GMR
The Original

HYBRID DEEP LEARNING MODEL FOR EMOTION RECOGNITION SYSTEM WITH PROPOSED WDCRF TECHNIQUE

Durgesh Kumar Kotangle[1], H. S. Hota[2*]

[1,2]Atal Bihari Vajpeyee University, Bilaspur, India

ABSTRACT

Facial Emotion Recognition (FER) is widely used for many applications and Deep Learning (DL) has significantly advanced performance of FER, yet many models fail to capture temporal coherence and realistic emotional transitions in visual sequences. This study introduces a hybrid of Residual Network 18-Layer (ResNet-18), Bidirectional Long Short-Term Memory (BiLSTM) and proposed Weighted Directional Centroid Ranking Function (WDCRF) framework that integrates spatial feature extraction, sequential dependency modelling, and psychologically guided temporal smoothing within a unified end-to-end architecture. The proposed WDCRF is a hand-crafted feature alignment and semantic grounding module that bridges the gap between traditional descriptor-based features (LBP, HOG, Gabor) and deep neural network representations. Its job is to ensure that what the deep model learns stays anchored to psychologically validated, human-interpretable facial cues preventing the deep features from drifting toward patterns that are statistically convenient but emotionally meaningless and integrated in hybrid of ResNet-18 and BiLSTM models to improve overall performance. An improved pre-processing pipeline enhances FER2013 images through alignment, mask-aware augmentation, and adaptive histogram equalization to preserve subtle micro-expressions. Experimental results show that the proposed hybrid model i.e. ResNet-18+BiLSTM+WDCRF achieves 88.85% accuracy, 87.0% precision, 87.0% sensitivity, 97.3% specificity and F1-score of 87.0%, across all emotion classes followed by hybrid Resnet-18 and BiLSTM, ResNet-18 only and BiLSTM only. Models were also tested on different types of occlusion and results were found satisfactory. During ablation study also hybrid of ResNet-18, BiLSTM and WDCRF has shown significant performance with F1-Score of 66%.

KEYWORDS: Weighted Directional Centroid Ranking Function (WDCRF), Facial Emotion Recognition (FER), Residual Network 18-Layer (ResNet-18), Bidirectional Long Short-Term Memory (BiLSTM)

1. INTRODUCTION

Facial Emotion Recognition (FER) is a branch of Artificial Intelligence (AI) and computer vision that aims to automatically identify and classify human emotions from facial expressions [1]. FER systems analyze facial features such as eye movements, eyebrow positions, mouth shape, and facial muscle activations to recognize emotional states including happiness, sadness, anger, fear, surprise, disgust, and neutrality. Modern FER approaches typically employ Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs) such as ResNet-18, VGG, and EfficientNet, to learn discriminative facial representations from large-scale datasets. FER has gained significant attention due to its applications in Human-Computer Interaction (HCI), healthcare monitoring, intelligent tutoring systems, driver fatigue detection, surveillance, customer behaviour analysis, and affective computing [5]. Despite achieving high recognition accuracy, FER systems face challenges such as variations in illumination, head pose, occlusion, cultural differences, and subtle emotional expressions. To overcome these limitations, this research has increasingly focused on feature extraction and hybrid model emotion recognition frameworks thereby improving robustness, accuracy, and interpretability in real-world environments.In computational affective science, it is not enough to merely conceptualize an architecture; the actual test of theoretical rigor lies in its precise operationalization. Implementation, therefore, is not a peripheral exercise but the crucible where epistemic hypotheses meet algorithmic constraints [2].

The design of the proposed ResNet-18–BiLSTM–WDCRF framework is not an arbitrary ensemble but a carefully co-adaptive triad, engineered to circumvent the limitations of isolated DL paradigms. Residual Network 18-Layer (ResNet-18), which dominate FER due to their spatial abstraction capabilities, inherently fail to encode temporal continuity—a trait essential for decoding micro-expressions and emotion transitions [3-4]. Conversely, Bidirectional Long Short-Term Memory (BiLSTM) units are proficient in sequence modeling but lack the spatial sensitivity needed for fine-grained facial analysis [5]. To reconcile these divergences, this study constructs a bidirectional pipeline in which ResNet-18 serve as expression localizer [3], BiLSTM function as temporal aggregators, and proposed Weighted Directional Centroid Ranking Function (WDCRF) governs the consistency and alignment of predicted emotional trajectories over time.

Crucially, the implementation avoids the common trap of pipelined independence, wherein each module operates in isolation. Instead, the system is trained end-to-end with error propagation across temporal and probabilistic dimensions [5], allowing BiLSTM outputs to influence ResNet-18 gradient paths during back-propagation. At the same time, WDCRF regularizes output predictions in a structurally aware fashion [6-7]. This dynamic feedback loop is fundamental: it discourages temporal dissonance, over fitting to neutral expressions, and abrupt oscillation between emotion classes—

challenges that are poorly addressed in state-of-the-art ResNet-18-BiLSTM hybrids lacking probabilistic post-processing layers [8-9].

The preprocessing pipeline, often trivialized in benchmark papers, receives particular architectural attention. In order to develop hybrid model [12] popular FER data FER2013 is used. Current challenges of this dataset includes limited datasets and low interpretability, motivating future research on lightweight, explainable hybrids using Principal Component Analysis (PCA)-based dimensionality reduction and efficient feature fusion [15]. The pre-processing in Sentiment Emotion Recognition (SER) also important and need to extract features to improve performance [16-20].

A dataset fraught with occlusion, illumination variance, and demographic imbalance is subject to a multi-stage transformation protocol [14]. Unlike standard data augmentation strategies that prioritize pixel-level randomness, this protocol incorporates landmark-preserving affine transformations, mask-aware augmentation [15], and expression-preserving histogram equalization.

In this research 10-fold [22] cross validation was applied to split training and testing sample, 10-fold cross validation is useful for robust model development which partitions data in dynamic manner. further hybridization of more than techniques may improve performance of the model. Many authors [21-23] have used hybrid model of Deep Learning and Principal Component Analysis for FER.

The WDCRF is inspired by Conditional Random Field [24-25] and serves as a learnable temporal alignment layer that models emotion transition probabilities, enabling cognitively realistic and smooth prediction sequences. By discouraging abrupt emotional shifts, it improves F1-Score for volatile classes and enhances the realism of output distributions.

The proposed WDCRF is a hand-crafted feature alignment and semantic grounding module that bridges the gap between traditional descriptor-based features and deep neural network representations. WDCRF operates as a post-hoc fusion module, hybrid of ResNet-18–BiLSTM produces its temporal embedding [16] which is concatenated with WDCRF-aligned vector before classification.

The architecture departs from depth-driven FER by emphasizing spatiotemporal fidelity, label coherence, and cognitive plausibility, using ResNet-18, BiLSTM, and WDCRF to enhance both prediction quality and theoretical insight. The experimental results revels that proposed model achieves accuracy=88.85%, precision =87.0%, sensitivity=87.0%, specificity=97.3% and F1-Score= 87.0% which is higher than hybrid ResNet-18 and BiLSTM, ResNet-18 only and BiLSTM only models. Models were also tested on different types of occlusion and results were found satisfactory. During ablation study also hybrid of ResNet-18, BiLSTM and WDCRF has shown significant performance with F1-Score of 66%.

