Classification of Human Facial Expressions

Group #80

Members: Santiago Becerra Cordoba, Sarah Kim, Chloe Seo, Kaylee Vo, Jie Zhao

Data Description and Summary

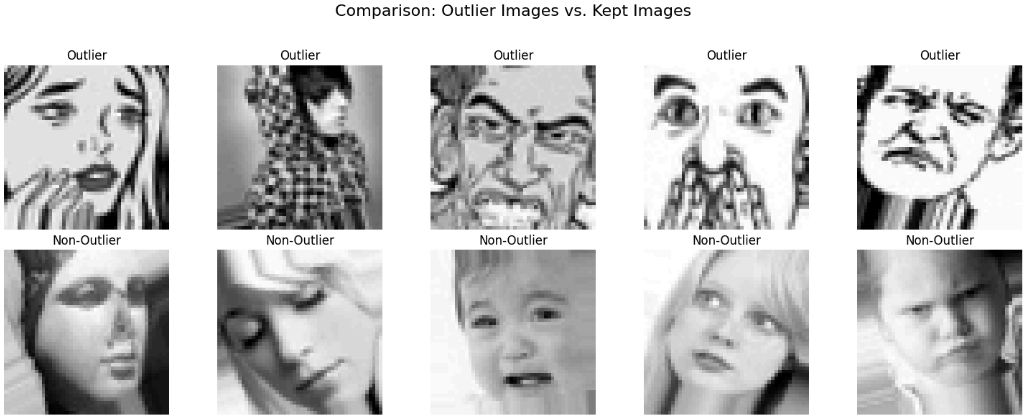
For our project, we aim to classify human facial expressions using the FER-2013 dataset, which contains 35,887 grayscale images of size 48x48 pixels. Each image is labeled with one of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The data itself came preformatted with a train and test set. We performed a validation split on the training data, with 20% of the data allocated to the validation set. After a 20% validation split, we have 22,968 images in the training set, 5741 images in the validation set, and 7178 images in the test set. Figure 1 shows a sample of the data along with the pipeline architecture.

We computed simple summary statistics of pixel values and checked for missing data. Both results are displayed in Table 1. For brevity's sake, this write-up will focus on new additions to the notebook. All referenced figures can be found in the Appendix along with figures from MS2 for the reader's reference.

Data Analysis

We will not cover class imbalance and denoising since both were covered in MS2. Figures 2 and 3 show examples each respectively. Instead, in this section, we will expand on outlier detection.

For outlier detection, we used an autoencoder and were able to successfully identify outliers based on an image's reconstruction error. We chose the 98th error percentile as the threshold for identifying outliers. The model flagged a total of 460 images, which is 2% of the training set by design. By removing them, we reduced our training set to 22,508 images. The sample of outliers show images that are stretched, drawn, occluded or have anomalous facial poses, as shown below.



Using the identified outlying indices we performed offline preprocessing to filter out the anomalous images. Performing this optimization greatly improves model runtime, since we no longer need to filter at every batch call.

