IDS 705 Final Report

# Feeling It: Emotion Classification with Machine Learning

*Using Convolutional Neural Networks to Detect and Analyze Human Emotions from Facial Images*

**Team Members**:
Andrew Kroening
Chloe (Ke) Liu
Wafiakmal Miftah
Jenny (Yiran) Shen

**Team Number**: 7

---

## Abstract

Computer imagery interpretation has advanced significantly with the arrival of deep learning and convolutional neural network architectures and techniques. We seek to extend these methods to interpreting emotional states in photos of people. We then compare to classical research in this area to examine the performance and limitations of machines. Faces are also obstructed in one experiment to explore the impact of facial coverings on emotional detection. These approaches yield insights into ways the human mind might adapt to a scenario where a full face is not visible, such as during a pandemic or in cultures where coverings are traditional.

# Introduction

A significant amount of historical psychological research has explored how humans interpret the emotional states of one another by examining facial expressions. Of particular interest, research has sought to map emotional states to the activation of specific facial muscle groups giving rise to popular phrases such as, “it takes fewer muscles to smile than frown, so be happy!” In recent years, expanding computational power and advanced methods have also brought computers and computer vision to this space. Numerous computational methods have attempted to recreate or expand upon this knowledge by examining pictures or videos of faces using various techniques.

The COVID-19 pandemic introduced a new wrinkle to identifying emotions from observation. As much of the world adopted facemasks as a standard practice, questions naturally emerged about the possible impact of obstructing a portion of a face on human interaction and emotional inference. While this challenge has already existed in certain cultures where the wearing of items that obstruct a part of a person’s face might be tradition, the COVID-19 pandemic brought this question to the masses. To date, various studies have explored this topic, and we will seek to build upon this knowledge as a starting point for our experiments.

With this foundation in mind, the research team will further explore the ability of computer vision models to classify human emotions. This project aims, in particular, to gain insights into how specific machine learning models might infer human emotions from facial representation, compare those results to traditional psychological research, and then challenge that framework by obstructing portions of the faces to estimate those impacts on emotional inference.

# Background

To date, considerable efforts have sought to explore the interpretation of facial emotions using modeling techniques. We aim to incorporate and leverage this literature by providing a baseline that will serve as a measure to compare certain aspects of our model’s performance. One area with keen interest is the regions of a face that are significant for predicting each emotional state. We already know that certain emotions trigger specific muscle groups, and we would like to incorporate a comparison of our model’s insights with traditional psychological research. Put simply; we want to compare what our model sees versus what a human sees.

In a cursory review of available literature, we found that most were worthy attempts to explore this space. However, some of the research we uncovered was outdated and would benefit from more advanced models and computing power developed in the years since publication, such as O’Toole and

Abdi's work[1] from 2001. We did find some promising contributions in recent years from Dajose[2] and Dores[3], and an exciting piece from Heaven[4] that we considered for potential application to our efforts.

Ultimately, the most insightful piece to the project was from Liu, Meng, et al.[5] This work sought to explore specific areas of facial activation during various emotional states. Of particular use, discussed in more depth later, was the high-quality mapping of these emotional activation areas. We will use these maps when interpreting the saliency maps generated by our models.

With this knowledge, we can expand upon the problem space by obstructing portions of the pictures that our model sees to ascertain an effect on the overall performance. We do this to address a question that extends from the first. Once we know what the model is seeing and basing decisions on, does it remain equally effective when we challenge it by being unable to observe some of those critical areas? We see this as a key to understanding the impact of covering portions of a face on emotional inference and being able to make an educated guess about how this may also impact a human's ability to do the same.

# Data

The facial expression dataset originated from Kaggle[6]. This dataset consisted of 7 facial expressions from various gender, race, and age. In the original state of the dataset, when downloaded, we found over 35,887 images. Our exploratory data analysis began by checking these emotional categories for balance and visual sampling to ensure that the categories were assigned correctly. The visualization of the category balance shows some skew, while the visual inspection found multiple instances that the team considered either an erroneous assignment or a potentially duplicated image.

One team member conducted an image-by-image review of the dataset to correct the erroneous categorization and re-classified images as best as possible. This resulted in a slight alteration to the balance of the emotional categories, as the overall size of the dataset outweighed the small but concerning number of misclassified images. This process also brought about a brief discussion about the nuances of emotional interpretation for persons raised in different cultures. While we consider our reassignment to be correct, it is certainly possible that another person examining our resigned labels may find also find issues.

---

[1] O'Toole, A.J., and H. Abdi. "Face Recognition Models." *International Encyclopedia of the Social & Behavioral Sciences*, 2001, pp. 5223-5226. *Science Direct*, https://www.sciencedirect.com/science/article/pii/B0080430767006902.

[2] Dajose, Lorinda. "Facial Expressions: How Brains Process Emotion." *Caltech*, 24 April 2017, https://www.caltech.edu/about/news/facial-expressions-how-brains-process-emotion-54800. Accessed 9 March 2023.

[3] Dores, Artemisa, et al. "Recognizing Emotions through Facial Expressions: A Largescale Experimental Study." *National Library of Medicine*, vol. 20, 2020, p. 7420. *National Library of Medicine*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7599941/.

[4] Heaven, Douglas. "Why faces don't always tell the truth about feelings." *Nature*, 2020, https://www.nature.com/articles/d41586-020-00507-5. Accessed 08 March 2023.

