Facial Emotion Recognition using CNN with Grad-CAM for Explainability

The model is deployed at: https://www.alfateh.tech

Prepared by: Mohamedalfateh Tagalsir Maroof Saeed

February 2025

Table of Contents

List of Tables

List of Figures

1. Introduction

Facial Expression Recognition (FER) is a pivotal area within computer vision and affective computing, focusing on the automated identification and categorization of human facial expressions. This technology has significant applications in human-computer interaction, psychological research, and security systems.

The human face is a rich source of emotional information, with expressions conveying essential cues about an individual's internal state. Automating the recognition of these expressions enables more intuitive interactions between humans and machines. Traditional FER systems often relied on handcrafted features and shallow classifiers, which were limited in their ability to capture the complex variations in facial expressions across different individuals and cultures.

FER systems traditionally relied on handcrafted features combined with classical machine learning methods like Support Vector Machines (SVMs). For instance, histogram-based features, such as Histogram of Oriented Gradients (HOG), have been widely used to describe facial contours and classify emotions. While effective under controlled conditions, these approaches often failed to generalize to diverse and unconstrained real-world scenarios [1]

Recent advancements in deep learning have markedly enhanced FER capabilities. Convolutional Neural Networks (CNNs), in particular, have demonstrated proficiency in learning hierarchical feature representations from facial images, leading to improved recognition accuracy. For instance, a study published in Scientific Reports adopted a deep neural network for facial emotion recognition, highlighting the effectiveness of deep learning approaches in this domain [1] [2].

The availability of large-scale annotated datasets has been instrumental in training robust FER models. Datasets such as the Extended Cohn-Kanade (CK+) and the Japanese Female Facial Expression (JAFFE) database have provided diverse facial expression samples, facilitating the development and benchmarking of FER algorithms [3].

Despite these advancements, FER systems face challenges in real-world applications. Variations in lighting, occlusions, head poses, and cultural differences can adversely affect recognition performance. To address these issues, researchers have explored techniques such as data augmentation, domain adaptation, and the integration of temporal information from video sequences to enhance system robustness. For example, a study on real-time facial expression recognition in the wild proposed disentangling 3D expression from identity to improve recognition accuracy under diverse conditions [4] [5] [6].

In summary, FER has evolved into a sophisticated interdisciplinary field, leveraging advances in machine learning, computer vision, and psychology. Ongoing research continues to address existing challenges, aiming to develop FER systems that perform reliably across varied and unconstrained environments.

2. Dataset

The dataset utilized in this study is the Facial Expression Recognition 2013 (FER-2013) dataset [7], a widely recognized benchmark in the field of facial emotion recognition. FER-2013 comprises approximately 35,887 grayscale images, each resized to a resolution of 48×48 pixels, depicting facial expressions categorized into seven distinct emotion classes:

1. Angry
2. Disgust
3. Fear
4. Happy
5. Sad
6. Surprise
7. Neutral.

The dataset is divided into three subsets: a training set containing 28,709 images and a test set also comprising 7,178 images, allowing for a structured evaluation of machine learning models.

To provide a comprehensive overview, Table (1) and Figure (1) illustrate the distribution of images across the seven emotion classes within each subset. Notably, the dataset suffers from class imbalance, with the "Disgust" category containing only 547 images compared to nearly 5,000 samples in other categories. Such imbalance can lead to biased model training, emphasizing the need for class rebalancing or weighting techniques during model optimization. Figure (2) depicts the class weight across the FER-2013 dataset.

Table 1: Distribution of Images Across Emotion Classes in the FER-2013 Dataset

Class	Train Dataset	Test Dataset	Total
Happy	1774	7215	8989
Sad	1247	4830	6077
Fear	1024	4097	5121
Surprise	831	3171	4002
Neutral	1233	4965	6198
Angry	958	3995	4953
Disgust	111	436	547
Total	7178	28709	35887

