due to its various applications in fields such as healthcare, customer service, and human-computer interaction. Using the FER2013 dataset, this study evaluates the effectiveness of various convolutional neural network (CNN) designs, including InceptionV3, ResNet50, and VGG19, in accurately predicting face emotions. The study required extensive data preparation, including data cleaning, augmentation, and normalization of all images to a consistent pixel size. The preprocessed data was then used to train and evaluate the selected CNN models with different epoch counts. The models' performance was evaluated based on their accuracy, efficiency, and capacity to tolerate faults in emotional recognition using facial expressions. The results proved the benefits and limits of each architecture, providing useful information for choosing the best model for real-time facial emotion identification tasks.

**Keywords**

Convolutional Neural Network (CNN), Emotions Detection, Image Classification, Computer Vision, Facial Expression Analysis.

Table of Contents

[1. Introduction [Neha & Ricardo] 1](#_Toc164885315)

[2. Literature Review 1](#_Toc164885316)

[Emotions [Ricardo] 1](#_Toc164885317)

[Facial Expressions [Neha] 2](#_Toc164885318)

[Methodologies Used in Literature [Greeshma] 3](#_Toc164885319)

[3. Methodology 5](#_Toc164885320)

[Dataset [Neha & Ricardo] 5](#_Toc164885321)

[Models 6](#_Toc164885322)

[VGG[Greeshma] 6](#_Toc164885323)

[Inceptionv3[Ricardo] 7](#_Toc164885335)

[ResNet [Neha] 8](#_Toc164885336)

[Implementation [Ricardo] 8](#_Toc164885337)

[4. Results [Ricardo] 9](#_Toc164885338)

[4.1 Real-time [Ricardo] 10](#_Toc164885339)

[5. Conclusion [Neha] 10](#_Toc164885340)

[6. Limitations and Future Work [Greeshma] 11](#_Toc164885341)

[References 11](#_Toc164885342)

# 1. Introduction [Neha & Ricardo]

The noticeable change in the face, that results from impulsive response, is known as facial expression. Automated facial expression is the process of identifying different facial expressions with artificial intelligence systems [1]. In nonverbal communication, facial expressions are extremely important since they provide an insight into someone's feelings and mental health. These involuntarily facial expressions can represent a variety of emotions, such as joy, sorrow, fear, anger, and more. Facial expressions are significant because they can convey emotions on their own and offer useful indicators for understanding and decoding social interactions. Due to its potential applications in domains including emotional well-being evaluation, cognitive computing, and human-computer interaction, facial expression recognition has grown in importance as a subject of study [2], [3].

A lot of work has gone into creating automatic facial expression analysis systems as they are useful in a variety of industries, including robotics, health sector with therapy, and customer services to name a few. This project aims to create an improved image classification model that is able to predict facial emotion recognition more correctly through finetuned convolutional neural network architectures InceptionV3, ResNet50, and VGG19 models with the FER2013 dataset provided by Google in 2013.

In the next chapters, we will review the current relevant research on our project topic. Besides this, we will discuss the data set and our experiment approach. After this, we will present our results and summarize our conclusions and provide limitations and recommendations for future work. This project research code can be found on GitHub: *https://github.com/Neatherblok/Facial\_Emotion\_Detec/*

# 2. Literature Review

Facial emotion recognition stands at the intersection of technology and human interaction which offers insights into non-verbal communication. This literature review is structured into three pivotal sections: emotion, facial expressions, and previous methodologies. Initially, it delves into a diverse range of emotions that can be detected by facial cues and subsequently examines how facial expressions manifest these emotions. Lastly, it assesses various previous methodologies that researchers have previously implemented. It also sets the stage for future advancements.

## Emotions [Ricardo]

There are six universal emotions, that regardless of culture, regions or ethnicity, are the same for all humans. These emotions include anger, disgust, fear, happiness, sadness / grief, surprise [4], [5], and some researchers might include contempt as a seventh universal emotion [5]. These emotions are of major influence on how humans act and behave [6], [7]. In fact, they have been of major influence on how humans have acted for the last two million years according to [7]. Emotions have helped humans survive as hunter-gatherers when we were not as evolved as the modern humans are. However, in our modern world, objectiveness and rational thinking is of high importance for humans to have. Emotions still impact our judgement and decision-making process a lot [8]. For example, it could result in unwanted bias in judgement while making informed decisions, or in some type of tunnel vision when a person is feeling frustrated. On the other hand, emotions can also play a complementary role to our rational thinking. Humans could also use emotions as a beneficial guide to a path with less risks or help to make big important ethical decisions which might be harder with a rational mindset [8].

There are distinct types of emotions, both positive and negative ones. Negative emotions, like anger, disgust, fear, and sadness will impact rational thinking a lot. This has a big influence on a lot of situations, for example when you are trying to do unbiased research, but also when trying to safely be driving a car [4]. On the other hand, positive emotions, like happiness and surprise can provide a lot of health benefits for a person in the long term [5] can also provoke a positive bias [8].