Meaningful Insights

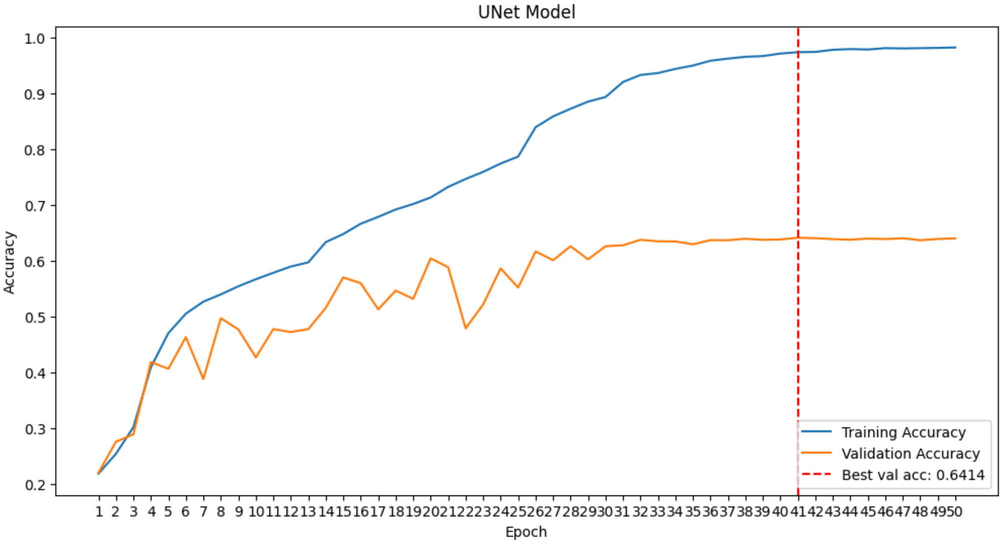
From our data analysis, we observed significant class imbalance in the dataset, especially with the disgust category, which had far fewer training examples compared to other classes like happy or neutral. We will apply methods such as class weighting or data augmentation to mitigate this risk. We found that denoising smoothed the input images while preserving important facial structures to a degree. Our autoencoder was successfully able to filter outliers. Using a threshold of 2%, we dropped images that were anomalous, noisy or uninformative. We exclude these outliers from all downstream training to reduce overfitting and enhance class specific feature clarity. Excluding outliers showed a performance increase in our baseline models of around 4%.

Research Questions

Now that we've established an exhaustive EDA, we present the research questions to guide our empirical analysis. 1) What is the best custom model architecture and strategy for classifying human facial expressions and how well does it perform compared to existing models in the literature? 2) How does the model perform for each emotion class and what are the implications of the results? 3) How do vision transformer models compare to CNN's in terms of performance on FER-2013?

Improved Baseline Model

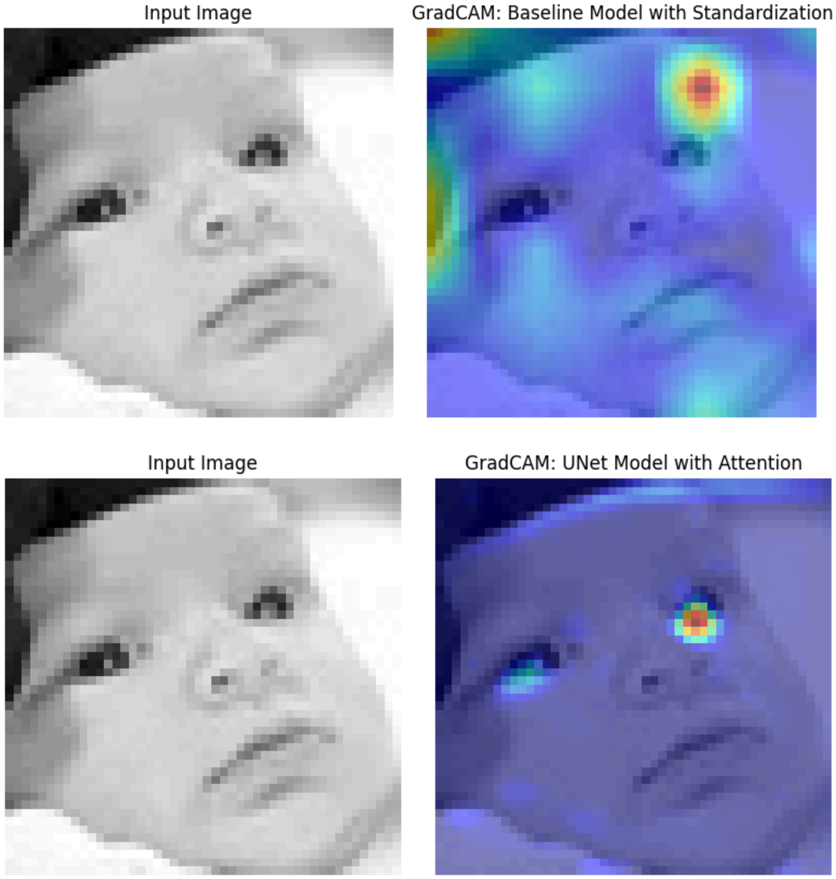
Building on our prior baseline model work, we applied data augmentation to our baseline model (Figure 4a). We used rescaling, rotation, translation, zoom and horizontal flipping. However the results under performed the model without augmentation (Figure 4b). Our breakthrough occurred when we constructed a U-Net model, which significantly improved performance over the baseline, achieving 64% accuracy compared to 52%. This improvement validates our approach of using a more sophisticated architecture with built-in attention mechanisms on skip connections, which helps the model focus on emotion-relevant facial features.



