Facial Emotion Recognition

Project Report

Participants:
Nati Shchiglik, Yaniv Kempler, Yuval Bar Levi, Wisam Salameh

Introduction

This is an image classification project where multiple state-of-the-art backbone models were trained to specifically recognize facial emotions.

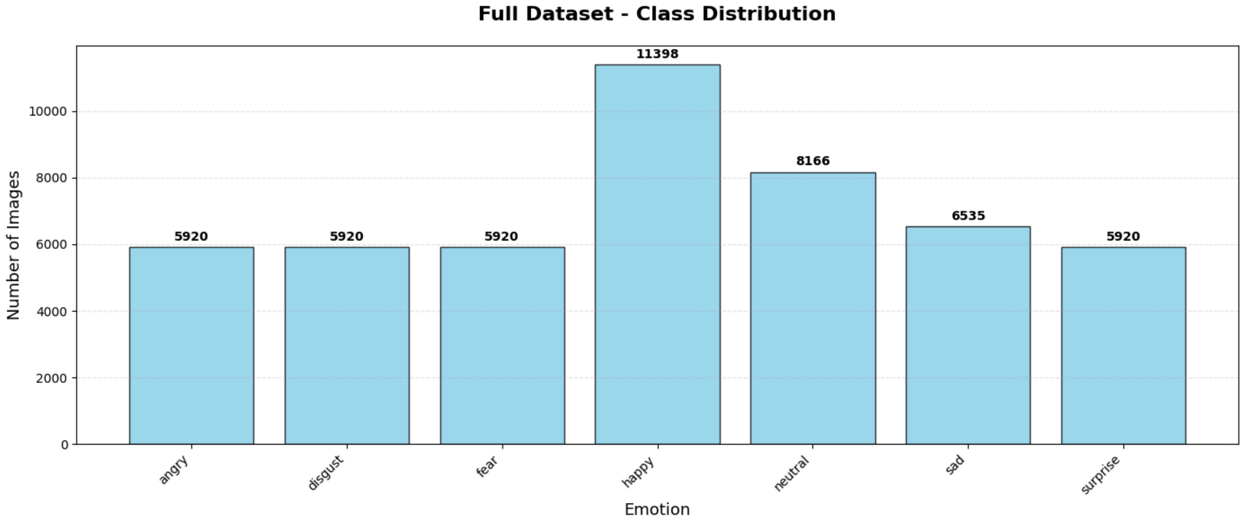
The models were trained using a pre-existing labeled dataset and a comprehensive performance analysis was later conducted to deepen our understanding of strong-weak points of each architecture.

Dataset Description

The dataset used for training is Facial Emotion Recognition dataset from kaggle. This is a preprocessed dataset combining both FER2013 and RAFDB datasets.

The preprocessing included filtering images to include faces only, converting to RGB, and labeling into 7 classes of emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

The following graph shows a count histogram of each class in the dataset, yielding a max/min imbalance ratio of 1.93x:



1. As a result of the imbalance, our first processing step included balancing each emotion class to include 5,920 images only (random sampling).

2. The second preprocessing step included a dataset split of 70% for training - 15% for validation -15% for testing.

3. The third and final preprocessing step included defining the on-the-fly augmentation and input size operators. The input size was left to be dynamic per each model's expected input, while the augmentation techniques used were universal across all trainings and included:

    a. Mandatory rescaling to [0,1] for all images.
    b. Random rotations up to 10 degrees.
    c. Random translation of height and width by a factor of 10%.
    d. Random Zoom by a factor of 10%.
    e. Random Horizontal flip.

The input size and rescaling operators were uniform across training, validation, and test set, while the augmentation operator was defined on the training epochs only.

Dataset Generation

Alongside the dataset from kaggle, we used the ComfyUi platform to generate synthetic data. ComfyUi is based on stable-diffusion, and specifically CyberRealistic XL model. First we wanted to only validate the correctness of the fine-tuned models, but after we started to generate the synthetic images, we thought that it can be an extra step to be able generate around 900 very high detailed and high resolution images, but when we started to generated in hundreds images of (fear, disgust or surprise), we started to see incorrect image generation, so in the end we decided because of the number of the images in the problematic emotions just to validate the inferences, so we don't mistakenly fine-tune our best model on unbalanced synthetic images.