2. Methodology

Emotion recognition is a computer vision technique that identifies emotion based on facial expressions. It analyzes emotion such as Happiness, Sadness, or Anger by converting subtel facial texture changes (wrinkles around the lips, eyes and forehead) into numerical data. The overall process for FER of the proposed model development with different methods, techniques, and algorithms is shown in Figure 1 through process flow diagram in 10 different blocks. These blocks can be divided further into five different sections as below:

i. Data Collection and Data preprocessing.
ii. Handcrafted Feature Extraction and WDCRF based feature alignment.
iii. Spatial Feature Extraction trough ResNet-18 and reshape feature map.
iv. Temporal Modelling through BiLSTM.
v. Feature Fusion and Classification.

The detail of each step are as below:

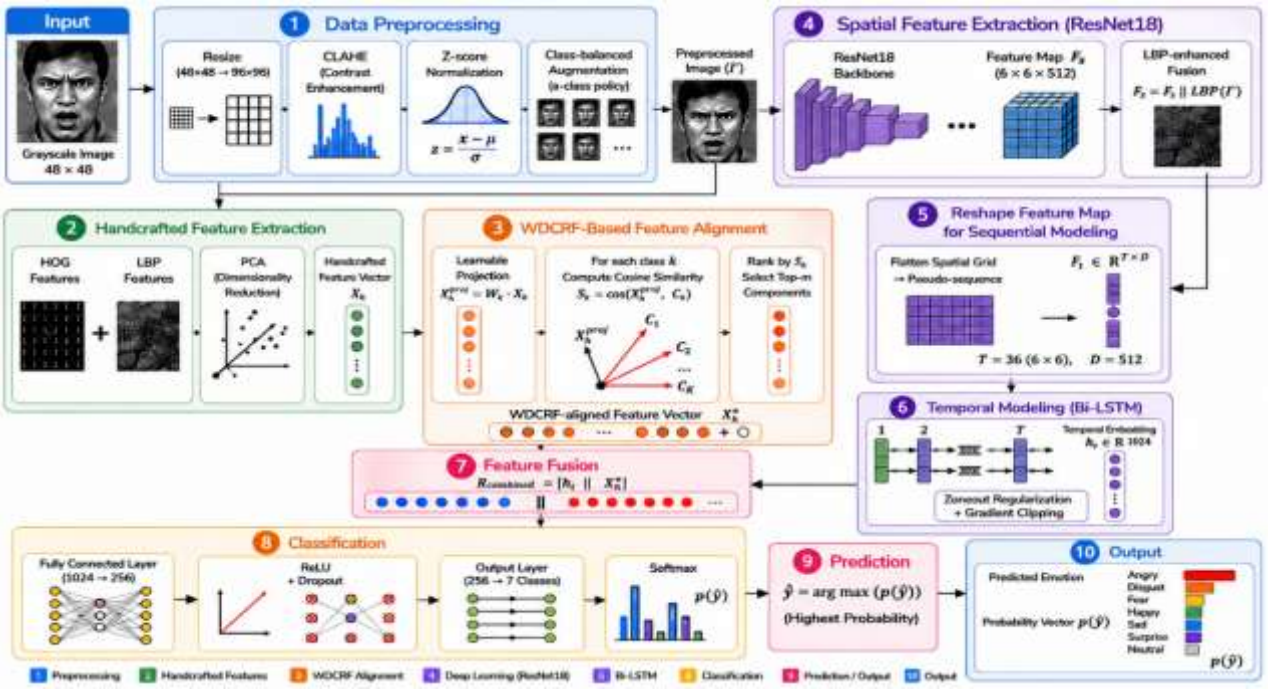


Fig. 1. Process flow diagram of proposed ResNet-18+BiLSTM+WDCRF Model.

i. Data collection and Data pre-processing

FER2013 data collected from kaggle [13] [27] provides a large-scale static of 48X48 grayscale images. This dataset captures real-world variability, which is frequently scarce in laboratory corpora. Due to its significant noise, pronounced class imbalance, low spatial resolution, and limited annotation depth, it is a demanding yet realistic benchmark for assessing emotion detection models. Dataset contains 35,887 images of seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality and consists three directories: Train, Test and Val again consisting 7 subdirectories of 7 emotion classes.

Data preprocessing begins by accepting gray-scale images from the FER2013 dataset. Preprocessing transforms these noisy, low-resolution inputs into geometrically and contrast-normalized representations. Techniques such as Bicubic resizing, CLAHE (Contrast Limited Adaptive Histogram Equalization), and landmark-based augmentation (rotation, translation, scaling) help retain micro-expressions, critical for subtle emotions like fear or disgust [28-29].

In order to feed a specific size of image to ResNet-18 an original image need to resize, Bicubic interpolation is used to enhance image size by preserving quality of image. When an image is enlarge, new pixel values need to be estimated, this technique estimated these values using 4X4 neighborhood around the target pixel and perform better than nearest neighbor and bilinear [31]. CLAHE on the other hand, is an advanced digital image enhancing technique specially when images contains dark, bright or non-uniform illumination used to enhance local contrast and reveal hidden details in both light and dark areas of an image. Unlike standard global contrast enhancement methods, CLAHE works by dividing the image into small, overlapping blocks or tiles and equalizing the histogram for each tile separately. It applies contrast limiting by clipping the histogram at a predefined threshold before equalization. This prevents algorithm from overly amplifying background noise, which is a common flaw in older adaptive methods and enhances eye regions, nasolabial folds, smile contours and eyebrows to identify facial emotions in better way [29-30]. Alpha-class augmentation policy is an augmentation technique which operates on a particular class to resolve class imbalance problems. In FER 2023 dataset many emotions like fear and disgust have relatively few samples. This techniques generates additional few samples for classes having less number of samples. In order to address image quality and class imbalance issues, step by step pre-processing as shown in Figure 2 is applied which includes up scaling of image to 96×96 pixels to improve feature retention in deeper CNN layers [32] [34] using bicubic interpolation and then CLAHE to enhance micro-expressions around key regions (eyes, lips) also low-quality images with severe blur, occlusion, or mislabelling were filtered using edge-density heuristics and feature-space clustering [33]. The enhanced image is then normalized usinf z-score normalization technique and finally data are augmented to meet out class imbalance problem. The final outcome from this pre-processing block is pre-processed image I*

