**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Vision Transformer with Local Binary Pattern: A Contrastive Learning-Based Local–Global Hybrid Model for Facial Expression Recognition

Meriem SARI [1], Abdelouahab MOUSSAOUI [1]

[1] Department of Computer Science, University of Ferhat Abbas Setif1, Setif, 19000, Algeria.

**ARTICLEINFO**

Received: 18 Nov 2025
Revised: 15 May 2026
Accepted: 10 July 2026

**ABSTRACT**

Facial expression recognition is a critical component in affective computing and human-computer interaction. Automated facial expression recognition systems have shown impressive results lately due to Deep Learning innovated techniques. On the other hand, handcrafted methods still offer a good ability to extract micro patterns. This paper proposes a novel hybrid framework that integrates Local Binary Pattern (LBP) texture features with a Vision Transformer (ViT) backbone as one of the most effective deep learning approachs in computer vision; optimized using a joint loss function combining cross-entropy and supervised contrastive learning. The LBP features enhance local texture representation, while the ViT captures global spatial dependencies. The supervised contrastive loss promotes discriminative feature embeddings by maximizing inter-class separation and minimizing intra-class variance. Experimental evaluation on two benchmark datasets namely CK+ and JAFFE, demonstrates the superiority of the proposed method, achieving 97.45% and 95.24% accuracy respectively. These results highlight the effectiveness and generalizability of combining handcrafted and deep features in automated facial expression recognition tasks.

**Keywords**: Vision Transformers ; Facial Expression Recognition ; Deep Learning ; Local Binary Patterns ; Supervised Contrastive Learning

## 1. INTRODUCTION

Facial expression is a non-verbal communication behavior for human beings that offers insight into human's emotional states and plays a key role in social interactions. Facial Expression Recognition (FER) has gained importance in various domains, including human-computer interaction, behavioral analysis, security systems, and healthcare monitoring. The foundational concept of interpreting facial expressions dates back to Darwin's early observations on the universality of emotions [1], later extended by Paul Ekman's identification of six primary emotions that are expressed consistently across cultures: happiness, sadness, anger, fear, disgust, and surprise [2].

Traditional FER systems generally consist of three major stages: detecting and capturing facial images, extracting informative features, and classifying these features into different emotion classes. Feature extraction methods traditionally fall into two groups: geometric-based techniques [3–8], which model the shape and relative position of facial landmarks, and appearance-based techniques [6, 8–11], which describe local texture changes on the face [8]. However, variations in lighting, occlusion, head pose, and individual facial morphology continue to pose challenges for conventional approaches [12, 13].

In the last decade, image classification field has progressively shifted toward deep learning techniques. While Convolutional Neural Networks (CNNs) have dominated the literature due to their ability to capture spatial patterns [14], they often struggle to model long-range dependencies within facial dynamics. This limitation has motivated the exploration of alternative architectures such as vision transformers (ViTs). ViTs have emerged as a powerful alternative to CNN in the field of image classification [15] taking advantage of self-attention mechanism to model long-range dependencies in image data [16]. Unlike CNNs, which focus on local receptive fields, ViTs divide images into patches and process them similarly to words in Natural Language Processing (NLP). Pre-trained ViT models, such as those trained on large-scale datasets like ImageNet, offer rich and transferable representations that can be fine-tuned for downstream tasks with limited data. To further enhance representation learning, contrastive learning

has proven effective as a self-supervised technique; it aims to bring semantically similar samples closer in the feature space while pushing dissimilar ones apart. Contrastive learning improves generalization and robustness by enhancing the model's ability to capture invariant and distinctive features especially when integrated with deep models [17]. Computer vision also takes benefit of handcrafted on-the-shelf methods like Local Binary Patterns (LBP) which is considered as a texture descriptor that helps capture micro-patterns [18]; it offers strong performance due to its simplicity, low computational cost and invariance to illumination changes.

In this paper, we propose a hybrid FER framework that integrates handcrafted LBP features with a pre-trained Vision Transformer backbone further optimized through contrastive learning. This model aims to capture fine-grained facial textures while modeling contextual dependencies over space. This integration enhances the discriminative power of the model and addresses common FER challenges such as variability in expressions and environmental conditions. The handcrafted LBP histograms serve as auxiliary input channels, complementing the visual features extracted by ViT; this fusion aims to achieve better accuracy, precision and recall.

The remainder of the paper is structured as follow: following section consists of the most recent works that have been conducted. Section three will explain Methodology used in this paper. Section four will show the Experimental Setup. Section five depicts the Experimental Results and Discussion, and at the end a Conclusion.