[5] Liu, Meng, et al. "Facial Expressions of Emotion Categories are Embedded within a Dimensional Space of Valence-arousal." https://psyarxiv.com/pw5uh/.

[6] Oheix, Jonathan. "Face expression recognition dataset." *Kaggle*, 2019, https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset. Accessed 9 March 2023.

Once the reassignment was completed, the team reexamined the dataset, this time inspecting for duplicate images. This was smoother than the manual reassignment process. The image contents were read into a data frame and converted to a hashable data type. The contents were then compared to find duplicates that were subsequently dropped from the dataset. In total, we identified 1,506 images that had at least one exact copy. Anecdotally, we later observed cases where a face was present in more than one image, but the aspect ratio, zoom, or other features were altered very slightly. These cases escaped our inspection because they were not identical copies of one another. Figure 1 (below) presents the dataset before and after these steps were taken.

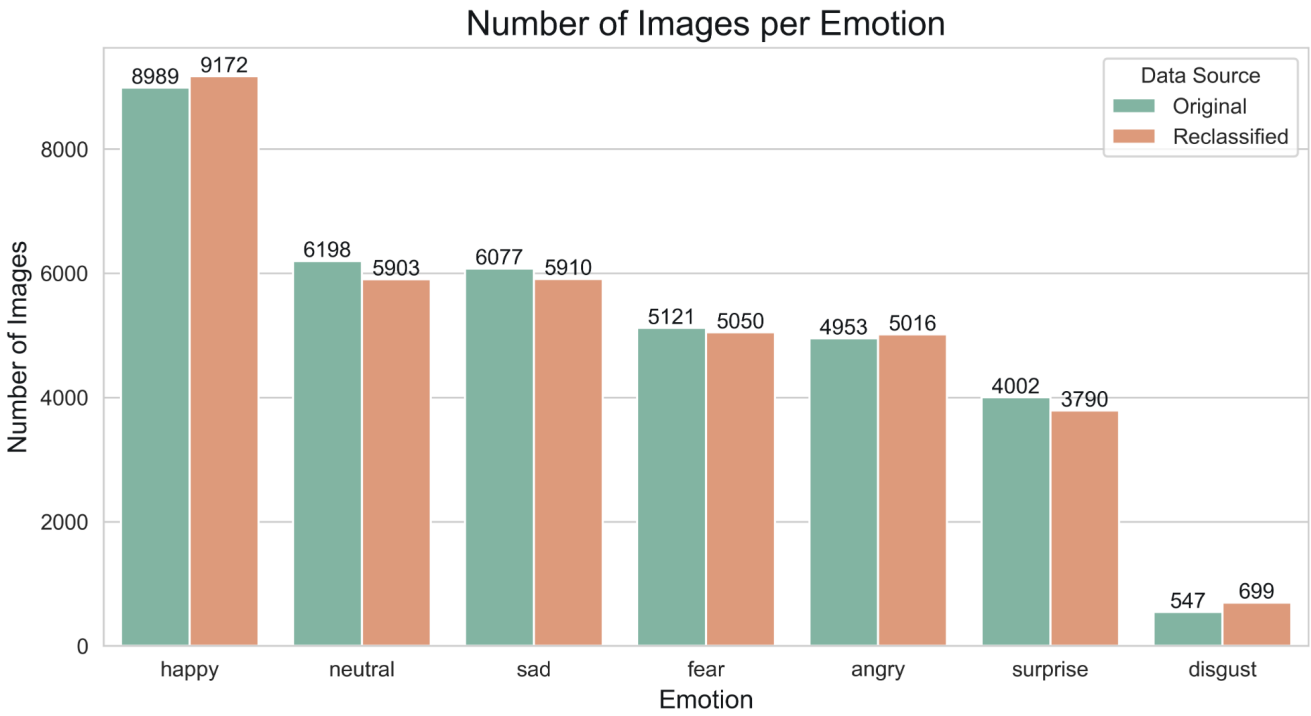


*Figure 1. Categories of the Dataset*

The list of facial expressions and the number of images for each train and validation data are stated in Table 1 (below).

**Categorical Summary**

| Facial Expression | Training Images | Validation Images |
|---|---|---|
| Angry | 3,977 | 888 |
| Disgust | 551 | 83 |
| Fear | 3,867 | 916 |
| Happy | 7,230 | 1,752 |
| Neutral | 4,580 | 1,181 |
| Sad | 4,696 | 1,088 |

| Facial Expression | Training Images | Validation Images |
|---|---|---|
| Surprise | 2,645 | 580 |

*Table 1. Dataset properties (After relabeling and cleaning)*

The images were also resized to a 48x48 pixel size to focus on the person's facial features, greyscaled, and labeled accordingly. Our group chose this dataset because it supports our goal of learning the features most important in facial expression recognition. This robust dataset will also allow us to experiment with data augmentation.

