Predicting Human Emotions

Using convolutional neural networks to recognize human facial emotions and determine the relative importance of the upper, lower, right, and left half of one's face in emotion recognition.

Team Members:

Aarya Desai
Gavin Li
Ofosu Osei
Katherine Tian

Team Name: Iguana

Table of Content

Abstract **2**
Introduction **4**
Background **4**
Data **6**
Experiments **8**
I. Dealing with the Data: 8
Preprocessing: 8
Selective Blackout Process: 9
II. Training the Models: 9
Full Facial Images: 9
Fine-Tuning: 9
III. Evaluation: 9
Results **10**
Baseline Model: Full Facial Images 11
Fine-tuned Model Results 11
Conclusion **12**
Strengths and Limitations: 13
Strengths: 13
Limitations: 13
Future Directions: 13
References **14**
Roles **16**
Appendix **17**
Appendix A: Class-wise Facial Emotion Recognition 17
Appendix B: Micro and Macro AUC Scores 19
Appendix C: Source Code 20

Abstract

Facial emotional recognition (FER) is an important and developing field in deep learning, with a wide range of applications, including informing human-computer interaction, mental health awareness, and more. This study aimed to further explore FER using ResNet50 on the FER2013 dataset. In particular, it sought to understand whether FER could be conducted on partial images, by using selective occlusion of the upper, lower, right, and left halves of facial images. The findings indicate that a baseline ResNet50 model performs optimally on full facial images, and that performance significantly deteriorates when attempting to predict emotions on masked images. However, when masked images are incorporated into the training set in fine-tuned models, the performance on the respective masked image (e.g., performance on upper-masked images for a baseline model further fine-tuned on upper-masked images) is comparable, if not marginally superior. The findings of this study indicate that fine-tuning using image augmentation via occlusion may be a promising avenue for the advancement of the field of FER. Additionally, the study's strengths and limitations are outlined, along with potential future directions for the research area.

Introduction

Combining concepts from psychology, anatomy, and machine learning, a field within deep learning called affective computing involves studying and developing systems to recognize and understand human emotions (Ballesteros et al., 2024). In recent years, many researchers have successfully trained models to accurately recognize up to 40 independent emotions from human faces (e.g., Borgalli et al., 2022) and continue to optimize deep learning approaches to improve the process.

However, there is still room for improvement in this field, as emotional recognition is highly nuanced and complex due to the intricacies of one's facial features (e.g., Sohail et al., 2021). Moreover, today's small facial expression datasets do not allow for performance improvement (Arora et al., 2022). To tackle both issues, the current study aims to look at emotional recognition based on specific halves of facial images. This will attempt to demonstrate the relative importance of the upper, lower, left, or right sections of faces. Furthermore, the selective occlusion of facial images used in this study will allow for additional, augmented images in training data to increase the size of an otherwise small, FER2013 dataset, and therefore, optimistically improve performance.

Understanding the relative importance of the different halves of faces compared to the full face in deep learning facial emotional recognition has wide-ranging applications as well. It can be used for any application that requires the computer to detect facial emotions, such as consumer research, sentiment analysis, medical applications, or even the justice system (EDPS, 2021). Furthermore, being able to achieve comparable performance in facial emotion recognition with partial information could be extremely useful in similar real-world applications as it would reduce the need for large training datasets and high computational power for facial emotion recognition. Overall, this study aims to provide additional insights into the field of affective computing and offer reliable and practical information to assist the development of innovation applications in various domains.

Background

The recognition of human emotions plays a critical role in many applications, from enhancing human-computer interaction to supporting mental health assessments (e.g., Al-Zanam et al., 2023; Chowdary et al., 2023). The advent of machine learning has significantly advanced this field, enabling more subtle and accurate interpretations of emotional expressions through complex algorithms and vast datasets (Monisha & Yogashree, 2023). Traditional emotion recognition methods have mainly relied on analyzing full facial expressions, exploiting facial features to discern underlying emotional states.

However, more recent research has indicated that different facial regions contribute variably to emotion recognition. For instance, studies have shown that the eyes and eyebrows are particularly expressive of emotions such as surprise, fear, and sadness, while the mouth and chin are more indicative of emotions like happiness, disgust, and anger (Eisenbarth and Alpers, 2011; Tanaka et al., 2012). As a result, the global outbreak of COVID-19 introduced unprecedented challenges to emotion recognition paradigms. The widespread use of face masks as a preventive measure against the virus covered a significant portion of the face, primarily the bottom half, which includes expressive features such as the mouth and chin. Previous studies have shown that the use of facial masks impaired emotion detection performance largely (Rinck, 2022). This brought a necessity to reevaluate traditional emotion recognition approaches, highlighting the need for innovative strategies that can adapt to these new constraints.

On the other hand, understanding the relative importance of the right or left side of one's face is more nuanced, but still important. Research has suggested that as humans, due to the right hemisphere of the brain being more emotion-dominant, the contralateral control of the brain on the body results in more intense expressions presented on the left side of one's face (Lindell, 2018). Therefore, this may present itself as an important factor to consider in deep learning-based emotional recognition as well, however, to the best of our knowledge, extensive research has not been conducted in that specific area. Furthermore, exploring this area of research would also provide insightful additions when considering FER of images of individuals that may suffer from medical conditions such as unilateral strokes that impair one's ability to have a full range of muscular movement on one side of their face (Konecny et al., 2011).