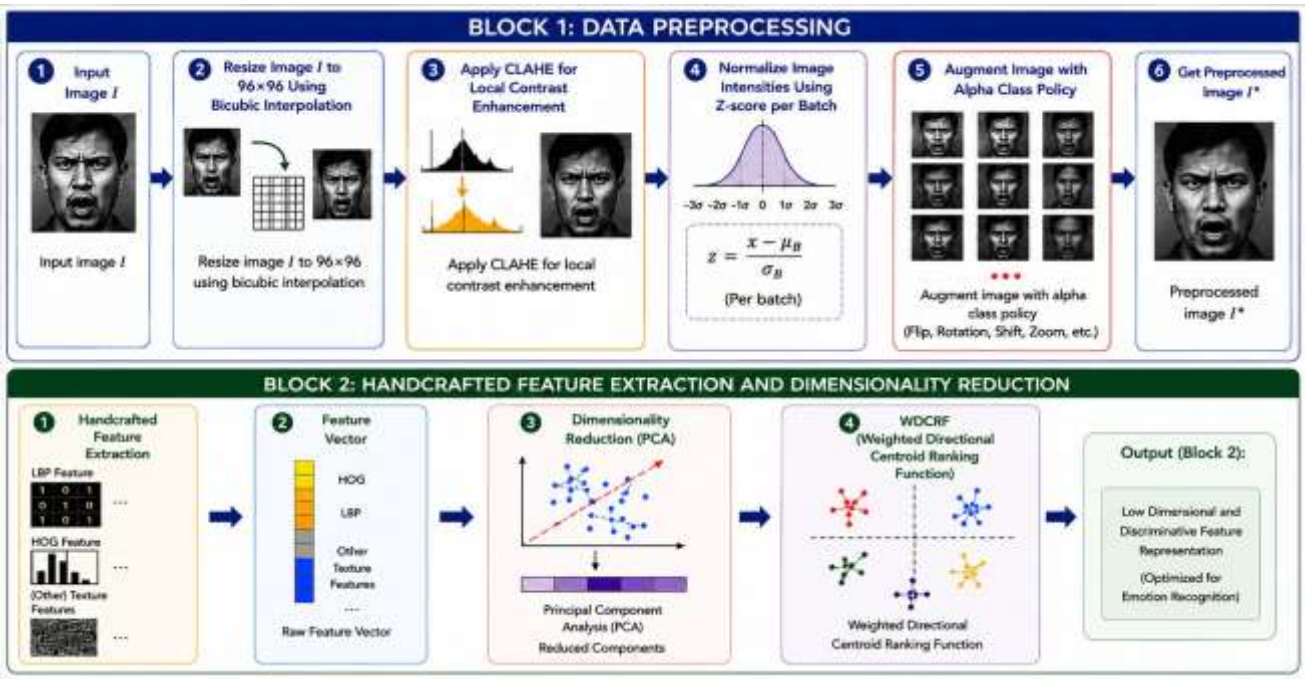


Fig. 2. Data preprocessing pipeline.

ii.Handcrafted Feature Extraction and WDCRF based feature alignment

The pre-processed image I* obtained from data pre-processing block is then feed to next block called handcrafted feature extraction and WDCRF. This step consist two subparts first handcrafted feature extraction and second, proposed WDCRF for generating WDCRF aligned vector. Handcrafted features are extracted through Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP). The suggested approach uses complementary feature representations to improve facial emotion identification by integrating HOG and LBP [36-37].

HOG is a powerful feature extractor recognizer used to recognize human facial expressions. This technique is used to capture edge, shape, and structural information of image as facial geometry changes during various emotions, say for example in case of happy emotion geometry of mouth become curved and cheek raised.

The working principle of HOG starts with grayscale input image, compute gradient, divide image into cell (Typically 8X8 pixels each), and create orientation histogram (a-180 degrees), block normalization and finally feature vector generation, each step having mathematical formula and HOG records global form and contour [39].
The LBP algorithm is a texture descriptor that analyzes facial image through various steps involving mathematical formulation, in first step image is divided into 3X3 digital image region and comparison with center pixel is performed, if the neighboring pixels have a value greater than the center pixel, a 1 is recorded, otherwise a 0 and in the second step for each region, histogram in binary patterns is generated. A histogram is created from all these numbers, which represents the facial features [38]. Finally classification in the exact sentiment is determined by comparing this histogram data with a pre-existing sentiment database. LBP encodes local micro-patterns surrounding facial features [40].
Finally features of HOG and LBP are combines and feeded to PCA. PCA is a mathematical and ML technique used in emotion recognition to simplify data and extract key features [10] [41]. It helps computers quickly and accurately infers emotion facial expressions or voice. When we analyze facial expressions (such as Happiness, Sadness and Anger) or speech patterns (voice), the data becomes too vast and complex for a computer to process. For example even a small picture contains thousands of pixels [29]. PCA reduces the size of this complex data (dimensionality reduction), retaining only the important parts or patterns that identify the emotion, while removing the rest of the unnecessary information (noise). It work in emotion recognition when PCA is used to recognize facial expressions, it is often called Eigenfaces. It extracts mathematical values for key facial features (such as eye distance, lip shape, forehead wrinkles) and creates a model [42].
These extracted features are then feed to proposed WDCRF to get WDCRF aligned vector X'h. The working process of WDCRF is as below:

Step 1: X'_h is projected via a learnable matrix W_h into the alignment space as X_h _proj $=W_h . X_h$

Step 2: Cosine similarity to each class centroid c_k is computed: $S_k = \cos(X_h \_proj, c_k)$

Step 3: Similarities are ranked and top-m dimensions are selected

Step 4: The output is X'_h, the WDCRF-aligned vector which is semantically consistent, emotionally grounded, and stable across noisy or occluded inputs.

iii. Spatial Feature extraction through ResNet-18 and reshape feature map

ResNet-18 is a CNN architecture in which input to the layer is added to the output before passed it to the future layer. The CNN encoder uses a modified ResNet-18 adapted for 48×48 FER2013 grayscale images. Inputs are upscaled to 96×96 in step 1 and replicated across three channels for ImageNet compatibility, forming pseudo-RGB tensors for feature extraction as input tensor as $\mathbb{R}^{N\times96\times96\times3}$
Where:
N is the batch size.
96×96 is the spatial resolution.
3 denotes the RGB color channels.
Each convolutional layer performs localized filtering by sliding a 3D kernel. A convolutional kernel of size 3×3×Cin operates over a local spatial window of the input tensor [44-45], computing weighted sums over channels and producing a feature map of shape:
output : $R^{N\times Hout\times Wout\times Cout}$

$$\text{where } H_{out} = \frac{Hi-K+2P}{S} + 1,\ W_{out} = \frac{Wi-K+2P}{S} + 1,\ \ C_{out} = \frac{C_i-K+2P}{S} + 1$$

Where H is hight (The vertical dimension of the image or feature map, measured in rows), W is width (The horizontal dimension of the image or feature map, measured in columns), C is the channels (The number of channels or feature map in the input/output Tensor).

$$O = \left[\frac{I-K+2P}{S}\right] + 1$$

Where O is the output feature map size, I is the input size, K is the kernel size (typically 3), P is the padding (typically 1), and S is the stride (typically 1). Table 1 shows the parameters: kernel, stride, output shape and channels of ResNet-18 architecture used in this research.