Emotions are often influenced by outside influences through stimulus, organism, and response (S-O-R model) [6]. Stimulus is about using stimuli in an environment [9], think of visual or auditory cues such as a beautiful sunset, cute dogs, or great music [6]. These are exploited by a lot of companies to influence decisions to buy a product, as a stimulus can activate something in us as organisms [9]. Organism in S-O-R is about the internal process of a human, where cognition, memory and knowledge come into play [6], [9]. Past experiences are used to determine if something is satisfying or deemed unpleasant [9]. After deeming a response from the organism on the stimuli, a response is given. Responses are significant changes after a response to an emotional stimulus. These reactions are normally cognitive and not well-thought through [9]. Minimizing these cognitive emotions can have a great positive effect on a person’s objectivity, and it is possible through training minimizations strategies [8].

There are different strategies that can be used to minimize emotional reactions. Strategies include using an opposing emotion to counteract one emotion; delaying a decision over a certain period of time, so a person can think more rationally about a decision; reappraising an emotional stimulus and think about it in a different light; or a person could crowd out someone’s emotions by giving it a wide variety of stimuli when presenting a decision, like a pros-cons list [8].

It is important to be aware of one’s emotions as they can be of major influence on one’s decisions and might result in wrong decisions or bad actions [4]. Computers have been helping us in detecting certain emotions by tracking our facial expressions, as these expressions are of high importance for humans to interpersonally interact with other humans [10].

## **Facial Expressions** [Neha]

Emotions are often expressed through facial characteristics. A recent study investigated the effects of facial asymmetries as an indication of age using the facial recognition technology (FERET) dataset [11]. Additional research has also been done on how behavior affects facial emotion recognition (FER), for example in people with autism spectrum disorders (ASD) [12]. To achieve successful social interactions, FER is found to be a critical component, and the difficulties that people with ASD encounter in this area are recognized [12]. Problems in recognizing and interpreting emotions are emphasized in the DSM-IV-TR (Diagnostic and Statistical Manual of Mental Disorders) diagnostic criteria for ASD, especially through nonverbal cues like facial expressions. Even so, the results on facial emotion recognition in autism spectrum disorders are characterized as mixed, with some studies showing normal FER and others showing notable abnormalities. [12] demonstrates that individuals with ASD exhibit abnormal eye gaze patterns, delayed event-related potential components, and abnormal activity in their emotion-processing circuitry, even though their behavioral performance during FER tasks remains unchanged.

Understanding facial expressions is crucial for computer and human communication in a variety of scenarios, including online gaming, displaying advertisements, assisting with medical treatment, receiving and customer feedback [13]. Strong emotion recognition accuracy has been attained in photographs taken in consistent and regulated contexts because of developments in computer vision.

Because of the significant intra-class variance and low inter-class variation, such as small variations in facial position and expressions, challenges in emotion recognition under naturalistic circumstances occur. FER2013 is an emotion recognition dataset that covers challenging naturalistic settings and obstacles. It was first presented in 2013 at the International Conference on Machine Learning (ICML), and quickly rose to the top when it came to model performance comparisons for emotion recognition. The estimate of human performance on this dataset is 65.5% [13].

Convolutional Neural Networks (CNNs) have demonstrated significant promise in image classification because of their ability to extract features and perform computation efficiently [14].

Facial muscles contract to produce facial expressions, which cause temporally deformed features such as the nose, lips, eye lids, eyebrows, and skin texture, which are frequently visible as wrinkles and bulges [15]. Muscular activity variations usually occur in brief periods of activity, usually lasting a few seconds, but rarely lasting longer than five seconds or less than 250 milliseconds. The placement, intensity, and dynamics of facial motions are significant factors. The geometric distortion of facial features or the density of wrinkles that occur in specific face regions can be used to evaluate the intensities of facial expressions.

## Methodologies Used in Literature [Greeshma]

From all the techniques used for FER, convolutional neural networks are found to show great results due to their efficiency and its feature extraction process. In [13] FER2013 is used by the VGG network to thoroughly tune the models’ hyperparameters. Data augmentation is applied to the training data by rescaling, shifting the images both horizontally and vertically by ±20% and by rotating it by ±10 degrees. After which the image is cropped to 255 pixels each. The experiments are run with 300 epochs and are evaluated, tested using validation accuracy and standard ten-crop averaging. Saliency map is used to understand the importance of every single feature in the image and to classify the emotion. The best results are achieved by reloading the best model and tuning for 50 epochs by using cosine annealing schedulers and achieved an accuracy of 73.28% while the estimated human performance is only 65.5% [13] . While testing on FER2013, the confusion matrix shows great results and classification on emotions like “happiness” and “surprise” while “fear” and “disgust” show lower accuracy. Additionally, among all the 6 optimizer algorithms used in [13], the combination of Stochastic Gradient Descent (SGD) with Nesterov momentum as the optimizer and Reduce Learning Rate on Plateau (RLRP) as the learning rate scheduler, produced the highest validation and testing accuracies. The suggested enhancement is to use various other deep learning techniques to improve the overall performance.

