National College of Ireland

# Analiz ing Limitation Of Pre-Trained Deep Learning Model For Facial Emotion Recognition

MSc Research Project
Data Analytics

Harshavardhan Kajipuram
Student ID: x23182717

School of Computing
National College of Ireland

Supervisor: Harshani Nagahamulla

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** ……. …Harshavardhan kajipuram……

**Student ID:** ……x23182717……..……

**Programme:** ……MSC Data Analytics…… **Year:** ……2024-2025…….

**Module:** ……MSC Research Project ……..……

**Supervisor:** ……**Harshani Nagahamulla** ……..……

**Submission Due Date:** ……11-12-2024……..……

**Project Title:** Analyzing Limitations of Pre-Trained Deep Learning Models for Facial Emotion Recognition……

**Word Count:** ……6123…… **Page Count**……20……..

I certify that the information in this (my submission) pertains to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** … Harshavardhan kajipuram ……

**Date:** …11-12-2024……

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |

| | |
|---|---|
| Date: | |
| Penalty Applied (if applicable): | |

# Analyzing Limitations of Pre-Trained Deep Learning Models for Facial Emotion Recognition

Forename Surname

Student ID

**Abstract**

Deep learning facial emotion recognition is one area of research that has attracted considerable interest in recent years because of its usability across healthcare, human-computer interaction, and security. This work evaluated the drawbacks of using pre-trained deep learning models, especially the VGG16, towards FER applications. The research hypothesis was to establish the rigidity factors that limit model generalization and examine how performance can be enhanced with more realistic test set restrictions, including class imbalance, diversity, and variations in facial expressions.

The VGG16 model was fine-tuned and extensively trained using the FERC dataset, which consists of seven emotion classes: militant, Disgust, fear, happiness, neutral, sadness, and surprise. The hyperparameters in the models were tuned using the learning rate, dropout rate, batch size, and the type of optimizer used (Adam, SGD, RMSprop). While the training accuracy exceeded 80%, the validation accuracy stagnated at 32.0%, indicating significant overfitting.

The study identified key limitations, including dataset quality, class imbalance, and the complexity of subtle emotional features. Recommendations include exploring alternative architectures such as ResNet, EfficientNet, and Vision Transformers (ViT), enhancing datasets through augmentation and balancing, and incorporating advanced evaluation metrics like precision and F1-score. This research highlights the need for further improvements to address overfitting and ensure robust performance in practical applications.

## 1. **INTRODUCTION**

### **1.1 Research Background**

Facial emotion recognition has grown to be a vital part of the subsegment of computer vision for many fields, including medicine, security, human-computer interaction, and robotics. During the past decade, deep learning, mainly with the use of Convolutional Neural Networks or CNNs, has dramatically enhanced the accuracy of facial emotion detection. For example, models including VGG16, ResNet, EfficientNet and EmotionNet have shown impressive performance by automatically extracting features from image data in a spatial hierarchical way, outperforming traditional methods, including HOG and LBP.

Despite these advancements, challenges remain in **generaliz**ing pre-trained deep learning models to real-world data. Variations in lighting conditions, facial occlusions (e.g., masks, glasses), and demographic diversity (gender, age, or ethnicity) limit the generalizability of these deep-learning models. Furthermore, these models tend to overfit when trained on specific datasets, making them unreliable when tested with diverse conditions. These challenges limit the real-world application of facial emotion recognition systems. Therefore, this study focuses on understanding the limitations of pre-trained models such as VGG16 for facial emotion

recognition, determining what causes those models to fail at generalization, and how to address this issue.
Recognition, determining what causes those models to fail at generalization, and how to address this issue.

### 1.2 Why VGG16 Was Chosen

In this paper, the VGG16 architecture was selected because it is a well-structured model with high efficiency. It has been widely adopted in many image classification tasks, including facial recognition and emotion detection. This model architecture has a deep and uniform architecture, consisting of 16 layers but mainly 3×3 convolutional layers and max-pooling layers, allowing effective hierarchical feature extraction. Compared to deeper models like ResNet, VGG16 is computationally more tractable and, therefore, is a good model to be implemented in practical applications with limited computational resources.

VGG16 has also shown awe-inspiring performance on benchmark datasets and often serves as a good baseline when checking facial emotion recognition tasks. Besides being practical, it has some conspicuous limitations: high memory consumption and vulnerability to overfitting, especially when trained on small and imbalanced datasets. Understanding such constraints allows targeted improvements; for example, regularization techniques and tuning hyperparameters will enhance generalization performance.

### 1.3 Research Question and Justification Research Question:

"What are the primary limitations in applying pre-trained deep learning models like VGG16 for facial emotion recognition, and how can these limitations be addressed to improve generalization?"

**Justification**
VGG16 has shown high accuracy in Benchmark data sets in many learning paradigms. However, when applied to diverse or real-world conditions, their performance deteriorates due to several key limitations:

Overfitting: Major issues include the models tending to memorize features specific to the training data rather than learning a general pattern and performing relatively poorly when new data is presented.

Dataset Imbalance: Many facial emotion datasets have some or significant class imbalance issues, wherein there are vast representations of some emotions, such as happiness, while others, such as Disgust, are rarely captured.

### 1.4 Contribution to Knowledge
The concept of this research is to improve the understanding of deep learning and facial emotion recognition in the following ways:

Identifying Model Limitations: This paper discusses the generalization difficulties of pre-trained models in detail, including the problems that deep learning models like VGG16 encounter in facial emotion recognition.
Proposing Solutions: Adopt and assess algorithms, like hyperparameter tuning. Two techniques were also studied that are commonly applied to prevent overfitting, namely data augmentation and regularization techniques, to use on the model.

### 1.5 Structure of the Dissertation

This dissertation is organized as follows:
Introduction: This section highlights the research background, research question, research justification, and the potential contribution of the study.

Literature Review: A survey of prior works that introduced deep learning models for facial emotion recognition, their advantages, drawbacks, and future directions.

Dataset Collection and Preprocessing: How the datasets were obtained, how the data were prepared for input to the algorithms (e.g., resizing, normalization), and how the imbalance problem was solved.

Model Development and Tuning: A talk on how to work with the VGG16 model, what strategies to fine-tune the model, how to select hyperparameters, and how to prevent overfitting.

