REAL – TIME HUMAN EMOTION RECOGNITION FROM FACIAL EXPRESSIONS USING DEEP LEARNING

By

Harshitha Polsani

A Project
Presented to the Faculty of
The Computer Science Department

In partial fulfillment
Of the requirements for the degree
Master of Science

Huntsville, Texas
November 20, 2023

Table of Contents

1 Introduction

Facial emotion recognition holds immense significance in contemporary human communication, finding applications in diverse domains like human-computer interaction, virtual reality, and mental health monitoring. This study aims to develop an efficient deep learning model for accurately discerning individuals' emotional states from their facial expressions, comprising two critical stages: face detection and emotion classification.

In the initial phase of face detection, we employ the RetinaFaceResNet50 model to swiftly identify and extract faces from real-time video inputs. Notably, our model is trained on a combination of three datasets: FER2013 (48 x 48 grayscale images), RAF (100 x 100 RGB images), and Affectnet (high-resolution RGB images with eight labeled emotions), ensuring a broad spectrum of labeled facial expressions for robust model training.

Subsequently, our deep learning model, trained on this diverse dataset, categorizes emotions into happiness, sadness, anger, surprise, disgust, neutrality, fear, and contempt (for Affectnet). Additionally, we introduce a customized MobileNet-based model equipped with patch extraction and attention mechanisms, achieving an impressive 80% accuracy on the RAF dataset. Alongside custom model development, we conduct a comparative analysis of various deep learning architectures, including CNN, VGG16, Vision Transformers, ResNet, DenseNet, and EfficientNet, assessing their performance with and without data augmentation techniques to gauge their efficacy in emotion classification.

This study represents a significant advancement in facial emotion recognition, especially in real-time applications. By seamlessly integrating face detection and emotion classification, it offers valuable insights with broad applicability across domains and human-computer interaction.

2 Related Work

In recent years, there has been significant research focused on facial emotion recognition using computer vision, machine learning, and deep learning techniques. Several studies have explored different approaches to identify and classify emotions from facial expressions.

In the field of Human Emotion Recognition from Facial Expressions, Chang et al. conducted a study, where they employed a Convolutional Neural Network (CNN) model based on ResNet to extract features from the Fer2013 and Cohn-Kanade Dataset (CK+) datasets [4]. They proposed a Complexity Perception Classification (CPC) algorithm and evaluated its performance using different classifiers such as Linear Support Vector Machine, and Random Forest. The combination of CNN + Softmax with CPC achieved impressive recognition accuracies of 71.35% for Fer2013 and 98.78% for CK+ [4]. This study highlights the efficacy of deep learning techniques in accurately identifying and classifying emotions from facial expressions.

H. Kishan et al. and Vishal et al. conducted a study on "Emotion Detection using Deep Facial Features" [5]. The study involved three main phases: pre-processing, feature extraction, and classification techniques. They explored the utilization of Transfer Learning from famous pre-trained models like VGG-16, ResNet152V2, InceptionV3, and Xception. The performance of these models was evaluated on a combined dataset comprising the Cohn-Kanade Dataset (CK+) and Japanese female facial emotion (JAFFE). The reported accuracies for the respective architectures were 83.16%, 82.15%, 77.1%, and 78.11% [5]. The study aimed to assess the effectiveness of these deep learning models for emotion detection by leveraging deep facial features and transfer learning.

In the paper titled "Emotion recognition from facial expression using deep convolutional neural network" by D.Y. Liliana [5], the author explores the task of facial expression recognition using deep Convolutional Neural Networks (CNNs). The study focuses on detecting facial Action Units (AUs) as part of the Facial Action Coding System (FACS) to represent human emotions. The CNN architecture incorporates fully-connected layers with a dropout regularization method to reduce overfitting. The research utilizes the Extended Cohn Kanade (CK+) dataset for facial expression recognition experiments. The proposed method achieves an average accuracy rate of 92.81% and successfully classifies eight basic emotion classes. This study offers two contributions: automatic feature extraction using deep learning CNNs to detect Action Units and the use of the CK+ dataset, which provides comprehensive data and ground truth for facial expressions and Action Units [6].

Nguyen et al. presented a novel approach for facial emotion recognition using an ensemble of multi-level convolutional neural networks (CNNs) [7]. Their proposed model, similar to VGG, incorporated both high-level and mid-level features to improve the accuracy of emotion recognition. The baseline CNN model achieved an accuracy of 69.21% on the Fer2013 dataset, while the proposed multi-level CNN model achieved an improved accuracy of 73.03%.

Ozdemir et al. [8] conducted a study on real-time emotion recognition from facial expressions using a Convolutional Neural Network (CNN) architecture based on LeNet. The researchers merged three distinct datasets, including JAFFE, KDEF, and a custom dataset, to create a comprehensive collection of face images representing seven different emotions. To ensure consistency, they employed the Haar Cascade library for accurate face detection and localization. The rectangular facial expressions were subsequently cropped and resized. Additional preprocessing steps involved converting the images to grayscale. By incorporating the Haar Cascade technique for face detection and leveraging the CNN architecture, the proposed method achieved remarkable classification accuracies, with an impressive 96.43% accuracy for emotion recognition and a validation accuracy of 91.81%. This study demonstrates the effectiveness of CNNs in accurately classifying facial emotions. [8]

3 Dataset

In our research, we utilized three datasets, each playing a pivotal role in the study of Human Emotion Recognition from Facial Expressions. FER2013 and RAF datasets are well-established in the literature, while Affectnet is included to provide a broader spectrum of emotional recognition through high-resolution images.

The FER2013 dataset, known for its standard use in facial expression recognition challenges, contains grayscale images at a resolution of 48 x 48 pixels. It presents a multiclass classification task across seven emotions: Angry, Fear, Disgust, Happy, Sad, Neutral, and Surprise. Notably, the FER2013 dataset exhibits class imbalance, with certain emotions being represented by significantly fewer examples than others. This imbalance is a common challenge that can affect the efficiency of training and the accuracy of the resulting model.[1]

The second dataset, Real-world Affective Faces (RAF), introduced a different dimension to our study. This dataset comprises 15,339 images, split into train and test sets, with 12,271 for training and 3068 images allocated for testing. To enhance our model's performance, we further subdivided the training set into train and validation sets. [2]

Affectnet, the third dataset employed in our study, enhances the diversity of facial emotion data with 41,553 images distributed across eight classes. The training set includes 31,919 images, the validation set comprises 5,634 images, and the test set contains 4,000 images. The Affectnet dataset is significant for its addition of an eighth emotion, 'contempt,' and for providing higher resolution images. To standardize input to our model, we explored two preprocessing approaches: resizing the images to 224 x 224 pixels and performing PCA to retain the most important features while reducing the dimensions to 224 x 224. This dual approach allowed us to compare the effectiveness of simple resizing against dimensionality reduction while maintaining essential features. The class names in Affectnet are anger, contempt, disgust, fear, happy, neutral, sad, and surprise, reflecting a comprehensive range of human emotions.[3]

Figure 1 showcases images corresponding to each emotion class from both the RAF and FER2013 datasets.

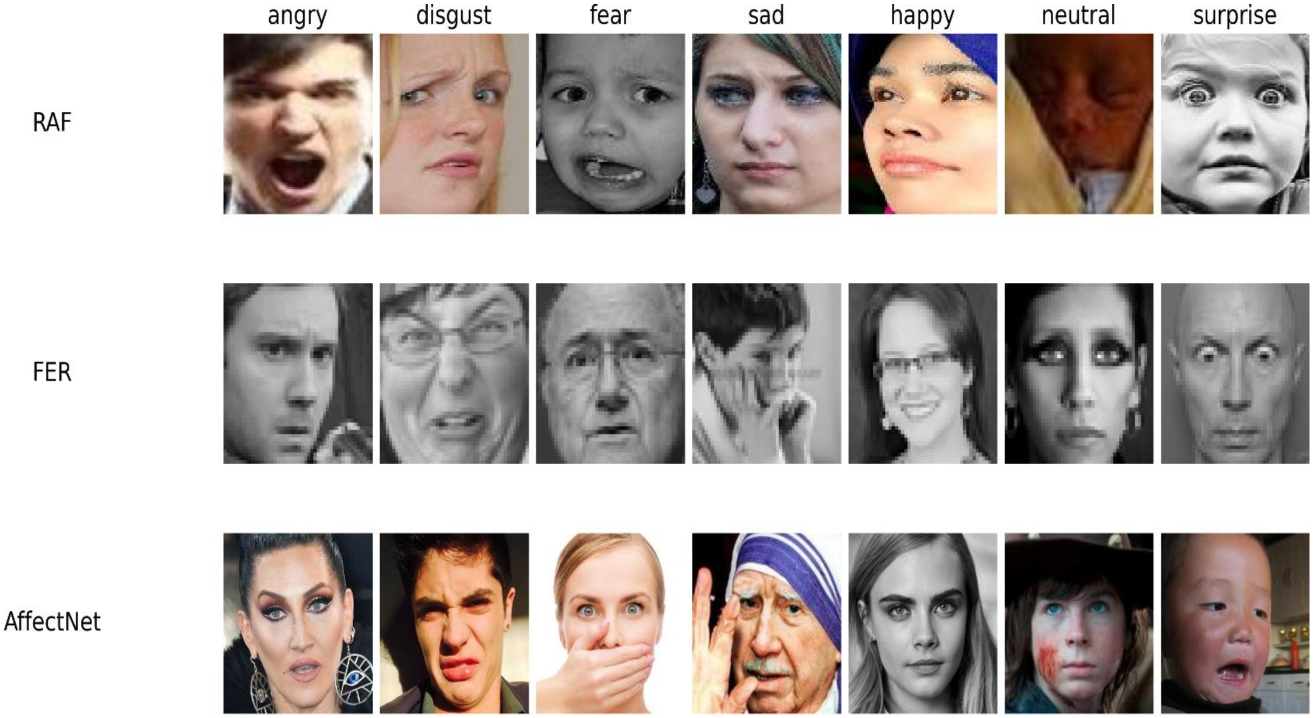


Figure 1 - Sample Images of Emotion Classes from RAF, FER2013 and Affectnet Datasets respectively

