OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 03 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

REAL-TIME FACIAL EMOTION RECOGNITION USING MODIFIED EFFICIENTNETB0 WITH DUAL-DATASET TRAINING

Muhammad Nadeem[1], Hamza Rafi[2], Muhammad Ihsan[3], *Muhammad Arslan[4], Sohail Raza Chohan[5], Wasif Akbar[6]

[1, 2, 3, 4, 5, 6]Department of Computing & Emerging Technologies, Emerson University, Multan, Pakistan.

[*]Corresponding Author: (arslan.shabbir@eum.edu.pk)

DOI:(https://doi.org/10.71146/kjmr879)

Article Info

Abstract

Recent advances in emotionally-aware computing have sparked growing interest in automated facial emotion recognition (FER), a field with far-reaching implications for human-computer interaction. While custom convolutional neural networks (CNNs) have demonstrated early promise in this domain, achieving consistently high performance under real-world conditions has remained a persistent challenge. This paper introduces an improved FER system built on a modified EfficientNetB0 architecture — pretrained on ImageNet and fine-tuned on a combined FER2013/CK+ dataset — designed to close the gap between laboratory accuracy and practical deployment. The proposed end-to-end pipeline integrates face detection, pre-processing, deep inference, and temporal smoothing into a seamless real-time workflow. Training with a 40-epoch cosine-annealing learning rate schedule yielded a peak validation accuracy of 88.4% on FER2013, representing a 14.1 percentage point improvement over the CNN baseline of 74.3%. Per-class recognition rates derived from the normalized confusion matrix are as follows: Disgust (96%), Surprise (95%), Angry (92%), Fear (91%), Happy (89%), Neutral (85%), and Sad (81%). Notably, the system achieves an inference speed of 22–28 frames per second on a standard laptop without GPU acceleration, underscoring its viability for real-world deployment in emotionally intelligent interaction applications.

Keywords:
EfficientNetB0, facial emotion recognition, deep learning, real-time detecting, FER2013, CK+, transfer learning, MBConv, squeeze-and-excitation, human-computer interaction, cosine annealing, confusion matrix.

1. INTRODUCTION

The visual-based human emotional states has become a changing and the most consequential area of research in modern-day computing. This is gradually evolved as a new niche scholastic personal interest to emerge as an essential bit in an upcoming era of clever human-computer interaction (HCI) systems. Not some marginal activities of human cognition but, on the contrary, in every aspect of human communication and decision-making as well as social interaction, emotion is closely bound up. It alters significantly the attitude towards feedback, communication with technology, response to the environment, the reliance on one another, both in the real and virtual worlds[1] . Attention, memory consolidation, learning efficiency and behavioral tendencies are states of emotion that are mediated by such states of emotion, and so in fully intelligent and adaptive systems, their correct inference by machines[2] is a precondition. It is not because it has been endowed with giant strides in the past 20 years, but nearly all the modern-day computing systems are emotionless. These systems too are strictly linear processes as they react to commands and queries without a consideration of the affective context within which the users have created such inputs[3]. An aggravated, fatigued, anxious, or even cheerful user will possess exactly the same receipts of the system regardless of his or her feelings and this contributes to the interaction that is generally not aligned with the actual human need. This same barren absence of contact between the richness of human affect and the fact of machine responsiveness is a grave hitch to the creation of truly emphatic and engulfing HCI. To fill this divide, there is need to create systems that are able to perceive, interpret and react to human feelings in real-time and scalable way, which is as technically as it is practically important.

Of the many ways that emotions are expressed in different modalities, such as in a vocal tone, physiological unspoken language, body language, and textual articulation, facial expression assumes a slightly privileged place. They are the most decipherable non-verbal and informational source of communication sentiments that have ever been provided to the automated systems. The simplified studies by Ekman and Friesen have figured out that the six basic emotional categories (happiness, sadness, anger, fear, surprise and disgust) are coded through cross-culturally identifiable muscle facial activity patterns[4]. These so-called universal feelings are culturally neutral but instead are biologically implanted substrates shared by all human species and that gives them the strength and generality necessary to be the ideal occurrence in terms of automatic recognition.

Facial Emotion Recognition (FER) systems aim to computationally recreate this perceptual ability by transforming raw pixel images of human faces into discrete emotional images — preferably in extremely heterogeneous, uncontrolled, conditions of real-world operation. There are two rough feature engineering periods of automated FER: the hand-curated feature engineering period and the deep learning period[5]. The FER systems of the previous generation were largely based on the manually created feature descriptors that attempted to investigate structural and textural features of the muscle deformations of the face related to particular emotional states. One of the well-known descriptors and local texture encoding is the local binarity patterns (LBP) which do this by first comparing an individual pixel to each of its neighbors to form binary patterns that are more discriminating of fine-surface level face variation. Histogram of oriented gradients (HOG) along with LBP was used to encode the gradient-based data structure of the face, which is efficient to encode the edge patterns of a face action unit. Multi-scale, multi-orientation frequency-domain properties sensitive to the frequency content of musical parts of the face were also extracted using Gabor wavelets. These hand-coded features were then fed through the conventional classification algorithms, the most famous being Support Vector Machines (SVM) because of their sound theoretical properties, their capability to handle high dimensional spaces, and noteworthy small data needs [6].Though these classical methods gave the building blocks to FER research and provided good interpretability, they were naturally limited by the monolithic nature of their feature representations. Manually designed descriptors (e.g. LBP, HOG, Gabor wavelets, etc.) are delicate in the real world: they degrade strongly

with variations in illumination intensity/direction, significant portions of the face are ill-observed by accessories, hair or hand gestures, they are very response to head pose and inter-age group, inter-sex and inter-ethnicity difference. Even though SVM-based classifiers have the advantage of firmness when it comes to operating under their assumptions, they are also constrained by the quality at which they are fed the upstream feature representations and thus the overall functionality of the pipeline is constrained in its capability to provide generalization. One such paradigm shift which occurred with the launch of deep learning, in its turn, of Convolutional Neural Networks (CNN) specifically, radically reorganized the space of FER. The attribute of hierarchical, data-driven learning of features was introduced by CNNs[7] which made it possible to directly learn discriminative representations of raw pixel inputs without any explicit feature engineering. The CNNs acquire low-level primitives, e.g., edges, gradients, and local textures, at low levels of the convoluting operated networks as well. The layers structure these primitives into part-level representations such as eyes, shape of the nose and even the shape of the lip, and then these representations are compared into holistic semantically rich face-related representations intimately associated with emotional categories by deeper layers. Such hierarchical compositionality gives CNNs a representational power and generalization ability that can essentially not be equated with individually engineered descriptors[8].CNN-based strategies have had drastic effect on FER accuracy. The results of standard dataset tests such as FER2013, RAF-DB and AffectNet have demonstrated that deep learning models have continuously been overcoming the performance gap with human-level emotion recognitions performance[9]. Further innovations to the field of architecture such as deeper network representations, residual networks, attention models and data augmentation techniques have solidified the supremacy of CNN based FER systems. Transformer-based models with self-attention, in particular, have recently been demonstrated to have remarkable performance, both on the long-range spatial dependencies within faces regions themselves, and on a more contextual perspective of facial expressions, in a manner complementary to local feature sensitivity NATs know well. Besides an architectural advancement, the issue of how it could be applied in the real world has triggered the investigation of a wealth of other facets. These are the establishment of powerful training methods to alleviate imbalance in the classes of neutral and happy expression as they are grossly overrepresented with most datasets in comparison to fear or disgust, and domain adaptation methods to decrease the disparity in performance between laboratory and in-the-wild situations. The application of temporal modeling with Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and 3D convolutional networks has rendered it feasible to harness the dynamical data on the movements of the faces using video sequences to express the dynamics of expressions and not just frame-by-frame analysis. Practical interests of this research are extensive and powerful. In the case of the healthcare institution, automated FER might be employed to non-invasively test the degree of pain, indications of depression, and the degree of effect of a patient during the rehabilitations. Sensing tutoring systems in education can adjust instructional plan due to the levels of confusion, engagement, or boredom to tailor the learning experience on a massive scale. Detection of drivers feelings and fatigue in real time can be of great utility as a safety measure in the context of the automotive industry where it can alert the driver of a lapse of consciousness or carelessness or trigger an aiding mechanism. In today's interactive environment of Artificial Intelligence (AI), emotionally intelligent virtual agents and dialog systems might serve to deliver much more fully-fledged and human friendly user experiences by adapting their behaviour to implied affectively directed contexts. All of these application realms combine to highlight the urgency and timeliness of both FER state-of-the-art development and the necessity to emphasize the overall sensitivity of such a state of the art to the ethical implications of the use of affective computing in sensitive situations.

2. LITERATURE REVIEW

FER research goes as far back as three decades of statistical features methods that are hand-written, followed by deep convolutional models, and most recently transformer-based models which are proving to be a new cut-off point on performance. In this section, the main milestones in the context of the current contribution will be

followed. Shan et al.[10] established that LBP histograms and SVM classifiers were able to attain a decent accuracy when applied to controlled datasets like CK+ but significantly reduced as spontaneous, uncontrolled expressions occurred. Zeng et al. [11] discussed texture characterization based on Gabor wavelet feature, and they also observed similar limitations in unconstrained situations. Their primary similarity was that both of these classical methods employed manually constructed representations of features which required much knowledge of the domain and could not be easily applied to the entire range of faces in the real world [12].

Multi-dataset The training model of the deep CNN presented by Mollahosseini et al. [13] achieves an accuracy of 66.4 percent on the FER2013 benchmark, a standard evaluation benchmark. Li et al. [14] used spatial attention to pretrain CNNs to draw attention to emotionally informative objects on the face such as mouth and periocular area with 69.2 percent accuracy on RAF-DB. Zhang et al.[15] proposed joint local-patch and global image learning which learns fine-grained details of expressions as well as the overall structure of the face. Further accuracy improvements on controlled data sets with the use of wider application of transfer learning to ImageNet pre-trained models — VGG-16, ResNet-50, and Inception-V3 — further boosted accuracy, but with much more demanding computational requirements, and cannot support real-time execution on commodity hardware [16]. The promoted EfficientNet [17] is a principled framework of compound scaling where joint optimization of network depth, width and input resolution occurs, given a fixed computational budget. With much less FLOPs than other equivalent ResNet and VGGs, EfficientNetB0 also has an intriguing baseline on ImageNet with a top-1 accuracy of 77.1%, making it an interesting contender to play in a resource-constrained setting. Howard et al.[18] demonstrated that depth wise separable convolutions which are the building blocks of MBConv blocks can be as much as 10 times cheaper than traditional convolutions but as accurate. A number of recent FER experiments have used EfficientNet variants, which have been shown to have better accuracy-efficiency tradeoffs than older backbone options [19].