## 2. RELATED WORKS

Several recent FER studies report competitive performance in terms of accuracy and computational complexity. For instance [19] proposed a hybrid model for facial emotion classification using an AdaBoost-based Random Forest (ARFEC). The approach employs the 2D Ortho-normal Stockwell Transform (DOST) to extract discriminative time-frequency features from facial images, followed by bivariate t-tests for feature selection. The ensemble classifier combines AdaBoost's adaptive learning with the robustness of Random Forests to improve accuracy. Evaluated on CK+, FER2013, and Flickr8k, the model achieved accuracies of 92.5% on CK+.

[20] introduced a lightweight four layer ConvNet (four convolutional layers + two fully connected layers) trained in minimal epochs, emphasizing the significance of data diversity across datasets. Their model detects seven emotions and converges rapidly, achieving 92.05% on JAFFE and 98.13% on CK+. The authors highlight that diversity in training data significantly enhances generalization and real-time applicability.

[21] Presented an enhanced CNN based FER model characterized by a novel deep convolutional architecture that integrates face landmark detection, affine transformations, attention modules, and circular derivative Local Binary Patterns (LBP) for robust feature extraction. Evaluated on MA MFD, JAFFE, FER2013, and CK+ datasets, it achieves an average accuracy of 95.63%, outperforming previous methods.

[22] proposed a facial expression recognition method that extracts Gabor filter features from facial regions of interest identified via landmarks and employs a genetic algorithm to simultaneously optimize SVM hyper-parameters and select the most discriminative features. Evaluated on JAFFE, CK, and CK+ datasets, the model achieves 96.30% on JAFFE dataset and 94.26% CK+.

[23] introduced SSAE FER, a lightweight facial expression recognition model using a stacked sparse auto-encoder for unsupervised pre-training and supervised fine tuning, followed by a softmax classifier. The approach autonomously extracts high-level features, mitigating typical challenges like illumination changes, occlusion, and overfitting. Tested across seven basic emotions, SSAE FER achieves 92.50% on JAFFE and 99.30% on CK+ datasets.

[24] introduced a novel lightweight CNN architecture for facial expression recognition, incorporating a feature boosting block with dense skip connections and translation modules to emphasize expression specific features and reduce redundancy. The model includes extensive data augmentation and transfer learning: pretrained on FER 2013 and fine tuned on smaller datasets like JAFFE and CK+. The architecture uses minimal trainable parameters, enabling real time application with low runtime cost compared to heavier CNNs. It achieved 96.16% on JAFFE, and 96.52% on CK+.

[25] introduced a hybrid geometric and deep learning framework that extracts both facial landmark geometry and convolutional features, fusing them and classifying with an SVM to enhance discriminative power in facial expression recognition. the method emphasizes geometric features derived from landmark point positions to capture spatial

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

movements while leveraging CNNs for rich texture representation. Experiments on benchmark datasets show significant accuracy, it achieved 96.19% on CK+ and 89.23% on JAFFE. introduced SCL FExR, a supervised contrastive learning framework tailored for facial expression recognition that effectively exploits label information to enhance inter-class separability and intra-class compactness, addressing limitations of standard cross entropy training. Designed to integrate with relatively lightweight CNN backbones, the method emphasizes robustness and noise resistance rather than competing with heavily parameterized deep models. The contribution underscores how supervised contrastive loss can boost feature discrimination in FER while keeping the model relatively simple. Its strengths lie in enhancing robustness under noisy labels and limited-complexity settings. Compared to classical CNN only models, SCL FExR's data-centric contrastive strategy offers a meaningful performance and stability gain without heavily increasing model complexity.

## 3. METHODOLOGY

In this paper we propose a novel and effective hybrid FER architecture that integrates LBP descriptor within a vision transformer based deep learning model. The system takes benefit of LBP feature descriptor power and the sequential modeling capabilities of ViTs in order to enhance the performance and increase the accuracy of the model.

The overall architecture depicted in Figure 1 consists of three main components: the Vision Transformer encoder, the LBP histogram encoder, and a fusion-based classification head. The process begins by feeding each grayscale facial image through a pre-processing pipeline, where it is resized and replicated across three channels to conform to ViT input requirements. The grayscale image is also processed using the standard uniform LBP operator while ViT encoder processes the same image (converted to RGB). The LBP histogram vector and the ViT [CLS] embedding are then concatenated to form a unified feature representation that encompasses both local texture cues and global semantic structure. This combined vector is passed through a fully connected feed-forward network composed of one hidden layer. Finally, an output layer produces the predicted emotion class.

