TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 DOI:
Jurayev U.S., Kurbanov A.A. № 3(22)2025

UDC 004.896

REAL-TIME ASSESSMENT OF HUMAN EMOTIONAL STATE FROM FACIAL IMAGES USING THE EFFICIENTNET MODEL AND THE FACEEMOCDS DATASET

Jurayev Umidjon Sayfullayevich
Associate Professor, Gulistan State University

Kurbanov Abdurahmon Alishboyevich *Doctoral student at the Jizzakh branch of theNati onal University of Uzbekistan named after Mirzo Ulugbek*

Abstract. The detection of emotional states from human facial expressions is a crucial research area in artificial intelligence and computer vision. This paper explores real-time emotion recognition using the EfficientNet-B0 model on the FaceEmocDS dataset. The FaceEmocDS dataset comprises 72,412 high-quality images, encompassing eight emotion classes: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise. Derived from multiple open-source datasets, it ensures diversity and balance after rigorous cleaning to remove low-quality or redundant samples.

The model was trained via transfer learning, leveraging pre-trained weights from ImageNet, and fine-tuned over 30 epochs on the PyTorch framework. It achieved 74.32% accuracy on the validation set. To enhance generalization, data augmentation techniques were applied, including random rotations (up to 10 degrees), color jittering (adjusting brightness, contrast, and saturation), random resized cropping, horizontal flipping, and random erasing (with probability 0.7). Class weights were incorporated into the CrossEntropyLoss function to address imbalances, particularly in underrepresented classes like contempt and disgust.

For real-time implementation, MediaPipe Face Detection was integrated for

efficient face localization, followed by emotion classification with the EfficientNet-B0 model. The system operates at 20-30 frames per second (FPS) on standard hardware, making it suitable for dynamic environments. Test results yielded an accuracy of 73.18% and a weighted F1-score of 0.7327, with highest performance on happiness (93.08%) and lowest on contempt (63.05%). Confusion matrix analysis revealed common misclassifications, such as contempt with anger or disgust, highlighting areas for improvement. Compared to baseline models such as VGG and ResNet, this approach provides superior efficiency in terms of parameter count and inference speed while maintaining competitive accuracy.

Keywords*:EfficientNet-B0, aceEmocDS, MediaPipe Face Detection, PyTorch, Transfer learning, CNN, Real-time, facial emotion.*

Introduction

Human facial expressions serve as one of the most immediate and intuitive visual cues for conveying emotions. The subtle movements of the eyes, eyebrows, lips, and other facial muscles reflect internal psychological states, acting as essential signals in interpersonal communication. These expressions not only facilitate social interactions but also provide insights into an individual's mental well-being, intentions, and responses to external stimuli. Recognizing emotions from facial images, known as Facial Emotion Recognition (FER), has emerged as a pivotal area in artificial intelligence (AI) and computer vision, establishing a link between human psychology and machine understanding. This capability enables machines to interpret human emotions in real-time, fostering more empathetic and responsive systems across various domains.

The roots of FER can be traced back to evolutionary biology and psychology. In 1872, Charles Darwin published "The Expression of the Emotions in Man and Animals," positing that facial expressions are innate and universal across cultures and species, serving adaptive functions for survival. Darwin's observations laid the groundwork for understanding emotions as biologically driven phenomena [1]. Building on this, psychologist Paul Ekman in the 1970s and 1980s developed the Facial Action Coding System (FACS), a comprehensive

framework that categorizes facial movements into action units (AUs) based on muscle contractions [2]. Ekman's research identified six basic emotions – anger, disgust, fear, happiness, sadness, and surprise – that are universally recognizable through facial cues, with contempt sometimes included as a seventh. FACS has since become a cornerstone for FER algorithms, providing a standardized way to decode expressions. However, Ekman's model also acknowledges cultural variations and contextual influences, emphasizing that while core expressions are innate, their interpretation can vary [3].

The technological evolution of FER began in the 1960s with early facial recognition experiments by Woodrow Bledsoe and his team, who manually processed images due to limited computational power. These efforts focused on identity recognition rather than emotions, but they highlighted the challenges of image analysis [4]. By the 1990s, advancements in machine learning automated FER processes. Initial models relied on simple linear algorithms and hand-crafted features like Histogram of Oriented Gradients (HOG) or Local Binary Patterns (LBP). The advent of Convolutional Neural Networks (CNNs) in the 2010s revolutionized the field by enabling automatic feature extraction through hierarchical layers – convolutional for detecting edges and textures, pooling for dimensionality reduction, and fully connected for classification [5]. CNNs offer robustness to variations in pose, lighting, and scale, making them ideal for complex image tasks.

Despite these progresses, effective FER requires vast, diverse datasets. Early datasets like CK+ (Cohn-Kanade) and JAFFE (Japanese Female Facial Expression) were limited to controlled laboratory settings, with small sample sizes (e.g., CK+ has 981 images) and imbalances across emotion classes. Larger datasets such as FER2013 (35,887 images) and AffectNet (over 1 million) introduced "in-the-wild" scenarios but suffered from annotation noise, class imbalances (e.g., happiness overrepresented), and limited ethnic diversity. These limitations hinder model generalization, leading to biases in real-world applications. To address this, the **FaceEmocDS** dataset was curated by merging multiple open-source datasets, including AffectNet (41,553 samples), RAF-DB (15,339 samples), ExpW (91,778 samples), FER-2013 (35,886 samples), CK+ (981 samples), JAFFE (213 samples), TFEID (336 samples), and others, totaling 277,923 images initially. After rigorous cleaning to eliminate low-quality, redundant, or ambiguous samples, it retains 72,412 high-quality images across eight emotion classes: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise [6]. This dataset ensures better class balance, diversity in age, gender, ethnicity, lighting, and poses, making it suitable for robust training.

