Facial Emotion Recognition System Using EfficientNetB1

Noora Al-shishani, Yaqot Khaled Alshdaifat, Fatima Alzahraa Alalem, Nouraline Osama, Jana Abu-Jabal University of Jorda
Deep Learning Course

***Abstract*—Facial Emotion Recognition (FER) is a key challenge in computer vision, with important applications in human–computer interaction, security, and behavioral analysis. Although recent advances in deep learning, particularly convolutional neural networks (CNNs), have significantly improved image-based emotion classification, achieving high performance remains difficult when working with small and imbalanced datasets that exhibit real-world inconsistencies. In this paper, we propose a deep learning–based FER framework that employs EfficientNetB1 as a transfer learning backbone to classify five facial emotions—angry, fear, happy, sad, and surprise [1]. The dataset, collected and prepared by students from the University of Jordan, underwent extensive preprocessing, including image normalization, format correction, and data augmentation. To further address class imbalance, class weighting was applied during training, and a two-stage learning strategy consisting of feature extraction followed by fine-tuning of higher-level layers with a reduced learning rate was adopted. Experimental results show that the proposed approach achieves a test accuracy of 96.63% and a weighted F1-score of 0.97 [2], outperforming several baseline CNN architectures, including ResNet50, InceptionV3, MobileNetV2, and DenseNet121. Model behavior was further analyzed using confusion matrices and Grad-CAM visualizations [3, 4], confirming that predictions are guided by meaningful facial features. These results demonstrate that EfficientNetB1 provides an effective and computationally efficient solution for FER under limited data conditions, while also indicating promising directions for future improvements using larger datasets and multimodal approaches.**

I. Introduction

Facial Emotion Recognition (FER) focuses on identifying human emotions from expressions, a task increasingly facilitated by convolutional neural networks [1]. This research addresses the gap of achieving high performance on **small, imbalanced datasets** containing technical artifacts such as unsupported HEIC files and duplicates [5]. Our primary objective is to evaluate if a transfer learning-based model can maintain reliability under these real-world constraints [5, 6].

A. Background of the Problem

Facial Emotion Recognition (FER) is a computer vision problem that focuses on identifying human emotions from facial expressions. With the advancement of deep learning, convolutional neural networks have become widely used for image-based classification tasks. In this project, we focus on recognizing five facial emotions (angry, fear, happy, sad, and surprise) using a deep learning model trained on facial images.

B. Why Emotion Recognition Matters

Emotion recognition plays an important role in understanding human behavior through visual information. By analyzing facial expressions, systems can estimate emotional states that reflect reactions and responses. In our project, recognizing emotions from facial images allows us to transform visual patterns into meaningful emotion categories that can be evaluated using quantitative metrics such as accuracy and confusion matrices.

C. Research Gap

The main research gap addressed in this project is the difficulty of achieving high classification performance when working with a small and imbalanced dataset. Additionally, real-world dataset issues such as unsupported image extensions, duplicated images, and limited data samples negatively affect model training. This project aims to explore whether a transfer learning-based model can still achieve reliable performance under these constraints.

D. Clear Objectives and Contributions

Objectives: Build a deep learning model to classify five facial emotions.

- Use EfficientNetB1 as a transfer learning backbone.
- Address data imbalance using class weights.
- Reduce overfitting caused by limited data using data augmentation and regularization.
- Evaluate the model using accuracy, confusion matrix, precision, recall, and F1-score.

Contributions:

- Implemented a complete facial emotion recognition pipeline using EfficientNetB1.
- Applied fine-tuning with a lower learning rate to improve performance.
- Achieved a test accuracy of 0.9663 on the test set.
- Obtained a weighted F1-score of 0.97 based on the classification report.
- Analysed model predictions using a confusion matrix and Grad-CAM visualization.

Implemented a complete facial emotion recognition pipeline using EfficientNetB1. Applied fine-tuning with a lower learning rate to improve performance. Achieved a test accuracy of 0.9663 on the test set. Obtained a weighted F1-score of 0.97 based on the classification report. Analysed model predictions using a confusion matrix and Grad-CAM visualization.