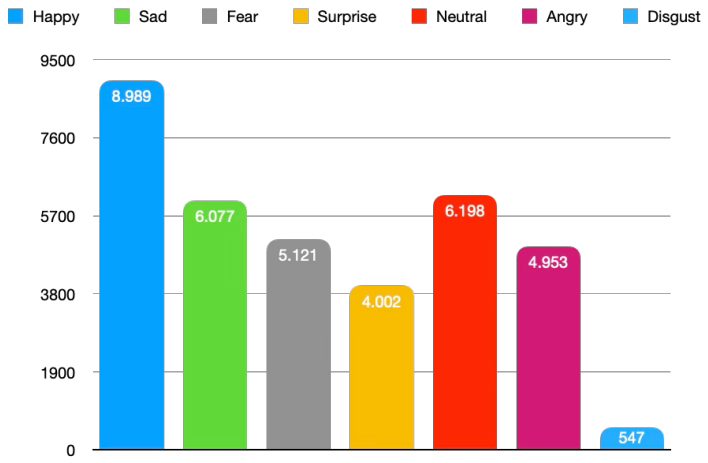

Figure 1: Class Distribution Analysis Across the FER-2013 Dataset

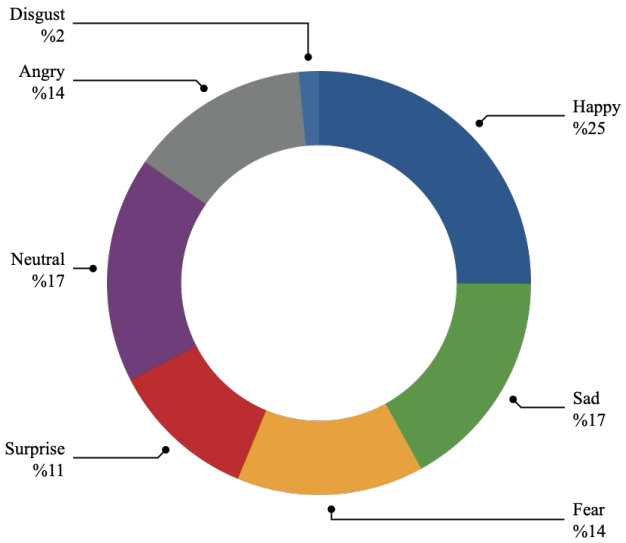

Figure 2: Class Weight Across the FER-2013 Dataset

The images in FER-2013 encompass diverse facial expressions captured under varying conditions, including differences in pose, lighting, and occlusions. This diversity enhances the dataset's utility for training robust models but also introduces challenges. Figure (3) showcases representative samples from each emotion class, emphasizing the visual complexity and inter-class variations inherent in the dataset.