Recent advancements in deep learning have propelled FER forward, particularly from 2023 to 2025. Surveys highlight the shift toward hybrid models combining CNNs with attention mechanisms and transformers for better feature focus on key facial regions like eyes and mouth. For instance, a 2024 study introduced improved deep learning models to mitigate ambient noise effects, achieving higher accuracy in noisy environments. Real-time FER using models like ResNet-50 and Dynamic Time-Spatial CNN (DTSCNN) has shown superior performance, with CNNs outperforming in speed and accuracy for live video feeds. Multimodal approaches, integrating facial with body cues, have advanced emotion classification, leveraging neural networks for realistic expression synthesis. Comprehensive reviews emphasize challenges like compound emotions, cross-domain inconsistencies, and ethical concerns, proposing self-supervised learning and domain adaptation as solutions. In 2025, surveys on FER using DL methods stress the importance of generalization across datasets, with techniques like generative augmentation via StyleGAN improving bias mitigation.

Model selection is critical for efficient FER. Traditional CNNs like VGG and ResNet are effective but parameter-heavy, limiting real-time deployment on edge devices. EfficientNet, introduced by Google in 2019, addresses this through compound scaling—balancing depth, width, and resolution for optimal performance with fewer parameters. EfficientNet-B0, the lightest variant, excels in transfer learning, where pre-trained ImageNet weights are fine-tuned for specific tasks like FER [7]. This approach reduces training time and enhances accuracy on smaller datasets. In this study, EfficientNet-B0 is employed on FaceEmocDS, incorporating data augmentation (rotations, flips, color jitter) and class weights to handle imbalances.

Challenges in FER include illumination variations, pose changes, occlusions (e.g., masks), and cultural differences. Real-time systems must

achieve high frames per second (FPS) while maintaining accuracy. MediaPipe Face Detection offers fast, accurate localization, complementing deep models for end-to-end pipelines. Ethical considerations, such as privacy in surveillance and bias in datasets, are increasingly addressed in recent literature, advocating for diverse training data and fairness metrics.

This paper investigates real-time FER using EfficientNet-B0 on FaceEmocDS, achieving competitive results. By overcoming dataset limitations and leveraging advanced DL, it contributes to practical AI applications, paving the way for future multimodal and unbiased systems.

Literature review

A. Akbar et al. proposed to classify the obtained representations with an L2-regularized SVM using EfficientNet-B0 as a feature extractor [8]. The main idea of the authors is to achieve high accuracy in small or resource-constrained environments using high-level representations obtained from pre-trained EfficientNet, without requiring extensive fine-tuning and extended augmentation. Experimental results demonstrate high performance on classic benchmarks such as JAFFE and CK+, with accuracies of 99–100% in some cases. However, one should be careful when interpreting such results: datasets such as JAFFE and CK+ were obtained in laboratory conditions and do not fully reflect real-world (in-the-wild) diversity, so the high results may be limited in terms of generalizability [8]. The advantage of the SVM classifier is interpretability and simplicity, but it is limited in terms of flexibility compared to end-to-end neural fine-tuning on large classes and large data sets. The practical aspect of the work is that the combination of EfficientNet representations + lightweight SVM can be an effective solution for resource-constrained edge devices or rapid prototypes.

In a paper published by Akhmedov Farkhod et al., a landmark-based graphical approach is proposed for emotion recognition in masked faces during the COVID-19 pandemic. In the experimental process, high-quality landmark detectors such as MediaPipe were used to extract points from the upper face (eyes, eyebrows), construct a graph based on their coordinates, and then classify them with a lightweight model [9]. The strength of this approach is the ability to detect emotions even in the absence of signal from the lower face (mouth) covered by the mask; in practical tests, an overall accuracy of ~91.2% was demonstrated for 7 emotions based on images. In real-time webcam tests, a slight decrease was observed compared to the image, which indicates a dependence on the accuracy and noise of the landmark detector in the pipeline. Landmark-based approaches may be less computationally intensive and easier to interpret than convolutional image-based methods, but may be limited in capturing micro-expressions or signals specific to the lower face (mouth). The proposed method demonstrates important and practical solutions for FER systems suitable for real-world, masked environments, but it is suggested to improve the stability of landmark accuracy and consider multimodal (mouth, voice) integration.

Amjad Rehman et al. have published a general review and survey-type article, which provides a broad analysis of machine learning, deep learning, and hybrid methods for emotion recognition from facial images, as well as how to preprocess, augment, and process temporal features [10]. The authors highlight the strengths and weaknesses of deep networks (CNN, 3D-CNN, RNN/LSTM, transformer), and provide a general description of practical benchmarks and evaluation criteria (accuracy, F1, and AUC). This type of work provides a systematic overview of the field for newcomers and methodological guidelines for experienced researchers. In particular, the article discusses the importance of temporal context (the relationship between video frames) and step-by-step strategies for solving real-world problems such as class mismatch and domain shift. This resource provides a the oretical framework for the this research and a comprehensive context for comparison.

"A Survey on Facial Expression Recognition of Static and Dynamic Emotions", SFER and DFER (static and dynamic FER) are analyzed in depth, dividing them into model-oriented and challenge-oriented categories. The authors clearly indicate eight SFER and seven DFER problems, for example, expression disturbance, compound emotions, cross-domain inconsistency, key frame sampling, expression intensity [11]. For each problem, existing approaches (attention mechanisms, temporal sampling, self-supervised learning, domain adaptation) and their limitations are discussed. The advantage of the work is that it is a copy that combines a wide range of recent techniques, proposes promising directions (self-supervision, multimodal fusion, fairness/ethics), and

provides a brigade of directions for new research. However, as a preprint, some recommendations have not yet been verified by peer review; nevertheless, the article is methodologically sound and the bibliographic base is also extensive.