Model Choice

Four backbone models were chosen for our task, VGG16, ResNet, EfficientNet, and a Vision Transformer.

- **VGG16**, **ResNet50**, **InceptionResNetV2 Overview**
    - Everything is done in one pipeline: data loading, training, fine-tuning, evaluating, and comparing.
    - Added a custom classification head:
        - Global average pooling
        - BatchNorm
        - Dropout
        - Dense(256)
        - Softmax(7) for the emotion classes

- **ViT Overview**
    - Specific model: Swin Tiny Patch4 Window7 244 Transformer.
    - The Model was loaded with ImageNet-1k classification weights.
    - The transfer learning head introduced was a single dense layer of 128 neurons with ReLU activation function and a classification layer with a softmax activation function.

Training

The transfer learning for all three models was done by the two step method, where in the first step a higher learning rate was used while freezing all backbone layers for feature extraction, while in the second step the learning rate was decreased and several layers were unfrozen from the backbone model for the final finetuning.

VGG16

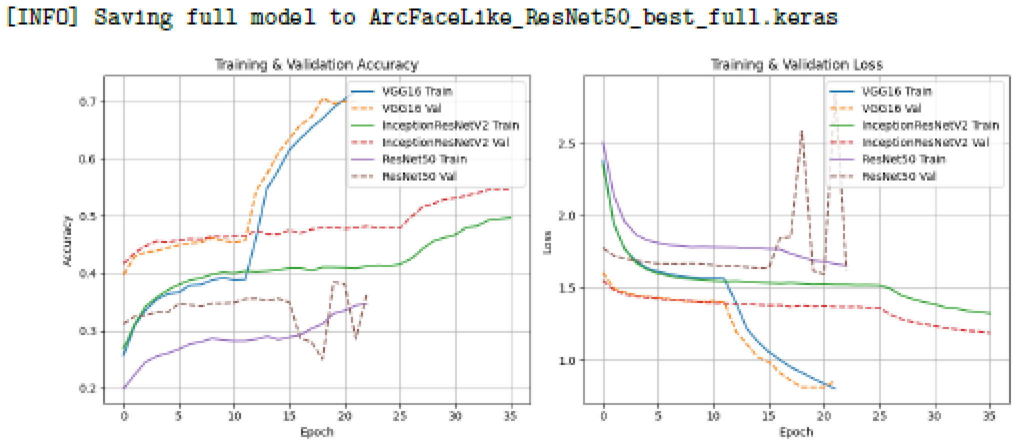
Fine-tuning boosted accuracy from ~45% → 70% on validation.
 This is a huge improvement — VGG16 responded very well.

InceptionResNetV2

Improved from ~47% → 54%
 Fine-tuning helped, but less dramatically.

ResNet50

Barely improved — indicating this architecture didn't match FER2013 patterns well.



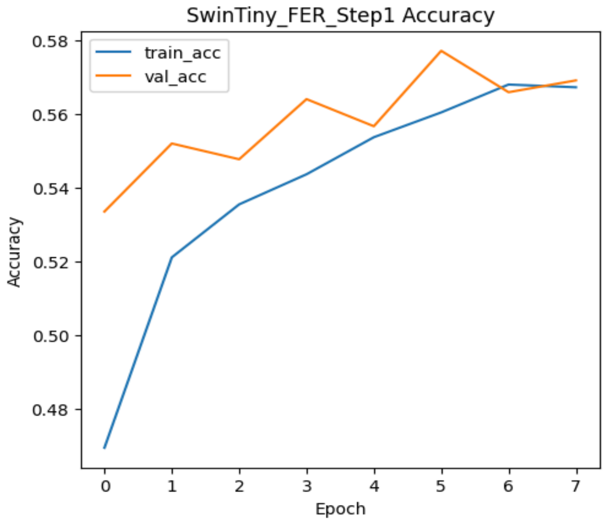
```
[INFO] Saving full model to ArcFaceLike_ResNet50_best_full.keras
```