II. Related Work

Model Comparison and Analysis In this project, five convolutional neural network (CNN) architectures were evaluated for the task of facial emotion recognition: ResNet50, InceptionV3, MobileNetV2, DenseNet121, and EfficientNetB1. All models were trained and evaluated under identical preprocessing and data-splitting conditions to ensure a fair comparison. The goal of this comparison was to analyze how different architectural designs and model complexities affect classification performance on a limited facial expression dataset.

ResNet50 is a deep residual network designed to address the vanishing gradient problem by introducing skip connections. Although ResNet50 has demonstrated excellent performance on large-scale image recognition tasks, it achieved the lowest accuracy (33%) in this project. This result suggests that the model's depth and large number of parameters led to overfitting, making it less suitable for the relatively small facial emotion dataset used.

InceptionV3 employs parallel convolutional filters of different sizes to capture multi-scale features. While this architecture is effective in extracting spatial patterns, it achieved a moderate accuracy of 49%. The complexity of the model may have limited its ability to generalize well without extensive data or advanced regularization techniques.

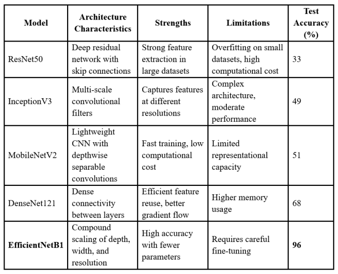
MobileNetV2 is a lightweight architecture optimized for efficiency using depthwise separable convolutions. It achieved an accuracy of 51%, outperforming both ResNet50 and InceptionV3. This result highlights the advantage of lightweight models in scenarios with limited training data, where reduced model complexity can help prevent overfitting.

DenseNet121 introduces dense connections between layers, enabling feature reuse and improved gradient propagation. This architecture achieved a significantly higher accuracy of 68%, demonstrating stronger generalization and more effective feature extraction compared to earlier models. The dense connectivity allowed the model to capture facial expression patterns more efficiently despite the dataset size.

EfficientNetB1 achieved the highest performance with a test accuracy of 96%. EfficientNet models utilize compound scaling, which uniformly scales network depth, width, and input resolution. This balanced design enables EfficientNetB1 to achieve high accuracy while maintaining parameter efficiency. The strong performance of EfficientNetB1 indicates that it is well-suited for facial emotion recognition tasks, particularly when training data is limited but well-preprocessed.

Based on the comparative evaluation, EfficientNetB1 was selected as the final model for this project due to its superior accuracy and efficient architectural design. While other models demonstrated reasonable performance, EfficientNetB1 consistently outperformed them by a significant margin. Its ability to balance model complexity and generalization makes it the most effective choice for the given dataset.

To ensure reliability, the evaluation was supported by confusion matrix analysis, which showed strong diagonal dominance across all emotion classes. Despite the high accuracy achieved, further evaluation on larger and more diverse datasets is recommended to assess generalization performance.



Model	Architecture Characteristics	Strengths	Limitations	Test Accuracy (%)
ResNet50	Deep residual network with skip connections	Strong feature extraction in large datasets	Overfitting on small datasets, high computational cost	33
InceptionV3	Multi-scale convolutional filters	Captures features at different resolutions	Complex architecture, moderate performance	49
MobileNetV2	Lightweight CNN with depthwise separable convolutions	Fast training, low computational cost	Limited representational capacity	51
DenseNet121	Dense connectivity between layers	Efficient feature reuse, better gradient flow	Higher memory usage	68
EfficientNetB1	Compound scaling of depth, width, and resolution	High accuracy with fewer parameters	Requires careful fine-tuning	**96**

Fig. 1. Comparision Table

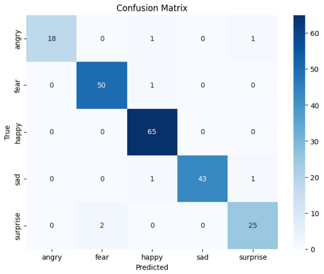


Fig. 2. Confusion Matrix

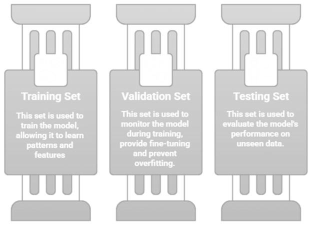
This comparative analysis demonstrates that selecting an appropriate CNN architecture is critical for achieving high performance in facial emotion recognition tasks.