In a research paper titled "Convolutional Neural Network Based Facial Expression Recognition Using Image Filtering Techniques" by Ritanshi Agarwal et al., it was shown that using a simple CNN architecture and classical filters (Gaussian, LoG) on the CK+ dataset, high performance was achieved [12]. The authors chose a strategy to improve the image quality by pre-filtering the signal (noise reduction) using a small and simplified network, thereby increasing the accuracy of the model. Their experience has been reported to give 100% accuracy results for CK+ in laboratory conditions, but it is typical that such results are difficult to replicate on larger and in-the-wild datasets. The practical conclusion of this work is that preprocessing (filtering) helps significantly on small, controlled datasets, but in real-world conditions, the role of robustness, augmentation and diverse training examples becomes greater. It is also worth noting that simple architectures can sometimes perform well on strict tasks, but more cross-dataset validation is needed to draw a final conclusion.

The article by Sarthak Aggarwal et al. provides an overview of the field of FER and its application areas (psychology, marketing, HCI, healthcare) and describes techniques for classifying emotions by analyzing facial features. It provides a simple explanation and practical application of the technology: indicators such as eyebrow movement, eye widening, lip curvature, and how they can be detected using algorithmic tools. The strengths are that it emphasizes the practical and multidisciplinary context, making FER understandable to non-tech audiences; the weaknesses are the lack of scientific experimental details and rigorous benchmarks. The article provides a good conceptual overview of the research, process, and practice, which may help to enrich the introduction and motivation sections of this study [13].

In his paper "The Study of Performance Related to Classical Convolutional Neural Networks in the Field of Facial Emotion Recognition," Yanxiao Liu compares architectures such as VGG, ResNet, MobileNet, and DenseNet for facial expression recognition. The authors show the trade-off between parameter efficiency, computational complexity, and accuracy of each architecture, which plays an important role in practical selection. Research shows that the highest accuracy does not always belong to the largest architecture; sometimes lightweight architectures (MobileNet) meet real-time requirements and provide sufficient accuracy. The results of the work emphasize the need to match the choice of architecture with the task requirements and consider the possibilities of applying the model to edge devices. The article also provides practical recommendations on training methods, optimizer options, and data augmentation strategies [14].

A paper by Akhand M.A. et al. provides a detailed description of how to adapt very deep pre-trained DCNN models for FER using a transfer learning approach [15]. The authors demonstrate the effectiveness of gradually optimizing the model by training the dense part first, then fine-tuning the blocks one by one. In experiments, high gains were achieved on the KDEF and JAFFE datasets using models such as VGG, ResNet series, DenseNet, and Inception; DenseNet-161 showed high performance on KDEF and JAFFE. The main conclusion is that fine-tuning the layers step by step improves the generalization of the model on small-to-medium-sized datasets. This approach provides a direct practical basis for your training strategy. However, it is important to extend these results to in-the-wild datasets and evaluate them in the context of real-time inference.

This paper describes a case study using EfficientNetB3 as a feature extractor and a classification result with GAP + FCL + Softmax. The results show 94.44% accuracy and 0.94 F1 score. The authors test the compound scaling and parameter-FLOP performance of the EfficientNet family on the FER task and find that EfficientNet shows a good balance compared to traditional deep models [16]. In particular, EfficientNetB3 is reported to give the best result in this case, where the model distinguishes between positive and negative emotions well. However, the authors emphasize the need to re-examine the compound scaling formulations for a specific domain (e.g., teacher faces). This work demonstrates the effectiveness of EfficientNet for FER and provides empirical evidence supporting your approach based on EfficientNet-B0.

This paper by T. Kusunose et al. demonstrates the generation of class-appropriate augmentation

based on StyleGAN2 and then emotion recognition using transfer learning with VGG16. The strength of the method is that it can increase the dataset size using synthetically generated data and strengthen the model against racial or other biases; experiments show a 75.10% accuracy on the CFEE dataset without augmentation and an 82.04% accuracy with StyleGAN2 [17]. This result confirms the practical utility of generative augmentation, especially on limited and low-variance datasets. However, the quality and realism of the generative samples are also important — low-quality synthesized images can force the model to learn the wrong signal. There may also be ethical and licensing issues with generative augmentation.

Agung, ES et al. 's "Image-based facial emotion recognition using convolutional neural network on emognition dataset" compares transfer learning models such as Inception-V3 and MobileNet-V2, and approaches to building an architecture from scratch using the Taguchi method using the Emognition dataset [18]. The paper reports its results on a large dataset of 10 emotions and shows a high accuracy of 96%. The strength of this work is the use of a large target emotion catalog and the use of non-traditional methods such as Taguchi hyperparameter optimization to stabilize the model through rigorous preprocessing + augmentation processes. Such approaches play an important role in working with real-world, large emotion clusters. The rigorous review and extensive statistical analysis of the Nature articles serve as a reliable benchmark for evaluating your results.

A research paper published by U. Rawat and CS Rai provides empirical evidence on how deep models (Inception-v3, VGG-19, ResNet-152, ResNet-50, DenseNet-161) improve in FER through transfer learning. The authors show improvements in recall, precision, and F1 parameters of each model on datasets such as JAFFE and KDEF; models such as DenseNet-161 are noted to have significantly improved over JAFFE. This paper demonstrates the practical implications of transfer learning and helps to determine which backbones are more effective on small or medium datasets. The paper also helps to identify which emotions are confused using per-class analysis and confusion matrices [19]. This type of detail is important when interpreting FaceEmocDS results and identifying model weaknesses.

The paper by WA Khan et al. is a work that covers deep learning and transfer learning practices and compares how new architectures (Xception, NASNet, EfficientNet) perform on FER. The authors note that Xception and EfficientNet models outperform classical models by showing training accuracies of around 87–89%, but that overall class accuracies still remain around 64% – a warning about class-level confusion and real-world difficulties. The conclusion of the work is that while new architectures have high potential, preprocessing, balancing, and domain adaptation strategies that lead them to real-world generalization are also important. This is important in understanding the limits of source model selection and generalization [20].