For future milestones, we see different promising avenues to further improve model performance. These include experimenting with ensemble methods combining multiple architectures, using class weights for class imbalance, exploring transfer learning with facial recognition pre-trained models, and investigating more advanced attention mechanisms like transformers.

Visualizations

To better understand how our vision models make predictions, we used GradCAM to generate class activation heatmaps for both the baseline_std_model and u-net model. These visualizations show us which regions of the image the model is attending to when making its classification decision. We choose GradCAM because it is agnostic to the model architecture compared to methods such as CAM. We also experimented with activation maximization techniques like DeepDream but the results were underwhelming.


Input Image
GradCAM: Baseline Model with Standardization
Input Image
GradCAM: UNet Model with Attention

From the images above, we see the baseline model did not learn strong, localized features for the task, relying on more general patterns in the image. In contrast, the u-net model was able to focus strongly on the eyes of the face. This shows that u-net has learned to attend to more relevant and discriminative regions

Summary of Findings

As a group, we built an autoencoder model to successfully detect outliers and filter them from all downstream training tasks. We revised our research question to be more ambitious, including vision transformers as a goal. We also implemented data augmentation on the baseline model to compare different preprocessing techniques. We were able to improve on the baseline with a u-net model, which currently stands at 64% accuracy. Finally, we implemented GradCAM to visualize the model's attention and understand how it makes predictions.

For future work, we hope to build on our momentum and shoot for a model that can predict 75% accuracy. The current SOTA model is able to predict 79.79% accuracy on the test set. We will also try implementing a vision transformer model to see how it compares to CNN architectures. With respect to each of these models, we will try more preprocessing strategies until we land at the optimal preprocessing, model combination.