In [16], the techniques used for face recognition are ResNet, AlexNet and VGG16. It also focuses on both geometric and statistical features. It uses a four-step process including template library, feature extraction and comparison. This not only increases efficiency but also reduces the running time. AlexNet, with 5-layer convolutional layers and three full layers, is used on the ImageNet dataset. ReLU is one of the most used activation functions for neural networks. AlexNet also uses ReLU as its convolutional activation function but uses the SoftMax layer as an activation function to predict the probability of emotion in seven classes. Five different epochs are used, and the best accuracy is 0.6424 with AlexNet. The accuracy of all the four models is compared and the average was between 0.55 to 0.6. [16] also mentions that there might be a chance of overfitting due to the size of the dataset, which might affect the accuracy.

Another effective method was developed by using CNN and VIT for both images and videos. The combination of CNN-10 and data augmentation like image rotation, scaling, flipping are used which enhanced the model's robustness by achieving high accuracy of 99.9% for Cohn-Kanade+ (CK+),84.3% for FER-2013 and 95.4% for Japanese Female Facial Expression Database (JAFFE) datasets [5]. The data is augmented by cropping to 75x75 and dividing the dataset into 64% training, 20% validation and 16% testing. After which cross validation is used to assess the performance of the models. The proposed model has been compared with GG19 and Inceptionv3 as it known for its accuracy and efficiency and its performance was conducted using metrics including accuracy, confusion matrix, macro average, precision, recall, F1-score and weighted average precision [5]. One of the limitations was the shortage of images of certain motions like disgust. Future studies will focus on selecting facial features through transfer learning and depicting mental health and health status applications [5].

The FER2013 dataset is used in smart car application experiments where the seven main emotions of the driver are identified using CNN-based model. Preprocessing was rigorously implemented which included correction of labels, cleaning the unnecessary images using crowd sourcing and converting the original size of the 1D image from 48x48 pixels to 2D array and is reshaped to (48,48,1). As FER2013 is an unbalanced dataset, balancing techniques such as over sampling, under sampling and hybrid methods are used and a new dataset is created, data augmentation and finally training the CNN model on the dataset [4]. Under over sampling, SMOTE is used which addresses the class imbalance by increasing minority class samples. The Neighborhood Cleaning Rule (NCR), ADASYN are some of the under-sampling methods used. For Hybrid type a combination of SMOTETomek and SMOTE-ENN is used to balance the data. After which, the parameters such as width, height, shift and rotation of the data set are modified and augmented. It was also observed that in the proposed model has effective results when compared to the pre-trained models with the accuracy, precision, F1, score and recall of 98.41%, 98%, 94% and 96% respectively [4].

From [17] Extraction of facial features is an important step for emotion recognition. To extract these features the placement and the change in the expression is analyzed. The proposed model is classified into Pre-processing, Face and Facial feature extraction and Emotion Classification. In preprocessing, the clarity and the scaling of the image is modified whereas in facial feature extraction methods such as Gabor filter are used which help in identifying the geometry of the features in the image. Euclidean and Hausdorff distance are used to classify emotions. Different methodologies such as Geometric based, Temporal and Appearance based are used while concentrating on eyebrows, nose and mouth to classify the emotion. The accuracy was found to be 88.9% after extracting features from frontal images and yaw angles from the non-frontal images.

Real time facial emotion detection using AI has been quite famous and [18] uses CNN model and FER 2013 and OpenCV for detecting emotions. The images are pre-processed by resizing and augmented to increase the data size for better accuracy and are trained using CNN model where SoftMax function is used for classifying the images into its various classes of emotions. The model is saved with .h5 file and then the image is captured from the camera while using Viola-Jones algorithm. The proposed CNN model is imported from Keras library and the image with shape (48,48,1). Maxpooling layers and ReLU are used and after 100 epoch the accuracy exceeded that of the pretrained models. When the model is trained with Inceptionv3 and VGG16 the accuracy obtained was 45.60% and 62.47% respectively while the proposed CNN model has an accuracy of 90.4%. Future work is to work on individual student emotion analysis using hardware like Raspberry Pi and integrate with a real time camera.

A framework has been proposed where models developed can be compared with one another for FER. AffectNet dataset is used, and CNN is used for training. Additionally, a web-based application has been developed and a standard model to evaluate deep learning models for real world applications. AffectNet is found to be one of the best data sets due to its high quality and quantity with RGB channels [19]. To achieve a balance between time and accuracy a light weight AlexNet variant CNN is implemented. The CNN model is embedded into the proposed application, and it has the flexibility that any fully trained model built in tensor flow can be converted and embedded well into the application while achieving an accuracy of 55.09% on AffectNet [19]. Although, the trained models such as VGGNet and MobileNet have achieved an accuracy of 58%. One of the flaws would be the imbalance of the class frequency. However, Class weighting helped in mitigating frequency bias by a 8-10% increase. The future work will be about utilizing the web framework to test alternative FER models. It also implies that as the application matures, it will have the potential to contribute to the field by generating valuable datasets through user interactions.