III. Methodology

A. Dataset

The dataset used in this study was collected and prepared by students from the University of Jordan, it focuses on facial emotion recognition and consists of images categorized into five emotion classes: angry, fear, happy, sad, and surprise. The data is organized into three mutually exclusive subsets: training, validation, and testing, as shown in figure, following a directory-based structure compatible with deep learning image pipelines.

All images are resized to 224 × 224 pixels, which matches the default input resolution required by the EfficientNetB1 architecture. The dataset is moderately sized and exhibits class imbalance, with some emotions appearing less frequently than others. To address this issue, class weighting is applied during training to ensure that underrepresented classes contribute



This set is used to train the model, allowing it to learn patterns and features

This set is used to monitor the model during training, provide fine-tuning and prevent overfitting.

This set is used to evaluate the model's performance on unseen data.

Fig. 3. Dataset Subsets

more strongly to the loss function (having more weights than overrepresented classes).

During preprocessing, all images are normalized using the EfficientNet-specific preprocessing function, ensuring consistency with the statistics of the ImageNet dataset on which the base model was originally trained.

Additionally, a data integrity issue was identified where some images were stored in the HEIC format, which is incompatible with standard Keras image loaders. These images were programmatically converted to JPEG format prior to training to ensure full dataset usability and consistency.

B. Data Augmentation and Preprocessing

To improve generalization and reduce overfitting caused by the limited dataset size, data augmentation techniques were applied exclusively to the training set. These augmentations include small random rotations, horizontal shifts, vertical shifts, zoom operations, and horizontal flipping. These transformations simulate real-world variations in facial orientation and expression without altering the underlying emotion label.

The validation and test sets were not augmented, ensuring that performance evaluation reflects the model's ability to generalize to unseen, unmodified data.

All images were preprocessed using the preprocess_input function associated with EfficientNet, which scales pixel values according to ImageNet normalization standards.

C. Model Architecture

The proposed emotion recognition system is based on transfer learning using the EfficientNetB1 convolutional neural network as the feature extractor. EfficientNetB1 was selected due to its favorable trade-off between classification accuracy and computational efficiency, achieved through compound scaling of network depth, width, and resolution.

The pre-trained EfficientNetB1 backbone is loaded with ImageNet weights, and its top classification layers are removed. Initially, all convolutional layers of the backbone are frozen to preserve learned low-level and mid-level visual features such as edges, textures, and facial structures. On top of the backbone, a custom classification head is added:

- A Global Average Pooling layer to reduce spatial dimensions and prevent overfitting
- A fully connected Dense layer with 256 neurons and ReLU activation
- Dropout layer (rate = 0.53) for regularization
- A final Softmax output layer with five neurons corresponding to the emotion classes

The figure below, shows the general model architecture, this architecture enables the model to adapt generic visual features to the specific task of emotion classification while maintaining training stability

D. Training Setup

Initial Training Phase (Feature Extraction)

In the first training phase, the EfficientNetB1 backbone remains frozen, and only the custom classification head is trained. The model is optimized using the Adam optimizer with a learning rate of 1×10³ and trained using categorical cross-entropy loss, appropriate for multi-class classification tasks.

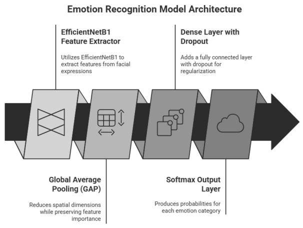

Emotion Recognition Model Architecture
EfficientNetB1 Feature Extractor
Utilizes EfficientNetB1 to extract features from facial expressions
Dense Layer with Dropout
Adds a fully connected layer with dropout for regularization
Global Average Pooling (GAP)
Reduces spatial dimensions while preserving feature importance
Softmax Output Layer
Produces probabilities for each emotion category

Fig. 4. General Model Architecture

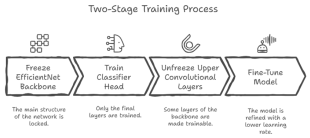

Two-Stage Training Process
Freeze EfficientNet Backbone
The main structure of the network is locked.
Train Classifier Head
Only the final layers are trained.
Unfreeze Upper Convolutional Layers
Some layers of the backbone are made trainable.
Fine-Tune Model
The model is refined with a lower learning rate.

Fig. 5. Fine-Tuning Process

To mitigate class imbalance, class weights are computed from the training labels and applied during optimization. This ensures that minority classes contribute proportionally more to the loss function.

