# FINAL REPORT

**Mark Qi**
Student# 1006764645
mark.qi@mail.utoronto.ca

**Richard Zhao**
Student# 1006750614
richardyz.zhao@mail.utoronto.ca

**Yulang Luo**
Student# 1005843740
yulang.luo@mail.utoronto.ca

**Linda Wang**
Student# 1007029270
lindaaa.wang@mail.utoronto.ca

## ABSTRACT

This is Team 40's Final Report. This document will present our final deliverable, including our project's background, data processing, primary model and baseline model's performances and design. The team will also discuss the results and analyze project's quality in details.

—-Total Pages: 8

## 1 INTRODUCTION

The team created a facial expression detection computer program that takes in face images as input and outputs the expression of that face, such as angry, happy, sad, etc. Detecting and classifying facial expressions is useful because it could help people suffering from Autism Spectrum Disorder correctly identify different people's emotions in online settings. Autism Spectrum Disorder patients are likely to suffer from deficits in expression recognition. The ability to interpret other people's emotions from their facial expressions could be crucial for understanding other people's feelings and for developing empathy and many aspects of social communication.

This project allows the group to implement a machine learning model to tackle this problem. Deep learning is a perfect method to solve our proposed problem because traditional hard coding can not account for the large sample space of human faces, such as lighting, skin tone, angles, and obstructions. In addition, humans have 43 muscles in the facial region Scheve (2021); the huge number of combinations further increases the difficulty of hard coding since slight muscle movements could result in a completely different facial meaning, making it hard to construct a program with manual tuning. In contrast, deep learning is capable of doing classification problems that require high precision after it has learned from a large data set, which can be obtained from the internet relatively easily. 1

## 2 BACKGROUND AND RELATED WORK

Facial expression recognition software is used commonly for research and medical purposes; it is particularly important in research fields such as psychology, affective neuroscience, and political science. There are many existing products in the facial emotion recognition field, and we will introduce five of them in this section to provide context. Küntzler et al. (2021)

FaceReader, an emotion analysis software created by Noldus, is one of the existing facial recognition software similar to our project. FaceReader can analyze facial expressions in live streams, videos, recordings, and images to determine the users' emotions. In addition, FaceReader can adjust the measure matrices based on the ages and genders of the test participants. According to Noldus' website, FaceReader was implemented using a deep learning-based face finding algorithm, an accurate face model and an artificial neural network. Based on this information, the team thinks it's reason-


Figure 1: Identify Facial Expression
Guillou (2018)

able to use a deep learning approach in our model and try out different modern CNN architectures to find the best model for our project. Noldus (2021)

ML Kit, a mobile SDK that brings Google's machine learning expertise to mobile apps, also has a face recognition feature. Similar to FaceReader, it can also recognize facial features and analyze emotions from pictures and videos. It can capture face contours and detect face landmarks, which are points of interest in a face (i.e. eyes, nose, mouth). The face detection API can classify emotions by detecting the angle of landmarks, but some classification only works for the frontal face. We encountered similar difficulties in our model, too; our model can predict frontal faces accurately but struggle to reach high accuracy on faces with obstructions.google (2021)

iMotions Facial Expression Analysis (FEA) Module is another famous engine in the field. The module provides 20 action units (e.g. chin raiser, lip stretcher), seven core emotions, facial landmarks, and behavioural indices such as head orientation. The most significant difference between the FEA module and our project is that this product uses a facial action coding system (FACS) instead of the deep learning approach to measure emotions. FACS pinpoints the exact action units that trigger a certain emotion and classify emotions based on these action units. However, based on our research, the FACS approach is too time-consuming with respect to both application and analysis compared to the deep learning approach. imo (2021)

Face API is a face detection AI developed by Microsoft Azure; this API embeds facial recognition into apps. It can identify face location, facial landmarks, and face attributes, including ages, gender, emotions and more, in images and videos. This API can accomplish many more tasks than our project; facial expression recognition is only one small part of its features. fac (b)