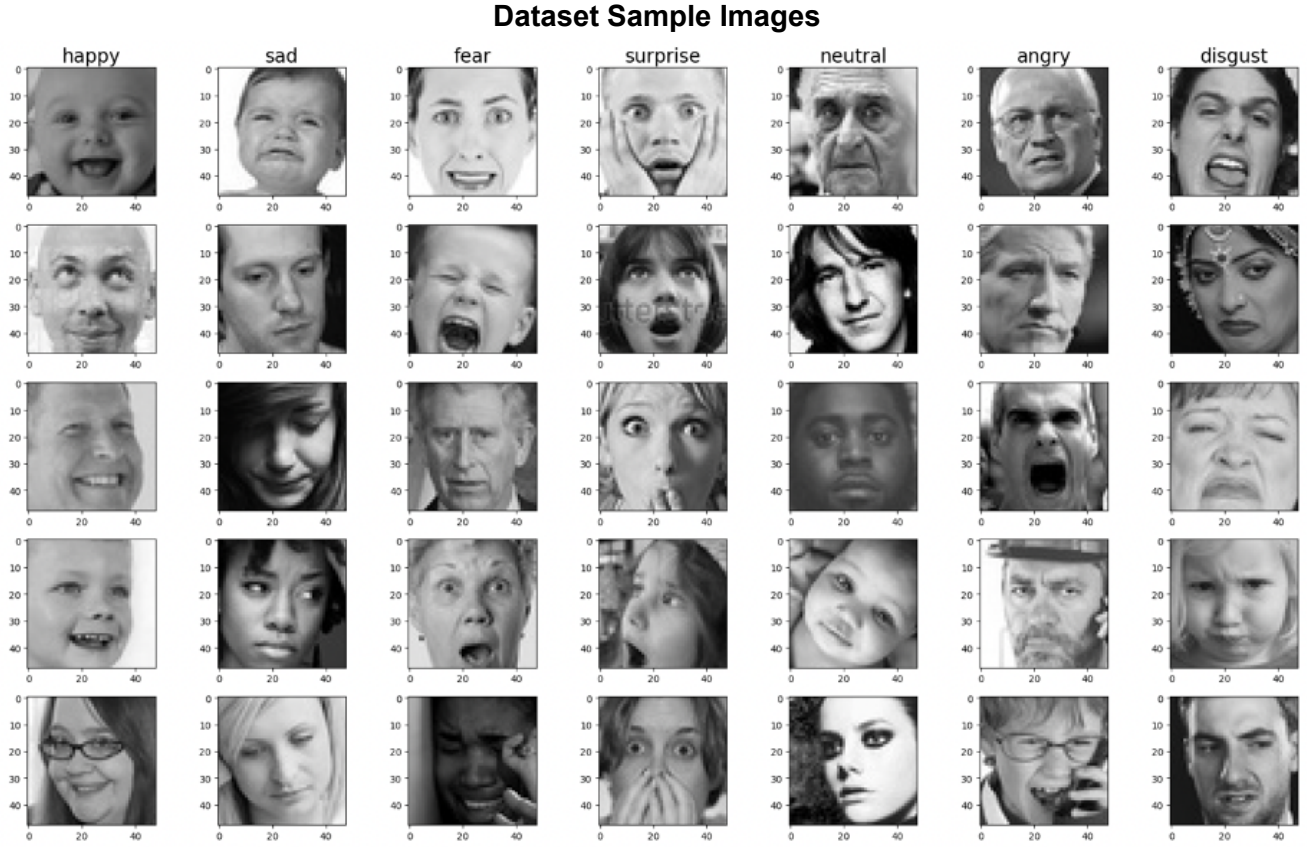


*Figure 2. Dataset Excerpt*

Above is an excerpt of the dataset, with five samples of each of the seven emotions. As seen from the sample images, there is considerable variation within each class that our model(s) will have to contend with. For instance, we can observe hands-on or covering parts of faces, young people and older adults, angled faces with respect to the photo, and even some watermarks. While the source of the data we derived from Kaggle is unclear, many easily identified personalities suggest these are images collected from various internet sites and manually cataloged. We further corroborated these suspicions when we observed the incorrectly classified images earlier in data exploration.

# Experiments

## Methodology

The experiment methodology the team uses is laid out below in the flowchart in Figure 3. In general, we utilized a parallel process utilizing the same cleaned dataset. A VGG-16 model was utilized in one arm, and the other was based on a 50 model. Both models were fit by comparing the performance of various hyperparameter configurations and validated against a version of the images with no masking and a version with simple masking covering portions of the face. We explored methods that relied on more advanced masking techniques but discovered unreliable performance in several cases, prompting a return to a more straightforward technique that relied on image regions rather than feature recognition.

