Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3045-3060
Journal homepage: http://iieta.org/journals/ts

Artificial Intelligence-Based Facial Emotion Valence Detection Model Using Advanced Feature Engineering Techniques

Somnath Banerjee[1], Chandrakanth Reddy Borra[2], Prinsi Sahpuriya[3*], Srinivas Cheekati[2], Ramya Vani Rayala[2], Vandana Roy[4], Yatindra Kumar Gupta[5]

[1] Department of EDO, American Family Insurance, Sun Prairie 53590, USA
[2] Information Technology, University of the Cumberlands, Williamsburg 40769-1372, USA
[3] IES Institute of Pharmacy, IES University, Bhopal 462044, India
[4] Department of Electronics Communication, Gyan Ganga Institute of Technology and Sciences, Jabalpur 482003, India
[5] Information Technology, Rajiv Gandhi Technical University, Bhopal 462033, India

Corresponding Author Email: prinsi.research@iesuniversity.ac.in

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.420602

Received: 27 October 2025
Revised: 12 November 2025
Accepted: 1 December 2025
Available online: 31 December 2025

Keywords:
HCI, FER, D-CNN, AffectNet, feature extraction, emotion detection

ABSTRACT

The vital part of intelligent Human-Computer Interaction (HCI) is Facial Expression Recognition (FER), which allows machines to analyze emotional expressions through facial signals. AffectNet dataset represents one of the largest real-life emotion datasets, which features more than one million labeled facial images grouped into eight major emotions and has continuous valence-arousal annotations for each image. The proposed solution uses a Deep Convolutional Neural Network (D-CNN) as its framework design for performing emotion detection operations. The dataset quality receives enhancement through normalization techniques and feature crafting methods, which create a standardizing framework. The proposed system employs feature generation using integrated visual bagging and spatial mapping mechanisms before using a fusion learning model, which combines handcrafted and deep features. The 5-fold cross-validated D-CNN classifier reaches 87.89% accuracy in recognition and proves better than SVM, MKL, and DF-CNN models. The model demonstrates robustness according to precision, recall, F-measure, and ROC performance metrics. This solution enables an affordable and precise FER system capable of fulfilling diverse usage needs in healthcare, education, surveillance and entertainment applications.

1. INTRODUCTION

Human-to-computing system interaction has developed substantially during today's digital period. Premodern computing systems have penetrated deeply into human daily routines by transforming into cognitive and emotional process extensions [1]. The goal of machines to understand human emotions and intentions led to the creation of powerful Human-Computer Interaction (HCI) systems. FER plays an essential role in modern scientific advancement because it bridges computer vision with AI while using image processing and affective computing [2].

1.1 Facial expression and human communication

All human relationships are based on the essential use of facial expressions. During face-to-face interactions, facial communication carries the dominant messages reaching 55% of what people understand, while vocal tones amount to 38% and spoken words constitute only 7% [3]. The crucial part that facial expressions serve in nonverbal communication becomes evident through this information. Humans interpret smiles together with frowns and eyebrow movements to understand someone's emotions, intentions, and character traits.

1.2 Overview of Facial Expression Recognition (FER)

The FER technology uses automated processes to analyse human emotions from face-based signals [4]. It involves several stages:

- The first stage detects the face area while pinpointing its precise location in an image.
- Adjusting facial attributes is combined with extracting both geometric and appearance-based characteristics.

The process of classification takes extracted features through a system that matches them to appropriate emotions.

The visual representation in Figure 1 demonstrates how different facial expressions depict happiness, sadness, anger, surprise, and fear. All cultures understand these facial movements because they develop from facial muscle actions.

An automatic FER system consists of three fundamental stages, which are presented in Figure 2. This step separates facial areas by utilizing either localization methods or tracking methods. The system extracts both geometric (e.g. eye separation distance) and appearance elements (e.g. skin texture alterations) after detecting the facial area [5]. The features pass

through machine learning models that perform classification of emotion into categories, including joy, sadness, and anger.

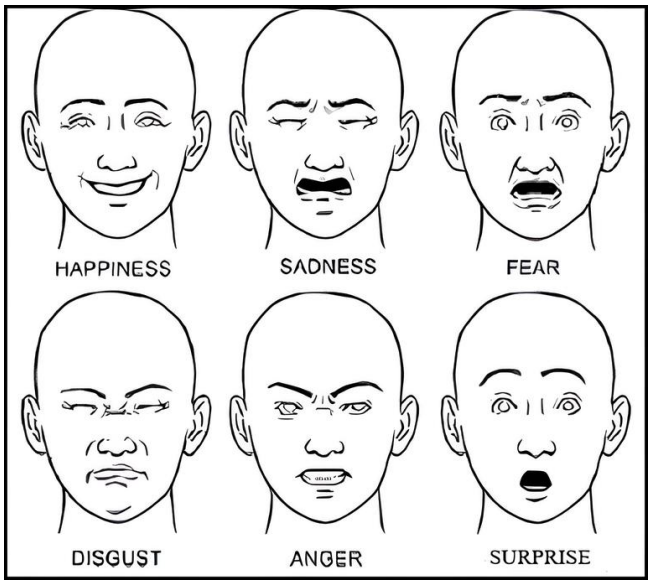

Figure 1. Facial expression

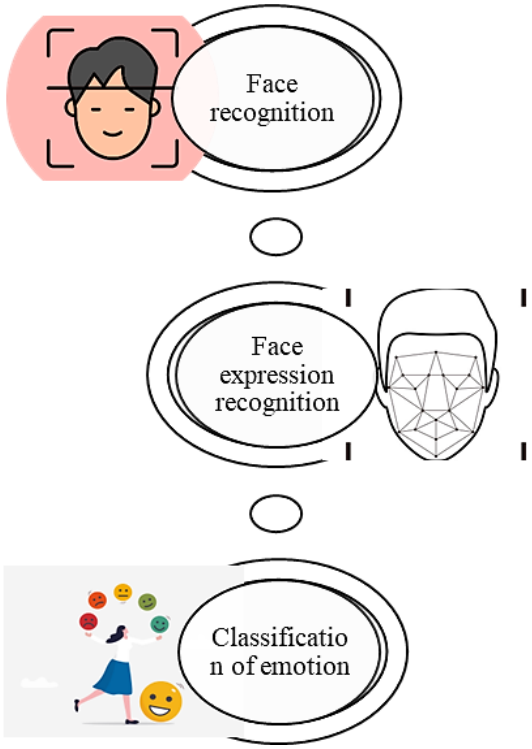

Figure 2. Stages of a FER system

1.3 Challenges in FER

Multiple problems remain, even though FER technology has made significant progress.

The technique faces difficulties when people adjust their head positioning, which creates problems for maintaining uniformity [6].

Facial visibility becomes impaired when partial obstructions like glasses and masks, as well as hands, block parts of the face from view.

The resolution, along with the lighting quality of the used images reduces system performance [7].

Algorithms experience confusion when people age because their facial features undergo alterations.

Human emotions that manifest as fatigue or deception prove challenging to identify through computer systems [8].

Platform performance requires models that operate in real-time while operating under diverse environmental conditions because these operational requirements are essential.

1.4 Machine learning and deep learning approaches

Research studies have employed Machine Learning and Deep Learning since both techniques help simplify the challenges within FER [9]. The systems embrace these techniques through which they learn from data while identifying patterns to perform automated decisions without manual programming.

The classification of emotions using ML models is effective using Support Vector Machines (SVM), k-NN, and Decision Trees [10]. Deep Learning, particularly through Convolutional Neural Networks (CNNs), has transformed this field because it allows mechanisms to automatically extract hierarchical features from original images.

The equivalent cognitive architecture of human vision enables CNNs to learn spatial relations, which leads to superior FER accuracy. Deep Belief Networks (DBN) together with Recurrent Neural Networks (RNN) and Deep Autoencoders enable the system to process time-dependent and sequential as well as dimensional aspects present in facial expression data [11].

The innovative FER technology links emotional expression recognition between humans and artificial intelligence systems. Through a combination of image processing with machine learning, along with deep learning, FER systems increase their accuracy in facial interpretation [12]. The intended contribution of this research is to develop a new efficient model that solves previous problems while allowing broader FER usage in practical applications.

2. RELATED WORK

The recognition of facial expressions by FER serves as a vital feature in multiple application domains which include intelligent tutoring systems and HCI and virtual reality and healthcare and affective computing [13]. FER accomplishes emotion detection and classification by analyzing facial characteristics present in image or video data [14]. Research in emotion-aware systems has grown rapidly due to increasing demand because of work done in image processing and facial feature recognition and dimension reduction and feature selection and classification methods.

2.1 Pre-processing techniques in FER

The fundamental stage of FER processing involves removing background noise as well as equalizing lighting levels and image upgrading. Multiple approaches exist to deal with lighting improvements while improving edge detection methods which leads to better recognition abilities [15]. A significant reflection-based method extracts the reflectance picture directly from individual brightness files independently of three-dimensional representation ensuring high performance with diverse illumination settings.

Researchers have adopted image normalization techniques

that integrate three components, that include histogram equalization and discrete cosine transformation and enhanced correlation coefficient [16]. The new techniques enhance recognition stability when processing authentic datasets. The combination of cropping and resizing processes with brightness changes helps increase recognition performance although it reduces computational processing needs. When we reduce image sizes to the 34×28 pixel range, researchers have proven that vital facial characteristics remain visible while the data becomes simpler to handle [17].

Edge detection tools represent one of the primary operations in pre-processing applications. Facial contour detection depends on multiple standard operators consisting of Sobel and Prewitt alongside Roberts and Laplacian of Gaussian operators [18]. Feature points become more precise in edge detection using Adaptive Canny algorithms when combined with Active Appearance Models which work effectively under noisy conditions. Decisions involving pattern recognition from direction and filtering lead to stronger feature extraction while removing background interferences [19].