Appendix

Figure 1a

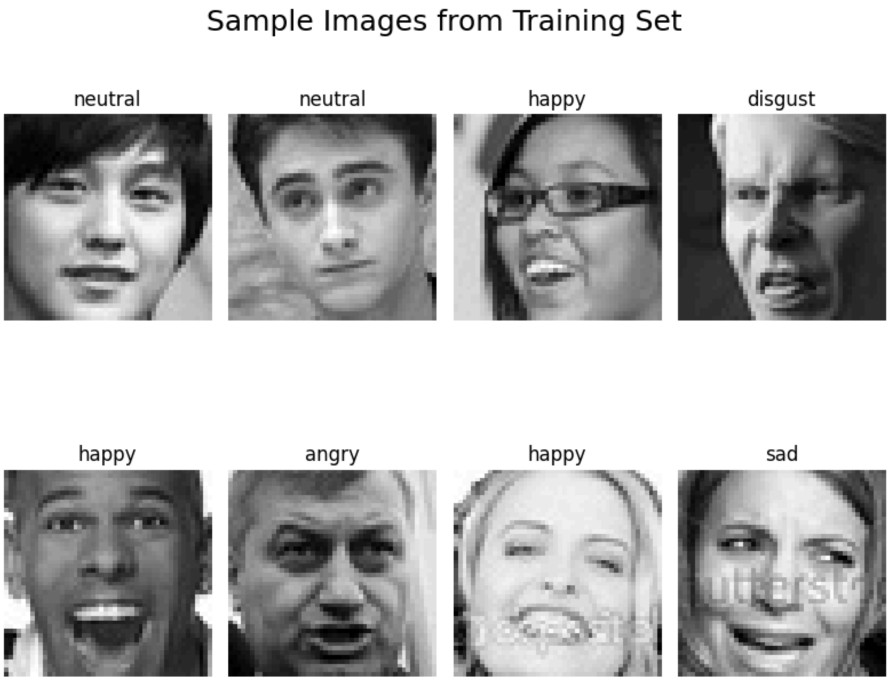


Figure 1b