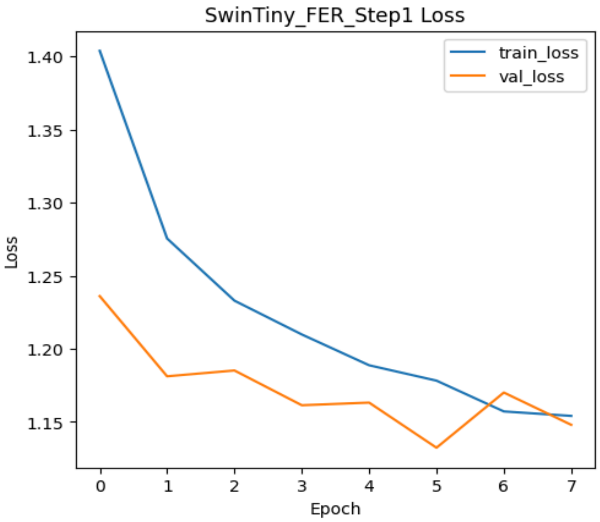
*

* VGG16 had the best, cleanest accuracy curves → best generalization
* InceptionResNetV2 improved but remained moderate
* ResNet50 showed signs of struggle and does not follow the expected curve shape

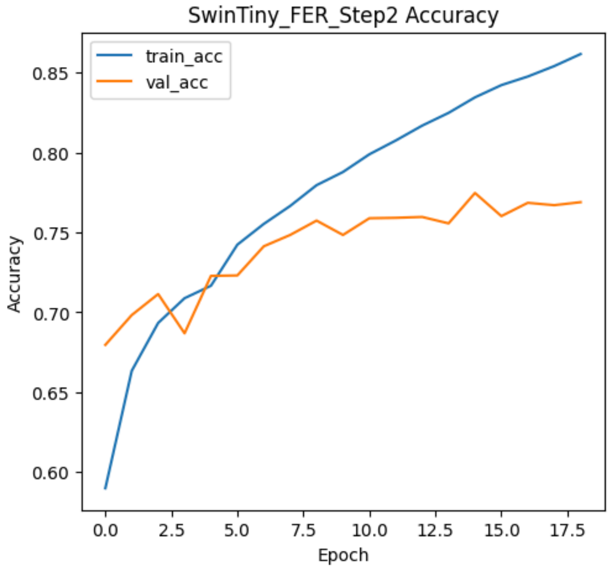
ViT Training

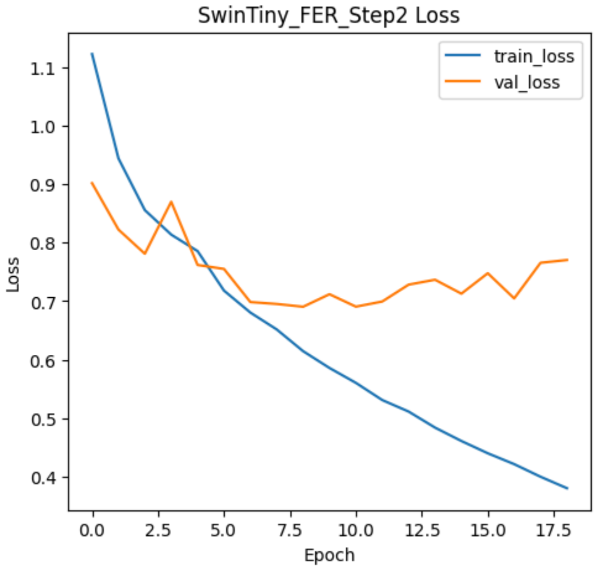
Step 1 - learning rate 0.001, model converged after 8/10 epochs, with val accuracy ~58.9%