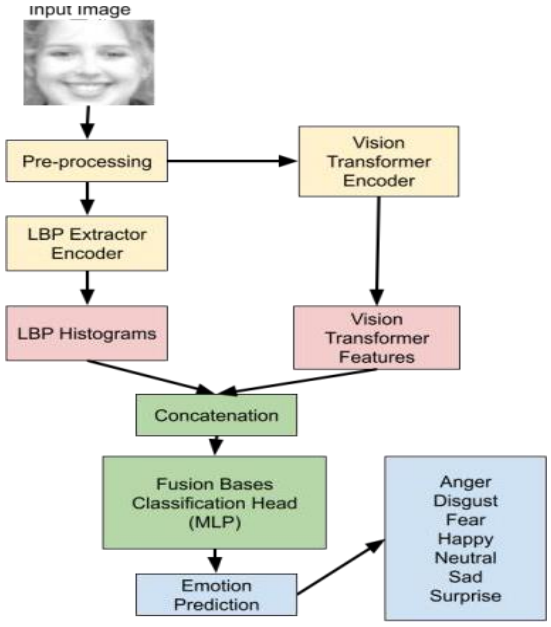


**Figure 1**. General Architecture of the Model

### 3.1 LBP Feature Extraction

In this work we adopt the standard uniform LBP operator, with $P = 8$ sampling points and a radius of $R = 1$, the system thresholds each 3*3 neighborhood around a central pixel resulting a series of values of consecutive 1 or 0, the resulting binary patterns are flattened and converted into normalized histograms computed over the entire image. These histograms serve as handcrafted feature vectors that are passed to the model alongside the image input as shown in Figure 2.

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

For a given pixel $I_c$ in a grayscale image, the LBP operator thresholds its $P$ neighbouring pixels $\{I_p\}_{\{p=0\}}^{\{P-1\}}$ at radius $R$, producing a binary code:

$$LBP_{\{P,R\}}(x_c, y_c) = \sum_{p=0}^{P-1} s(I_p - I_c) * 2^p \quad (1)$$

Where

$$s(x) = \{^{1\ if\ x \geq 0}_{0\ if\ x < 0}\}$$

And $(x_c, y_c)$ denote the coordinates of the central pixel.

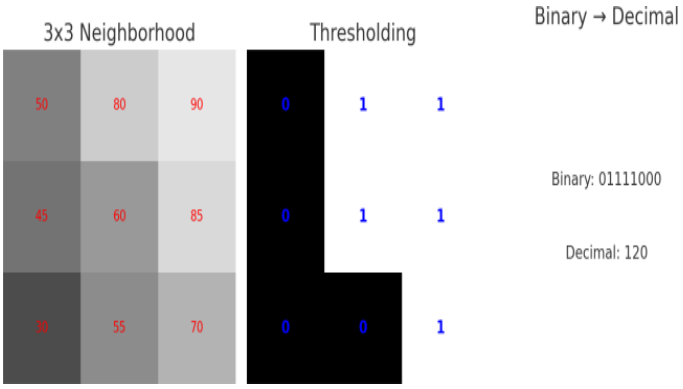


**Figure 2.** CK+ and JAFFE Dataset Samples

### 3.2 Vision Transformer Backbone

The core of the proposed model leverages the Vision Transformer (ViT) architecture; in our implementation, we utilize the ***vit_base_ patch16_224*** model from the ***timm*** library, which has been pre-trained on the ImageNet-1K dataset.

Given an input face image $I \in R^{\{H*W*C\}}$, the image is partitioned into a sequence of fixed-size non-overlapping patches:

$$x_p \in R^{\{N*(P^2.C)\}},\ N = \frac{HW}{P2} \quad (2)$$

Where $P$ denotes the patch size ($P = 16$ in our case), and $N$ is the number of patches. Each patch is flattened and linearly projected into a dimensional embedding. The embeddings are then passed through $L$ layers of transformer encoders, each composed of a multi-head self-attention (MSA) module and a feed-forward network. This high-dimensional feature vector captures global contextual relationships across the face, such as coordinated changes between eyebrows, eyes, and mouth regions.

For the purpose of hybrid feature fusion, we discard the final classification head of the pre-trained ViT and retain only the encoder output to extract the high-level feature embeddings. These embeddings represent the global visual characteristics of the face image and are concatenated with handcrafted LBP features to enrich the representation.

### 3.3 Fusion-Based Classification Head

The fusion-based classification head is the final stage of the proposed hybrid facial expression recognition system. Its primary role is to integrate the feature representations from two distinct encoders: the Vision Transformer (ViT) and the Local Binary Pattern (LBP) histogram encoder; and to predict the facial expression class with high accuracy. At this stage the system concatenates deep global features obtained from the [CLS] token $f_{Vit}$ of the ViT encoder and the

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

local handcrafted features represented by the normalized LBP histogram $f_{LBP}$. This results in a single fused feature vector combining both learned and handcrafted information.

