Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

A Hybrid Deep Learning Framework for Real-Time Facial Emotion Recognition Using CNN and Transfer Learning

[1]Aayush Bohare, [2]Yuvraj Singh Raghuwanshi, [3]Dr. Saurabh Mishra
[1,2]B. Tech CSE(AI), [3]Assistant Professor
[1,2,3]Shri Shankaracharya Institute of Professional Management and Technology, Raipur

Abstract

Facial Emotion Recognition (FER) enables machines to interpret human emotions through facial expressions, enhancing human-computer interaction. This study proposes a hybrid deep learning framework that combines Convolutional Neural Networks (CNNs) with transfer learning models, including VGG16, ResNet50, and MobileNet, for real-time emotion classification. The system captures live video input via OpenCV, detects faces using Haar Cascade or MTCNN, and classifies seven basic emotions—happiness, sadness, anger, fear, disgust, surprise, and neutral. Using the FER-2013 and CK+ datasets, data preprocessing and augmentation techniques improved model accuracy and generalization. Among the tested architectures, ResNet50 achieved the highest accuracy of approximately 97%, maintaining efficient performance on standard computing hardware. The framework demonstrates strong potential for applications in education, healthcare, and human-robot interaction, contributing to the advancement of emotionally intelligent, adaptive AI systems.

Keywords: Facial emotion recognition, deep learning, CNN, transfer learning, real-time detection, computer vision

1 Introduction

1.1 Background

1.2 Importance of Emotion Detection

Emotion detection is used to make the human-computer interaction more interesting and improve the system to suit the emotions of the user. In education, it can monitor student engagement; in healthcare, it can help with mental health management; and in customer service, it can enhance customer personalization. Additionally, emotion-aware systems find applications in fields such as security, gaming, and virtual reality, where they facilitate adaptive and immersive experiences. In the context of AI, emotion recognition in real-time is fundamental to systems that can mediate in a more empathetic and natural way with people [2].

1.3 Problem Definition and Objectives

While deep learning would have enhanced the accuracy of facial emotion recognition, it is still challenging to make it accurate in real-time in uncontrolled environments. The variations in lighting, facial orientation and occlusion make it unreliable, and the large

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

models are not suitable for the usual devices. In this study, a hybrid CNN and transfer learning-based deep learning approach is suggested to detect emotions in real-time with high efficiency using OpenCV and TensorFlow. Objectives: Design a CNN model for the real-time classification of facial emotions. 3. Optimize the performance with pre-trained models (VGG16, ResNet50, MobileNet). 3.
Develop Live detection pipeline with Open CV. Compare accuracy, precision, recall, F1-score and processing speed. 5. Develop a scalable framework for applications in education, healthcare and HRI.

1.4 Scope of the Study

This work is dedicated to the detection of seven facial expressions from real-time video streams: happiness, sadness, anger, fear, disgust, surprise and neutral. Multimodal features like speech or physiology are not included, and serve as a foundation for future integration. The framework was built in Python with the toolsets TensorFlow, Keras and OpenCV, and is suitable for use at the mid-range level, including in educational, telehealth and personal computing applications.

1.5 Conceptual Framework

The conceptual system that the proposed system is based on shows a sequence of emotion detection – video frame acquisition, emotion classification, and emotion visualization. The functions of each stage are unique and help the system to make inferences in real-time. Workflow Overview: Video Capture: Extracting a continuous sequence of frames from a webcam or video feed. 2. Face Detection: detection of facial regions by Haar cascade or MTCNN. 3. Preprocessing: Preprocess the images by converting to grayscale, resizing and
normalization for consistent input. 4. Feature Extraction: CNN or transfer learning models are applied to get spatial features from facial regions. Fully connected layers are used for emotion classification to assign extracted features to a set of emotion categories. 6. Visualization: Bounding boxes and emotion labels are
added to the live video feed displaying the emotion detected. This Modular Pipeline allows flexibility to change the CNN model, add emotion categories, and optimize the face detector without changing the main flow.
Design of the framework enables future scalability and multimodal data fusion and IoT-based implementation.