A real-time facial emotion recognition system is designed with optimal complexity and time. The dataset is used from Kaggle and is divided into seven classes and normalized by taking the mean and augmented by flipping the images horizontally from the training set. Initially VGG and ResNet50 are used to classify the images. Later, a new model was implemented to get better results. The new CNN model is implemented in a way that more or fewer layers can be added. The three important aspects while training a CNN model are LR Scheduler which are used mostly for object detection and NLP while ensuring the model converges to get the optimal solution. Fine-tuning is employed to adjust the final weights of the pre-trained model to improve accuracy. This is done by unfreezing some or few layers of the pretrained model and using a lower learning rate than the initial training [20]. To evaluate the model accuracy of all the classes a confusion matrix is used, and the results show high accuracy for few emotions and low accuracy for “disgust” and “fear” as they have smaller number of samples. A comparison of other alternate pretrained models is used and the proposed model resulted in an higher accuracy of 74.57% [20]. The future work would be to analyze the expression of the individual and the truth by using pulse measuring devices.

In [21], an analysis of the available pretrained CNN models is done. Future scope would be usage of some of the other statistical and data processing techniques to analyze the feature emotions. Additionally, work on negative emotions and differentiating with neutral ones will be implemented.

# 3. Methodology

## Dataset [Neha & Ricardo]

The [FER2013](https://www.kaggle.com/datasets/msambare/fer2013) (Facial Expression Recognition 2013) dataset includes images and categories that indicate the person's emotions. The 48 by 48-pixel grayscale images in the collection depict seven distinct emotions: angry, disgust, fear, happy, sad, surprised, and neutral. The dataset contains 28709 examples in the training set, 7178 examples in the testing set. It contains a total of 35887 images. The FER-2013 dataset contains a few drawbacks, including noisy photos, incorrect labeling, improper cropping, non-face images, and imbalanced classes. The most occurring

Before training any model, the noisy images were first removed, and the incorrect labeled images were corrected to their right label. After this, the image pixels were normalized to a [0,1] range and further normalized to the normalization based on mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], which were the parameters for the ImageNet-1K dataset which was used by the three pre-trained models this experiment will use for classification [22].

## Models

### **VGG[Greeshma]**

VGG stands for Visual Geometry Group which is also known as VGGNet is developed by Andrew Zisserman and Karen Simonyan in 2013 at Oxford to increase the overall model performance of CNN. It is a classical CNN architecture. It has multiple layers and is one of the most popular image recognition architectures. It is also considered as the best model for feature extraction [23]. VGG16 and VGG19 are two commonly used models for Facial Emotion Recognition. The numbers 16 and 19 refer to the number of layers used. It is constructed using small convolutional filters as it improves the decision functions. It consists of 4 components:

Input: ImageNet dataset is used for VGG and the creators of VGG standardized the input size to 224x224 to maintain consistency.

#### *Convolutional Layers*

These layers use a small 3X3 receptive field. It captures both the horizontal and vertical features of the input data despite the small size. Besides that, they also employ 1x1 filters which increases the network power and reduces the computational cost. It uses ReLU a rectified linear unit which is an activation function and is used to improve training time. Additionally, they use fixed convolution stride of size 1 pixel which ensures that the fin-grained features are retained in the output [24].

#### *Hidden Layers*

The hidden layers use ReLU similar to AlexNet. They do not use Leverage Local Response Normalization which is usually used to improve memory consumption and training time. This layer is not involved in improving accuracy.

#### *Fully Connected Layers*

It consists of 3 fully connected layers, where the first two layers of 4096 channels each and the final layer has 1000 channels.

#### *VGG19*

As per the name it refers to 19 deep neural network layers out of which 3 are fully connected layers. It is extensive and has around 138 million parameters. However, its architecture is quite simple and uniform. Every few convolution layers are followed by pooling layer. In this first set of convolutional layers, 64 filters are used, and it is doubled to 128 and 256 in the next set of layers. The last layer uses 512 filters. The limitation of VGG is that it trains slowly and takes huge disk space as ImageNet weights are close to 528 MB. In [24], a modified VGG is proposed. A small 3\*3 convolutional filter which extracts minute details from the training data is used and the max pooling layer is removed. This has led to decreasing training loss.

#### *Limitations*

The model is very slow to train. It also takes a lot of disk space and bandwidth. Despite having numerous diverse applications and solving complicated problems, VGG suffers from computational cost as they have multiple convolutional layers and will lead to inefficient hardware implementation. It also suffers from high latency. According to [25], the newly proposed VGG can process convolutional layers with an increase of 9.5 times faster than state of art accelerators. In [26], it was mentioned that VGG has a localization problem and an extra layer of VGG network is added which led to improving accuracy.

Difference between VGG16 and Resnet50: VGG 16 was found in the year 2014 and consists of 138M parameters and has 16 convolutional layers while achieving an accuracy of 92.07% on ImageNet dataset. However, Resnet-50 was found in the year with 25.6M parameters and consists of 50 convolutional layers with residual connections and achieve an accuracy of 93.30% on the ImageNet dataset. Mostly, Resnet50 outperforms VGG in terms of accuracy especially when dealing with huge datasets. In terms of architecture Resnet50 uses residual connections also known as skip connections which allows the model to learn with reference to layer inputs [27]. VGGNet is used in those image applications which require deeper network whereas Resnet is used for much extreme deep structures and in applications where object detection and picture segmentation is required. From [27], the accuracy achieved with VGG16 and Resnet is 59.2% and 65.1% respectively.