$f_{Fusion} = [f_{Vit} \| f_{LBP}]$

The fused feature vector is then passed through a lightweight Multi-Layer Perceptron (MLP) which acts as the classifier. The final prediction is obtained by applying a softmax function. This allows the model to leverage both global semantic information and local texture details for more accurate facial expression recognition.

## 4. EXPERIMENTAL SETUP

### 4.1 Data Collection and Pre-processing

In this research, datasets are collected from two public and widely used databases: the Extended Chon Kanade (CK+) and the Japanese Female Facial Expression Database (JAFFE). CK+ dataset contains 981 image frames from 123 subjects, each transitioning from a neutral to a peak emotional expression, seven basic emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. JAFFE dataset consists of 213 grayscale frontal facial images of 10 Japanese female subjects, each subject posed seven basic expressions.

All images are first converted to grayscale and resized to 224*224 pixels to extract LBP features; simultaneously, to align with the ViT encoder input requirements, each image is transformed into RGB format and fed to the ViT branch. Datasets are split into training and validation sets in an 80:20 ratio using stratified random sampling. Figure 3 depicts sample images from both datasets.


***Figure 3.*** *CK+ and JAFFE Dataset Samples*

### 4.2 Hyper-Parameter Tuning and Training Phase

In this research, we utilized a hybrid FER model integrating pre-trained Vision Transformer backbone with Local Binary Pattern (LBP) descriptors. Specifically, we employ the **vit_base_ patch16_224** variant from the **TIMM** library, which consists of 12 transformer layers, 12 self-attention heads, and an embedding dimension of 768. The ViT is pre-trained on ImageNet and used as a fixed feature extractor by freezing all of its parameters during training. The output CLS token embedding from the ViT is concatenated with a 10-bin uniform LBP histogram ($P = 8$). Adam optimizer was adopted, and the network was trained for 30 epochs. The loss function combines Cross-Entropy loss for classification with Supervised Contrastive loss to encourage discriminative feature embedding learning.

### 4.3 Classification and Contrastive Learning

The concatenated feature vector is passed through a multi-layer perceptron (MLP) classifier comprising a hidden layer with a $ReLu$ function, a dropout layer and a final linear layer mapping to the number of emotion classes. During the training phase, we combine two loss functions for more optimization results and to encourage feature discrimination.

**Cross-Entropy Loss (LCE):** For standard supervised multi-class classification, it measures the divergence between the predicted class probabilities and the true labels according to Equation 3 where $C$ is the number of classes,

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

$yi$ is the true label (one hot encoded), and $pi$ is the predicted probability for class $i$. This loss encourages the model to assign high probability to the correct class.

$$L_{\{CE\}} = -\sum_{i=1}^{C} y_i \log(p_i) \quad (3)$$

**Supervised Contrastive Loss (LSupCon):** it encourages the model to learn class aware feature embeddings by minimizing intra-class distances and maximizing interclass distances in the embedding space. It is calculated according to Equation 4 where $zi$ is the anchor embedding, $P(i)$ is the set of positives (samples with the same label as $i$), $A(i)$ is the set of all samples except the anchor, $sim(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature scaling parameter.

$$L_{\{SupCon\}} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log\left( \frac{\exp\left(\frac{sim(z_i, z_p)}{\tau}\right)}{\sum_{a\in A(i)} \exp\left(\frac{sim(z_i, z_a)}{\tau}\right)} \right) \quad (4)$$

The overall function is the sum of both loss functions (Equation 5) where λ1 and λ2 are hyper-parameters that balance the contribution of each term. These weights are chosen empirically to ensure stable convergence and enhance both classification accuracy and embedding discriminability.

$$L_{\{total\}} = \lambda_1 L_{\{CE\}} + \lambda_2 L_{\{SupCon\}} \quad (5)$$

## 5. EXPERIMENTAL RESULTS AND DISCUSSION

This paper investigates the effectiveness of a hybrid framework for Automated Facial Expression Recognition combining two powerful approaches which are handcrafted feature extractor represented by LBP and a deep learning model represented by a pre-trained vision transformer. The model is assessed using accuracy, precision, recall, F1-score in addition to t-distributed Stochastic Neighbor Embedding (t-SNE) visualization. Accuracy and Loss curves of CK+ and JAFFE datasets are shown in Figures 3, 4 respectively. Classification reports are depicted in Tables 1, 2 respectively. Confusion matrices are also provided in Figures 5.