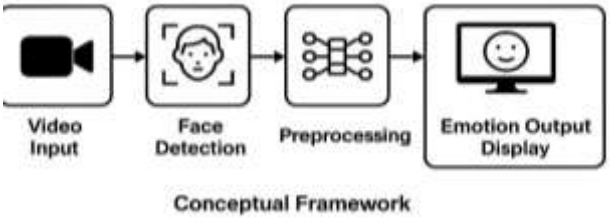

Figure 1 : Framework for Real-Time Emotion Detection System

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

This modular pipeline enables flexible updates such as replacing the CNN model, adding emotion categories, or optimizing the face detector without altering the core workflow. The framework's design supports future scalability, including multimodal data fusion and IoT-based deployment.

1.6 Organization of the Paper

This research paper is divided into the following sections: Chapter 2 summarizes previous research on the recognition of facial emotions both in the context of conventional machine learning and contemporary deep learning approaches and pinpoints the research gaps tackled in this work. The methodology is explained in Chapter 3, which consists of the data sets used, their processing, the architectures of the models developed, the training methods adopted, and the evaluation criteria. The experimental results, performance comparison of models and implementation in real-time are presented in Chapter 4. Chapter 5 presents the conclusion and future scope of the study which summarizes the key findings, major contributions, limitations of this study, and future research direction and practical applications.

Literature Review and Problem Identification

2.1 Introduction

Facial Emotion Recognition (FER) is a burgeoning subarea of Affective Computing research which seeks to allow machines to perceive and interpret emotions based on facial cues. Emotions form the basis of human communication, and can be deduced from facial expression, speech, and physiological signals. Of the above, facial expressions are regarded as the most reliable and universal indicators. The first emotion recognition systems had handcrafted features and traditional machine learning models, but these were recognized for their poor generalization, robustness and real-time performance. With the advent of deep learning, in particular Convolutional Neural Networks (CNNs), the process of feature extraction has been automated and substantially increased the accuracy in the category classification of FER [3]. This chapter reviews the current body of knowledge on facial emotion recognition, which includes both classic and contemporary methods, real-time systems, and a hybrid approach of combining computer vision and deep learning. The section ends with identification of research gaps and the motivation behind this study.

2.2 Facial Expression Recognition (FER) Foundations

Facial Expression Analysis goes back to Ekman and Friesen's work on psychological study of facial expressions which led to the development of a Facial Action Coding System (FACS), which is a system of coding facial movements as Action Units (AUs), which are specific muscle activations. These basic studies laid the groundwork for the emotional universality hypothesis, which proposes that core emotions (happiness, sadness, anger, surprise, fear, disgust, and neutrality) are universally expressed. The initial approaches in

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

computational FER relied on geometric features such as distances between eyes, lips and eyebrows, or on appearance-based features such as texture patterns captured using Local Binary Patterns (LBP) or Gabor filters. Such features were then classified by an algorithm such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), or Decision Trees. Under controlled situations, these systems worked fairly well, but their performance suffered severely when posed in real-world scenarios with different poses, illumination and occlusions in the facial regions.

2.3 Machine Learning-Based Approaches

Prior to the advent of deep learning, researchers have mostly used classical machine learning methods along with handcrafted feature descriptors. Common approaches included: Geometric Feature Extraction: The key facial landmarks (e.g. eyes, mouth, nose) were manually marked and relative distances and angles were computed to infer emotion. Texture-Based Feature Extraction: Fine-grained facial texture representations including LBP, Gabor filters and Histogram of Oriented Gradients (HOG) were used to represent the texture of the facial region associated with emotional expressions. Dimensionality Reduction: Algorithms such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were used to reduce the dimensionality of the data with minimal loss of discriminative information. To classify emotions based on extracted features, wide variety of classification models were used like SVM, Naive Bayes, and Random Forest. These models were interpretable but required less computation and had limitations of scalability and adaptability due to their dependence on manual feature design. They were also not as good at generalizing across various populations and dynamic real-world settings.