K. Desai et al., in their paper "A Transfer Learning Framework for Facial Emotion Recognition: Leveraging Pre-Trained Convolutional Neural Networks", present a transfer learning approach (ResNet50, DenseNet121, EfficientNetB5, Xception) for FER and report that EfficientNetB5 performs best (training 86.23%, test 84.39%). This result confirms that the high-order variants of the EfficientNet family perform well on small-to-medium FER datasets, but the high-order variants are computationally more demanding. The paper also provides practical advice on per-class analysis and model configuration [21]. This resource will be useful in comparing your choice of EfficientNet-B0 with high-order variants: the lightness of B0 and the potential accuracy advantage of variants such as B5/B3 can be analyzed.

In the research paper "Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning Resnet-50" by Istiqomah, AA et al., ResNet-50 shows high performance (around 99%) on transfer learning and preprocessed low-resolution (48×48 grayscale) images [22]. This result shows that transfer learning can be very effective under fixed and uniform preprocessing (grayscale, resolution) conditions. However, such high performance may also depend on the dataset size, image quality, and laboratory conditions; there is a possibility that the results will degrade in real-world conditions. This work provides evidence for the applicability of transfer learning strategies to small or low-resolution data, but generalization needs to be verified.

Slimi S. et al., in their paper "Transfer learning-based facial emotion detection for enhanced driver safety: a hybrid approach", propose a hybrid architecture based on ResNet50 and show high

performance on the Keimyung University driver emotion dataset [23]. This work demonstrates the effectiveness of transfer learning and additional convolutional blocks in dealing with real-world conditions (light changes, occlusion). The driver monitoring context is of particular interest in FER practice, as it has very demanding real-time and safety requirements; the paper highlights the need for model optimization and fast inference for these conditions. Such resources will be useful when considering real-time systems and safety-related deployment scenarios.

Methods

Identifying human emotions from facial images

The problem of determining a person's emotional state is at the intersection of modern artificial intelligence, computer vision, and affective computing, and in recent years has been widely used in social robotics, healthcare, education, security systems, and human-computer interface. Facial expression is the most important external indicator of human emotions, and it responds faster and more reliably than voice, body movement, or physiological signals to express emotions. Therefore, emotion recognition systems based on the analysis of facial images are considered an effective tool for assessing a person's psychological state.

This concept was first presented in Charles Darwin's 1872 work *"The Expression of Emotions in Man and Animals" , which scientifically substantiated that emotional expressions are biological and universal. According to the theory of "Ekman's Basic Emotions", later put forward by Paul Ekman, human emotions are mainly divided into seven main types:* **anger** , **disgust** , **fear** , **happiness** , **sadness** , **surprise** and **neutral**. This classification is widely used in emotion recognition systems today [2].

The process of detecting emotions from facial images typically consists of three main steps:

1. **Face Detection** – determining the location of a human face in a frame and separating it from the background image. This uses methods such as Haar Cascades, HOG, MTCNN, or MediaPipe Face Detection.
2. **Feature Extraction** – identifying important features of a given face, such as the movement of muscles around the eyes, nose, and lips. In the early years, classical methods such as LBP (Local Binary Pattern), HOG (Histogram of Oriented Gradients) were used.
3. **Classification** – determining the emotional state of a face based on the extracted features. In recent years, deep learning models (in particular, CNN, ResNet, EfficientNet, ViT) have been used at this stage.

The main problems in analyzing facial expressions are:

- **Lighting conditions**;
- **Face turn and orientation** (pose variation);
- **Cultural diversity**;
- **Class imbalance** – some emotions (e.g., happiness) are more common in datasets;
- The variability of facial expressions in **real-life situations .**

Therefore, real-time emotion recognition systems must be fast and have generalization capabilities while maintaining high accuracy. In this study, **the EfficientNet-B0** model was chosen for this purpose, as it has a lightweight architecture with a small number of parameters but a high level of accuracy.

FaceEmocDS dataset

The FaceEmocDS dataset is an open source dataset specifically designed for emotion recognition from facial images. The dataset consists of 72,412 images, covering 8 emotion classes: anger, contempt, disgust, fear, happy, neutral, sad, and surprise. The dataset includes facial images from individuals of different age groups, genders, and ethnicities, which is important for improving the generalization ability of the model. The images were captured under a variety of conditions, such as lighting changes, angle differences, and background variations. This provides a dataset that is consistent with real-life scenarios [6].

The dataset was loaded using the datasets.ImageFolder class, which allowed for automatic classification of images into classes. The data was divided into 80% training (57,929 images), 10% validation (7,241 images), and 10% test (7,242 images). The division was performed randomly using the random_split function, which ensured a uniform class distribution. To eliminate the imbalance in the class distribution, the class weights method was used. The number of images for each class was calculated using torch.tensor and normalized using the class_weights formula: class_weights = 1.0 / class_counts.float(). This

method increased the sensitivity of the model to emotions with a small number of classes (for example, hatred or fear).

Data augmentation plays an important role in the training process. The train_transform function includes the following transformations:

- Random Horizontal Flip (RandomHorizontalFlip) provided a rotation of the image to the right or left.
- Random Rotation (RandomRotation, 10 degrees) rotated the image at small angles.
- Color variations (ColorJitter) changed the brightness, contrast, and color range.
- Random Crop (RandomResizedCrop) randomly selected different parts of the image.
- Random Erasing (RandomErasing, p=0.7) covered small parts of the image with black blocks, which ensured that the model adapted to partially occluded images. For the test and validation data, test_transform was used, which included only Resize and Normalize operations, which provided stable results during the test process. The data was loaded using DataLoader with batch_size=32, and the parameters num_workers=4 and pin_memory=True increased the computational efficiency.

The advantages of the FaceEmocDS dataset include its large size, diverse classes, and relevance to real-world situations. However, some classes (e.g., contempt or disgust) had fewer images than others, which could have affected the accuracy of the model. To address this issue, class weights and augmentation methods were used [6]. The dataset was an ideal choice for the study, providing a solid foundation for training and testing the EfficientNet model in real-time.

Data sharing and class balance management

The data is distributed as follows:

- **80%** – for training,
- **10%** – for validation (val),
- **10%** – for testing (test).