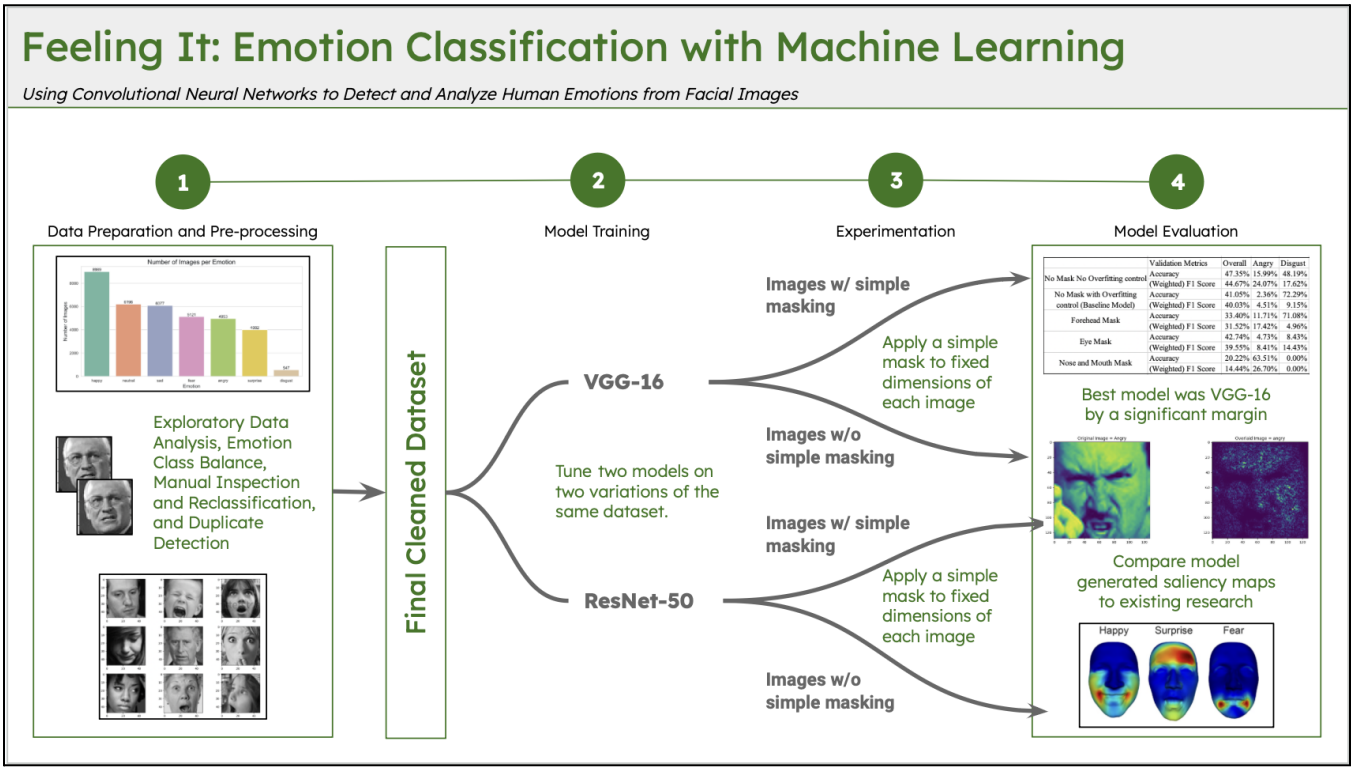


*Figure 3. Experiment Flowchart*

## Model Architecture and Hyperparameters:

We tried two common model architectures for the Face Emotion Classification model. One followed the study by Porcu et al. and used VGG-16 as the model architecture.[7] The other followed Li and Lima and utilized ResNet50.[8]

[7] Porcu, Simone. "(PDF) Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems." ResearchGate, 18 November 2020, https://www.researchgate.net/publication/345940392_Evaluation_of_Data_Augmentation_Techniques_for_Facial_Expression_Recognition_Systems.

[8] Li, B., & Lima, D. (2021). Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering*, *2*, 57-64.

## VGG-16

The VGG-16 model architecture was utilized to create a baseline model and three masking layer augmentations, all with a learning rate of 1e-4. An early stopping callback function was implemented to stop training after three epochs with no improvement in validation loss. We further experimented with a dropout rate of 0.3 and batch normalization layers to account for model overfitting, which increased the overall model accuracy and decreased loss after controlling for overfitting. Therefore, we choose to use the model with batch normalization layers and dropout as the baseline model, and all models with masking layers use the same architecture. The model architecture is depicted below in Figure 4.

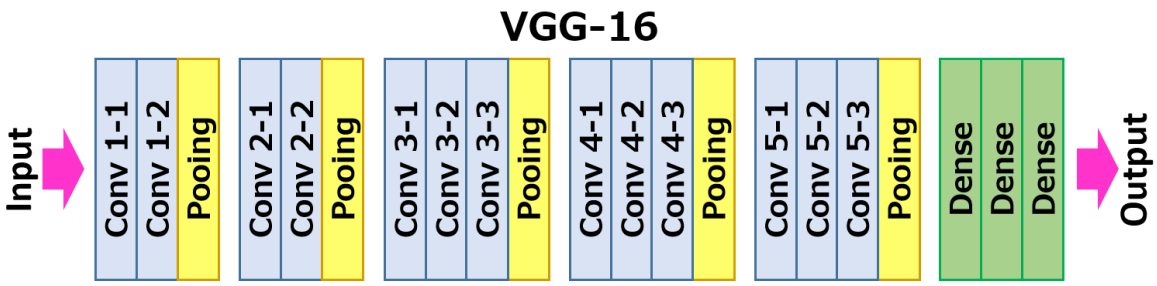


*Figure 4. VGG-16 Architecture*[9]

## ResNet-50

The ResNet-50 model was utilized to create an alternative baseline model. The learning rate for ResNet to achieve its yet best performance is 1e-4, with an 'Adam' optimizer, Cross-entropy loss, and the same stopping callback logic as VGG-16. We also experimented with batch normalization and dropout. However, no improvement was found based on the performance of the evaluation metric. Hence, we choose to move forward with the model without batch normalization layers and dropout as the baseline model for ResNet-50.