Model Training and Evaluation: Discuss the results obtained in the models by providing the observed accuracy and validation accuracy while addressing the observed shortcomings.

Results and Discussion: Report the experimental outcomes by assessing the model failings and the influence of incorporated enhancements.
Conclusion and Future Work: Finally, generalizations, suggestions for future studies, and implications of the studies are briefly discussed as a conclusion.

## 2. LITERATURE REVIEW

### 2.1 Introduction

With deep learning techniques, especially pre-trained CNNs such as VGG16, ResNet, and MobileNet, facial emotion recognition performance has considerably improved. These have increased the possibility of detecting facial expressions and genuine emotions more accurately. Nevertheless, limitations persist in generalizing them into natural data due to challenges such as overfitting, class imbalance, demographic bias, and environmental changes. This chapter reviews the literature on deep learning by highlighting the current advancements in the field, the challenges faced in pre-trained models, and the techniques proposed to overcome such limitations.

## 2.2 Deep Learning for Facial Emotion Recognition

Most recent developments in facial emotion recognition have been due to deep learning methods, particularly CNNs. Pre-trained models such as VGG16 by (Simonyan Zisserman, 2014) employ hierarchical feature extraction for emotion classification. For example, (Parkhi et al., 2015) demonstrated that VGGFace could cope well with different variations, such as lighting and pose. Likewise, Mollahosseini et al., 2016 proposed EmotionNet, a CNN-based model that yielded impressive results in FER but required large-scale labelled data and computational resources.

While pre-trained CNNs allow the reduction of training time based on transfer learning, they generalize poorly. To address this challenge, it was in (He et al., 2016) that the development of what is called Residual Networks (ResNets) became important, leading to deeper models being improved with residual links for stability in training. Applications of such architectures still often face difficulties in managing occlusions, articulations, and expressiveness of faces. Therefore, a new concept of Capsule Networks (CapsNets) was proposed by (Sabour et al., 2017), where spatial hierarchies are protected and robust features to facial occlusions, misalignments or other variations are improved.

While these improvements were obtained, the generalization for pre-trained models such as VGG16 is still a problem while testing on datasets containing real-world variability, and more investigation is needed.

## 2.3 Overfitting and Generalization Challenges

Overfitting is a severe limitation of deep learning models on the facial emotion recognition problem, especially when the data is scarce or imbalanced. (Mollahosseini et al., 2016) Observe that CNNs most commonly memorize training data, producing extremely high training accuracy but yielding poor performance on unseen data. This problem still increases in emotion detection tasks, where rare expressions like "disgust" are underrepresented.
(Zhang et al.,2019) stated that the VGG16 models tend to overfit when exposed to some datasets lacking in diversity. The data augmentation technique of rotation, flipping, and brightness adjustment assist models in generalizing reality by simulating the variability of the data (Kortylewski et al., 2019). In addressing the overfitting issue, (Luo et al.,2020) suggested regularization techniques like dropout and weight decay on deep learning models.

## 2.4 Dataset Limitations and Preprocessing

It was learned that credibility, variability, and richness of the database are pivotal factors for facial emotion recognition model generalization. FER-2013 and LFW often include biases in emotions or demographics that do not represent wide ranges of variations from one another (Martono et al., 2016) examined higher-level emotions. AffectNet was released recently as one

of the largest, with more than a million images labelled. The models trained on the affected did worse in the several conditions tested.

Data preprocessing is vital in handling challenges on the dataset's side. Normalization, resizing, and data augmentation enhance the input consistency and the model immune system. For example, (Mollahosseini et al.,2016) showed that normalization makes an algorithm less sensitive to lighting conditions, and augmentation enhances the algorithm's generalization. However, utilizing Generative Adversarial Networks for data synthesis helps balance the underrepresented classes (Kortylewski et al., 2019).

### 2.5 Recent Advances in Model Architectures

Some of the new developments in deep learning applications are related to efforts to search for alternatives to traditional CNNs to avoid generalization and overfitting issues. In a more recent study (Sahan et al., 2021), CNN-RNN models were employed, and the hybrid models included both spatial and temporal to enhance the recognition of dynamic emotions. Likewise, Graph Convolutional Networks (GCNs) (Yan et al., 2020) modelled relationships between the facial landmarks; the additional accuracy is prominent for fine-grained emotional expression.

It has been a common approach to implement attention mechanisms in facial emotion recognition to attend to the most important features. (Zhang et al.,2021) introduced attention mechanisms with CNNs to overcome problems such as occlusion arising during feature extraction of the facial regions. Further, transformer-based models, for instance, the Vision Transformer (ViT) by (Dosovitskiy et al.,2021), showed more striking improvements compared to CNNs, particularly in solving those tasks that require long-range dependencies. Besides, the model (Le, 2019) named EfficientNet leverages neural architecture search coupled with compound scaling to raise the bar higher with computational efficiency. EfficientNet yielded better accuracy with higher generalization than traditional CNNs, and it can be considered one of the promising alternatives for facial emotion recognition. Vision Transformers (ViTs) use in facial emotion recognition is also gaining traction, as they leverage self-attention mechanisms to capture complex spatial relationships between facial features, improving recognition performance in challenging conditions.

These are promising architectures, yet their potentials are still restricted due to overfitting and imbalance in the data on pre-trained datasets such as VGG16. The necessity of hyperparameter tuning and robust evaluation supports such facts.

### 2.5 Ethical and Practical Challenges

One of the major ethical concerns in using pre-trained models in facial emotion recognition problems is bias and data privacy. (Buolamwini and Gebru, 2018) Demonstrated significant racial and gender biases in commercial models, where performance declined for darker-skinned individuals. (Gong et al.,2021) suggested using domain adaptation methods as remedies for these biases to achieve equal performance across demographics.
Besides, practical deployment also faces the challenge of handling privacy. Differential privacy (Abadi et al., 2016) techniques and federated learning (McMahan et al., 2017) aim to tackle

data privacy issues so that the models are securely trained. However, most of the above approaches have remained underexplored for emotion recognition using pre-trained models.