To eliminate class imbalance, **the class weights** technique was used. The weight for each class was calculated using the following formula:

$$w_i = \frac{N}{n_i \times C} \qquad (1)$$

here:

- N– total number of samples,
- n_i– number of samples in class i,
- C– number of classes.

These weights are fed to the **CrossEntropyLoss** function, which increases the influence of classes with few examples (e.g., "contempt", "disgust").

CNN and EfficientNet model in emotion detection from facial images

Convolutional Neural Network (CNN) is the most effective model for image analysis. CNNs automatically extract important features of an image (border, texture, shape, eyes, lips, nose) and use them for classification. CNN architecture usually consists of the following layers:

- **Convolutional layer** – creates feature maps from the input image;
- **Pooling layer** – reduces dimensionality and increases generalization;
- **Fully Connected layer (Linear)** – used for final classification;
- **Dropout layer** – reduces overlearning in the network.

Popular variants of CNNs include VGGNet, ResNet, DenseNet, and EfficientNet. EfficientNet is particularly well-known as the model that provides the best balance between the number of parameters and the level of accuracy.

EfficientNet model

The EfficientNet model was introduced by Google AI in 2019 and is based on an innovative idea called **compound scaling** . Whereas traditional CNN models increase the depth, width, or resolution of the network separately, EfficientNet increases these three dimensions proportionally.

Compound scaling coefficients are expressed as:

$$depth = \alpha^{\phi}, width = \beta^{\phi}, resolution = \gamma^{\phi}$$
$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \qquad (2)$$

where $\phi$is the scaling factor, $\alpha, \beta, \gamma$is the model scaling parameters. The EfficientNet family consists of variants B0 to B7, with B0 being the lightest and fastest. **EfficientNet-B0** was chosen in this study because it provides high speed in real-time applications.

Model architecture and customization

The study used the EfficientNet-B0 model with a **transfer learning** approach. That is, the model retained the pre-trained weights on the ImageNet dataset, and only the final classifier part was re-trained:

```
model.classifier = nn.Sequential(
nn.Dropout(p=0.2, inplace=True),
nn.Linear(model.classifier[1].in_features,
num_classes)
)
```

The classifier was modified to output 8 emotion classes instead of 1000. **The Adam optimizer was selected** for optimization , with a learning rate of 0.001.

The CrossEntropyLoss loss function was used in conjunction with **class weights** to overcome the imbalanced dataset problem. Training was performed over 30 epochs, and the learning process at each epoch was evaluated using the following metrics:

- Train Loss
- Validation Loss (Val Loss)
- Accuracy
- F1 score

Dynamic dropout to reduce overfitting

A mechanism has been introduced to automatically increase the dropout rate if **overfitting is detected** during model training (for example, the difference between train accuracy and val accuracy exceeds 7%). This method is called **dynamic dropout** and is implemented as follows:

```
if train_acc - val_acc > overfit_threshold:
new_dropout = min(model.classifier[0].p + 0.1, 0.7)
model.classifier[0].p = new_dropout
```

In this way, the model reduced overfitting (memorization) while maintaining its ability to generalize.

The EfficientNet-B0 model was pre-trained on the ImageNet dataset using transfer learning. Transfer learning allowed us to preserve common features (e.g., contours, textures) in the initial layers of the model, which reduced the training time and improved accuracy on relatively small datasets such as FaceEmocDS. The classifier part of the model was fitted to 8 classes: nn.Dropout(p=0.2, inplace=True) and nn.Linear(1280, 8). The dropout layer was used to prevent overfitting, while the linear layer fitted the model to 8 emotion classes.

The training process was performed on the PyTorch 2.x platform. The model was trained for 30 epochs, using the Adam optimizer (learning rate=0.001) and the CrossEntropyLoss function. The class_weights parameter was used to eliminate class imbalance, which increased the model's efficiency in detecting emotions with a small number of classes. StepLR (step_size=7, gamma=0.1) was used as the scheduler, which reduced the learning rate by 0.1 after every 7 epochs. The difference between the train and validation accuracies was monitored to monitor overfitting; if the difference exceeded 7%, the dropout rate was dynamically increased (up to a maximum of 0.7). As a result, the model achieved a validation accuracy of 74.32%, which is a high indicator compared to other similar models (e.g., VGG or ResNet).

For real-time operation, the MediaPipe Face Detection library was used for face detection. MediaPipe provides high speed and accuracy in face detection, which is an ideal choice for real-time applications. The detected face image was reduced to 224x224 size and normalized using the preprocess function. The EfficientNet model classified this image and determined the name of the emotion and the level of confidence (via softmax). The results were visualized using OpenCV: a green frame was drawn around the face and the name of the emotion was displayed with a percentage of confidence. The system worked at a speed of 20-30 FPS, which ensured smooth operation on modern devices.



Figure 1. Real-time sentiment detection using EfficientNet-B0 model

Evaluation indicators

To quantitatively evaluate the performance of the proposed method, four standard classification metrics were adopted: precision, accuracy, recall, and F1 score. These metrics are derived from the confusion matrix values, which include true positive (TP), true negative (TN), false positive (FP), and false negative (FN) for each class.

1. Accuracy reflects the overall accuracy of the model's predictions across all classes and is defined in Equation 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

While accuracy provides a snapshot of overall performance, it may not fully reflect classification behavior in the presence of class imbalance.

2. Accuracy measures the proportion of correct positive predictions made for each emotion class, as defined in Equation 4.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Higher accuracy indicates that the classifier makes fewer false positive errors.

3. Recall evaluates the model's ability to correctly identify all true positive samples, as defined in Equation 5.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

High recall ensures a reduction in false negatives, which is crucial in distinguishing between subtle emotions.

