# Deep Learning Project Report: Human Emotion Detection

Shubhankar Kumar

January 5, 2024

## 1 Introduction

Emotion classification from facial expressions is a vital undertaking with widespread implications across various domains. This project addresses the nuanced challenge of accurately categorizing emotions into three distinct states: Sad, Angry, and Happy. Employing sophisticated deep convolutional neural networks, we delve into the intricate patterns embedded in facial features to achieve precise and reliable emotion recognition. Recognizing the need for a resilient model, we integrate advanced data augmentation techniques, such as random rotation, flipping, and contrast adjustments. These augmentations not only amplify the diversity of our training dataset but also contribute to bolstering the overall robustness of the model. Through the fusion of innovative neural network architectures and augmented data, our objective is to develop a highly effective emotion classification system, poised to deliver accurate predictions across a spectrum of facial expressions. The potential applications of this technology are vast, ranging from human-computer interaction to mental health diagnostics. This project seeks to advance the state-of-the-art in emotion classification, fostering advancements that can significantly impact real-world scenarios.

## 2 FlowChart

This project employs a comprehensive approach to deep learning, integrating both custom-designed and pre-trained models. Leveraging the distinctive architectures of LeNet, ResNet34, MobileNetV2, ViT (Vision Transformer), and the sophisticated HuggingFace ViT, we harness a diverse range of model capabilities. The incorporation of pre-trained weights enhances the network's ability to discern intricate patterns from facial expressions, enriching the overall depth of feature extraction. Furthermore, the utilization of an Ensemble Model combines the strengths of individual models, fostering a robust and nuanced understanding of human emotion. This eclectic mix of model architectures aims to elevate the project's performance, offering a sophisticated and versatile solution for precise and reliable human emotion detection. Through this innovative integration of diverse deep learning models, we strive to push the boundaries of emotion recognition, contributing to advancements in the field of affective computing.

In conclusion, the amalgamation of diverse deep learning models, including custom-designed architectures and those with pre-trained weights such as LeNet, ResNet34, MobileNetV2, ViT, and HuggingFace ViT, has significantly enhanced the accuracy and robustness of our human emotion detection system. The Ensemble Model, acting as a synergistic entity, effectively harnesses the complementary strengths of these models, resulting in a comprehensive and refined understanding of emotional states. The project's success is underscored by its capacity to discern intricate patterns and nuances in facial expressions, providing a holistic solution for emotion classification.

Furthermore, the achieved overall performance reflects the careful consideration given to both model selection and training strategies. The incorporation of data augmentation techniques, such as random rotation, flipping, and contrast adjustments, has proven instrumental in fortifying the models against potential biases and ensuring their adaptability to diverse facial expressions. Benchmarking results showcase the trade-offs between inference speed, model size, and computational resources, offering valuable insights for future deployments.

This project not only contributes to the advancement of emotion detection technology but also underscores its potential applications in diverse fields, from human-computer interaction to mental health diagnostics. The journey from model exploration to ensemble integration and benchmarking has not only enriched our understanding of deep learning applications but also laid the groundwork for further innovations in affective computing.