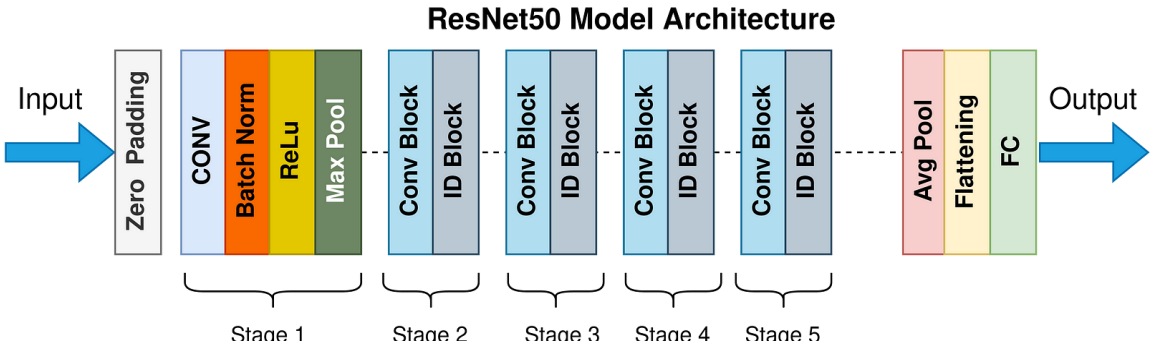


*Figure 5. ResNet50 Architecture*[10]

[9] http://www.renom.jp/notebooks/tutorial/image_processing/neural-style-transfer/notebook.html
[10]https://towardsdatascience.com/the-annotated-ResNet-50-a6c536034758

# Models with Masking Layers

In many real-life scenarios, facial recognition technology must face various challenges. For example, individuals may not always show their entire faces. They may be wearing glasses or sunglasses, masks in hospitals, hats for fashion, or hijabs due to religious practices. Additionally, cameras may capture only a portion of an individual's face or may be affected by changes in lighting. To address these challenges, we conducted experiments by obstructing regions of the images used by our models.

To compare the effect of different masking techniques on the model's predictive performance, we used a baseline model that was trained on the entire 128x128 image and then compared it against the masking layer augmented models. We explored 3 different masking layers, with each corresponding to real-life scenarios.

- Top 1/4 mask: corresponds to the covering of the forehead.
- Middle 1/4 mask: corresponds to the covering of the eyes.
- Bottom half mask (1/2): corresponds to the covering of the mouth and lower face.

Since face positions in our training data vary, we further refined masking by adopting dlib and OpenCV to detect facial landmarks in images for more sophisticated masking. We would gain specific (x, y)-coordinates of regions surrounding each facial structure and then crop the region out correspondingly. Below are samples and explanations for the simple masking procedure.

**Simple Masking Samples**

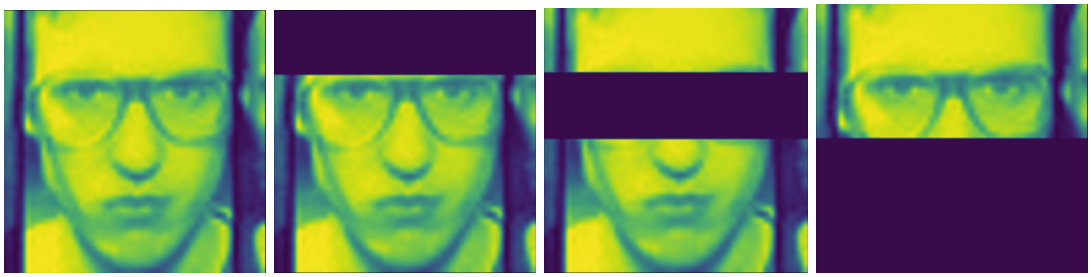

*Figure 5. Simple Masking Samples*

## Top 1/4 Mask (Forehead Mask)

The Forehead Mask is a masking layer augmentation technique that involves reducing the number of pixels incident on the camera lens, i.e, we block all pixels incident on the camera lens in the top 1/4 of the image. Given that our original image is of size 128x128, removing the top 1/4 pixels results in an image of 96x128. In Figure 5, the left image above represents the original non-masked 128x128 image, and the center-left represents the same image with a forehead mask.

## Middle 1/4 Mask (Eyes and Eyebrows Mask)

A Middle Mask is a physical layer augmentation technique that involves reducing the number of pixels incident on the camera lens in the middle 1/4 of the image. Since our original image is 128x128, removing the middle 1/4 pixels results in a two-part image with a top size of 32x128 and the bottom size

of 64x128. The center-right image in Figure 5 represents the base image with a eyes and eyebrows mask applied to it.

### Lower Face Mask (Nose and Mouth Mask)

Lower Face Mask is a physical layer augmentation technique that involves reducing the number of pixels incident on the camera lens in the bottom half of the image. Given that our original image is of size 128x128, removing the bottom-half pixels results in an image of 64x128. The right image in Figure 5 represents the base image with a lower face mask applied to it.

## Training and Evaluation

### Performance Comparison

Each model was trained on 9-13 epochs with a batch size of 32 images and tested after each epoch with a batch size of 32 images. We had a total of 7,230 happy, 4,580 neutral, 4,696 sad, 3,867 fearful, 3,977 angry, 2,645 surprised, and 551 disgusted images. The training data was split into an 80-20 stratified training validation split that preserved the percentage of samples for each class. Once the early stopping criteria were met or after 30 epochs, the test data was used to evaluate the model's out-of-sample performance and calculate metrics. Tables 2 and 3 (below) depict the outcomes for the VGG-16 model and ResNet50, respectively.