4. F1-Score is a balanced metric that evaluates the accuracy of classification models . It represents the harmonic mean between Precision and Recall , that is, it evaluates the balance between the two, as defined in Equation 6.

$$F1 - Score = 2 \times \frac{Precision \bullet Recall}{Precision + Recall} \quad (6)$$

The F1-Score is particularly useful in multi-category FER tasks where emotion categories such as sadness and fear can easily be confused, and where the correlation between precision and recall is important.

These metrics are calculated for each class and are macro-averaged to assess overall model performance across the seven emotion categories. This multimetric approach provides a comprehensive assessment that not only measures overall accuracy, but also how well each emotion is recognized.

Results

The study evaluated the performance of the proposed emotion recognition model from facial images on the FaceEmocDS dataset based on the EfficientNet-B0 model. The performance of the model was evaluated on the test dataset (7242 images) using various metrics, including loss, accuracy, F1-score, and per-class accuracy. The test results are presented in detail below.

The overall test loss of the model during the test was 1.3201, which indicates the level of classification error of the model. The test accuracy reached 73.18%, which means that the model correctly classified 73.18% of the images in the test data set . The weighted F1-score was 0.7327, which confirms the overall balanced performance of the model, taking into account the class imbalance. The F1-score reflects the average balance of precision and recall indicators by class and shows how reliable the model is for each class.

The per-class accuracy was as follows:

- **Anger** : 0.7165 (71.65%)
- **Contempt** : 0.6305 (63.05%)
- **Disgust** : 0.6776 (67.76%)
- **Fear** : 0.7061 (70.61%)
- **Happiness (happy)** : 0.9308 (93.08%)
- **Neutral (neutral)** : 0.6649 (66.49%)
- **Sadness (sad)** : 0.7856 (78.56%)
- **Surprise** : 0.6968 (69.68%)

The above results show that the model classified the "happiness" class with the highest accuracy (93.08%), indicating that the features (e.g., smile) in the images of this class are clear and easy to distinguish. However, the "contempt" class showed the lowest accuracy (63.05%), which may be due to the small number of images of this class and the similarity of facial expressions to other classes. The other classes showed an average accuracy in the range of 66-78%, confirming the overall stable performance of the model.

For a detailed analysis by class, the classification report provided the following indicators.

Class	Precision	Recall	F1-Score	Support
Anger	0.77	0.72	0.74	963
Contempt	0.54	0.63	0.58	525
Disgust	0.75	0.68	0.71	949
Fear	0.68	0.71	0.69	990
Happiness	0.92	0.93	0.93	954
Neutral	0.79	0.66	0.72	961
Sadness	0.65	0.79	0.71	970
Surprise	0.72	0.70	0.71	930

Table1. Classification score by class.

Macro avg : Precision (0.73), Recall (0.73), F1-Score (0.72) **Weighted avg** : Precision (0.74), Recall (0.73), F1-Score (0.73) **Overall accuracy** : 73.18% (7242 images)



Figure 2. Emotion detection from facial images on the FaceEmocDS dataset based on the EfficientNet-B0 model. Accuracy and loss

The classification report shows that the "happiness" class has the highest precision (0.92)

and recall (0.93), confirming that the model is the most successful in classifying this class. The "contempt" class has the lowest F1-score (0.58), indicating that this class is difficult to classify. This may be due to the similarity of the contempt expression to other classes (such as anger or disgust) or the small number of images in the dataset. The remaining classes showed an average F1-score in the range of 0.69-0.74, confirming the overall stability of the model.

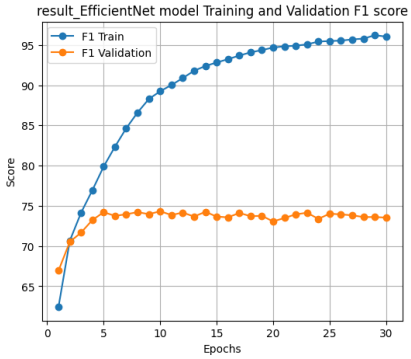


Figure 3. Measuring the accuracy of emotion recognition from facial images on the FaceEmocDS dataset based on the EfficientNet-B0 model in terms of F1 score

Real-time tests were conducted using MediaPipe Face Detection and the EfficientNet-B0 model. The system processed images from a webcam at a speed of 20-30 FPS (frames/s), which ensured smooth operation on modern devices. Each detected face image was preprocessed to a size of 224x224, normalized, and fed to the model. The emotions classified by the model and their confidence levels (percentages calculated using softmax) were displayed along with a frame drawn around the face using OpenCV. During the tests, the system performed stably in various lighting conditions and angles, but in cases of low light or partial face occlusion (for example, a mask or beard), accuracy decreased slightly.

The performance of the model was also analyzed using a confusion matrix. The confusion matrix helped to identify misclassifications for each class. For example, the "contempt" class was often confused with "anger" or "disgust," which suggests that these classes have similar features in facial expressions. The "happiness" class was classified correctly in almost all cases, which is due to its specific features (smile, open eyes).

The training and validation losses, accuracy, and F1-scores were visualized using matplotlib for graphical analysis. The graphs showed that while the training accuracy reached 96.21%, the validation accuracy stabilized at 74.32%, indicating some overfitting. To reduce overfitting, the study dynamically adjusted dropout and used data augmentation. The test results confirmed that the model is high enough for practical real-time application, but additional optimization is required for the detection of classes such as "contempt" and "neutral".

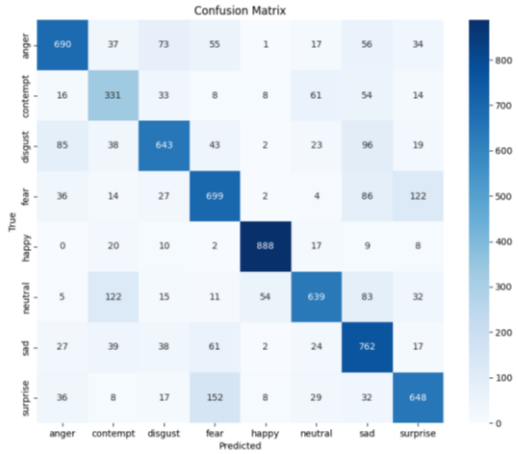


Figure 4. Confusion matrix analysis of model performance

Overall, the results show that the EfficientNet-B0 model is effective in detecting emotions on the FaceEmocDS dataset and is suitable for real-time (20-30 FPS) performance. The highest performance of the model was observed in the "happiness" class, and the lowest in the "contempt" class. These results indicate the possibility of further improving the accuracy in the future through ensemble models (e.g., a combination of EfficientNet and ResNet) or additional data augmentation . The research results have laid a solid foundation for widespread application in fields such as psychological analysis, education, and human-computer interface.