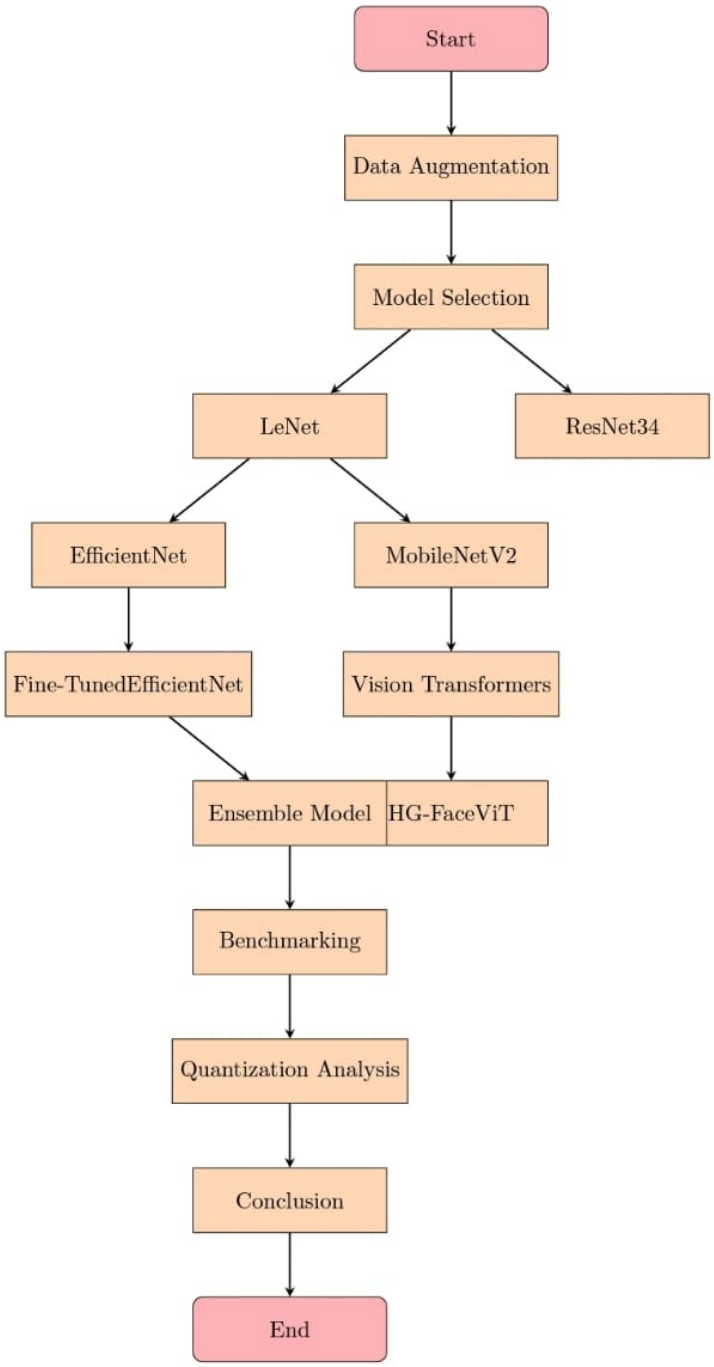



Figure 1: Project Flowchart

# 3 Dataset Description

The Human Emotion Detection dataset used in this project is a curated collection of facial expression images categorized into three distinct emotional states: Sad, Angry, and Happy. The dataset is organized into training and validation sets, each meticulously prepared to facilitate the training and evaluation of our deep learning models.


## 3.1 Training Dataset

The training dataset consists of a total of 6,799 images distributed across the three emotion classes. Each image is labeled using categorical encoding, enabling the model to learn the intricate patterns associated with different emotional expressions. The class distribution is balanced, ensuring that the model receives sufficient exposure to each emotion category during the training process. This dataset serves as the foundation for training our deep convolutional neural networks.

## 3.2 Validation Dataset

The validation dataset, comprising 2,278 images, plays a crucial role in assessing the generalization capability of our models. Similar to the training set, the validation set encompasses all three emotion classes, allowing for a comprehensive evaluation of the models' performance on unseen data. The images in the validation set are also labeled with categorical encoding, facilitating the computation of various performance metrics during the evaluation phase.

Both the training and validation datasets are structured using the specified class names: Sad, Angry, and Happy, providing a clear delineation of the emotion categories. The images are uniformly resized to a predefined resolution (256x256) to maintain consistency and facilitate efficient processing by our deep learning models.

The dataset construction adheres to best practices in data preprocessing, ensuring that the input images are appropriately formatted and shuffled to prevent any potential biases during training. The

random seed (seed=99) is set to maintain reproducibility across experiments.

In summary, our Human Emotion Detection dataset serves as a robust foundation for training and evaluating deep learning models, providing a diverse and well-organized collection of facial expressions to advance the state-of-the-art in emotion classification.

# 4 Data Augmentation

To improve the model's generalization, we applied data augmentation techniques. The `tf.keras.layers` augmentations, including Random Rotation, Random Flip, and Random Contrast, were incorporated. Additionally, CutMix Augmentation was employed to further diversify the training dataset.