The Figure 2 and 3 visually represents this class distribution within the FER2013 dataset and RAF dataset, indicating the number of images available for each emotion class in both the training and test sets. This class imbalance can potentially affect the performance of the model during training and evaluation, as the model may be biased towards the majority classes

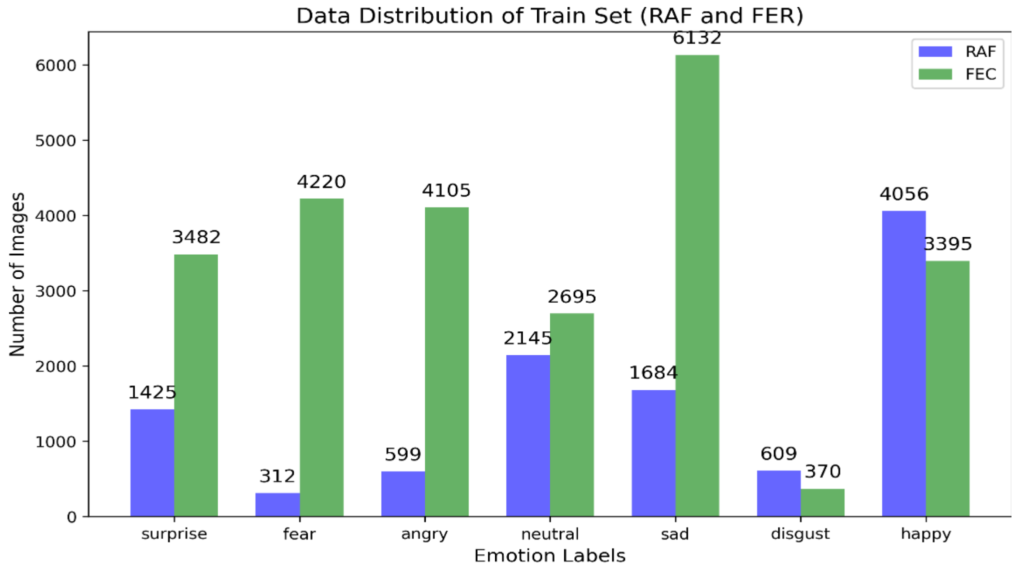


Figure 2 - Data Distribution of Training set on RAF and FER2013 Dataset

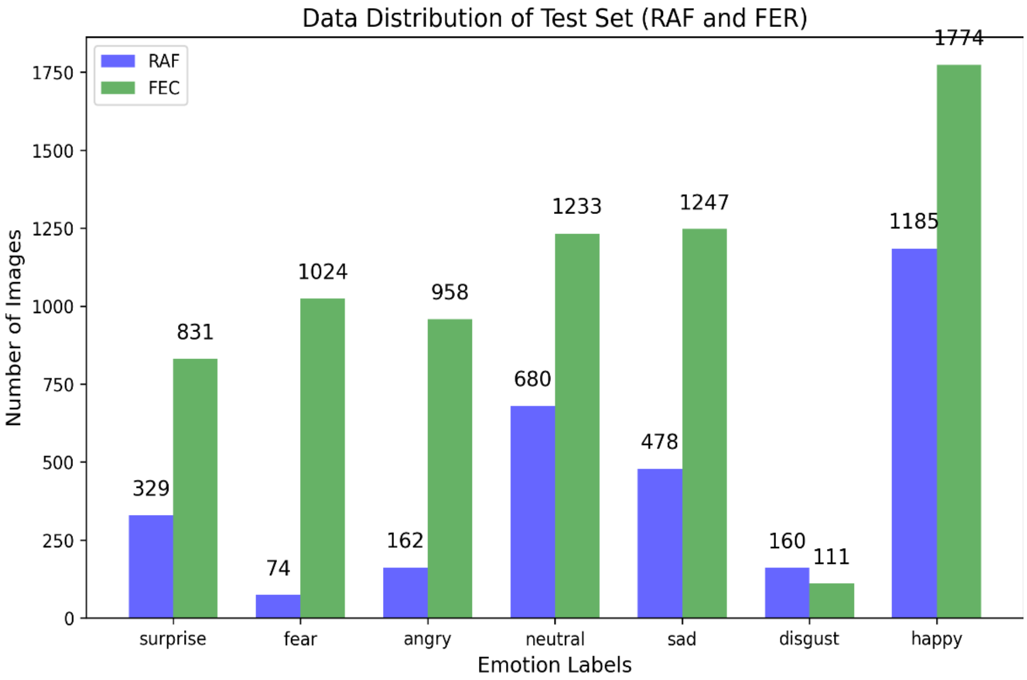


Figure 3 - Data Distribution of Test set on RAF and FER2013 Dataset

Both the FER2013 and RAF datasets exhibit imbalanced class distribution, with specific classes having a relatively higher number of images. This imbalance poses a significant challenge in training robust and accurate emotion recognition models. To address the issue of class imbalance within the RAF and FER2013 datasets, our approach primarily involved data augmentation for the minority classes. The objective of data augmentation is to artificially expand and diversify the training dataset, with a particular focus on creating a balanced class representation.

Data augmentation methods commonly used for facial expression recognition tasks include random rotation, scaling, flipping, or adding noise to the images. These techniques introduce variations to the existing data, enhancing the model's ability to generalize and learn from diverse examples. By applying data augmentation selectively to the minority classes, we can effectively address the class imbalance issue and improve the model's performance on underrepresented emotions.

The Figures 4 and 5 visually present the data distribution before and after augmentation for the FER2013 and RAF datasets. These figures illustrate the number of images available for each emotion class within the training set, emphasizing the impact of data augmentation on achieving a more balanced class distribution.

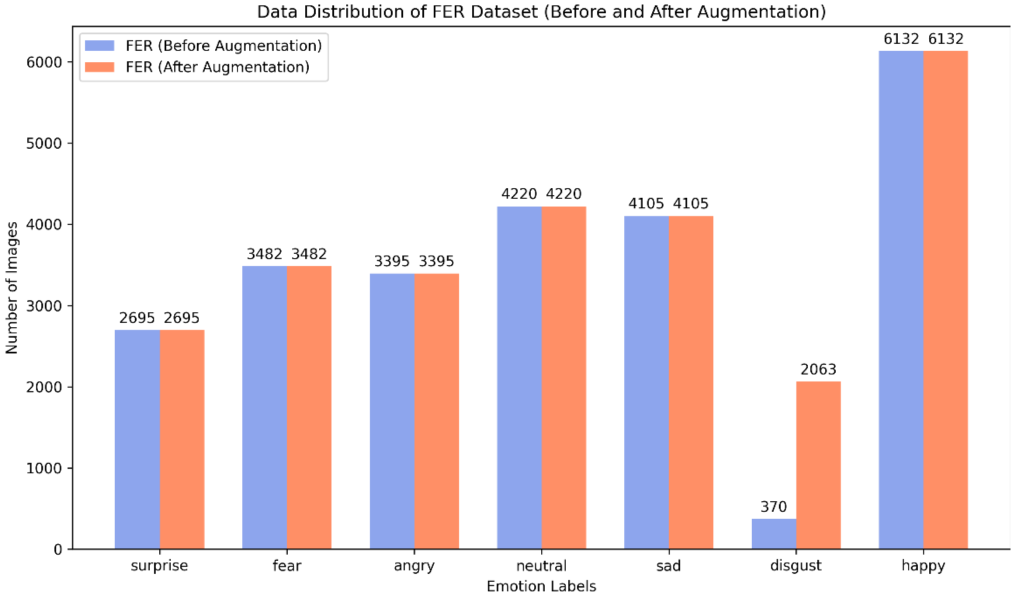


Figure 4 - Data Distribution Before and After Augmentation on FER2013 Dataset

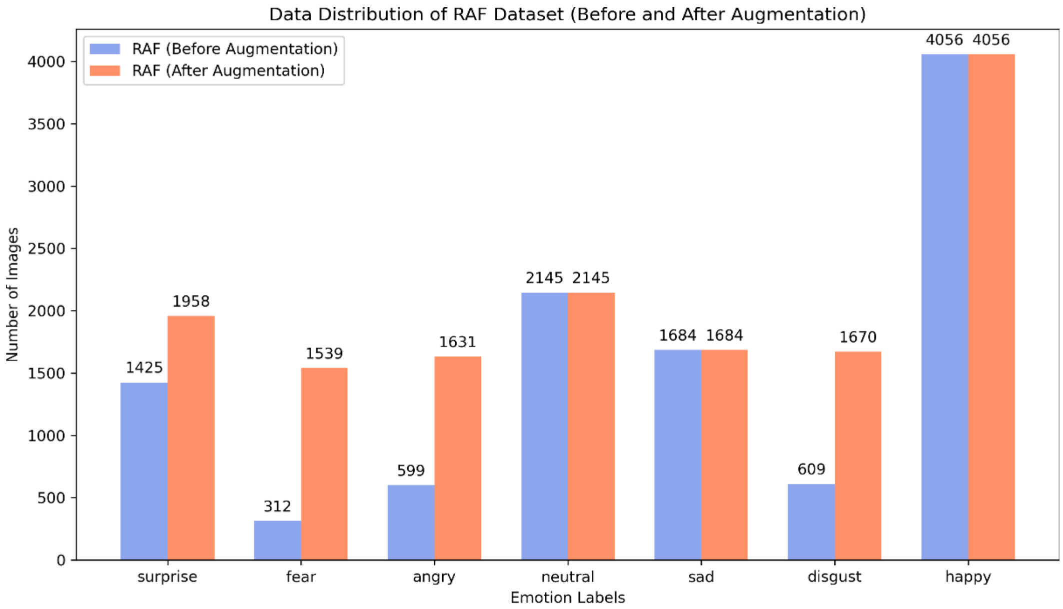


Figure 5 - Data Distribution Before and After Augmentation on RAF Dataset

The Figures 6 visually present the data distribution of the AffectNet dataset. These figures illustrate the number of images available for each emotion class within the train, test, and validation sets, providing a comprehensive view of the dataset's class distribution.