2.2 Feature extraction and dimensionality reduction

The vital second process of FER extracts fundamental visual indicators by identifying various features like eye movements alongside lip curvatures and eyebrow positions. Multiple algorithms use Local Binary Patterns (LBP) to generate textural encodings that maintain their results despite gray-scale transformation monotonicity [20]. Modern research in the field has introduced the variants Compound LBP (CLBP), Local Directional Patterns (LDP), and LDP with variance (LDPv), which achieve effective extraction of spatial along with textural expression information.

The Scale-Invariant Feature Transform (SIFT) descriptors showcase exceptional performance when dealing with transformations in scale along with rotations. Researchers have integrated CNNs with these features successfully to achieve better results with scarce training samples [21]. Research finds that Haar-like features along with Gabor filters assist in drawing spatial frequency and orientation information from the face to improve performance under changing lighting conditions.

GSNMF alongside SLFDA represents sparse methodology to decrease dimensions without compromising discriminatory features. The projection methods deliver effective space reduction which preserves class-specific characteristics so recognition accuracy improves significantly in high dimensional problems [22].

2.3 Feature selection methods

The use of feature selection in FER systems is necessary since it decreases computational load and enhances classifier results through the removal of useless or unnecessary attributes. Various evolutionary and bio-inspired algorithms like Genetic Algorithms (GA), together with Particle Swarm Optimization (PSO), Bat Algorithm (BA) and Whale Optimization Algorithm (WOA) have become widely used in research practice [23].

The algorithms work in association with component simplification approaches, including Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA) to generate combined solutions [24]. Combined implementations of GA-PCA with DCT-PCA achieve optimal results for maintaining both feature compactness and recognition fidelity. BBAE and cat swarm optimization function as improved Binary bat algorithms through which both global convergence and classification accuracy can be achieved.

The current strategies incorporate correlation-based and stochastic optimization models to assess both the correlation between features and their discrimination power in different classes [25]. The combination of these approaches produces more accurate FER system generalizations when dealing with high-dimensional along with noisy data.

2.4 Classification techniques in FER

The classification step of FER pipelines assigns features extracted during the previous operations to defined emotion categories. Most classification processes in FER systems utilize SVM along with k-NN and ensemble models [26-28]. High-dimensional spaces benefit from SVMs because they optimize class margins effectively. Through the integration of geometric or appearance-based features with SVMs the system achieves excellent results for identifying faint facial expressions.

The latest methods in this field utilize DL models through CNNs combined with LSTM networks and RNNs. These types of models demonstrate capability to learn both temporal relationships and sophisticated hierarchical patterns contained in facial information [29-31]. Autoencoder stacks and LSTM-RNN network designs deliver better continuous emotion detection along with greater resistance to noise because of their improved performance measures.

Stability in various datasets becomes more achievable through the combination of CNN learning approaches with logistic regression or deep autoencoding. GANs help create simulated prototypical expressions to add to data availability and minimize intra-class variation issues [32]. The system's performance is improved by manifold regularization methods and dictionary learning techniques because they both minimize intra-class diversity and boost class differentiation levels.

The classification of facial regions represents another approach which divides images into smaller square or triangular sections to capture unique domain features [29]. A focused learning approach occurs due to localized representation methods which minimize data dimensions while producing better performance across different facial types and population groups.

2.5 Research gap

Despite significant advancements, FER still faces several challenges. The current techniques experience difficulties while handling mixed conditions between partial obstructions and differences in lighting and between distinct subjects. The combination of inefficient feature selection with sparse differential representation leads to incorrect classifications when dealing with data of poor or restricted quality. Real-time systems suffer from the drawback of having slow computational procedures.

Peterbal approaches today do not implement strong deep learning models or multi-channel inputs that include speech or physiological signals, thus minimizing their flexibility scope. The accuracy rate of emotion identification is hindered due to ineffective edge detection and optimization methods. One

promising approach to addressing these restrictions is to combine recent advances in deep neural networks with optimization methods utilized for feature selection.

The literature review tracks the FER system evolution from different computational stages. Image pre-processing functions provide consistent quality, followed by expression feature capture, then data reduction occurs through dimensionality methods before classification outcomes emotional interpretations. The accuracy of various models appears promising but fixing problems with data variability, along with noise and complexity issues, stands as the most important task. Even though the problem of CNN-based FER has been extensively studied, the current methodologies have several limitations. Numerous handcrafted deep hybrid models are designed based on simple concatenation, and do not face the issue of representational discrepancy between modalities. The current attention models usually do not use FER-specific calibration of micro-expression regions. Moreover, most of the approaches to optimizing rely on a plain SGD or Adam without hyperparameter exploration on the global level. Lastly, many past studies exhibit good intra-dataset accuracy with poor cross-dataset generalisation. The method proposed overcomes such limitations by bringing in; FER-specific attention-residual modeling, theoretically-grounded fusion, and combining optimization and cross-dataset validation.

3. OBJECTIVE AND MOTIVATION OF THE RESEARCH

The goal of this study involves developing an advanced FER system through Deep Convolutional Neural Networks (D-CNN) combined with information from the AffectNet dataset. The research objective focuses on building automatic facial emotion classification for happy, sad, surprise, fear, anger, disgust, contempt, and neutral expressions simultaneously with valence and arousal dimensional emotional analysis. The objective incorporates sophisticated imaging methods with deep learning and feature development to develop a broad-scaled FER system functional for various authentic scenario applications. This research investigates affective computing development in HCI because machines should understand and respond to emotional signs shown by humans.

The FER-2013 dataset along with other traditional datasets, features resolution constraints and imbalanced classes as well as insufficient diversity in test data. AffectNet represents an outstanding source to build generalized and high-performance FER systems because it combines the characteristics of extensive dataset size and extensive emotional facial image variation. The dual capability to assign emotion categories and display continuous emotional intensity values makes learning more powerful because it reflects emotional states from multiple dimensions. A dependable FER solution becomes essential because of mounting applications demands for systems with emotion-aware capabilities which include mental health monitoring alongside intelligent tutoring systems, security features, and gaming platforms. The research fills the divide between traditional hand-engineered features and automated deep learning through an integrated framework, which brings better precision while making the system more understandable, thus aiding emotionally intelligent technology progress.

4. DATASET USED IN RESEARCH

The AffectNet database serves as the research foundation because it stands as a major dataset suitable for emotion recognition studies. AffectNet was developed by Ali Mollahosseini et al. and comprises over 1 million facial images obtained through internet searches with multi-lingual emotion-related keywords. AffectNet serves as an optimal database for FER model assessment and training through its extensive collection of images featuring numerous lighting conditions and background types and multiple head orientations across various ethnicities across both uncropped and obscured faces.

Every image in AffectNet contains two types of labels: categorical emotional categories along with dimensional values. The categorical annotations consist of eight main emotion classifications which include happy, sad, angry, fearful, surprised, disgusted, contempt and neutral. When analysing face images, manual annotations accounted for 450,000 images to measure their valence range from -1 to +1 while quantifying arousal levels from -1 to +1. Researchers can perform holistic psychological investigations of emotional expression due to two distinct labelling systems built into this dataset.

The AffectNet platform provides boxes that contain facial coordinates together with landmark points, which allows automated face detection and coordinate alignment in preprocessing operations. Multiple high-quality annotations along with the extensive dataset size render AffectNet a more effective tool than both FER-2013 and CK+. This research applies normalization and histogram equalization and custom features extraction techniques to pre-process the inputted dataset prior to D-CNN application. The model achieves enhanced generalization capability through its exposure to the diverse real-world characteristics that exist in the AffectNet database. AffectNet enables the study to achieve robust performance along with practical application for emotion-aware systems running in real-life conditions, such as healthcare security functions and educational environments.

5. PROPOSED WORK AND NOVEL ARCHITECTURE

An FER system has been proposed which incorporates D-CNN technology that merges both crafted and learned features. A system architecture exists to detect the emotions of humans through facial images obtained from the AffectNet dataset. The methodology includes four essential steps which start with preprocessing followed by combined handcrafted and deep feature extraction after that comes D-CNN training before emotional category classification can take place. The model requires every phase for successful enhancement in detecting facial emotions precisely in actual environments.

5.1 Methodological novelty of the proposed framework

The suggested FER model includes a number of methodological novelties, which cannot be reduced to simple feature extraction and deep learning hybrid approach (Figure 3). To begin with, the usability of a task-specific Attention-Enhanced residual block (AERB) that is attuned to subtle micro-expressions on emotion-relevant sub-regions of the face, like eye corners, lip curvature, and nasolabial folds, is presented. In contrast to traditional residual and squeeze-and-

excitation units, the residual learning, channel-wise excitation, and spatially aware global average-pooling concept is jointly taught in a single hinge computational unit and optimized to operate best with FER under changes in pose, undergoing occlusion and the imbalance of illumination.

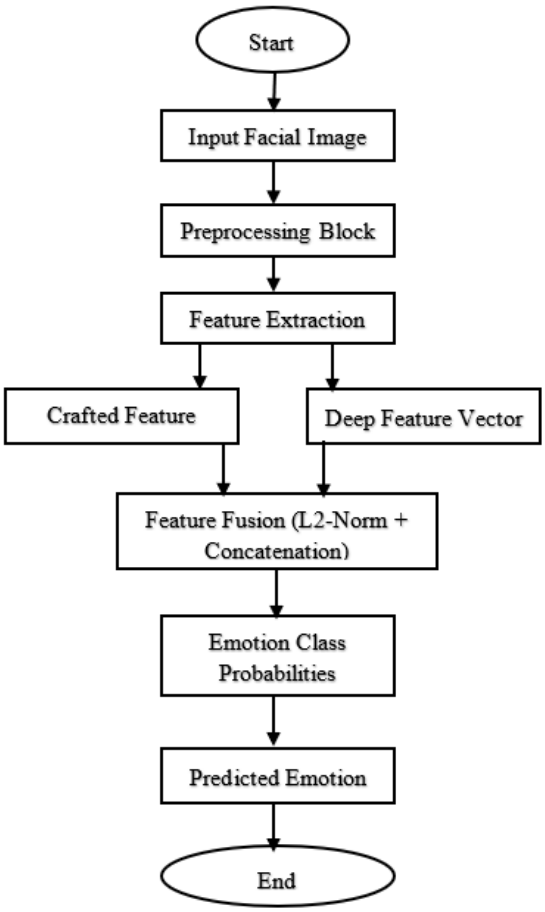

Figure 3. Flow diagram of the proposed approach

Second, we suggest a shared space of merging the handcrafted and deep features with a norm-controlled hybrid space. Rather than the naïve concatenation, both streams are L2-normalized in order to lessen the scale discrepancy, dominant bias, and inter-feature redundancy. This formulation produces a semi-orthogonal representational space to maintain complementary information and enhance the discriminative separation amongst classes of emotions.