The model demonstrates excellent performance on the CK+ validation set, achieving an overall accuracy of 97.45% with high precision, recall, and F1-scores across all emotion classes. Emotions such as Anger and Disgust are classified perfectly, with scores of 1.00 across all metrics, indicating that these expressions are well represented and easily distinguishable by the model. Happy, sadness, and surprise also show strong results with F1-scores close to or above 0.95. While Contempt achieves perfect recall (1.00), its slightly lower precision (0.83) suggests occasional misclassification of other emotions as Contempt. Similarly, Surprise has perfect precision but a slightly reduced recall (0.92), indicating some missed detections. The macro and weighted averages of the F1-score (0.96 and 0.97, respectively) further confirm the model's strong and balanced performance, even on classes with fewer samples such as Fear and Contempt.

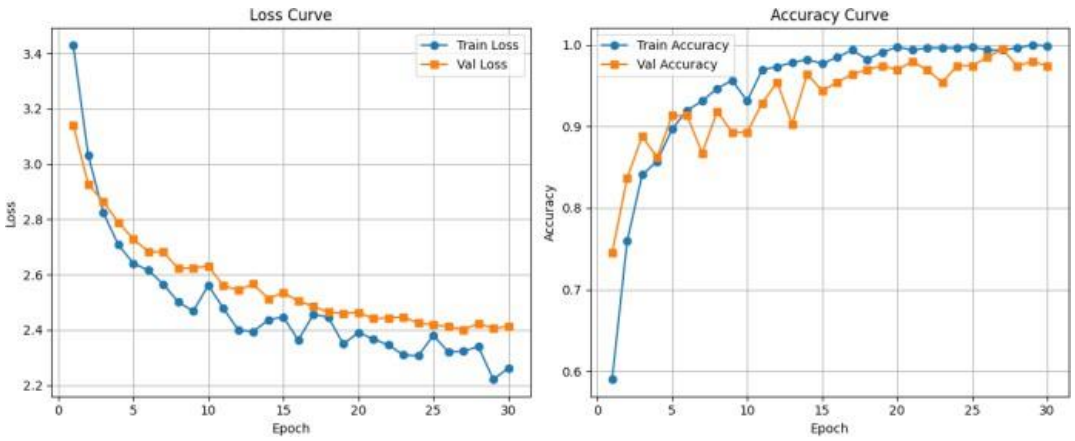


***Figure 4.*** Accuracy and Loss Curves Over Epochs for CK+

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

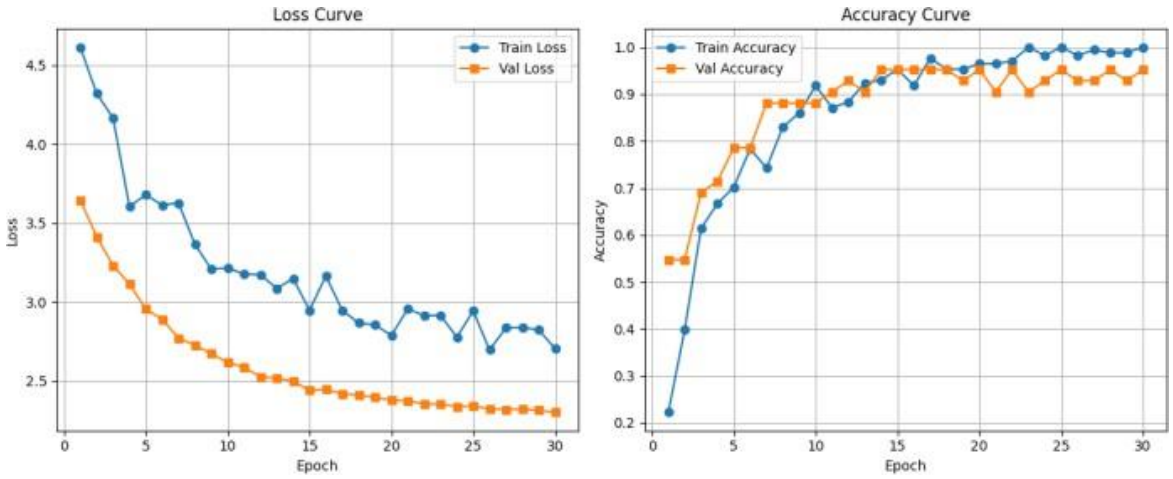


**Figure 5.** Accuracy and Loss Curves Over Epochs for JAFFE

***Table 1.*** *Classification Report for CK+*

| | **Precision** | **Sensitivity** | **F1-Score** |
|---|---|---|---|
| Anger | 1.00 | 1.00 | 1.00 |
| Disgust | 1.00 | 1.00 | 1.00 |
| Fear | 0.90 | 1.00 | 0.95 |
| Happiness | 1.00 | 0.98 | 0.99 |
| Neutral | 0.83 | 1.00 | 0.91 |
| Sadness | 0.90 | 1.00 | 0.95 |
| Surprise | 1.00 | 0.92 | 0.96 |