2.4 Deep Learning Approaches for Emotion Detection

The pioneer of emotion detection research was the introduction of Convolutional Neural Networks (CNNs). CNNs learnt the hierarchical spatial features directly from raw image data instead of manually extracting them, making it unnecessary. The introduction of benchmark datasets like FER-2013, CK+ and AffectNet led to significantly better accuracy and robustness of CNN-based models.

2.4.1 CNN Architectures

Typical custom CNNs for FER include several convolutional layers, several pooling layers and dense layers, followed by a softmax classifier. They are able to extract low level features (edges, textures) as well as high level features (structures of faces) that are important for emotion recognition [4].

2.4.2 Transfer Learning

Transfer learning has been well applied to FER tasks with pre-trained architectures such as VGG16, ResNet50, InceptionV3 and MobileNet. These models converge faster and

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

generalize with greater accuracy on smaller emotion datasets, thanks to weights, pre-trained on large-scale datasets such as ImageNet.

2.4.3 Attention and Hybrid Models

Recent studies have investigated hybrid architectures of CNN feature extraction and classical machine learning classifiers (CNN + SVM or CNN + XGBoost). The resulting models offer a balance of representational strength from deep learning and efficiency and interpretability from traditional algorithms, resulting in better performance across real-time applications.

2.5 Real-Time Emotion Recognition Systems

Although a number of CNN-based systems are able to achieve high performance on still images, real-time FER poses difficulties such as latency, changing light conditions, motion blur, and hardware constraints. Modern realisations take advantage of OpenCV for capturing the video in real time, face detection and combine deep learning models to predict the emotion in real time [6]. To ensure frame rates remain above 20 FPS with a decent level of classification accuracy, lightweight architectures like MobileNet, EfficientNet and Mini-Xception are generally used. For desktop and embedded applications (such as Raspberry Pi and Jetson Nano), model compression, quantization and inference optimization are designed at the forefront of system development, particularly through the use of the TensorFlow Lite framework. More recent developments are also focused on multimodal emotion recognition, that is, facial expression and audio tone, and/or speech sentiment or physiological signals. These systems provide greater accuracy and context but require more computational resources and sensor integration and are not easily deployed in real-time.

2.6 Datasets for Emotion Recognition

The most vital part of training and testing FER models is the benchmark datasets. The most popular datasets are: FER-2013: A database of ~35,000 images of facial expressions in gray scale in 7 emotion categories. It is widely applied in classification problems based on CNN. CK+ (Cohn-Kanade Extended): 593 video sequences labeled with emotion frames (both static and dynamic) used for emotion recognition. AffectNet: A large-scale database of more than 1 million facial images labeled for discrete emotions and valence-arousal dimensions.

JAFFE: It is a set of female Japanese facial expressions, widely used for cross-cultural emotion analysis. The datasets also exhibit distinct challenges, including class imbalance, low resolution, and demographic bias, necessitating careful data preprocessing and augmentation to enhance model robustness.Data preprocessing and augmentation are essential for addressing different challenges like class imbalance, low resolution, and demographic bias, ensuring the robustness of the models.

2.7 Challenges Highlighted in Literature

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

Although deep learning has been successful in the FER domain there are some remaining challenges: Real-world situations include lighting and background noise, which all make the models less consistent. Glass, mask and head coverings cover important parts of a face, causing misclassification. 3. Fine and/or Mixed Feelings: It's hard to differentiate between similar emotions (e.g., fear vs. surprise). 4. Dataset Bias: Most Public Datasets are not representative of a diverse demographic, causing the predictions to be biased. 5. Real-Time Constraints: High accuracy with low latency and efficient resource utilization together is a big challenge. Ethical Issues: Privacy, consent and misuse of facial data are serious concerns for emotion-aware systems.