### **Inceptionv3[Ricardo]**

Inceptionv3, is proposed in 2016 by researchers from Google at the Conference on Computer Vision and Pattern Recognition (CVPR). It is a deep learning algorithm that can be used for image classification, object detection [28] or feature extraction [29]. It uses previous image classification approaches such as VGG, GoogLeNet and AlexNet as inspiration for its architecture. However, this architecture addresses limitations such as early network bottlenecks such as hard data compressions where data gets drastically reduced in size at once; Increasing feature disentanglement through larger convolutions kernels; adding spatial aggregation over low-dimensional embeddings to improve faster learning without a loss of model performance and balancing the network width and height to improve the network performance. It does this by utilizing convolutional, pooling, concat, and fully connected layers [28].

#### Early network bottlenecks

Commonly, in a convolutional network, an activation layer is followed by a pooling layer to decrease the feature map size of the image. However, in the Inceptionv3 architecture, two different stride blocks are used on a similar feature map. One uses a convolutional function with a 2x2 stride that reduces the feature map by half, and one that uses a maximum or average pooling with a 2x2 stride that has the same effect. The two filter maps are then concatenated into one shared filter bank that contains a high representation of the image with a significant drop in computational cost [28].

#### Feature Disentanglement Increasement

Large kernels used in convolutional networks are usually more expensive than smaller convolutional kernels such as 3x3. Normally, a 5x5 kernel is able to capture more correlations between neighboring pixels, but the cost outweighs the gains. To still give more weight to potentially important missed relationships, the Inceptionv3 network uses two 3x3 layers with a convolutional and fully connected layer. This setup is achieving similar model performance but is less computationally expensive [28].

#### Spatial Aggregation

Besides using two 3x3 kernels, the researchers in [28] have concluded that it is also possible to replace these n x 1 kernels with n x 1 kernels asymmetrically. This means that in the system, a 3 x 1 convolution can be used, as long as it is followed by a 1 x 3 kernel. This principle is able to capture a similar receptive field as the 3x3 kernel when concatenated together. However, this principle is way more computationally cheaper, and becomes even larger the more n grows. This results from the fact that the system has an easier time doing matrix calculations on two vectors instead of two matrices. And although a little information might be missing, this information is significantly small so that the performance does not reduce. Because this technique is cheaper, this promises that a convolutional kernel could potentially be replaced with larger sized convolutional kernels to capture even more related connections [28].

#### Balancing Network

According to [28], auxiliary classifiers have been used to improve the performance of very deep neural networks. It was originally thought that this happened because the auxiliary classifiers pushed useful gradients to the first couple of layers, making immediate changes to the model as it was able to combat vanishing gradients at the start. However, it was found by these researchers that these auxiliary classifiers didn’t improve convergence early in training and that it was used more as a regularizer. It was not a very good regularizer for the Inceptionv3 network, and it was determined that the network would be better balanced out if different regularizers like batch normalization would be used [28].

Eventually, the InceptionV3 uses a multitude of convolutional, pooling, Inception modules, linear and SoftMax layers. The Inception modules are sections where multiple convolutional layers create feature maps, and their outcomes are eventually concatenated together as described using the feature disentanglement and spatial aggregation method. This method results in the model performing well on lower resolution input and is able to perform a 18.77% top-1 error on the ILSVRC 2012 classification benchmark. Which is better than for example compared to VGG which achieved a score of 24.4% [28].

### ResNet [Neha]

ResNet, which stands for Residual Network, is a special kind of deep learning system developed by Microsoft researchers in 2015. It is well known for its exceptional ability to train deep networks. It resolves a significant issue with "vanishing gradients," which made it hard to train deep networks before ResNet [30]. One particular version of the ResNet architecture is called ResNet-50. There are 50 layers total, comprising skip connections, fully connected layers, pooling layers, and convolutional layers. In a number of computer vision tasks, including object identification and image classification ResNet-50 has demonstrated better performance.

#### ResNet Blocks

The residual block is the fundamental ResNet construction block. Every residual block comprises two convolutional layers using ReLU activation algorithms and batch normalization, in addition to a shortcut connection. The input is added to the two convolutional layers' outputs using these shortcut connections [30]. ResNet-50 uses residual connections to increase computational efficiency even if it is deeper than previous architectures like VGGNet [31]. This indicates that ResNet-50 consumes less processing power and trains more quickly even with more layers.  ResNet-50 are frequently used for transfer learning because of their effectiveness. Transfer learning is the process of optimizing a pre-trained model for a particular task or dataset. This can result in improved performance and faster convergence, particularly when there is limited data.

## Implementation [Ricardo]

This experiment will be using these three model architectures, as they have been shown to be very good at extracting features and more particularly facial features. The implementation of these models is done through PyTorch [32] and will be using model finetuning and transfer learning to make InceptionV3, Resnet50, and VGG19 generalize the concepts of our relatively small data quicker and better. Finetuning will help for small data set samples, as the models’ configuration and weights to extract features has been configured on the larger ImageNet-1K dataset [33].