**Table 2.** Classification Report for JAFFE

| | **Precision** | **Sensitivity** | **F1-Score** |
|---|---|---|---|
| Anger | 1.00 | 1.00 | 1.00 |
| Disgust | 0.80 | 1.00 | 0.89 |
| Fear | 1.00 | 1.00 | 1.00 |
| Happiness | 1.00 | 0.92 | 0.96 |
| Neutral | 1.00 | 0.86 | 0.92 |
| Sadness | 0.83 | 1.00 | 0.91 |
| Surprise | 1.00 | 1.00 | 1.00 |

The model also performs remarkably well on the JAFFE validation set, achieving an overall accuracy of 95.24%. Most emotion categories are classified with perfect or near-perfect precision and recall. Emotions such as anger, fear, and surprise exhibit flawless classification with an F1-score of 1.00, indicating that these expressions are consistently and

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

correctly identified. Happy and Sad maintain high F1-scores (0.96 and 0.91, respectively), reflecting the model's strong generalization even with moderate class imbalance. Disgust shows slightly lower precision (0.80) but perfect recall, suggesting the model occasionally misclassifies other emotions as Disgust. Similarly, Neutral is recognized with high precision but slightly lower recall (0.86), which may be due to the subtle nature of the expression or limited sample size. The macro and weighted averages of the F1-score (both at 0.95) indicate balanced performance across all classes despite the small dataset size, highlighting the model's robustness in recognizing facial expressions even in lower-data scenarios like JAFFE.

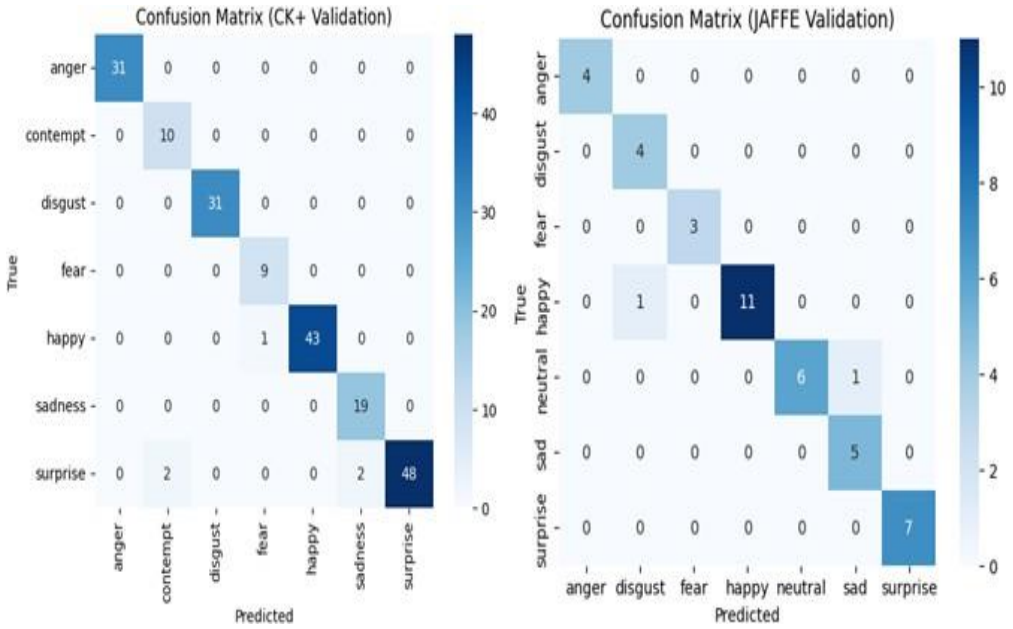


**Figure 6.** Confusion Matrices of CK+ and JAFFE Datasets

Figure 6 represents t-SNE of ViT Feature Embeddings for CK+ and JAFFE datasets respectively. The t-SNE plot provides a 2D projection of the high-dimensional feature embeddings extracted by the Vision Transformer (ViT). Each point represents a sample, colored according to its ground-truth emotion class.

The plot suggests that the model has learned reasonably discriminative features for several emotion classes.

For CK+ dataset, well-separated clusters are evident for Surprise (pink) and Happy (purple), indicating that the model effectively distinguishes these emotions with high confidence, Anger (blue) and Disgust (green) also exhibit distinguishable clusters, although with some overlap, suggesting a few shared facial features in their expression patterns, Fear (red) and Contempt (yellow) appear in smaller, scattered clusters. This is consistent with the small number of samples for these classes in the CK+ dataset and possibly more ambiguous visual cues, Sadness (brown) shows moderate clustering, with some points overlapping with Anger and Neutral-like expressions, reflecting natural similarities in facial expression features.