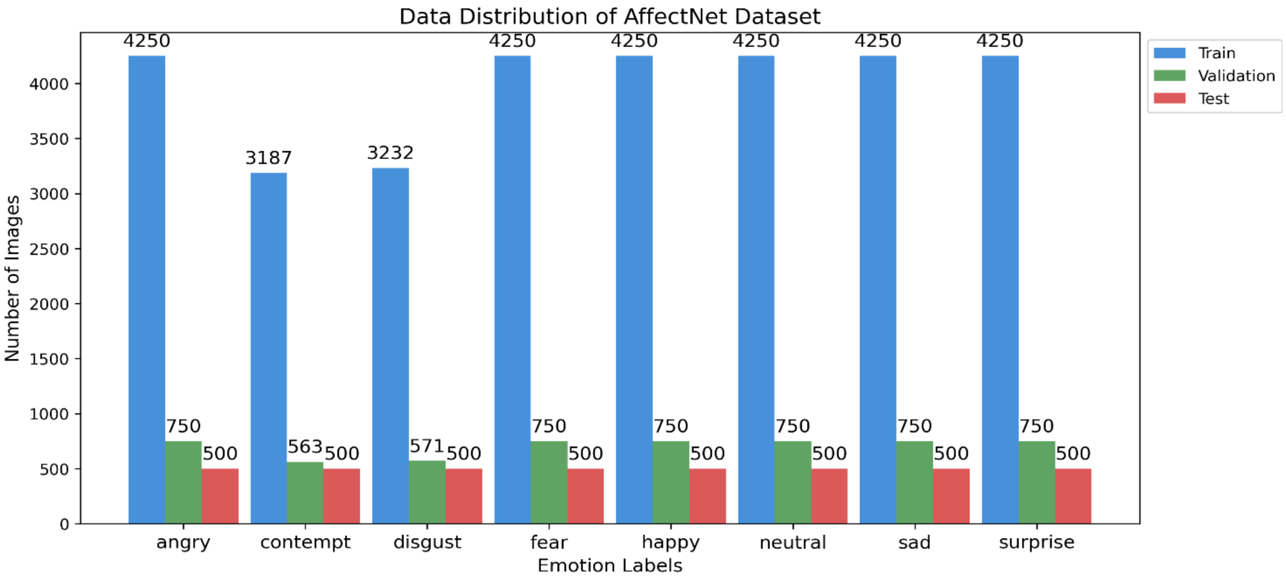


Figure 6 - Data Distribution of Affectnet set

Our study included a comparative analysis of model performance both before and after data augmentation on FER2013 and RAF datasets, evaluating the impact of these techniques on model efficiency and accuracy.

4 Proposed Methodology

In this section, we present a comprehensive methodology for human emotion recognition from facial expressions. Our approach is grounded in the use of three distinct datasets: FER2013, RAF and Affectnet. Notably, our deep learning models were not preceded by face detection during training, given that the datasets inherently comprise pre-extracted facial images. Instead, face detection is executed at the prediction stage, specifically when processing sequences of video frames. To detect faces in the images, we utilize the RetinaFace model pre-trained with the ResNet50 backbone. This model provides accurate face detection along with bounding box coordinates. By applying the RetinaFace model to the FER2013 dataset, we can obtain the location and size of each face present in the images.[13]

This approach allows us to seamlessly integrate face detection into the real-time emotion detection pipeline. The utilization of these datasets enables a thorough exploration of our proposed models' robustness and adaptability across varying image resolutions and class distributions. Our methodology is structured around key steps, encompassing face detection, data augmentation, normalization, deep learning model training, and the extension of our models for real-time emotion detection within video sequences.[4]

4.1 Data Collection:

We acquired three distinct datasets for our study: FER2013, RAF, and Affectnet. The FER2013 dataset consists of grayscale images with a resolution of 48 x 48 pixels, categorized into seven emotion classes: Happy, Sad, Fear, Angry, Neutral, Disgust, and Surprise. The RAF dataset, in contrast, includes RGB images, each of 100 x 100 pixels resolution, featuring the same range of seven emotion classes. The Affectnet dataset further enriches our study with its high-resolution RGB images and an additional eighth emotion class, 'contempt.' A key aspect of these datasets is that they contain pre-extracted facial images, eliminating the need for separate face detection processes in our workflow. This advantage has streamlined our approach, allowing us to concentrate on analyzing diverse image resolutions and class distributions, thereby enabling a robust and comprehensive comparative analysis.

4.2 Principal Component Analysis on Affectnet:

Principal Component Analysis (PCA) was applied to the Affectnet dataset, aimed at optimizing the handling of its high-resolution RGB images. This process involved a detailed analysis of each image channel (Red, Green, Blue), where the images were first reshaped into a two-dimensional array. PCA was then employed to extract and retain the most significant features, ensuring that 95% of the data variance was preserved. The main objective of this approach was to reduce the dimensionality of the images while maintaining their essential visual information. After applying PCA, the images were reconstructed and resized to a uniform dimension of 224 x 224 pixels. This dimension reduction via PCA was an essential step in enhancing the computational efficiency of our models, enabling us to process the rich visual data from the Affectnet dataset more effectively while preserving the critical elements necessary for accurate emotion recognition.

4.3 Data Splitting:

In this section, we focused on the splitting of the FER2013, RAF, and Affectnet datasets. Each dataset was split into two primary subsets: a training set comprising 85% of the data, and a validation set making up the remaining 15%. This division is fundamental to evaluate the performance of our models during training and validation phases. Additionally, each dataset includes a separate test set, which is utilized alongside the training and validation sets.

4.4 Data Augmentation:

In response to the significant class imbalance in the FER2013 and RAF datasets, we implemented data augmentation techniques to enhance the diversity of our training examples. Utilizing KerasImageDataGenerator, we made random adjustments to the images, including rotation, width and height shifts, shearing, zooming, and horizontal flipping. This process was crucial in creating a more balanced dataset by improving the representation of the less frequent emotional states, especially focusing on enhancing the minority classes in both the FER2013 and RAF datasets.

4.5 Normalization:

Normalization was a key preprocessing step applied across all three datasets: FER2013, RAF, and Affectnet. For the FER2013 dataset, consisting of grayscale images, and the RAF dataset, containing RGB images, as well as the high-resolution images of Affectnet, we normalized the pixel values to a range of [0, 1]. This standardization was achieved by dividing each pixel value by 255, the maximum pixel value for these images. Standardizing pixel values across all datasets ensures that our models train more effectively, preventing any single feature from disproportionately influencing the learning process. These preprocessed face images from FER2013, RAF, and Affectnet then serve as input to our deep learning model, which is engineered to learn and classify emotions based on the extracted features.

5 Model Building

5.1 *Multi-level Convolutional Neural Network (CNN)*

Our primary model for emotion recognition is based on a Convolutional Neural Network (CNN) architecture. This model is composed of several layers designed to extract salient features from the input images. Its structure is as follows:

The model begins with a Convolutional Layer of 64 filters, each with a kernel size of (3,3), followed by Batch Normalization and ReLU activation. This layer is designed for initial feature extraction. Subsequent Convolutional Layers with 128, 256, and two sets of 512 filters are included, each followed by Batch Normalization and ReLU activation. These layers progressively capture more complex features. After each Convolutional Layer, Max Pooling with a pool size of (2,2) is applied to reduce the dimensionality, and Dropout at a rate of 0.25 is employed to mitigate overfitting. The model incorporates a Flatten Layer to convert the multi-dimensional feature maps into a flat vector. Following the Flatten Layer are two Dense Layers, each with 512 units and equipped with Batch Normalization, ReLU activation, and a Dropout rate of 0.5. The Output Layer consists of seven units (corresponding to the seven emotion classes) with a softmax activation function for multi-class classification. The model uses the RMSprop optimizer with a learning rate of 0.001 and is trained on categorical cross-entropy loss. Training is conducted over 40 epochs, and the model's effectiveness is evaluated on the test set, focusing on metrics like accuracy. This CNN architecture, with its specific configuration of layers and parameters, forms the foundation of our analysis on the FER2013, RAF and Affectnet datasets. The model achieved an accuracy of 77% before and 78% after augmentation on the RAF dataset, 63% before and 62% after augmentation on the FER dataset, and an accuracy of 50.9% on the Affectnet dataset.

5.2 ResNet18 Model

We employed the ResNet-18 architecture as a key component in our study of emotion recognition from facial expressions. The ResNet-18 model, known for its deep residual blocks, offers the capability to capture intricate features and patterns from input images. The architecture is designed with specific considerations to effectively process the given grayscale images of size 48x48 and RGB images of 224x224. The model begins with an initial convolution layer, followed by a series of residual blocks. These residual blocks are grouped into four stages, each containing two residual blocks. In each stage, the model makes use of convolutional layers and batch normalization, followed by ReLU activation functions. A critical aspect of the ResNet architecture is the inclusion of shortcut connections, enabling the flow of gradient information during training. These connections, along with the residual units, help mitigate the vanishing gradient problem and facilitate the training of deep networks.

After processing through the residual blocks, the model employs global average pooling to summarize the learned features. The final dense layer with softmax activation serves for multi-class classification, allowing the model to predict one of the seven distinct emotion classes.

The ResNet-18 model is compiled with the Adam optimizer and categorical cross-entropy loss. Training occurs over 30 epochs. Upon training completion, the model with the best validation accuracy is selected, and its performance is assessed on the test set, with evaluation metrics.

5.3 DenseNet121

In our pursuit of accurate and efficient emotion recognition from facial expressions, we harnessed the power of the DenseNet-121 architecture. DenseNet models are renowned for their ability to capture intricate patterns and features in images. In our specific case, we tailored the model to process grayscale images of size 48x48pixels of FER2013 dataset and 224x224 pixels of Affectnet dataset. The DenseNet-121 base model, pre-trained on the ImageNet dataset, forms the foundation of our architecture.

To adapt the base model for grayscale images, we implemented a Lambda layer to replicate the single-channel grayscale image into three channels. This adjustment allowed seamless integration with the pre-trained DenseNet architecture. The model then incorporates custom top layers, including a Global Average Pooling layer, followed by a dense layer with 512 units and ReLU activation. Batch normalization is employed for regularization, and the final dense layer uses softmax activation for multi-class classification across the seven distinct emotion classes.

The model is compiled with the Adam optimizer and categorical cross-entropy loss. Training occurs over 50 epochs. and the model's performance is assessed on the test set. Performance evaluation on the test set revealed that the model achieved an accuracy of 61% before augmentation and 60% after augmentation on the FER dataset, and it accomplished an accuracy of 54% on the Affectnet dataset."