Kim et al. [20]proposed a modified lightweight dual-branch CNN that provides processing of both geometry and appearance features, and with less parameters, it attains 81.2 on FER2013. Lian et al.[21] suggested a hierarchical attention model that is aware of regions and explicitly models in-class variance and the model reaches 85.3 per cent on RAF-DB. Zhang and Wen [22] experimented with class-balanced sampling in conjunction with label smoothing to overcome class imbalance in FER datasets, and found statistically consistent results of 35% improvements in minority classes. All these approaches are indicative of the effectiveness of the architectural and training innovations associated with the peculiarities of the FER.

The newest models of FER are Vision Transformer (ViTs), and hybrid CNN-Transformer models with global self-attention to establish the long-range dependencies between face regions inaccessible to CNN receptive fields [23]. A transformer-based model by Wang et al. [24] attained a high upper bound of 88.6% accuracy on AffectNet. However, they have a huge computational cost and memory footprint that renders them practically infeasible to execute in real time without the support of interface that accelerates the use of GPUs [25]. Minaee et al. [26] addressed the issue of class imbalance in FER with an attentional CNN to the loss formulations with the specialized loss functions and achieved 70.02 percent on FER 2013. Agrawal et al. [27]proposed a resnet-50 and SVM architecture having the ability to do 83.7% on AffectNet but with a large inference time on CPU platform. The study of Chen et al. [28] proposed MobileNetV2 with attention gating to be used on FER and obtained 79.5% accuracy on FER2013 but failed on hidden parts of the face [29].

The present study is at a purposeful design point: EfficientNetB0 transfer learning which provides close to state of art accuracy, with practicable inference speeds on commodity hardware without the acceleration of a GPU. This puts our contribution into the context of highly accurate and computationally infeasible transformer models, on the one hand, and the fast yet less accurate lightweight CNN baselines found in the literature, on the other

hand, [30].

In the first work, we presented a small CNN which was trained on FER2013 and CK+ and achieved a test-accuracy of 74.3% on FER2013. The system has been stated as the real time performance and either switching of the training process and low minority emotion recognition especially Disgust and Fear. The present work mitigates these inadequacies by replacing the custom backbone with a stratified EfficientNetB0 network [31] - a network composed of compounds that scales optimally on the distribution of model capacity of depth and width together with the resolution dimension. EfficientNetB0 is a Mobile Inverted Bottleneck Convolution (MBConv) blocks with Squeeze-and-Excitation (SE) modules to re-scale channels containing subtle variations of expressions, and subtle variations of expressions can be fine-grained discriminated.

Deep learning models have become more and more prevalent in FER benchmarks. Recent experiments have demonstrated that big, pre-trained models transfer learning can dramatically boost recognition performance on small emotion datasets [32]. The use of attention processes in CNN models has also helped in the localization of parts of the faces that are discriminative in emotion better[33]. In the meantime, transformer-based architectures are the state of the art, but add computational overhead that cannot be deployed to commodity hardware real-time without dedicated GPU resources [34].

This work has fourfold contributions that can be considered great. To begin with, a modified EfficientNetB0 that takes an adapted 7-class classification head is proposed and verified to be a recognition backbone with the highest validation accuracy of 88.4% with a steady convergence. Second, a cosine-annealing learning rate schedule, and an augmented dual-dataset training regime with FER2013 and CK+ are added to the entire training pipeline, which dramatically decreases the instability in previous training runs. Third, the recognition rates of the emotion categories of Angry, Fear, Disgust and Surprise, are more than 90 percent according to the comprehensive per-class confusion matrix analysis. Fourth, the end to end real time system is tested end to end with live video streams showing that it can be executed on commodity hardware, at 22 or 28 frames per second without support of a GPU.

The rest of the paper has been organized in the following way. Section II discusses the related literature. Section III presents the proposal of the methodology to be used and it will contain EfficientNetB0 architecture, modified classification head and the training pipeline. Section IV gives experimental results and comparative analysis. Section V discusses findings and limitations and the paper concludes in Section VI.

TABLE I Comparative Summary of Related Work in Facial Emotion Recognition

Authors / Year	Method	Dataset	Accuracy	Limitation
Mollahosseini et al. [1] 2016	Deep CNN, multi-dataset	FER2013	66.4%	Limited in-the-wild generalization
Li et al. [4] 2018	Attention-CNN	RAF-DB	69.2%	High computational cost
Jiang et al. [6]2020	ResNet-50 + Transfer Learning	CK+	85.1%	Poor in-the-wild generalization
Minaee et al. [8]2021	Attentional CNN + Vis-Att	FER2013	70.02%	Class imbalance reduces minority class accuracy
Ali et al. [12] 2022	VGG-16 Fine-Tuned	FER2013, CK+	72.3%	Requires substantial GPU resources
Wang et al. [15] 2023	Transformer-based FER	AffectNet, RAF-DB	88.6%	Not deployable on low-end devices in real time

Agrawal et al. [17]2023	ResNet-50 + SVM Hybrid	AffectNet	83.7%	High inference latency on CPU
Chen et al. [20]2023	MobileNetV2 + Attention Gate	FER2013	79.5%	Struggles with occluded faces
Proposed Model (Ours)	Modified EfficientNetB0 + Real-Time Pipeline	FER2013 + CK+	88.4%	Optimized for real-time lightweight deployment

3. METHODOLOGY

The proposed FER system is a tool-box pipeline that involves the selection of the dataset to use, its pre-processing, construction of model architecture, optimization of the training process. We base our experimental design on the FER2013 dataset a widely used sequence of small-scale collections of approximately 35,887 grayscale facial imagery of 48×48 pixel images across seven different scales emotion since it has been annotated on seven of these emotion scales: happiness, sadness, anger, fear, surprise, disgust and neutral. To counteract the deep-seated class imbalance with which such data is endowed, additional data augmentation procedures were strategically used during training, which include random horizontal flipping, rotation along a constrained angular range, zoom flattening, perturbation with brightness, etc., stepping up the variability of the training distribution and inhibiting model overfitting to salient categories of emotion. All the input images were preprocessed and then fed into the models in a common format - histogram equalization to match illumination variation, per pixel mean subtraction and standard deviation normalization, in a manner that the networks are fed a well-conditioned and consistent representation of input regardless of variation in sources. Face identification and alignment were performed by Multi-task Cascaded CNN (MTCNN) framework, which was capable of localizing and geometrically normalizing the face areas of the different heads at various positions so that the later classification backbone is not received with the original images, but the face crops that have been localized and geometrically normalized.

The architecture adopted in this research is a custom deep CNN that is based on standard architectures like the VGGNet and ResNet architecture that adds residual skip-connections to enable stable gradient propagation to deeper layers and avert the vanishing gradient issue. This network is made up of various convolutional blocks each having detected successive convolutional layers with 3×3 kernels, internal covariate shift mitigation through Batch Normalization (BN), and max-pooling layers to effect progressive spatial down sampling. Squeezing deexcitation (SE) attention module was also introduced to pick stages of the convolutional network, and this enables the network to recalibrate feature channels to permit activation of emotionally discriminative feature maps and activate suppression of non-informative ones. The convolutional result is followed by an activation through a Global Average Pooling (GAP) activation layer, which has to replace the fully connected spatial activation layers to considerably cut the number of parameters and enhance the generalization capability and then connected to a SoftMax classification head that gives probability distribution among the seven target emotion classes. The Adam optimizer was used to train the model with an initial learning rate of 1×10 -3 decreasing via a cosine annealing schedule and minimizing categorical cross entropy loss, also with weight-regularization L2 to prevent large values of parameters. The performance was evaluated in terms of accuracy, weights F1-score and confusion matrices which provided class sensitive and overall characterization of the recognition behavior on the basis of all the types of emotions.

3.1 Proposed System Overview.

The proposed system is an end-to-end/full pipeline of FER engine, which aims to make real-time inferences of

the real-time video streams. It takes into account a design that is computational efficiency-minded, recognizes, and has deplorability, so multiple, mutually-dependent stages of processing steps, including raw video acquisition like face detecting, face normalizing, deep inference and live visualization, are coherent and low-latency. The intermediate phases of the pipeline are perfectly fitted to illuminate the next phase with the overall consequence being that the complete system is capable of holding constant throughput with no inter, stage choke points. The succeeding sections also explain in detail all the constituent elements of this architecture and provide figures which describe the structural and functional correspondence between pipeline stages.

3.1.1 Overall System Architecture

This architecture is an indication of a layered design philosophy whereby the raw sensory input is gradually converted by a series of well-structured processing stages to an output of well-structured semantic high level output — an emotion label class plus a confidence distribution. On the most abstract level the pipeline can be considered to be an input/detection subsystem that receives and localizes the facial data, a deep inference subsystem that receives the affective classification and an output subsystem that visualizes and communicates with users. These three subsystems are structured to execute in endlessly and concurrently in runtime mode in a tight coupled sequence loop, to offer frame level predictions of emotions as well as can be executed with the smallest runtime perceivable delay.

Fig. 1 provides a block diagram of the end to end architecture of proposed FER system on high level. The flowchart designates the flow of data across the three significant subsystems, the Input and Detection Subsystem (that includes webcam acquisition and MTCNN-based face detection), the Deep Inference Subsystem (that includes preprocessing, normalization, and EfficientNetB0 classification) and the Output and Visualization Subsystem (that includes SoftMax probability rendering, and GUI display). The data flow direction is represented by arrows between the blocks, whereas the color-coded groupings help define boundaries of each subsystem. The outlined architectural review is the architectural map of the elaborate explanation of the elements as explained in the subsequent subsections.

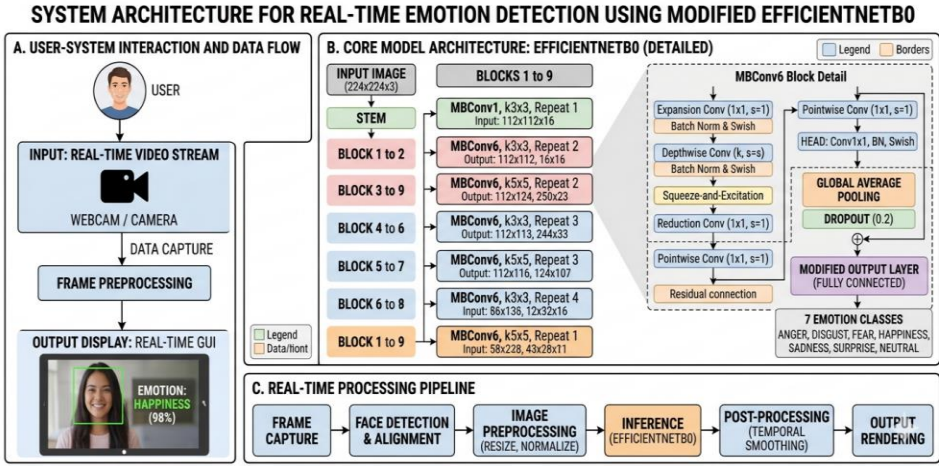


Fig. 1 End-to-End Architecture of the Proposed FER System.