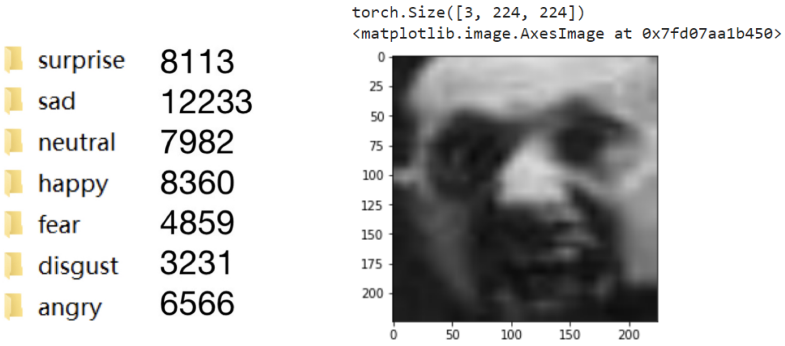
Another product similar to Face API is Face++, developed by Megvii. It can also detect face landmarks and many face attributes, including emotion recognition. Face++ can classify emotions into seven categories: angry, happy, sad, surprise, neutral, fear, and disgust; it can output the possibility of each emotion as a percentage. Both Face API and Face ++ use the deep learning approach; hence, we believe that the deep learning approach is the most efficient method for this problem. fac (a)

Overall, there are many existing successful facial emotion recognition products; we want to implement a model that works well on new data and has similar functionalities to those successful products. Based on the information about these products, we decided to use CNN architecture as our model.

## 3 Data Processing

The group collected 127680 labelled images in total. These images were mainly obtained from the MMA Facial Expression Dataset from Kaggle, which was already split into training, validation, and

testing data. The group has found an imbalance in the amount of data in each category of facial expressions. The most abundant one is ”neutral,” with 29,000 more, and the least is ”disgust,” with about 3,000. To ensure the network is not learning the imbalance of the data and to reduce the data to a more manageable amount, the group decided to use only a section of the more abundant category (e.g. only kept 7982 images of the ”neutral” class in the training data ). After this cleaning process, the number of training images was reduced to 51332, and the number of both validation and testing images was reduced to 10956. Figure 32 below shows the number of samples in each class in the training data. The size of these raw images were all [3, 48, 48], and the team adjusted the resolution of images based on which transfer learning model to use. For example, when we worked on transfer learning with ResNet, we adjusted the size to [3, 224, 224]. One sample of the cleaned images is shown in the figure 2 below.




(a) Samples in Each Class (b) sample image from data-set label: fear

Figure 2: Data Processing

The baseline model is trained on the raw dataset. However, we augmented our training data to improve accuracy while maintaining balance among categories in the training phase. Data augmentation can avoid overfitting the model and make sure that our model performs well on new data. We applied random horizontal flip 3, random rotation between -35 and 35 degrees, random affine 4, and the addition of Gaussian blur in combination; we obtained a total of 6000 augmented images and concatenated them with the training data. Additionally, we collected, cleaned, labelled, and edited pictures of our facial expressions and added these images to our testing data, which helped reflect the generalization of our model. While training the primary model, the group saved checkpoints and used the stored weights to compute the test accuracy.

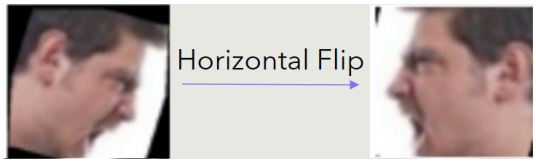



Figure 3: Random Horizontal Flip


Figure 4: Random Affine

# 4 ARCHITECTURE

The group used transfer learning with ResNet as our primary model because we learned that a pre-trained CNN model works well with image classification problems. Moreover, CNN can recognize low-level features such as deformed face parts, which are keys to identifying facial expressions. The group has tried AlexNet, ResNet50, ResNet152, and our version of ResNet; transfer learning with ResNet50 produces the best result out of all models. The structure of ResNet50 is shown in figure below 5. ResNet50 consists of 50 layers and 23,522,375 trainable parameters; the group loaded ResNet50 from torchvision with its default pre-trained weights (ImageNet_1k.V2). We used all 50 layers of ResNet50, including convolutional layers, batch-normalization layers, max-pooling as well as average-pooling layers; at the end, we changed the output number of the final fully-connected layer from 1000 to seven to fit the context of our project. We trained the model with categorical cross-entropy as our loss function and SGD with momentum and weight decay as our optimizer.