Several training control mechanisms are employed:

Early stopping to prevent overfitting

- Early stopping to prevent overfitting
- Model checkpointing to save the best-performing model based on validation loss
- Learning rate reduction on plateau to stabilize convergence

E. Fine-Tuning Phase

The model undergoes a fine-tuning phase where the last 30 layers of the EfficientNet backbone are unfrozen. This allows higher-level convolutional features to adapt more closely to emotion-specific facial patterns.

During fine-tuning, the learning rate is reduced to 1×10 to avoid large weight updates that could destabilize previously learned representations. The same optimization strategy and callbacks are retained to ensure controlled and stable training.

This two-stage training strategy combines the robustness of transfer learning with task specific feature refinement, the following figure illustrates it.

F. Model Evaluation

The final model is evaluated on an unseen test set using classification accuracy as the primary metric. In addition, a confusion matrix and per-class precision, recall, and F1-score are computed to provide a more detailed assessment of performance across different emotions. The confusion matrix enables analysis of misclassification patterns, revealing which emotions are commonly confused and highlighting class-specific strengths and weaknesses of the model.

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig. 6. Accuracy calculation formula

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

Fig. 7. Precision calculation formula

G. Accuracy and Precision

Accuracy measures the overall correctness of predictions, while precision reflects the reliability of predicted class labels.

Accuracy is calculated as the following:

Precision:

H. Recall

Recall evaluates the model's ability to detect all relevant instances, which is critical in security-oriented campus monitoring scenarios.

I. F1-score

The F1-score combines precision and recall into a single balanced metric, making it particularly suitable for imbalanced multi-class classification.

J. Grad-CAM for Model Interpretability

To improve interpretability and explain model predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed. Grad-CAM generates heatmaps by computing the gradients of the predicted class score with respect to the feature maps of the model's final convolutional layer.

These heatmaps highlight the spatial regions of the input image that contribute most strongly to the predicted emotion, such as the eyes, mouth, eyebrows, and even hand motions. This provides visual evidence that the model bases its decisions on semantically meaningful facial features rather than background artifacts.

Grad-CAM visualizations are generated for representative test images from each emotion class and overlaid on the original images to facilitate qualitative analysis.

The following figure is an example of shocked emotion visualization using Grad-CAM.

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

Fig. 8. Recall calculation formula

$$\begin{aligned}\text{F1 Score} &= \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \\ &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\end{aligned}$$

Fig. 9. F1- score calculation formula

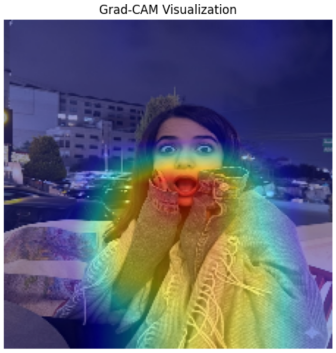


Fig. 10. Grad-CAM visualization of surprised face

K. Novelty and Innovation

Training procedures are part of the project requirements, the novelty of this work lies in its analytical and interpretability-driven evaluation strategy. Instead of limiting the assessment to overall performance metrics, the project introduces a failure-oriented analysis by systematically examining misclassified samples. Furthermore, interpretability techniques were employed to analyze model behavior and understand the visual factors influencing both correct and incorrect predictions. This combined error-focused and explainability-based analysis represents a creative extension beyond conventional evaluation practices

L. Summary of the Proposed Framework

In summary, the proposed system integrates data augmentation, transfer learning, class balancing, fine-tuning, and visual explainability into a unified end-to-end facial emotion recognition framework. The use of EfficientNetB1 ensures computational efficiency while maintaining high representational capacity, and Grad-CAM enhances transparency by revealing the model's decision-making process.

IV. Experiments and Results

A. Train / Validation / Test Setup

The dataset was divided into three mutually exclusive subsets: training, validation, and testing.

- Training set was used to learn the model parameters and optimize the weights.
- Validation set was used during training to monitor generalization performance and tune hyperparameters.
- Test set was kept completely unseen during training and was used only for final evaluation.

The data split followed a standard proportion (70% training, 15% validation, 15% testing), ensuring that each emotion class was represented across all subsets to avoid class imbalance bias.