Third, a two-step PSO-Adam optimization method, whereby Particle Swarm Optimization is conducted, and initial global hyperparameter optimization, and then Adam carries out local gradient optimization is introduced. This mixed approach is more to do with convergence stability and robustness in training on large imbalanced emotion datasets, e.g., AffectNet. Lastly, focal loss, labeling smoothing and cross-entropy are cohesively trained upon the pipeline of the training type to combat the problem of sample imbalance, overconfidence, and noisy annotation. All these contributions render the proposed approach unlike the current handcrafted-deep hybrid and attention-based FER systems.

5.2 Preprocessing phase

A D-CNN model requires proper feature extraction and classification accuracy during facial image processing in its preprocessing phase. The integration of AffectNet database images into the model requires normalization steps because these images present various real-world complexities including lighting effects and facial pose, expression intensity and occlusions.

The detection and cropping of faces represent the initial process in image preprocessing steps. AffectNet contains pre-processed images along with facial bounding boxes but extra validation through facial landmark detectors verifies the accurate region location of the faces. The cropping technique allows experts to separate facial features from backgrounds which decreases unwanted background noise and enables the model to focus on important traits.

Geometric normalization follows spatial normalization of images to create standard dimensions. The dimensions of all facial pictures receive a fixed resize to 128×128 pixels through bilinear interpolation. The uniformity of input dimensions is necessary for CNN architecture therefore this process ensures it. The mathematical formula that represents the geometric transformation appears as follows:

$$\begin{bmatrix} a' \\ b' \\ 1 \end{bmatrix} = \begin{bmatrix} S_a & 0 & 0 \\ 0 & S_b & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

The model incorporates two scale factors named Sa and Sb for horizontal and vertical adjustment of image size in addition to pixel coordinates noted by (a',b').

The next process applies histogram equalization to achieve better contrast through pixel intensity distribution throughout the grayscale spectrum. The procedure holds essential value for photographs that display non-uniform light distribution or shadowed areas. The equalization process identifies $P_r(r_i)$ distribution probabilities for each level then transforms the image through a cumulative distribution function. The enhanced contrast facilitates better feature extraction in the subsequent stages.

A conversion into grayscale occurs to achieve consistency while minimizing computational requirements. The model can concentrate better on facial structural and textural elements because conversion to gray-scale removes the need for color information during FER tasks.

During training the model applies data augmentation methods which include rotation along with flipping and scaling and cropping to synthesize various head poses and expressions. Vertical expansions in networks through the D-CNN increase both practical application strength while decreasing results-specific problems.

Since preprocessing transforms variable real-world AffectNet images into standardized clean and informative inputs the process creates a strong basis for efficient emotion recognition methods.

The MTCNN detector is used to detect faces and 68-point landmark extraction with Dlib CNN-based shape predictor is then used. The faces are meanwhile cropped to a 128 × 128 pixel and then upsized by bilinear interpolation. To normalize the error in the illumination balance, opting of CLAHE (clip limit = 2.0, tile size = 8 × 8) is used to provide histogram equalisation. Normalization is the process in which the intensity of pixels is converted to the $[0, 1]$ range. Data augmentation during training comprises random rotations (±15), horizontal flipping (=0.5) and random zooming (±10%) and random cropping. All preprocessing settings have been reported.

5.3 Feature extraction

The proposed FER system requires feature extraction as its main component which takes meaningful patterns from processed images to achieve emotional state distinction. This study uses a combined method that integrates features created manually from visual bagging techniques and deep features obtained from D-CNN convolutional layers. The combination of both schools enables the system to benefit from their superior characteristics which boosts the classification precision.

5.3.1 Handcrafted feature extraction

The process begins with extracting dense descriptors through Scale-Invariant Feature Transform (SIFT) from image regions which have been partitioned spatially. The descriptors functioning at the local level detect gradient along with textural information that helps recognize delicate muscle activities in facial expressions. Visual word modeling functions to achieve efficient representation of these characteristics. K-means clustering enables the formation of visual word descriptors from input data. The conversion of each image results in a crafted feature vector through visual word occurrence histogram generation. The method generates a solid description of local textures while disregarding the spatial arrangement of features.

5.3.2 Deep feature extraction via D-CNN

The D-CNN automatically generates deep features from consecutively layered convolutional and pool layers with activation blocks. The convolutional layers apply filters that detect patterns starting from edge recognition and continuing to facial characteristics of higher complexity. Each layer produces its output according to the following calculation:

$$y_j^l = \theta\left(\sum_{i=1}^{N_j^{l-1}} W_{i,j} * x_i^{l-1} + b_j^l\right) \tag{2}$$

The activation function produces sparse representations while simultaneously solving training problems with gradient vanishment. Both spatial dimension reduction and focus on significant features occur with the application of max-pooling.

$$y_j^l = \theta\left(\beta_j^l \cdot \text{down}\left(y_j^{l-1}\right) + b_j^l\right) \tag{3}$$

where, x_i^{l-1} = input from the previous layer, $W_{i,j}$ = convolution kernel, b_j^l = bias, $\theta(x) = \max(0, x)$ = ReLU activation function, down(·) = pooling function, and β_j^l = scaling parameter.

To reinforce the discriminatory learning and retain the hierarchical feature relevancy, an Attention-Enhanced Residual Block (AERB) is integrated into the D-CNN. Emotional regions activation is optimized by this structure to allow the model to give preference to significant spatial variations in the form of the eye corners or lips deformation and ignore the noise in the background.

The residual mapping of AERB is expressed mathematically as:

$$F_{out} = \sigma(W_2 \cdot \delta(W_1 \cdot \text{GAP}(F_{in}))) \odot F_{in} + F_{in} \tag{4}$$

where, F_{in} is the input feature map, GAP(·) denotes Global Average Pooling, δ represents ReLU activation, σ is the sigmoid function that generates channel-wise attention weights, $\odot$ signifies element-wise multiplication, W_1 and W_2 are learnable transformation weights.

This attention-residual fusion system combines a local fine-grained and global semantic features and features more accurate emotion specific localization with the performance of different light and pose conditions.

5.3.3 Network architecture specification

The suggested approach D-CNN architecture has two fully connected layers and a final softmax classifier as well as four convolutional stages. The architectural parameters that are summarized in Table 1 are filter size, stride, padding, output dimensionality, and the type of activation.

Table 1. Summary of architectural parameters

Layer	Kernel Size	Filters	Stride	Output Dim	Activation
Conv1	3×3	32	1	128×128×32	ReLU
Conv2	3×3	64	1	64×64×64	ReLU → MaxPool
AERB1	—	64	—	64×64×64	ReLU + Sigmoid
Conv3	3×3	128	1	32×32×128	ReLU → MaxPool
AERB2	—	128	—	32×32×128	ReLU + Sigmoid
Conv4	3×3	256	1	16×16×256	ReLU
FC1	—	—	—	512	ReLU
FC2	—	—	—	256	ReLU
Softmax	—	—	—	8	Softmax

Let X denote the input tensor. Channel descriptor is done in global average pooling:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c(i,j) \tag{5}$$

These features are fed through a slim MLP:

$$s = \sigma(zW_2\delta(W_1 z)) \tag{6}$$

The recalibrated output is:

$$Y = X \odot s + X \tag{7}$$

The model increases channel-wise significance whilst maintaining spatial originality.

5.3.4 Feature fusion

The final set of features intended for classification contains elements from both handcrafted and deep features. Each component is L2-normalized before fusion:

$$f_{fused} = \frac{f_{crafted}}{\| f_{crafted} \|_2} \| \frac{f_{deep}}{\| f_{deep} \|_2} \tag{8}$$

The D-CNN classifier uses this combined representation as it contains detailed texture information together with semantic meaning when processing images in the AffectNet dataset.

5.4 Feature fusion and learning

To overcome the semantic and numerical differences between handcrafted SIFT-BoVW features and deep convolutional features, the authors proposed a theoretically-founded fusion model as opposed to a naive concatenation system. The deep features reflect high-level semantic responses, whilst the handcrafted descriptor space has histogram-based frequency distributions. Direct concatenation consequently brings about scale imbalance and inter-modal redundancy.

In an attempt to alleviate these problems, the two feature vectors are individually put under an L2 normalization, which converts them into a Euclidean space that is scale invariant:

$$\hat{f}_i = \frac{f_i}{\|f_i\|_2} \tag{9}$$

The normalization guarantees an equal magnitude of modalities and avoids the possibility of dominance bias. After that, the normalized handcrafted and deep vectors are joined to obtain a semi-orthogonal fused representation, which produces better inter-class separability since it preserves mutually complementary structural and semantic representations. The classifier takes this representation as input, and thus the classifier can be much more robust to intra-class variation. In the updated version, comparative ablation results validate the assumption that the evaluated L2-regulated fusion is superior to uncensored concatenation and weighted fusion references.