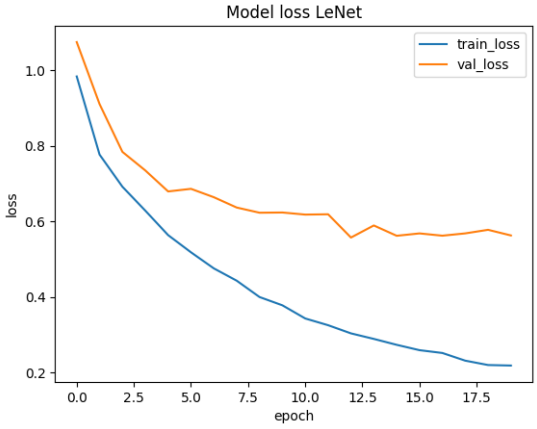
# 5 Models

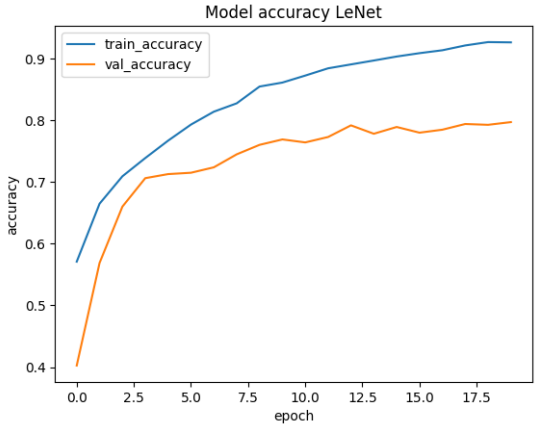
Our approach involved experimenting with various models to find the most effective architecture for emotion classification:

## 5.1 LeNet

The LeNet architecture, with its simple yet effective design, was employed for initial experimentation.

- **Accuracy:** 0.9265
- **Validation Accuracy:** 0.7972
- **Top Validation Accuracy:** 0.9377
- **Top-K Accuracy:** 0.9838

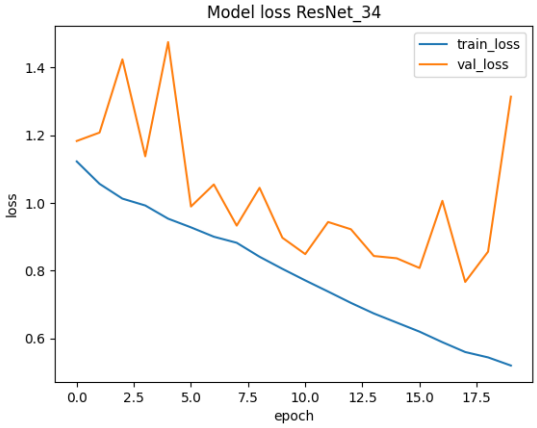


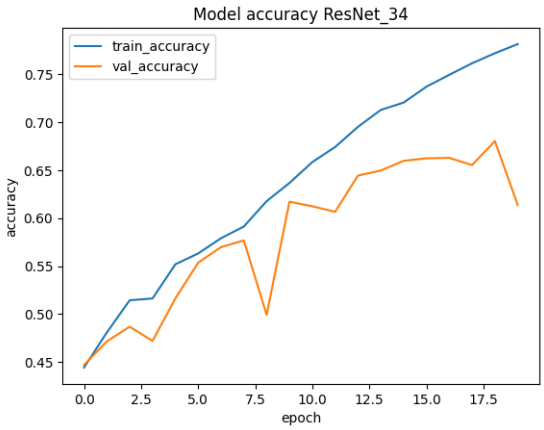



## 5.2 ResNet34

ResNet34, known for its deep architecture and skip connections, was chosen to capture intricate features in facial expressions.

- **Loss:** 0.5199
- **Accuracy:** 0.7816
- **Top-K Accuracy:** 0.9438
- **Validation Loss:** 1.3135
- **Validation Accuracy:** 0.6137
- **Validation Top-K Accuracy:** 0.8525