**VGG-16 Performance Metrics**

| | Validation Metrics | Overall | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|---|---|
| No Mask No Overfitting control | Accuracy | 56.41% | 42.00% | 36.14% | 40.83% | 82.53% | 50.97% | 44.94% | 59.66% |
| | (Weighted) F1 Score | 56.28% | 45.38% | 33.90% | 38.70% | 79.89% | 51.81% | 44.76% | 63.31% |
| No Mask with Overfitting control (Baseline Model) | Accuracy | 58.25% | 55.52% | 21.69% | 42.14% | 79.74% | 51.99% | 44.39% | 66.90% |
| | (Weighted) F1 Score | 58.58% | 50.85% | 32.14% | 39.31% | 82.13% | 54.72% | 44.58% | 67.60% |
| Forehead Mask | Accuracy | 57.48% | 49.66% | 20.48% | 30.35% | 86.02% | 76.88% | 25.55% | 51.72% |
| | (Weighted) F1 Score | 55.78% | 49.58% | 29.83% | 36.92% | 81.92% | 55.10% | 34.01% | 62.05% |
| Eye Mask | Accuracy | 54.93% | 43.13% | 19.28% | 44.87% | 83.50% | 53.26% | 29.04% | 59.66% |
| | (Weighted) F1 Score | 54.21% | 42.67% | 30.77% | 39.90% | 79.84% | 50.48% | 36.57% | 61.13% |
| Nose and Mouth Mask | Accuracy | 43.34% | 23.31% | 8.43% | 39.30% | 47.43% | 61.30% | 20.68% | 78.97% |
| | (Weighted) F1 Score | 42.89% | 33.99% | 14.89% | 34.75% | 58.81% | 47.35% | 28.70% | 42.80% |

*Table 2. VGG-16 Model Results*

**ResNet50 Performance Metrics**

| | Validation Metrics | Overall | Angry | Disgust | Fear | Happy | Neutral | Sad | Surprise |
|---|---|---|---|---|---|---|---|---|---|
| No Mask No Overfitting control | Accuracy | 47.35% | 15.99% | 48.19% | 3.28% | 72.20% | 56.65% | 60.02% | 47.07% |
| | (Weighted) F1 Score | 44.67% | 24.07% | 17.62% | 6.24% | 72.22% | 47.80% | 42.05% | 56.12% |
| No Mask with Overfitting control (Baseline Model) | Accuracy | 41.05% | 2.36% | 72.29% | 45.52% | 58.73% | 56.39% | 4.69% | 72.24% |
| | (Weighted) F1 Score | 40.03% | 4.51% | 9.15% | 34.45% | 70.50% | 48.49% | 8.60% | 57.32% |
| Forehead Mask | Accuracy | 33.40% | 11.71% | 71.08% | 18.12% | 82.76% | 1.52% | 10.29% | 44.48% |
| | (Weighted) F1 Score | 31.52% | 17.42% | 4.96% | 22.45% | 67.65% | 2.99% | 16.30% | 48.73% |
| Eye Mask | Accuracy | 42.74% | 4.73% | 8.43% | 62.12% | 78.77% | 40.14% | 21.23% | 12.07% |
| | (Weighted) F1 Score | 39.55% | 8.41% | 14.43% | 32.88% | 71.37% | 42.45% | 27.88% | 21.15% |
| Nose and Mouth Mask | Accuracy | 20.22% | 63.51% | 0.00% | 40.83% | 1.43% | 0.00% | 22.89% | 17.24% |
| | (Weighted) F1 Score | 14.44% | 26.70% | 0.00% | 26.81% | 2.75% | 0.00% | 23.95% | 25.03% |

*Table 3. ResNet50 Model Results*

## Saliency Mapping

Once the model evaluation was complete, we generated saliency maps of the outcomes to compare the areas of importance for detecting specific emotions in our images. Below is a comparison of the saliency map for “anger” determined by VGG-16 and ResNet-50. Both models keyed into the cheekbone/nasal region of the face for queues for this emotion. VGG-16 also found importance in the central forehead region.

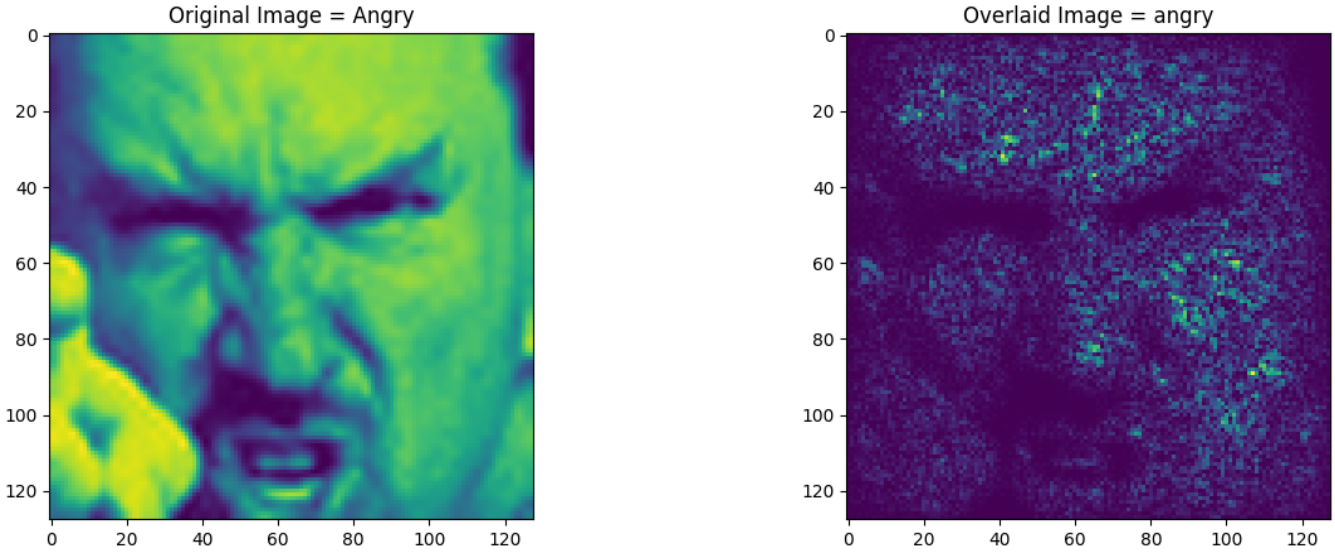


*Figure 6. VGG-16 with Overfitting Control Saliency Map on No Mask No Overfitting Control (Highest Overall Accuracy)*