# 4. Results [Ricardo]

Three models' architectures have been trained under different epoch counts and learning rate settings. Each model has been trained once in 25 epochs. However, due to the complexity of the InceptionV3 model, this model performed significantly better results when the epoch count has been increased to 90 epochs. A comparison in this epoch count can be found in Table 1.

Results improved even more after a learning rate step scheduler (LR Step) got introduced. This scheduler had the purpose of reducing the learning rate after each epoch to optimally optimize the model to the smallest possible prediction error. This improvement is especially visible for the VGG19 model and ResNet50 models, in which the LR Step scheduler got introduced to the training session with 20 and 25 epochs, respectively. The increase in accuracy shows that this technique is a powerful technique, which eventually even lead to the fact that the ResNet50 model with 20 epochs outperformed a usually powerful model such as InceptionV3.

Table 1. Summary of training results under different training settings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy Train** | **Accuracy Test** | **Loss Train** | **Loss Validation** |
| **InceptionV3**  - 25 epochs | 37.49% | 15.48% | 2.03 | / |
| **InceptionV3**  - 90 epochs - LR step with 0.90 momentum | 40.48% | 38.87% | 1.66 | 1.55 |
| **InceptionV3**  - 90 epochs  - LR step with 0.85 momentum | 41.75% | 13.54% | 1.62 | 1.54 |
| **ResNet50**  - 10 epochs | 14.09% | 14.01% | 46.74 | 46.66 |
| **ResNet50**  - 20 epochs  - LR step with 0.90 momentum | 47.68% | 44.11% | 1.41 | 1.43 |
| **ResNet50**  - 25 epochs  - LR Step with 0.85 momentum | 46.58% | 42.51% | 1.81 | 1.99 |
| **VGG19**  - 15 epochs | 41.35% | 39.57% | 28.25 | 25.18 |
| **VGG19**  - 20 epochs | 39.50% | 37.16% | 25.82 | 25.82 |
| **VGG19**  - 25 epochs  - LR Step with 0.85 momentum | 42.39% | 40.02% | 24.67 | 24.67 |

However, this ResNet50 model still showed significant limitations as well. The accuracy of 44.11% was not as equally distributed over all classes as would be the case in an ideal situation. As a matter of fact, the most common occurrence in the model’s prediction was a negative classification for all classes, as shown in table 2. An exception to this is the class ‘Happy’, which predicts a higher percentage of true and false positive values than negative predictions.

Table 2. Results of the best performing ResNet50 model (20 epochs; LR step with 0.9 momentum) per class.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **True Positives (%)** | **False Positives (%)** | **False Negatives (%)** | **True Negatives (%)** |
| Anger | 5.43 | 33.64 | 94.57 | 66.36 |
| Disgust | 0.00 | 43.77 | 100.00 | 56.23 |
| Fear | 12.89 | 43.11 | 87.11 | 56.89 |
| Happy | 52.54 | 60.14 | 47.47 | 39.86 |
| Neutral | 11.46 | 37.62 | 88.54 | 62.38 |
| Sad | 1.66 | 5.28 | 98.34 | 94.72 |
| Surprise | 6.67 | 36.85 | 93.33 | 63.15 |

This could be a direct result of the class imbalance in the FER2013 dataset, which was insufficiently addressed in the data preparation. This resulted that the model had the opportunity to train on more instances where a person is happy, as compared to a person who has a disgusted facial expression. This unequal distribution led to a heavily biased model that is more prone on accurately prediction happy facial expressions and correctly predicting when a facial expression doesn’t show a certain class (true negatives), while also predicting false negatives more often.

## 4.1 Real-time [Ricardo]

Even though some classes are performing insufficiently, this project moved forward in creating a real-time classification program that was able to accurately tell if a person has a happy facial expression or not. This was shown to not be a big problem, as true positives for happiness and true negatives for the other classes are quite high in table 2.

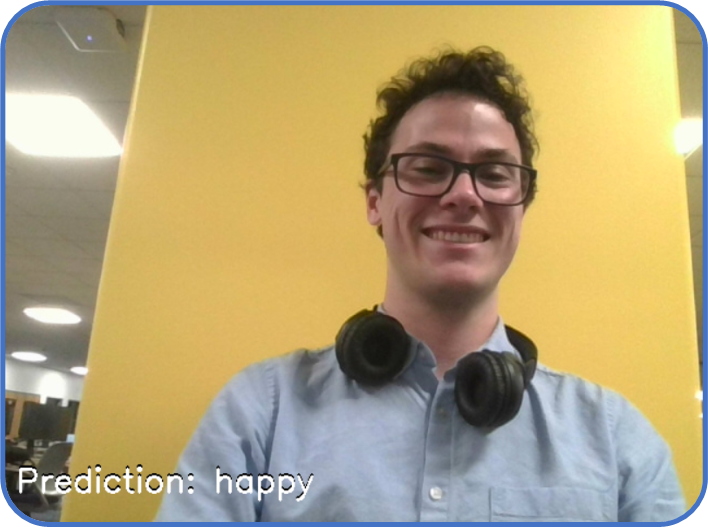


Figure 1. A screenshot of the facial emotion classification program

This program got created in Python, using packages such as CV2 which enabled the Python code to capture a live webcam feed. The frames which got captured were augmented to make sure that they follow a standard format that the model expects based on the training data. After preparing the frame, it gets pushed to the model and a prediction returns. This prediction gets captured and shown on the screen.