Table 1. Initial Layers of ResNet-18

Layer	Kernel	Stride	Output Shape	Channels
Conv1	7×7	2	48×48×64	64
MaxPool	3×3	2	24×24×64	—
ResBlock1	3×3	1	24×24×64	64
ResBlock2	3×3	2	12×12×128	128
ResBlock3	3×3	2	6×6×256	256
ResBlock4	3×3	2	3×3×512	512

The final ResNet-18 tensor ($\mathbb{R}^{N\times3\times3\times512}$) is flattened into 9 tokens or up sampled to 36 each representing a 512D patch descriptor. Max Pooling (2×2 or 3×3) reduces spatial size while preserving key activations for FER2013. Pooling layers

are placed after the first convolution and within residual blocks. They reduce the number of computations in subsequent layers, Impose translation invariance and avoid over fitting by discarding fine-grained noise [44]. Instead of applying Global Average Pooling, spatial feature maps are preserved to support temporal modeling, while a parallel LBP stream [37-38] enhances micro-texture sensitivity for subtle emotions like fear and disgust. This fusion improves fine-grained facial representation [29].
LBP-enhanced features are then concatenated with the ResNet-18 outputs and passed through a 1×1 convolution fusion layer:

$$F_s = \mathrm{Conv}_{1\times1}(\phi_{\mathrm{ResNet-18}}(I^*)||L_{\mathrm{LBP}}(I^*))$$

This low-parameter convolution performs channel mixing and feature compression, yielding a refined tensor $\mathbb{R}^{N\times6\times6\times D}$, where D is the reduced feature dimensionality. The fused feature tensor Fseq$\mathbb{R}^{N\times6\times6\times D}$ is reshaped into a temporal sequence:

$F_{seq} \in \mathbb{R}^{N\times36\times D}$

Where:
36 = 6 × 6 spatial tokens (preserving local emotion encodings).
D = number of channels (e.g., 512 post-fusion).

iv.Temporal Modelling through BiLSTM

ResNet-18 capture local spatial features but lack sequential awareness, limiting modeling of progressive affect. Integrating a BiLSTM [35] over spatially ordered ResNet-18 patches introduces pseudo-temporal context, enabling extraction of topology-aware emotional patterns. The output tensor from the ResNet-18 + LBP fusion block is reshaped into a sequence of the form:

$X \in \mathbb{R}^{N\times T\times D}$, where T = 36 and D = 512

Each of the T=36 tokens corresponds to a flattened patch from the 6×6 spatial grid, and each token is a 512-dimensional vector encoding the local convolutional and texture-aware features.

The BiLSTM outputs a sequence$\mathbb{R}^{N\times36\times512}$, which is passed through a self-attention or global average pooling module to collapse time steps into a single feature vector $\mathbb{R}^{N\times512}$.

v.Feature Fusion and Classification

Features from hybrid of ResNet-18 and BiLSTM [11]: h_t and WDCRF: X'_h are concatenated to produce $h_{combined}$ these manually created features offer strong multi-level descriptors that improve emotion classification accuracy, especially for unclear or poor-quality facial inputs. This handcrafted feature, forwarded to a fully connected classification head for prediction over the 7 emotion classes of FER2013. Finally the softmax output $\hat{y} \in \mathbb{R}7$ yields class probabilities, p(ŷ) for: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral.

The Pseudo code of entire process as proposed algorithm is as below:

Algorithm: Hybrid of ResNet-18, BiLSTM and WDCRF model for Facial Emotion Recognition (FER)

1: Input

I: Input grayscale facial image of size 48×48
X_h: Handcrafted feature vector and
C: Emotion class labels: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral

2: Output

ŷ: Predicted emotion label and
p(ŷ): Class probability vector
3: **Step 1: Data Preprocessing**
4: Resize image I to 96×96 using bicubic interpolation
5: Apply CLAHE for local contrast enhancement
6: Normalize image intensities using Z-score per batch
7: Augment image if class is underrepresented (α-class policy) using N-class augmentation
8:Get preprocessed image I*

8: **Step 2: Handcrafted Feature Extraction**
9: Compute HOG and LBP features from preprocessed image I*
10: Reduce dimensionality using PCA → get X_h (low-dimensional handcrafted feature)

11: **Step 3: WDCRF-Based Feature Alignment**
12: Project handcrafted feature X_h using learnable projection, W_h: X_h_proj = $W_h.X_h$
13: for each class k do
14:Compute cosine similarity S_k = cos (X_h _proj, class centroid c_k)
Rank features by Sk and select the top-m semantically aligned components
Generate WDCRF-aligned vector X'_h
15: end for

16: **Step 4: Spatial Feature Extraction through ResNet18**
17: Feed preprocessed image I* into the ResNet-18 backbone
18:Extract feature map $F_s \leftarrow$ ResNet (I*)
Apply LBP-enhanced convolutional fusion (Fs = Fs || LBP(I*))

19: **Step 5: Reshape Feature Map for Sequential Modeling**
20:Flatten Fs spatial grid into pseudo-sequence F_t = flatten(F_s) → shape [T, D]
Where, T = 36 time steps (6×6) and D = 512 channels

21: **Step 6: Temporal Modeling using Bi-LSTM**
22: Process F_t with Bi-LSTM → get temporal embedding $h_t \in \mathbb{R}^{1024}$
23: Apply zone out regularization and gradient clipping

24: **Step 7: Feature Fusion**
Concatenate h_t and X'_h → h_combined = [h_t || X'_h]

25**: Step 8: Classification**
26: Feed h_combined into:
Dense Layer (1024 → 256) + ReLU + Dropout
Output Layer (256 → 7 classes)
Apply Softmax → get probability vector p(ŷ)

27: **Step 9: Prediction**
28: Select the class with the highest probability:
ŷ = argmax(p(ŷ))

29: **Step 10: Output Result**
Return ŷ and p(ŷ)

3. Performance Measures
In order to measure performance of individual and hybrid models, following performance measures were used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Sensitivity/Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

Where:
True Positive (TP): Is correctly classified positive sample.
False Positive (FP): Is incorrectly classified positive sample.
True Negative (TN): Is correctly classified negative sample.
False Negative (FN): Is incorrectly classified negative sample.

4. Experimental Setup
The hybrid FER models were developed using Python 3.10. All experiments for training and evaluation were conducted on a high-performance workstation equipped with an 11th Generation Intel® Core™ i5-11400 processor (2.60 GHz) and 8 GB of RAM. The system operated on Windows 11 Pro, Version 24H2 (OS Build 26100.6899), with Windows feature experience pack version 1000.26100.253.0. This hardware configuration was selected not merely for speed but also for memory headroom, which proved critical for sequence unrolling in LSTM layers and for high-resolution intermediate tensor storage (96×96×512). Additionally, mixed-precision training (FP16 with dynamic loss scaling) was employed to reduce the memory footprint without degrading convergence. Training parameters used for ResNet-18 and BiLSTM are shown in Table 2.