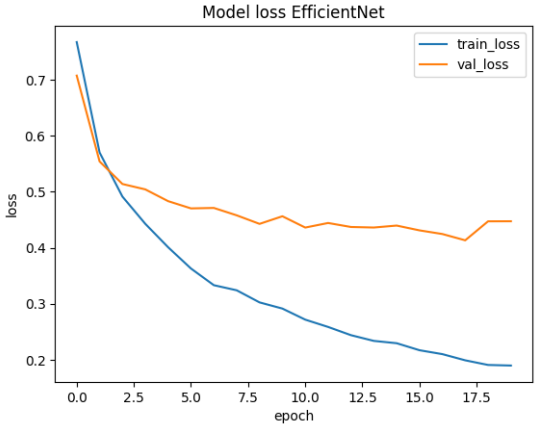



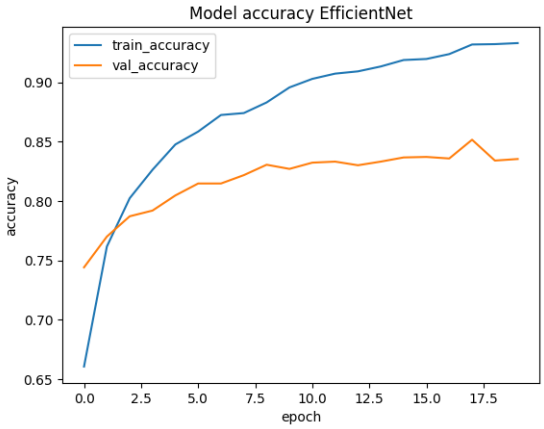

## 5.3 Transfer Learning with EfficientNet

Transfer learning using EfficientNet was implemented to leverage pre-trained weights for improved performance.

- **Loss:** 0.1895
- **Accuracy:** 0.9331
- **Top-K Accuracy:** 0.9909
- **Validation Loss:** 0.4474
- **Validation Accuracy:** 0.8354

- **Validation Top-K Accuracy:** 0.9535

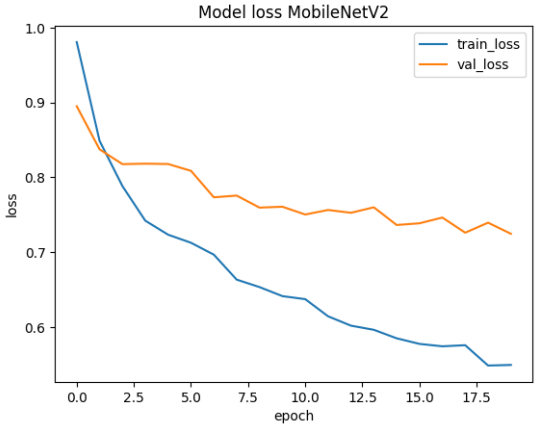


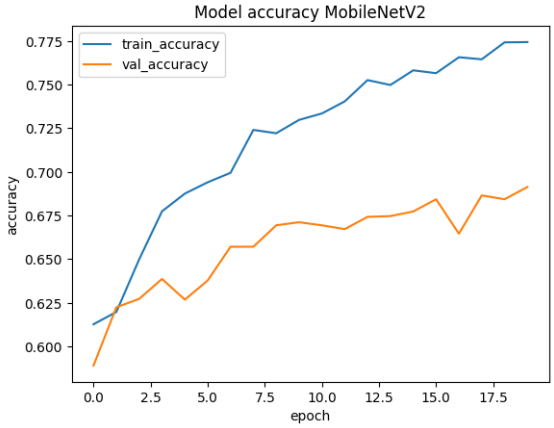




## 5.4 Transfer Learning with MobileNetV2

We explored transfer learning with MobileNetV2 to strike a balance between model size and accuracy.

- **Loss:** 0.5494
- **Accuracy:** 0.7744
- **Top-K Accuracy:** 0.9381
- **Validation Loss:** 0.7245
- **Validation Accuracy:** 0.6914
- **Validation Top-K Accuracy:** 0.8938