For the JAFFE dataset, even though it has small number of images, the visualization still provides insights into how well the model separates emotional expressions: Happy (red) and Surprise (pink) classes form relatively distinct clusters, indicating that the model captures their expressive features effectively even with limited data. Sad (brown) and Neutral (purple) expressions show partial overlap, suggesting some ambiguity in facial features between calm and subdued emotional states. Anger (blue) and Disgust (orange) appear close and somewhat entangled, possibly due to their similar localized facial cues (e.g., furrowed brows or tense mouth). Fear (green) samples are few and scattered, which is expected given the small number of examples and the subtlety of fear related expressions.

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

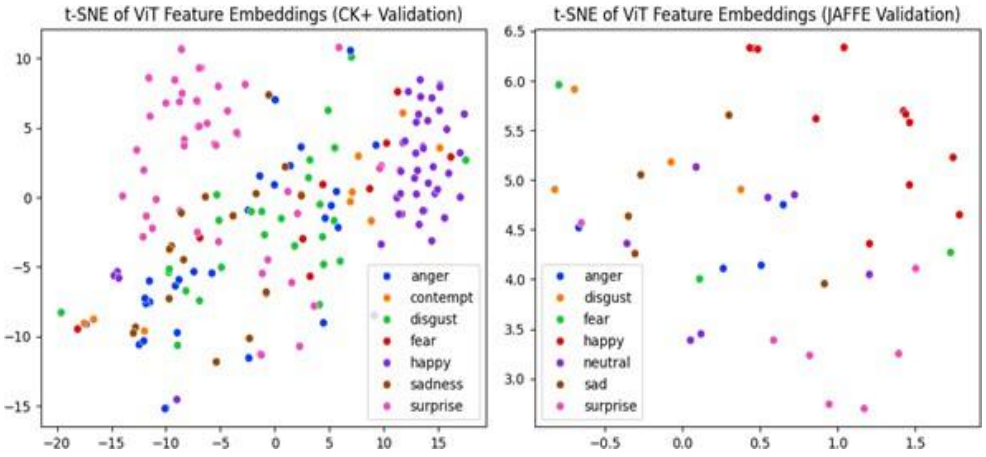


**Figure 7.** t-SNE plot of CK+ and JAFFE Datasets

Compared to recent state-of-the-art facial expression recognition (FER) models, our proposed method demonstrates competitive performance across standard benchmarks, particularly on the CK+ and JAFFE datasets. For instance, while ARFEC [19] achieved 92.5% on CK+, and SSAE-FER [23] reported 92.5% on JAFFE, our model surpasses these results with improved generalization. Debnath et al. [20] achieved strong results on CK+ (98.13%) but reported a lower performance on JAFFE (92.05%), indicating limitations in cross-dataset generalization. Similarly, [24] and [25] attained higher CK+ scores (96.52% and 96.19%, respectively), but their JAFFE performance (96.16% and 89.23%) remains inconsistent. Although [22] obtained 96.3% on JAFFE using Gabor features and genetic optimization, their CK+ performance (94.26%) is lower than ours. In contrast, our model maintains a strong balance between both datasets without relying on complex preprocessing or heavy architectures. Furthermore, unlike SCL-FExR [26], which emphasizes robustness under label noise, our approach achieves higher accuracy while maintaining architectural simplicity. These results collectively highlight the effectiveness and stability of our method, especially in scenarios requiring both accuracy and efficiency across diverse datasets. Table 3 shows a comparison between state-of-the-art and our results.

**Table 3.** Comparison with state-of-the-art Methods

| **Reference** | **CK+** | **JAFFE** |
|---|---|---|
| [19] | 92.50% | - |
| [20] | 98.13% | 92.05% |
| [21] | 95.63% | - |
| [22] | 94.26% | 96.30% |
| [23] | 99.30% | 92.50% |
| [24] | 96.52% | 96.16% |
| [25] | 96.19% | 89.23% |
| [26] | 76% (FER2013) | - |
| ***Ours*** | ***97.45%*** | ***95.24%*** |

**Journal of Information Systems Engineering and Management**
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

## 6. CONCLUSION

In conclusion, the proposed hybrid framework combining Vision Transformer architecture with LBP texture features and supervised contrastive learning demonstrates strong and consistent performance across multiple benchmark datasets. The model achieves a validation accuracy of 97.45% on CK+ and 95.24% on JAFFE, alongside high precision, recall, and F1-scores for most emotion classes.