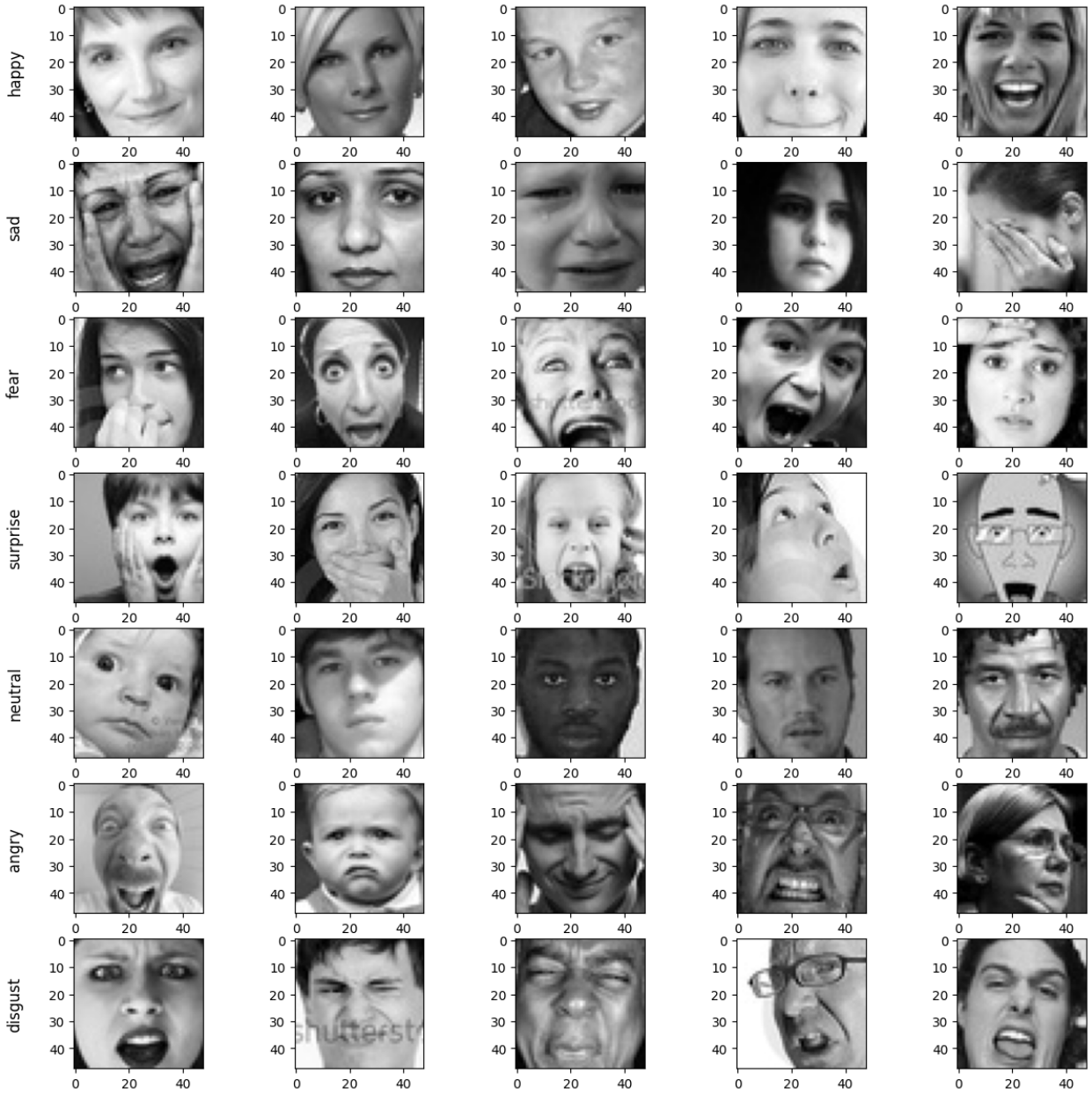

Figure 3: Image Samples from each Class.

Despite its extensive use, FER-2013 includes certain limitations, such as annotation errors and low-resolution images, which can hinder model performance. Researchers have highlighted instances of mislabeling and inconsistent quality across samples [8]. For instance, images with ambiguous expressions are sometimes incorrectly categorized, as depicted in Figure (4), which provides examples of mislabeled images. Figure (5) also includes some non-face images. Such inconsistencies necessitate data cleaning and augmentation techniques to mitigate their impact on model accuracy.

Figure 4: Non-face images

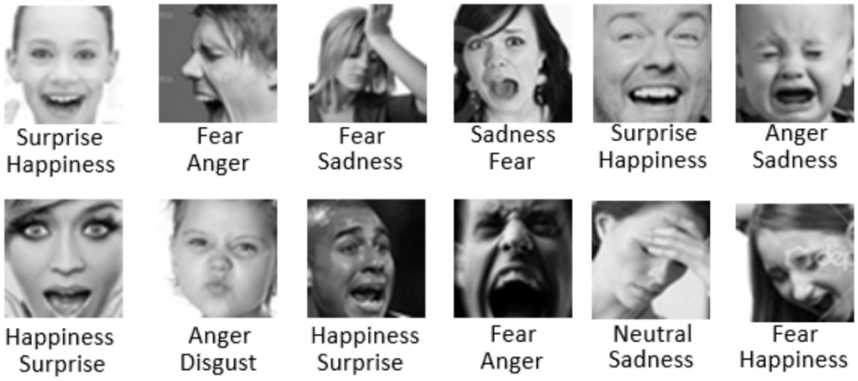

Figure 5: Mislabeled Images

To address these challenges, this study incorporates advanced data preprocessing strategies, including image augmentation (e.g., rotations, flips, and intensity adjustments) and class balancing using oversampling techniques. Additionally, Figure (6) demonstrates the effect of augmentation on the dataset.

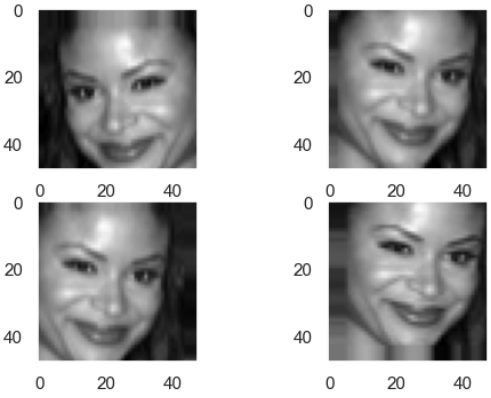

Figure 6: Augmented Data Samples

In conclusion, while FER-2013 serves as a robust foundation for training and evaluating facial expression recognition models, its inherent challenges underline the importance of preprocessing and augmentation strategies to ensure model robustness and reliability. By addressing these limitations, this study aims to build upon the strengths of FER-2013 while mitigating its drawbacks to enhance the overall accuracy and generalizability of the proposed model.

3. Preprocessing

Preprocessing is a fundamental step in preparing the Facial Expression Recognition 2013 (FER-2013) dataset for training deep learning models. This stage ensures that the input data is properly formatted, normalized, and augmented to maximize model performance and generalizability. Given the FER-2013 dataset consists of grayscale images of size 48 * 48, the following preprocessing steps were applied.

3.1 Resizing and Rescaling

All images were resized to 48 * 48 pixels to ensure uniformity across the dataset, which aligns with the original image dimensions. To facilitate faster convergence during training, pixel values were normalized to the range [0, 1] by dividing by 255. Normalization reduces the variance in input values, stabilizing the training process.