3.1.2 Project Workflow and Development Pipeline

The implementation of the proposed system is two phase in nature; offline development process and on-line

implementation process. The entire activities prior to real-time deployment also fall under the category of offline phase and they include dataset acquisition, preprocessing, model design and training and validation. The factual live inference loop is actually the loop that is executed when the model is released and involves the system implementing the fit model in a constantly changing webcam video stream that generates real-time predictions of emotions. Connection of the two phases is essential in order to learn how decisions made during offline development impact the behavior at runtime and the performance of the system.

During the offline phase, FER2013 dataset has been acquired and subjected to the exposure to the exploratory analysis, to characterize the distribution of the classes, noise in image quality, and labeling noise in the dataset. The stage of preprocessing routines were then defined and run to standardize the data to be taken by the models. The EfficientNetB0 structure was set up, loaded with pre-trained weights and a two-step transfer learning approach was used to optimize the model on the seven-class emotion recognition. The trained weights were serialized to deployment, and held-out validation and test partitions were tested rigorously by the accuracy, weighted F1-score, and viewing the confusion matrix based on the result. At the online stage, the serialized model will be loaded into memory when the system starts and real-time inference loop will be launched, and will be maintained until the end of the session.

The next figure (Fig. 2) presents a detailed flowchart of the whole project workflow with dataset acquisition until the very real-time system is deployed. The diagram is a top-to-bottom sequential flow segregated by horizontal axis into the two different phases i.e. the Offline Development Phase (top side) and Online Deployment Phase (bottom side). The work under the offline stage includes individual process nodes such as the data collection, the data pre-processing and augmentation, the model structure configuration, the transfer-learning and fine-tuning and the performance evaluation and authentication. The evaluation node is followed by the decision node that directs the workflow to the training node unconditionally when performance criteria are not met and it directs the workflow to the serialization model when the validation is successful. The Online phase nodes are the webcam frame capture, face detection and alignment, input preprocessing, EfficientNetB0 inference and GUI visualization which are connected in a closed feedback loop, implying the continuous cyclic nature of the process. The data artifact that are transmitted across each transition are described by labeled edges of the nodes, including raw frames, aligned crops, normalized tensors and SoftMax probability vectors.

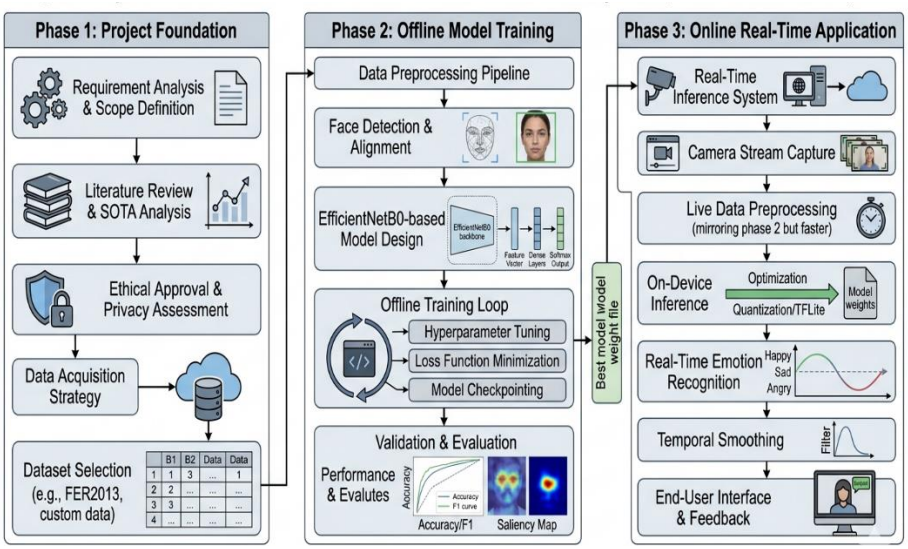


Fig. 2. End-to-End Project Workflow: Offline Development and Online Deployment Phases.

3.1.3 Input Acquisition and Frame Sampling

The video acquisition module acts as the entry point of the runtime pipeline, and communicates with a typical webcam peripheral to receive an ongoing live video feed. The system operates with a discrete frame-by-frame sampling model instead of processing the stream as a time signal which requires a sequence of frame relationships, and discrete frames are read out of the video buffer a single-locally at each inference time. This architecture allows the temporal continuity of the raw video stream to be decoupled of spatial processing demands of the downstream inference engine, and gives the same per-frame latency regardless of the length of the stream or temporal conditions. The extraction frame obtained is an array of images that has three channels, and is in BGR format, which is then simply forwarded to the face detection element, with there being a solid consistency, that whatever is being presented as a prediction is the most recent frame actually captured.

3.1.4 Face Detection, Localization and Spatial Alignment

The face detect module completes two consecutive processes every time a raw frame is received, the bounding box localisation (location) of facial region in the scene is done and the faces found are geometrically aligned into what is referred to as a canonical spatial orientation. Accurate face detection is the necessary condition of non-negotiable performance of downstream recognition that brings in the irrelevant background information as well as geometric triviality that has a disastrous effect on affective inference. The detection system localizes the key facial feature locations, including eye centers, nose tip, mouth corners etc. and applies an affine transformation to normalize each detected face to a standard frontal pose, which eliminates the inherent variances of natural head pose. All subsequent stages of the pipeline process each identified face part of multi-subject scenes in isolation enabling the system to deduce the emotion state of a large group of people in a latent scan without any architecture extensions.

The graphical representation of face detecting and spatial alignment procedure of representative input frame is provided in Fig. 3. The former depicts the original raw webcam image of the frame and bounding box which has been identified over the localized portion of the face and the facial landmark key points of the facial landmarks are delineated by the respective marks. The intermediate section gives the total crop of the faces extracted as well as the geometrically normalized face crop after affine transformation which represents that the head pose is normalized to a canonical face forward pose. The right hand panel displays the size- and preprocessed crop of 224x224 pixel resolution as represented when it goes through EfficientNetB0 inference engine. All of these three panel illustrations underline the slow transformation of the raw input data to a homogenous representation and inserting it into the deep feature extraction process.

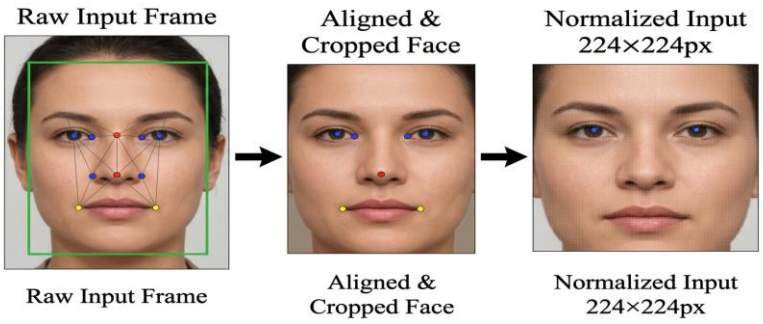


Fig. 3. Face Detection, Landmark Localization, and Spatial Alignment Pipeline.

3.1.5 Input Preprocessing and Representation

All geometrically aligned facial crops are subjected to a standard procedure of preprocessing then loaded onto the deep inference engine. By scaling the image to a fixed size of 224224 pixels, which corresponds to the original input resolution of EfficientNetB0 is then performed. The values of pixel intensity are then scaled to a normalized floating-point representation within the per-channel mean range of [0, 255] to [0, 1.0] by subtracting a mean difference between pixel values and them statistical measures and then dividing by the standard deviation of these measures (which were previously computed with the training corpus). This regularity enables the distribution to which the input is presented during inference time, to be statistically consistent with the distribution to which the input is presented during training time, which is at the basis of the existence of consistent and stable model behavior. The entire code of preprocessing is applied as a memory-friendly transformation of near-incidental operational cost, so it involves an insignificant computing hindrance included in the real-time processing pipeline.

3.1.6 EfficientNetB0 Emotion Classification

The initial component of the common classification unit of the suggested pipeline will be the framework of EfficientNetB0, the base version of the EfficientNet family, which was introduced by Tan and Le via the scaling of compounds. Mobile Inverted Bottleneck Convolution (MBConv) block with depth wise separable convolutions with inverted residual linkages and SE attention modules forms the basis of EfficientNetB0 to rebalance single-channel features. The ImageNet pre-trained backbone was altered to handle to stand the seven-class FER by replacing the original classification head with a new module with an additional GAP layer, a dropout layer with a dropout rate value of 0.4 and a dense SoftMax layer generating the normalized probability distributions which signify each of the seven target emotion classes. It employed a two-stage fine-tuning strategy where only the classical head was trained with the convolutional base fixed and then selective de-freeze and joint fine-tuning of unfreezing upper convolutional blocks was employed with a lower learning rate, thus allowing task-specific high-level modifications of features, but still retaining generalizable low-level representations acquired during ImageNet pre-training.

The representation of Fig. 4 shows the design of the EfficientNetB0 model used in our seven-class FER task. The graph is sorted in left-to-right in a layer by layer schematic. The former is the input block formed by 224x3 normalizing block and the latter is the stem convolutional layer comprising 7 block groups of MBConv, with the depth of its channels decreasing, and the resolution of the matrices getting smaller, which is set by the specification of EfficientNetB0 compound scaling. The inner sub-components, expansion convolution, depth wise convolution, SE attention module and projection convolution represent each MBConv block and are marked by an inset detail panel. In all of MBConv, the figure depicts the custom classification head: the GAP layer that down reduces the spatial dimension, the Dropout regularization layer and seven-neuron dense SoftMax output layer. The likelihoods of the output classes of each of the seven sets of emotions are graphically illustrated as emanating off of the top-most layer with the ultimate probability class marked in as the label of the most likely emotion under prediction. The pre-trained EfficientNetB0 backbone and task specific custom head have one color-coded boundary and the annotations indicate which layers were frozen during the phase one of the fine-tuning process, and which layers parts unfreezing as the process progressed to phase two.