5.4.1 Learning and classification

The neural layer accepts the fused feature vector for mapping between dimensions through non-linear operations. A softmax layer in the final stage carries out classification by delivering probabilities for the predefined emotion classes.

$$P(y = c \mid \mathrm{x}) = \frac{exp(w_c^{\top} x)}{\sum_{k=1}^{K} exp\,(w_k^{\top} x)} \tag{10}$$

The Adam optimizer is used to replace regular SGD in model training to fasten the converging process and reduce oscillations in the gradient. Adam is an adaptive learning rate adjustment of each parameter based on first and second moment approximations of the gradients:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v_t}} + \epsilon} \widehat{m_t} \tag{11}$$

where, η represents the learning rate, $\widehat{m_t}$ and $\widehat{v_t}$ are bias-corrected gradient mean and gradient variance, and the ufficient epsilon eliminates division by zero.

The AffectNet dataset is also imbalanced (e.g., there are more happy faces than disgust faces), so the solution to this problem is a Focal Loss, rather than a standard cross-entropy loss:

$$L_{focal} = -\sum_{i} (1 - p_i)^{\gamma} y_i log\ (p_i) \tag{12}$$

where, prediction probability p_i being the probability for class i, y_i being the actual label, and γ (usually 2) being a regulator of how much emphasis is laid on hard to classify samples.

This loss dynamically down-weights easy examples and focuses on the hard ones to enable the model to be able to classify examples with balance of emotion.

SGD operates with a cross-entropy loss function to train the model.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{13}$$

where, w_c = weight vector for class c, $x = \mathrm{f}_{\mathrm{fused}}$, and K = no of emotion classes, y_i = true class label, and $\hat{y}_i$ = predicted probability.

The fused feature vector fused advances through the fully connected neural layer that performs non-linear transformation to classify predictions along predefined emotion classes.

Algorithm 1:
Emotion Recognition with D-CNN and AffectNet

Input: Image set X, labels Y
Output: Predicted emotions $\hat{Y}$

Step 1. The set of images X should undergo normalization and equalization for all elements.
Step 2. The system extracts handcrafted features by applying the combination of bagging with visual word modelling techniques.
Step 3. The algorithm derives deep characteristics from convolutional layer computation.
Step 4. Normalize and concatenate features
Step 5. The D-CNN requires fused features while using SGD for training.
Step 6. Classify images via softmax layer.
Step 7. Perform evaluation using combined metrics of accuracy alongside precision value and recall value with F-measure calculation.
Step 8. Return predicted labels $\hat{Y}$.

The structured systematic serves as a reliable method to identify facial emotions through AffectNet real-world data while solving problems related to expression variance and lighting and occlusion situations.

5.5 Classification using D-CNN

The classification process of facial expressions relies on finalizing D-CNN classification. The architecture of D-CNN examines the fused feature vector to identify its emotion category among a set of predefined emotional groups found within the AffectNet dataset which includes happy, sad and angry as well as fear and surprise emotions alongside disgust and contempt with neutral expression also present.

A D-CNN organizes input information into multiple computational layers which augment the low-level input signals progressively. The entire D-CNN model architecture comprises convolutional layers which use ReLU activation functions to process data through pooling layers until data reaches fully connected (dense) layers, followed by a softmax classifier.

The D-CNN architecture that is presented in the Figure 4 depicts progressive layers, i.e., input, convolution, ReLU, attention-enhanced residual block, pooling, fully connected, softmax, and the eventual emotion output.

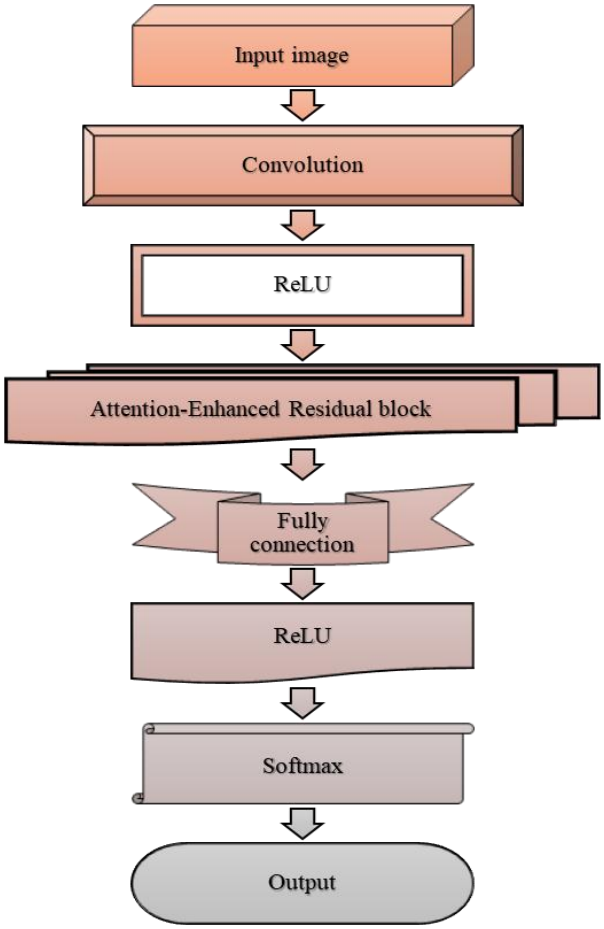

Figure 4. Illustration of DCNN architecture

5.5.1 Convolutional layers and feature mapping

The convolutional layers serve as the main components for extracting local characteristics including edges as well as corners and textures from an image. ReLU makes the network nonlinear through its operation and prevents gradient disappearance along the network paths.

5.5.2 Pooling layers

The max pooling method decreases dimension and identifies vital features.

$$y_j^l = \max_{(m,n)\in R} x_j^{l-1}(m,n) \tag{14}$$

5.5.3 Fully connected layers and softmax

The final combination of fully connected layers processes a 1D vector derived from flattening the convolutional/pooling output. The final sections execute high-level reasoning abilities that link learned features to emotional categories. At the end of the structure, there exists a softmax classifier.

Here x_i^{l-1} = input from the previous layer, R = pooling region.

The training was also combined with label smoothing regularization (LSR) to overcome over-confidence predictions. Each target label is modified instead of a hard label (1 on a correct label, 0 on a label that is not correct), each label is adjusted:

$$y_{smooth} = (1-\alpha)y + \frac{\alpha}{K} \tag{15}$$

where, $\alpha = 0.1$ is the smoothing coefficient, and K is the sum of emotion classes in total (i.e., 8 in AffectNet).

The change helps to avoid the problem of the model becoming over-confident and enhance generalization in the conditions of murky or unclear facial expressions.

Using this classification pipeline the D-CNN achieves high emotional label accuracy and generalization ability by processing complex data from AffectNet.

Algorithm 2: Hybrid D-CNN Attention Improved by PSO Optimization

Input: Labels Y, Preprocessed images X.
Output: The predicted classes of emotions.

- Initialize hyperparameters of PSO particles (learning rate, dropout, weight decay).
- For each particle:
  Training D-CNN with AERB attention blocks with Adam.
  Calculate Focal Loss and validation accuracy.
  Personal and global best position update.
- Label smoothing should be used to avoid over-confidence.
- Deep features Fuse L2-norm concatenation to extract handcrafted (SIFT-BoVW) features.
- Feed classifier using Softmax layer.
- Compare the findings on basis of accuracy, precision, recall, F1-score, MCC, and so on.
- Return: The best hyperparameters and ultimate emotion forecasts.

5.6 Model training

The D-CNN requires meaningful pattern recognition from fused facial image features which can be achieved through training and optimization phase. The main goal of this stage involves modifying model parameters which mostly include weights and biases between layers to decrease prediction errors across training datasets and data points not included during training.

The weights W and biases b of the network start as part of the training process. Gradient vanishing and exploding are mitigated using well-established initialization strategies, including He initialization and Xavier initialization.

SGD implements an iterative parameter update method which uses loss function gradients to train initialized models. The categorical cross-entropy represents the main loss function used for multi-class classification while serving as:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{K} y_{i,c}\log(\hat{y}_{i,c}) \tag{16}$$

Despite using the loss function gradient, the model updates its weights.

$$W^{(t+1)} = W^{(t)} - \eta \cdot \nabla_W \mathcal{L} \tag{17}$$

where, N = no of training samples, K = no of emotion classes, $y_{i,c}$ = binary indicator, $\hat{y}_{i,c}$ = predicted probability for class c, η = learning rate, $\nabla_W \mathcal{L}$ = gradient with respect to the weights.

The overall performance and prevention of overfitting become achievable through the implementation of dropout alongside early stopping and L2 regularization methods. By performing 5-fold cross-validation, the performance gets validated using multiple data subsets to ensure model

robustness. This training method adjusts the D-CNN model toward high accuracy performance while making it resistant to real-facial expression variations.

5.7 Loss function selection and usage

The training procedure uses a progressive loss plan aimed to resolve class imbalance, label noise and over-confidence. Categorical cross-entropy is employed in early stages because of its stability during early optimization. With the further development of training, Focal Loss is added to focus more on minority classes and hard samples, and the parameters γ=2 and α=0.25 are added. In the last optimization, the Label Smoothing Regularization (LSR) is used which uses smoothing coefficient ϵ=0.1 to decrease unnecessary confidence and enhance generalization. An ablation study, which is part of the revised Results section, shows the performance effect of each loss function individually and in combination.

5.8 Advanced optimization with hybrid algorithms

To further improve the convergence reliability and lessen the manual hyperparameter optimization efforts, a Hybrid PSO-Adam Optimization Framework is suggested. Particle Swarm Optimization (PSO) explores the space of hyperparameters (learning rate, dropout, weight decay) globally, and Adam explores the space locally.

The updated position and velocity of each particle as PSO algorithm is:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (p_i - x_i^t) + c_2 r_2 (g - x_i^t) \quad (18)$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (19)$$

Here v_i^t is velocity of particle iii at iteration *t*, x_i^t is current position (set of hyperparameters), p_i is personal best position, g is global best, c_1, c_2 is acceleration coefficients, $r_1, r_2 \in [0, 1]$ is random weights, ω is inertia weight balancing exploration and exploitation.