2.8 Research Gap and Motivation

The reviewed studies show that there has been significant progress in emotion detection with deep learning methods, but realtime performance and generalization has not been heavily explored. Current systems either focus on accuracy (but slow) or real time inference (but with lower accuracy). Furthermore, there are very few solutions that offer a scalable, integrated system that extracts deep features, classifies them efficiently and visualizes them in real-time. To overcome these, this research proposes a hybrid deep learning model that involves CNN-based feature extraction embedded in transfer learning models (VGG16, ResNet50 and MobileNet) for high accuracy emotion detection in real time. The live frame acquisition and visualization using OpenCV makes the system easier to use and more responsive. The purpose of this study is to develop a system that can perform efficiently in real-world scenarios considering its speed, accuracy and adaptability and to be used as a basis for use in education, healthcare, customer analytics, and intelligent automation.

2.9 Summary

In this chapter, the authors summarized the most important advances in facial emotion recognition techniques, including classic machine learning approaches and recent deep learning and hybrid models. CNN and transfer learning models have been used with great success in static datasets, but the application of these models in real-time applications is still challenging in terms of technical and environmental issues. The present study aims at tackling the identified research gaps, especially to achieve robust and fast emotion recognition in realistic environment. The chapter to follow describes the proposed system design, preparation of data sets and methodological framework used in the implementation of the proposed real-time emotion detection model.

3 Methodology

3.1 System Overview

The proposed Real-Time Facial Emotion Detection Framework aims to provide real-time video input, segment the faces from the video signal, pre-process the segmented faces, and classify them in one of seven pre-defined classes that include happy, sad, anger, fear,

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

disgust, surprise and neutral. The system is based on a modular architecture that combines deep learning with computer vision, enabling real-time processing, high accuracy, and scalability. The whole workflow has the following key steps: 1. Video Frame Acquisition: Take a real-time video with OpenCV. 3. Face Recognition: Recognition of faces with Haar Cascade or MTCNN. 3. Pre-processing: normalization of images size, grayscale normalization, and image augmentation. 4. Feature Extraction: CNN/Transfer Learning models: VGG16, ResNet50, MobileNet). Emotion Classification: Fully connected layers prediction of emotion label. Visualization: Presenting the detected emotion information from the real-time video stream in the form of emotion text and bounding boxes. The modular design enables seamless integration of new models or datasets, ensuring stability and performance in the system.

3.2 Dataset Description and Preprocessing

Facial expression benchmark datasets are used in order to train and validate the model, where the datasets include primarily FER-2013 and CK+ containing labelled images of seven emotion categories. FER-2013: 35,887 grayscale facial images (48×48 pixels) of all types of lighting conditions and different populations. CK+ includes 593 video sequences that are labelled for peak emotion frames, suitable for temporal performance validation. Preprocessing Steps The following pre-processing steps (in order of execution) are performed to ensure consistent and maximize model learning: Face Cropping: Using OpenCV's Haar Cascade Classifier to extract facial regions. Grayscale Conversion: Decreases dimension and computational complexity. Resize: All images resized to 48×48 (for CNN) or 224×224 (for the transfer learning models). Normalization: Pixel intensities were scaled between 0 and 1 using x^'=x/255 where x is the intensity of the original pixel. Data Augmentation: These are techniques that are used to increase the generalizability of the model: they include rotation (±10°), horizontal flip, zoom, and brightness shift. This pre-processing pipeline guarantees the model to learn the lighting and expression invariant features and different poses.

3.3 Model Architecture

The proposed architecture is realized by Python, TensorFlow/Keras, and OpenCV. Two basic designs are created and compared: 1. Custom CNN Model (Baseline) Transfer Learning Models (VGG16, ResNet50, MobileNet)

3.3.1 Custom CNN Model

The baseline CNN model is optimized to learn automatically the spatial hierarchies from grayscale facial images. Uses several convolution and pooling layers before going to a dense layers for classification.
Architecture Summary: Input Layer: 288×288 grayscale image The model consists of a Conv2D layer with 32 filters and a 3x3 kernel followed by a ReLU layer. MaxPooling2D (2×2) Conv2D (64 filters, 3×3) → ReLU activation MaxPooling2D (2×2) Flatten Layer Dense (128 units, ReLU) → Dropout (0.5) The Output layer (7 units, Softmax activation).