3.2 Channel Conversion

The FER-2013 dataset consists of grayscale images, while pre-trained models like VGGNet are designed to operate on 3-channel RGB images. To address this incompatibility, grayscale images were converted to pseudo-RGB format by duplicating the single grayscale channel across the three color channels. This conversion ensured compatibility without altering the semantic information of the images.

3.3 Data Augmentation

Data augmentation was employed to enhance the diversity of the training dataset and to reduce overfitting. The following augmentation techniques were applied:

- Rotation: Images were randomly rotated within a range of -30° to $+30^{\circ}$ to simulate variations in head pose.
- Horizontal Flipping: Images were horizontally flipped to account for symmetry in facial expressions.
- Zooming: A zoom range of 0.8 to 1.2 was applied to simulate variations in camera distance.
- Brightness Adjustments: Variations in brightness were introduced to mimic different lighting conditions.

These augmentations improved the model's ability to generalize across diverse real-world scenarios without requiring additional data collection.

3.4 Class Balancing

The FER-2013 dataset exhibits significant class imbalance, particularly with the "Disgust" class being underrepresented. For example, the "Disgust" class contains only 547 images compared to nearly 5,000 images in the "Happy" and "Neutral" classes. This imbalance can

lead to biased predictions favoring majority classes. To mitigate this, class weights were assigned during model training to penalize misclassifications of minority classes more heavily. Additionally, oversampling techniques were considered to ensure a more balanced representation of classes.

3.5 Batch Processing and Real-Time Augmentation

To efficiently process large amounts of data, images were organized into batches using the Keras ImageDataGenerator. This method reads images directly from directories and applies real-time preprocessing, reducing memory overhead. Each batch consisted of 32 images, with preprocessing steps such as resizing, rescaling, and augmentation applied on-the-fly.

3.6 Verification of Preprocessed Data

To ensure correctness, preprocessed images were visualized, and their dimensions, pixel intensity ranges, and color channels were verified. The shape of input batches was consistently confirmed to be (32, 48, 48, 3), indicating 32 images per batch, each resized to 48 * 48 pixels and converted to RGB format.

Preprocessing of the FER-2013 dataset involved resizing, normalization, channel conversion, data augmentation, class balancing, and batch processing. These steps addressed challenges such as the low resolution of images, class imbalance, and compatibility with pre-trained models like VGGNet. The preprocessing pipeline ensured the dataset was optimized for training deep learning models, improving the model's robustness and generalizability for facial expression recognition tasks.

4. Model Implementation

The implemented Convolutional Neural Network (CNN) is designed for efficient and accurate emotion classification of grayscale images (48×48), categorizing them into seven emotions: Happy, Sad, Angry, Neutral, Surprise, Fear, and Disgust. Below is a concise overview of the architecture and training strategy:

4.1 Model Architecture

• **Convolutional Layers:** Sequential layers with filter sizes of 32, 64, 128, and 256 progressively extract spatial features. Each uses a kernel size of 3×3, padding to preserve dimensions, and ReLU activation.

• **Batch Normalization:** Applied after each convolutional layer to normalize activations, improving stability and speeding convergence.

• **Max-Pooling:** Alternate convolutional layers are followed by max-pooling (2×2), reducing spatial dimensions while retaining essential features.

• **Dropout:** Dropout layers (rate = 0.25) reduce overfitting by deactivating random neurons during training.

• **Flattening Layer:** Converts 2D feature maps into a 1D vector for fully connected layers.

• **Dense Layers:** Includes a dense layer with 256 neurons, Batch Normalization, and Dropout (rate = 0.25), followed by a 7-neuron output layer with Softmax activation for probabilistic classification.

Figure 7 depicts the model architecture.

4.2 Training Strategy

The model uses categorical cross-entropy as the loss function and the Adam optimizer for efficient convergence. Data augmentation (e.g., rotations, shifts, flips) is applied to improve generalization. Figure (8) depicts the training and validation accuracy and loss during the training stage, while Table (2) shows the hyperparameters used.

4.3 Design Justification

• **Increasing Filter Depth:** Captures progressively complex features.

• **Batch Normalization:** Enhances stability and reduces training time.

• **Dropout Regularization:** Mitigates overfitting.

• **Softmax Output:** Provides interpretable probabilities for multi-class classification.

4.4 Output

The model predicts one of seven emotions based on the highest output probability.

This CNN architecture balances depth, regularization, and computational efficiency, ensuring robust performance in emotion recognition tasks and adaptability for additional datasets or classes.