In this context, our project aims to investigate the relative importance of facial features in FER, with a specific focus on comparing the contributions of the top half (including the eyes and eyebrows) and the bottom half (including the nose, mouth, and chin) of the face. We present the following hypothesis:

> *H_1: Facial emotion detection performance will be higher on the top half of the face compared to the bottom half.*

To gain a more holistic understanding of the relative importance of facial regions, this study will also compare the performance of emotion recognition based on the left versus right halves of faces. As no previous studies have looked into this difference according to our research, we put forth the following exploratory question:

> *Question 1: Is there a difference in predicting emotions based on the right versus left half of one's face?*

The COVID-19 pandemic has catalyzed a crucial shift in the development of FER research, urging the need for a reassessment of the relative importance of facial features in conveying emotional states, particularly in scenarios where mask-wearing is common, such as in healthcare or enhanced security systems (EDPS, 2021). Our research aims to contribute to the development of flexible emotion recognition models by applying existing theories of emotion-specific diagnostic regions for the six basic emotions to advanced machine learning techniques (Kim & Cho, 2022). We plan to quantitatively analyze and compare the emotional expressiveness of the top and bottom halves of the face. Our project aims to utilize machine learning to uncover new understandings that could contribute to the future of FER technologies. The following sections will elaborate on our approach to this project, including details about the data and methods used, the findings, and a discussion of said results and potential avenues for future research.

Data

This project uses the open-source Facial Emotion Recognition 2013 (FER2013) dataset which was originally created by researchers Carrier and Courville for an ongoing deep-learning project (Goodfellow et al., 2013). This dataset consists of a comprehensive and diverse set of 48x48 pixel grayscale facial expression images. This dataset is available in a pre-cleaned form, where the faces have been automatically registered, hence each face is centered in the image and occupies approximately the same amount of space (Wang et al., 2019). It features 35887 (28,709 in the public training set and 3,589 in the test set) facial images collected using Google Image search API. The images showcase a wide variety of characteristics, including differences in age, gender, and ethnicity. This dataset contains labels with 7 basic emotional categories, namely, angry, disgust, fear, happy, neutral, sad, and surprise (Figure 1). The distribution of these images in these categories is presented in Figure 2, where it is noticeable that there are considerably fewer images in the ‘disgust’ emotional category (approximately only 1.5% of the dataset) and a higher number of images present in the ‘happy’ emotional category (approximately 25% of the dataset; Table 1).

Figure 1. Sample of images for each emotional label in the training set.

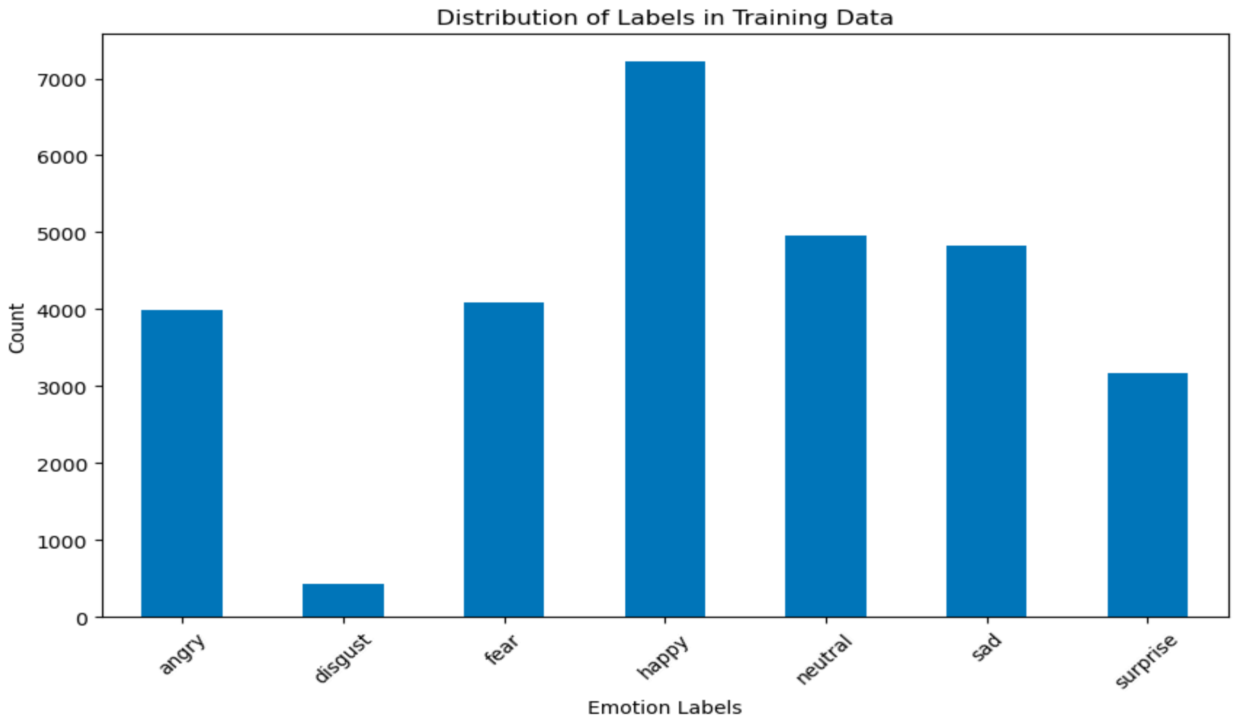

Figure 2. Distribution of the number of images present in the training set for each emotional label.