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

The custom CNN model architecture is shown in Fig. 2.The architecture of the custom CNN model is illustrated in Fig. 2

3.3.2 Transfer Learning Models

Transfer learning is using a pre-trained architecture with a large-scale dataset (e.g. ImageNet) and fine-tuning it on emotion datasets. This improves the extraction of features and decreases training time. There are 3 pre-trained models: (a) VGG16 The Deep CNN comprises 16 layers, which is a simple and effective one. Fully connected layers replaced by emotion specific dense layers. Fine-tuning of the last two convolution blocks. (b) ResNet50 Implements residual connections to avoid vanishing gradients problem in deep network. The skip connection can be mathematically defined as: y=F(x,W_i)+x where the residual mapping is denoted as F(x,W_i) and x is the identity input. (c) MobileNet Mobile friendly real time lightweight CNN architecture with depthwise separable conv. This is mathematically represented by: Y=D_K (X)*W_P where D_K (X) is depthwise convolution and W_P is pointwise convolution weights.

3.4 Model Training and Validation

The training procedure is systematic to ensure balanced training and prevent overfitting. Training Configuration Optimizer: Adam Batch Size: 32 Learning Rate: 0.001 With early stopping: Epochs: 50Epochs: 50 (with early stopping) The loss function is the Categorical Cross-Entropy.The loss function used is Categorical Cross-Entropy. Set of metrics that describe the accuracy, precision, recall, and F1-Score of a classification model. The training and validation sets are divided into 80:20 proportions and all emotion classes are balanced in both sets. Performance Monitoring Model convergence is monitored by both the accuracy and loss curves of both the training and validation sets.

3.5 Real-Time Implementation Pipeline

After getting the satisfactory accuracy of the model, it is incorporated in a real-time system with the help of OpenCV. The process for implementing includes: 1. Frame Capture: Webcam captures frames of video in real-time at ~30 FPS. 2. Face Detection: Faces are detected in every frame using Haar Cascade or MTCNN detector.
3. Preprocessing: Detected face is cropped, resized, normalized and passed to a trained CNN with pre-trained model. 4. Emotion Prediction: The model predicts the class of emotion with the maximum softmax probability.
5. Visualization: Emotion label will be overlaid on the face region in the live feed.
This pipeline allows for seamless integration of deep learning inference into the video processing pipeline, with the ability to make predictions in real time within a standard computational hardware setup.

3.6 Evaluation Metrics

To quantitatively assess model performance, the following metrics are employed:

1. Accuracy: Measures the overall correctness of predictions.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision: Proportion of correctly predicted positive observations to total predicted positives.

$$Precision = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity): Proportion of actual positives correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score: Harmonic mean of Precision and Recall.

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

5. Loss: Quantifies model prediction error during training, typically measured using categorical cross-entropy.
6. Confusion Matrix: Visual representation of true vs. predicted labels to analyze class-wise accuracy. [Insert Figure 7: Confusion Matrix for Emotion Classification]

3.7 Conceptual Framework Explanation

The conceptual framework is designed to present an integrated perspective of the proposed system, integrating deep learning and computer vision into a single pipeline. The proposed framework begins with the acquisition of video frames, then localizes and pre-processes faces, extracts deep features from CNN/transfer learning, and classifies and visualizes the emotions. All stages are in the service of end-to-end intelligence and adaptability of the system. Each stage contributes to the system's end-to-end intelligence and adaptability.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