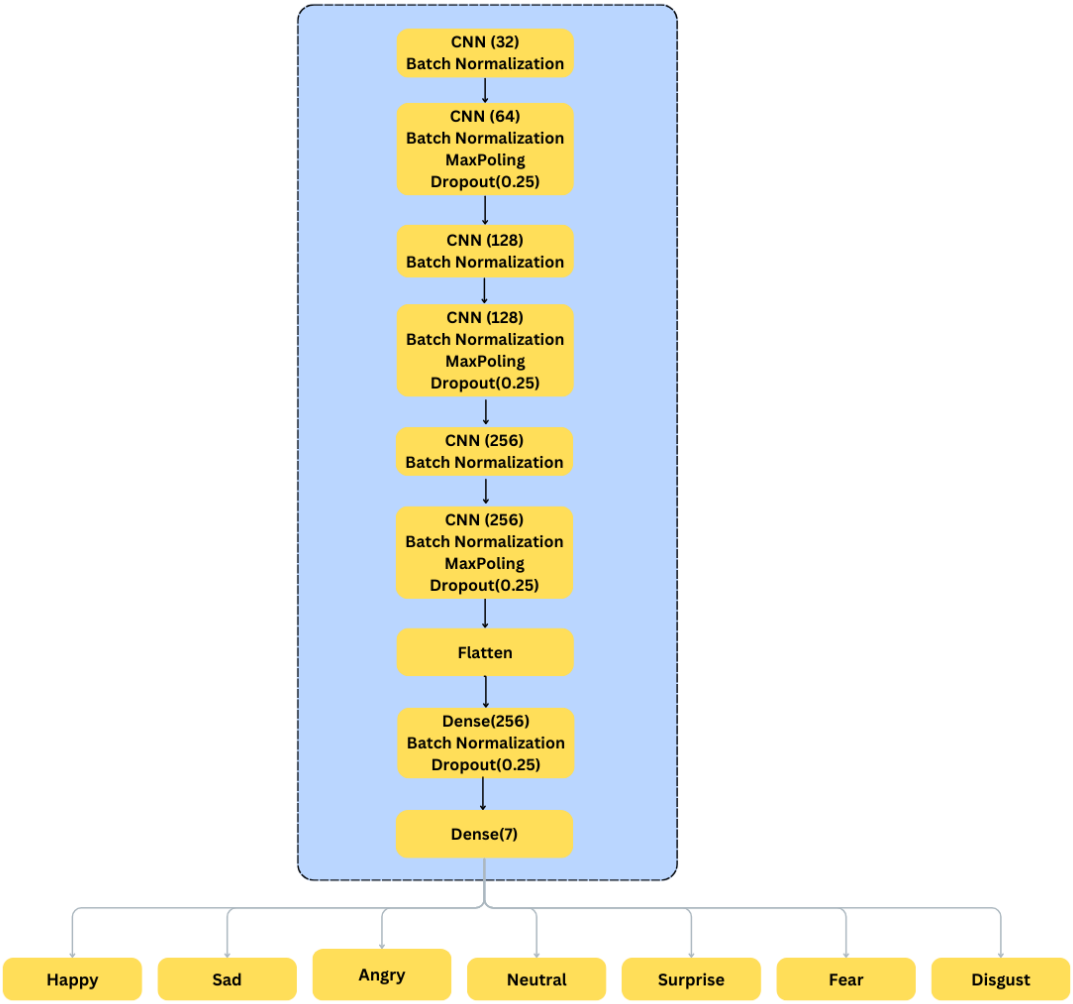

Figure 7: Model Architecture

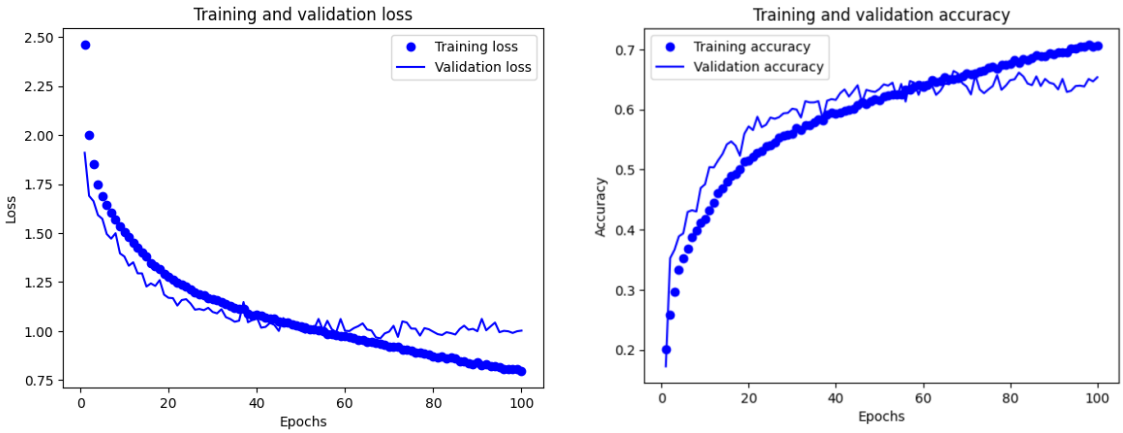

Figure 8: Training and Validation Accuracy and Loss During Training

Table 2: Model's hyperparameters

Hyperparameter Value	Value
Batch size 32	64
Optimizer Adam	Adam
Learning rate 0.001 (with decay)	Learning rate 0.0001 (with decay)
Epochs 20	100