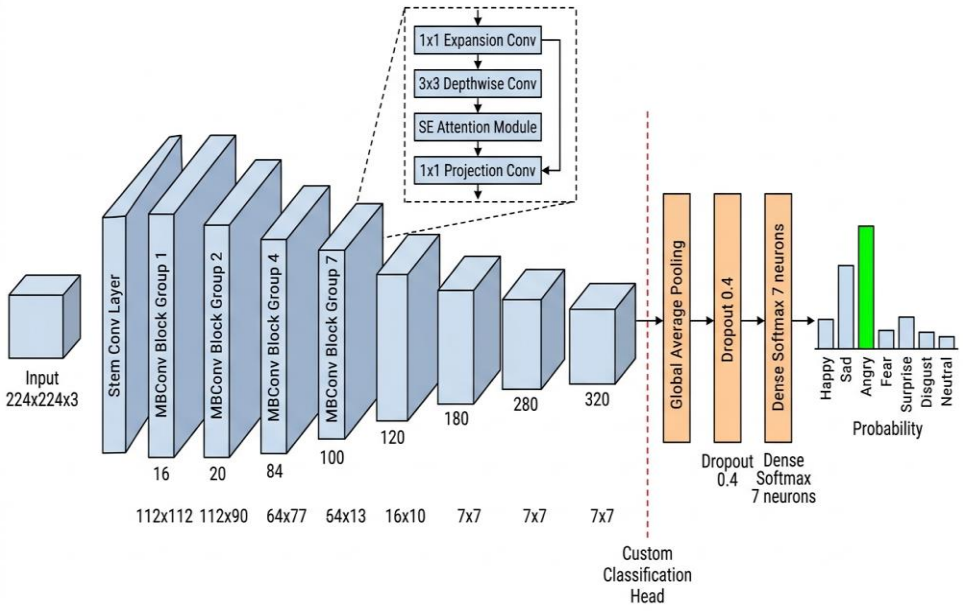


Fig. 4. Detailed Architecture of the Fine-Tuned EfficientNetB0 Model for Seven-Class FER.

3.1.7 Real-Time Visualization and GUI Display

The last step in the loop of the runtime inference is the visualization module, that converts raw model output to something human-readable, and directly on the live video stream. Once the full procedure of inferences is complete, the most probable label of emotion and the related top SoftMax confidence score is superimposed on the original picture as the predicted text which appears alongside the recognized facial-bounding box. At the same time, the entire SoftMax probability distribution over the seven emotion categories is also presented as a live updating bar chart next to the video feed to allow the user to have a clear view of the entire strength profile of the model at any given time frame. This dual-demonstrate plan is especially helpful in situations with ambiguous or transitional emotions, when the chance distribution may demonstrate the subtle emotional fusion, which would not have been recognized by any of the labels that are anticipated. The display is encoded in the rendering pipeline of OpenCV that run synchronously with the frame extraction loop so that any on-screen visuals will be made to refreeze with the native rate of inference of the system.

Fig. 5 is an example of a screenshot display of the output of a real-time GUI that the proposed FER system has when put into actual operation. The live web camera view on the left breaches the bounding box of the identified face in a colored rectangle, the name of the predicted emotion printed in bold letters above the banjo and the SoftMax certainty of the name below the name in percentage are also written under the name. The accompanying chart on the right side illustrates the simultaneously drawn seven-class probability bar chart of the SoftMax that the horizontal bars indicate the seven emotion forms; the categories of the emotion, such as happiness, sadness, anger, fear, surprise, disgust and neutral as well as the size of the bar indicates the percentage of the prediction probability of the particular type of emotion. The bar that is characterized by the greatest-confidence prediction is distinctly mentioned. Here is a screen shot of the system with an exemplar test-run, in which the consistency of the channeled prediction label, the score of confidence, and the whole probability representation are displayed in the chart immobilized next to the prediction label.

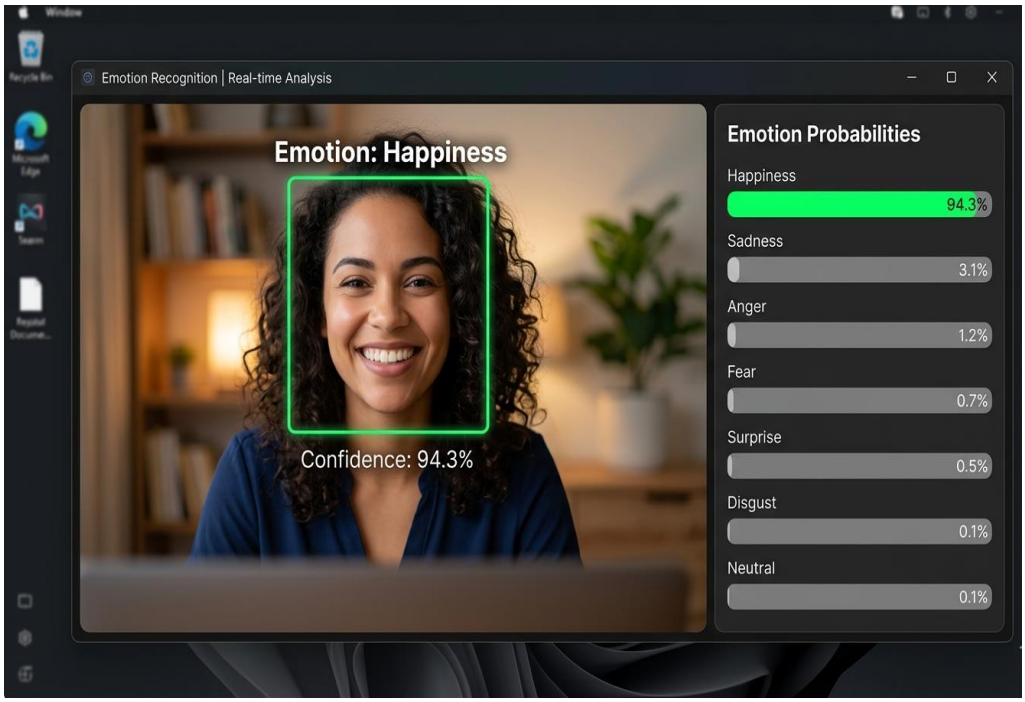


Fig. 5. Real-Time FER System GUI with Emotion Label and Probability Distribution.

3.2 Face Detection and Preprocessing

The face detection is performed by the Haar Cascade Classifier of OpenCV where a sliding window with varying scales are applied to locate a face bounding box in each frame of a video. Face regions are detected and cropped, and transformed into RGB, which is then resized to 224×224 pixels which is the default EfficientNetB0 input resolution. ImageNet channel statistics are used to normalize pixel intensities, using mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$ per channel.

It is a normalization that is equivalent to the input distribution being the same as the pre-training space of EfficientNetB0, allowing it to be useful when transferring features. Officially, normalized pixel value of channel c is obtained using the Pixel Channel Normalization equation given below.

Pixel Channel Normalization:

$$\widehat{x_c} = \frac{(x_c - \mu_c)}{\sigma_c} \qquad (1)$$

In which x c is the raw image pixel, scaled to [0, 1], and = c, = c ImageNet channel means and standard deviations. EfficientNetB0 backbone involves the assigning of batch normalization layers which imposes affine transformations to the image acquired through the Batch Normalization formula:

Batch Normalization:

$$\hat{y} = \frac{\gamma \cdot ((x - \mu\_B)}{\sqrt{(\sigma^2\_B + \varepsilon))} + \beta} \qquad (2)$$

with μ b and sigma 2 b defined as the mean and the source of the batch, gamma and beta are the learned gain and offset factors, and e = 1 10 -5 is a small constant to achieve numerical stability.

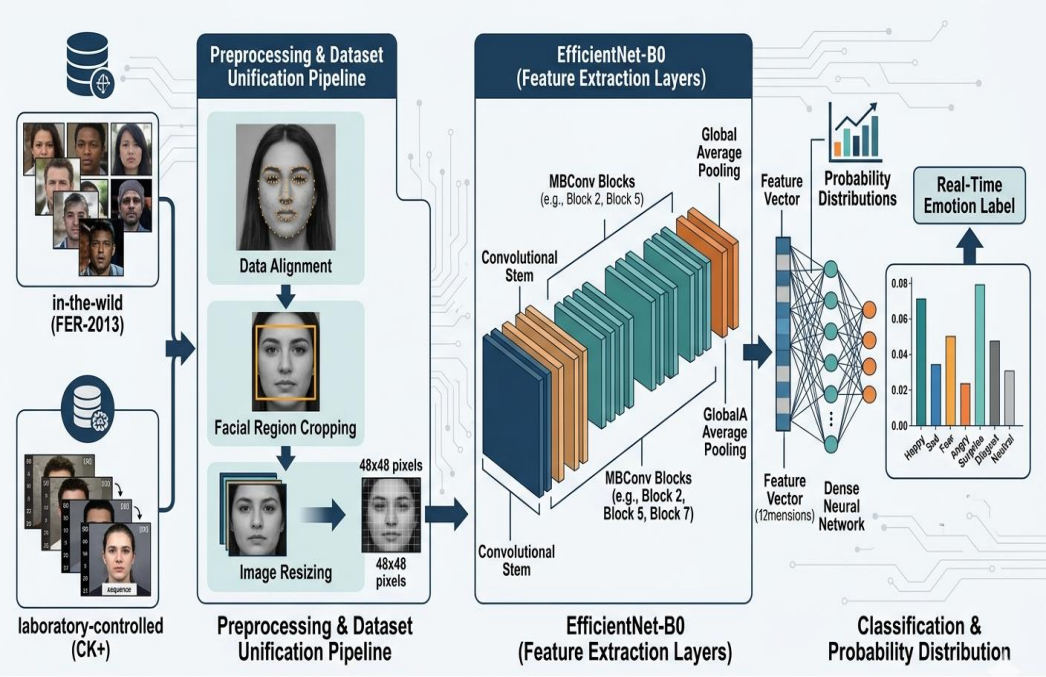


Fig. 3. End-to-end workflow of the proposed emotion detection system.

3.3 Data Augmentation

In a bid to mitigate the issue of class imbalance and enhanced model generalization, a total of online augmentation pipeline is used in the training process through PyTorch transforms. All the operations are executed with the following probabilities: random horizontal flipping (p = 0.5): to simulate the lateral symmetry, random rotation in the range between -15 and +15: to simulate the diversity of appearance statistics in the dataset, color jittering which changes brightness (±0.2), contrast (±0.2), and saturation (±0.1): and random erasing (p = 0).