5.4 VGG16

To achieve precise emotion recognition from facial expressions, we employed the VGG16 architecture, widely recognized for its performance in image classification tasks. By leveraging the VGG16 model pre-trained on the ImageNet dataset, we set the stage for accurate emotion prediction from grayscale facial images sized at 48x48 pixels and 100 x 100 RGB images.

To adapt the VGG16 model for our specific task, The grayscale input images were transformed into a format compatible with VGG16 by replicating the single-channel grayscale image into three channels. Our custom model extends from this pre-processed base with a sequence of layers designed to capture relevant information from the facial expressions. A Flatten layer is employed to transform the multi-dimensional feature maps into a one-dimensional vector. Two dense layers follow, featuring 128 units each and ReLU activation functions to promote feature learning.

To enhance the model's generalization and prevent overfitting, we incorporated a Dropout layer with a rate of 0.3. The final dense layer employs softmax activation, making it suitable for multi-class classification across the seven distinct emotion classes. The model is compiled with the Adam optimizer and categorical cross-entropy loss. It's important to note that the model might have the potential for further enhancement with extended training across a greater number of epochs. This extended training duration holds promise for refining the model's performance and advancing the precision of emotion recognition from facial expressions. Upon evaluation, the VGG16 model demonstrated an accuracy of 72% before data augmentation and 74% following augmentation on the RAF dataset.

5.5 EfficientNetV2B0

We explored the EfficientNetV2B0 architecture, renowned for its efficiency and effectiveness in various computer vision tasks. With our grayscale facial images, sized at 48x48 pixels, 100 x 100 and 224 x 224 RGB images we used the capabilities of the EfficientNetV2B0 model, fine-tuning its pre-trained layers to adapt to the specifics of our task. Notably, to accommodate grayscale input, we transformed single-channel grayscale images into a format compatible with the model by replicating the channel.

The custom architecture we employed builds upon this pre-trained foundation, adding layers to capture pertinent information from the facial expressions. A Conv2D layer with ReLU activation was introduced to adjust the input channel dimensions. Following this, the EfficientNetV2B0 model was incorporated, providing a strong feature extraction backbone. To further enhance the model's performance and mitigate the risk of overfitting, a Global Average Pooling layer was applied, followed by two Dense layers, each with 128 units and ReLU activation functions. A Dropout layer with a rate of 0.3 was included to improve generalization.

The model's final Dense layer, with softmax activation, makes it well-suited for multi-class classification across the seven distinct emotion classes. We opted for the SGD optimizer and categorical cross-entropy loss for compilation. Training was conducted over 60 epochs, allowing the model to evolve and improve its prediction accuracy. Upon evaluation, the model achieved an accuracy of 54% before augmentation and 51% after on the FER dataset. On the RAF dataset, it achieved 73% accuracy before augmentation and showed improvement to 79% post-augmentation. For the Affectnet dataset, the model maintained an accuracy of 51%.

5.6 Vision Transformers (ViT) for Emotion Classification

In our exploration of advanced architectures for emotion recognition, we implemented two configurations of Vision Transformers (ViT): ViT-B32 and ViT-B16. These models are part of the cutting-edge ViT family, known for their efficacy in processing image data through self-attention mechanisms.

5.6.1 ViT-B32 Model

The ViT-B32 model was employed to analyze high-resolution RGB images of size 224x224 pixels. It leverages a pre-trained base with the sigmoid activation function, adapted to our dataset consisting of 8 classes. The model includes a custom sequential architecture comprising an input layer, the ViT-B32 layer, a Flatten layer, Batch Normalization, Dense layers, and a final classification layer with softmax activation. Compiled with the Rectified Adam optimizer and Categorical Crossentropy loss, this model emphasizes label smoothing for improved generalization. The training process is guided by callbacks like ReduceLROnPlateau, EarlyStopping, and ModelCheckpoint for optimization.

The model underwent training for 30 epochs, and its performance was validated on a separate test set. On the Affectnet-7 dataset, it achieved 54% accuracy.

5.6.2 ViT-B16 Model

Similar to the ViT-B32, the ViT-B16 model is designed to process 224x224 pixel RGB images. It incorporates a pre-trained base tailored for our dataset of 7 classes. The architecture follows a similar pattern to the ViT-B32, including layers for flattening, normalization, dense layers, and softmax activation for classification. The training and optimization techniques are similar to those of the ViT-B32, ensuring consistency in model evaluation.

In summary, our exploration of Vision Transformers, particularly the ViT-B32 and ViT-B16 models, has enriched our understanding of emotion recognition techniques. These models utilize the transformer's self-attention mechanism to extract deep features from facial images. This exploration of ViT models provided a different perspective from traditional CNN-based approaches, potentially enhancing our ability to capture subtle details in human emotions. On the Affectnet-7 dataset, it achieved 60% accuracy.

5.7 Custom MobileNet Model with Patch Extraction and Attention Mechanism

In our study, we developed a custom model based on the MobileNet architecture, integrating patch extraction and attention mechanisms to enhance emotion recognition from facial expressions. This model was applied across all datasets in our study, including FER2013, RAF, and Affectnet, showcasing its versatility and adaptability.

The core of our model is the MobileNet base, originally pre-trained on ImageNet to leverage its robust feature extraction capabilities. On top of this, we added a patch extraction block that uses a series of convolutional layers to create multiple image patches. The patch extraction block consists of convolutions with different kernel sizes to extract various aspects of the facial features. This block captures intricate details by creating multiple patches of the image, each processed through its own convolutional pathway. These patches are then concatenated to form a comprehensive feature map.

Following this, an attention mechanism is employed. This mechanism focuses on the most significant features of the extracted patches through global average pooling and dense layers. It amplifies relevant features while suppressing less useful ones, enabling the model to concentrate on crucial aspects of the facial expressions.

The model is completed with global average pooling and dense layers for classification. The final architecture includes dropout layers to prevent overfitting and an early stopping mechanism to halt training at the optimal point, ensuring the model does not overfit to the training data.

To validate the model's effectiveness, we performed a 5-fold cross-validation on the RAF dataset. This resulted in the model achieving an accuracy of 80% on RAF dataset.

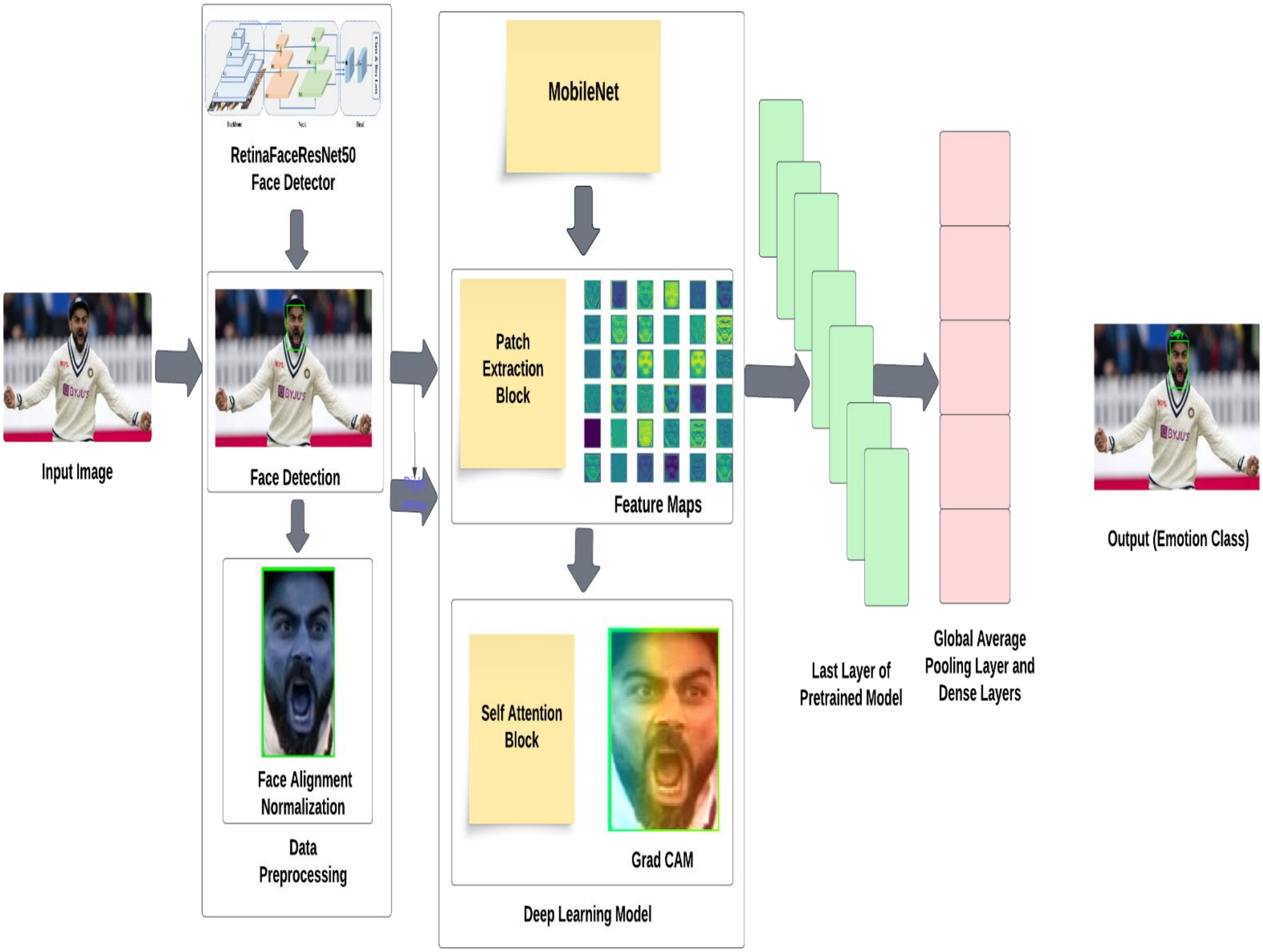


Figure 7 – Overview of Proposed Methodology

Figure 7 provides an overview of the proposed methodology for emotion recognition from facial expressions. It depicts the sequential process from face detection using RetinaFaceResNet50 to emotion classification using a custom MobileNet model enhanced with patch extraction and attention mechanisms. The flowchart illustrates the transformation of input data through each stage, resulting in the identification of emotion from facial features.