Table 1. Distribution of the number of images present in the training set and test sets for each emotional label, along with the percentage of the dataset taken up by each category.

	Train Count	Train Percent	Test Count	Test Percent
labels				
angry	3995	13.92%	958	13.35%
disgust	436	1.52%	111	1.55%
fear	4097	14.27%	1024	14.27%
happy	7215	25.13%	1774	24.71%
neutral	4965	17.29%	1233	17.18%
sad	4830	16.82%	1247	17.37%
surprise	3171	11.05%	831	11.58%
Total:	**28709**	**100%**	**7178**	**100%**
Total Split:	**Train**	**75%**	**Test**	**25%**

Experiments

Figure 3 depicts the overall methodology used for this project. This included pre-processing the data and using selective occlusion to mask images. After this, a baseline ResNet50 model trained on full facial images was evaluated based on its ability to predict emotions on full face, upper-, lower-, right-, and left-masked images. Four fine-tuned models were then created, each of which was the baseline ResNet50 model additionally trained on one form of the masked image. These fine-tuned models were again evaluated on all five types of images. The following sections elaborate on this experimental design, along with the evaluation metrics used.

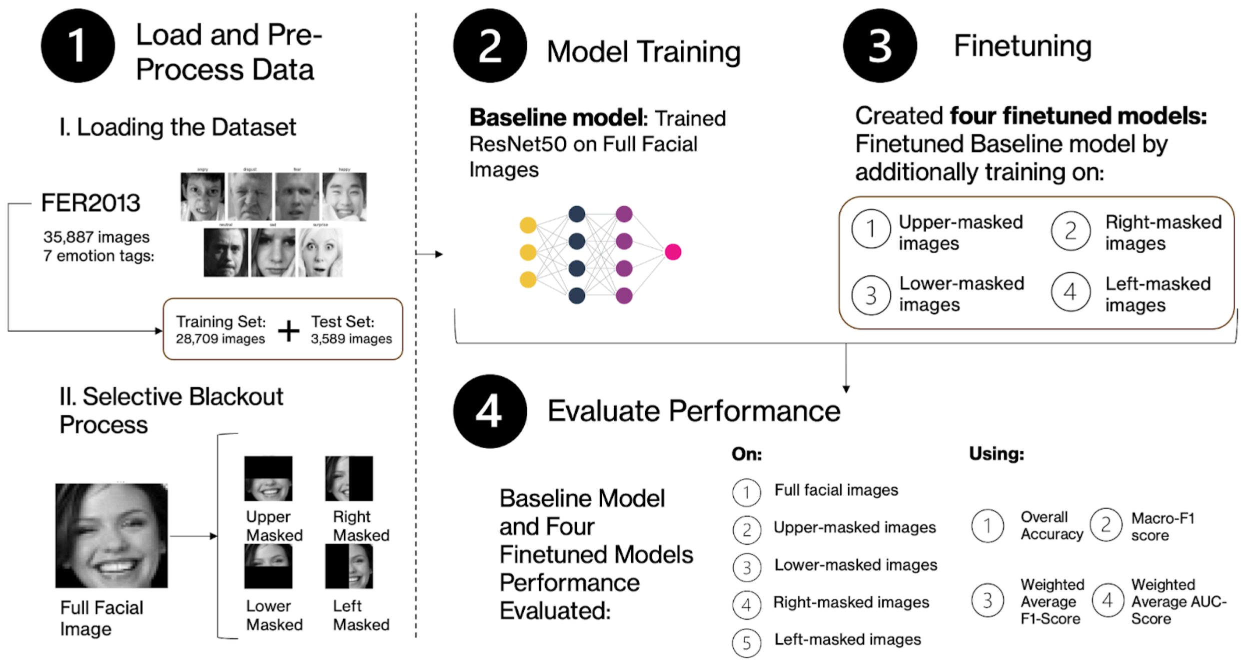

Figure 3. Flowchart depicting the methodology of this experiment.

I. Dealing with the Data:

Preprocessing:

The dataset required minimal preprocessing, as the images were already standardized. This included pre-labeling, pre-centering of faces in each image, and the use of grayscale images with a resolution of 48 x 48 pixels. To integrate this dataset into PyTorch for efficient training and validation, we developed a tailored class. This custom class has been designed to provide PyTorch with straightforward access to the dataset, facilitate data transformations, and enable seamless data sampling during both training and validation phases.