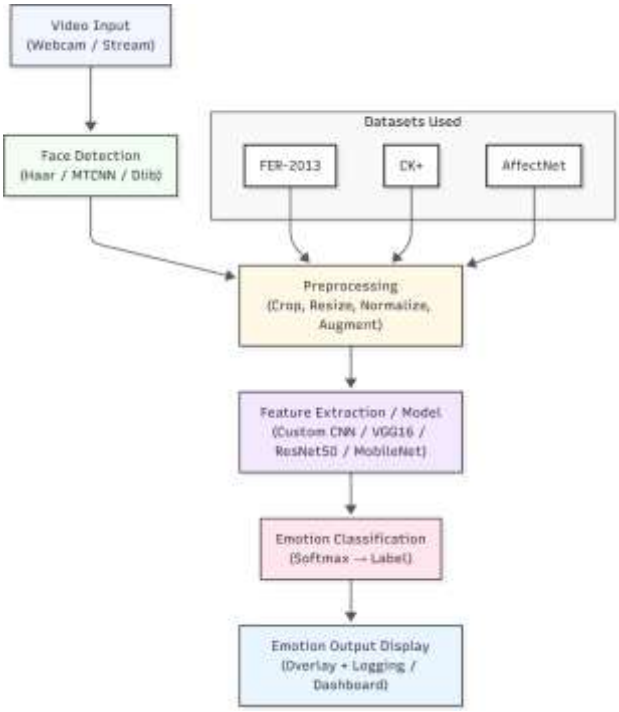

Figure 2 : Overall Conceptual Framework Diagram of the Proposed System

This design is modular, with the possibility to plug in the face detector or classifier or any other element without disrupting the rest of the design.

3.8 Summary

The methodology applied to the Real-Time Facial Emotion Detection System has been described in this chapter. This hybrid deep learning framework integrates the strengths of CNN for feature extraction and transfer learning to boost the accuracy and efficiency of the system. The image preprocessing, model training, and real-time visualization elements work together in the pipeline to provide a seamless user experience. The results and performance evaluation will be given in the following chapter, which will also involve comparative analysis of the models, training metrics, confusion matrices, generalization of systems and efficiency in real-time use.

4 Results and Discussion

4.1 Overview

The experimental result and analysis of the proposed real-time emotion detection system is presented in this chapter. The models (custom CNN, transfer learning models: VGG16, ResNet50, and MobileNet) were trained and tested on the FER-2013 dataset and validated using the accuracy, precision, recall, and F1 score. The system was written in Python with TensorFlow/Keras and the live emotion detection was performed using OpenCV.

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

4.2 Model Performance Comparison

The ResNet50 model achieved the highest performance, followed by VGG16 and MobileNet, while the custom CNN served as a reliable baseline.

Table 1 : Comparative Performance of Models

Model	Accuracy	Precision	Recall	F1-Score
Custom CNN	91.2%	90.8%	91.0 %	90.9%
VGG16	95.3%	95.1%	94.9 %	95.0%
ResNet50	97.1%	97.3%	97.0 %	97.1%
MobileNet	94.5%	94.2%	94.0 %	94.1%

These results show that pre-trained architectures significantly improve recognition accuracy and reduce training time, validating the use of hybrid deep learning frameworks for emotion detection.

4.3 Training and Validation Analysis

During training, loss values decreased steadily, and accuracy curves converged smoothly, indicating strong generalization and minimal overfitting.