### 2.7 Summary

This review, therefore, focuses on the advantages and disadvantages related to the use of VGG16 and other pre-trained deep-learning models for facial emotion recognition. These models are excellent for feature extraction from an image, but over-fitting, unbalanced datasets and biases still pose challenges. Some solutions are given by recent developments of hybrid architectures, data augmentation, and regularization, but these issues should be researched. Besides, newly emerging architectures such as EfficientNet and Vision Transformers are promising alternatives that could help generalize and be robust. Overcoming such shortcomings remains indispensable for enhancing applicability and generalization in deep learning models.

## 3. RESEARCH METHODOLOGY

This research, therefore, attempts to assess the deficiency of the pre-training methods, which in this context uses deep learning models, VGG16 for FER, and point out some overfitting problems and develop much better generalization. A tight workflow was put forward for this study, including data gathering, preparation, modelling, and optimization, with assessment.

### 3.1 Data Gathering

#### 3.1.1 Dataset Description

This dataset was chosen for this study due to its structured nature and labelled categories of facial expressions, including Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise in the FERC Dataset. The data is organized into two folders: train and test, containing 35,887 images. The class representing Disgust is underrepresented, possibly leading to prejudice during model formation.

#### 3.1.2 Data Collection and Storage

The dataset was downloaded from the repository Kaggle, renowned for public data sets. Great precautions were taken to adhere to licensing ordinances and the ethical issues of using face images. Making data publicly available allowed data origin and usage to be open, enhancing the work's reproducibility.

The format of the downloaded data set was definite, containing two different folders for training and test data. Each folder was subdivided into subfolders corresponding to seven emotion classes: anger, Disgust, fear, happiness, neutral, sadness, and surprise—even the preliminary labelling of images allowed convenience in sorting and categorizing according to the classification hierarchy.

The structured organization facilitated efficient data loading and processing during the training phase. For example:

Training Data: Images to train the model were kept in a different folder so that the model could catch the features that belong to each class.

Testing Data: Testing was carried out by creating an independent test folder. It allows independent judgment by a model of things it had never seen before.

### 3.2 Data Preprocessing

Proper preprocessing was critical to ensuring data consistency, improving model performance, and addressing class imbalances. The following steps were implemented:

#### 3.2.1 Data Cleaning

**Image Integrity Check:** All corrupted, mislabeled, or unreadable images were removed from the dataset to ensure that only clean and usable data was used for training. This eliminated noisy or incomplete data that could negatively affect the learning process.

**Normalization:** All images' pixel values were normalized to range 0 – 1. Normalization helps scale values so that the distribution of data is more balanced. When training a model, the input values do not grow very large, allowing the model to converge faster.

**Resizing:** The images were also scaled to reduce the size of 224 X 224 pixels to fit the entry size of the VGG16 cast stone model. Resizing brings the dimensions of input into a norm that is compatible with the model size of the pre-trained model and adequate for computations.

#### 3.2.2 Data Augmentation

To improve generalization and simulate real-world conditions, dynamic augmentations were applied during training:

Rotation: Regarding rotation angles, I chose random angles between -20° and +20° to cover head tilts and different poses. This ensured the model's realism when tested on authentic images.

Horizontal Flipping: Different mirrored images were produced to create appearances of routine modifications to test their ability to discern symmetrical facial expressions.

Zooming: Random zoom within a range of 0.2 was introduced as a variant of the model, which slightly influenced the pixel-level details and helped the model identify the features at different zoom levels.

Shifts: While translating the faces in the x and y directions, up to 10% horizontally and vertically healthy displacement assured that minimal positional alterations did not impede the model.

### 3.3 Model Development

### 3.3.1 Model selection

The decision on which model to implement was based on the desire to identify the seven emotions correctly and do so in real-world scenarios. The following architectures were chosen and implemented:

**Convolutional Neural Networks (CNNs)**

The well-known artificial three-layered perception network, CNNs CNNs, was chosen because of its high accuracy in image recognition tasks. The fact that CNNs can train raw image data to differentiate between higher-level features was also an advantage in this research. This study adopted a deep CNN approach relying on the VGG16 model, and the pre-trained model was fine-tuned for the emotion recognition task. CNN comprised more convolutional layers fol, lowed by several pooling layers to minimize the dimensionality, and fully connected layers completed the final classification of emotions.

**Justification for CNNs**
CNNs are particularly effective at capturing spatial hierarchies in images, making them the most suitable architecture for facial recognition tasks. Using a pre-trained model (VGG16), we leveraged transfer learning to accelerate training and improve model performance on the relatively small FERC dataset.

### 3.3.1 Model Architecture
**Base Layers**: The convolutional layers from VGG16, pre-trained on the **ImageNet dataset**, were frozen. These layers retained their ability to extract general features such as edges, shapes, and textures without retraining.

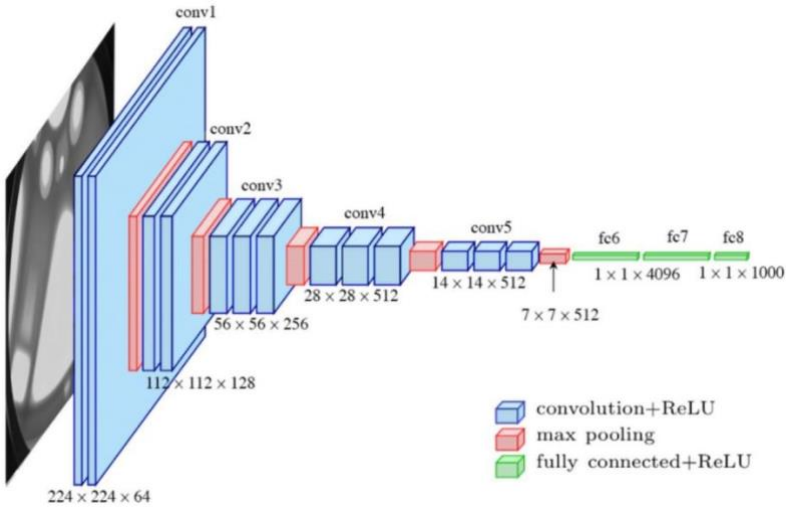
**Custom Top Layers**: To adapt VGG16 for facial emotion classification, additional layers were added:
A **flattened** layer to convert spatial features into a one-dimensional vector.

A **Dense** layer with **512 units** and **ReLU activation** to enable higher-level feature extraction.

A **dropout** layer with rates between **0.3 and 0.7** prevents overfitting by randomly turning off neurons during training.