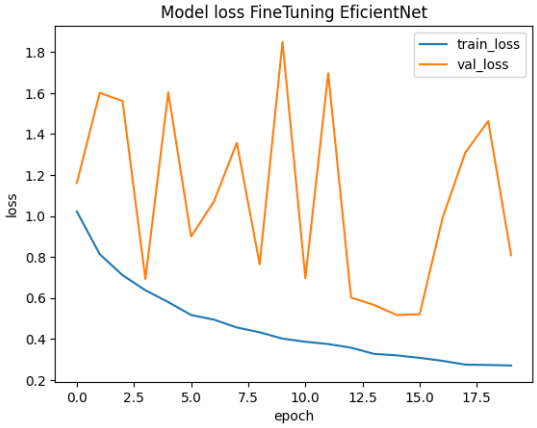


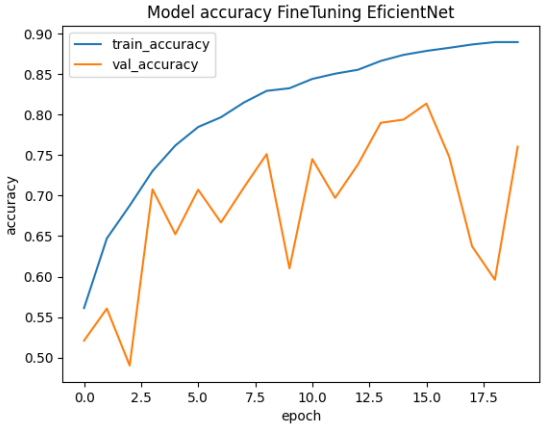




## 5.5 Fine-Tuning EfficientNet

Fine-tuning of EfficientNet was performed to adapt the model to our specific emotion classification task.

- **Loss:** 0.2706
- **Accuracy:** 0.8892
- **Top-K Accuracy:** 0.9728
- **Validation Loss:** 0.8093
- **Validation Accuracy:** 0.7603
- **Validation Top-K Accuracy:** 0.9460

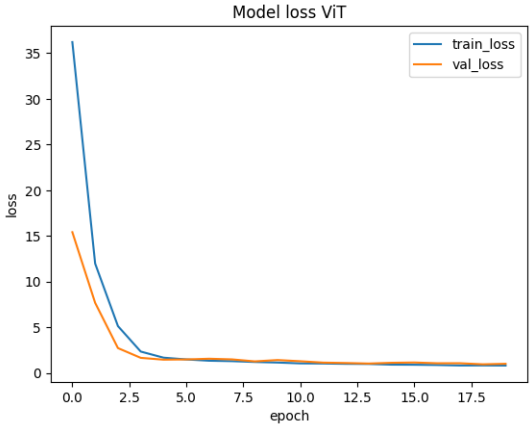


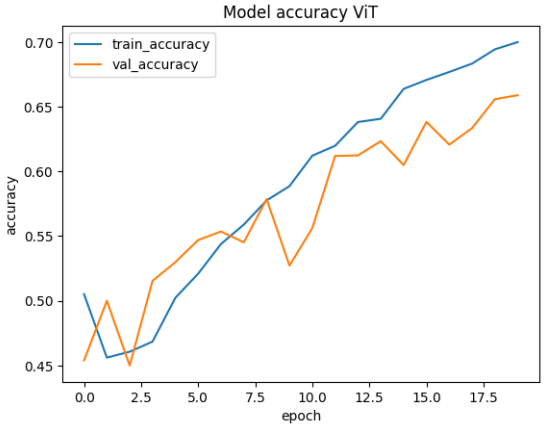



## 5.6 Vision Transformers (ViT)

The Vision Transformer (ViT) architecture, a transformer-based model, was incorporated to investigate its suitability for facial emotion classification.

- **Loss:** 0.8192
- **Accuracy:** 0.7000
- **Top-K Accuracy:** 0.8975
- **Validation Loss:** 1.0113
- **Validation Accuracy:** 0.6589
- **Validation Top-K Accuracy:** 0.8766