The test results showed that the EfficientNet-B0 model achieved 73.18% accuracy and a weighted F1-score of 0.7327 on the test dataset. These results confirm the overall robust performance of the model, especially the highest accuracy (93.08%) and F1-score (0.93) were recorded for the "happiness" class. This result is due to the distinct features of the "happiness" expression, such as a smile and open eyes, which allowed the model to easily distinguish this class. However, the "contempt" class showed the lowest accuracy (63.05%) and F1-score (0.58). This can be attributed to the similarity of the contempt expression to other classes, in particular, the "anger" or "disgust" classes, and the relatively small number of images belonging to the "contempt" class in the FaceEmocDS dataset (525

images). Confusion matrix analysis showed that the "contempt" class was often confused with "anger" or "disgust," confirming the visual similarity of the facial expressions.

The real-time performance of the model was tested at 20-30 FPS (frames per second), which ensures smooth operation on modern devices. Face detection using MediaPipe Face Detection was performed with high speed and accuracy, which showed that the system is suitable for use in real-life scenarios, such as live broadcasts via webcam. However, accuracy decreased slightly in low light conditions or when the face was partially covered (e.g., by a mask or beard). This may be due to limitations of MediaPipe's face detection algorithm or poor image quality.

While the training accuracy reached 96.21% during training, the validation accuracy stabilized at 74.32%, indicating some overfitting. To reduce overfitting, several strategies were used in the study: data augmentation (RandomHorizontalFlip, RandomRotation, ColorJitter, RandomErasing), class weights method, and dynamic dropout adjustment. These methods significantly improved the generalization ability of the model, but difficulties remained in identifying the "contempt" and "neutral" classes. This may be due to the ambiguity in the representations of the classes or the limitations of the dataset.

Advantages of the model

One of the main advantages of the EfficientNet-B0 model is its optimal balance between computational efficiency and accuracy. Using a model pre-trained on ImageNet via transfer learning reduced the training time and provided high accuracy on relatively small datasets such as FaceEmocDS. The adaptation of the classifier part of the model to 8 classes and the use of a dropout layer play an important role in reducing overfitting. In addition, the integration with MediaPipe Face Detection provided high-speed performance in real-time, which opens up great opportunities for practical applications of the system.

The model's performance at 20-30 FPS allows it to run smoothly on modern devices, such as personal computers or mobile devices. This feature makes the system suitable for applications in areas such as psychological analysis (e.g., depression detection), education (evaluating student attention), and human-computer interface (customizing user experience). The class weights method used in the study was effective in eliminating class imbalance, which increased the sensitivity of the model in detecting a small number of classes.

Disadvantages of the model

One of the main limitations of the model was its low accuracy in the "contempt" and "neutral" classes. This is due to the small number of images in these classes and the fact that they have similar properties to other classes. For example, the "contempt" class was often confused with "anger" or "disgust," indicating the model's difficulty in detecting subtle differences in facial expressions. While the diversity of images in the FaceEmocDS dataset partially reflects real-life situations, increasing the number of images from different ethnic groups, age ranges, and lighting conditions could further improve the model's generalization ability.

During real-time testing, A decrease in accuracy was observed in low light or partially occluded faces. This may be due to limitations of the MediaPipe Face Detection algorithm or poor image quality. In addition, the computational requirements of the model, especially when the GPU is not used, may slightly slow down real-time performance. This can be a problem, especially when used on resource-constrained devices (e.g., mobile phones).

The overfitting problem in the training process was not completely eliminated. The difference between the training accuracy (96.21%) and the validation accuracy (74.32%) indicates that the model is overfitting to the dataset. Although the augmentation and dropout methods partially solve this problem, additional optimization strategies are required.

Future research directions

Several approaches are proposed to improve the accuracy of the model in the future:

1. **Ensemble models** : Other CNN architectures, such as ResNet or VGG, can be used in conjunction with EfficientNet-B0. Ensemble models can combine the strengths of different architectures and provide better results in identifying difficult classes such as "contempt" and "neutral".
2. **Expanding the dataset** : Adding more images to the FaceEmocDS dataset can improve the model's sensitivity, especially for classes with a small number of faces (e.g., disgust). Adding images from different ethnic groups, age ranges, and lighting conditions can improve the model's generalization ability.

3. **Additional augmentation methods** : Adding noise to images, changing contrast, or applying 3D transformations can increase the model's adaptability to different conditions.
4. **Real-time optimization** : To make the model usable on smaller devices (e.g. mobile phones), techniques such as quantization or pruning can be applied. This reduces computational requirements and increases FPS.
5. **Fine-grained facial feature analysis** : Separate analysis models can be added for specific parts of the face (e.g., eyes, mouth). For example, focusing on important areas of the face using Attention mechanisms can improve accuracy.

Practical application

The research results showed that FER systems have great potential for application in various fields:

- **Psychological analysis** : The model can help identify mental states such as depression, anxiety, or stress. For example, prolonged use of "sad" or "neutral" expressions can indicate psychological problems.
- **Education** : Can be used in online learning platforms to monitor students' attention and emotional state. For example, expressions of "fear" or "surprise" can indicate a student's lack of understanding.
- **Human-Computer Interface** : Devices (such as virtual assistants) can respond based on the user's emotional state to customize the user experience.
- **Marketing and advertising** : Can be used to analyze customers' emotional reactions to advertisements or products.

