Facial Expression Recognition Report

Evelyn Liang, Huei Liu, Shelly Wei,Tzuyang Lin, Yuqin Yang

Motivation

Human facial expressions are key to non-verbal communication, yet traditional recognition systems using handcrafted features struggle to generalize across diverse settings. This project aims to develop a scalable deep learning model that classifies emotions from facial images, enabling emotionally aware applications in fields like human-computer interaction, mental health, education, and security.

Data Preprocessing

Data Collection

- FER-2013 publicly available on the FER-2013 Dataset
- **35,000** facial images categorized into **seven** basic emotions, including happy, sad, angry, surprised, fearful, disgusted, and neutral

Data Cleaning

- Grayscale Conversion: Convert all images to **grayscale**
- Face Detection & Resize: Use **MTCNN** to detect face, and resize them to **48x48 pixels**
- Normalization: Normalizes the pixel values to [0, 1]

Data Augmentation

Used **ImageDataGenerator** to generate 5 augmented samples per image with transformations including:

- Random rotation (±10°)
- Width and height shifts (±10%)
- Zoom (±10%)
- Horizontal flipping

Data Preparation

- Splitting Ratio: The augmented dataset was split into training (60%), validation (20%), and test (20%) sets
- Class Balancing: Stratified splitting was applied within each emotion class to maintain label distribution

Dataset Splitting and Class Balance Result:

Emotion	Train	Validation	Test	Total
Happy	23367	7789	7789	38945
Sad	14094	4698	4698	23490

Fear	12213	4071	4071	20355
Surprise	10326	3442	3442	17210
Neutral	16623	5541	5541	27705
Angry	12558	4186	4186	20930
Disgust	1296	432	432	2160
Total	**90477**	**30159**	**30159**	**150795**

Model Building

CNN Model

The implemented CNN processes 48×48 grayscale facial images through three sequential convolutional blocks (32, 64, 128 filters with 3×3 kernels), each incorporating batch normalization, max pooling, and dropout for regularization. The network culminates in a 512-unit dense layer before the 7-class softmax output representing distinct emotional states. Training metrics demonstrate progressive improvement from 25% to 57% accuracy (training) and 30% to 63% (validation), with convergent loss curves indicating effective learning without significant overfitting. Performance analysis via a confusion matrix reveals pronounced class disparities. The model excels at recognizing 'happy' (86.65%), 'surprise' (74.17%), while struggling with 'disgust' (22.92%) and 'fear' (27.09%). Notably, 'disgust' samples frequently register as 'angry', 'neutral', or 'sad', while 'fear' demonstrates substantial confusion with 'angry' and 'sad'.

ResNet Model

The ResNet model was constructed using a series of convolutional layers with residual connections, which help mitigate the vanishing gradient problem and enable the training of deeper networks. The architecture includes batch normalization and activation functions (typically ReLU) after each convolutional layer, promoting stable and efficient learning. The final layers consist of global average pooling, followed by dense (fully connected) layers, and a softmax activation to output class probabilities for each emotion category. The model was trained using cross-entropy loss and the Adam optimizer, with callbacks for early stopping, learning rate reduction, and model checkpointing to save the best-performing weights.

MobileNetV2 Model

The MobileNetV2 model was also built using transfer learning, starting with a pre-trained MobileNetV2 backbone. Initially, only the custom head (top layers) was trained, while the base was frozen. After a few epochs, the top N layers of the base model were unfrozen for fine-tuning with a lower learning rate, allowing the model to adapt more closely to the emotion dataset. The architecture is characterized by depthwise separable convolutions, which make it lightweight and efficient, suitable for deployment on resource-constrained devices. The classification head mirrors the ResNet approach, ending with a softmax layer.

Vision Transformer (ViT) Model

Vision Transformer (ViT) served as the main architecture for our image classification task. We used the vit_tiny_patch16_224 model from the timm library, which splits each 224×224 input image into 196 fixed-size patches (16×16 pixels) and processes them as tokens through a 12-layer Transformer encoder. Each layer uses 192-dimensional embeddings and 3 attention heads, totaling about 5.7 million parameters, making it a lightweight yet expressive model. To adapt it to our task, we replaced the original classification head with a custom NN.Linear layer matching the number of emotion categories. The model was initialized with ImageNet-pretrained weights to support transfer learning and improve performance on our limited dataset.