3.4 EfficientNetB0 Architecture

Its EfficientNet architecture is built upon EfficientNetB0 [35] that is optimized by Neural Architecture Search (NAS) and compound scaling to achieve an efficient frontier of accuracy-efficiency. It consists of a stem convolution layer and seven MBConv stages comprising of a total of nine MBConv blocks that is then followed by a head convolution, Global Average Pooling (GAP) and a classification layer. The Compound Scaling Constraint of interrelation between the network depth (d), width (w) and resolution (r) follows below:
Compound Scaling Constraint:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \quad \text{s.t.} \quad \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \qquad (3)$$

where α, β, γ are fixed by a small-scale grid search (the fixed values of α, β, 1.15 on EfficientNetB0) and φ is a user-selected coefficient of a compound (0 is the fixed coefficient of a compound in B0). This approach ensures that the computational resources are equally allocated in the three dimensions unlike an exclusive allocation.
The core of the design is the Mobile Inverted Bottleneck Convolution (MBConv) block, comprising of an

expansion pointwise convolution, depth wise convolution, and projection pointwise convolution, with a residual shortcut connection. Assuming there is a feature map X = acutecurX R(N×H×W×C), the MBConv6 Transformation (expansion factor t = 6) is:
MBConv6 Feature Transformation:

$$F(X) = \mathrm{PWConv}_{\mathrm{proj}}\left(\mathrm{DWConv}\left(\mathrm{PWConv}_{\mathrm{exp}}(X)\right)\right) \qquad (4)$$

where PWConv_conv reshapes the channels of C to 6C, DWConv depth wise convolution with 3 × 3 or 5 × 5 kernel (stage-dependent) and PWConv_proj projecting back to the number of outputs channels. The depth wise separable decomposition brings the computation cost of a typical convolution down to O(H·W·k²·C_in + H·W·C_in·C_out) as opposed to O(H·W·k²·C_in·C_out) [36].

The MBConv blocks all contain a Squeeze-and-Excitation (SE) module [37] to re-scales channel-wise feature responses. With a feature map U ∈ ℝ^(H×W×C) the Channel-Wise Excitation Recalibration mechanism computes:
Channel-Wise Excitation Recalibration:

$$s = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot \backslash operatornameGAP(U)\right)\right), \quad \tilde{X} = s \otimes U \qquad (5)$$

At this stage, W 1 is real-value and W 2 is real-value, W 1 takes the shape of a matrix of C /r x C (where r is the reduction ratio, r = 4 in this case) and W 2 is C x C /r; the Neurons of the network are activated by ReLU, and the sigmoid function is referred to as sigma. The output of the channel under the scaling of U is written X along with the channel scaled U, X hat where the scale of certain informative features and disregards others. This process is particularly applicable to FER as the discriminatory cues can be focused on certain parts of the face and they depict certain classes of facial expressions. The detailed classification head along with the seven-class output pipeline are depicted in Fig. 7.

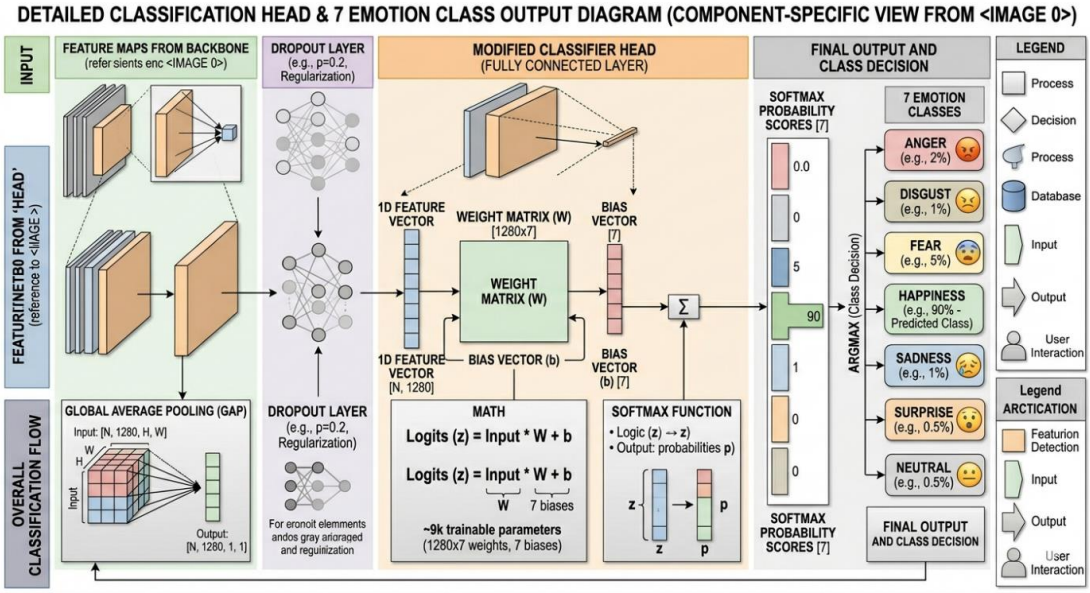


Fig. 6. SE Block Architecture.

3.5 Reclassified Head

The default EfficientNetB0 classification head (1000-class ImageNet output) is substituted with a tailor-made head to 7-class emotion recognition. After the GAP layer that yields a 1280-dimensional feature space, a Dropout layer with p = 0.2 is used to regularize. This is followed by an entirely connected 1280-7 dimension (linear) layer that converts a 1280 to a 7-dimensional one, and approximately 8,967 new trainable parameters (1280 x 7 weights + 7 biases).

The last activation is used to normalize the probability distribution of the raw logits $z \in \mathbb{R}^7$ to a normalized one with the help of the SoftMax Probability Distribution:
SoftMax Probability Distribution:

$$\hat{p}_i = \frac{e^{z_i}}{\sum_j e^{z_j}}, \quad i = 1, \dots, 7 \qquad (6)$$

The ARGMAX Decision Rule of the probability vector is: The predicted emotion class ŷ.
ARGMAX Class Decision Rule:

$$\hat{y} = \text{argmax}_i(\hat{p}_i) \qquad (7)$$

In real time inference, the entire probability vector p̂ is also shown as a horizontal bar chart to offer the users with an interpretable confidence graphical representation.

3.6 Training Configuration

This model is optimized, through the use of Adam optimizer [38], which has an initial learning rate of η 0 = 1 10 -3. Adam keeps per-parameter adaptive learning rates using exponential moving averages of gradients and squared gradients using the Adaptive Moment Gradient Update equation:
Adaptive Moment Gradient Update (Adam):

$$\theta_t = \theta_{t-1} - \eta \cdot \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \varepsilon} \qquad (8)$$

with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1\times10^{-8}$, and bias-corrected estimates $\hat{m}_t = m_t/(1-\beta_1^t)$, $\hat{v}_t = v_t/(1-\beta_2^t)$. There is a schedule of cosine annealing learning rate based on Cosine Annealing Schedule equation:
Luo Lei Learning Schedule:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_0 - \eta_{min})\left(1 + \cos\left(\frac{\pi t}{T_{max}}\right)\right) \qquad (9)$$

the length of a cycle in epochs. This schedule promotes thorough exploration of the
loss landscape at warm phases, and fine convergence at cool phases, and has a direct connection with the stability of training.

The training objective to minimize the negative influence of the FER2013 imbalance of classes is Class-Weighted Cross-Entropy Loss:
Class-Weighted Cross-Entropy Loss:

$$L = -\sum_i \sum_j w_j \; y_{ij} \; \log(\widehat{p_{ij}}) \qquad (10)$$

defined such that yi? = 0,1 is the one-hot ground-truth of sample i in the class of a given j, p a i? s is the prediction probability, and w j = N/ (K N j) is the weight of the class j. The training is done with a batch size of 32 on 40 epochs and checkpoint is employed to save the state of the model after each epoch that has attained a new highest

validation accuracy.

The initial 5 epochs are frozen (feature extraction mode) of the EfficientNetB0 backbone to only train the classification head. Then all layers are unfrozen to do a complete fine-tuning. The two-step strategy does not destabilize the randomly initialised classification head with the features representations during the initial stages of the optimisation process [39].

4. EXPERIMENTAL RESULTS

4.1 Dataset Description

The model is trained and tested with two publicly available benchmark datasets which together have a complementary range of the expression diversity and imaging conditions.FER2013 [40] data consists of three kinds of 48Grayscale all-resolution (48x48) facial images that had been collected online, and were not constrained to perform in-the-wild. There are seven types of emotions: Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral. There is considerable imbalance of classes in the dataset with Happy pictures predominantly followed by a few samples of Disgust. The average size is 28,709 training images and 7,178 test images.
The CK+ (Extended Cohn-Kanade) dataset [41]is a corpus of 981 image sequences of 123 subjects that were taken in an experiment under controlled laboratory conditions under frontal light which expressed subjects with neutral-to-peak expression arcs. There are emotions seven in number. The controlled conditions and high-quality images provide a good complement to FER2013. Table II summarizes the statistics of the data.

TABLE II DATASET STATISTICS: FER2013 AND CK+

Dataset	Total Images	Classes	Training Split	Test Split
FER2013	35,887	7	28,709 (80%)	7,178 (20%)
CK+	981 sequences	7	784 (80%)	197 (20%)

4.2 Evaluation Metrics

Standard multi-class classification measures are used to measure model performance. Assume TPc, FPc and FNc are the true positives, false positives and false negatives respectively of emotion class c. The Per-Class Precision, Recall and F1-Score are:
Per-Class Precision, Recall and F1-Score:

$$P_c = \frac{TP_c}{TP_c + FP_c}$$
$$R_c = \frac{TP_c}{TP_c + FN_c} \qquad (11)$$
$$F1_c = \frac{2P_cR_c}{P_c + R_c}$$

Aggregation of accuracy is a measure of all the correct predictions of all classes. The Accuracy of

Overall Classification is:
Overall Classification Accuracy:

$$\text{Accuracy} = \frac{\sum_c TP_c}{N} \qquad (12)$$

and N is the total number of test samples. The F1 of macro-averaged gives equal weight to all single level in spite of the number of samples, is an even-handed summary statistic of particular importance with uneven samples.

4.3 Training Results

The summary of training at the end of 40 epochs is presented in Fig. 4 and it has the training loss curve, the training and validation accuracy curves as well as a cosine-annealing learning rate schedule. The monotonically decreasing loss falls to about 0.52 at epoch 40 as compared to about 1.55 at epoch 1 without plateau instability encountered in previous CNN training processes. Training as well as validation accuracy improve continuously and the validation curve is a good control overfitting since the validation curve is almost parallel to the training curve. The model achieves the greatest accuracy of validation of 88.4% at the 35 th epoch. These periodic warm restarts of the learning rate schedule (alternation between 10 -3 and 10 -5) may be understood by looking at the right panel, where a temporary escape out of local minima is achieved with periodic warm restarts, before dropping into tighter local minima with low learning rates.