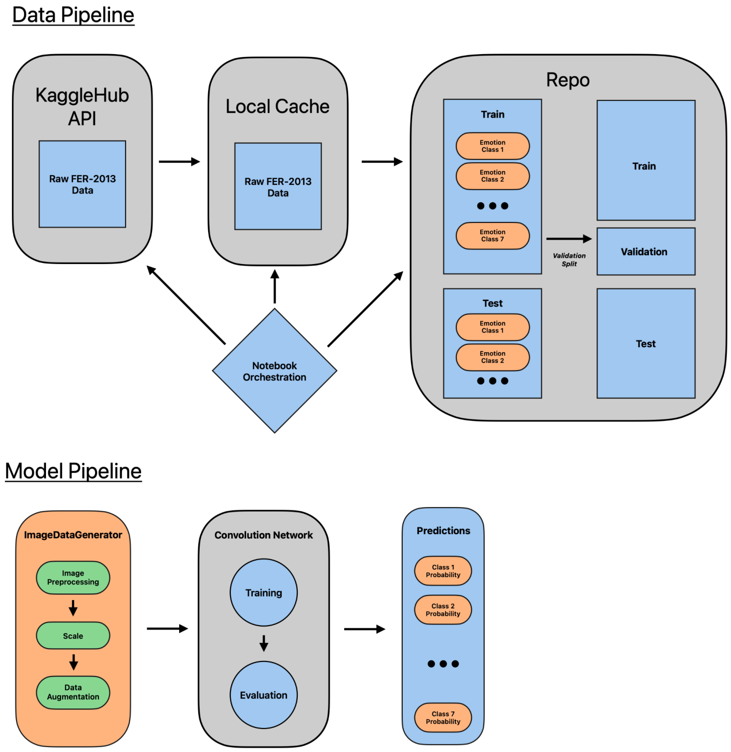
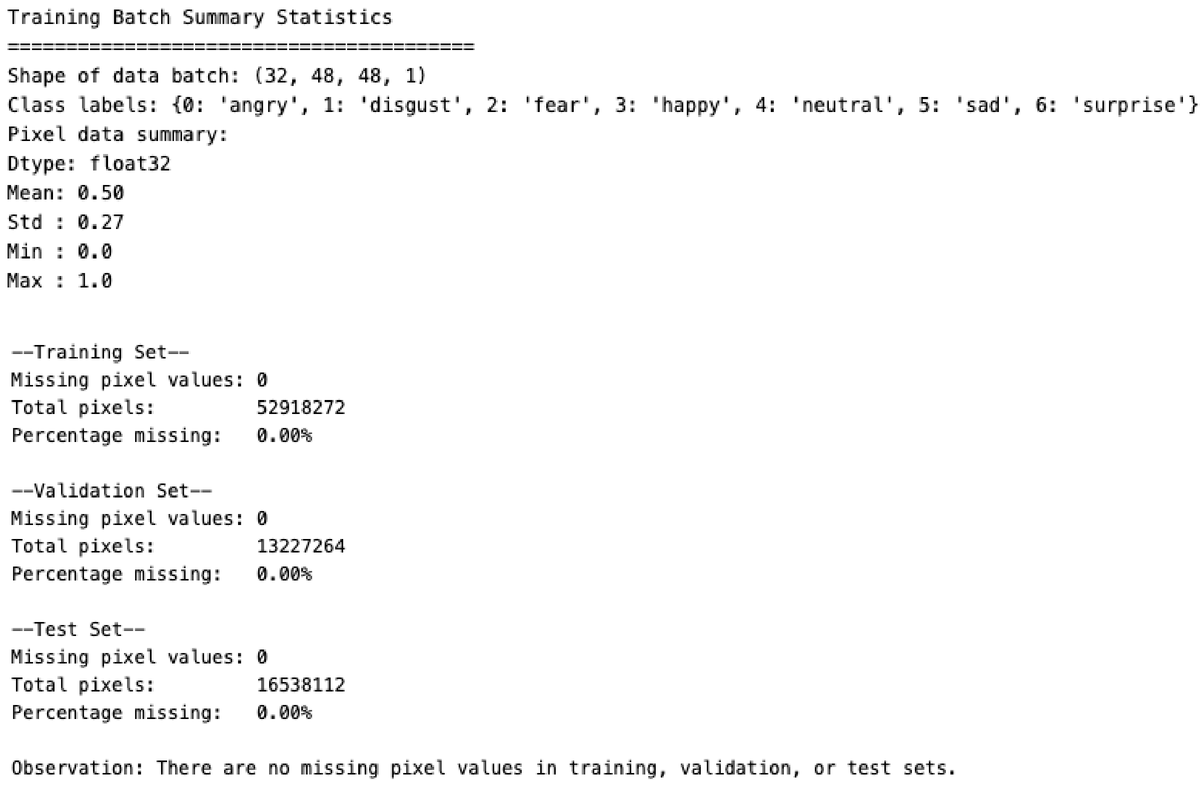