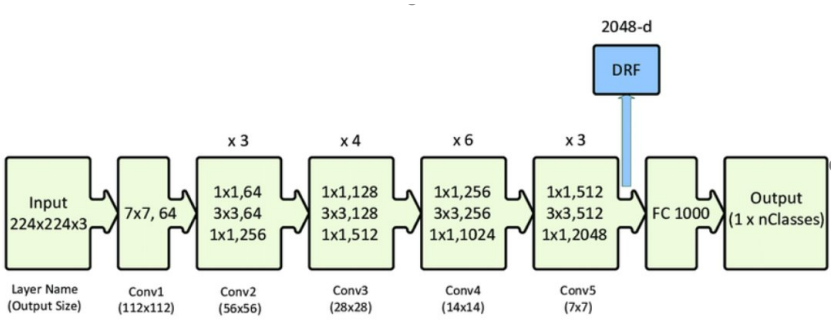



Figure 5: ResNet50 Architecture

# 5 BASELINE MODEL

## 5.1 LARGENET

The group selected a model similar to the LargeNet model from lab 2. The reason for this decision is that the group found that LargeNet was very simple to implement and achieved average accuracy in recognizing hand gestures with minimal tuning. It is the baseline model because the group wants to achieve a result that is better than the model provided in labs. Additionally, since LargeNet is a convolutional neural network, which is one of the best networks for image processing problems, the group believed that it's a reasonable baseline model for our project. This model consists of three convolutional layers, three max-pooling layers, and three linear layers, each with a 0.4 dropout except for the output layer; an example of our network is shown below 6. We trained this model using cross-entropy as the loss function and SGD with momentum to optimize weights.

The best result is a validation accuracy of 35.5%. The model is overfitting since the validation accuracy drops after 60 epochs as the training accuracy increases. We have tried many hyperparameters, but the improvement was not significant.

One qualitative fact that the team found is that while checking whether the LargeNet can overfit a small dataset, the group randomly selected three images for each group and got a validation accuracy of 30%, which was not much lower than that of going through all 50,000 labelled data. In short, training on a large amount of data did not significantly improve the accuracy of our baseline model.

## 5.2 HUMAN

Besides the LargeNet model, the group also wanted an upper-level reference. We decided to identify facial expressions ourselves.

Each member of our team individually did a human facial expression test on 28 randomly selected images from our processed data set. After tallying up our results, the average accuracy of our team is exactly 50%. To further analysis, we have computed a confusion matrix as shown in figure7below.

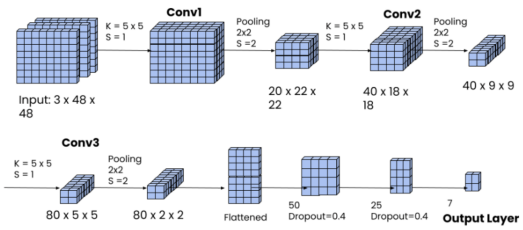



Figure 6: Baseline Network Layer Configuration

Since our human baseline is quite low, it shows that classifying facial expression is a complex problem; we won't expect the model to produce very high accuracy.