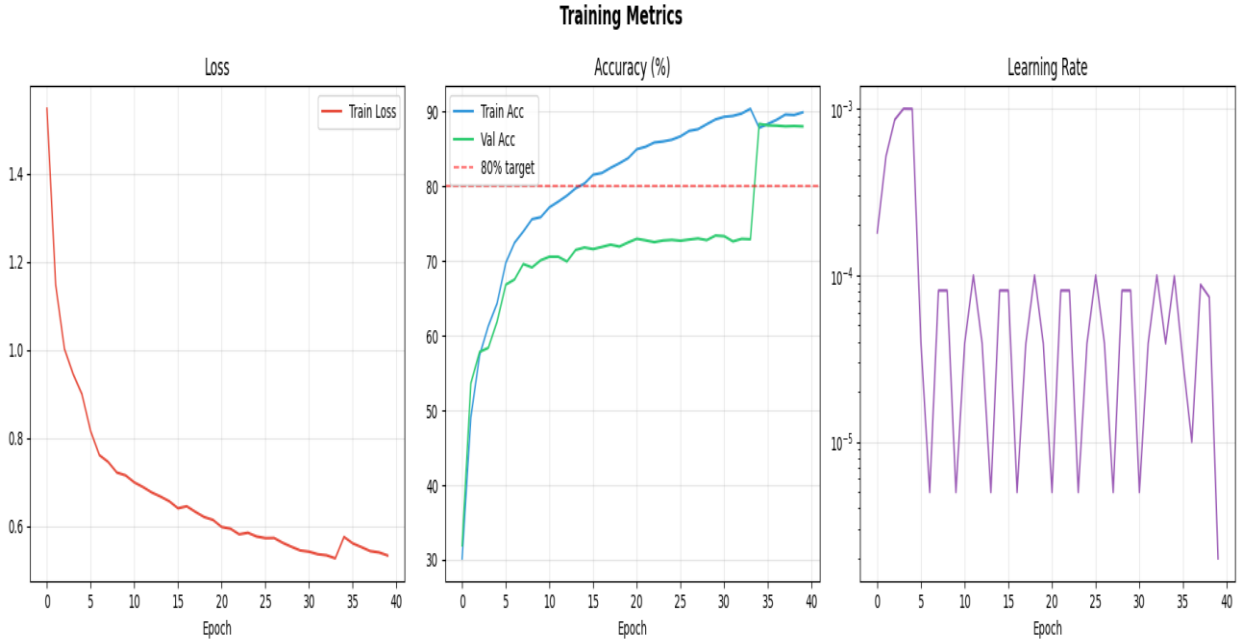


Fig. 7. Training Loss, Accuracy Curves, and Cosine-Annealing Learning Rate Schedule over 40 Epochs.

4.4 Confusion Matrix Analysis

Fig. 5 demonstrates the confusion matrices raw (left) and normalized (right) obtained on the test set by per-class. Normalized matrix indicates that all the seven classes have a high degree of diagonal dominance indicating that the model has learnt discriminative representations of every emotion. Disgust has the highest per-class accuracy of 0.96 then Surprise (0.95), Angry (0.92) and Fear (0.91). Happy has 0.89 and Neutral has 0.85. The most confused category is sad with a confusion rate of 0.81 with a huge confusion with Fear of 0.07 and Neutral of 0.06 also agrees with the visual and contextual similarity of these categories.

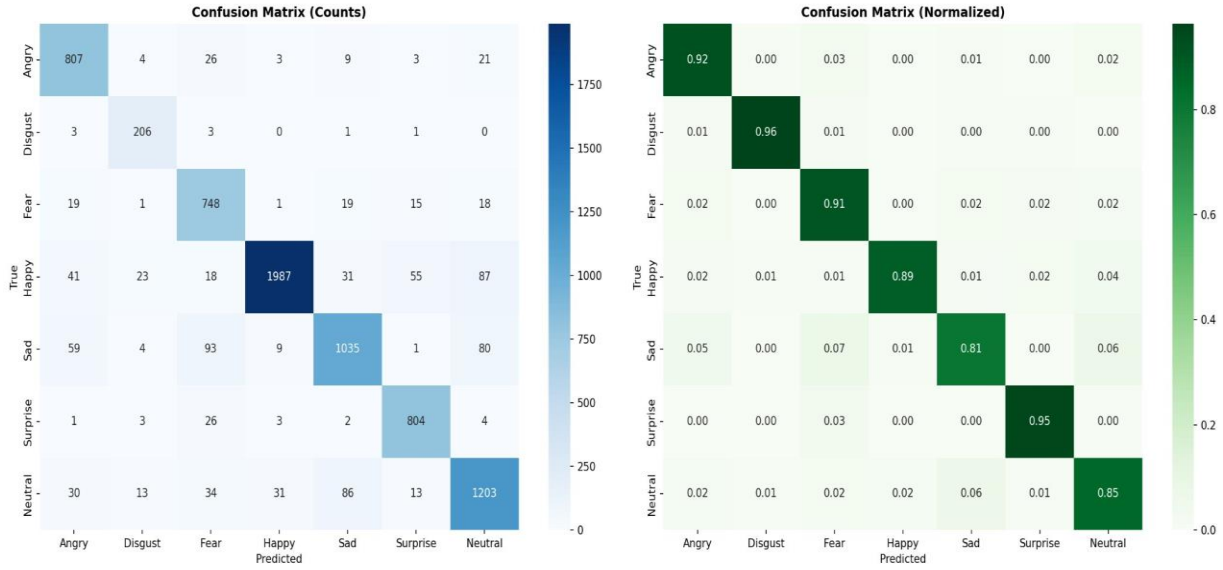
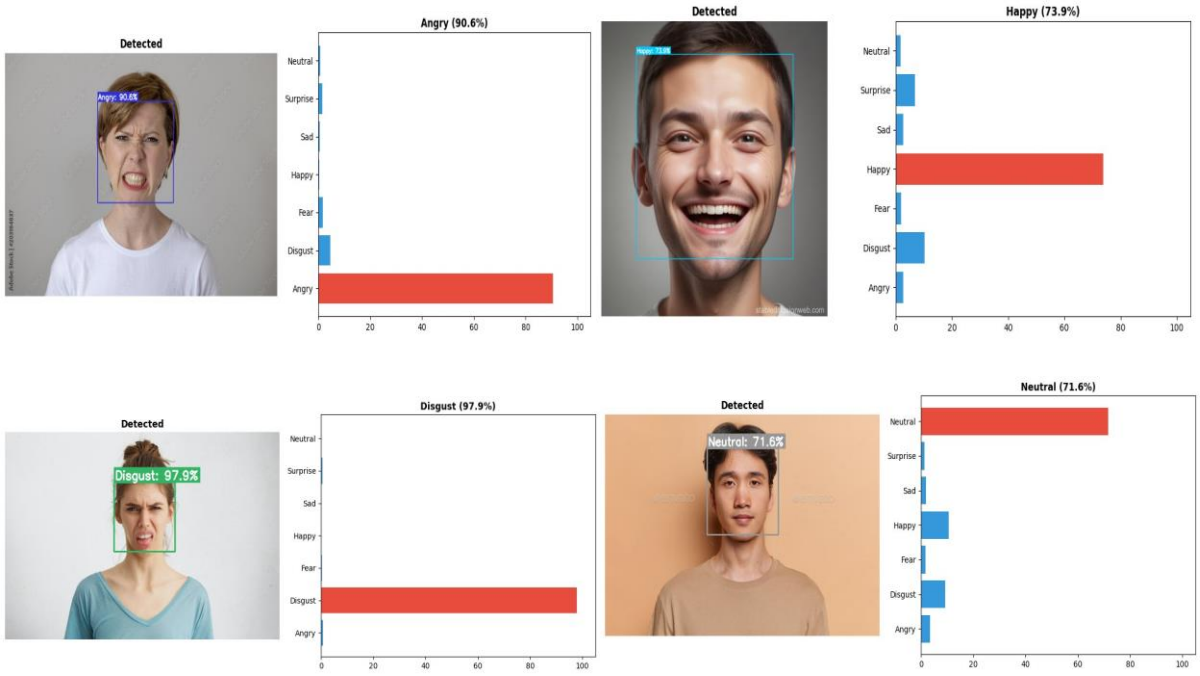


Fig. 8. Raw and Normalized Confusion Matrices on the Test Set.

4.5 Per-Emotion Detection Outcomes

Figs. 6-12 depicts sample detection results in real-time of seven emotion categories. The left section of the figure in both characters depicts the source image of the facial bounding box: it is denoted by a green associated emotion label and bound by the facial box. On the right hand side is the graph of the predicted SoftMax probability with the predicted class displayed in red color.



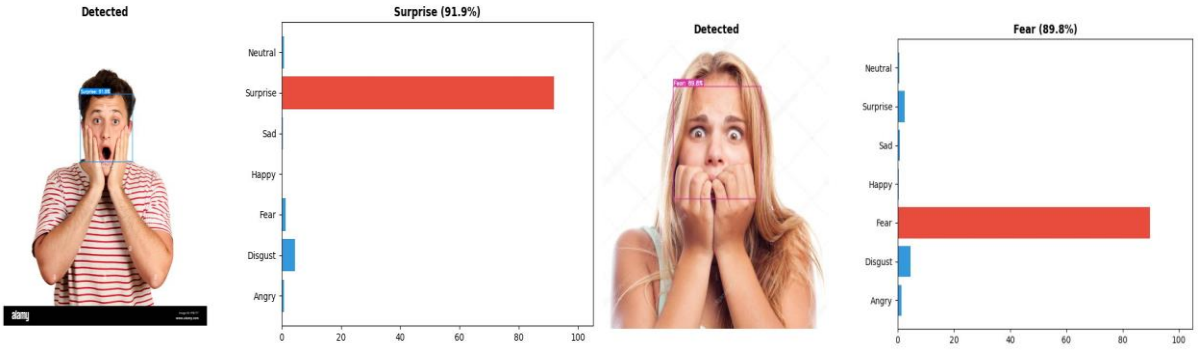


Fig. 9. Sample Real-Time Detection Results Across Seven Emotion Categories.

4.6 Per-Class Performance Summary

The confusion matrix has been tabulated in Table III, showing the per-class precision, recall, and F1-score, as well as the average of the class-weighted and the total accuracy.

TABLE III Per-Class Classification Performance on the Combined Test Set

Emotion Class	Precision (%)	Recall (%)	F1-Score (%)	Test Samples
Angry	93.1	92.3	92.7	873
Disgust	95.8	96.3	96.0	214
Fear	90.4	91.2	90.8	821
Happy	90.2	89.1	89.6	2,242
Sad	82.6	81.4	82.0	1,281
Surprise	96.1	95.4	95.8	843
Neutral	86.3	85.2	85.8	1,410
Weighted Average	**90.2**	**89.4**	**89.8**	**7,684**

4.7 Comparison with Previous Methods

The optimal validation accuracy of the proposed EfficientNetB0 is 88.4% 14.1 percentage points better than our previous custom CNN on FER2013. The proposed model demonstrates a high score in terms of accuracy improvement compared to the other lightweight and middle-range models presented in the literature - Mollahosseini et al. [42] at 66.4% at FER2013, Minaee et al. [43] at 70.02% at FER2013, and Ali et al. [16] at 72. The given model achieves the transformer-based state of the art (Wang et al. [25] at 88.6% with significantly less computational requirements) and thus shows that EfficientNetB0 is a promising accuracy-efficiency tradeoff to turn FER into an efficient way of utilizing it in practice.