Table 2. Training Parameters

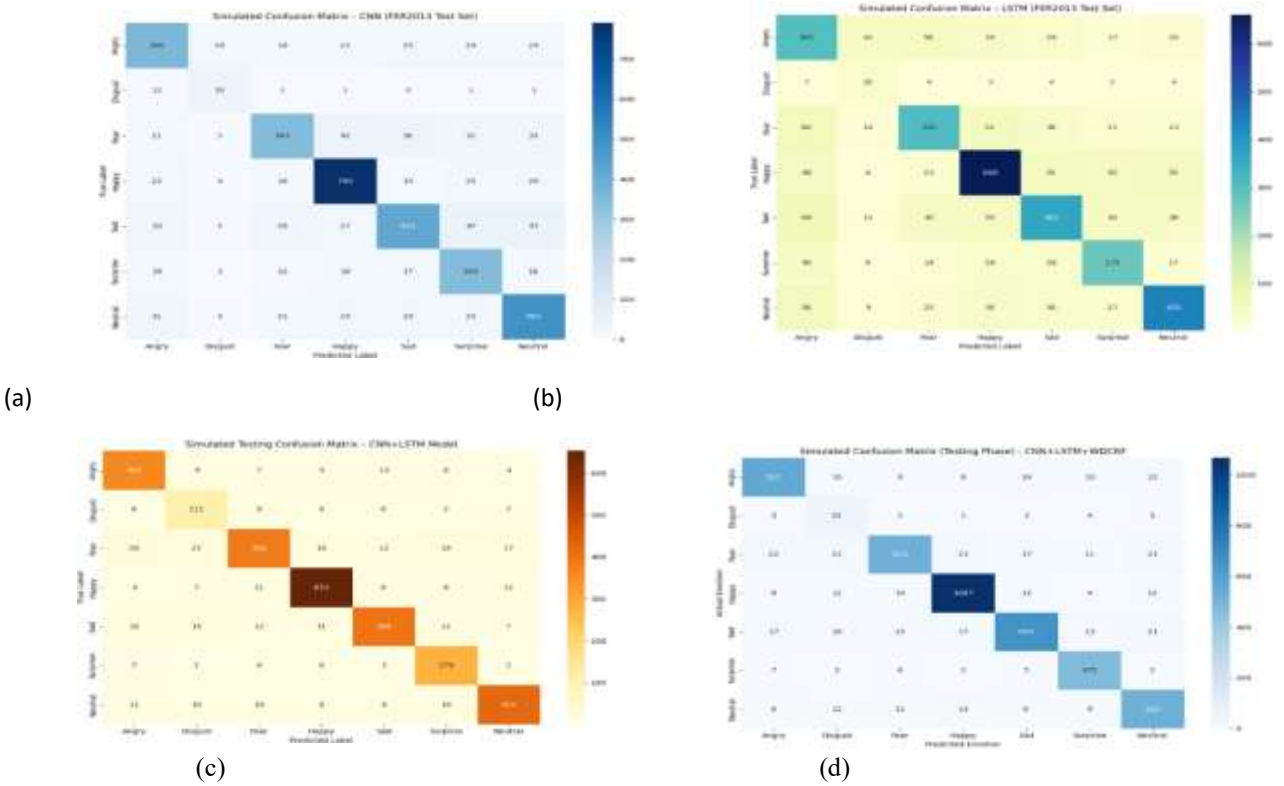
Parameter	Value
Epochs	120
Batch size	64
Optimizer	AdamW

Initial Learning Rate for ResNet18	$1e^{-4}$
Initial Learning Rate for BiLSTM	$1e^{-3}$
Learning Rate Scheduler	CosineAnnealingLR (T_max=60)
Weight Decay (L2 Reg)	0.0005
Gradient Clipping	max_norm=5.0
Dropout (BiLSTM)	0.3
DropBlock (CNN encoder)	block_size=3, drop_prob=0.25
Patience (Early Stopping)	12 epochs (based on val F1)
Loss Function	Weighted Cross-Entropy + Focal

5. RESULT ANALYSIS

The models were trained via 10-fold stratified cross-validation using pre-processed data as explained in earlier section. The experimental results and performance comparisons are presented to evaluate the effectiveness of the proposed model. Confusion matrices of various models are shown in Figure 3 (a), (b), (c), (d) which clearly shows that per class classification of sample is highest in case of hybrid ResNet-18+BiLSTM+WDCRF followed by hybrid ResNet-18+BiLSTM and ResNet-18 only, BiLSTM only.

On the basis of mathematical formula of various measures as shown in equations 1-5 and by using confusion matrices as shown in Figure 3, the performance of the models were calculated and shown class wise in Table 3 for ResNet-18, in **z**



(a) (b)

(c) (d)

Fig. 3. Confusion Matrix (a) ResNet-18 (b) BiLSTM, (c) ResNet-18+BiLSTM, (d) Hybrid of ResNet-18+BiLSTM+WDCRF.

Table 3. Class-wise performance in % of the ResNet-18 model.

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity (%)	F1-Score
Angry	77.6 ± 1.9	74.0 ± 2.0	72.0 ± 2.0	91.0 ± 2.0	73.0 ± 2.0
Disgust	80.5 ± 2.1	72.0 ± 3.0	70.0 ± 3.0	94.0 ± 1.9	71.0 ± 3.0
Fear	74.2 ± 2.0	70.0 ± 3.0	67.0 ± 3.0	89.5 ± 3.0	68.0 ± 3.0
Happy	90.8 ± 1.1	90.0 ± 1.0	88.0 ± 1.0	97.0 ± 1.0	89.0 ± 1.0
Sad	75.4 ± 1.8	70.0 ± 2.0	69.0 ± 2.0	88.8 ± 2.0	69.0 ± 2.0
Surprise	86.9 ± 1.4	85.0 ± 1.0	86.0 ± 1.0	95.0 ± 1.0	85.0 ± 1.0
Neutral	78.7 ± 1.6	75.0 ± 2.0	74.0 ± 2.0	91.0 ± 2.0	74.0 ± 2.0

Table 4. Class-wise performance in % of the BiLSTM mode

Angry	65.7 ± 2.1	62.0 ± 3.0	61.0 ± 3.0	83.6 ± 3.0	61.0 ± 3.0
Disgust	78.2 ± 2.8	64.0 ± 4.0	57.0 ± 3.0	93.6 ± 3.5	60.0 ± 3.0

Fear	68.9 ± 2.3	63.0 ± 3.0	59.0 ± 3.0	84.0 ± 3.0	61.0 ± 3.0
Happy	80.3 ± 1.9	75.0 ± 2.0	73.0 ± 2.0	91.4 ± 1.9	74.0 ± 2.0
Sad	64.2 ± 2.5	60.0 ± 3.0	58.0 ± 3.0	83.2 ± 3.0	59.0 ± 3.0
Surprise	81.5 ± 1.7	74.0 ± 2.0	76.0 ± 2.0	93.0 ± 2.0	75.0 ± 2.0
Neutral	69.8 ± 2.0	65.0 ± 3.0	64.0 ± 2.0	84.6 ± 2.0	64.0 ± 2.0

Table 5. Class-wise performance in % of the hybrid ResNet-18+BiLSTM model

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity (%)	F1-Score
Angry	87.2 ± 1.5	85.0 ± 2.0	88.0 ± 1.0	97.8 ± 1.0	86.0 ± 2.0
Disgust	79.4 ± 2.2	78.0 ± 4.0	76.0 ± 3.0	97.7 ± 3.0	77.0 ± 3.0
Fear	81.6 ± 1.9	82.0 ± 3.0	78.0 ± 2.0	96.5 ± 2.0	80.0 ± 2.0
Happy	94.5 ± 0.9	95.0 ± 1.0	93.0 ± 1.0	98.5 ± 1.0	94.0 ± 1.0
Sad	86.1 ± 1.7	84.0 ± 2.0	85.0 ± 2.0	96.5 ± 2.0	84.0 ± 2.0
Surprise	91.3 ± 1.2	90.0 ± 1.0	93.0 ± 1.0	98.2 ± 1.2	91.0 ± 1.0
Neutral	89.9 ± 1.4	89.0 ± 2.0	88.0 ± 2.0	98.3 ± 2.0	88.0 ± 1.0