During training, the model showed a consistent decline in loss, dropping from around 1.90 to 1.25 in the first epoch and reaching 0.2976 by the fifth, with a final training accuracy of 89.31%. This indicates that the model effectively learned meaningful patterns from the data. However, a major drawback was the long training time—completing five epochs took approximately 1,522 minutes (over 25 hours). This highlights the computational demands of ViT models, even in lightweight variants like vit_tiny, making efficiency a key consideration in practical applications.

Model Comparison

Metric / Aspect	CNN	ResNet	MobileNetV2	ViT (Vision Transformer)
Test Accuracy	**62.12%**	58%	42%	**80.43%**
Best Performing Class	Happy (F1: 0.855)	Happy (F1: 0.82)	Happy (F1:0.61)	Happy (F1:0.91)
Poor Performing Classes	Disgust (F1: 0.340), Fear (F1: 0.350)	Disgust(F1: 0.39), Fear: (F1:0.39)	Disgust(F1: 0.39), Fear: (F1:0.39)	Sad (F1:0.72)
Macro F1-score	0.553	0.53	0.31	**0.79**
Weighted F1-score	0.613	0.58	0.38	**0.80**
Computational Time (per epoch)	~29 min	~2min	**~1.3min**	~304 min

Our baseline model, the CNN, achieved a test accuracy of **62.12%**, outperforming both ResNet (~58%) and MobileNetV2 (~42%). This superior performance is likely due to CNN's stronger ability to generalize across common facial expressions, despite challenges in accurately classifying less represented emotions such as *disgust* and *fear*, which showed low F1-scores. While ResNet demonstrated similar strengths in identifying *happy* expressions, it also struggled with the same underrepresented classes. MobileNetV2, though computationally lightweight and

efficient, showed the lowest performance overall, reflecting its limitations in capturing the complexity of nuanced emotions.

On the other hand, the ViT (Vision Transformer) model significantly outperformed all others, achieving **80.43%** validation accuracy along with the highest macro and weighted F1-scores (0.79 and 0.80, respectively). Its strength lies in leveraging global attention mechanisms, which allow it to capture fine-grained patterns in facial expressions. However, this comes at the cost of increased computational demand and longer training time.

Considering these trade-offs, **CNN** offers a strong balance of accuracy and efficiency for resource-constrained environments, while **ViT** stands out as the top performer when computational resources are ample and the highest recognition accuracy is required. Thus, the choice between CNN and ViT should be guided by the specific scenario and system limitations.

Explainability

Grad-CAM visualizations reveal critical insights into the model's decision-making process, showing focused attention on anatomically relevant regions—mouth and cheeks for 'happy' expressions, eyebrows and eyes for 'angry' and 'sad'—while exhibiting diffuse, unfocused heatmaps for 'disgust' and 'fear' that correlate with their lower classification performance and reflect model uncertainty in feature identification; these interpretability findings, complemented by varying confidence levels in predicted class probabilities, demonstrate the model's successful learning of discriminative features for certain emotions while struggling with more ambiguous expressions, suggesting future optimization through targeted augmentation for underrepresented classes and integration of specialized attention mechanisms.

Impact

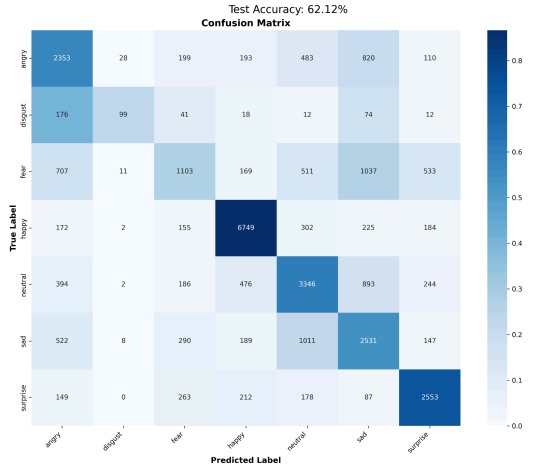
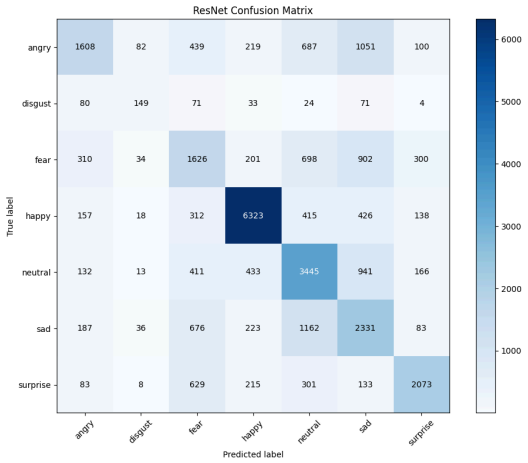
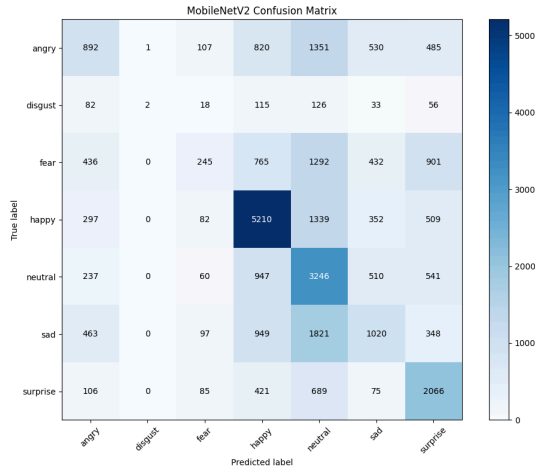
Our selected **Vision Transformer (ViT) Model**, achieves strong performance across multiple emotion classes, demonstrating both high accuracy and interpretability. By dividing images into patches and applying self-attention across the entire image, ViT captures both **fine-grained features** and **global facial context**, such as symmetry, expression structure, and subtle emotional cues. This enables the model to detect nuanced patterns with robustness across a wide range of facial variations, making it highly effective for real-world emotion recognition.