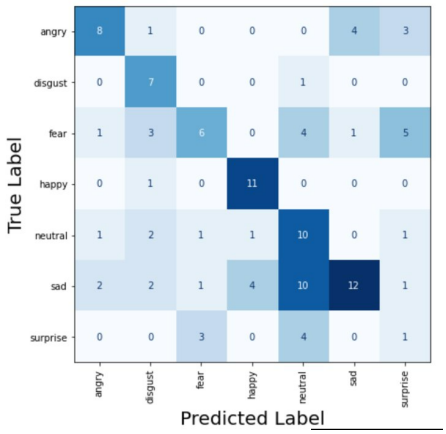


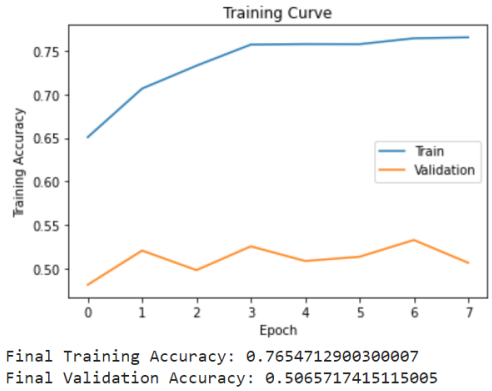

Figure 7: Human testing result confusion matrix

## 6 Quantitative Results

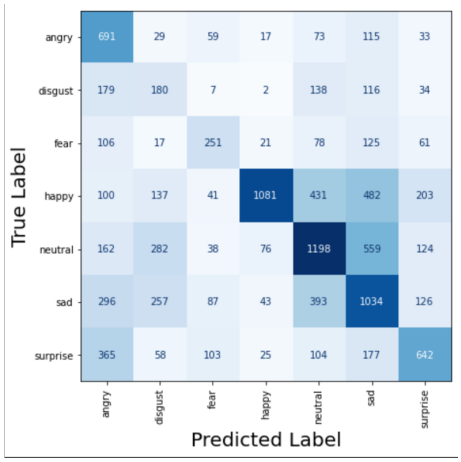
The best validation and test accuracy the group got was 53.3% and 53.8%, respectively. The group considered it an acceptable result because it is higher than both the human baseline accuracy (50%) and LargeNet's validation accuracy (35.5%). The following is an illustration of a confusion matrix on test data 8. The group worried that due to slight imbalanced data, the model might be only categorizing images into categories with a higher number of data points. However, after examining the confusion matrix, except for the category "disgust," the model has learned each category relatively well. Therefore, the model did not take advantage of data but learned to recognize emotions.

## 7 Qualitative Results

Here are a few analyses of our self-collected data with the output vector going through soft-max in figure 9 below. Figure 9(a) is a straightforward example of a confident and accurate prediction of a happy face. On the other hand, in figure 9(b), the prediction is completely wrong since this image was shot at an angel with obstruction. Plus, the group categorized it as surprise because, in the context of this image, the person in it was watching a live soccer game. Overall, our model can predict the emotion of a frontal face relatively accurately but often fail to identify faces with obstruction or being shot at an angle.

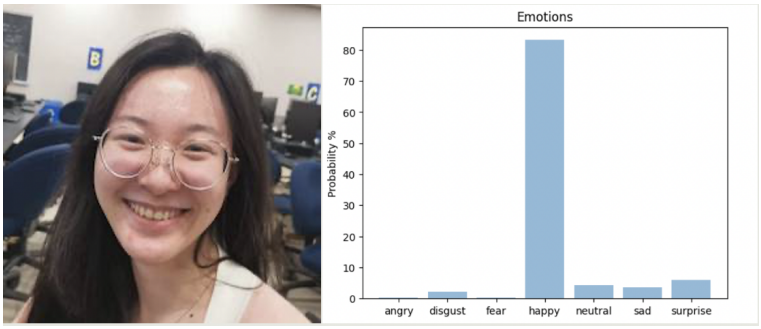



(a) Training Curve

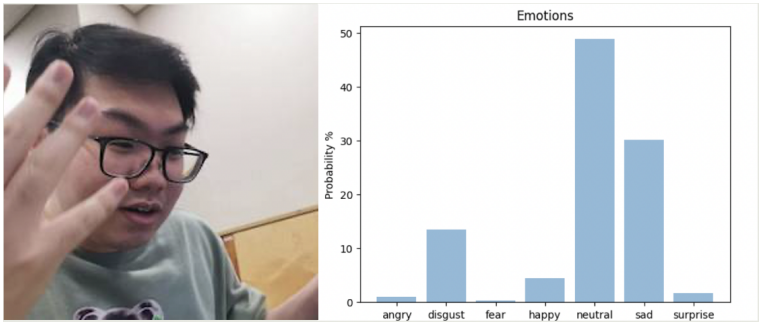



(b) Testing Data Confusion Matrix

Figure 8: Testing Data Confusion Matrix




(a) Correct Prediction




(b) Incorrect Prediction

Figure 9: Qualitative Result of Collected Data

## 8 Evaluate Model On New Data

The data were split at a ratio of 70% training data and 15% each for validation and testing data. (See Data Processing For More Details). All testing data is never seen during the training and validation process during machine learning. The testing data were fed to the final model, and the testing accuracy was 53.8%.