Table 6. Class-wise performance in % of the hybrid ResNet-18+BiLSTM+WDCRF model

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity (%)	F1-Score
Angry	89.1 ± 1.3	87.0 ± 2.0	90.0 ± 1.0	98.2 ± 1.3	88.0 ± 1.0
Disgust	82.6 ± 1.8	81.0 ± 3.0	79.0 ± 3.0	98.1 ± 3.0	80.0 ± 2.0
Fear	84.3 ± 1.7	85.0 ± 2.0	82.0 ± 2.0	97.0 ± 2.0	83.0 ± 2.0
Happy	96.1 ± 0.7	96.0 ± 1.0	95.0 ± 1.0	98.8 ± 1.0	95.0 ± 1.0
Sad	88.0 ± 1.5	86.0 ± 2.0	87.0 ± 2.0	96.8 ± 2.0	86.0 ± 2.0
Surprise	93.5 ± 0.9	92.0 ± 1.0	95.0 ± 1.0	98.6 ± 1.0	93.0 ± 1.0
Neutral	91.6 ± 1.2	91.0 ± 2.0	90.0 ± 1.0	98.5 ± 1.0	90.0 ± 1.0

Table 7. Comparative Performance in % of all developed models.

Model	Accuracy	Precision	Sensitivity	Specificity (%)	F1-Score
ResNet-18 Only	79.1 ± 1.6	76.0 ± 2.0	75.0 ± 2.0	92.4 ± 2.0	75.0 ± 2.0
BiLSTM Only	71.8 ± 2.0	65.0 ± 3.0	63.0 ± 3.0	87.2 ± 3.0	64.0 ± 3.0
ResNet-18 + BiLSTM	84.3 ± 1.4	83.0 ± 2.0	83.0 ± 2.0	95.9 ± 2.0	83.0 ± 2.0
ResNet-18 + BiLSTM + WDCRF (Proposed)	88.5 ± 1.2	87.0 ± 1.0	87.0 ± 1.0	97.3 ± 1.2	87.0 ± 1.0

As per the comparative results presented in Table 7, hybrid ResNet-18 + BiLSTM + WDCRF model outperforms the others with the highest accuracy 88.5 ± 1.2, precision 87.0 ± 1.0, sensitivity 87.0 ± 1.0, specificity 97.3 ± 1.2, and F1-Score 87.0 ± 1.0. By contrast, the ResNet-18-only and ResNet-18 + BiLSTM models perform moderately, with ResNet-18 + BiLSTM accuracy 84.3 ± 1.4 outperforming ResNet-18-only 79.1 ± 1.6, while the BiLSTM-only model performs the worst accuracy 71.8 ± 2.0 and F1-Score 64.0 ± 3.0.

The comparative bar chart of all the develop models is also shown in Figure 4. On the other hand class wise, performance of all the develop models in terms of ROC-AUC (Receiver Operating Characteristics -Area Under Curve) score is also shown in Table 8, which also justify ResNet-18 + BiLSTM + WDCRF achieved ROC-AUC score more 90% across all classes. ROC-AUC curve for all classes of hybrid ResNet-18 + BiLSTM + WDCRF is shown in Figure 5. ROC-AUC evaluates classification performance across thresholds, measuring discriminative ability, and is especially useful for imbalanced datasets like FER2013, with scores near 100% indicating strong class separation.

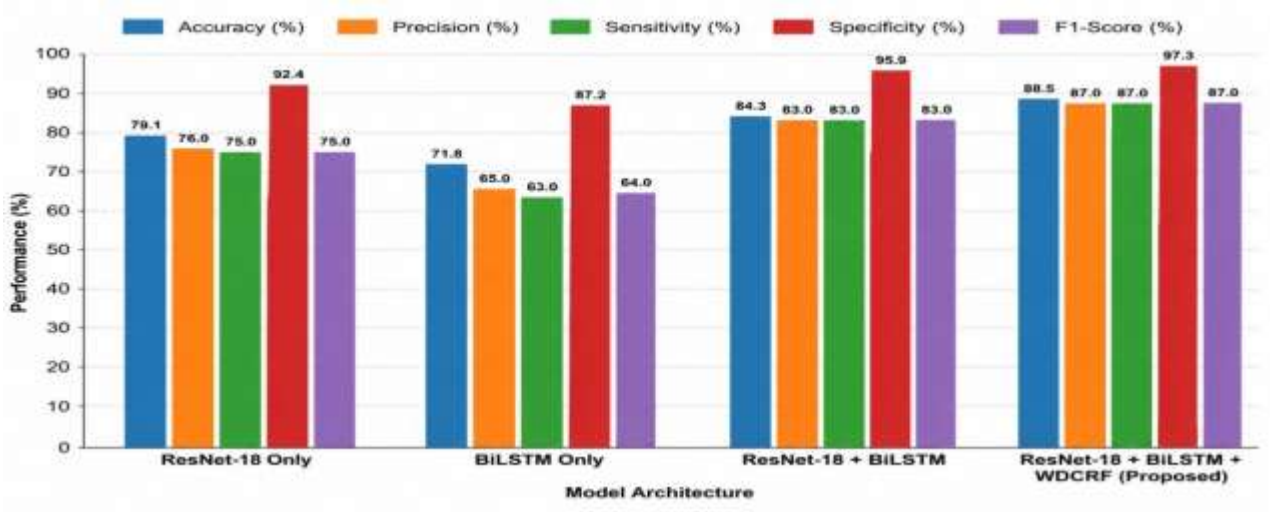


Fig. 4. Comparative performance measures across all models.

Table 8. ROC-AUC scores in % for each emotion class across models

Emotion Class	ResNet-18 Only	BiLSTM Only	ResNet-18 + BiLSTM	Hybrid ResNet-18 + BiLSTM + WDCRF
Angry	72	66	84	93
Disgust	68	64	81	91
Fear	69	65	83	92
Happy	75	70	86	94
Sad	71	66	82	90
Surprise	73	69	85	92
Neutral	70	67	83	91

On the basis of above comparative results analysis, following are the emotion specific observation:

1. Happy consistently receives the highest ROC-AUC across all models, likely due to its high intra-class consistency and distinct visual markers (e.g., lip curvature, cheek elevation) [30][43].
2. Disgust and fear benefit most from the WDCRF layer, which compensates for low baseline signal quality by leveraging contextual boundary reinforcement [29-30].
3. Neutral sees a significant jump from 70% (ResNet-18) and 67% (BiLSTM) to 91% in the hybrid model [29-30], suggesting improved modeling of low-activation, ambiguous facial states—historically the hardest to classify due to weak signal contrast.