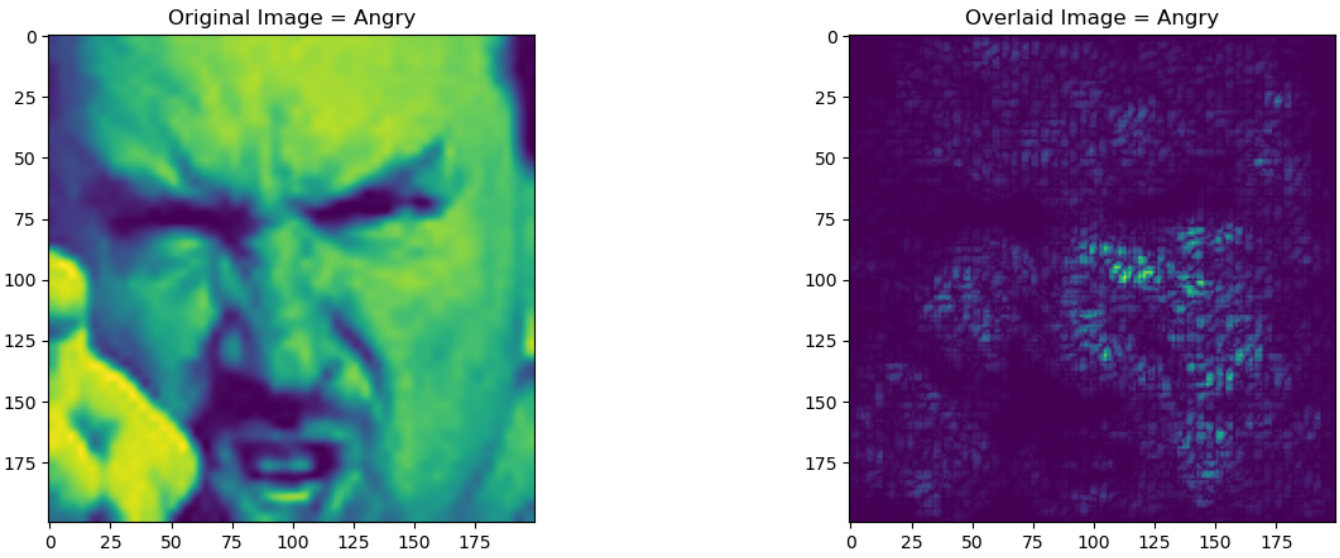


*Figure 7. ResNet-50 Saliency Map on No Mask No Overfitting Control (Highest Overall Accuracy)*

With these maps generated, we can compare with the mappings generated by Liu, Meng, et al. Below, in Figure 8, we show the mapping generated in their paper. A visual comparison shows that both models generally weighted the cheekbone/central face region as a human would when determining anger. The addition of the forehead region by VGG-16 is not expected using a human-focused interpretation. Interestingly, the overlap between disgust and anger is significant in human interpretation, which may contribute to our models' struggles in classifying these emotions. However, the baseline RENET-50 model was very good at detecting disgust (>72%) but atrocious at seeing anger (<3%).

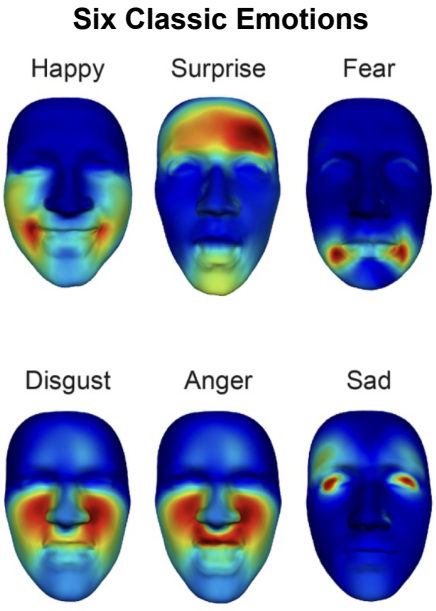


*Figure 8. Classic Emotion Mapping*[11]

[11]Liu, Meng, et al. "Facial Expressions of Emotion Categories are Embedded within a Dimensional Space of Valence-arousal." https://psyarxiv.com/pw5uh/.

# Results

The above experiment provides important insights into the performance of two popular deep learning models, VGG-16 and ResNet-50, in facial expression recognition tasks. Both models demonstrate significant improvements over random guessing, with VGG-16 achieving a higher accuracy and F1 score than ResNet-50, even when applying masks. This suggests that the VGG-16 model, with batch normalization and dropout, may be better suited for facial expression recognition tasks.

A possible reason for the performance difference could be VGG-16's simpler architecture compared to ResNet-50, which can be advantageous when dealing with smaller datasets. The VGG-16 architecture consists of 13 convolutional layers, while ResNet-50 has 50 layers. The simpler architecture of VGG16 can help prevent overfitting on smaller datasets, which is important when working with facial expression datasets that may be limited in size.