5. Results and Discussion

The model achieved accuracy of 70.13% .The normalized confusion matrix reveals the model's strengths and limitations in emotion classification. It performs well in detecting "fear" (80%), "happy" (73%), and "sad" (67%), reflecting effective feature extraction for these emotions. However, significant misclassification is observed for "disgust" (only 25% accuracy), with overlap across multiple categories, and "neutral," where 27% of samples are misclassified, indicating feature ambiguity. "Angry" also shows confusion with "neutral" (25%), while "surprise" achieves perfect separation but may be underrepresented in the dataset. Figure (9) shows the normalized confusion matrix

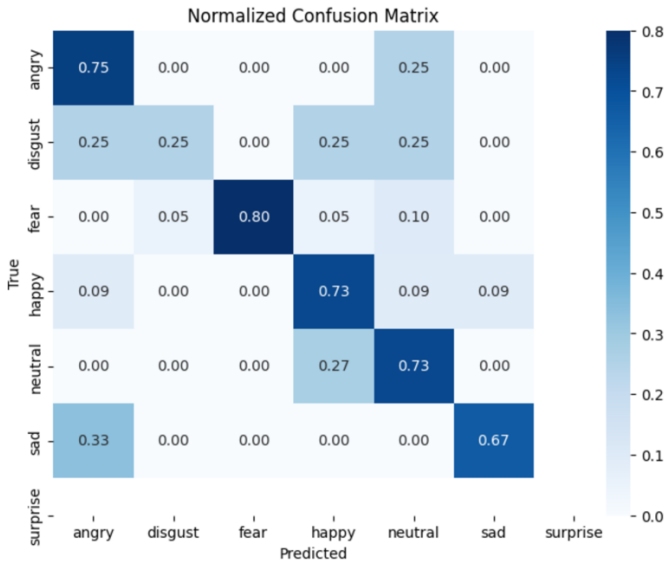

Figure 9: The Normalized Confusion Matrix

The performance comparison highlights the effectiveness of the proposed method, which achieves an accuracy of 70.13%, outperforming several existing approaches. While VGGNet [9] attains the

highest accuracy of 73.28%, the proposed method demonstrates superior results compared to CNN [1] (66.67%), SVM [1] (45.95%), CLCM [10] (63%), MobileNetV2 [10] (58%), and ShuffleNetV2 [10] (65%). This indicates that the proposed method offers a strong balance between accuracy and computational efficiency, positioning it as a competitive alternative to traditional and lightweight architectures for emotion classification tasks. Table (3) and Figure (10) are employed to present the literature comparison.

Table 3: Literature Comparison

Method	Accuracy
CNN[1]	66.67
SVM[1]	45.95
VGGnet [9]	73.28
CLCM [10]	63
MobileNetV2 [10]	58
ShuffleNet V2 [10]	65
Proposed Method	70.13

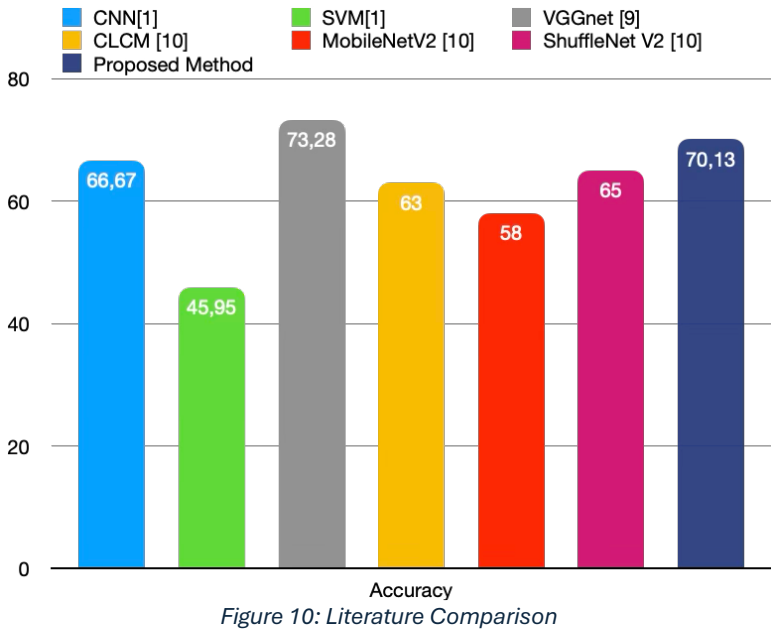

Figure 10: Literature Comparison

The Figure (11) demonstrates Grad-CAM visualizations for the emotions "Happy," "Fear," and "Angry," highlighting the areas of the image that the model focuses on