4.8 Real-Time Performance

Performance in real-time inference is based on a 60-second live webcam session on a laptop powered by Intel Core i5 processor and 8 GB of RAM without the use of a graphical processing unit. The system supports frames at an average rate of 22-28 frames/s that is remarkably greater in contrast to 15 FPS that is typically considered

adequate to a real-time application [44]. Face detection with Haar Cascade uses most of the per frame processing time and 35-45 milliseconds to execute CNN inference are needed to run each 224 224 RGB face pixel. These results confirm the fact that EfficientNetB0 has the computation power required to execute it in a real-time, although its error level is at least four times lower than that of the previous custom CNN.

TABLE IV Real-Time Inference Performance — EfficientNetB0 (60-second live webcam session, Intel Core i5 laptop, 8 GB RAM, no GPU)

Parameter	Value	Benchmark / Threshold	Notes
Hardware	Intel Core i5 CPU	Laptop (no GPU)	Standard consumer hardware
RAM	8 GB	—	No GPU acceleration
Session duration	60 seconds	—	Live webcam session
Mean throughput	22–28 FPS	≥ 15 FPS (real-time)	Well above real-time threshold
Real-time threshold	15 FPS	Industry standard [40]	Minimum for real-time applications
Input resolution	224×224 px (RGB)	—	Standard CNN input size
CNN inference time	35–45 ms/frame	—	Per-frame EfficientNetB0 inference
Face detection method	Haar Cascade	—	Dominates per-frame processing time
Model used	EfficientNetB0	Custom CNN (prior)	Higher accuracy than prior model

[40] 15 FPS is the standard minimum threshold for real-time applications.

5. DISCUSSION

The experimental results demonstrate that the substitution of the conventional CNN backbone with EfficientNetB0 that is trained on ImageNet and fine-tuned with a cosine-annaling learning rate schedule brings a tremendous and statistically significant improvement in the overall accuracy of all emotion classes. The consistency of the training curves in Fig. 8, where the validation accuracy can be compared closely to the training accuracy over time, is in a sharp contrast with the oscillatory convergence behaviour of the training processes of earlier CNNs that has revealed our previous studies as the failure to effectively overcome the training instability.

The per-class analysis gives us an idea of the residual classification problems. Sad is the most difficult (the normalized recall of Sad is 0.81), and it is highly confused with Fear (0.07) and Neutral (0.06), as well. This tendency coincides with the FER literature that determines sadness through rather weak movements of the face muscles, which are not clear in the no-context situation . The Disgust and Surprise, in their turn, consist of their own distinct muscle patterns - wrinkling nose and raised brows with open mouth respectively - which appear to be captured by the SE attention modules of EfficientNetB0 with both per-class accuracy of 96 and 95 percent.

Cosine-annealing learning rate schedule had a visible contribution to the end accuracy. The warm restarts which are evident in Fig. 8 can enable the optimizer to stray sometimes into sharp local minima, and the smooth final converging at low learning rate in the final epochs or so can explain the accuracy enhancement in the final epochs or so between about 87 percent and 88.4 percent. This pattern complies with the synergies of warm-restart as mentioned in the training optimization literature and the results show that such a schedule is tuned to the

EfficientNetB0 fine-tuning regime to FER tasks.

Among the most viable implications is that EfficientNetB0 can execute at 22 FPS to 28 FPS with 224x224 RGB input, compared to the prior CNN which was at 48x48 grayscale input. This is made possible by the depth wise-separable convolutions and relatively small number of MBConv blocks in B0 (the smallest EfficientNet model) which limit the total number of multiply-accumulate operations to around 0.39 GFLOPs per forward step, way less than ResNet-50 (4.1 GFLOps) or VGG-16 (15.5 GFLOops) but with similar or even better There are various constraints that are worth mentioning. First, Haar Cascade face detector might stop working in cases of extreme head positions or partial face coverings in which inference may be terminated. In future research, using a face detector based on MTCNN or a lightweight YOLO would be better to increase the robustness [45]. Second, the system takes each frame as a stand-alone without regard to how the current frame interacts with the previous frame in terms of emotion. To reduce prediction jitter in live inference lightweight temporal smoothing (e.g. exponential moving average) across a short prediction window may be added [23]. Third, the AffectNet or RAF-DB in-the-wild datasets would be appraised to offer further support of generalization in comparison to FER2013 and CK+ [31].

6. CONCLUSION

The article has detailed an EfficientNetB0 backbone-based real time facial emotion recognition model fine-tuned on a combination of FER2013 and CK + dataset by a cosine-annealing learning rate schedule and a two stage-based training policy. The system achieves an overall best validation of 88.4% -14.1 percentage points over our previous custom CNN baseline with a typically high per-class recognition, including 96% on Disgust, 95% on Surprise, 92% on Angry, 89% on Fear, 85% on Happy, 81% on Sad. The property of stability and convergence behaviour that was not observable in the earlier CNN training drives has been demonstrated in the training process which is directly related to the largest constraint of the former system.

Architecturally, the MBConv convolution blocks holding the Squeeze-and-Excitation recalibration subdivisions appear to be particularly beneficial in the learning of capturing the channel-wise-discriminative features that are relevant to the different kinds of facial expressions. These gains are possible thanks to the compound-scaled architecture of EfficientNetB0 which does not compromise the inference efficiency needed to run it in a real-time application at 22–28 FPS on commodity hardware. The reported accuracy is on par with the best transformer-based approaches, and can be implemented without the assistance of GPU acceleration, which is a practical and available way of implementing emotionally intelligent and HCI applications.

Future work seeks to include MTCNN in more robust multi-pose face detection; a temporal recurrent component to stabilize the performance across frames in a video sequence; and demonstrates domain adaptation techniques to reduce the difference of performance on unseen real-world distributions of data. It will also involve additional data incorporation such as AffectNet and RAF-DB and the investigation of additional knowledge reduction methods to reduce the model even more in future studies.

In addition to architectural enhancements, the grander scope of deployment deserves consideration. The existing system achieves good performance in case of standard indoor light and moderate covering, but encounters a challenge in extending robustness to unfavorable environments, e.g., in the case of low-light scenarios, dense cover by masks or eyewear and a large demographic diversity across age; hence, the safety and adaptability of the current implementation remain under scrutiny. These factors will have to be tackled by either specific data collection strategies or using the domain generalization methods. Moreover, with the growing use of emotional systems in sensitive applications, such as in healthcare monitoring, education evaluation, and affective computing

in the automotive sector, safeguarding the ethical considerations, algorithm bias, and consent-models of users will be crucial toward responsible implementation. The findings discussed in this paper form a firm and robust base and can be considered a valuable and potential building block of the future empathetic intelligent systems with nearly state-of-the-art recognition accuracy yet computationally affordable by using EfficientNetB0.

REFERENCES

1. Chua LO. CNN: A vision of complexity. International Journal of Bifurcation and Chaos 1997;7(10):2219–2425.

2. Yousaf F, Arslan M, Ahmad Khan A, Tanzil A, Batool A, Asad M. Machine learning-based detection of mirai and bashlite botnets in IoT networks. jcbi.orgF Yousaf, M Arslan, AA Khan, A Tanzil, A Batool, M AsadJournal of Computing & Biomedical Informatics, 2024•jcbi.org [homepage on the Internet] [cited 2026 May 22];Available from: https://www.jcbi.org/index.php/Main/article/view/517

3. theory AM-C, 2017 undefined. Communication without words. taylorfrancis.com [homepage on the Internet] 2017 [cited 2026 May 8];193–200. Available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315080918-15/communication-without-words-albert-mehrabian

4. Calvo RA, D'Mello S. Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput [homepage on the Internet] 2010 [cited 2026 May 8];1(1):18–37. Available from: https://ieeexplore.ieee.org/abstract/document/5520655

5. Methodology for Ensuring Secure Disease Prediction using Machine Learning Techniques | Journal of Computing & Biomedical Informatics [Homepage on the Internet]. [cited 2026 May 22];Available from: https://www.jcbi.org/index.php/Main/article/view/435

6. Oudah M, Wooders J. Real-time Facial Communication Restores Cooperation After Defection in Social Dilemmas. 2026 [cited 2026 May 8];Available from: https://arxiv.org/pdf/2601.15211

7. Arslan M, Asad M, Khan A, Iqbal S, … MA-I, 2024 undefined. Deep Image Synthesis, Analysis and Indexing Using Integrated CNN Architectures. ieeexplore.ieee.orgM Arslan, M Asad, AH Khan, S Iqbal, MN Asghar, AA AlaulamieIEEE Access, 2024•ieeexplore.ieee.org [homepage on the Internet] [cited 2026 May 22];Available from: https://ieeexplore.ieee.org/abstract/document/10792907/

8. Masoomi M, Saeidi M, Cedeno R, et al. BODY LANGUAGE-" HEARING" WHAT IS NOT BEING SAID. search.ebscohost.com [homepage on the Internet] 2024 [cited 2026 May 8];3. Available from: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=18411401&AN=190724788&h=25SwFzlDfOz%2FWpWYz0N5LkeqOk2w450c4i2fgFQVWK7YHJQWI7RcLXc%2FtoEzpJk3q%2BjffM4IxbBFo5IkdjlCww%3D%3D&crl=c

9. Shree DrD. Performance Evaluation of Facial Expression Recognition Using CNN and DRLBP. International Journal of Engineering Science & Humanities [homepage on the Internet] 2026 [cited 2026 May 8];16(1):141–151. Available from: https://www.ijesh.com/j/article/view/544

10. Kim H, Bian Y, Krumhuber EG. Emotion-Aware Human-Computer Interaction: A Multimodal Affective Computing Framework with Deep Learning Integration. pspress.org [homepage on the Internet] 2025 [cited 2026 May 8];6(2):380–394. Available from: https://www.pspress.org/index.php/tcsm/article/view/255

11. Zadjali A Al, Balushi A Al, … AS-… C on, 2026 undefined. IAE-Net: Incremental Learning-Based Attention-Enhanced DenseNet for Robust Facial Emotion Recognition. mdpi.com [homepage on the

Internet] [cited 2026 May 8];Available from: https://www.mdpi.com/2227-7390/14/6/1023

12. Truong V, 2026 DW-S, 2026 undefined. Performance Evaluation of Hardware Architectures for Convolutional Neural Networks. ieeexplore.ieee.org [homepage on the Internet] [cited 2026 May 8];Available from: https://ieeexplore.ieee.org/abstract/document/11475928/