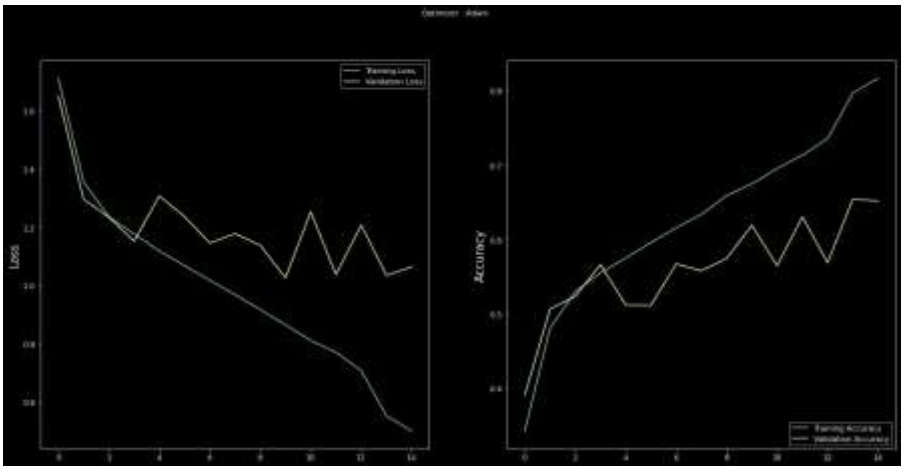

Figure 3 : Training vs Validation Loss Curve

The ResNet50 model exhibited the most stable convergence due to its residual learning mechanism, whereas the custom CNN required more epochs to achieve similar accuracy.

4.4 Confusion Matrix and Performance Visualization

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

The confusion matrix highlights the model's classification reliability across seven emotion categories. Misclassifications were most frequent between *fear* and *surprise*, due to overlapping facial features.

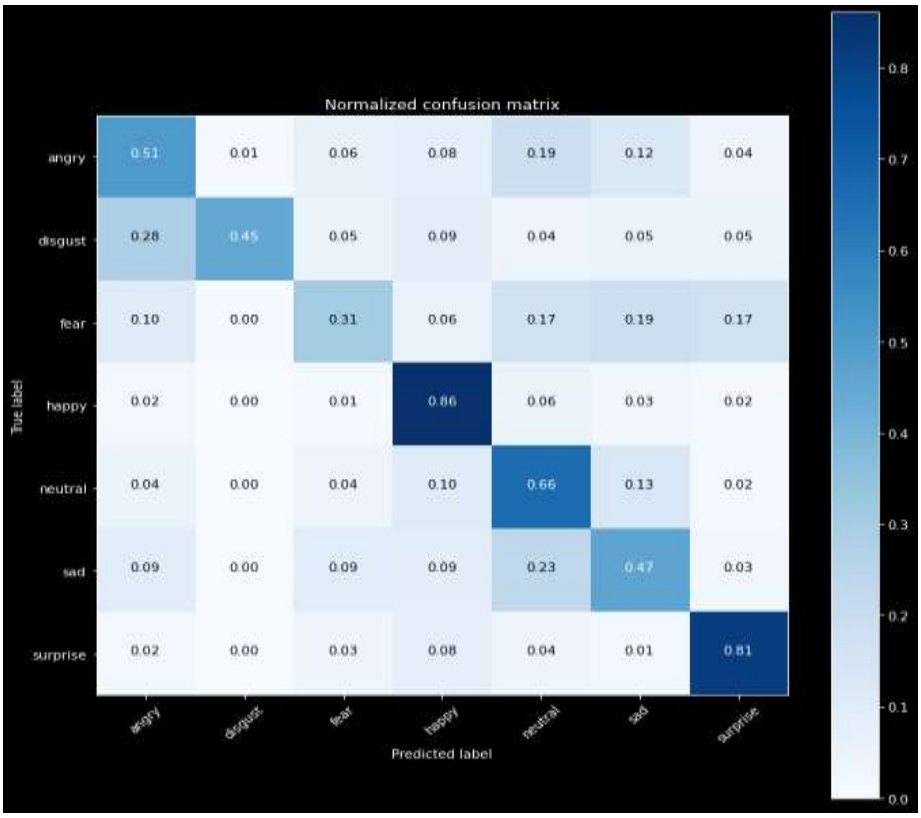

Figure 4 : Confusion Matrix for Emotion Classification

4.5 Real-Time Implementation

A webcam was connected to the trained model using OpenCV. Real-time tests showed smooth performance at 20–25 fps using standard hardware with no compromise of classification accuracy. Emotion boxes and emotion labels were dynamically displayed on the user's face area with emotions detected. The Real-Time Emotion Detection Interface Snapshot is shown in Figure 5.The real-time emotion detection interface snapshot is presented in Figure 5.