during its predictions. The heatmaps, overlaid on the original images, show that the model primarily attends to critical facial regions such as the eyes, mouth, and eyebrows, which are essential for distinguishing emotions. For "Happy," the model emphasizes the smile, while for "Fear," it focuses on the furrowed brows and tense facial features. Similarly, the "Angry" heatmap highlights the intense gaze and tight mouth region. These visualizations provide interpretability, validating that the model's predictions align with human understanding of emotion-related facial expressions.

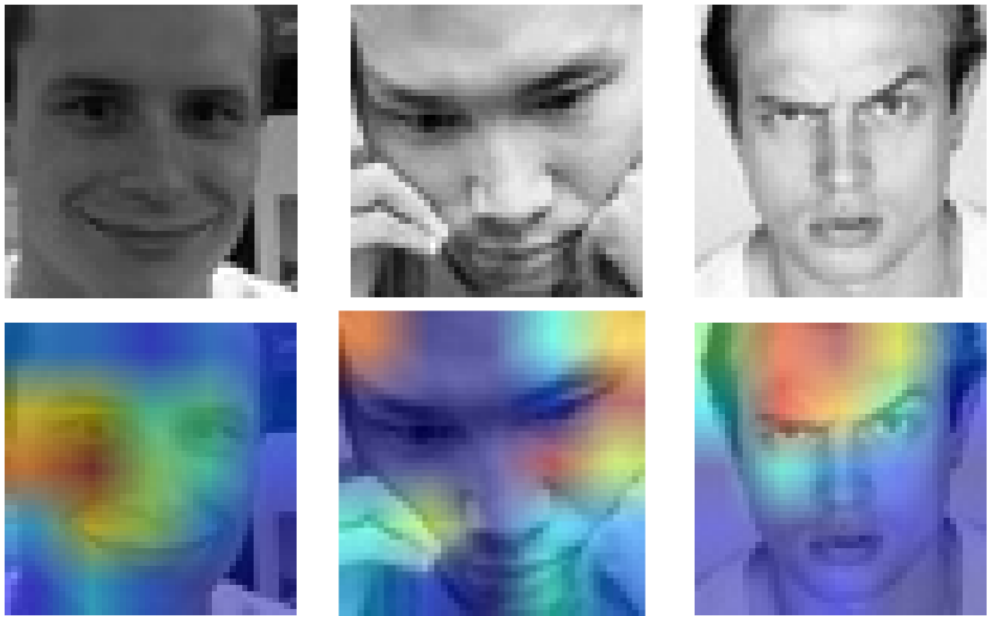

Figure 11: Grad-Cam Decision Explainability

6. Deployment

The model has been successfully deployed utilizing a robust stack of technologies, ensuring both performance and scalability:

• **Flask**: A lightweight web framework for building and managing the API.

• **JavaScript**: To provide dynamic and interactive functionality in the web interface.

• **HTML/CSS**: For a clean, user-friendly, and responsive front-end design.

• **Gunicorn**: A high-performance WSGI server for efficient handling of HTTP requests.

• **Nginx**: Serving as a reverse proxy to manage load balancing and ensure optimal traffic distribution.

• **Google Cloud**: Leveraging Google Cloud services for secure, scalable, and reliable deployment.

Figure (12) demonstrates the model's performance evaluation using various images. The model can be accessed and tested through the live deployment at: https://alfateh.tech/.