Selective Blackout Process:

The full images were utilized for both the baseline and iterative models. However, to fulfill the objective of this project, a masking technique was employed to selectively occlude either half of the facial images. This was accomplished by setting the RGB values to 0 for one-half of the image while preserving its structural integrity using image-processing-specific libraries. This allowed for the isolation and analysis of the impact of each half of one's face on facial emotion recognition.

II. Training the Models:

For this project, ResNet50 was selected due to its well-documented effectiveness in image classification tasks, including those related to FER studies (e.g., Bukhari et al., 2022; Li & Lima, 2021). Previous research has indicated that full facial emotion recognition using ResNet50 on FER2013 achieves an accuracy of approximately 57% (Gan, 2018; Zhou et al., 2020). This provides a reference point for the initial training of our baseline model with full facial images. Subsequently, progressive models were then trained on upper- or lower-masked images, and hyperparameter-tuned ResNet50 models.

Full Facial Images:

A ResNet50 model was initially trained on the entirety of the initial training data. This served to establish a baseline for subsequent models that focused on the specific regions of the face. As previously stated, an approximate accuracy of 57% was anticipated for this initial model training.

The hyperparameters employed were a learning rate of 1e-2, momentum of 0.9, and L2 regularization of 1e-4. The learning rate was decayed by 0.1 at five-epoch intervals to facilitate the achievement of higher accuracies.

Fine-Tuning:

Following the training of a baseline model on full facial images, the model was subjected to further fine-tuning by additional training on masked images. The objective of this process was to increase the size of the dataset through the use of augmented images (through occlusion). The performance of the fine-tuned models was then evaluated on both full facial and masked images.

The hyperparameters employed for training the four fine-tuned models were analogous to those utilized for the baseline model, with the exception the L2 regularization term is set to 5e-4.

III. Evaluation:

The performance of each model, distinguished by the image data it was trained on, was evaluated based on the following metrics: overall accuracy, macro F1-score, weighted average F1-scores, and weighted average AUC score. These metrics were employed to gain a more comprehensive understanding of this study, which was a multiclass classification problem with imbalanced classes (the disgust class having significantly fewer images). While overall accuracy provides an accurate assessment of the proportion of correctly classified instances, it may be misleading in the current case where the classes are imbalanced. Therefore, the additional metrics were employed, as they account for varying class sizes and provide a more comprehensive evaluation of the baseline and fine-tuned models' performance (Leung, 2022).

Results

This section will undertake a comprehensive evaluation of our baseline model and subsequent fine-tuned models. Each section will present the most significant findings and possible explanations. A detailed summary of the findings is presented in Table 2 and Appendices A through B.

Table 2. The results of training the baseline and subsequent fine-tuned models (the baseline model was additionally trained on full face, upper-, lower, right-, or left-masked images) and evaluating their performance on full facial, upper-, lower-, right-, and left-masked images are presented. All models demonstrated the highest performance on full facial images; therefore, the best performances on masked images are also highlighted in the table. The metrics employed were overall accuracy, macro F1 scores, weighted average F1 score (W.A. F1 score), and weighted average AUC score (W.A. AUC).

Trained On	**Evaluation Metric**	**Evaluated On**				
		Full Face	**Upper Mask**	**Lower Mask**	**Right Mask**	**Left Mask**
Full Face	*Accuracy*	0.6649	0.4444	0.2701	0.4064	0.4200
	Macro F1 score	0.6374	0.3111	0.2328	0.3475	0.3481
	W.A. F1 score	0.6612	0.3804	0.2370	0.4085	0.4115
	W.A. AUC	0.9119	0.7939	0.6744	0.7762	0.7775
Full Face + Upper Mask	*Accuracy*	0.6076	0.6051	0.2010	0.3958	0.4015
	Macro F1 score	0.5773	0.5867	0.1235	0.3321	0.3550
	W.A. F1 score	0.6012	0.6062	0.1387	0.3986	0.4073

	W.A. AUC	0.8813	0.8748	0.5530	0.7525	0.7492
Full Face + Lower Mask	*Accuracy*	0.5892	0.3011	0.5433	0.3426	0.3318
	Macro F1 score	0.5604	0.1374	0.5400	0.2887	0.2268
	W.A. F1 score	0.5811	0.2063	0.5397	0.3313	0.2961
	W.A. AUC	0.8752	0.6858	0.8301	0.7158	0.7125
Full Face + Right Mask	*Accuracy*	0.6257	0.4338	0.2952	0.6219	0.3289
	Macro F1 score	0.6009	0.3286	0.2516	0.5977	0.2939
	W.A. F1 score	0.6279	0.3929	0.2571	0.6213	0.3199
	W.A. AUC	0.8936	0.7860	0.6710	0.8856	0.7503
Full Face + Left Mask	*Accuracy*	0.6314	0.4209	0.2843	0.3753	0.6131
	Macro F1 score	0.5948	0.3087	0.2358	0.3228	0.5881
	W.A. F1 score	0.6220	0.3719	0.2545	0.3799	0.6092
	W.A. AUC	0.8952	0.7815	0.6595	0.7291	0.8846

Baseline Model: Full Facial Images

The baseline model yielded findings that were superior to those reported by Gan (2018) and Zhou et al. (2020), with an overall accuracy of approximately 66% (see Table 2). The similarity of the macro and weight average F1-scores further validates these results, despite the class imbalances present in the dataset. Notably, performance on masked images is significantly lower across all types of masked images. However, when comparing across only the occluded images, the baseline model performs the best on left-masked images, as indicated by the F1-scores. This aligns with previous research suggesting a stronger presentation of emotions on the left side of the face.