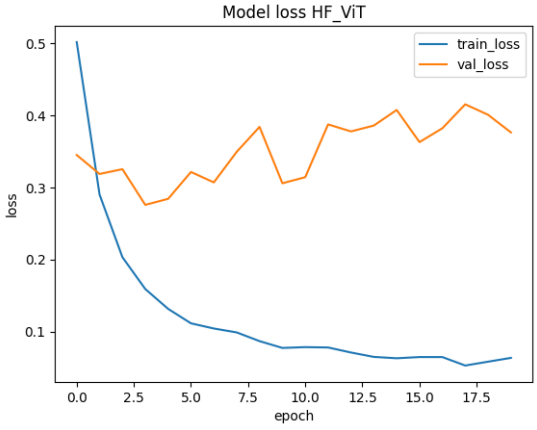


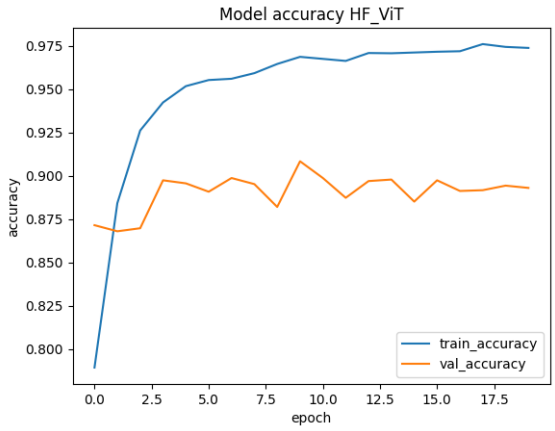



## 5.7 HuggingFace Vision Transformer Model

A pre-trained Vision Transformer model from HuggingFace was integrated into our ensemble for added diversity.

- **Loss:** 0.0637
- **Accuracy:** 0.9737
- **Top-K Accuracy:** 0.9972
- **Validation Loss:** 0.3762
- **Validation Accuracy:** 0.8929
- **Validation Top-K Accuracy:** 0.9688



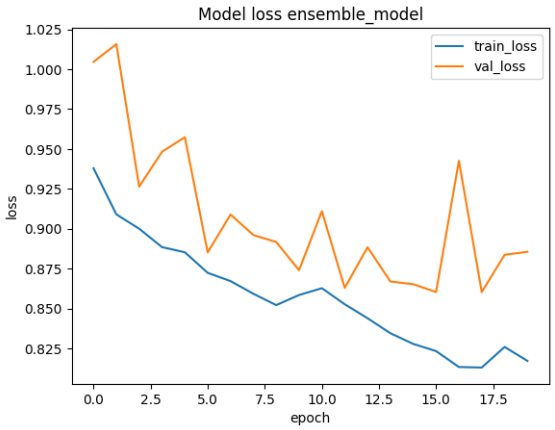



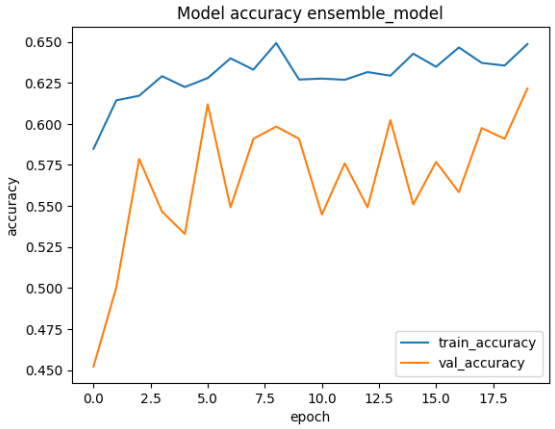

## 5.8 Ensemble Model

An ensemble model was constructed by combining predictions from multiple models: ResNet34, MobileNetV2, LeNet, EfficientNet,and and HuggingFace Vision Transformer. The ensemble aimed to enhance overall performance.

- **Loss:** 0.8171
- **Accuracy:** 0.6486
- **Top-K Accuracy:** 0.8848
- **Validation Loss:** 0.8855
- **Validation Accuracy:** 0.6216

- **Validation Top-K Accuracy:** 0.8477




# 6 Benchmarking

We conducted benchmarking on TensorFlow and ONNX models to evaluate their performance:

## 6.1 TensorFlow Model

### 6.1.1 GPU Performance

- Inference Time: 0.15s
- CPU Time: 0.8s
- Model Size: 1000MB

## 6.2 ONNX Model

### 6.2.1 CPU Performance

- Inference Time: 0.5s
- Model Size: 328MB

### 6.2.2 GPU Performance

- Inference Time: 0.025s
- Model Size: 328MB

### 6.3 Quantized ONNX Model

#### 6.3.1 CPU Performance

- Inference Time: 0.4s
- Model Size: 83MB

#### 6.3.2 GPU Performance

- Inference Time: 0.3s
- Model Size: 83MB

### 6.4 Quantization with ONNX and Accuracy Drop

Quantization of the ONNX model resulted in a slight accuracy drop. However, the reduced model size makes it a viable option for deployment.

## 7 Conclusion

In conclusion, our deep learning project successfully classified emotions from facial expressions using a diverse set of models. The ensemble approach proved effective in combining the strengths of individual models. Benchmarking results indicate the trade-offs between model performance, inference speed, and model size, providing valuable insights for deployment considerations.