The study also reveals the importance of the nose and mouth regions in recognizing facial expressions. When images with masks are used, occluding these regions leads to the lowest accuracy and F1 score compared to applying forehead or eye masks. This indicates that the nose and mouth regions are critical for correctly identifying facial expressions and that occlusion of these regions can significantly impair the ability of a facial expression recognition model to detect emotions accurately. These findings may help guide the development of more effective facial recognition models that prioritize these critical regions.

# Conclusions

We find that the VGG-16 model architecture outperforms ResNet-50 in detecting the emotional state of persons in images. When the dataset is augmented to obstruct portions of the face, we see alterations in the performance of both models to detect specific emotions. This is not surprising, although we find additional insights into the cause of this drop-off when we examine the saliency maps of both models and compare them to existing psychology research in the area.

There were some surprises when examining the saliency maps, such as VGG-16's addition of the forehead region for detecting anger. However, these maps generally followed our expectations. The facial obstruction and the model's performance provide insights into how the human brain might adapt to detecting emotions when other people in the environment are wearing face masks, for instance. We observed model performance increase in some cases, such as VGG-16 at detecting neutral or sad emotions. This suggests that the human brain might theoretically be able to adapt to detecting emotions in scenarios where a full face is not visible.

# Roles

Andrew Kroening:

- Exploratory Data Analysis
- Report Compilation and Editing
- CVV Facial Masking
- Task Management

Chloe (Ke) Liu

Wafiakmal Miftah:

Jenny (Yiran) Shen

# References

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, *32*.

Dajose, Lorinda. “Facial Expressions: How Brains Process Emotion.” *Caltech*, 24 April 2017, https://www.caltech.edu/about/news/facial-expressions-how-brains-process-emotion-54800. Accessed 9 March 2023.

Dores, Artemisa, et al. “Recognizing Emotions through Facial Expressions: A Largescale Experimental Study.” *National Library of Medicine*, vol. 20, 2020, p. 7420. *National Library of Medicine*, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7599941/.

Goodfellow, Ian J., et al. “Challenges in representation learning: A Report on Three Machine Learning Contests.” *Neural Information Processing: 20th International Conference*, vol. 8228, 2013, pp. 117-124. *Springer*, https://link.springer.com/chapter/10.1007/978-3-642-42051-1_16.

Hase, Peter, Chaofan Chen, Oscar Li, and Cynthia Rudin. "Interpretable image recognition with hierarchical prototypes." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 32-40. 2019.

Heaven, Douglas. “Why faces don’t always tell the truth about feelings.” *Nature*, 2020, https://www.nature.com/articles/d41586-020-00507-5. Accessed 08 March 2023.

Jeon, J., et al. “A Real-time Facial Expression Recognizer using Deep Neural Network.” *International Conference on Ubiquitous Information Management and Communication*, 2016, pp. 1-4.

Kim, B. K., et al. “Hierarchical committee of deep convolutional neural networks for robust facial expression recognition.” *Journal on Multimodal User Interfaces*, vol. 10, 2016, pp. 173-189.

Liu, Meng, et al. “Facial Expressions of Emotion Categories are Embedded within a Dimensional Space of Valence-arousal.” https://psyarxiv.com/pw5uh/.

Minaee, S., et al. “Deep-emotion: Facial expression recognition using attentional convolutional network.” *Sensors*, vol. 21, no. 9, 2021, p. 3046.

Minaee, Shervin, and Amirali Abdolrashidi. “Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Networks.” 2019, https://arxiv.org/pdf/1902.01019v1.pdf.

Oheix, Jonathan. “Face expression recognition dataset.” *Kaggle*, 2019, https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset. Accessed 9 March 2023.

O'Toole, A.J., and H. Abdi. “Face Recognition Models.” *International Encyclopedia of the Social & Behavioral Sciences*, 2001, pp. 5223-5226. *Science Direct*, https://www.sciencedirect.com/science/article/pii/B0080430767006902.

Pei, Zhao, et al. “Face Recognition via Deep Learning Using Data Augmentation Based on Orthogonal Experiments.” *Electronics*, 2009. *Semantic Scholar*, https://www.semanticscholar.org/paper/Face-Recognition-via-Deep-Learning-Using-Data-Based-Pei-Xu/8449af7f2950e1beacdb5d759ca743815bb59748.

Porcu, Simone. “(PDF) Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems.” *ResearchGate*, 18 November 2020, https://www.researchgate.net/publication/345940392_Evaluation_of_Data_Augmentation_Techniques_for_Facial_Expression_Recognition_Systems. Accessed 5 April 2023.

Ranganathan, G. “A study to find facts behind preprocessing on deep learning algorithms.” *Journal of Innovative Image Processing (JIIP)*, vol. 3, no. 01, pp. 66-74.

Shin, M., et al. “Baseline CNN structure analysis for facial expression recognition.” *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 724-729.

Vepuri, Ksheeraj Sai. “Improving Facial Emotion Recognition with Image processing and Deep Learning.” *SJSU ScholarWorks*, 2021. *San Jose State University*, https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=2029&context=etd_projects.

Zeiler, Matthew, and Rob Fergus. “Visualizing and Understanding Convolutional Networks.” 2013. *Visualizing and Understanding Convolutional Networks*, https://arxiv.org/pdf/1311.2901.pdf.