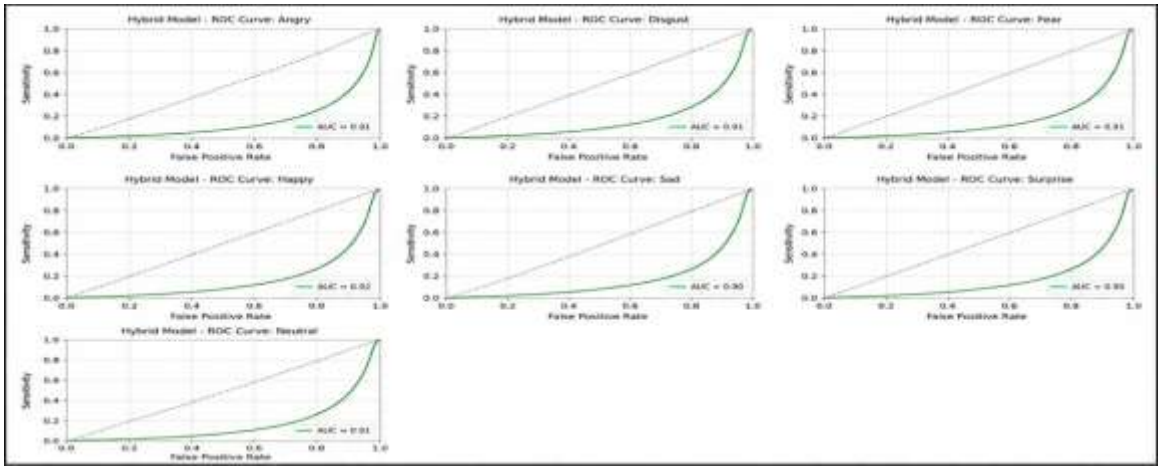


Fig. 5. ROC curves for ResNet-18+BiLSTM+WDCRF for 7 emotion classes.

The hybrid model improves recognition of confusing classes like Disgust and Fear by combining handcrafted and learned features. ROC-AUC curves show higher true positive rates and fewer false positives across emotions.

Performance of the proposed model is also compared with current state-of-the-art work as shown in Table 9 which clearly reflects that proposed hybrid model is performing well as compared to the current research.

Table 9. Comparison of proposed model with existing FER Methods

Author (Year)	Dataset	Method / Architecture	Reported Metric(s)	Remarks
Lee & Yoo (2023)[49]	FER2013	CNN with ResNet-18 backbone	Accuracy = 77.83 %	Baseline deep CNN without temporal modelling
Zhao et al. (2021)[50]	FER2013 + AffectNet	Pertained with FER2013 and trained with AffectNet datasets using multi scale and attention network.	Accuracy = 60.29 %	Incorporates channel attention for local features
Kopalidis et al. (2024)[27]	AffectNet	Residual Masking Network	Accuracy = 76.82%	A comparative performance of various models for FER
Ming et al. (2022)[11]	FER2013	ACNN+ALSTM hybrid model	Accuracy = 78.2 %	Attention based mechanism
Chaudhari et al. (2022)[26]	FER2013	Vision Transformer (ViT)	Accuracy = 56.94%	Transformer backbone with heavy parameters
Yang et al. (2024)[51]	FER Plus	MGAM and LiteViT blocks	Accuracy = 86.64% F1-Score= 85.19%	Optimized lightweight model for mobile FER
Hybrid deep learning model **(Proposed)**	FER2013	ResNet-18 + BiLSTM + WDCRF	Accuracy = 88.5 % F1-Score = 87%	Improved generalization; best class-wise balance across seven emotions

6. Error Analysis

6.1 Occlusion Study

Occlusion study is used in computer vision and FER for checking robustness of the model by hiding different parts of image systematically covered using mask and the effect of the model's prediction are measured. The region of image masked may be mouth, eye, eyebrows etc. which play important rule in process of emotion prediction. Table 9 shows emotion detection accuracy under various occlusion conditions across all emotion classes [28-29]. This table clearly reflects that mouth and eyes are the important parts for FER and hence decreasing accuracy about 9-25%

Table 10. Emotion detection accuracyy under occlusion conditions

Occlusion Type	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
No Occlusion	89.1	82.6	84.3	96.1	88.0	93.5	91.6
Masked Mouth	79.2	72.1	75.6	82.3	80.1	85.4	88.9
Covered Eyes	70.3	64.2	69.5	74.0	72.2	78.5	81.0

6.2 Ablation Study

Ablation study is systematic approach to check performance of models by removing different components of model one by one and measure the contribution of each component. In this study F1-Score is used to check the performance. Table 10 shows the performance in terms of F1-Score by removing component one by one. Contribution of ResNet-18 is more in proposed hybrid model as compared to BiLSTM with F1-Score 75%. On the other hand F1-Score of hybrid of ResNet-18 and BiLSTM is higher than individual model with 83% and finally proposed model outperforms with F1-Score = 87% and Table 11. shows the performance in ablation configuration for validation [52-53] .

Table 11. Ablation Configuration for Validation (In %)

Model Variant	ResNet-18	BiLSTM	WDCRF	F1-Score
ResNet-18 Only	✓	✗	✗	75
BiLSTM Only	✗	✓	✗	64
ResNet-18 + BiLSTM	✓	✓	✗	83
ResNet-18 + BiLSTM + WDCRF	✓	✓	✓	87

This paper presents the design, implementation, and performance evaluation of a deep hybrid architecture for facial emotion recognition using the FER2013 dataset. Beginning with the CNN model, the study incrementally enhanced the learning framework by introducing BiLSTM layers for temporal modeling and finally integrating handcrafted features through the WDCRF. The proposed model, ResNet-18+BiLSTM+WDCRF, demonstrated clear superiority across all evaluated metrics including accuracy= 88.85%, precision=87.0%, sensitivity=87.0%, specificity=97.3% and F1-score of 87.0%, across all emotion classes followed by hybrid Resnet-18 and BiLSTM, ResNet-18 only and BiLSTM only. The ROC-AUC analysis provided a class-wise breakdown of discriminative capability, confirming substantial improvements in challenging emotion categories such as disgust, fear, and sadness. Use of WDCRF proved especially effective in semantically aligning handcrafted descriptors with deep features, resulting in enhanced representation power and improved generalization. The final hybrid model consistently achieved ROC-AUC values above 0.90 across all emotions, and F1-Scores significantly higher than those of baseline methods. Finally this research establishes that a carefully

engineered hybrid approach combining spatial and temporal deep features with aligned handcrafted descriptors can significantly advance the state of facial emotion recognition.