A final **Dense layer** with **softmax activation** to classify facial expressions into the **seven emotion categories**.

**Figure 1: Showing vgg166 model architecture**




### 3.4 Model Training

The models will be trained using high-performance computing resources such as GPUs to train acceleration and leverage online platforms like Google Collab.

Why HPC? Training a deep learning model like VGG16 involves many computations since the dataset is large and the model architecture is complex.

GPU Usage: The training involved employing graphics processing units, which support parallel computation, hence decreasing training time for a factor compared to the central processing units.

Platform: Training was done using Google Colab, a popular cloud-based platform with free GPU access, enabling training without highly computationally powerful hardware.

The following are the training parameters:

**Optimizer (Adam)**

The Adaptive Moment Estimation-Adam optimizer was used because it automatically adjusts the learning rates for each parameter at runtime. The Adam optimization algorithm combines the ideas of adaptive learning rates with those of momentum, which smooths gradient updates. This allows it to converge faster and more efficiently. By its nature, it is well-suited for sparse gradients. For this reason, it will be very effective for our application in recognizing facial emotion. Other optimizers, such as RMSprop and SGD, were tried before choosing the best parameters for this model.

**Learning Rate Scheduling**

Here, the initial learning rate was selected as 0.001 so that the model can learn fast in the initial stages of training. The approach followed then is a strategy of learning rate decay, which is fed with a reduction factor of 0.1 after every ten epochs. Decreasing the learning rate helps fine-tune the model to converge well and avoid overshooting during training.

**Batch Size: 4**

A batch size of 4 was used, meaning four images were processed before updating the model weights. A moderate batch size keeps the memory load on the GPU in check. Smaller batches mean more minor updates more frequently, speeding up training with a stable gradient descent.

**Loss Function: Cross-Entropy Loss**

The loss to be used measured the categorical cross-entropy between the predicted probabilities and accurate class labels. It is a default choice of loss for multi-class problems because it tries to make models penalize predictions when those turn out wrong, thus improving their accuracy with iterations.

### 3.5 Model fine-tuning

As the vgg16 model initially showed overfitting in training, extensive hyperparameter tuning and several techniques were used to enhance its generalization capability. This section discusses the methods applied and the reasoning for each approach.

#### 3.5.1 Hyperparameter Tuning

In this process, a series of combinations were methodically checked to determine the best hyperparameters for the VGG16 model. The tuned hyperparameters were:

1. **Learning Rate**

A learning rate grid was tested to find a good balance between convergence speed and stability: 0.00001, 0.001, 0.01. While overshooting the optimal loss could be avoided with lower learning rates, higher rates were watched for quick convergence.

2. **Batch Size**

Batch sizes (4, 8, 16, 32) were tested to balance computational efficiency and gradient stability. Smaller batch sizes provided more frequent updates but could lead to noisier gradients, while larger batches offered smoother convergence. Since the model from the first epoch showed overfitting signs, the batch size was gradually reduced to a batch size of 4. This lower batch size helped slightly reduce the rate of overfitting.

3. **Dropout Rate**

Dropout rates (0.3, 0.5, 0.7) were introduced to regularize the network by randomly turning off neurons during training. This helped prevent overfitting by forcing the model to learn robust features rather than memorizing the training data. With a lower dropout rate, the model is less

likely to overfit. Since our model was overfitting, a lower dropout rate of 0.3 was finally selected to reduce the rate at which the model overfitted.

4. **Dense Units**

Fully connected layers with 512 and 256 units were evaluated to test the network's ability to extract higher-level features before classification. When the model was trained on 256 units, it did not perform well; it was likely underfitting, so 512 units were selected.

5. **Optimizers**

Different optimizers (Adam, SGD, RMSprop) were tested to evaluate their performance in minimizing the loss function. Adam's adaptability to sparse gradients was complemented by SGD's ability to generalize and RMSprop's ability to handle non-stationary objectives. Among these three optimizers, the Adam optimizer gave better results than RMSprop and SGD, thus being selected as our final optimizer.

### 3.5.2 **Techniques Implemented to Reduce Overfitting**

1. **Data Generators**:

Custom data generators were used to load training and testing data in batches, ensuring memory efficiency and scalability for large datasets. These generators also applied augmentations like random flips, rotations, and brightness adjustments, enriching the training data to improve model robustness. The data argumentation was done during the preprocessing using the data generator functions.

2. **Layer Freezing**:

The convolutional layers of the pre-trained VGG16 model were frozen to retain their learned features, reducing the risk of overfitting on a small dataset. Only the newly added dense layers were trainable, which limited parameter updates to task-specific layers.

3. **Dropout Regularization**:

Dropout layers were introduced after the dense layers to reduce the co-adaptation of neurons and promote generalization. With a lower dropout rate, the model is less likely to overfit. Since our model was overfitting, a lower dropout rate of 0.3 was finally selected to reduce the rate at which the model overfitted.

4. **Mixed Precision Training**:

Mixed precision (float32) was used to optimize memory usage and improve computational efficiency, allowing faster experimentation without compromising model performance.

5. **Learning Rate Scheduling**:

A ReduceLROnPlateau callback was implemented to lower the learning rate when the validation loss plateaued dynamically. This ensured finer adjustments during later stages of training, preventing oscillations or divergence. The model was tested on all three learning rates; a higher learning rate caused the model to underfit, and when the learning rate was 0.001, the model gave a training accuracy of 20 – 26% from the first to the $10^{th}$ epoch. We then decided to lower the learning rate to 0.00001, which at least allowed the model to have a training accuracy of 80% on the $10^{th}$ epoch and a validation accuracy of 32%. Despite overfitting, the lower learning rate gave better results than the higher learning rate.

6. **Early Stopping**:

The training was terminated early if the validation loss did not improve for five consecutive epochs, preventing overfitting by stopping training once the model's performance on unseen data stopped improving.

7. **Model Checkpointing**

The best-performing model for each hyperparameter configuration was saved based on validation loss, ensuring that subsequent analysis used the most effective model.

**8. Class balancing**

Only the disgust category was underrepresented. It only had 493 instances, which would have introduced some biases in the model. Class balance was achieved through sampling, where the minority class of Disgust was upsampled to 3500 for equal class distribution.