The results highlight the effectiveness of integrating handcrafted features with deep learning representations in enhancing discriminative power, particularly for subtle or low-sample expressions. These findings suggest that the proposed method is both accurate and generalizable, making it a promising approach for real-world facial expression recognition tasks.

Despite these promising results, several directions remain open for future work. For example, extending the framework to more diverse in-the-wild datasets and enabling real-time deployment could further improve the recognition of micro-expressions and spontaneous emotions.

**Author Statements:**

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** All datasets used in this paper to support the findings are publicly available. They are cited and reported in the bibliography.

## REFERENCES

[1] Darwin, C., Prodger, P.: The Expression of the Emotions in Man and Animals. Oxford University Press, USA, ??? (1998)
[2] Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology 17(2), 124 (1971)
[3] Ghimire, D., Lee, J.: Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. Sensors 13(6), 7714–7734 (2013)
[4] Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L.: Static and dynamic 3d facial expression recognition: A comprehensive survey. Image and Vision Computing 30(10), 683–697 (2012)
[5] Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. IEEE Transactions on Affective Computing 2(4), 219–229 (2011)
[6] [6] Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE transactions on Affective Computing 6(1), 1–12 (2014)
[7] Kumari, J., Rajesh, R., Pooja, K.: Facial expression recognition: A survey. Procedia Computer Science 58, 486–491 (2015)
[8] Tian, Y., Kanade, T., Cohn, J.F.: Facial expression recognition. In: Handbook of Face Recognition, pp. 487–519. Springer, ??? (2011)
[9] Yang, J., Zhang, D., Frangi, A.F., Yang, J.-y.: Two-dimensional pca: a new approach to appearance-based face representation and recognition. IEEE transactions on pattern analysis and machine intelligence 26(1), 131–137 (2004)
[10] Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition
[11] of facial actions and their temporal models. IEEE transactions on pattern analysis and machine intelligence 32(11), 1940–1954 (2010)
[12] Jafri, R., Arabnia, H.R.: A survey of face recognition techniques. Jips 5(2), 41–68 (2009)
[13] Zhao, X., Zhang, S.: A review on facial expression recognition: Feature extraction and classification. IETE Technical Review 33(5), 505–517 (2016)

[14] Meriem, S., Moussaoui, A., Hadid, A.: Automated facial expression recognition using deep learning techniques: an overview. International Journal of Informatics and Applied Mathematics 3(1), 39–53 (2020)
[15] Karpathy, A., et al.: Cs231n convolutional neural networks for visual recognition.Neural networks 1 (2016)
[16] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Dovision transformers see like convolutional neural networks? Advances in Neural Information Processing Systems 34, 12116–12128 (2021)
[17] Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62 (2021)
[18] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 18661–18673 (2020)
[19] Ahonen, T., Hadid, A., Pietik¨ainen, M.: Face recognition with local binary patterns. In: Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8, pp. 469–481 (2004). Springer
[20] Gubbala, K., Kumar, M.N., Sowjanya, A.M.: Adaboost based random forest model for emotion classification of facial
[21] Debnath, T., Reza, M.M., Rahman, A., Beheshti, A., Band, S.S., Alinejad-Rokny, H.: Four-layer convnet to facial emotion recognition with minimal epochs and the significance of data diversity. Scientific Reports 12(1), 6991 (2022)
[22] Elsheikh, R.A., Mohamed, M., Abou-Taleb, A.M., Ata, M.M.: Improved facial emotion recognition model based on a novel deep convolutional structure. Scientific Reports 14(1), 29050 (2024)
[23] Boughida, A., Kouahla, M.N., Lafifi, Y.: A novel approach for facial expression, recognition based on gabor filters and genetic algorithm. Evolving Systems 13(2), 331–345 (2022)
[24] Ahmad, M., Alfandi, O., Khattak, A.M., Qadri, S.F., Saeed, I.A., Khan, S., Hayat, B., Ahmad, A., et al.: Facial expression recognition using lightweight deep learning modeling. Mathematical Biosciences and Engineering 20(5), 8208 (2023)
[25] Podder, T., Bhattacharya, D., Majumder, P., Balas, V.E.: A feature boosted deep learning method for automatic facial expression recognition. PeerJ Computer Science 9, 1216 (2023)
[26] Jabbooree, A.I., Khanli, L.M., Salehpour, P., Pourbahrami, S.: Geometrical facial expression recognition approach based on fusion cnn-svm. Int. J. Intell. Eng. Syst 17(1), 457–468 (2024)
[27] Vasudeva, K., Dubey, A., Chandran, S.: Scl-fexr: supervised contrastive learning approach for facial expression recognition. Multimedia Tools and Applications 82(20), 31351–31371 (2023)