The prediction process, from capturing to prediction, takes on average about **400** milliseconds. This means that our project is able to predict if a person is happy in near real-time.

# 5. Conclusion [Neha]

The project started with properly preparing data and training models to produce accurate results. This includes cleaning our dataset to assure that it was accurate, which includes not only finding and correcting the errors, but also increasing the dataset's variety using approaches such as data augmentation. By increasing the dataset in this way, we hoped to present the models with a broader variety of examples to learn from. Moreover, we implemented methods such as normalization to ensure that the data was consistent.

For model training, three common models were ResNet50, VGG19, and InceptionV3. Each model was trained throughout multiple epochs, or iterations of the training process, to assess its performance under different conditions. It was observed that ResNet50 outperformed VGG19 and InceptionV3 in terms of predicting best accuracy at epoch 20 with a learning rate step scheduler set on momentum 0.9. The resulting model was able to achieve an accuracy on the unseen data of 44.11%, which is significantly higher than the maximum accuracy for InceptionV3 and VGG19.

Overall, the project highlights the need for data preparation and careful model selection in generating accurate results. By focusing on these essential characteristics, we hope to contribute to the evolution of best practices in the field, particularly for practical applications across multiple domains.

# 6. Limitations and Future Work [Greeshma]

The project encountered several limitations during the initial phase of development. The primary constraint was the utilization of a small grayscale image dataset for facial emotion recognition, which inherently limits the diversity and variance within the training data, potentially leading to overfitting and reduced generalizability of the model. Additionally, the grayscale nature of the dataset precluded the use of color information, which could be significant in the accurate detection of emotional states. Computational resources were also limited, constraining the complexity and scalability of the models that could be trained and evaluated within our project scope. The dataset also encountered a significant limitation due to the class imbalance within the dataset. This manifested in the varying degree across different emotional classes, leading to disparities in model performance. “Happy” exhibited a higher frequency of occurrence compared to others like “Disgust”. This disparity in class distribution can severely impact the model's ability to learn and generalize effectively across all emotional states. The metrics, including true positives, false positives, false negatives and true negatives showcased considerable variables among the emotional categories. This leads to not achieving uniform accuracy and reliability across all classes within the dataset.

Future work is aimed at mitigating these limitations to enhance the robustness and accuracy of the emotion recognition system. Expanding the dataset is of utmost priority, introducing a larger and more varied set of images to improve the model's ability to generalize across different facial expressions. There is also an intention to explore the efficacy of leveraging other pre-trained models. This would enable the application of deep learning architectures proven to be successful in similar tasks. Furthermore, building models from scratch is anticipated, fostering the development of bespoke solutions tailored to the specific nuances of emotion recognition in grayscale imagery. These steps will be instrumental in advancing the project towards a more sophisticated and reliable facial emotion recognition system. Mitigating the imbalance within the dataset stands as a critical objective for future work. To improve the model’s generalizability and performance consistency across all emotional classes, future endeavors will focus on refining the training strategies. Additionally, we might transition from grayscale to RGB datasets which may provide extra discriminative features for the enhancement of emotion recognition accuracy.

# References

[1] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo, and O. O. Oluwagbemi, “Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10),” *Applied Computational Intelligence and Soft Computing*, vol. 2023, 2023, doi: 10.1155/2023/2457898.

[2] B. Fasel and J. Luettin, “Automatic facial expression analysis: a survey,” 2003. [Online]. Available: www.elsevier.com/locate/patcog

[3] Y. Khaireddin and Z. Chen, “Facial Emotion Recognition: State of the Art Performance on FER2013.”

[4] A. Soultana, F. Benabbou, and N. Sael, “Enhanced CNN-Based Model for Facial Emotions Recognition in Smart Car Applications,” *Arab J Sci Eng*, 2023, doi: 10.1007/s13369-023-08507-z.

[5] E. G. Dada, D. O. Oyewola, S. B. Joseph, O. Emebo, and O. O. Oluwagbemi, “Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10),” *Applied Computational Intelligence and Soft Computing*, vol. 2023, 2023, doi: 10.1155/2023/2457898.

[6] H. Yang, Y. Fan, G. Lv, S. Liu, and Z. Guo, “Exploiting emotional concepts for image emotion recognition,” *Visual Computer*, vol. 39, no. 5, pp. 2177–2190, May 2023, doi: 10.1007/s00371-022-02472-8.

[7] J. T. Cacioppo and W. L. Gardner, “Emotion,” *Annu Rev Psychol*, vol. 50, no. 1, pp. 191–214, 1999.

[8] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, “Emotion and Decision Making,” *Annu Rev Psychol*, vol. 66, pp. 799–823, Jan. 2015, doi: 10.1146/annurev-psych-010213-115043.