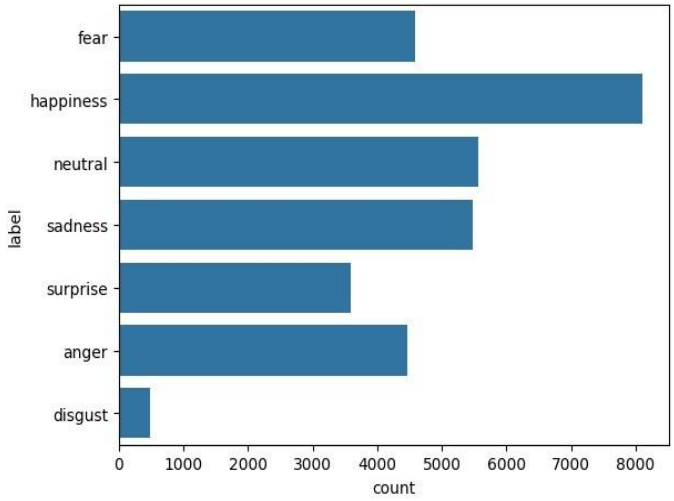



*Figure 2: Class distribution before sampling*

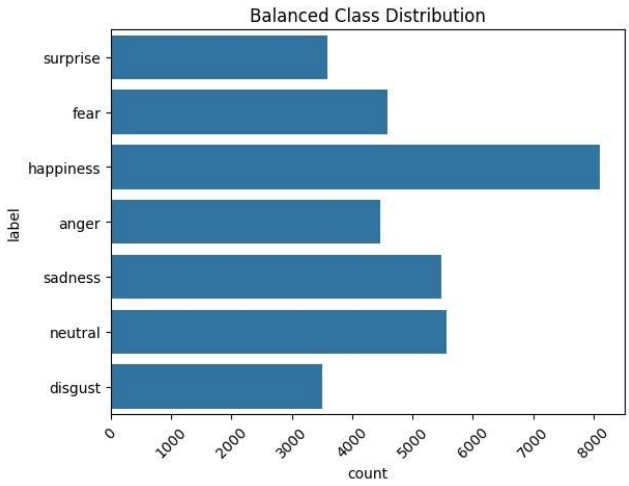



*Figure 3: Class distribution after sampling*

### 3.6 Model Evaluation

#### 3.6.1 Quantitative Metrics

Model performance was assessed using the validation accuracy metric, which is the proportion of correctly classified images on unseen data. This metric effectively illustrated the model's capability for generalization. The model was further evaluated with precision, recall, and f1-score metrics. Training curves for training accuracy/loss and validation accuracy/loss were also used to assess the model overfitting trend.

### 3.8 Summary

Resizing images, normalization, and augmentation were done to prepare the dataset for training by making the input data consistent and diverse. The architecture hyperparameters-learning rate, dropout rate, and batch size were fine-tuned systematically to enhance performance by avoiding overfitting. This included dropout regularization, freezing of layers, dynamic learning rate scheduling, and early stopping. Overfitting remained a challenge even with all these techniques, reflected in the gap between training and validation performances. Limitation: Therefore, This will remain a topic of future research involving more advanced architectures, different regularization methods, and varied datasets.
In summary, though the methodology improved aspects of the model's performance, the challenges faced underscore the complexity of achieving robust and generalizable results in facial emotion recognition, mainly when using pre-trained deep learning models. These issues need to be addressed in further research to develop more reliable systems in real-world applications.

## 4. RESULTS AND CRITICAL ANALYSIS

Accordingly, extensive hyperparameter tuning and regularization techniques were used to avoid overfitting and enhance the generalization of the VGG16 model. This section critically reviews model performance and the effectiveness of applied methods.

### 4.1 model performance before tuning

During the first training set, the VGG16 model showed severe overfitting, steadily improving training accuracy with near-flat validation accuracy. It showed that it learned features from the training data but failed to generalize on unseen data. The high validation loss indicates that better regularization and finer tuning are required.

**Key Results Best Configuration (Learning Rate: 0.001, Batch Size: 4, Dropout: 0.3, Dense Units: 512, Optimizer: Adam)**

**Table 1: Showing initial model results before hyperparameter tuning**

| Epoch | Training Accuracy | Validation Accuracy | Training Loss | Validation Loss |
|---|---|---|---|---|
| Epoch 1 | 25.94% | 24.0% | 3.89 | 1.82 |
| Epoch 3 | 25.69% | 24.58% | 1.97 | 1.82 |
| Epoch 5 | 25.68% | 24.50% | 1.84 | 1.81 |
| Epoch 8 | 25.93% | 24.53% | 1.79 | 1.81 |

*Table 1: Showing model performance before hyperparameter tuning*

### 4.2 After Hyperparameter Tuning

A large-scale hyperparameter grid search was performed. The tuned parameters include learning rate, batch size, dropout rate, optimizer, and dense layer units.

**Learning Rate:** The learning rate was kept very low, 1e-5, for stable and smooth convergence. Batch Size: Smaller batch sizes, 4, allowing for more frequent weight updates and hence better efficiency in training on the case of a limited dataset

.

**Dropout Rate:** A moderate dropout rate of 0.3 was applied to regularize the fully connected layers.

**Optimizers:** Three optimizers, Adam, SGD, and RMSprop, were tested. The Adam optimizer demonstrated better stability and faster convergence than SGD and RMSprop.