Additionally, it is important to note that performance was much better for upper-masked images compared to lower-masked images. This is at odds with previous research that has found that lower-masked images are better suited for emotion detection due to the relative importance of upper facial features such as the eyes and eyebrows on emotional recognition (Eisenbarth and Alpers, 2011).

Fine-tuned Model Results

It is crucial to assess the effectiveness of fine-tuned models by comparing their performance to baseline models. Notably, each model displays a strong ability to closely approximate the baseline model's performance for its designated image segment. For instance, a model fine-tuned for the upper half of an image outperforms in tasks specifically involving the upper segment, maintaining a performance level near that of the baseline model trained on full images.

Moreover, these fine-tuned models demonstrate remarkable flexibility by extending their proficiency to full-face images. This adaptability is illustrated through the comparable outcomes achieved by models trained on distinct facial halves (e.g., left and right models) when tested on full-face images. Such results underscore the potential of specialized, segment-focused training in achieving high performance only with limited facial information.

Conclusion

The aim of this study was to examine the feasibility and efficacy of utilizing ResNet50, for FER with a particular focus on partial facial images. The key findings from the experiments include:

1. **Baseline Model Performance**: The baseline ResNet50 model, trained on full facial images from the FER2013 dataset, achieved an overall accuracy of approximately 66%, demonstrating the effectiveness of ResNet50 in recognizing emotions from full facial expressions.

2. **Impact of Masking on Performance:** The study revealed significant variations in performance when the model was tested on masked images. The lower half of the face, which includes critical expressive features like the mouth and chin, plays a significant role in emotion recognition.

3. **Performance of Fine-Tuned Models:** Models fine-tuned on specific facial halves demonstrated superior performance on the corresponding masked test images relative to the baseline model. This indicates that fine-tuning on partial images can compensate for the loss of information when parts of the face are occluded.

4. **Comparative Analysis of Masked and Full Images**: Despite the specialized training, the fine-tuned models still exhibited the highest performance on full facial images. This highlights the challenge of FER with occluded faces, even when the models have been trained to recognize these types of images.

Overall, the study demonstrates that while full-face images yield the best FER performance, there is substantial merit in exploring and improving techniques for FER from partially occluded faces. This is contingent on the implementation of appropriate training adjustments and model enhancement.

Strengths and Limitations:

Strengths:

This study replicated previous research that used ResNet50 trained on FER2013 for full facial images, rendering a comparable overall accuracy to previous research of approximately 66% (Gan, 2018; Zhou et al., 2022). Furthermore, this study provides novel insights by investigating the FER of different halves of faces and contributes to future studies and real-world applications that would need this information, e.g., healthcare practices.

Limitations:

Certain limitations must be noted as well. First, our study had counterintuitive results in terms of the FER performance on upper- versus lower-masked images. This finding is at odds with previous research which found an accuracy of 81.36% on lower-masked images and only 55.52% accuracy on upper-masked images (Kim et al., 2023). However, it is important to note that MS-Celeb-1M was a Microsoft-curated dataset consisting of 10 million facial images (Guo et al., 2016). Consequently, the relative importance of upper-face features may have been extracted due to the longer training period on a larger dataset with higher-resolution images and greater demographic diversity.

Future Directions:

1. **Data Improvement:** The utilization of superior datasets, such as MS-Celeb and RAF-DB, can enhance the quality of training.

2. **Model Upgrades:** The implementation of more sophisticated models such as ResNet152 and the exploration of other cutting-edge deep learning architectures can improve the efficacy of the model. However, these enhancements may necessitate the availability of significantly more computational resources.

3. **Advanced Techniques**: The employment of ensemble methods can also be considered for the purposes of boosting the model's accuracy and robustness.

4. **Real-World Applications:** The application of FER techniques in areas such as consumer research, mental healthcare, human-computer interaction, fraud detection, transaction analysis, sentiment analysis, and the judicial system.

References

Al-zanam, A. A. A., Alhomery, E. H., Abdou, O. H., & Tan, C. P. (2023). Mental Health State Classification Using Facial Emotion Recognition and Detection. *International Journal on Advanced Science, Engineering & Information Technology*, *13*(6).

Kumar Arora, T., Kumar Chaubey, P., Shree Raman, M., Kumar, B., Nagesh, Y., Anjani, P. K., ... & Debtera, B. (2022). Optimal facial feature based emotional recognition using deep learning algorithm. *Computational Intelligence and Neuroscience*, *2022*.