Table 1



```
Training Batch Summary Statistics
========================================
Shape of data batch: (32, 48, 48, 1)
Class labels: {0: 'angry', 1: 'disgust', 2: 'fear', 3: 'happy', 4: 'neutral', 5: 'sad', 6: 'surprise'}
Pixel data summary:
Dtype: float32
Mean: 0.50
Std : 0.27
Min : 0.0
Max : 1.0

 --Training Set--
 Missing pixel values: 0
 Total pixels:         52918272
 Percentage missing:   0.00%

 --Validation Set--
 Missing pixel values: 0
 Total pixels:         13227264
 Percentage missing:   0.00%

 --Test Set--
 Missing pixel values: 0
 Total pixels:         16538112
 Percentage missing:   0.00%

 Observation: There are no missing pixel values in training, validation, or test sets.
```

Figure 2

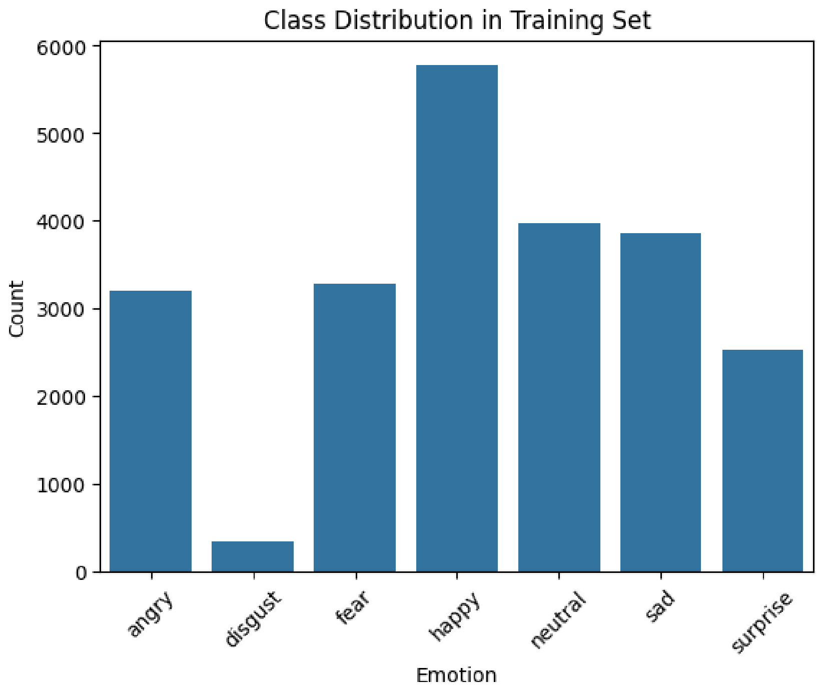


Figure 3

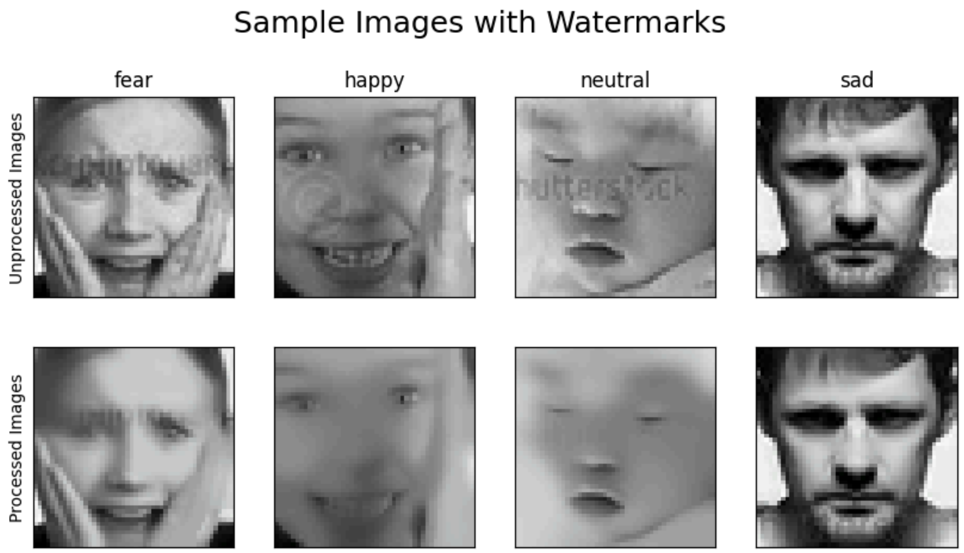


Figure 4a

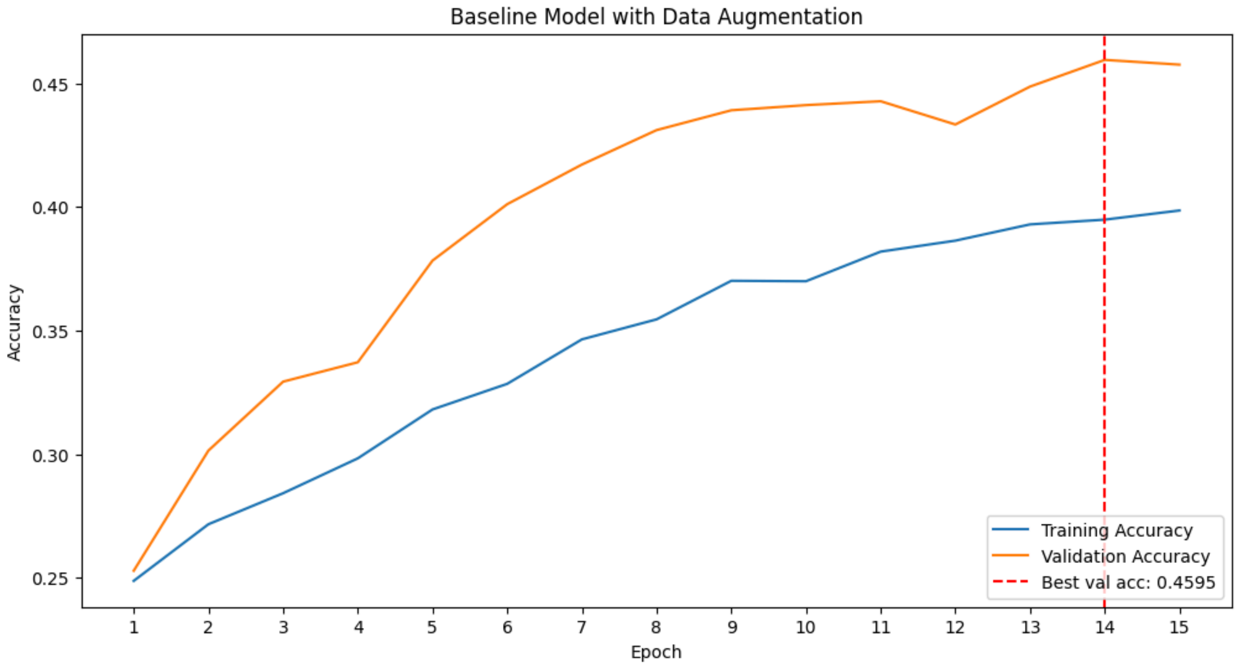
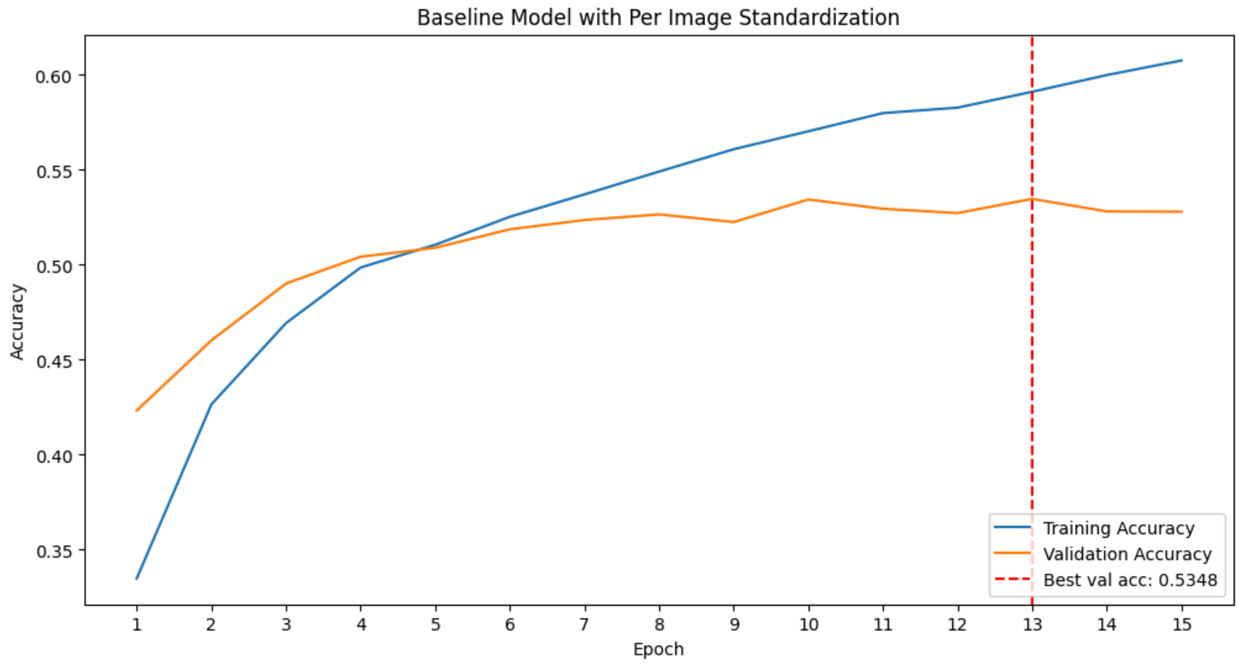


Figure 4b



Sources

Wikipedia contributors. (2025, January 24). *Non-local means*. Wikipedia. https://en.wikipedia.org/wiki/Non-local_means

OpenCV: Denoising. (n.d.). https://docs.opencv.org/3.4/d1/d79/group__photo__denoise.html#ga03aa4189fc3e31dafd638d90de335617

Papers with Code - FER2013 Benchmark (Facial Expression Recognition (FER)). (n.d.). https://paperswithcode.com/sota/facial-expression-recognition-on-fer2013