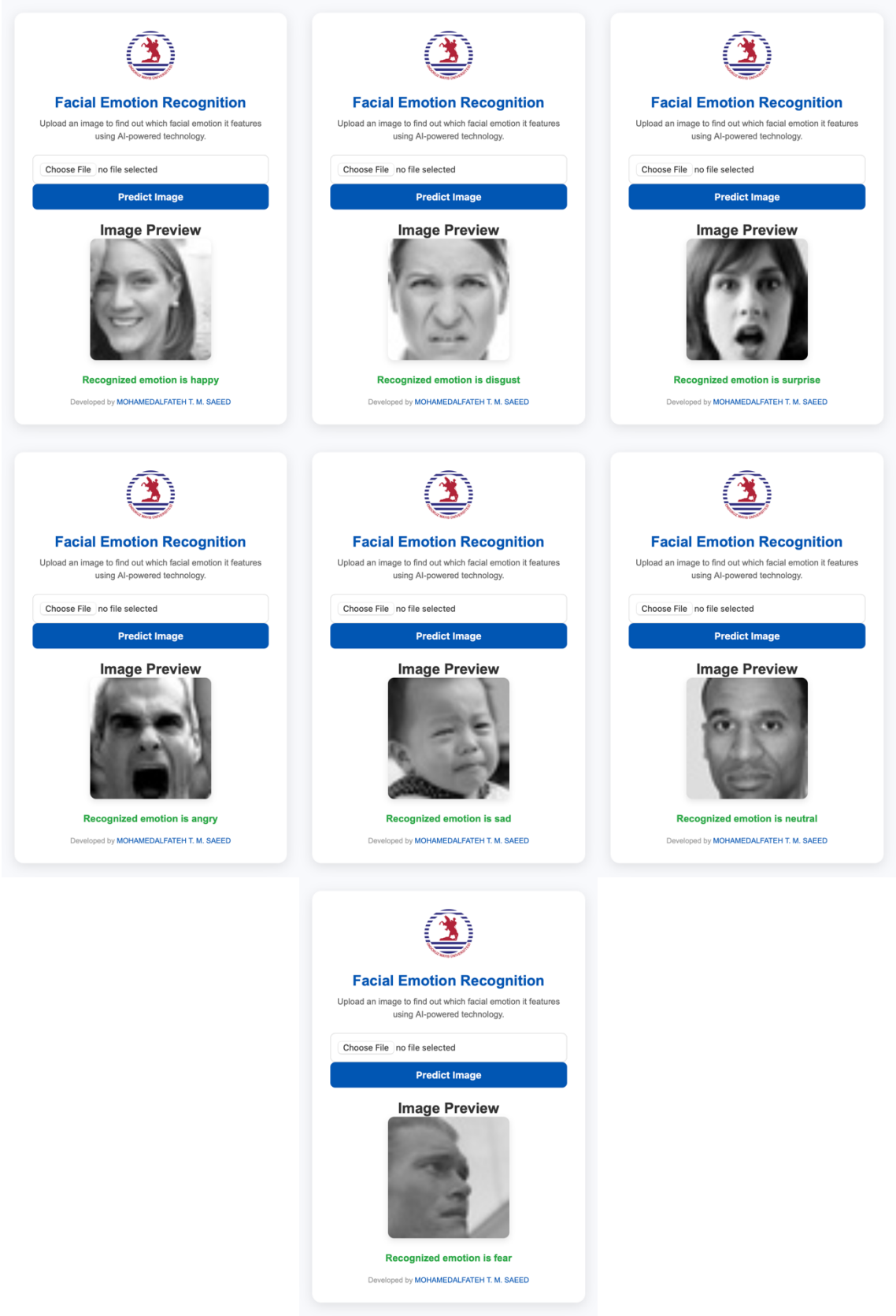

Figure 12: Deployment Test Using Samples from All Classes

7. References

[1] Minh-An Quinn, Grant Sivesind, Guilherme Reis, Real-time Emotion Recognition From Facial Expressions, Standford University, 2017.

[2] Corina Petean, Virginia Sandulescu, Ovidiu Bica, Emotion Detection from Face Images Using Deep Learning Techniques, Iasi: The 12th International Conference on E-Health and Bioengineering - EHB 2024, 2024.

[3] B. C. Ko, A Brief Review of Facial Emotion Recognition Based on Visual Information, Daegu: sensors, 2018.

[4] Jeniffer Xin-Ying Lek and Jason Teo, Academic Emotion Classification Using FER: A Systematic Review, Human Behavior and Emerging Technologies, 2023.

[5] Yuhang Zhang , Xiuqi Zheng, Chenyi Liang, Jiani Hu, and Weihong Deng, Generalizable Facial Expression Recognition, European Conference on Computer Vision - Springer, Cham., 2025.

[6] N. Mehendale, Facial emotion recognition using convolutional neural networks (FERC), SN Applied Sciences, 2020.

[7] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John S, Challenges in Representation Learning: A report on three machine learning contests, Neural information processing: 20th international conference, ICONIP 2013,, 2013.

[8] Christian Białek, Andrzej Matiola ´nski, and Michał Grega , An Efficient Approach to Face Emotion Recognition with Convolutional Neural Networks, Electronics, 2023.

[9] o. K. a. Z. Chen, Facial Emotion Recognition: State of the Art Performance on FER2013, Boston: arXiv preprint arXiv:2105.03588., 2021.

[10] S. L. M. D. L. B. A. G. A. A. L. MUSTAFA CAN GURSESLI, Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets, Florence: IEEE Access, 2024.