**Figure 4: Showing model performance after hyperparameter tuning**

**Learning Rate: 1e-05, Batch Size: 4, Dropout: 0.3, Dense Units: 512, Optimizer: Adam**

| Epoch | Training Accuracy | Validation Accuracy | Training Loss | Validation Loss |
|---|---|---|---|---|
| Epoch 1 | 27.18% | 25.89% | 3.62 | 1.81 |
| Epoch 2 | 34.11% | 28.37% | 1.66 | 1.76 |
| Epoch 3 | 47.20% | 30.55% | 1.37 | 1.75 |
| Epoch 4 | 58.35% | 31.19% | 1.11 | 1.76 |
| Epoch 5 | 67.96% | 31.72% | 0.89 | 1.78 |
| Epoch 6 | 74.59% | 31.80% | 0.71 | 1.84 |
| Epoch 7 | 81.10% | 31.83% | 0.57 | 1.85 |
| Epoch 8 | 85.19% | 32.08% | 0.47 | 1.90 |

*Table 2: show model performance in accurately predicting*

**Confusion matrix showing model performance**

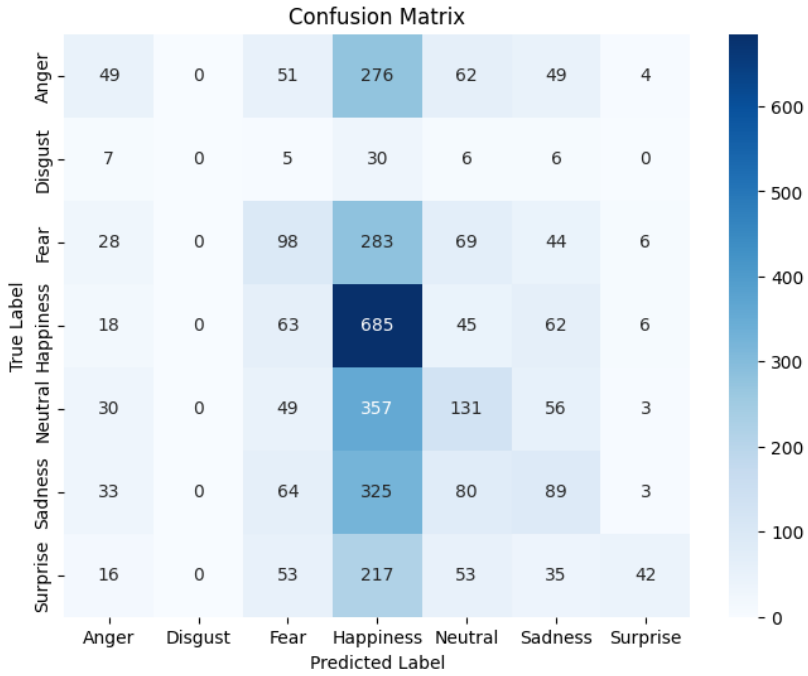



*Figure 1: Confusion matrix*

The above matrix visually represents how well the model classifies each emotion. Darker squares along the diagonal indicate correct classifications, while lighter off-diagonal values show misclassifications. The model performs well in detecting "Happiness" but struggles with other emotions like "Disgust" and "Surprise."

**Most and Least Accurately Classified Emotions**

**Most Accurately Classified Emotion**

- **Happiness**

The model correctly classified **685** instances of "Happiness," the highest number among all emotions. This is reflected in the **high recall (0.78)**, meaning most actual instances of "Happiness" were correctly identified. Precision is moderate at **0.32**, indicating some false positives, but overall, the model performs best for this class.

**Least Accurately Classified Emotion**

- **Disgust**

The model struggles significantly with "Disgust," correctly classifying **only seven instances** while misclassifying the rest. This is reflected in the **precision, recall, and F1-score of 0.00**, meaning the model barely identifies "Disgust" correctly and often confuses it with other emotions. Given the small number of correctly classified samples, this class likely suffers from **data imbalance** or poor feature representation.

**Other Observations**

**Fear, Neutral, and Sadness** have moderate accuracy but are often misclassified as "Happiness" or other emotions. **Surprise** has a low recall (0.10), meaning most true instances of "Surprise" are misclassified. **Anger** also struggles with low precision (0.27) and recall (0.10), meaning it is frequently misclassified.

The model is **most potent at identifying "Happiness"** but **struggles with other emotions, particularly "Disgust"**. Addressing **class imbalance, refining feature extraction, or using more advanced models** could improve classification for poorly predicted emotions.

**Classification report**

**The overall model accuracy was 30%.**

| expression | precision | recall | F1-score |
|---|---|---|---|
| Anger | 0.27 | 0.10 | 0.15 |
| Disgust | 0.00 | 0.00 | 0.00 |
| Fear | 0.26 | 0.19 | 0.22 |
| Happiness | 0.32 | 0.78 | 0.45 |
| Neutral | 0.29 | 0.21 | 0.24 |
| Sadness | 0.26 | 0.15 | 0.19 |
| Surprise | 0.66 | 0.10 | 0.17 |

*Table 3: Showing model performance in the classification task*

**Precision** measures the accuracy of optimistic predictions (e.g. when the model predicts "Anger," how often is it correct?). **Recall** measures how well the model captures all instances of a given emotion—the **F1-score** balances precision and recall, showing overall classification

performance. The model has the highest F1-score for "Happiness" (0.45), while "Disgust" has the lowest (0.00), meaning it was not detected well.

### Performance with SGD and RMSprop Optimizers

**SGD** optimizer: The model achieved a peak validation accuracy of **27.51%**, with a slower learning curve and relatively stable loss reduction. The **RMSprop**: While training accuracy tried to reach **28.09%**, the validation accuracy stagnated at **24.94%**, and the model displayed unstable loss values.

## 4.3 Critical Analysis of Performance

1. **Training Behavior**:

The VGG16 model demonstrated strong learning capability during training, with training accuracy exceeding **80–90%** in later epochs. However, validation accuracy remained low (peaking at ~32%), which indicates significant **overfitting** despite applying dropout regularization and learning rate schedule.

### Training vs. Validation Accuracy and Loss

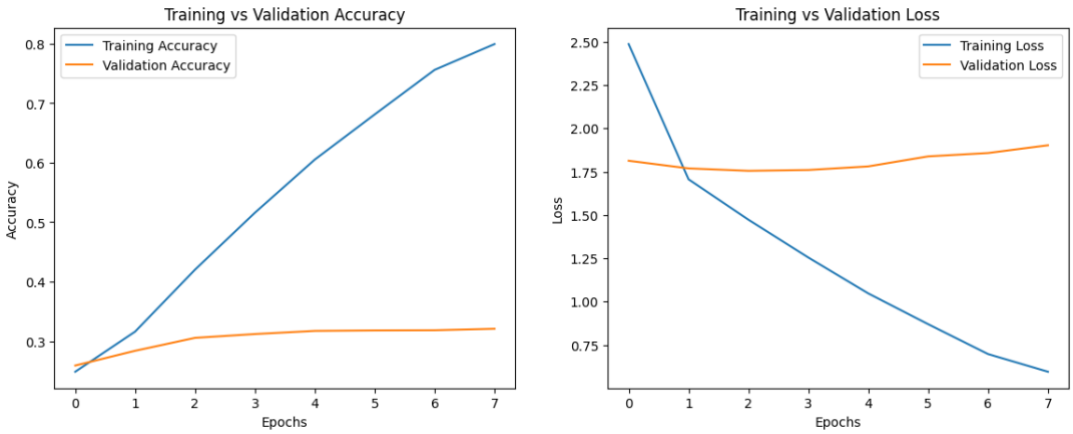



*Figure 2: Training and validation loss curves*

The accuracy plot shows how the model's performance improves on the training set over epochs while validation accuracy stagnates, suggesting potential overfitting.