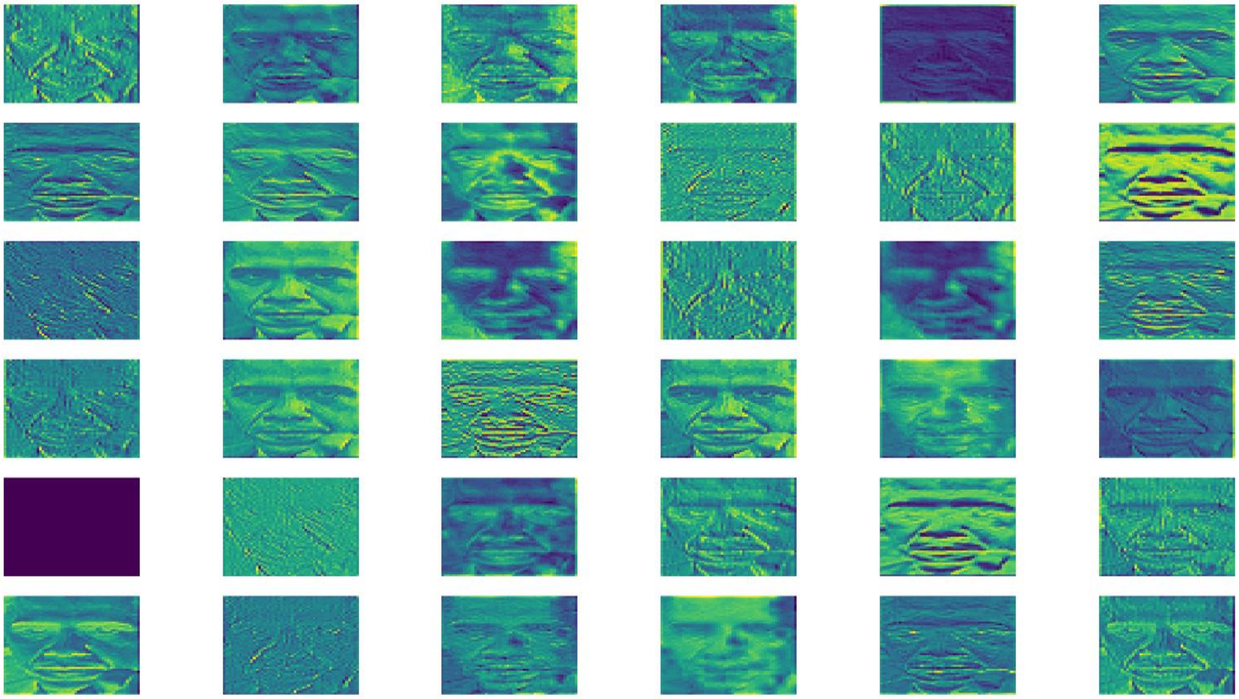


Figure 7.1 - Visualization of Feature Maps of from Custom MobileNet Model on RAF Dataset for Angry Emotion

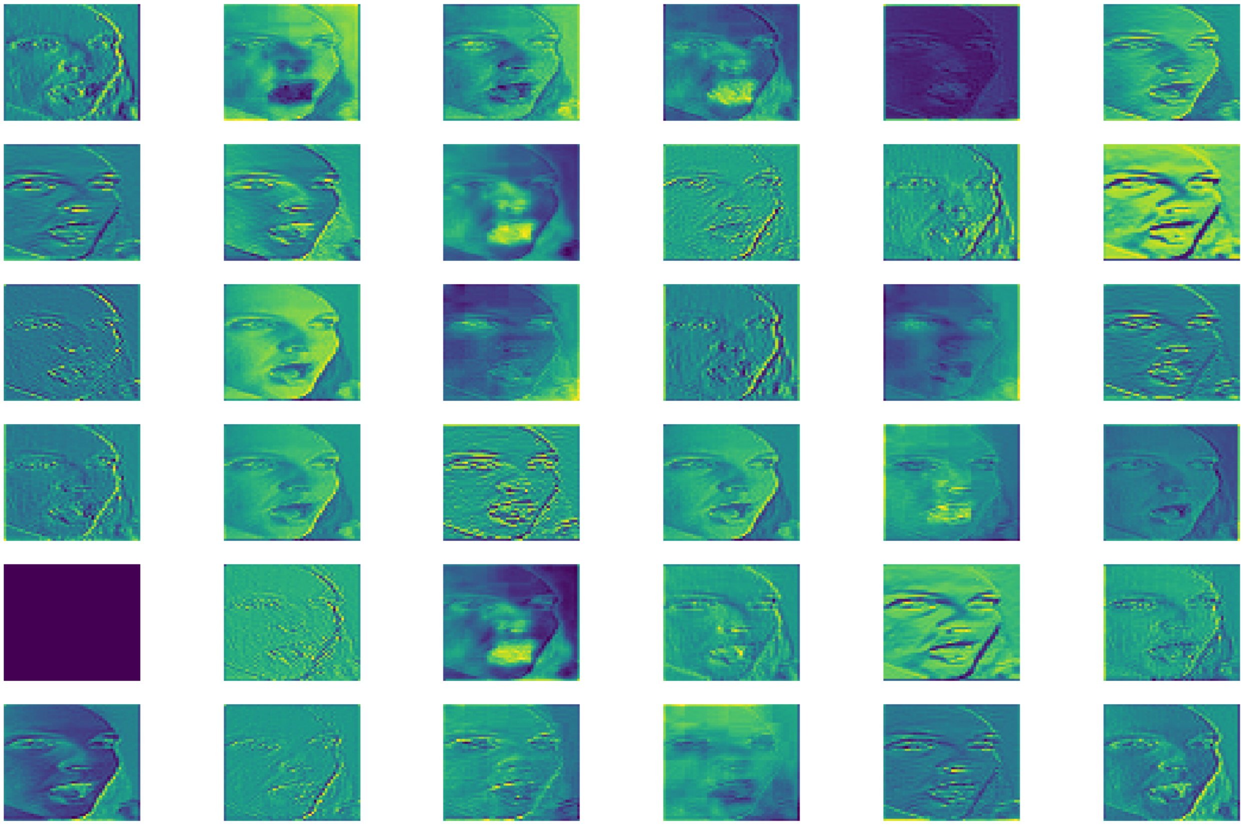


Figure 7.2 - Visualization of Feature Maps of from Custom MobileNet Model on RAF Dataset for Surprise Emotion

The figure 7.1 and figure 7.2 presents a visualization of feature maps extracted from our custom MobileNet model. Specifically, it showcases the first 36 feature maps from the 8th to the 16th layers of the model, generated in response to an ‘angry’ and ‘surprise’ facial expression image from the RAF dataset. Each subplot within the figure represents a unique feature map, highlighting various patterns and details captured by the model. These visualizations offer a window into the intricate processing carried out by the neural network, illustrating how different layers and filters within the model respond to specific features within the facial image. This insightful display aids in understanding the model’s interpretive capabilities, particularly in discerning subtle emotional cues from facial expressions, thus contributing to our comprehension of the model's internal workings in emotion recognition tasks.

The Figure 7.3 presents a collection of Grad-CAM visualizations that highlight the areas within facial expressions deemed most influential for emotion classification by a custom MobileNet model. Grad-CAM, or Gradient-weighted Class Activation Mapping, is a technique used to visualize the regions of an input image that are important for predictions from Convolutional Neural Networks. In these visualizations, warmer colors indicate regions with higher importance for the model's decision-making process. The technique effectively uncovers the facial features such as the eyebrows, eyes, mouth, and nose contours that the model finds most indicative of each emotion. This nuanced insight into the model's focus areas underscores the complex interplay of facial features in conveying emotions and demonstrates the model's ability to attend to key aspects of the human face for accurate emotion detection. This figure would be aptly placed in the sections of the document discussing model interpretability or within the analysis of the model’s performance.

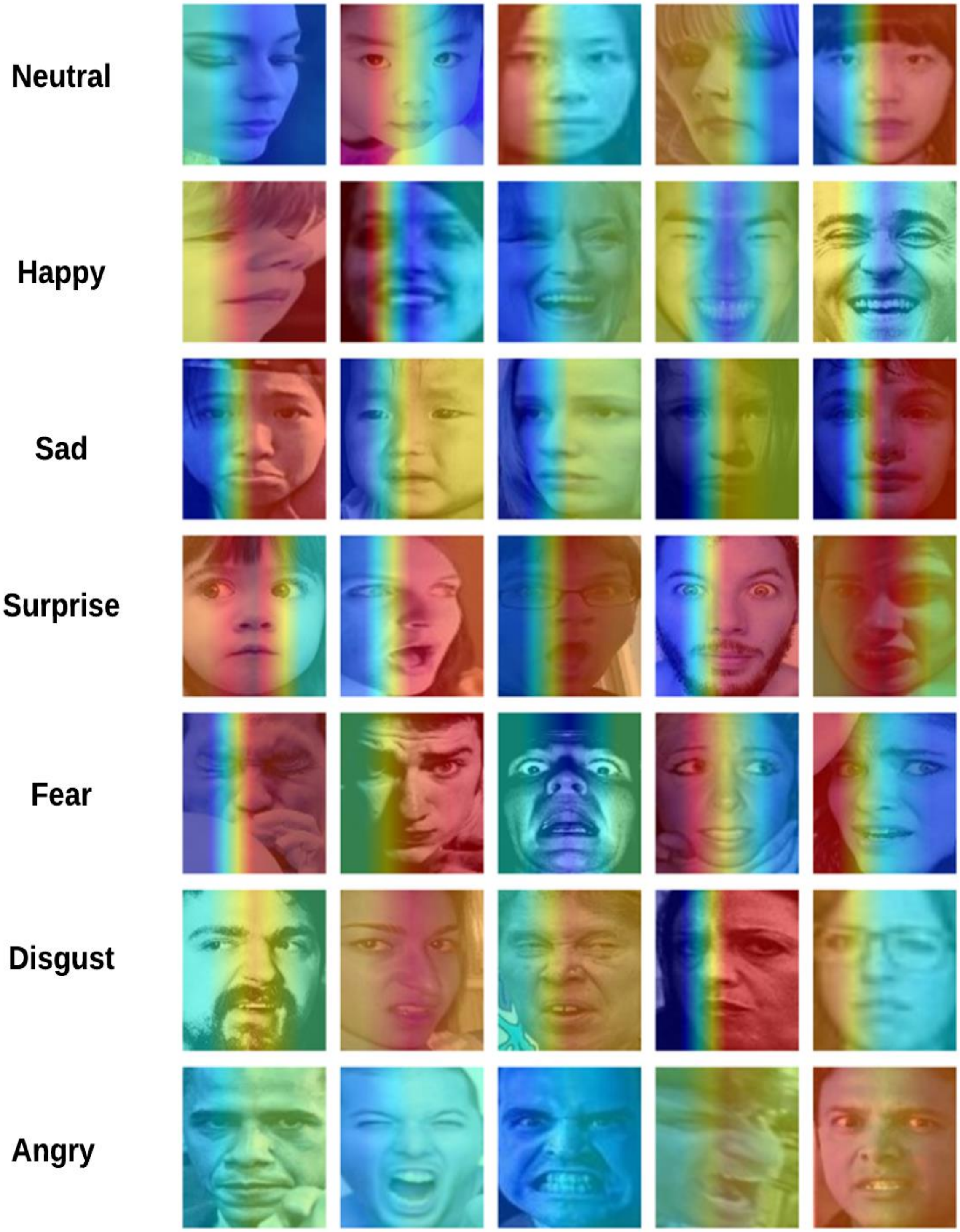


Figure 7.3 - Feature Importance Visualization for Emotion Recognition Using Grad-CAM