To further evaluate the trained model, we have collected each team member's facial expression in each category to test its capability and accuracy. We organized a new dataset with 28 images in total and fed it to our final model. The model outputs the result with an accuracy of 67.8%, which is higher than our validation and testing accuracy. The figure below is the confusion matrix of our self-

collected data 10. The model can predict Happy, Sad, and Surprise well but fail to predict disgust accurately. One thing that we need to mention is that the team was unable to pose angry and fear confidently; therefore, there is a slight imbalance in the self-collected data, which is possibly the reason for the much higher accuracy. Nevertheless, the final model performs as expected.

From the performance of testing and self-collected dataset, we can clearly see that our model performs as expected, above the human benchmark but still not high enough to be used in real life.

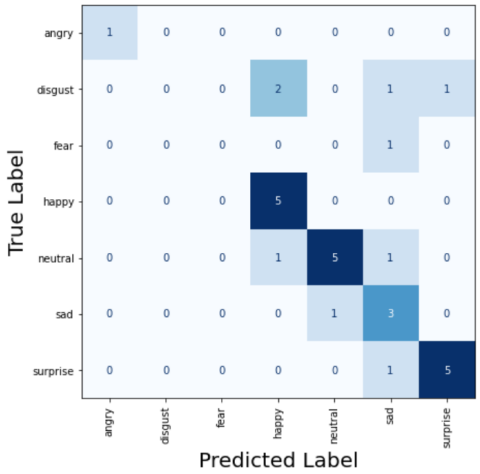



Figure 10: Self Collected Data Confusion Matrix

## 9 Further Discussion

The model is performing well and produces meaningful results for the majority of the given inputs. However, there are areas of improvements that we can make. Looking at Figure 8 8, our model can predict Happy, Neutral and Angry better than other emotions, with Angry extremely well predicted achieving an accuracy of 67.94%.

Here is a list of accuracy for each emotion - Angry: 67.94%, Disgust: 27.43%, Fear: 38.09%, Happy: 43.68%, Neutral: 49.12%, Sad: 46.24%, Surprised: 43.55%.

Disgust has a high possibility of being classified as Neutral, and Fear has a high possibility of being mis-labelled as Surprise. Intuitively, fear and surprise are similar since they both contain the element of unexpected and therefore, facial expressions could look similar but are labelled as different emotions. However, we want to further understand the cause. After manually viewing the mis-labeled images, we also found them to be hard to distinguish. Then, we tried investigating the output using softmax and found that the correct labels of some mis-labeled images are the second highest accuracy in the output. We believe it is because human emotion could be mixed and one could be both happy and surprised at the same time. Additionally, as mentioned in section 7 7, the model's predictions are correct for frontal faces but are often wrong when the faces are obstructed, which reduces our accuracy significantly. More interestingly, we found that some existing emotion recognition products have difficulty identifying obstructed faces too; for example, we upload one of our images to Face++ which is introduced in section 22, it also predicted the wrong answer although the second highest accuracy is the correct answer as shown in the image below 11. This shows that obstruction removal is one of the biggest challenges in this field.

## 10 Ethical Considerations

There are many mature databases of imposed or spontaneous facial expressions with labels assigned by more than one human. A free example would be the MMA Facial Expression Dataset from Kaggle. Such a dataset can be used to train the model since the project is for educational purposes, and we will not commercialize the model trained by those datasets. To avoid conflicts and copyright