The project enables emotion-aware technologies across industries. In mental health, it supports passive mood monitoring during teletherapy. In Human-Computer Interaction, it enhances virtual assistants and learning platforms by adapting responses based on user emotions. For UX testing, it offers real-time feedback from facial cues, helping designers improve user interfaces more effectively than manual surveys. In future work, we plan to integrate temporal information from video streams and add modalities like voice or posture for improved multimodal accuracy.

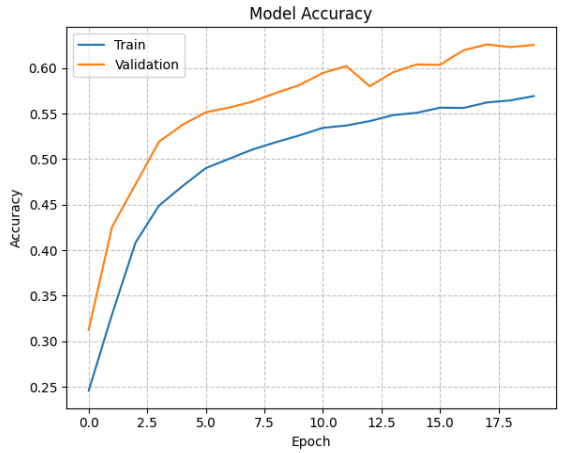
See the attached code at this link for implementation details.

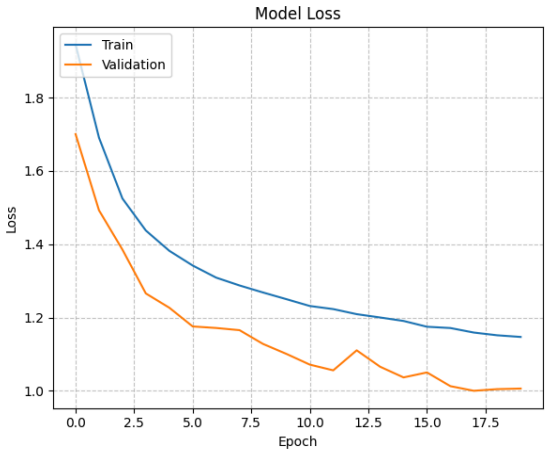
Appendix

Models' Confusion Matrix:

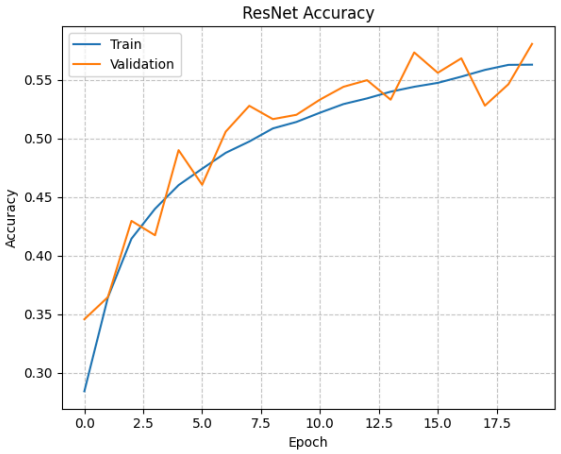
CNN model	ResNet model
	
MobileNet	**ViT model**
	

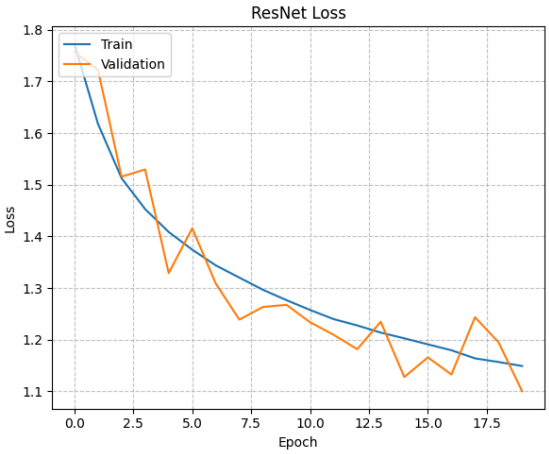
CNN model's Training / Loss curves



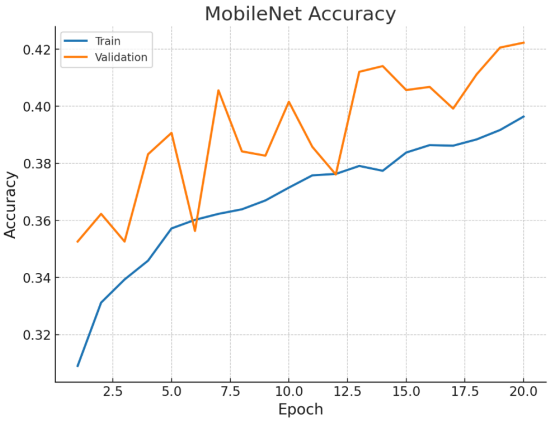


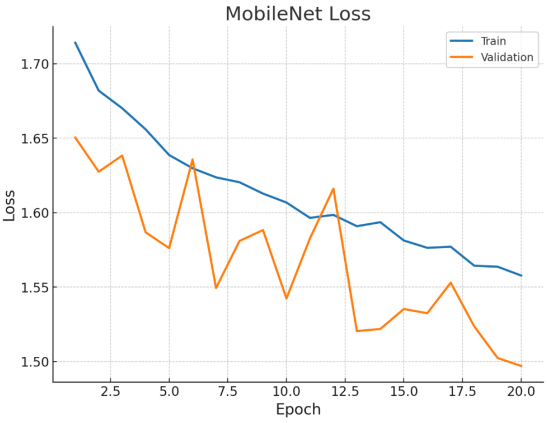