By comparing the performance of different architectures, we aim to identify the most suitable approach for precise human emotion recognition. Additionally, we remain open to considering other promising deep learning models that have demonstrated effectiveness in similar tasks. This comprehensive exploration will enable us to determine the optimal deep learning framework for our Human Emotion Recognition from Facial Expressions.

5.8 Evaluation Metrics

Evaluation metrics such as accuracy, loss, precision, recall, and F-score are used to assess the performance of a model in predicting emotions. These metrics provide valuable insights into the model's ability to classify emotions correctly.

1) Accuracy:

Accuracy measures the proportion of correct predictions made by the model. It is calculated by dividing the number of correctly classified emotions by the total number of predictions.

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ emotions}{Total\ number\ of\ predictions}$$

2) Precision:

Precision measures the proportion of correctly predicted positive emotions out of all the emotions predicted as positive. It is calculated by dividing the number of true positive predictions (correctly predicted positive emotions) by the total number of predictions classified as positive.

$$Precision = \frac{Number\ of\ true\ positive\ (TP)}{Number\ of\ predicted\ positives\ (TP + FP)}$$

3) Recall:

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive emotions out of all the actual positive emotions. It is calculated by dividing the number of true positive predictions (correctly predicted positive emotions) by the sum of true positive predictions and false negative predictions (missed positive emotions).

$$Recall = \frac{Number\ of\ true\ positives\ (TP)}{Number\ of\ Actual\ positives(TP + FN)}$$

4) F - Score:

The F-score is the harmonic mean of precision and recall. It is given by

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

6 Real Time Emotion Recognition from Facial Expressions

After the comprehensive development and training of deep learning models for emotion recognition from static facial expressions, our focus shifted towards applying these models to real-time video sequences. This advancement is pivotal in translating the static emotion recognition capabilities into dynamic, real-time applications, significantly broadening the scope of potential applications in fields like interactive media, telehealth, and human-computer interaction.

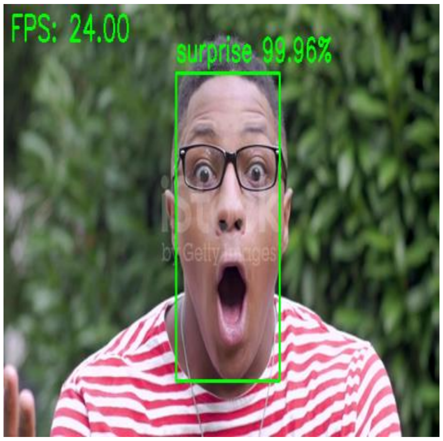
The ability to recognize human emotions in real-time from facial expressions in videos has vast applications, ranging from enhancing user experience in digital platforms to supporting mental health assessments. Leveraging advanced deep learning models, this study presents a robust model capable of processing live video feeds, detecting human faces, and accurately classifying their emotional states.

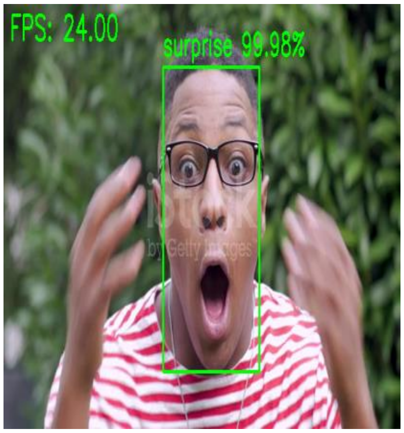
The real-time processing of video involves handling a continuous stream of frames, where each frame is subject to face detection and emotion classification. The RetinaFace ResNet50 model plays a crucial role in the initial phase, accurately detecting faces in diverse and dynamic conditions. Once the faces are detected and extracted from each frame, they are resized and preprocessed to align with the input requirements of our custom MobileNet-based emotion recognition model.

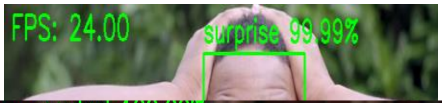
Our implementation employs TensorFlow for model execution, OpenCV for video processing, and ipywidgets for creating an interactive interface within Jupyter Notebooks.

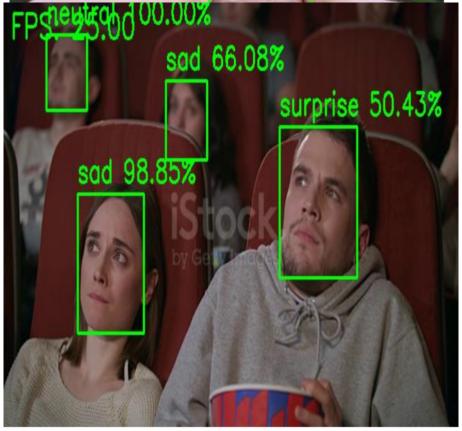
This setup enables the real-time display of video frames annotated with detected faces and their corresponding emotion labels. Additionally, we incorporated features like frame rate display and confidence scores for each emotion prediction to enhance the system's transparency and user experience.

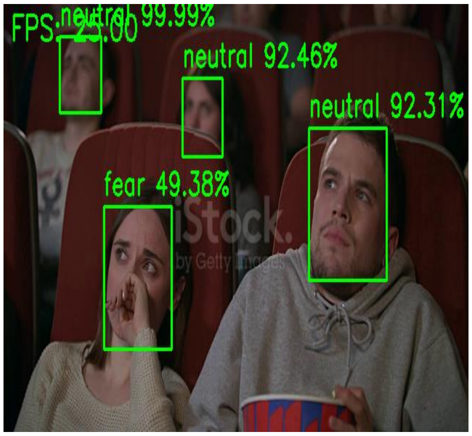
In summary, our project successfully demonstrates the capability to recognize and classify human emotions in real time from video sequences, employing a face detection and deep learning-based emotion classification techniques.

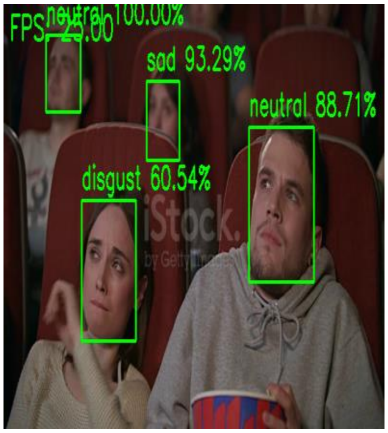












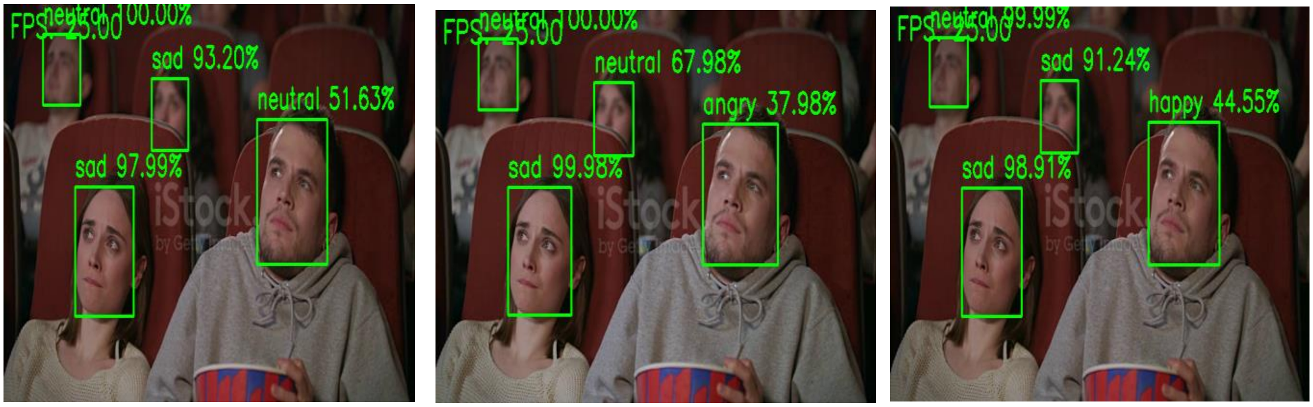


Figure 8 - Sample Images from Real-Time Video Feed.

Figure 8: Sample Images from Real-Time Video Feed Demonstrating Emotion Recognition Performance. This collection showcases various frames extracted from live video streams, each annotated with detected facial expressions and corresponding emotion labels as identified by our deep learning model.

7 Results and Discussion

In this section, we present the results of our experimentation on emotion classification using various deep learning models. The models were trained, validated, and tested on three distinct datasets: FER2013, RAF, and Affectnet. For the FER2013 and RAF datasets, we conducted evaluations both before and after data augmentation. We will discuss performance of each model, examining classification metrics.

7.1 Performance Evaluation of CNN on RAF Dataset Before and After Augmentation

For the RAF dataset, we evaluated the CNN model both before and after data augmentation. Below are the classification reports for the CNN model before and after augmentation.