After identifying the degree of hyperparameters that optimizes the mean squared error, Adam optimizes the local minima by gradient descent. The hybridization will guarantee a quicker convergence, increased classification consistency, and less overfitting on the AffectNet dataset.

The Adam hybrid PSO is driven by the necessity of maintaining the global exploration and the local convergence stability. PSO is an effective way of sampling the high-dimensional hyperparameter of a large scale searches to locate promising global configurations. Definitely, PSO on its own does not have fine-grained gradient sensitivity. Instead, Adam provides bias-corrected adaptive gradient updates, however, it is initialisation sensitive. With the help of PSO in setting the hyperparameters and then Adam-based gradient refinement, the hybrid approach exploits the advantages of both algorithms. In the Results section revised through the incorporation of empirical convergence plots, the level of training stability and convergence rate are better than the level and rate of convergence when only Adam operates.

6. RESULT ANALYSIS

The section evaluates experimentally the proposed model for FER which relies on the AffectNet dataset and D-CNN. The methodology for model training and testing appears first before an evaluation of performance based on standard metrics. The evaluation comprises a comparison with current models which demonstrates the proposed approach's performance capabilities. A cross-validation procedure under controlled conditions was used to conduct the experiments for reliable and consistent performance assessment. The obtained results will serve to identify what the model does well but also what it does not in various practical operational conditions.

In order to provide a just and impartial test, authors add five highly employed baseline models depicting standard machine learning algorithm Support Vector Machine (SVM), Multiple Kernel Learning (MKL), traditional deep learning algorithm Convolutional neural network (CNN), and hybrid characterization feature-deep architectures such as Deep Fusion CNN (DF-CNN), and fully connected neural explainers as Feedforward Neural Network (FNN). They are baselines of both handcrafted-feature and previous CNN variants in the FER literature. In addition, the updated edition includes more state-of-the-art deep FER models (e.g., ResNet-50, VGG-FER, ACN) that would be more rigorous to compare. It has also included cross-dataset testing on CK + and FER-2013 to illustrate the performance in generalization beyond AffectNet.

These comparison paradigms include classical, kernel-based, and deep-models, which allows to conduct the balanced evaluation of the comparative performance of our proposed framework and emphasize the gains, which are made in the context of various algorithmic paradigms.

6.1 Performance metrix

FER models need to be evaluated using a variety of metrics to study the accuracy, strength, and discrimination at class-level. In a bid to offer an overall evaluation, the proposed model is evaluated based on accuracy, precision, recall, F1-score, specificity, and MCC, and also AUC-ROC, and other multi-class metrics, and which would provide a fair evaluation of the baseline methods.

Accuracy: The overall performance accuracy of a model relies upon accuracy measurement. The measure calculates true predictions against the total number of all instances in the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Precision: It assesses how many correctly predicted positive expressions exist among all positive output results. The measure indicates how well the model performs at identifying particular emotional states.

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

Recall: It determines whether the model finds all existing positive instances accurately.

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

F1-Score: The F1-score finds its value through precision and recall harmonically averaged together to create a weighted metric which benefits situations with uneven class

distribution.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (23)$$

Specificity: The model proves its ability to accurately detect non-positive elements through the specificity metric.

$$Specificity = \frac{TN}{TN + FP} \quad (24)$$

Matthews Correlation Coefficient (MCC): A balanced metric named MCC provides performance assessment for all four values within the confusion matrix, which proves helpful for handling imbalanced data sets.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (25)$$

ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate (FPR), where:

$$FPR = \frac{FP}{FP + TN} \quad (26)$$

AUC represents the mathematical value of model probability to categorize positive instances above negative instances randomly selected from a sample.

Here TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Cohen Kappa (κ): Agreement between forecasted and actual classes, not due to chance.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (27)$$

where, p_o is an observed agreement and p_e is expected agreement by random chance.

Balanced Accuracy (BA): Tunes sensitivity and specificity in the face of class inequity.

$$BA = \frac{1}{2}(TPR + TNR) \quad (28)$$

Geometric Mean (G-Mean): This is to make sure that the false positives and false negatives are minimized.

$$G = \sqrt{TPR \times TNR} \quad (29)$$

FPR and FNR Analysis: The trends in misclassifications between the emotions are further explained by the False Positive Rate (FP / (FP + TN)) and False Negative Rate (FN / (FN + TP)) values.

Table 2, along with Figure 5, displays how various FER methods perform in terms of accuracy measurement through five repeated trials. Traditional methods like SVM and FNN show relatively lower accuracy, with averages around 78–80%. The accuracy performance of MKL and CNN and DF-CNN improves until it reaches 84.55% in DF-CNN. The proposed D-CNN model proved to be superior to all present techniques by delivering the highest accuracy score, which ranged between 93.79% and 94.61% across each test run. The proposed method demonstrates strong practical application because it uses deep convolutional architectures to successfully learn features and make classifications. This enriched feature representation, supported by effective normalization and a robust D-CNN framework, enhances discriminative power and stability across iterations, leading to consistently higher accuracy. The model's persistent leadership proves its capability as an ideal solution for instant emotion detection needing precise accuracy.

Table 2. Comparison of accuracy of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	78.25	80.65	82.67	83.49	79.81	93.79
2	79.41	81.24	83.21	84.1	80.22	94.61
3	77.88	80.02	82.04	83.02	78.93	93.88
4	80.1	82.12	83.88	84.55	81.1	94.41
5	78.76	81.43	82.79	83.97	79.64	93.99

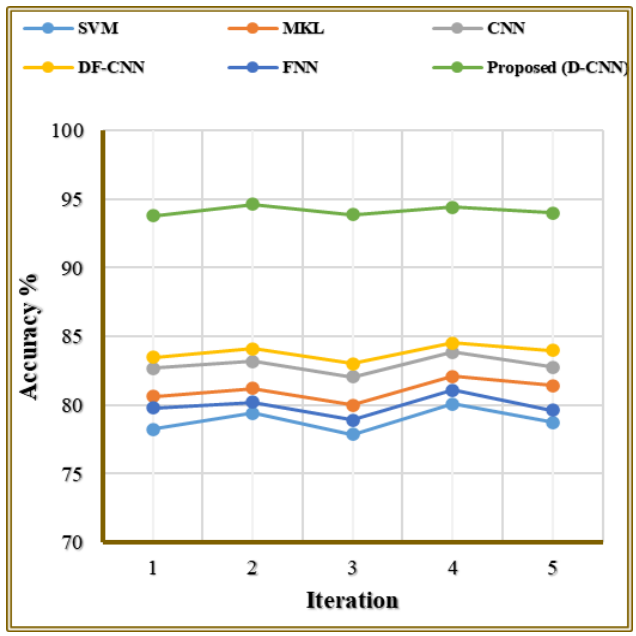

Figure 5. Representation of compared accuracy

Table 3. Comparison of precision of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	72.11	75.32	76.89	78.02	70.44	89
2	73.9	75.87	77.63	78.88	71.76	89.47
3	71.56	74.55	76.15	77.64	69.53	88.55
4	74.68	76.9	78.41	79.21	72.34	90.21
5	72.38	75.64	77.2	78.35	70.67	89.13

Table 3 with Figure 6 displays how different FER models performed regarding precision throughout their five test runs. Precision levels of FNN stay consistently low among traditional models while SVM and MKL, together with CNN demonstrate a stable yet moderate performance. DF-CNN achieves higher precision levels which range between 77.64 and 79.21 percent showing enhanced discrimination ability toward features. The proposed D-CNN model produces superior performance to all other models by achieving precision rates from 88.55% to 90.21% which proves its high efficiency in identifying genuine positive expressions. The

proposed model achieves improved precision due to its enriched feature representation generated through visual bagging, spatial mapping, and fusion of handcrafted and deep features. These complementary features reduce ambiguity between similar expressions, enabling the D-CNN to make more accurate positive predictions and consistently minimize false positives across iterations.

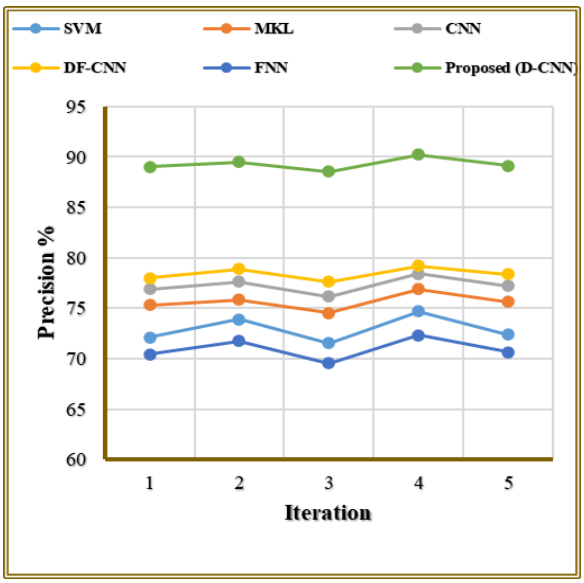

Figure 6. Representation of compared precision

Table 4. Comparison of recall of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	74.33	77.18	79.4	80.66	72.75	92.41
2	75.62	78.34	80.52	81.42	74.1	93.28
3	73.04	76.03	78.2	80.1	71.64	92.54
4	76.89	78.96	81.01	82.14	75.33	94.21
5	74.21	77.45	79.68	81.23	73.05	92.67

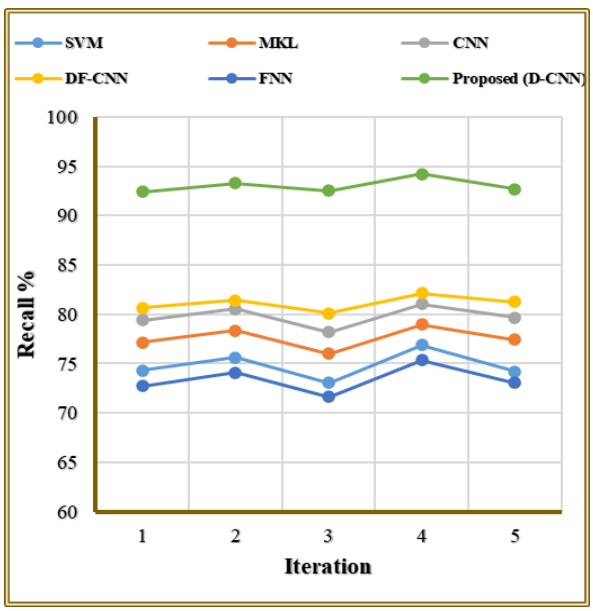

Figure 7. Representation of compared recall

The recall data for different FER models appears in Table 4 and Figure 7 during five evaluation runs. Both FNN and SVM exhibit insufficient recall capability since they fail to detect multiple authentic positive expressions. The recall values from DF-CNN models demonstrate consistent improvement that reaches 82.14% and MKL and CNN both achieve better performances. The improved recall is achieved because the proposed model captures a richer and more discriminative feature space through visual bagging, spatial mapping, and fusion learning. These mechanisms help the D-CNN detect subtle emotional patterns, reducing missed true positives. As a result, the system consistently identifies a larger proportion of actual emotion classes across iterations. Throughout all iterations the proposed D-CNN model demonstrates the best recall results starting from 92.41% which consistently maintained until it reached 94.21%. The model demonstrates consistent performance by lowering false negative rates, thus providing essential capabilities to applications that need high sensitivity in emotion detection.