[9] M. T. Goi, V. Kalidas, and N. Yunus, “Mediating roles of emotion and experience in the stimulus-organism-response framework in higher education institutions,” *Journal of Marketing for Higher Education*, vol. 28, no. 1, pp. 90–112, Jan. 2018, doi: 10.1080/08841241.2018.1425231.

[10] J. Pao, “Emotion Detection Through Facial Feature Recognition,” 2016.

[11] M. Sajid *et al.*, “The impact of asymmetric left and asymmetric right face images on accurate age estimation,” *Math Probl Eng*, vol. 2019, 2019, doi: 10.1155/2019/8041413.

[12] M. B. Harms, A. Martin, and G. L. Wallace, “Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies,” *Neuropsychology Review*, vol. 20, no. 3. pp. 290–322, Sep. 2010. doi: 10.1007/s11065-010-9138-6.

[13] Y. Khaireddin and Z. Chen, “Facial Emotion Recognition: State of the Art Performance on FER2013,” 2021.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017, doi: 10.1145/3065386.

[15] B. Fasel and J. Luettin, “Automatic facial expression analysis: a survey,” 2003. [Online]. Available: www.elsevier.com/locate/patcog

[16] Y. Gan, “Facial expression recognition using convolutional neural network,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2018. doi: 10.1145/3271553.3271584.

[17] V. Upadhyay and D. Kotak, “A Review on Different Facial Feature Extraction Methods for Face Emotions Recognition System,” 2020.

[18] T. Kishore Kumar and D. S. Tummala, “Artificial Intelligence-Based Real-Time Facial Emotion Monitoring System,” in *Proceedings of the 9th International Conference on Computer and Communication Engineering, ICCCE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 361–366. doi: 10.1109/ICCCE58854.2023.10246038.

[19] N. Siddiqui, R. Dave, T. Bauer, T. Reither, D. Black, and M. Hanson, “A Robust Framework for Deep Learning Approaches to Facial Emotion Recognition and Evaluation,” in *Asia Conference on Algorithms, Computer and Machine Learning*, IEEE, 2022.

[20] A. Nautiyal, D. K. Bhardwaj, R. Narula, and H. Singh, “Real Time Emotion Recognition Using Image Classification,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2023, pp. 8–12. doi: 10.1145/3607947.3608295.

[21] A. Atanassov and D. Pilev, “Pre-trained Deep Learning Models for Facial Emotions Recognition,” in *International Conference Automatics and Informatics (ICAI)*, Varna: IEEE, Oct. 2020.

[22] Torch Contributors, “Models and Pre-Trained Weights,” PyTorch. Accessed: Feb. 23, 2024. [Online]. Available: https://pytorch.org/vision/stable/models.html

[23] S. Tammina, “Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, p. p9420, Oct. 2019, doi: 10.29322/IJSRP.9.10.2019.p9420.

[24] M. F. Haque, H.-Y. Lim, and D.-S. Kang, “Object Detection Based on VGG with ResNet Network,” in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, IEEE, Jan. 2019, pp. 1–3. doi: 10.23919/ELINFOCOM.2019.8706476.

[25] A. Ardakani, C. Condo, M. Ahmadi, and W. J. Gross, “An Architecture to Accelerate Convolution in Deep Neural Networks,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 4, pp. 1349–1362, Apr. 2018, doi: 10.1109/TCSI.2017.2757036.

[26] M. F. Haque, H.-Y. Lim, and D.-S. Kang, “Object Detection Based on VGG with ResNet Network,” in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, IEEE, Jan. 2019, pp. 1–3. doi: 10.23919/ELINFOCOM.2019.8706476.

[27] P. Dhankhar, “ResNet-50 and VGG-16 for recognizing Facial Emotions,” *International Journal of Innovations in Engineering and Technology*, 2019, doi: 10.21172/ijiet.134.18.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[29] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, “The Role of ImageNet Classes in Fréchet Inception Distance,” Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.06026

[30] S. Targ, D. Almeida, and K. Lyman, “Resnet in Resnet: Generalizing Residual Architectures,” Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.08029

[31] P. Dhankhar, “ResNet-50 and VGG-16 for recognizing Facial Emotions,” *International Journal of Innovations in Engineering and Technology*, 2019, doi: 10.21172/ijiet.134.18.

[32] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” in *Proceedings of the 33rd International Conference of Neural Information Processing Systems*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, and E. B. Fox, Eds., Vancouver: ACM, Dec. 2019, pp. 8026–8037. doi: 10.5555/3454287.3455008.

[33] M. F. Mohamedon, F. Abd Rahman, S. Y. Mohamad, and O. Omran Khalifa, “Banana Ripeness Classification Using Computer Vision-based Mobile Application,” in *Proceedings of the 8th International Conference on Computer and Communication Engineering, ICCCE 2021*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021, pp. 335–338. doi: 10.1109/ICCCE50029.2021.9467225.

1. Department of Computer Information Science  
   Minnesota State University, Mankato  
   United States of America [↑](#footnote-ref-2)