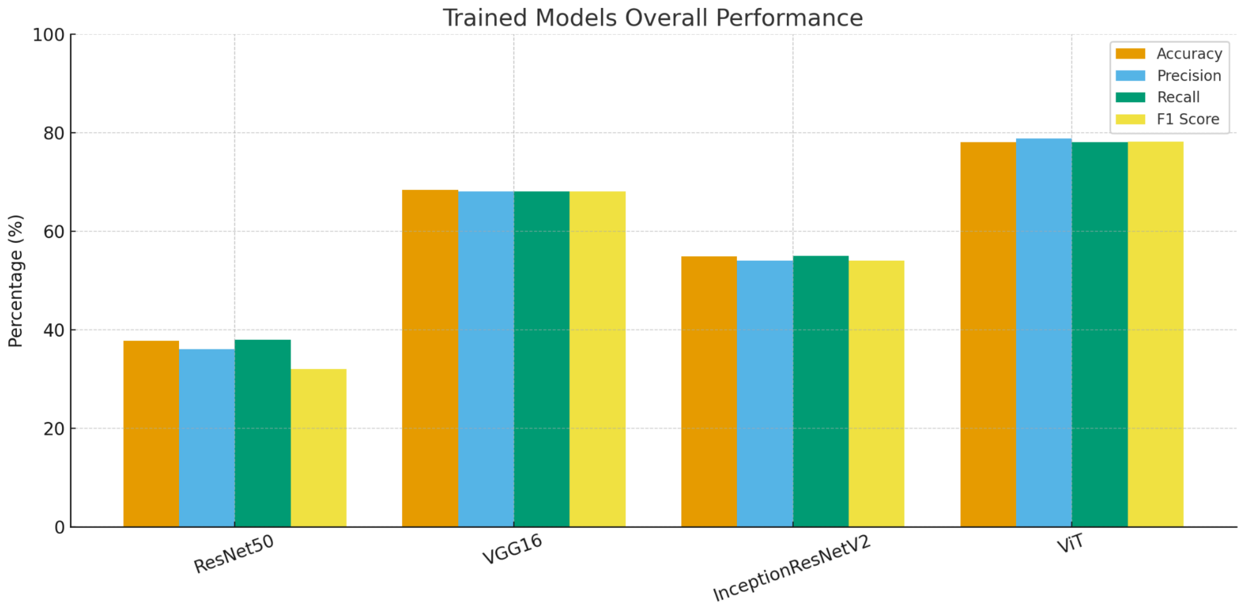
Step 2 - unfrozen 5 backbone layers, learning rate 0.0001, model converged after 15/30 epochs, with val accuracy ~76.9%



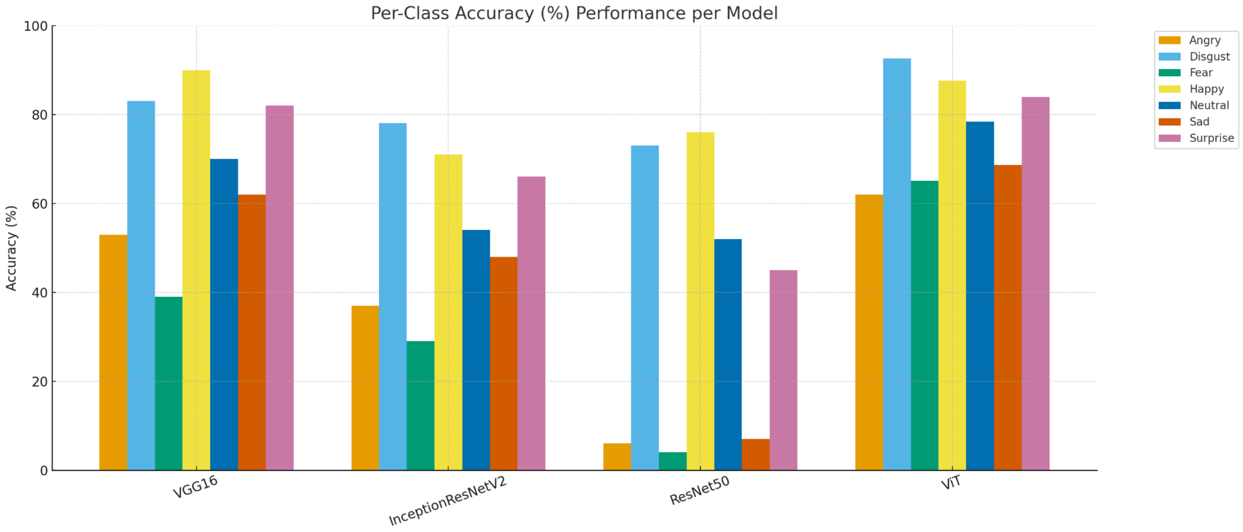


Evaluation

All three models were evaluated on the same test split, resulting in the following table:



Trained Models Overall Performance				
Model	**Accuracy[%]**	**Precision[%]**	**Recall[%]**	**F1 Score [%]**
ResNet50	37.7%	36.0%	38.0%	32.0%
VGG16	68.4%	68.0%	68.0%	68.0%
InceptionResNetV2	54.9%	54.0%	55.0%	54.0%
ViT	78.0%	78.8%	78.0%	78.2%



Per-Class Accuracy[%] Performance

Model	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
VGG16	53%	83%	39%	90%	70%	62%	82%
InceptionResNetV2	37%	78%	29%	71%	54%	48%	66%
ResNet50	6%	73%	4%	76%	52%	7%	45%
ViT	62.0%	92.6%	65.1%	87.6%	78.4%	68.7%	83.9%

Analysis

Summary of key findings

ViT — Best Overall Model

- Highest accuracy: 78%
- Strongest precision, recall, and F1 among all models
- Learns global face structure very well
- Significantly better generalization than CNN-based models

VGG16 — Best CNN Model

- 68.4% accuracy, stable precision/recall/F1
- Very balanced across all emotions
- Surprisingly strong despite being a simpler architecture
- Best-performing convolutional model

InceptionResNetV2 — Moderate

- 54.9% accuracy
- More complex architecture did not help on FER2013
- Reasonable results but worse than VGG16 and ViT
- Sensitive facial expressions (fear, angry) remain hard

ResNet50 — Worst Performer

- 37.7% accuracy
- Struggles heavily without proper ArcFace loss
- Fails on challenging emotions (fear, angry, sad)

Per-Class Accuracy Analysis

ViT dominates across almost all emotions

- Highest accuracy for Angry, Disgust, Fear, Neutral, Sad, Surprise
- Competitive on Happy as well
- Shows that Transformers capture subtle facial cues better

VGG16 performs consistently and strongly

- Best CNN performance for Happy (90%)
- Very solid across all emotions
- Handles facial muscle patterns effectively
- InceptionResNetV2 is mid-range
- Good at Disgust, Surprise
- Weak in Fear, Angry, Sad

ResNet50 clearly underperforms

- Extremely low for Angry (6%), Fear (4%), Sad (7%)
- Indicates the model is mismatched for FER2013 unless redesigned for embeddings

General training discussion

Training Strategy

All models followed a two-phase training approach: Phase 1 – Frozen Backbone for feature extraction and Phase 2 – Fine-Tuning

Model Behavior During Training

- Training accuracy consistently increased
- Validation accuracy improved until convergence
- Fine-tuning brought significant gains, especially for VGG16 and ViT

ViT Specific Behavior:
In ViT training, **no overfitting was detected during Step 1 (initial training)**—training and validation accuracy increased smoothly and remained aligned.
However, in **Step 2 (fine-tuning)**, a **slight overfitting began appearing toward the end of training**.
Because early stopping used a patience of **4 epochs**, training continued briefly after the optimal point.
The best validation accuracy occurred at **epoch 15/30**, after which the model started to overfit.
The **final saved weights were those from epoch 15**, ensuring optimal generalization.

Per-Class Performance

Emotions with strong facial cues (happy, disgust, surprise) consistently produce high accuracy across models while more subtle emotions (fear, sad, angry) remain challenging, especially for CNN-based architectures.

Conclusion

ViT is the new state-of-the-art in our tests, outperforming all CNNs with a large margin.
VGG16 is still extremely strong, simple, stable, and very reliable — best classical CNN.
InceptionResNetV2 is good but not optimal for emotion recognition on low-resolution FER2013 images.