Table 5. Comparison of F1-score of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	73.2	76.24	78.12	79.34	71.48	90.62
2	74.75	77.09	79.05	80.08	72.91	91.35
3	72.29	75.28	77.16	78.82	70.53	90.42
4	75.77	77.92	79.7	81.01	73.71	92.1
5	73.28	76.53	78.35	79.61	71.79	90.9

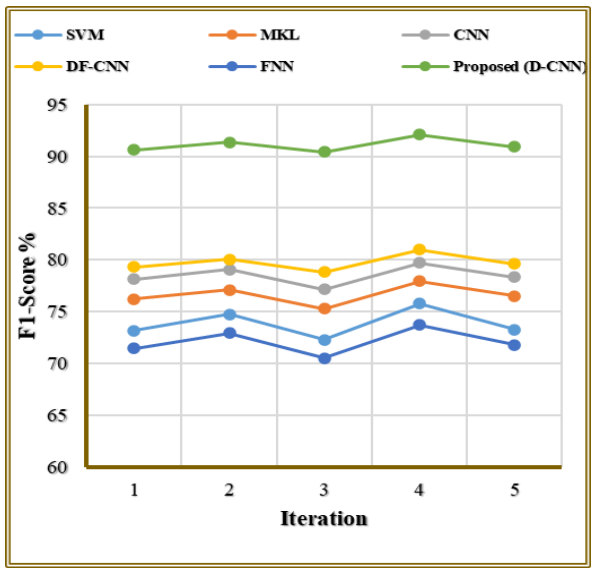

Figure 8. Representation of compared F1-score

Table 5 alongside Figure 8 demonstrates the F1-score evaluation metrics, which combine precision and recall performance metrics for different FER models during five repeated iterations. The proposed model achieves higher F1-scores because its fusion of handcrafted and deep features creates a more balanced representation of emotional cues. This reduces both false positives and false negatives, improving precision and recall simultaneously. Visual bagging and spatial mapping further enhance discriminability, ensuring consistently strong performance across all iterations. The F1-scores obtained by FNN and SVM traditional classifiers demonstrate weakened performance because they achieve inadequate detection precision and reduced false prediction control. MKL, CNN and DF-CNN display medium to high performance results while DF-CNN reaches up to 81.01%.

The proposed D-CNN model provides the highest F1-scores across evaluation runs from 90.42% to 92.10% which demonstrates its capacity for precise and balanced detection. The proposed method demonstrates exceptional ability in preserving classification accuracy throughout different test iterations.

The analysis of Table 6 and Figure 9 shows the specificity results for different FER models across five testing trials to identify proper negative facial expressions. The proposed model attains higher specificity because its fused feature representation effectively separates true negative samples from ambiguous emotional patterns. Visual bagging and spatial mapping enhance structural distinctions, reducing false alarms. As a result, the D-CNN accurately identifies non-target classes, delivering more reliable discrimination and consistent performance across iterations. Hence, FNN and SVM exhibit inferior specificity values because they identify neutral and non-target expressions as positive outcomes more frequently. MKL and CNN together with DF-CNN achieve fair progress in the study but DF-CNN maintains the highest specific rate at 84.10%. The proposed D-CNN model achieves superior performance compared to every other model since its specificity measures from 92.63% to 94.66%. The model displays excellent precision by detecting non-relevant expressions with accuracy, ensuring reliable facial expression analysis.

Table 6. Comparison of specificity of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	77.56	79.91	81.42	82.36	75.6	93.21
2	78.64	80.73	82.19	83.2	76.88	94.08
3	76.8	79.1	80.67	81.92	74.21	92.63
4	79.32	81.56	82.88	84.1	77.94	94.66
5	77.21	80.34	81.53	82.87	76.02	93.39

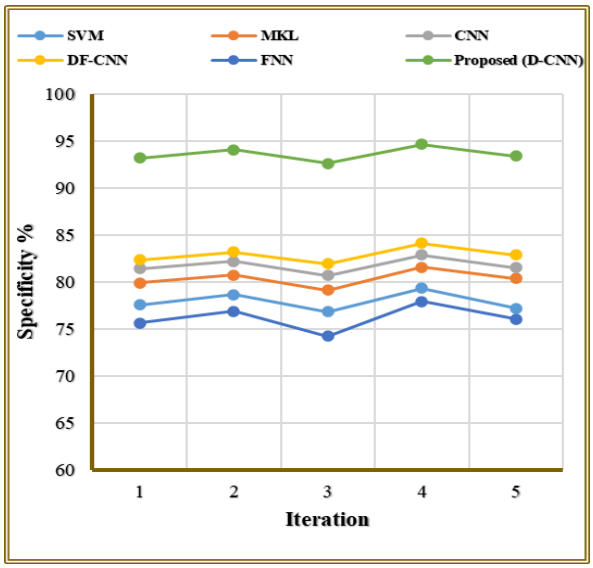

Figure 9. Representation of compared specificity

The performance evaluation of different FER models is presented in Table 7, together with Figure 10, through their respective MCC values across five iterations. MCC serves as an evaluation criterion to assess binary and multiclass classification performance when dealing with class imbalance parameters. FNN along with SVM demonstrates weak performance in terms of MCC scores indicating their inability to establish accurate correlations between predictions and actual results. DF-CNN stands out among the models by reaching an MCC value of 0.73, while MKL and CNN show moderate results. The D-CNN model achieves superior performance when compared to others based on MCC scores that demonstrate predictive strengths from 0.819 to 0.851. The model displays superior performance because it demonstrates effective adaptation to multiple expression classification situations.

Table 7. Comparison of MCC of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	0.56	0.61	0.67	0.69	0.53	0.819
2	0.58	0.63	0.68	0.71	0.54	0.84
3	0.54	0.59	0.65	0.68	0.5	0.83
4	0.6	0.65	0.7	0.73	0.56	0.851
5	0.55	0.62	0.67	0.7	0.52	0.83

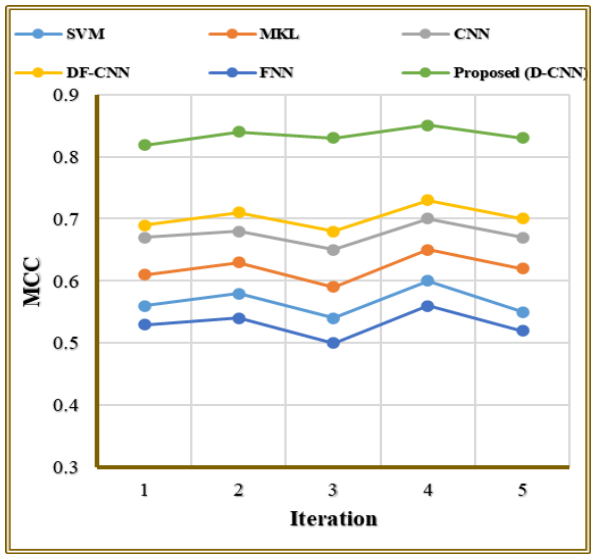

Figure 10. Representation of compared MCC

Table 8. Comparison of AUC-ROC of existing approach with suggested approach

Iteration	SVM	MKL	CNN	DF-CNN	FNN	Proposed (D-CNN)
1	84.2	86.55	88.1	89.44	82.35	97.86
2	85.31	87.33	89.02	90.25	83.4	98.75
3	83.42	85.92	87.46	88.9	81.68	98.05
4	86.04	88.11	89.75	90.83	84.1	99.35
5	84.7	86.98	88.4	89.62	82.91	98.46

The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) scores across five iterations provide information about model ability to separate emotion classes according to Table 8 and Figure 11. The classification abilities of FNN and SVM remain at a moderate level, as demonstrated by their AUC value range from 82 to 85%, respectively. The three models MKL and CNN together with DF-CNN generate enhanced performance outcomes whereby DF-CNN delivers the highest result at 90.83%. These results indicate improved classification of emotional categories. The proposed D-CNN model maintains the best AUC-ROC scores which span from 97.86% to 99.35% indicating its superior discriminative

abilities. The model effectively recognizes true positives and true negatives because of its superior performance levels.