TABLE I
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CNN MODEL ON RAF DATASET BEFORE AUGMENTATION

Emotion Class	Precision	Recall	F1 Score
Angry	0.62	0.69	0.65
Disgust	0.49	0.28	0.36
Fear	0.88	0.28	0.43
Happy	**0.86**	**0.94**	**0.90**
Neutral	0.68	0.82	0.75
Sad	0.81	0.54	0.65
Surprise	**0.76**	**0.78**	**0.77**

The CNN Model achieved an accuracy of 77% on the RAF dataset before augmentation, as depicted in Table VII. This robust accuracy suggests the model's proficiency in classifying emotions from images within the RAF dataset. Analyzing the classification metrics for each emotion class, it is evident that the model performed exceptionally well in recognizing "Happy", "Surprise” and "Neutral" emotions, with F1 scores of 0.90, 0.77 and 0.75 respectively. It also exhibited high precision for "Angry" and "Sad" emotions. However, the model faced challenges in accurately classifying "Disgust" and "Fear" emotions, as reflected by lower precision, recall, and F1 scores for these classes.

TABLE II
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CNN MODEL ON RAF DATASET AFTER AUGMENTATION

Emotion Class	**Precision**	**Recall**	**F1 Score**
Angry	0.77	0.54	0.63
Disgust	0.43	0.39	0.41
Fear	0.67	0.42	0.52
Happy	**0.89**	**0.92**	**0.90**
Neutral	0.72	0.77	0.74
Sad	0.68	0.74	0.71
Surprise	0.83	0.73	0.78

After data augmentation, the CNN Model showed mixed results: high performance on "Happy" emotions (F1 score of 0.90), moderate on "Surprise" (0.78) and "Sad" (0.71), but lower on "Angry" (0.63), "Disgust" (0.41), and "Fear" (0.52). The "Neutral" category had a recall of 0.77 but a reduced precision, leading to an F1 score of 0.74. These outcomes highlight the varied impact of data augmentation on emotional recognition accuracy.

7.2 Performance Evaluation of Custom MobileNet with Patch Extraction and Attention Mechanism on RAF Dataset Before Augmentation

TABLE III
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CUSTOM MOBILENET MODEL ON RAF DATASET BEFORE AUGMENTATION

Emotion Class	Precision	Recall	F1 Score
Angry	0.66	0.77	0.71
Disgust	0.52	0.45	0.48
Fear	0.67	0.42	0.52
Happy	**0.93**	**0.91**	**0.92**
Neutral	0.74	0.86	0.79
Sad	0.82	0.76	0.80
Surprise	0.82	0.75	0.79

The custom MobileNet with patch extraction and attention mechanism achieved an accuracy of 82% on the RAF dataset with 5-fold cross-validation. This model demonstrated strong performance in classifying various emotions. Notably, it excelled in recognizing "happy", "surprise" and "sad" emotions, achieving high precision, recall, and F1-scores for these classes. It also exhibited a balanced performance for other emotions, including "neutral," and "angry." This robust accuracy and comprehensive emotion classification highlight the effectiveness of the custom MobileNet architecture with patch extraction and attention mechanism on the RAF dataset.

.

7.3 Performance Evaluation of Models on RAF Dataset Before and After Augmentation

For the RAF dataset, the efficacy of various deep learning models was assessed, taking into account their performance before and after the implementation of data augmentation strategies. The findings are encapsulated in Table IV and Table V below.

TABLE IV

PERFORMANCE METRICS OF MODEL ON RAF DATASET BEFORE AUGMENTATION

Model	Accuracy
CNN	0.77
VGG16	0.72
EFFICIENTNETV2B0	0.73
INCEPTIONV3	0.78
CUSTOM MOBILENET WITH PATCH EXTRACTION AND ATTENTION MECHANISM	**0.82**

Prior to augmentation, the Custom MobileNet model, enhanced with patch extraction and attention mechanism, attained the highest accuracy of 82%. The CNN model followed closely with a commendable accuracy of 77%. The EfficientNetV2B0 demonstrated its robustness with an accuracy of 73%, while InceptionV3 displayed a competitive accuracy of 78%. VGG16, however, lagged slightly behind with an accuracy of approximately 72.62%.

TABLE V
PERFORMANCE METRICS OF MODEL ON RAF DATASET AFTER AUGMENTATION

Model	Accuracy
CNN	0.78
VGG16	0.74
EFFICIENTNET	0.79
INCEPTIONV3	0.78
CUSTOM MOBILENET WITH PATCH EXTRACTION AND ATTENTION MECHANISM	**0.81**

Post-augmentation, a nuanced shift in performance was noted. The CNN model saw a slight increase in accuracy to 78%. EfficientNetV2B0's accuracy improved to 79%, showcasing the benefits of augmentation. The VGG16 model's performance remained relatively stable. InceptionV3 maintained its accuracy at 78%, demonstrating consistency in performance. Notably, the Custom Model still led the group with an accuracy of 81%, albeit a marginal decreases post-augmentation.

These comparative insights underscore the Custom MobileNet model's superior performance, achieving the highest accuracy both before and after augmentation. The augmentation process has contributed to performance enhancements in most models, with the CNN and EfficientNetV2B0 models displaying significant improvements.

7.4 Performance Evaluation of CNN on fer2013 Dataset Before and After Augmentation

For the FER2013 dataset, we evaluated the CNN model both before and after data augmentation. Below are the classification reports for the CNN model before and after augmentation.

TABLE VI
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CNN MODEL ON FER2013 DATASET BEFORE AUGMENTATION

Emotion Class	**Precision**	**Recall**	**F1 Score**
Angry	0.56	0.53	0.55
Disgust	0.73	0.50	0.59
Fear	0.55	0.40	0.46
Happy	**0.84**	**0.81**	**0.83**
Neutral	0.57	0.58	0.57
Sad	0.47	0.58	0.52
Surprise	0.71	0.84	0.77

Table VI presents emotion classification metrics achieved by the CNN model when applied to the FER2013 dataset before the implementation of data augmentation techniques. The table includes precision, recall, and F1 score values for each emotion class, providing insights into the model's performance in accurately classifying different emotional states. The CNN achieved an accuracy of 63% on the FER2013 dataset before augmentation. The model displayed a mixed performance in emotion recognition on the FER2013 dataset before any data augmentation. It excelled in identifying happiness and surprise, with high precision of 0.81 for happy, 0.71 for surprise and recall of 0.81 for happy and 0.84 for surprise respectively, reflecting its ability to capture the distinctive facial expressions associated with these emotions. However, the model faced challenges in recognizing sad and fear, as indicated by lower precision and recall values, suggesting that it might struggle with the subtler cues linked to these emotions. Additionally, while the model achieved a balanced performance for neutral expressions, it demonstrated room for improvement in differentiating anger and disgust. These findings underscore the importance of fine-tuning the model to enhance its performance, especially for complex and less distinct emotions like sad and fear.

TABLE VII
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CNN MODEL ON FER2013 DATASET AFTER AUGMENTATION

Emotion Class	Precision	Recall	F1 Score
Angry	0.55	0.50	0.52
Disgust	0.36	0.59	0.45
Fear	0.49	0.40	0.44
Happy	**0.84**	**0.82**	**0.83**
Neutral	0.58	0.54	0.56
Sad	0.46	0.58	0.51
Surprise	**0.78**	**0.75**	**0.76**

Post-augmentation, the CNN model's ability to classify emotions on the FER2013 dataset presented a mixed outcome. While the model retained a high degree of precision and recall in detecting "Happy" and "Surprise" emotions, yielding F1 scores of 0.83 and 0.76 respectively, the classification of "Disgust" significantly improved in recall to 0.59, yet the precision remained low at 0.36, resulting in an F1 score of 0.45. "Fear" and "Neutral" emotions saw a moderate decline in classification accuracy, and "Sad" emotions experienced a noticeable improvement in recall, leading to an F1 score of 0.51. However, the most marked decrease was observed in the "Angry" emotion class, with a modest precision of 0.55 and an F1 score of 0.52. Overall, these after-augmentation results, achieving an accuracy of 62%, reveal that the augmentation process can have complex effects on a model's performance, necessitating a careful balance to enhance model robustness without compromising classification accuracy.

7.5 Performance Evaluation of Custom MobileNet with Patch Extraction and Attention Mechanism on FER2013 Dataset Before Augmentation

TABLE VIII
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CUSTOM MOBILENET MODEL ON FER2013 DATASET BEFORE AUGMENTATION

Emotion Class	**Precision**	**Recall**	**F1 Score**
Angry	0.46	0.53	0.49
Disgust	0.69	0.43	0.53
Fear	0.41	0.42	0.42
Happy	**0.82**	**0.81**	**0.81**
Neutral	0.58	0.48	0.53
Sad	0.48	0.46	0.47
Surprise	**0.67**	**0.81**	**0.73**

Table VIII provides a detailed breakdown of the classification metrics for the Custom MobileNet model on the FER2013 dataset before augmentation. This model is evaluated based on its performance in classifying various emotion classes.

In this context, the Custom MobileNet model achieved an accuracy of 59% on the FER2013 dataset before augmentation. It demonstrates varying levels of precision, recall, and F1-Score for different emotion classes. For example, it exhibits higher performance in classifying "Happy" and "Surprise" emotions, as indicated by their higher F1-Scores, while it faces challenges in accurately classifying "Angry" and "Fear" emotions, as reflected by their lower scores. The table provides a comprehensive view of the model's classification capabilities across different emotions in the dataset.

7.6 Performance Evaluation of Models on FER2013 Dataset Before and After Augmentation

TABLE IX
PERFORMANCE METRICS OF MODEL ON FER2013 DATASET BEFORE AUGMENTATION

Model	Accuracy
CNN	**0.63**
EFFICIENTNET	0.54
RESNET18	0.55
DENSENET121	0.61
CUSTOM MOBILENET	0.59

Before augmentation, the CNN model stood out with the highest accuracy of 63%, slightly ahead of DenseNet121 at 61%. The Custom Model also performed commendably with an accuracy of 59%. EfficientNetV2B0 and ResNet18 were in a similar range with accuracies of 54% and 55%, respectively.