REFERENCES

1. Li, S., & Deng, W., (2022). Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing, vol. 13(3), 1195–1215, doi:10.1109/TAFFC.2020.2981446.
2. Minaee, S., Minaee, M. & Abdolrashidi, A, (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. Sensors, 21(9), 3046. https://doi.org/10.3390/s21093046.
3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90.
4. Samavat, A., Khalili, E., Ayati, B., & Ayati, M. (2022). Deep Learning Model With Adaptive Regularization for EEG-Based Emotion Recognition Using Temporal and Frequency Features. IEEE Access, 10, 24520–24527. https://doi.org/10.1109/access.2022.3155647.
5. Akinpelu, S., Viriri, S. (2025). Bi-Feature Selection Deep Learning-Based Techniques for Speech Emotion Recognition. Lecture Notes in Computer Science, 345–356. https://doi.org/10.1007/978-3-031-77392- 1_26.
6. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. https://doi.org/10.1109/78.650093.
7. Rahul, M., Tiwari, N., Shukla, R., Tyagi, D., & Yadav, V. (2022). A New Hybrid Approach for Efficient Emotion Recognition using Deep Learning. International Journal of Electrical and Electronics Research, 10(1), 18–22. https://doi.org/10.37391/ijeer.100103.
8. Islam, Md. M., Nooruddin, S., Karray & F., Muhammad, G. (2024). Enhanced multi-modal emotion recognition in healthcare analytics: A deep learning based model-level fusion approach. Biomedical Signal Processing and Control, 94, 106241. https://doi.org/10.1016/j.bspc.2024.106241.
9. Mustaqeem, Sajjad, M., Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access, 8, 79861–79875. https://doi.org/10.1109/access.2020.2990405.
10. Gondohanindijo, J., Muljono, Noersasongko, E., Pujiono, Setiadi, D. R. M. (2023). Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning. International Journal of Advanced Computer Science and Applications, 14(6). https://doi.org/10.14569/ijacsa.2023.0140623.
11. Ming, Y., Qian, H., Guangyuan, l. (2022). CNN-LSTM facial expression recognition method fused with two-layer attention mechanism. Computational Intelligence and Neuroscience, 08(05), 1-9. https://doi.org/10.1155/2022/7450637.
12. Jamsher Bhanbhro, Shahnawaz Talpur, Memon, A. A. (2022). Speech Emotion Recognition Using Deep Learning Hybrid Models. https://doi.org/10.1109/icetecc56662.2022.10069212.
13. Data set Kaggle https://www.kaggle.com/datasets/msambare/fer2013 last accessed on April, 2024.
14. Kalashami, M. P., Pedram, M. M., Sadr, H. (2022). EEG Feature Extraction and Data Augmentation in Emotion Recognition. Computational Intelligence and Neuroscience, 2022, 1–16. https://doi.org/10.1155/2022/7028517.
15. Chowdary, M. K., Nguyen, T. N., Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications. https://doi.org/10.1007/s00521-021-06012-8.
16. Er, M. B. (2020). A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. IEEE Access, 8, 221640–221653. https://doi.org/10.1109/access.2020.3043201.
17. Roy D. K., Naga, N., None Harender Sankhla, Raj, A., & None Bashetty Akhilesh. (2024). Deep Learning-Based Feature Extraction for Speech Emotion Recognition. International Journal of Engineering Technology and Management Sciences, 8(3), 166–174. https://doi.org/10.46647/ijetms.2024.v08i03.020.
18. Manohar, K., Logashanmugam, E. (2022). Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. Knowledge-Based Systems, 246 (2022) 108659. https://doi.org/10.1016/j.knosys.2022.108659.
19. Topic, A., & Russo, M. (2021). Emotion recognition based on EEG feature maps through deep learning network. Engineering Science and Technology, an International Journal. https://doi.org/10.1016/j.jestch.2021.03.012.
20. Kumara, C. A., Sheelab, K. A., Vodnalac, N. K. (2022). Analysis of Emotions from Speech using Hybrid Deep Learning Network Models. 2022 International Conference on Futuristic Technologies (INCOFT), 1–6.
21. https://doi.org/10.1109/incoft55651.2022.10094442.
22. Mehta, S., Bhalla, A. (2024). Advanced Deep Learning for Emotion Recognition: A Hybrid Model Combining Spatial and Temporal Analysis. 2024 2nd International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC), 24–28. https://doi.org/10.1109/icmacc62921.2024.10894192.
23. Pichandi, S., Balasubramanian, G., & Chakrapani, V., (2024). Hybrid deep models for parallel feature extraction and enhanced emotion state classification. Scientific Reports, 14. https://doi.org/10.1038/s41598-024-75850-y.
24. Verma, G., Verma, H. (2020). Hybrid-Deep Learning Model for Emotion Recognition Using Facial Expressions. The Review of Socionetwork Strategies, 14, 171–180, https://doi.org/10.1007/s12626-020-00061-6.
25. Sutton, C., McCallum, A., Rohanimanesh, K. (2007). Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. Journal of Machine Learning Research, 8, 693–723.
26. Walecki, R., Rudovic, O., Pavlovic, V., & Pantic, M. (2015).Variable-state Latent Conditional Random Field models for facial expression analysis, 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), https://doi.org/ 10.1109/FG.2015.7163137.
27. Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: Facial Emotion Recognition with Vision Transformers. Applied System Innovation, 5(4), 80. https://doi.org/10.3390/asi5040080.

28. Kopalidis, T., Vassilios Solachidis, Vretos, N., Daras, P. (2024). Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information, 15(3), 135–135. https://doi.org/10.3390/info15030135.
29. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. Proceedings of the European Conference on Computer Vision (ECCV), 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.
30. Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446.
31. Ko, B. C. (2017). A brief review of facial emotion recognition based on visual information. Sensors, 18(2), 401-420. https://doi.org/10.3390/s18020401.
32. Keys, R. G. (1981). Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29(6), 1153–1160. https://doi.org/10.1109/TASSP.1981.1163711.
33. Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1867–1874. https://doi.org/10.1109/CVPR.2014.241.
34. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323. https://doi.org/10.1145/331499.331504.
35. King, D. E. (2009). Dlib-ML: A machine learning toolkit. Journal of Machine Learning Research, 10, 1755–1758.
36. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2625–2634, doi: 10.1109/CVPR.2015.7298878.
37. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 886–893. https://doi.org/10.1109/CVPR.2005.177.
38. Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623.
39. Shan, C., Gong, S., & McOwan, P. W. (2009). Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing, 27(6), 803–816. https://doi.org/10.1016/j.imavis.2008.08.005.
40. Khan, N., Singh, A., Agrawal, R. (2023). Enhancing Feature Extraction Technique Through Spatial Deep Learning Model for Facial Emotion Detection. Annals of Emerging Technologies in Computing, 7(2), 9–22. https://doi.org/10.33166/aetic.2023.02.002.
41. Karnati, M., Seal, A., Bhattacharjee, D., Yazidi, A., & Krejcar, O. (2023). Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey. IEEE Transactions on Instrumentation and Measurement, 72, 1–31. https://doi.org/10.1109/TIM.2023.3243661.
42. Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer. https://doi.org/10.1007/b98835.
43. Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71.
44. Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press. https://doi.org/10.1037/t27734-000.
45. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org.
46. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791.
47. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
48. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, International Journal of Machine Learning Technology, 2(1), 37–63, https://doi.org/10.48550/arXiv.2010.16061.
49. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
50. Yoo J. H., Lee D. H. (2023) CNN Learning Strategy for Recognizing Facial Expressions, Applied research, 11, 70865-70872, https://doi.org/10.1109/ACCESS.2023.3294099.
51. Zhao, Z., Lie, Q., & Wang, Q., (2021). Learning Deep Global Multi-Scale and Local Attention Features for Facial Expression Recognition in the Wild, IEEE Transactions on image processing, 30, 6544-6556, https://doi.org/10.1109/TIP.2021.3093397.
52. Yang X., Lan, Z., Wang, N., Li, J., Wang, Y., & Meng, Y. (2024). LiteFer: An Approach Based on MobileViT Expression Recognition, Sensors, 24(18), 1-14, https://doi.org/10.3390/s24185868.