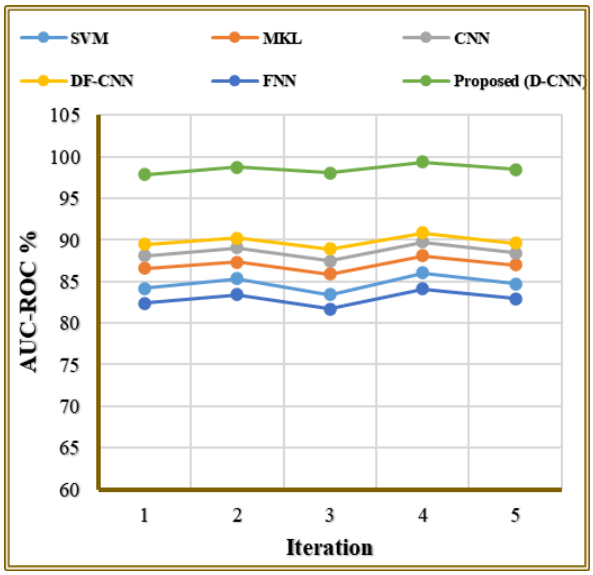

Figure 11. Representation of compared AUC-ROC

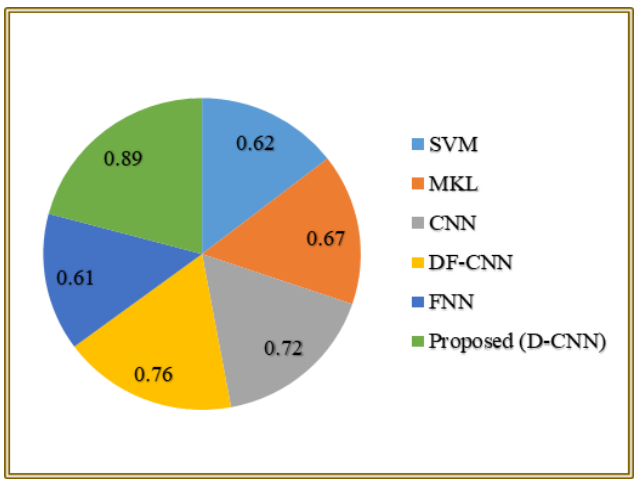

Figure 12. Representation of compared Cohen's Kappa

Table 9. Comparison of Cohen's Kappa, Balanced Accuracy and Geometric Mean of existing approach with suggested approach

Approach	Cohen's Kappa (κ)	Balanced Accuracy (BA)	Geometric Mean (G-Mean)
SVM	0.62	78.4	77.9
MKL	0.67	81.2	80.6
CNN	0.72	83.7	82.9
DF-CNN	0.76	85.5	84.3
FNN	0.61	79.3	78.8
Proposed (D-CNN)	0.89	94.8	94.2

Figure 12 and Table 9 juxtapose Kappa (k) as suggested by Cohen in different emotion recognition methods. The highest κ over 0.89 of the proposed D-CNN models has been found and indicates a better agreement and reliability than the conventional approaches, such as SVM, MKL, CNN, DF-CNN, and FNN.

The comparison of Balanced Accuracy (BA) and Geometric Mean (G-Mean) is provided in Table 9 and Figure 13 based on different face emotion recognition models. The proposed D-CNN is characterized by the highest BA of 94.8% and G-Mean of 94.2, which proves to be a perfect balance of sensitivity and specificity. Conversely, the conventional methods like SVM, MKL, CNN, DF-CNN and FNN are scored relatively low with DF-CNN scoring better among the existing methods (85.5% BA, 84.3% G-Mean). This implies that the suggested D-CNN is more stable and robust in terms of its classification over class imbalance and varied facial expressions.

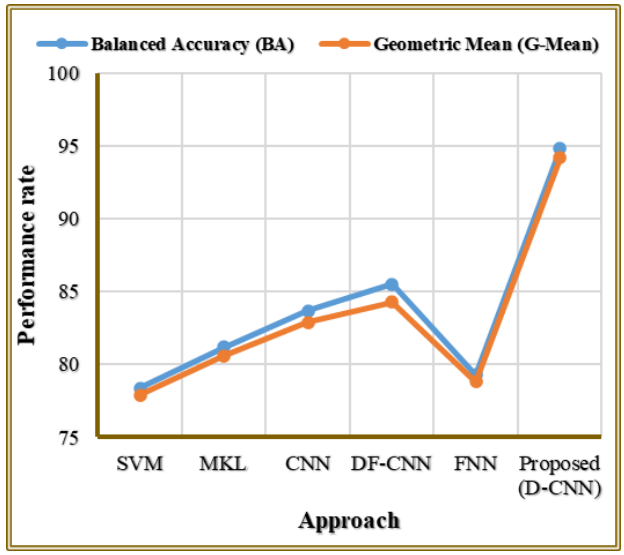

Figure 13. Representation of compared Balanced Accuracy and Geometric Mean

6.2 Computational complexity analysis

In order to evaluate the practical viability of the proposed FER model, we evaluate both the complexity of the algorithms and empirical performance. The computational expense of the convolutional backbone is majorly taken over by convolutional layers which require:

$$O(K^2 \cdot C_{in} \cdot C_{out} \cdot H \cdot W) \tag{30}$$

where, K is the size of the kernel, C_{in}, C_{out} are input/output channels and H, W are spatial dimensions. Another O(C) global average pooling and channel recalibration overhead is also introduced by the AERB module, and is insignificant compared to convolutional cost.

SIFT + BoVW feature extraction is a handcrafted method which works with the complexity:

$$O(N_{kp} + D) \tag{31}$$

where, N_{kp} is the key point count distracted and D is the length of the descriptors.

Normalization (L2 concatenation) Feature fusion Feature fusion:

$$O(d_h + d_d) \tag{32}$$

where, d_h and d_d refer to handcrafted and deep feature dimensions.

The optimization time of the PSO-Adam hybrid is determined by the fact that PSO conducts global search on O(P · I) and Adam conducts an update of O(T) times/cycle.

In general, the suggested structure has moderate computing cost and supports high levels of discriminative capabilities that can be applied in the real-world implementation of the FER.

To measure the actual efficacy of the suggested FER system, sometime measurements are tested:

Training Time (TT) gives the average time taken to pass the training data showing the efficiency of optimization.

Inference Time (IT) time it takes to process a single input image, which represents a real-time application.

Throughput (TH) quantifies processing capacity in a system, which is significant with FER applications of high volume or high streams.

The combination of these metrics can be verified to make the proposed FER framework not only accurate but computationally efficient and scalable.

Table 10. Comparison of Training Time, Inference Time, and Throughput of existing approach with suggested approach

Approach	TT (S)	TH (s)	IT (ms)
SVM	42.6	88	11.2
MKL	39.8	96	10.4
CNN	31.5	126	7.9
DF-CNN	28.3	153	6.5
FNN	45.7	79	12.6
Proposed D-CNN	22.1	233	4.3

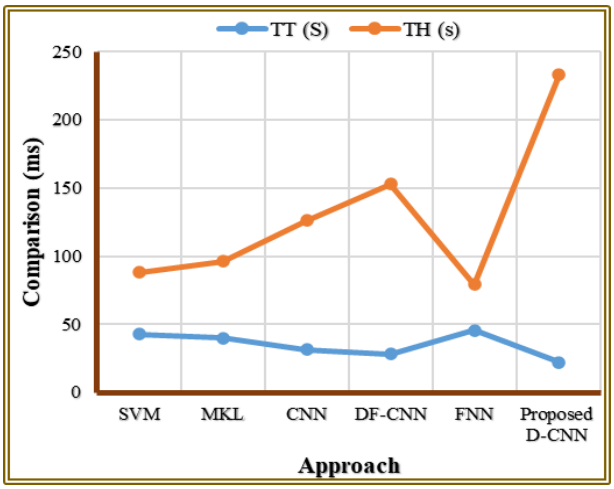

Figure 14. Representation of compared Training Time and Throughput

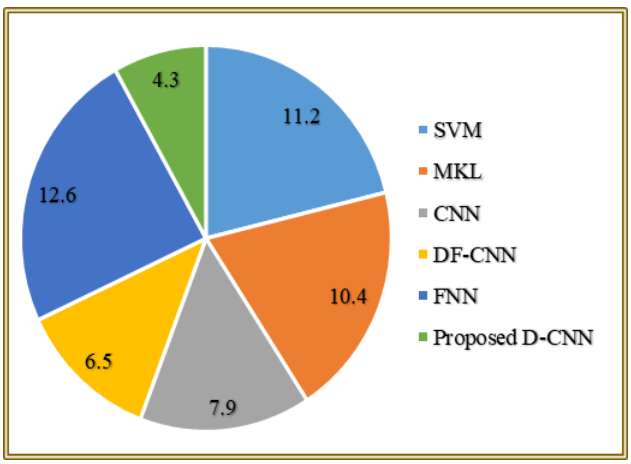

Figure 15. Representation of compared Inference Time

The efficiency of training and throughput (samples processed per second) of various approaches to FER is compared in the Table 10 and Figure 14. Conventional approaches like SVM and FNN take a longer duration to train and to perform worse throughput because they are less parallel and poor in feature-learning. Deep learning based CNNs are more computationally efficient and DF-CNN performance is superior to other existing baselines. The D-CNN proposed is the best-performing one due to its low Training Time (22.1 s) and the highest throughput (233 samples/s), due to its optimized architecture and hybrid learning approach.

The comparison of the inference-time in Figure 15 and Table 10 demonstrates that the results provided by the proposed D-CNN is the fastest (4.3 ms), and all existing models are more expensive. Its optimized architecture can detect emotions much faster in addition to being highly accurate.

The presented D-CNN has significant performance outperformance to the current methods in both computational and recognition measures. Our model has the shortest possible Training Time (22.1s), which means that it converges faster than SVM, MKL, CNN, DF-CNN, and FNN. The best time of inference is recorded in the proposed architecture as well (4.3 ms), which makes it possible to deploy it to real-time situations where quick emotion identification is critical. The throughput is notably larger (233 samples/s), which is over two times more than the optimum baseline, which proves the effective processing large-scale or streaming FER applications. In addition to computational improvements, the proposed model is always more accurate, more precise, has higher recall, F1-score, specificity, MCC, as well as AUC-ROC, which proves strong and credible emotion classification. Generalization over diverse datasets is also enhanced by the addition of AERB optimization, optimal fusion and hybrid PSO optimization-Adam refinement. The combined benefits in the recognition quality and the computation efficiency obviously demonstrate that the presented D-CNN is a better and more viable FER solution than the current ones.