Ballesteros, J. A., Ramírez V, G. M., Moreira, F., Solano, A., & Pelaez, C. A. (2024). Facial emotion recognition through artificial intelligence. *Frontiers in Computer Science*, *6*, 1359471.

Barabanschikov, V. A. (2015). Gaze dynamics in the recognition of facial expressions of emotion. *Perception*, *44*(8-9), 1007-1019.

Borgalli, R. A., & Surve, S. (2022, March). Deep learning framework for facial emotion recognition using cnn architectures. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1777-1784). IEEE.

Bukhari, N., Hussain, S., Ayoub, M., Yu, Y., & Khan, A. (2022). Deep learning based framework for emotion recognition using facial expression. *Pakistan Journal of Engineering and Technology*, *5*(3), 51-57.

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, *35*(32), 23311-23328.

EDPS, (2021). *Facial Emotion Recognition*, TechDispatch, https://www.edps.europa.eu/system/files/2021-05/21-05-26_techdispatch-facial-emotion-recognition_ref_en.pdf

Eisenbarth H., Alpers G. W. (2011). Happy mouth and sad eyes: scanning emotional facial expressions. Emotion 11, 860–865.

Einstein, G., & McDaniel, M. A. (2022). Investigating the impact of face masks on the recognition of basic emotions. Cognitive Research: Principles and Implications, 7(1), 58.

Gan, Y. (2018, August). Facial expression recognition using convolutional neural network. In *Proceedings of the 2nd international conference on vision, image and signal processing* (pp. 1-5).

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20* (pp. 117-124). Springer berlin heidelberg.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14* (pp. 87-102). Springer International Publishing.

Kim, S., An, B. S., & Lee, E. C. (2023). Comparative analysis of AI-based facial identification and expression recognition using upper and lower facial regions. *Applied Sciences*, *13*(10), 6070.

Konecny, P., Elfmark, M., & Urbanek, K. (2011). Facial paresis after stroke and its impact on patients' facial movement and mental status. *Journal of Rehabilitation Medicine*, *43*(1), 73-75.

Johnson, K. L., & Roberts, L. D. (2022). Emotion detection in masked facial expressions: A comparative study. PMC, 13(2), 1-12.

Leung, K. (2022, September 13). *Micro, Macro & weighted averages of F1 score, clearly explained*. Medium.
https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f

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

Lindell, A. (2018). Lateralization of the expression of facial emotion in humans. *Progress in brain research*, *238*, 249-270.

Savchenko, A. V., L. V., & Makarov, l. (2022). *Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network.* IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 13, NO. 4.

Smith, J. R., & Doe, A. B. (2023). A study on the effectiveness of machine learning algorithms in emotion recognition with obscured facial features. Procedia Computer Science, 200, 123-130.

Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S. H., AlGhamdi, M. A., ... & Masood, K. (2021). Racial identity-aware facial expression recognition using deep convolutional neural networks. *Applied Sciences*, *12*(1), 88.

Tanaka, J. W., Kaiser, M. D., Butler, S., & Le Grand, R. (2012). Mixed emotions: Holistic and analytic perception of facial expressions. Psychology of Aesthetics, Creativity, and the Arts, 6(3), 241-247.

Wang, L., Xu, S., Wang, X., & Zhu, Q. (2019). Eavesdrop the composition proportion of training labels in federated learning. *arXiv preprint arXiv:1910.06044*.

Zhou, N., Liang, R., & Shi, W. (2020). A lightweight convolutional neural network for real-time facial expression detection. *IEEE Access*, *9*, 5573-5584.

Roles

Project Manager:
Aarya Desai
Setting up schedules and deadlines, coordinating meetings, and ensuring the team sticks to the agenda and manages their deliverables was **Aarya Desai**'s responsibility. She also completed the final presentation for this project. She helped streamline the project objectives and re-assign deliverables based on the suggestions from Kyle's office hours and draft comments.

Deep Learning Model Developers:
Lead: Gavin Li.
Gavin Li brought his deep learning expertise to our model development process and led the team through all the steps that were necessary to build and finetune the model. He walked all members through the results of the write-up well and was the point of contact for all of Kyle's office hours.
Ofosu Osei was the idea behind this project and hence, the other deep learning model developer. He helped in training the baseline models and rendering results that were used in the final presentation.

Reporters:
Aarya Desai and **Katherine Tian** headlined the write-up of this project. Ofosu Osei helped in writing up the results and appendices sections for this report as well.

Appendix

Appendix A: Class-wise Facial Emotion Recognition

We begin by presenting our results from the training and validation stages, followed by a confusion matrix and the results from the test dataset, which offers a detailed breakdown of the model's predictions across the various emotion classes. This visualization allows us to observe not just the successes but also the intricacies of the model's behavior, such as its ability to distinguish between classes and where it tends to make errors.

Following the confusion matrix, we explore additional metrics—Accuracy, ROC-AUC, F1 Score, and Precision-Recall AUC—that give us a more rounded understanding of the model's performance. Each metric sheds light on different aspects of the model's predictions, from the overall correctness of the classifications (accuracy) to the balance between sensitivity and specificity (ROC-AUC) and the trade-off between precision and recall (F1 Score and Precision-Recall AUC).