The loss plot indicates that while training loss decreases significantly, validation loss remains almost constant or slightly increases, indicating overfitting.

2. **Impact of Hyperparameter Tuning**:

Using a low learning rate (1e-5) ensured stable convergence, but the model struggled to escape local minima, limiting improvements in validation accuracy. While dropout slightly reduced the risk of overfitting, it was insufficient to generalize well on unseen data. The Adam optimizer outperformed SGD and RMSprop regarding training speed and stability, but no optimizer significantly improved validation accuracy.

3. **Generalization Gap**

The persistent gap between training and validation accuracy reflects the challenges posed by the dataset. The VGG16 model appears to have memorized training patterns but failed to generalize due to the following factors:

**Dataset Size** Although the dataset had 35k samples, it may be too small to train a deep architecture like VGG16 effectively.

**Dataset Quality**: Emotion recognition is inherently challenging due to subtle and ambiguous facial expressions.
**Class Imbalance**: If certain emotion classes dominate, the model may need to learn underrepresented courses effectively. The disgust class was underrepresented, so to avoid biases in our model, the class needed to be balanced.

4. **Regularization Techniques**

**Dropout** and **ReduceLROnPlateau** helped stabilize the model, as evidenced by the gradual decrease in training loss. Data augmentation further reduced the level of overfitting. 5. **Comparison to Literature**

Studies using VGG16 on benchmark datasets like **AffectNet** often achieve 60–75% validation accuracy. The underperformance observed in this study suggests that the dataset used may lack the diversity or size needed to achieve similar results.

### 4.4 Implications of Results

**1. Model Suitability**

While VGG16 is strong in architectures for image classification tasks, such depth and complexity demand more extensive and diverse datasets. This constraint of the current dataset inhibits generalization.

**2. Dataset Limitations**
These results indicate that the quality and balancing factor affects emotion detection. Increasing the dataset and using advanced techniques for data augmentation could improve generalization.

**3. Hyperparameter tuning**

The tuning process was enlightening, showing how different learning rates, batch sizes, dropouts, and optimizers influence the performance. Further experiments with smaller architectures or more advanced models could be done using EfficientNet, which might result in better performance.

**Conclusion**

After extensive hyperparameter tuning, VGG16 models became more notorious in training performance but still overfitted; the best validation accuracy reached 32.0%. Also, the enhancement in quality and augmentation of the dataset is required for more vigorous regularization techniques to reduce overfitting and explorations of alternative architectures better suited for emotion detection. While VGG16 could learn features from the dataset, limited generalization clarifies the main issues to be pursued in this work.

## 4. DISCUSSION

The results presented here critically analyze the limitation inherent in using a VGG16 pre-trained deep learning model in identifying facial emotion recognition. Even with large-scale hyperparameter tuning and several regularization methods, it faces significant overfitting, as evidenced by the great difference in validation and training performance. While the accuracy during the training of later epochs was well beyond 80%, the accuracy over validation remained at 32.0%, thereby showing a model's inability to generalize to the unseen data effectively.

### 5.1 Causes of Overfitting in VGG16

While VGG16 works well for feature extraction on big and diverse datasets, it has some specific architectural limitations when working with smaller or imbalanced datasets, such as FERC. Some significant reasons for overfitting include:

**Model Complexity**

VGG16 has 138 million parameters and is incredibly computationally expensive; it easily overfits when trained on datasets that are not large or diverse enough. A likely scenario is that the model memorized the features rather than learned to generalize them. The VGG16 architecture, though adequate for large and diverse datasets like ImageNet, proved overly complex for the current dataset. Its depth and high number of parameters led to the following:

**Overfitting:** The model memorized the training data by its high training accuracy while failing to generalize to new samples. This overfitting VGG16 might not be appropriate for relatively small or imbalanced datasets.

The learning was inefficient, and even a very low learning rate, 1e-5, generally used for stability, struggled to get it out of the local minima. Further, VGG16 is a computationally intensive model; hence, training problems with limited data are also contributing here.

**Dataset Size and Class Balance**

Even though there were 35,887 images in the dataset, the class "Disgust" was poorly represented. These class imbalances might make the model biased toward classes like Happiness or Neutral, diminishing its ability to identify the minority classes correctly.

**Lack of Real-World Variability**

The dataset did not have enough variations in terms of facial expressions, lighting conditions, occlusions, and demographic diversity to further restrict the model from generalizing on unseen real-world conditions.

**Transfer Learning Limitations**

VGG16 was pre-trained on the ImageNet dataset, mainly comprised of generic object categories. The features learned may not fully align with the subtle nuances required for facial emotion recognition, leading to suboptimal performance.

**Limitations of Hyperparameter Tuning and Regularization**

Extensive hyperparameter tuning provided some insight into the model's behaviour but did not significantly improve the validation accuracy.

Some of the key strategies included:

Dropout Regularization: This was applied to offer more generalization ability and avoid overfitting; the rates ranged from 0.3–0.7, though the dataset constrained the impact.

Optimizers: Adam outperformed the other two, SGD and RMSprop, regarding convergence speed and training stability; none of the optimizers tested significantly improved the performance on the validation set.

Learning Rate Scheduling: The ReduceLROnPlateau callback helped smoothen the training by dynamically adjusting the learning rate in the event of a plateau in loss, but the difference in generalization remained.

**Nature of the Task**

Facial emotion recognition is fundamentally a complex problem due to emotional expression's inherent subtlety and dynamic characteristics. Relatively straightforward emotions, such as those corresponding to states of happiness and sadness, are very recognizable, whereas more subtle states like Disgust or fear require better refinement in feature extraction. In processing such finer features, the VGG16 configuration would call for adding attention mechanisms or using a more advanced structure of CNNs, such as Vision Transformers or Capsule Networks.