TABLE X
PERFORMANCE METRICS OF MODEL ON FER2013 DATASET AFTER AUGMENTATION

Model	Accuracy
CNN	**0.62**
RESNET18	0.54
DENSE121	0.60
EFFICIENTNET	0.51
CUSTOM MOBILENET	0.60

TABLE X provides an overview of the model's performance on the FER2013 dataset after augmentation. After augmentation, the CNN model maintained the highest accuracy, slightly decreasing to 62%. The Custom Model saw an improvement, achieving an accuracy of 60%. DenseNet121's performance remained stable at 60%, while EfficientNetV2B0 saw a decrease to 51%, and ResNet18 showed a minor drop to 54%. The augmentation process had a varied impact on the models, with some showing resilience and others facing decreases in performance metrics. These results indicate that while some models like the Custom Model and CNN demonstrated robustness against the challenges posed by augmentation, models such as EfficientNetV2B0 were more susceptible to performance drops. It is crucial to consider these effects when applying data augmentation to improve model generalization and to ensure that the models remain effective across various emotional states.

7.7 Performance Evaluation of Models on Affectnet Dataset

TABLE XI
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY CUSTOM MOBILENET MODEL ON AFFECNET-7 DATASET

Emotion Class	Precision	Recall	F1 Score
Angry	0.52	0.57	0.54
Disgust	0.51	0.50	0.50
Fear	0.54	0.61	0.57
Happy	**0.78**	**0.77**	**0.77**
Neutral	0.48	0.42	0.45
Sad	0.50	0.56	0.53
Surprise	0.59	0.47	0.52

Table XI above presents the performance metrics for Custom MobileNet with Patch Extraction and Attention Mechanism on the Affectnet-7 dataset. The model achieved an overall accuracy of 0.56 on this dataset. Looking at the individual emotion classes, the model exhibited varying levels of performance. It performed best in recognizing "Happy" emotions with a precision of 0.78 and an F1 score of 0.77. However, it faced challenges in accurately classifying "Neutral" and "Sad" emotions, as indicated by lower precision and recall values.

TABLE XII
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY VIT B16 MODEL ON AFFECTNET-7 DATASET

Emotion Class	Precision	Recall	F1 Score
Angry	0.52	0.58	0.55
Disgust	0.64	0.51	0.57
Fear	0.63	0.58	0.57
Happy	**0.78**	**0.86**	**0.82**
Neutral	0.50	0.53	0.51
Sad	0.57	0.61	0.59
Surprise	0.59	0.55	0.57

Vit B16 was evaluated on the Affectnet-7 dataset, achieving an overall accuracy of 0.6019. The model demonstrated relatively strong performance in recognizing "Happy" emotions with a precision of 0.78 and an F1 score of 0.82. It also performed well in classifying "Fear" emotions with an F1 score of 0.60. However, it faced challenges in accurately recognizing "Neutral" and "Sad" emotions.

TABLE XIII
EMOTION CLASSIFICATION METRICS WHEN CLASSIFIED BY VIT B32 MODEL ON AFFECTNET-8 DATASET

Emotion Class	Precision	Recall	F1 Score
Angry	0.47	0.52	0.49
Disgust	0.60	0.44	0.51
Fear	0.62	0.51	0.56
Happy	0.61	0.77	0.68
Neutral	0.40	0.48	0.43
Sad	0.50	0.52	0.51
Surprise	0.52	0.55	0.53
Contempt	0.55	0.43	0.48

Table XIII above illustrates the performance metrics of Vit B32 on the Affectnet 8 dataset. This model achieved an accuracy of 0.53, indicating its ability to classify emotions across eight distinct categories. Looking at individual emotion classes, Vit B32 performed well in recognizing "Happy" emotions, achieving a precision of 0.61 and an F1 score of 0.68. It also exhibited relatively strong performance in classifying "Fear" emotions with an F1 score of 0.56. However, it faced challenges in accurately classifying emotions such as "Neutral" and "Contempt," as indicated by lower precision and recall values.

TABLE X1V
PERFORMANCE METRICS OF MODEL ON AFFECTNET DATASET

Model	Accuracy
CNN	**0.509**
RESNET18	0.45
EFFICIENTNET	0.51
DENSENET121	0.51
DENSENET121(PCA)	0.54
INCEPTIONV3	0.46
VIT B32	0.53
VIT B16 (Affectnet 7)	0.60

Table XIV summarizes the performance metrics of various models on the Affectnet dataset. These models were evaluated based on their accuracy in classifying emotions within the dataset. The table provides a comparative view of their performance, with each model's accuracy score presented. The models include CNN, ResNet18, EfficientNet, DenseNet121, DenseNet121 with PCA, InceptionV3, Vit B32, and Vit B16 (Affectnet 7). These accuracy scores offer insights into the effectiveness of each model in recognizing and classifying emotions within the Affectnet dataset.

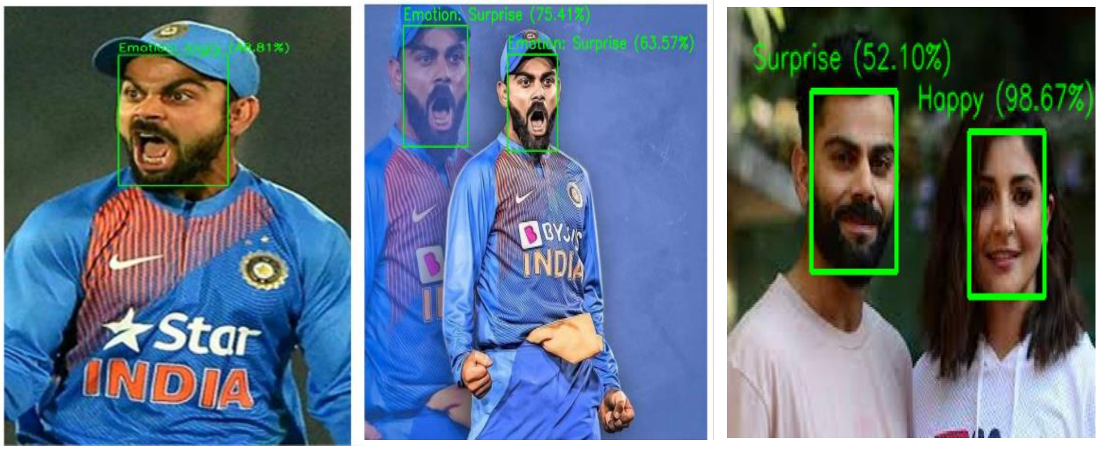


Figure 9 - Sample Images with Predicted Emotions

8 Conclusion

In the field of real-time emotion recognition from live video feeds, our study explored various methods and models thoroughly. We used the RetinaFace ResNet50 for precise face detection and extraction. We also looked into different deep learning models, like EfficientNet, ResNet, VGG16, and Vision Transformers, for classifying emotions across three distinct datasets.

We developed a specialized Custom CNN model with patch extraction and an attention mechanism to better understand key features for emotion prediction.

During our research, we faced certain challenges. The FER2013 and RAF datasets had imbalanced data, so we used data augmentation techniques. Interestingly, we noticed that, in some cases, the models performed better before augmentation. On the other hand, the Affectnet dataset, with its high-resolution images, presented challenges when emotions were mixed within a class.

In summary, our study provides a detailed comparison of various models for emotion recognition, highlighting their strengths and weaknesses. The potential for fine-tuning and improving these models remains an avenue for future research.

In conclusion, the field of real-time emotion recognition from facial expressions continues to evolve. The challenges encountered during this study highlight the potential for further refinement and fine-tuning of methods in this area. Such efforts can contribute to advancements in emotion recognition technology, with possible applications in human-computer interaction and areas like mental health monitoring.

Table of Figures

Bibliography

[1] Facial Emotion Recognition (FER2013), from Kaggle, https://www.kaggle.com/c/challenges-learning-facial-expression-recognition-challenge/-

[2] Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: CVPR. pp. 2584–2593 (2017)

[3] Mollahosseini, Ali & Hasani, Behzad & Mahoor, Mohammad. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing. PP. 10.1109/TAFFC.2017.2740923.

[4] T. Chang, G. Wen, Y. Hu, and J. Ma, "Facial Expression Recognition Based on Complexity Perception Classification Algorithm," arXiv e-prints, Accessed on: February 01, 2018, https://arxiv.org/ftp/arxiv/papers/1803/1803.00185.pdf

[5] H. Kishan Kondaveeti and M. Vishal Goud, "Emotion Detection using Deep Facial Features," 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI), Buldhana, India, 2020, pp. 1-8, DOI: 10.1109/ICATMRI51801.2020.9398439.

[6] D.Y. Liliana. "Emotion recognition from facial expression using deep convolutional neural network," Journal of Physics: Conference Series, vol. 1193, 012004, 2019. DOI: 10.1088/1742- 6596/1193/1/012004.

[7] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I. Na, and S. H.Kim, "Facial Emotion Recognition Using an Ensemble of MultiLevel Convolutional Neural Networks," International Journal ofPattern Recognition and Artificial Intelligence, 2018, DOI: https://doi.org/10.1142/S0218001419400159

[8] Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A. (2019). Real Time Emotion Recognition from Facial Expressions Using CNN Architecture. 2019 Medical Technologies Congress (TIPTEKNO).

[9] Yahui Nan, Jianguo Ju, Qingyi Hua, Haoming Zhang, Bo Wang, A-MobileNet: An approach of facial expression recognition, Alexandria Engineering Journal, 2022, DOI: 10.1016/j.aej.2021.09.066

[10] M. A. H. Akhand, Shuvendu Roy 1, Nazmul Siddique, Md Abdus Samad Kamal, Tetsuya Shimamura, Facial Emotion Recognition Using Transfer Learning in the Deep CNN, 2021, from https://www.mdpi.com/2079-9292/10/9/1036

[11] K Soumya Kuruvayil, Suja Palaniswamy, Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning, Journal of King Saud University – Computer and Information Sciences 34 (2022)7271–7282

[12] Khan, A.R. Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. Information 2022, 13, 268. https://doi.org/ 10.3390/info13060268

[13] Jia Guo, Jiankang Deng, Alexandros Lattas, Stefanos Zafeiriou, Huawei, Insight Face, Imperial College London, "Sample and Computation Redistribution for Efficient Face Detection", International Conference on Learning Representations (ICLR) 2022