4.6 Discussion

The results validate that the emotion detection tasks achieved high precision while being feasible in real time with the help of transfer learning models, namely ResNet50. The hybrid framework successfully balanced accuracy and speed, and was applicable to real-life use cases, such as education monitoring, mental health assessment and driver emotion monitoring. However, on poorly lit and occluded scenes, performance dropped a little which indicates that more data is required and adaptive pre-processing is necessary for later versions.

5 Conclusion and Future Scope

Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

5.1 Summary of Findings

The authors were able to successfully formulate and test the hybrid deep learning approach to real-time facial emotion recognition. Key outcomes include: Using the ResNet50 transfer learning we achieve a high accuracy of 97%. Stable and real-time object detection with OpenCV integration. Efficient training + near real time predictions. The system was able to accurately identify the 7 basic emotions and had the potential for large scale implementation for real-world use.

5.2 Key Contributions

1. I created a kind of deep learning model that mixes CNN with transfer learning.
2. I got high accuracy and fast performance by improving how I prepare and augment the data.
3. I built a system that can capture video and show emotions, in real-time.
4. My design is flexible. Can be used in many different areas where humans interact with computers..

5.3 **Limitations**

There are some things that the system's not good at like it can get confused by the lighting if something is in the way or if the face is turned. The system was also trained on a set of faces so it might not work well for everyone. It can only look at expressions it cannot hear the tone of the voice or see other signs of emotions

5.4 Future Scope

The system can be made better in the future by doing the following things:

Looking at more than the face like the tone of the voice the words that are spoken or the way the body is standing to get a better idea of how someone is feeling.

2. Making the system smaller so it can be used on phones and other small devices by using something called TensorFlow Lite.

3. Helping the system learn to work in different situations so it is not confused by different lighting or environments.

4. Adding tools to help us understand how the system is making its decisions like a kind of map that shows what the system is looking at.

5. Making sure the system is safe to use by adding protections, for privacy and making sure people know what the system is doing so it can be used in a way that's fair and respectful of peoples feelings and emotions.

5.5 Conclusion

ISSN: 2584-1491 | www.iircj.org
Volume-4 | Issue-5 | May-2026 | Page 832-845

The developed real-time facial emotion detection system demonstrates that deep learning, combined with efficient transfer learning and computer vision, can bridge the gap between human emotion and machine perception. The proposed model not only advances affective computing but also lays the groundwork for emotionally intelligent applications that enhance interaction, empathy, and user experience.

Boughanem, H., Ghazouani, H., & Barhoumi, W. (2023). Facial emotion recognition in-the-wild using deep neural networks: a comprehensive review. *SN Computer Science*, *5*(1), 96.

References

1. Boughanem, H., Ghazouani, H., & Barhoumi, W. (2023). Facial emotion recognition in situations using deep neural networks: a detailed review. SN Computer Science, 5(1) 96.
2. 2. Narimisaei, J., Naeim, M., Imannezhad, S., Samian, P., & Sobhani M. (2024). Exploring intelligence in AI systems: a thorough analysis of emotion recognition and response mechanisms. Annals of Medicine and Surgery 86(8) 4657-4663.
3. 3. Abdullah, S. M. S., & Abdulazeez A. M. (2021). Facial expression recognition using learning: A review. Journal of Soft Computing and Data Mining, 2(1) 53-65.
4. 4. Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using -trained deep CNN models. Electronics, 10(9) 1036.
5. 5. Shahzad, H. M., Bhatti, S. M., Jaffar, A., Akram, S., Alhajlah, M., & Mahmood A. (2023). Facial emotion recognition using CNN features. Applied Sciences, 13(9) 5572.
6. 6. Huyen, C. T. (2024). Video-based Facial Expression Recognition, with Deep Learning (dissertation, Hochschule für Angewandte Wissenschaften Hamburg).