## 5.2 Addressing the Limitations

The challenges in this study have been highlighted with the view of taking into consideration certain limitations that future research should try to address:

**Dataset Improvement**
Collect or augment an extra diverse and balanced dataset, including all emotion classes. Use the most advanced techniques, such as GAN, to generate synthetic samples for the underrepresented classes.

**Model Simplification**
Smaller and lighter architectures, such as MobileNet or EfficientNet, should be generalized better to smaller datasets.
Leverage methods such as transfer learning by fine-tuning the convolutional layers to learn from the task-specific features.
**Advanced Regularization Techniques**
The robustness of the model could further be improved by employing data augmentation techniques in an even more aggressive manner, simple manipulations of brightness, and simulation of occlusions up to rotation.
Employ attention mechanisms or hybrid models like CNN-RNN that can more powerfully model minute variations of facial features.

**Alternative Evaluation Metrics**
Besides validation accuracy, metrics such as precision, recall, and F1-score also indicate the performance of each emotion class.

**Addressing Class Imbalance**
The observed results were strongly related to the class imbalance, especially for the underrepresented class of Disgust. The following strategies could be applied in order to handle this:
Data Resampling: Upsampling the minority class, as in this study, ensures that all classes have equal importance during the training. Performing other ways of generating further synthetic data using GANs could be better for class balancing.

**Exploring Alternative Model Architectures**

While this work takes VGG16 as the backbone, there is better architecture for generalized performance with fewer parameters; hence, it would be doing a better job for facial emotion recognition:

ResNet Residual Networks introduce skip connections to combat vanishing gradients in deeper architectures, which enables stable training and improves generalization on smaller datasets.

EfficientNet represents the right balance between accuracy and efficiency; it scales the depth and width of the model while keeping a lower number of parameters, thus being less subject to overfitting.

Hybrid Models Combining CNNs with an attention mechanism or RNN may give better representations, especially in capturing spatial and temporal relationships in facial expression studies.

## 1.1 5.2 Recommendations

Consider exploring alternative architectures: Modern deep learning architectures like EfficientNet, ResNet, and ViT provide improved generalization with fewer parameters. These models leverage advanced design and feature extraction capabilities to perform better, especially on smaller or imbalanced datasets.

Hybrid Models: One can integrate CNN with RNN or attention mechanisms to model facial expressions' spatial and temporal characteristics. This type of approach could find subtle changes in facial features over time.

Fine-Tuning Strategy: Instead of freezing all base layers, fine-tune part of deeper layers from the pre-trained VGG16 model to adapt better to the task-specific features yet preserve pre-learned general features.

Robustness Testing: The task entails conducting systematic robustness testing of the models trained using the enhanced dataset creation. The testing has to be performed regarding occlusions like masks and glass and variable lighting conditions, and demographic variations that prove the real-world reliability of the model must be taken into account.
Cross-Validation: This is done via k-fold cross-validation to estimate model performance better, reduce the variance due to a particular data split, and make the results more reliable.

### 5.3 Conclusion

Conclusion

This work critically assessed the performance of the VGG16 model on facial emotion recognition and aimed to highlight its shortcomings and defects for further improvements. Extensive hyperparameter tuning and regularisation techniques against overfitting were performed by applying dropout, data augmentation, and learning rate scheduling. While it achieved more than 80% in training accuracy, it resulted in merely a 32.0% validation accuracy with this model, which shows the big generalization gap and poor performance on unseen data.

The findings underscore three significant challenges:
**Limitations of Dataset:** Class imbalance, limited diversity, and subtlety of emotions impeded the generalisation model's generalisation capability.

**Model Complexity:** While the VGG16 architecture was to deal with larger and more complex datasets, this became a problem while overfitting the generally small and imbalanced dataset presented in this work.

**Performance Metrics:** Relying exclusively on validation accuracy obscured performance differences across individual emotion classes.

Future research should emphasise better quality and diversity of the datasets, advanced architectures like ResNet, EfficientNet, or Vision Transformers, and robust evaluation strategies. This can be further improved by the integration of hybrid models and refinement of fine-tuning techniques that enhance the generalization capability of the model, hence improving the performance of the facial emotion recognition system in real-world applications.

## References


Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., (2016, October). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308–318).

Bhoomika, G., Pujitha, V.D., Sindusha, M., Rekha, C.S. and Suvarna, B., 2024, July. Facial Emotion Recognition: A Comparative Study of Pre-trained Deep Learning Models. In *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)* (pp. 377-382). IEEE.

Buolamwini, J. and Gebru, T., (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale—*arXiv preprint arXiv:2010.11929*.

Gong, L., Chen, W. & Zhang, D. (2024). An attention-based multi-domain bi-hemisphere discrepancy feature fusion model for EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Kollias, D. and Zafeiriou, S., 2021. Affect analysis in the wild: Valence-arousal, expressions, action units, and a unified framework—*arXiv preprint arXiv:2103.15792*.

Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster and T. Vetter, "Training deep face recognition systems with synthetic data" in arXiv:1802.05891, 2018.

McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A., (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273–1282). PMLR.

Mollahosseini, A., Hasani, B. and Mahoor, M.H., (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), pp.18–31.

Parkhi, O., Vedaldi, A. and Zisserman, A., 2015. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

Sabour, S., Frosst, N. and Hinton, G.E., 2017. Dynamic routing between capsules. *Advances in neural information processing systems*, *30*.

Simonyan, K. and Zisserman, A., (2014). Intense convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, M. & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on machine learning* (pp. 6105–6114). PMLR.

Ye, Y., Pan, T., Meng, Q., Li, J. and Lu, L., 2022, July. Online ECG Emotion Recognition for Unknown Subjects via Hypergraph-Based Transfer Learning. In *IJCAI* (pp. 36663672).

Zhang, Y. & Deng, W. (2020). Class-balanced training for deep face recognition. In *Proceedings of the, i.e.,/cvf conference on computer vision and pattern recognition workshops* (pp. 824–825).

Zheng, Y., Zhao, Y., Ren, M., Yan, H., Lu, X., Liu, J. & Li, J. (2020, October). Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2264–2272).