ResNet model's Training / Loss curves



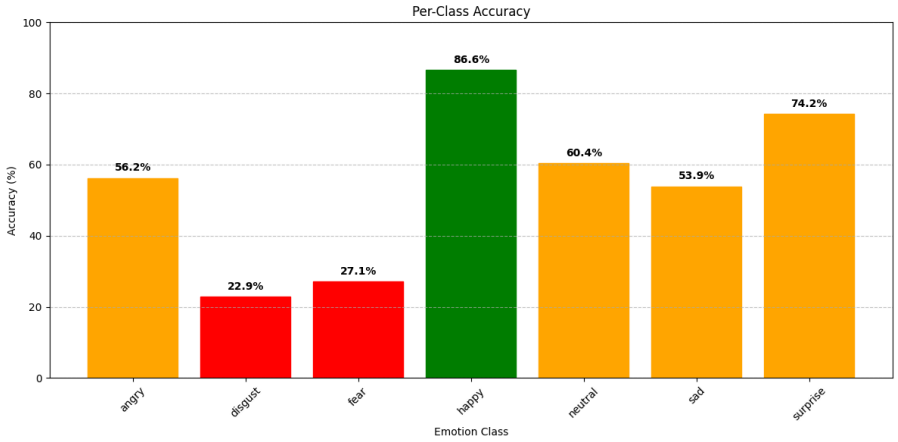


MobileNet model's Training / Loss curves

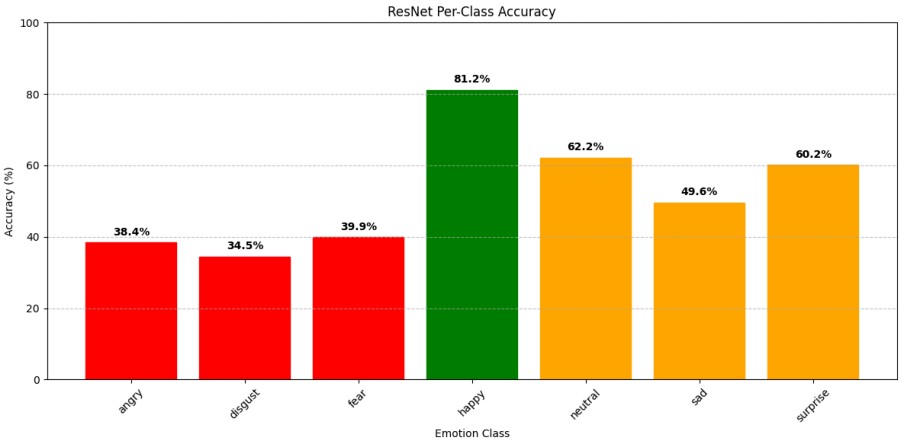




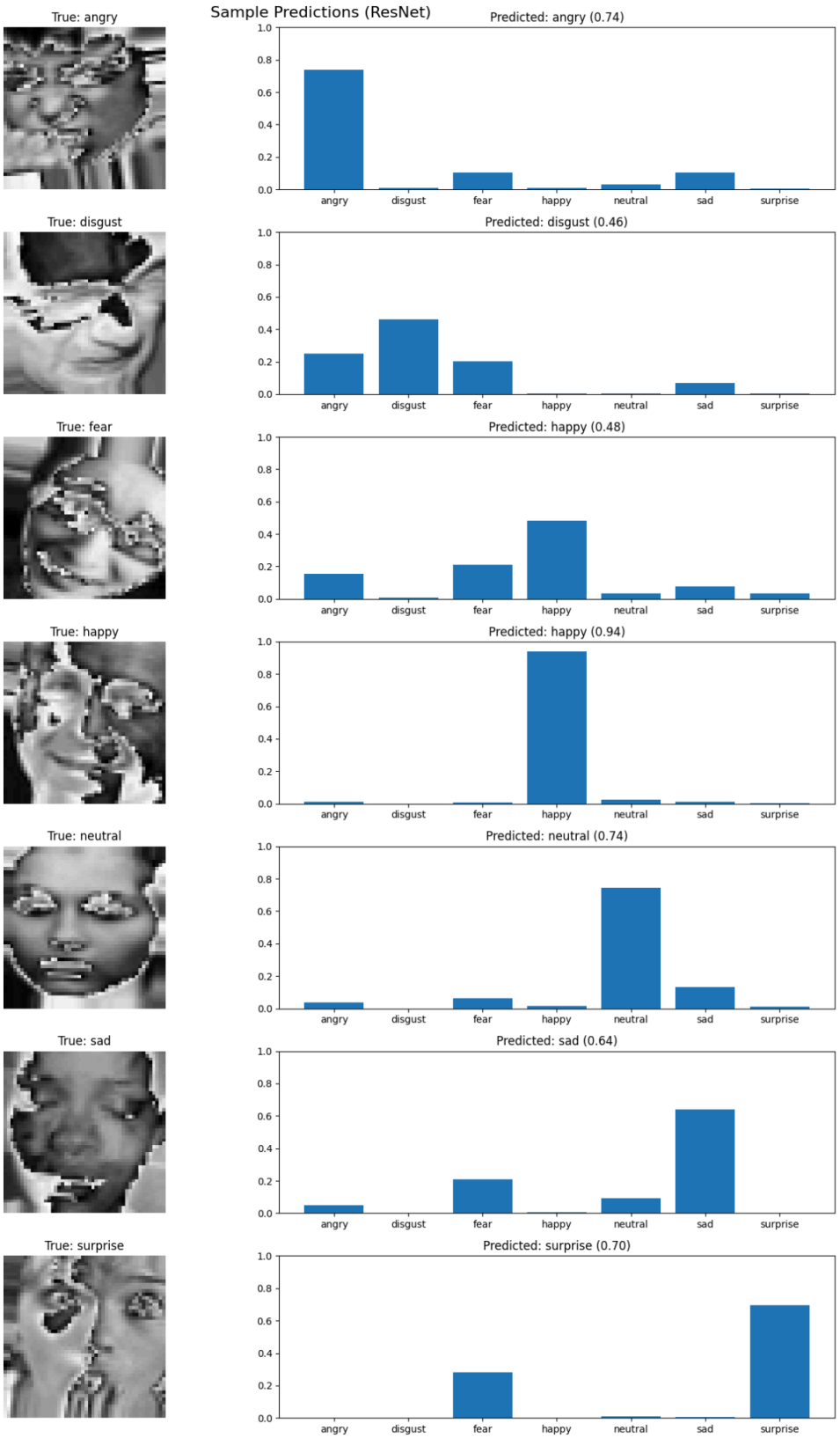
Overall CNN Model Performance by Class:



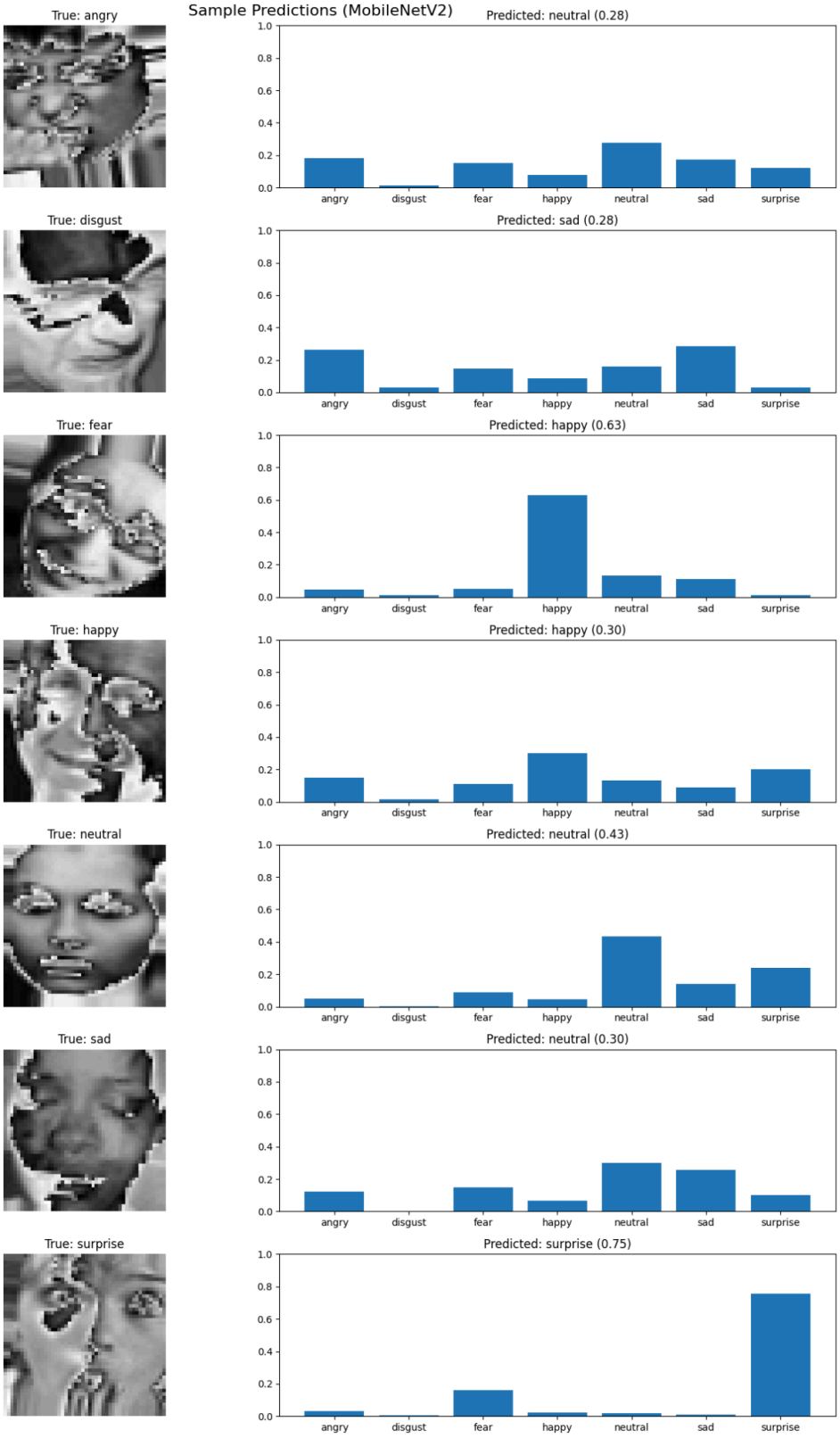
Overall ResNet Model Performance by Class:



ResNet Sample Predictions



MobileNet Sample Predictions



Grad-CAM on CNN