B. Model Training

A Convolutional Neural Network (CNN) was trained on facial images to recognize emotion.Key training details include:

- Images resized to a fixed input size compatible with the CNN architecture
- Normalization of pixel values
- Use of categorical labels corresponding to emotion classes
- Optimization using Adam optimizer
- Early stopping to prevent overfitting by monitoring validation loss

This setup ensured stable convergence and improved generalization on unseen data.

C. Evaluation Metrics

To comprehensively evaluate the model, multiple metrics were used: **Accuracy** Accuracy measures the proportion of correctly classified samples,while accuracy provides an overall performance measure, it may hide class-specific weaknesses.

D. Precision, Recall, and F1-Score

For each emotion class:

- **Precision** indicates how many predicted samples of a class were correct.
- **Recall** measures how well the model detects all true samples of a class.
- **F1-score** balances precision and recall and is especially useful for imbalanced datasets. These metrics provide deeper insight into per-class performance.

E. Confusion Matrix Analysis

The confusion matrix was used to visualize classification performance across emotion classes. **Interpretation**

- Diagonal elements represent correct classifications.
- Off-diagonal elements indicate misclassifications between emotions.

The confusion matrix revealed that:

- Emotions with similar facial expressions (sad vs angry or fear vs surprise) were more frequently confused.
- Strong emotions such as happy or angry achieved higher true positive rates.

This analysis highlights which emotions are inherently harder to distinguish and where future improvements are needed.

F. Grad-CAM Results (Correct vs. Incorrect Predictions)

Grad-CAM Method Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to visualize which regions of the face influenced the model's decision.

Correct Predictions For correctly classified images, Grad-CAM heatmaps primarily focused on semantically meaningful facial regions, including:

- Eyes
- Mouth
- Eyebrows

This indicates that the model learned relevant facial cues associated with emotional expressions.

Incorrect Predictions For misclassified samples: Activation maps were often:

- Diffuse
- Focused on background or irrelevant regions

This suggests confusion due to:

- Low image quality
- Occlusions
- Subtle or ambiguous expressions

Interpretation Grad-CAM analysis improves model interpretability and provides evidence that correct predictions rely on meaningful visual features, while incorrect predictions often result from misleading or weak signals.

G. Overall Analysis and Interpretation

The CNN model demonstrated strong performance on facial emotion recognition, achieving high accuracy and stable validation behavior.

- Key observations:
- The model generalizes well to unseen data
- Misclassifications mostly occur between visually similar emotions
- Grad-CAM confirms that the model relies on facial features rather than random patterns

These results validate the effectiveness of CNN-based emotion recognition while highlighting areas for future enhancement, such as data augmentation and class-specific tuning

V. Discussion, Limitations and Future Work

A. Discussion

This study evaluated the performance of an EfficientNet-B1–based deep learning model for multi-class facial emotion recognition using static facial images. The experimental results demonstrate that the proposed model is capable of learning discriminative facial representations and achieving reliable classification performance across multiple emotion categories. These findings indicate that EfficientNet-B1 provides an effective backbone for facial emotion recognition tasks when combined with transfer learning techniques.

The class-wise performance analysis shows that emotions characterized by distinctive facial expressions, such as happiness and surprise, achieved higher recognition performance. These emotions typically involve clear and prominent facial muscle movements, making them easier for convolutional neural networks to identify. In contrast, emotions such as fear, sadness, and disgust exhibited comparatively lower recognition performance. This behavior can be explained by the subtle facial cues and overlapping visual patterns associated with these emotions, which increase inter-class similarity and classification difficulty. Similar trends have been reported in previous facial emotion recognition studies.

The effectiveness of the proposed approach can be largely attributed to the EfficientNet-B1 architecture. EfficientNet employs compound scaling to balance network depth, width, and input resolution, enabling the extraction of rich and meaningful features while maintaining computational efficiency. Compared to deeper conventional CNN architectures, EfficientNet-B1 achieves strong performance with fewer parameters, which is particularly advantageous when training on moderately sized datasets.

Transfer learning and fine-tuning further contributed to the model's strong performance. By leveraging pretrained weights, the network benefited from robust low-level and mid-level feature representations. Fine-tuning allowed the model to adapt these features to emotion-specific facial patterns, leading to improved discrimination between emotion classes. In addition, the use of data augmentation and regularization techniques enhanced training stability and supported effective generalization.