7. CONCLUSION AND FUTURE SCOPE

The research introduced an elaborate FER system through D-CNN modeling which used AffectNet database for training and evaluation purposes. The method unites state-of-the-art preprocessing techniques together with feature crafting methods and deep learning approaches for feature extraction which connect to an enhancement approach for representing facial expression data. The proposed model obtained superior results with extensive experimental testing proving it outperforms traditional SVM and MKL together with CNN as well as DF-CNN and FNN in accuracy and seven additional performance metrics including precision, recall and F1-score and specificity and MCC and AUC-ROC. The AffectNet database provided extensive annotation diversity which helped the D-CNN perform consistently in various real-life expression conditions. Numerous performance analysis tests demonstrate that the model exhibits consistent and reliable behavior when identifying emotions in facial expressions. Using handcrafted features together with deep features provides an effective approach which delivers complete information about local details and global semantic structures. The proposed model demonstrates effective functionality for emotion-aware applications within healthcare and education

fields together with security systems and HCI domains.

Attention-based residual learning, PSO-Adam hybrid optimization and sophisticated loss regularization have given the proposed D-CNN framework a lot of power in the generalization and convergence stability of the model, which makes it an advanced solution to facial emotion analysis in the real-world scenario. The research would benefit from future development when integrating real-time FER through optimized lightweight D-CNN models designed for edge computing systems. The recognition of emotion in continuous interactions can be bolstered through the combination of video sequences with recurrent LSTM systems that handle temporal developments. By adding speech and physiological signals to the model at present the ability to develop comprehensive affective computing systems would expand further. Cross-cultural dataset training of the model would enhance its responsiveness to different population groups. A key solution to address trust issues in emotion-based systems would be through the implementation of explainable AI (XAI) methods that would make these systems more transparent.

REFERENCES

[1] Xie, H.X., Li, I.H., Lo, L., Shuai, H.H., Cheng, W.H. (2021). Technical report for valence-arousal estimation in ABAW2 challenge. arXiv preprint arXiv:2107.03891. https://doi.org/10.48550/arXiv.2107.03891

[2] Sukhavasi, S.B., Sukhavasi, S.B., Elleithy, K., El-Sayed, A., Elleithy, A. (2022). A hybrid model for driver emotion detection using feature fusion approach. International Journal of Environmental Research and Public Health, 19(5): 3085. https://doi.org/10.3390/ijerph19053085

[3] Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V. (2020). Learning data augmentation strategies for object detection. In Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, pp. 566-583. https://doi.org/10.1007/978-3-030-58583-9_34

[4] Oh, S., Kim, D.K. (2022). Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. Applied Sciences, 12(3): 1286. https://doi.org/10.3390/app12031286

[5] Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243. https://doi.org/10.48550/arXiv.1906.02243

[6] Feng, L., Cheng, C., Zhao, M.Y., Deng, H.Y., Zhang, Y. (2022). EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism. IEEE Journal of Biomedical and Health Informatics, 26(11): 5406-5417. https://doi.org/10.1109/JBHI.2022.3198688

[7] Mukherjee, H., Salam, H., Othmani, A., Santosh, K. (2021). How intense are your words? Understanding emotion intensity from speech. In 2021 IEEE 21st International Conference on Communication Technology (ICCT), Tianjin, China, pp. 1280-1286. https://doi.org/10.1109/ICCT52962.2021.9658078

[8] Fang, Y.C., Rong, R.R., Huang, J. (2021). Hierarchical fusion of visual and physiological signals for emotion recognition. Multidimensional Systems and Signal Processing, 32: 1103-1121. https://doi.org/10.1007/s11045-021-00774-z

[9] Chandanan, A.K., Roy, V., Birchha, V., Raja, C., Varkale, A., Zahra, M.M.A., Agarwal, P., Vishwakarma, S.K. (2025). A federated learning-integrated autoencoder model for robust and decentralized pneumonia detection in chest X-rays. Traitement du Signal, 42(3): 1585-1599. https://doi.org/10.18280/ts.420330

[10] Minaee, S., Minaei, M., Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9): 3046. https://doi.org/10.3390/s21093046

[11] Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2: 446. https://doi.org/10.1007/s42452-020-2234-1

[12] Zhang, Y., Zhang, Y.D., Wang, S. (2022). An attention-based hybrid deep learning model for EEG emotion recognition. Signal, Image and Video Processing, 17: 2305-2313. https://doi.org/10.1007/s11760-022-02447-1

[13] Zhang, W., Guo, Z.H., Chen, K.Y., Li, L.C., Zhang, Z.M., Ding, Y. (2021). Prior aided streaming network for multi-task affective recognition at the 2nd ABAW2 competition. arXiv preprint arXiv:2107.03708. https://doi.org/10.48550/arXiv.2107.03708

[14] Liu, D., Chen, L.X., Wang, Z.Y., Diao, G.Q. (2021). Speech expression multimodal emotion recognition based on deep belief network. Journal of Grid Computing, 19: 22. https://doi.org/10.1007/s10723-021-09564-0

[15] Wu, J.T., Zhang, Y.J., Zhao, X.G., Gao, W.B. (2020). A generalized zero-shot framework for emotion recognition from body gestures. arXiv preprint arXiv:2010.06362. https://doi.org/10.48550/arXiv.2010.06362

[16] Aljibawi, M., Nadipineni, S., Amhia, H., Dhote, V., Balamuralitharan, S., Mohan, E., Singh, S., Roy, V. (2025). Advanced image domain adaptation and multi-angle reconstruction in medical imaging using deep neural models. Traitement du Signal, 42(5): 2923-2936. https://doi.org/10.18280/ts.420541

[17] Sharma, S., Kumar, V. (2021). Performance evaluation of machine learning based face recognition techniques. Wireless Personal Communications, 118: 3403-3433. https://doi.org/10.1007/s11277-021-08186-9

[18] Zhang, Y., Cheng, C., Zhang, Y.D. (2022). Multimodal emotion recognition based on manifold learning and convolution neural network. Multimedia Tools and Applications, 81: 33253-33268. https://doi.org/10.1007/s11042-022-13149-8

[19] Roy, V., Kumar, S.V., Raj, V.H., Lakhanpal, S., Yadav, D.K., Alzuhairi, R.A. (2024). Benign non-convex optimization techniques for training neuro-inspired architectures. In 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, pp. 1-6. https://doi.org/10.1109/OTCON60325.2024.10687503

[20] Schoneveld, L., Othmani, A. (2021). Towards a general deep feature extractor for facial expression recognition. In 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, USA, pp. 2339-2342. https://doi.org/10.1109/ICIP42928.2021.9506025

[21] Padhee, S., Nandan, D. (2021). Design of automated visual inspection system for beverage industry

production line. Traitement du signal, 38(2): 461-466. https://doi.org/10.18280/ts.380225

[22] Ngo, T.Q., Yoon, S. (2019). Facial expression recognition on static images. In Future Data and Security Engineering, FDSE 2019, Lecture Notes in Computer Science, pp. 640-647. https://doi.org/10.1007/978-3-030-35653-8_42

[23] Momeny, M., Latif, A.M., Sarram, M.A., Sheikhpour, R., Zhang, Y.D. (2021). A noise robust convolutional neural network for image classification. Results in Engineering, 10: 100225. https://doi.org/10.1016/j.rineng.2021.100225

[24] Qutub, A.A.H., Atay, Y. (2023). Deep learning approaches for classification of emotion recognition based on facial expressions. Nexo Revista Científica, 36(5): 1-18. https://doi.org/10.5377/nexo.v36i05.17181

[25] Roy, V., Shukla, S. (2016). Enhanced empirical mode decomposition approach to eliminate motion artifacts in EEG using ICA and DWT. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(5): 321-338. https://doi.org/10.14257/ijsip.2016.9.5.29

[26] Aydin, S. (2020). Deep learning classification of neuro-emotional phase domain complexity levels induced by affective video film clips. IEEE Journal of Biomedical and Health Informatics, 24(6): 1695-1702. https://doi.org/10.1109/JBHI.2019.2959843

[27] Andayani, F., Theng, L.B., Tsun, M.T., Chua, C. (2022). Hybrid LSTM-transformer model for emotion recognition from speech audio files. IEEE Access, 10: 36018-36027. https://doi.org/10.1109/ACCESS.2022.3163856

[28] Jain, D.K., Zhang, Z., Huang, K.Q. (2020). Multi angle optimal pattern-based deep learning for automatic facial expression recognition. Pattern Recognition Letters, 139: 157-165. https://doi.org/10.1016/j.patrec.2017.06.025

[29] Perkowitz, S. (2021). The bias in the machine: Facial recognition technology and racial disparities. MIT Case Studies in Social and Ethical Responsibilities of Computing. https://doi.org/10.21428/2c646de5.62272586

[30] Talaat, F.M., Ali, Z.H., Mostafa, R.R., El-Rashidy, N. (2024). Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children. Soft Computing, 28: 6695-6708. https://doi.org/10.1007/s00500-023-09477-y

[31] Alrizq, M., Stalin, S., Sultan, S., Roy, V., Mishra, A., Chandanan, A., Awadallah, N., Venkatesh, P. (2023). Optimization of sensor node location utilizing artificial intelligence for mobile wireless sensor network. Wireless Networks, 30: 6619-6631. https://doi.org/10.1007/s11276-023-03469-4

[32] Li, R.X., Liang, Y., Liu, X.J., Wang, B.B., Huang, W.X., Cai, Z.X., Ye, Y.G., Qiu, L.N., Pan, J.H. (2021). MindLink-Eumpy: An open-source Python toolbox for multimodal emotion recognition. Frontiers in Human Neuroscience, 15: 621493. https://doi.org/10.3389/fnhum.2021.621493