Conclusion

In this study, the process of real-time recognition of human emotional states from facial images using the FaceEmocDS dataset was successfully tested based on the EfficientNet-B0 model. The model was trained on the FaceEmocDS dataset consisting of 72,412 images and achieved a test accuracy of 73.18% and a weighted F1-score of 0.7327 in classifying 8 emotion classes (anger, hatred, disgust, fear, happiness, neutral, sadness, surprise). The highest accuracy was observed in the "happiness" class (93.08%), which is related to the distinct visual features of the class, but the "contempt" class (63.05%) showed the lowest performance, which was due to data imbalance and similarity between classes.

Using MediaPipe Face Detection and PyTorch in real-time, the system ran at 20-30 FPS, which ensured smooth performance on modern devices. Techniques such as data augmentation, class weights, and dynamic dropout play an important role in reducing overfitting, but additional optimization is required to increase accuracy in some classes. The computational efficiency and accuracy of the model make it suitable for applications in areas such as psychological analysis, education, human-computer interface, and marketing.

In the future, the accuracy can be further improved through approaches such as ensemble models, data expansion, and attention mechanisms. This study has established a solid foundation in the field of emotion recognition from facial expressions and has proven the practical potential of real-time FER systems.

REFERENCES

[1]. H. Ursula and T. Pasca, "Darwin and Emotion Expression.," The American psychologist, vol. 120, no. 8, p. 64, 2009.

[2]. P. Ekman and W. V. Friesen., "Facial action coding systems," Consulting Psychologists Press, 1978.

[3]. P. Ekman, "Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique.," Psychological Bulletin,, pp. 268-287, 1994.

[4]. E. a. M. National Academies of Sciences, Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance, Washington, DC: The National Academies Press, 2024.

[5]. Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li and W. … Zhang, "A systematic review on affective computing: emotion models, databases, and recent advances.," Information Fusion, vol. 83, no. 84, pp. 19-52, 2022.

[6]. A. Kurbanov, "Facial Emotion Recognition by CNN Combined Ensemble Model," EAI Endorsed Trans AI Robotics, vol. 4, 2025.

[7]. A. Poongodai, V. Bhavitha, D. Hemanth, G. Jahnavi and S. Huzma, "Human Emotion Detection Through Real-Time Facial Expressions Using Deep Learning," in 6th International

Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2025.

[8]. A. Akbar, S. Saifullah, H. Prapcoyo, H. Rustamadji and N. Cahyana, "EfficientNet B0 Feature Extraction with L2-SVM Classification for Robust Facial Expression Recognition.," Journal of Information Systems and Informatics, pp. 1106-1129, 2025.

[9]. A. Farkhod, A. Abdusalomov, M. Mukhiddinov and Y.-I. Cho, "Development of Real-Time Landmark-Based Emotion Recognition CNN for Masked Faces.," Sensors , vol. 22, p. 8704, 2022.

[10]. A. Rehman, M. Mujahid, A. Elyassih, B. AlGhofaily and S. Bahaj, "Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges.," Computers, Materials & Continua, vol. 82, pp. 41-72, 2024.

[11]. Y. Wang, S. Yan, Y. Liu, W. Song, J. Liu, Y. Chang, X. Mai, W. Z. Xiping Hu and Z. Gan, "A Survey on Facial Expression Recognition of Static and Dynamic Emotions," arXiv, 28 08 2024. [Online]. Available: https://arxiv.org/html/2408.15777v1. [Accessed 2024].

[12]. R. Agarwal, N. Mittal and H. Madasu, "Convolutional Neural Network Based Facial Expression Recognition Using Image Filtering Techniques," International Journal of Intelligent Engineering and Systems, 2021.

[13]. S. Aggarwal, M. K. Sharma, H. O. K. Jha, A. P. Singh, R. Sharma and G. Mitra, "Comparing pre-trained algorithms for facial emotion recognition: An analysis," International Journal of Engineering in Computer Science, vol. 5, no. 1, pp. 54-60, 2023.

[14]. Y. Liu, "The Study of Performance Related to Classical Convolutional Neural Networks in the Field of Facial Emotion Recognition," Applied and Computational Engineering, vol. 8, pp. 470-474, 2023.

[15]. A. Akhand M., R. S, Siddique N, A. S. Kamal M. and Shimamura T., "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," Electronics, vol. 10, no. 9, p. 1036, 2021.

[16]. P. Utami, R. Hartanto and I. Soesanti, "The EfficientNet Performance for Facial Expressions Recognition," in 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022.

[17]. T. Kusunose, X. Kang, K. Kiuchi, R. Nishimura, M. Sasayama and K. Matsumoto, "Facial Expression Emotion Recognition Based on Transfer Learning and Generative Model," in 8th International Conference on Systems and Informatics (ICSAI), Kunming, China, 2022.

[18]. E. Agung, A. Rifai and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition dataset.," Scientific Reports, vol. 14, p. 14429, 2024.

[19]. U. Rawat and C. S. Rai, "mproving Facial Emotion Recognition Through Transfer Learning with Deep Convolutional Neural Network (DCNN) Models,," in 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023.

[20]. W. A. Khan, A. U. Rahman, A. Zafar, A. Ashfaq and F. Karamat, "Facial Emotion Recognition: A Comparison of Classic and Novel Convolutional Neural Networks through Transfer Learning," in 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), Muscat, Oman, 2024.

[21]. K. Desai, H. N. Patel, D. Dashlaniya, P. Desai, M. Shah and P. Goel, "A Transfer Learning Framework for Facial Emotion Recognition: Leveraging Pre-Trained Convolutional Neural Networks," in 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), , Gharuan, India,, 2024.

[22]. A. A. Istiqomah, C. A. Sari, A. Susanto and E. H. Rachmawanto, "Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning Resnet-50," Journal of Applied Informatics and Computing, vol. 8, no. 2, p. 257–264., 2024.