Through this multifaceted evaluation, we aim to present a clear and transparent picture of how the model performs, ensuring we have all the necessary information to make informed decisions about future improvements and applications of our emotion recognition system.

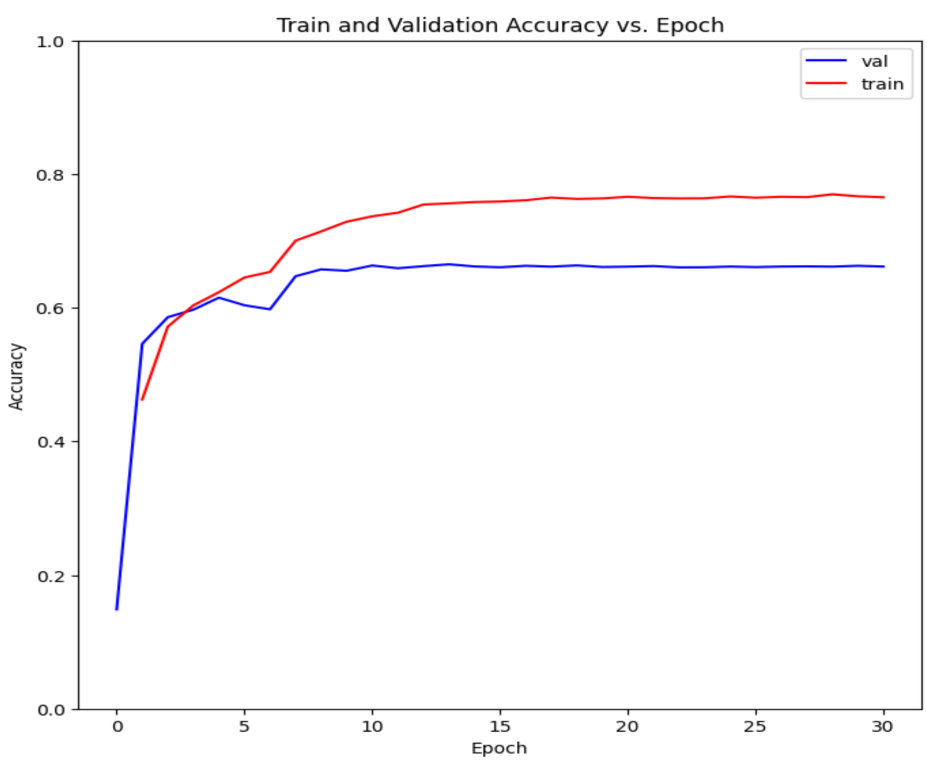

Figure 4: Initial base model performance over epochs - Training and Validation

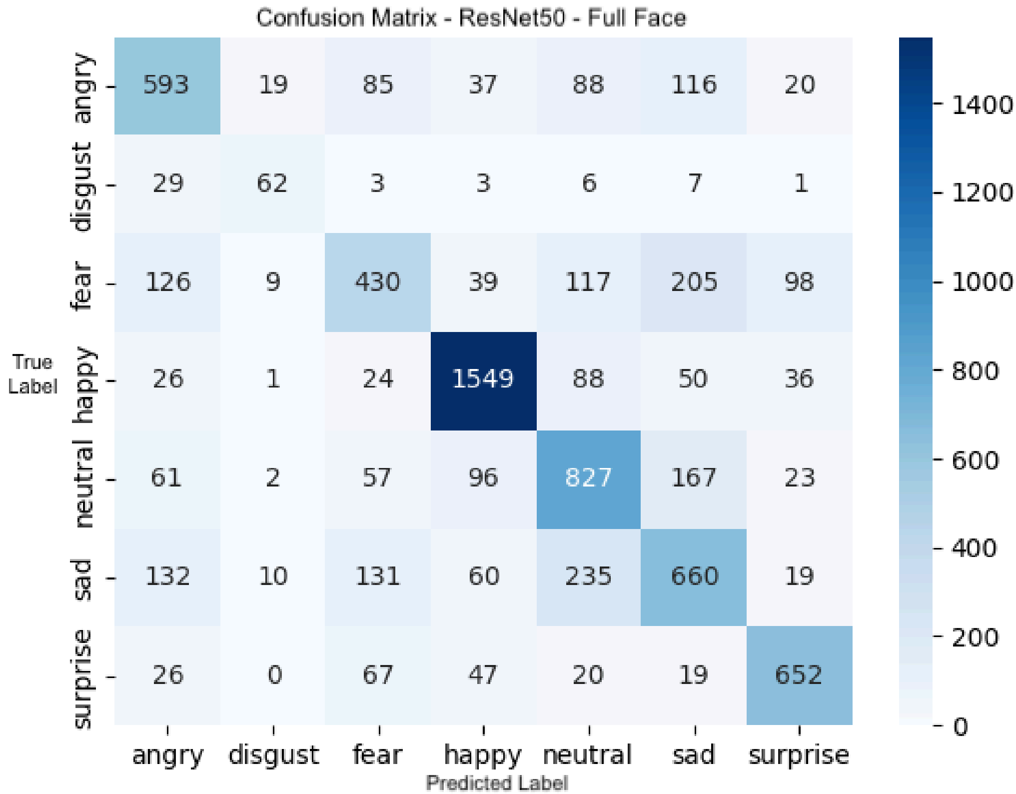

Figure 5: Confusion matrix of RestNet50 model on full face - Testing Results

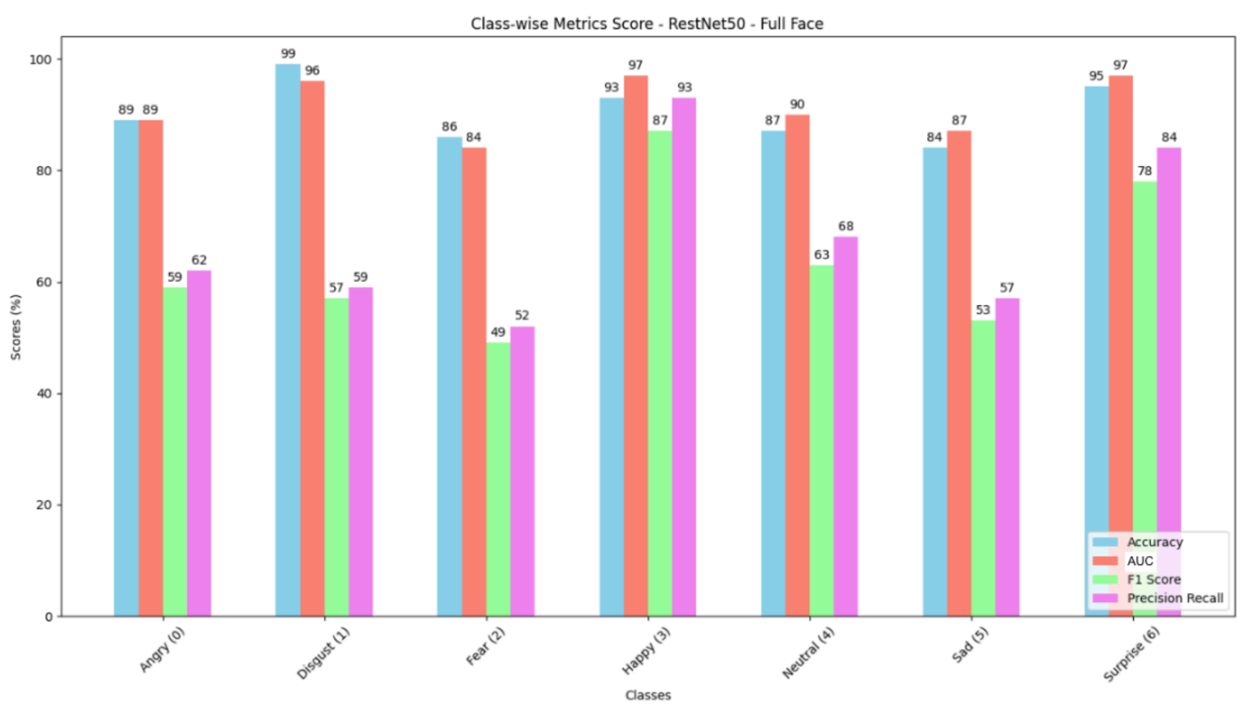

Figure 6: Class-wise metric scores of RestNet50 model on full face - Testing Results

Appendix B: Micro and Macro AUC Scores

Trained On	Evaluation Metric	Evaluated On				
		Full Face	Upper Mask	Lower Mask	Right Mask	Left Mask
Full Face	Micro AUC	0.9306	0.8041	0.6840	0.8011	0.8005
	Macro AUC	0.9140	0.7665	0.6852	0.7761	0.7700
Full Face + Upper Mask	Micro AUC	0.9013	0.8994	0.6088	0.7751	0.7834
	Macro AUC	0.8830	0.8781	0.5713	0.7423	0.7490
Full Face + Lower Mask	Micro AUC	0.8934	0.6824	0.8618	0.7499	0.7220
	Macro AUC	0.8799	0.6718	0.8430	0.7227	0.6981
Full Face + Right Mask	Micro AUC	0.9132	0.7988	0.6913	0.9081	0.7378
	Macro AUC	0.8974	0.7856	0.6850	0.8894	0.7496
Full Face + Left Mask	Micro AUC	0.9138	0.7972	0.6934	0.7699	0.9069
	Macro AUC	0.8977	0.7793	0.6747	0.7322	0.8891

Appendix C: Source Code

The source code and related resources for the computational methods discussed are available in our GitHub repository. The repository includes scripts, datasets, and documentation that illustrate the machine-learning techniques employed.

GitHub Repository: https://github.com/AaryaDesai1/Facial_Emotion_Recognition.git

You can access the repository directly via the link above, or clone it using the following command in your terminal:
git clone https://github.com/AaryaDesai1/Facial_Emotion_Recognition.git

This repository is organized into folders corresponding to each of the main sections of the analysis, ensuring that the materials are easy to navigate.