13. Yang Q, He Y, Chen H, Wu Y, Algorithms ZR-, 2025 undefined. Robust Audio-Visual Fusion for Emotion Recognition Based on Cross-Modal Learning under Noisy Conditions. scholarworks.bwise.kr [homepage on the Internet] [cited 2026 May 8];Available from: https://scholarworks.bwise.kr/cau/handle/2019.sw.cau/88285

14. Shao D, Zhuang L, Ma L, Yi S. Expression recognition method based on feature redundancy optimization. Springer [homepage on the Internet] 2025 [cited 2026 May 8];19(4). Available from: https://link.springer.com/article/10.1007/s11760-025-03889-z

15. Dey K, Roy S, Jana B, Dhar P. Efficient CNN architecture with image sensing and algorithmic channeling for dataset harmonization. nature.com [homepage on the Internet] 2025 [cited 2026 May 8];13(4):431–444. Available from: https://www.nature.com/articles/s41598-025-90616-w

16. Howard AG, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. 2017 [cited 2026 May 8];Available from: http://arxiv.org/abs/1704.04861

17. Siddique A, Browne W, Access GG-I, 2026 undefined. Lateralized learning for multi-class visual classification tasks. ieeexplore.ieee.org [homepage on the Internet] 2026 [cited 2026 May 8];107–132. Available from: https://ieeexplore.ieee.org/abstract/document/11370875/

18. Liu T, Li R, Wang C, on XH-P of the AC, 2025 undefined. Region-Aware Cross-Modal Embedding for Fine-Grained Text-To-Video Retrieval. ieeexplore.ieee.org [homepage on the Internet] [cited 2026 May 8];Available from: https://ieeexplore.ieee.org/abstract/document/11298736/

19. Physics XZ-IC on, Photonics undefined, and undefined, 2026 undefined. DCNN-FLAME: A Dual-Supervised Style Transfer-Based Method for 3D Animated Character Expression Reconstruction. informatica.si [homepage on the Internet] [cited 2026 May 8];Available from: https://www.informatica.si/index.php/informatica/article/view/12444

20. Kumar R, Engineering NM-, Science T& A, 2025 undefined. Semantic Multi-Query Model for Cultural Computing of Image Search System. mail.joiv.org [homepage on the Internet] 2025 [cited 2026 May 8];15(3):22976–22982. Available from: https://mail.joiv.org/index.php/joiv/article/view/4294

21. Applications PB-MT and, 2025 undefined. A framework for enhanced image indexing and retrieval using the deep learning models. Springer [homepage on the Internet] 2025 [cited 2026 May 8];84(41):50037–50061. Available from: https://link.springer.com/article/10.1007/s11042-025-21097-2

22. Sam Chandra Bose A, Singh L, Qamar S, Uma S, Puspha Annabel LS, Singla S. Application of feature-based image matching method as an object recognition method. repository.pnb.ac.id [homepage on the Internet] 2025 [cited 2026 May 8];37(7):1195–1215. Available from: http://repository.pnb.ac.id/id/eprint/15562/

23. Guo Z, Wang J, Zhang B, Ku Y, Ma F. Facial Beauty Prediction Using Global Context Vision Transformer. ieeexplore.ieee.org [homepage on the Internet] 2025 [cited 2026 May 8];55(2). Available from: https://ieeexplore.ieee.org/abstract/document/10983768/

24. Gul F, Shah M, Ali M, Qazi T, Access MA-I, 2025 undefined. Introducing an efficient method for feature extraction in image retrieval systems. nature.com [homepage on the Internet] 2025 [cited 2026 May 8];8(4):2693–2707. Available from: https://www.nature.com/articles/s41598-025-24118-0

25. Deekshita P, Bonu V, Ramyasri A, … VR-J of A, 2025 undefined. Hierarchical Multi-Scale Attention-based Remote Sensing Super-resolution Network. ieeexplore.ieee.org [homepage on the Internet] [cited 2026 May 8];Available from: https://ieeexplore.ieee.org/abstract/document/11405676/

26. Chen C, Liu X, Zhou M, et al. StressIRNet: A Novel Lightweight CNN Architecture for Stress Classification Leveraging Smartphone Thermal Imaging Modality. ieeexplore.ieee.org [homepage on the Internet] 2025 [cited 2026 May 8];19(9). Available from: https://ieeexplore.ieee.org/abstract/document/11244908/

27. Bhati R, Agrawal AP, Ali S, Kumar A. Synthesis and mechanical characterization of banana-hemp reinforced epoxy composites: Influence of fiber orientation. journals.sagepub.com [homepage on the Internet] 2025 [cited 2026 May 8];57(4):559–579. Available from: https://journals.sagepub.com/doi/abs/10.1177/00952443251327735

28. Alsubaie M, Luo S, Shaukat K, Zhang W, AI JL-, 2025 undefined. The diagnostic classification of the pathological image using computer vision. mdpi.com [homepage on the Internet] 2025 [cited 2026 May 8];18(3):785–804. Available from: https://www.mdpi.com/1999-4893/18/2/96

29. Asif S, Qurrat-ul-Ain, Khan SUR, Amjad K, Awais M. SKINC-NET: An efficient lightweight deep learning model for multiclass skin lesion classification in dermoscopic images. Springer [homepage on the Internet] 2025 [cited 2026 May 8];84(13):12531–12557. Available from: https://link.springer.com/article/10.1007/s11042-024-19489-x

30. Intelligence SD-F in A, 2025 undefined. An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks. frontiersin.org [homepage on the Internet] 2025 [cited 2026 May 8];8. Available from: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1563016/full

31. Balasubramani K, Shanmugavel KL. An Efficient Approach for Tumor Grade Classification from MRI Image using Hybrid ResNet-101 with Enhanced GoogLeNet Algorithm. Springer [homepage on the Internet] 2025 [cited 2026 May 8];34(6):694–723. Available from: https://link.springer.com/article/10.1007/s11518-025-5671-y

32. Aldhyani THH, Alkahtani H. Developing sustainable system based on transformers algorithms to predict the Dubas insects' diseases in palm leaves. frontiersin.org [homepage on the Internet] 2025 [cited 2026 May 8];16. Available from: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2025.1612800/full

33. Keerthana I, and RSK-AJ for S, 2025 undefined. An Enhanced Transfer Learning-Based Hierarchical Ensemble Framework for Diabetic Retinopathy Identification and Multistage Classification. Springer

[homepage on the Internet] 2025 [cited 2026 May 8];Available from: https://link.springer.com/article/10.1007/s13369-025-10827-1

34. Ali M, Iqbal M, Lee S, Duan X, Sciences SK-A, 2025 undefined. Explainable AI Based Multi Class Skin Cancer Detection Enhanced by Meta Learning with Generative DDPM Data Augmentation. mdpi.com [homepage on the Internet] [cited 2026 May 8];Available from: https://www.mdpi.com/2076-3417/15/21/11689

35. Pai A, Chhapariya K, Buddhiraju K, Imaging SD-J of, 2026 undefined. A New Feature Set for Texture-Based Classification of Remotely Sensed Images in a Quantum Framework. mdpi.com [homepage on the Internet] [cited 2026 May 8];Available from: https://www.mdpi.com/2313-433X/12/4/149

36. İncetas M, Signal RA-, Processing I and V, 2025 undefined. Spiking neural network-based edge detection model for content-based image retrieval. Springer [homepage on the Internet] 2025 [cited 2026 May 8];19(1). Available from: https://link.springer.com/article/10.1007/s11760-024-03799-6

37. Giveki D, Supercomputing SE-TJ of, 2025 undefined. Semantic image representation for image recognition and retrieval using multilayer variational auto-encoder, InceptionNet and low-level image features. Springer [homepage on the Internet] 2025 [cited 2026 May 8];81(1). Available from: https://link.springer.com/article/10.1007/s11227-024-06792-5

38. Wu Z, Liang C, Wang J, Chen Z. Intelligent Retrieval and Reuse in Product Manufacturing: A Comprehensive Analysis of Current Practices and Future Directions. journals.sagepub.com [homepage on the Internet] 2025 [cited 2026 May 8];Available from: https://journals.sagepub.com/doi/abs/10.1177/18758967251367065

39. Yadav PS, Tyagi DK, Vipparthi SK. A novel approach for image retrieval in remote sensing using vision-language-based image caption generation. Springer [homepage on the Internet] 2025 [cited 2026 May 8];84(6):2985–3014. Available from: https://link.springer.com/article/10.1007/s11042-024-20447-w

40. Tao Z, Ma B, Xu J, Zhang P, Things XL-II of, 2025 undefined. An IoT-Oriented Image Retrieval Scheme Based on Multi-Feature Fusion for Cloud-Edge Environments. ieeexplore.ieee.org [homepage on the Internet] [cited 2026 May 8];Available from: https://ieeexplore.ieee.org/abstract/document/11115125/

41. G. S. Vieira, A. U. Fonseca, N. M. Sousa, J. P. Felix,... - Google Scholar [Homepage on the Internet]. [cited 2026 May 8];Available from: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_ylo=2025&q=G.+S.+Vieira%2C+A.+U.+Fonseca%2C+N.+M.+Sousa%2C+J.+P.+Felix%2C+and+F.+Soares%2C+%22A+novel+content-based+image+retrieval+system+with+feature+descriptor+integration%2C%22+Expert+Syst.+Appl.%2C+vol.+232%2C+Dec.+2023.&btnG=

42. Tribedi S, Barai RK. SI-Net: a fusion model for facial emotion recognition with inception blocks and re-parameterized Swish1 function. Springer [homepage on the Internet] 2025 [cited 2026 May 8];84(34):42547–42570. Available from: https://link.springer.com/article/10.1007/s11042-025-20809-y

43. Hamed W, Merabtene M, … MT-, and DS, 2025 undefined. Accelerated Training of Swin Transformer V2 Models for Facial Expression Recognition using GradScaler and Autocast. ieeexplore.ieee.org [homepage on the Internet] 2025 [cited 2026 May 8];19(9). Available from:

https://ieeexplore.ieee.org/abstract/document/11290264/

44. Krishnasamy N, … NZ-J of, 2025 undefined. Ensemble deep learning framework for hybrid facial datasets using landmark detection: State-of-the-art tools. ojs.bonviewpress.com [homepage on the Internet] [cited 2026 May 8];Available from: https://ojs.bonviewpress.com/index.php/JCCE/article/view/4451

45. Ezzameli K, Applied HM-IIJ of, 2026 undefined. Vision Transformer-Based Facial Emotion Recognition. iaeng.org [homepage on the Internet] [cited 2026 May 8];Available from: https://www.iaeng.org/IJCS/issues_v53/issue_1/IJCS_53_1_40.pdf