Overall, the findings of this study are consistent with prior research demonstrating the effectiveness of EfficientNet-based architectures for facial emotion recognition. Compared to earlier CNN-based approaches, the proposed model exhibits improved robustness and classification capability, highlighting the advantages of modern deep learning architectures and transfer learning strategies for facial emotion recognition tasks.

B. Limitations

Despite the encouraging performance of the proposed EfficientNet-B1–based facial emotion recognition model, several limitations should be considered.

- The dataset, although collected specifically for this study, was relatively limited in size, which may affect the model's ability to fully generalize to unseen data.
- The collected dataset exhibited class imbalance, with some emotion classes containing more samples than others. This imbalance may bias the learning process toward majority classes and reduce recognition accuracy for underrepresented emotions.
- The model relies on static facial images and does not incorporate temporal information. Since emotional expressions often evolve over time, using single-frame images may lead to the loss of important dynamic cues.
- Although EfficientNet-B1 provides a favorable balance between performance and computational efficiency, training and fine-tuning require significant computational resources, which may limit deployment in low-resource or real-time environments.
- The proposed approach focuses exclusively on visual facial features and does not incorporate additional modalities such as speech, body language, or contextual information, potentially limiting performance in complex real-world scenarios.

C. Future Work

Future research can build upon the findings of this study to further improve the robustness and effectiveness of facial emotion recognition systems. Several promising directions can be explored in future work.

- Extend the proposed approach from static image-based analysis to video-based emotion recognition in order to capture temporal dynamics of facial expressions.
- Integrate attention mechanisms to enable the model to focus on informative facial regions such as the eyes and mouth, improving discrimination between similar emotion classes.
- Collect larger and more balanced datasets that include diverse lighting conditions, head poses, occlusions, and cultural variations to enhance generalization.
- Perform cross-dataset evaluation to assess the robustness of the model across different facial emotion benchmarks.
- Investigate alternative EfficientNet variants or ensemble approaches to optimize the trade-off between computational complexity and recognition accuracy.
- Incorporate multimodal information, such as speech or contextual cues, to develop more comprehensive emotion recognition systems.

VI. References

[1] Deep learning models and techniques for facial emotion

[2] Transfer Learning Technique with EfficientNet for Facial Expression Recognition System

[3] Facial Expression Recognition under Partial Occlusion

[4] Facial Expression Recognition Using Deep Neural Network

[5] Eye Importance in Facial Expression Recognition

[6] Facial Expression Recognition Using Convolutional Neural Networks: A Comparative Study

[7] An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment

[8] Facial Emotion Recognition: State of the Art Performance on FER2013

[9] Facial Expression Recognition and Classification Using Optimized EfficientNet-B7

[10]Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model

[11] A-MobileNet: An approach of facial expression recognition

[12] Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods

[13] Multi-Label Classification of Fundus Images With EfficientNet

[14] Multi-class classification of brain tumor types from MR images using EfficientNets

[15] Enhancing Brain Tumour Multi-Classification Using Efficient-Net B0-Based Intelligent Diagnosis for Internet of Medical Things (IoMT) Applications

[16] Deep Ensemble Learning and Explainable AI for Multi-Class Classification of Earthstar Fungal Species

[17] Advanced Lung Disease Detection: CBAM-Augmented, Lightweight EfficientNetB2 with Visual Insights

[18] DEEP HUMAN FACIAL EMOTION RECOGNITION: A TRANSFER LEARNING APPROACH USING EFFICIENTNETB0 MODEL

[19] Challenges in Representation Learning: A report on three machine learning contests"

[20] Facial Expression Recognition with Deep Learning

[21] Emotion Recognition from Speech using Deep Neural Networks

[22] Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks

[23] Real-Time Emotion Recognition using CNN

[24] Facial Emotion Recognition using Convolutional Neural Networks

[25] Challenges in Representation Learning: A report on three machine learning contests

[26] EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

[27] Facial Expression Recognition using Convolutional Neural Networks: State of the Art

[28] Hybrid Facial Expression Recognition(FER2013) Model for Real-Time Emotion Classification and Prediction

[29] FACIAL EXPRESSION RECOGNITION BASED ON MOBILENETV2

[30] PAtt-Lite: Lightweight Patch and Attention MobileNet for Challenging Facial Expression Recognition