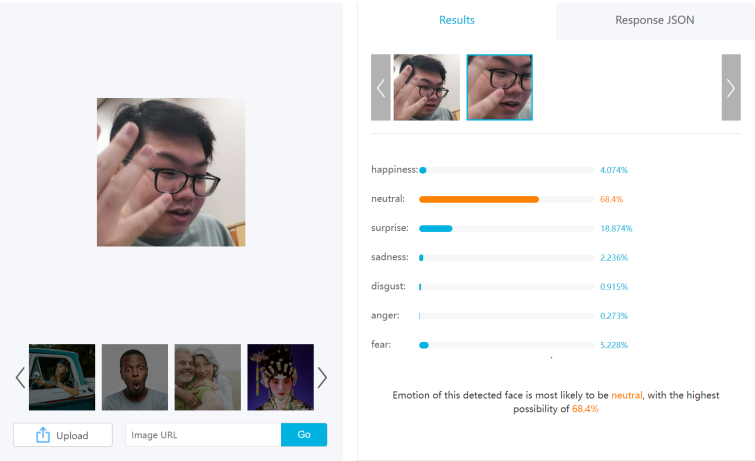



Figure 11: Face++'s Prediction

infringements, if any individual or organization reaches the group over the use of those datasets, the group will immediately remove related content from training.

The problem with collecting datasets from real humans is that they might be unaware of their facial expressions being recorded, and some people might be unwilling to share their pictures. It is unknown that the dataset we used was given full permission by all people recorded. Thus, using unidentified sources of data might result in a violation of portraiture rights.

Our project aims to help people with Autism Spectrum Disorder in their daily communications. A false prediction of the model can result in uncontrollable consequences for a person with Autism Spectrum Disorder. Before allowing anyone to use the product, the group would explain the potential fault of the model so the user can proceed with caution and know its risks.

## 11 Project Difficulty

The overall difficulty of our project is high since we have done a lot of research and have learnt beyond our lectures and lab materials. From reading papers on facial expression recognition to understand how they built their models, we have tried our best to use advanced architectures and techniques to enhance our model's performance and accuracy further.

After our research and going through different iterations of model change, from the knowledge we learned in class for using transfer leaning on AlexNet to building our ResNet with different layers. We believe that our model performed beyond expectation for facial expression recognition. Our model achieved 53.8% accuracy on our testing dataset. The number might not seem impressive at first. In comparison, our human benchmark constructed by our team's average only achieved 50% accuracy for the same dataset. Using the classification methods from labs with transfer learning of AlexNet, CNN and ANN, the accuracy was only 33%, which is almost 20% worse compared to our final model.

Additionally, based on our research, it is challenging to reach a high accuracy (i.e. 80% to 100%) in emotion recognition. The current highest accuracy we find for human expression recognition is 71% on the Kaggle face emotion recognition competition, but they have used a different dataset. This is due to our deliberate choice of selecting a difficult-to-recognize and categorized dataset. We believe that we should add spontaneously and posed expression faces as well as faces that do not have clear boundaries between different emotions. It will better simulate an everyday setting where expressions occur spontaneously and do not have clear boundaries between each expression. In real-life applications, a person would not pose a standardized happy face for you to recognize.

## REFERENCES

Emotion recognition, a. URL https://www.faceplusplus.com/emotion-recognition/.

Facial recognition microsoft azure, b. URL https://azure.microsoft.com/en-us/services/cognitive-services/face/.

Fea - facial expression analysis, Oct 2021. URL https://imotions.com/biosensor/fea-facial-expression-analysis/.

google. Ml kit nbsp;—nbsp; google developers, 2021. URL https://developers.google.com/ml-kit/guides.

Pierre Guillou. Face expression recognition with fastai v1, Nov 2018. URL https://medium.com/@pierre_guillou/face-expression-recognition-with-fastai-v1-dc4cf6b141a3.

Theresa Küntzler, T. Tim Höfling, and Georg W. Alpers. Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Frontiers in Psychology*, 12, 2021. doi: 10.3389/fpsyg.2021.627561.

Noldus. Facial expression recognition software: Facereader, 2021. URL https://www.noldus.com/facereader.

Tom Scheve. How many muscles does it take to smile?, Apr 2021. URL https://science.howstuffworks.com/life/inside-the-mind/emotions/muscles-smile.htm#:~:text=There%20are%2043%20muscles%20in,%2C%20buccal%2C%20mandibular%20and%20cervical.