Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

Facial Expression Recognition Using Fused Features: A Comparison of Deep and Machine Learning

Abbas Issa Jabbooree[1)*], Hussein Ala'a Alkaabi[2)], Ali Nadhim Kamber[3)]

[1)*]Ministry of Communication, ITPC, High Institute of Telecommunication & Post, Baghdad, Iraq
[2),3)]Ministry of Education, General Directorate of Education in Najaf, Najaf, Iraq

[1)*]it22abass@gmail.com , [2)]hussain.njf7@gmail.com, [3)]ali.nice1986@gmail.com

ABSTRACT

Facial expression recognition (FER) is a highly active field with applications in computer vision, human-computer interaction, security, and computer graphics animation. Recent advancements in deep learning and machine learning have increased interest in utilizing these techniques for accurate facial expression classification. This paper presents a comparative study that evaluates the performance of deep learning and machine learning as classifiers in FER systems, specifically after data fusion. Data fusion techniques combine and integrate multiple sources of information, aiming to enhance the overall classification accuracy by extracting two types of features using geometrical and appearance features trained using two types of convolutional neural networks. The feature outputs of these networks are fused to create a final feature vector for the classification process. The study evaluates the performance of deep learning on two benchmark datasets, the extended Cohn-Kanade (CK+) and Oulu-CASIA datasets, to assess the performance of deep learning. As a point of comparison, the traditional machine learning approach based on the support vector machine (SVM) is also evaluated on the same datasets. Performance metrics such as classification accuracy, precision, recall, and F1-score are utilized. The results obtained from the study highlight the strengths and limitations of both deep learning and machine learning techniques when employed as classifiers in FER systems. Notably, the experimental results demonstrate that the deep learning approach significantly outperforms the baseline methods, achieving an increase in recognition accuracy of 5.22% for the CK+ and 3.07% for the Oulu-CASIA dataset.

Keywords: Comparative study; deep learning; machine learning; data fusion; L2-SVM

1. INTRODUCTION

The proposed system enables practical applications in driver monitoring for fatigue detection, adaptive e-learning platforms, and mental health assessment through mobile apps, demonstrating its societal relevance. Facial expression recognition is vital for daily emotional communication and nonverbal signal exchange among humans. In recent years, Facial expression systems have garnered significant attention in various applications across various research domains, including computer vision approaches (Suraiya , Refat Khan , Munmun , Mayeen, 2020), driving behavior analysis (Sukhavasi et al., 2022), mobile applications (Nan et al., 2022), and understanding human behavior (Sajjad et al., 2020).

However, irrelevant facial information, such as background clutter, hair, glasses, and poses, challenges accurate FER performance. To overcome this obstacle, it is necessary to capture and accurately portray the most distinguishing features of facial expressions (C. Liu et al., 2021). Consequently, researchers have been exploring different techniques to enhance the performance of facial expression recognition systems. Two prominent techniques that have gained substantial attention are deep learning and machine learning. Deep learning, a subfield of machine learning, has exhibited remarkable success in various tasks such as image classification, object detection, and natural language processing (Sukhavasi et al., 2022). Its ability to automatically learn hierarchical representations from raw data has

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

made it a popular choice for many computer vision applications. On the other hand, traditional machine learning algorithms continue to be widely used in various domains due to their robustness and interpretability.

The conventional FER methods often emphasize local features. To train SVM classifiers, these approaches frequently use methods like Gabor filters with Genetic algorithms or Local Binary Patterns (LBPs) features (Boughida et al., 2022)(Niu et al., 2021). However, due to their reliance on simplistic facial characteristics, these techniques often yield low recognition accuracy. Recent studies have indicated that deep learning-based approaches, such as CNN and VGG16, outperform conventional methods (Porcu et al., 2020)(Ngoc et al., 2020). Nevertheless, the facial feature extraction algorithms used to date have focused solely on the pixel level (C. Liu et al., 2021). To enhance FER accuracy, this paper proposes a method that combines the results of two CNN models, considering the face's geometric and perceptual features. It has been observed that appearance-based traits are more accurate and require less data for processing (Barman & Dutta, 2021). In the proposed methodology, facial images are utilized as input data, and a series of pre-processing steps are performed to accurately determine the Region of Interest (ROI). Various pre-processing techniques are used to enhance the process of feature extraction.

After performing face detection and resizing the images to a consistent size, accurately locating and identifying facial landmarks becomes feasible. To develop a method for grouping facial landmarks, it is necessary to categorize them into seven distinct groups according to their specific locations on the face (C. Liu et al., 2021). A 1D-CNN, trained on undirected graphs called β-skeletons, can more effectively extract geometric features by grouping these landmarks. Simultaneously, a 2D-CNN is trained on the same image. The results from these two separate CNNs are combined, and the input features are concatenated along the axis to create a vector for classification.

Therefore, this paper aims to conduct a comprehensive comparative study that evaluates the performance of deep learning and machine learning as classifiers in facial expression recognition systems after data fusion. Deep learning models can capture complex patterns and representations, whereas machine learning algorithms offer better interpretability and require fewer computational resources. Understanding the relative advantages and limitations of these approaches can aid researchers in selecting the most suitable classifier for a given facial expression recognition task. This study employs seven categorical emotion labels: surprise, happiness, sadness, fear, disgust, neutral, and anger. (González-Lozoya et al., 2020). The study evaluates the proposed technique using two widely recognized datasets, CK+ [12] and Oulu-CASIA [13]. Various performance criteria, such as F1-score, precision, recall, and validation accuracy, are used to assess the suggested method's performance (Yan & Zhou, 2022). This fusion-based approach is particularly valuable for real-time applications like automotive safety systems and emotion-aware human-computer interfaces, highlighting its potential for real-world deployment. The paper's remaining sections are as follows: Section 2 provides an overview of related work in image expression recognition, deep learning, machine learning, and data fusion. Section 3 presents the methodology, including the dataset, pre-processing techniques, and the deep learning and machine learning algorithms used for classification. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper by summarizing the results, discussing their implications, and suggesting directions for future research.

2. METHOD

Related Work

Traditional FER systems approaches typically involve CNN and similar methods to extract features from image information for expression classification. Although these models can learn various expression-specific features from the facial image, their accuracy may suffer in exceptional cases, such as changes in illumination and facial occlusion. However, models based on multimodal data fusion can effectively address this issue by leveraging other sources of facial features for prediction when the complete set of features is unavailable (Han et al., 2021). This section provides an overview of relevant studies that have explored the use of deep learning and machine learning techniques as classifiers in FER systems, particularly after data fusion. Han Yi et al. (Han et al., 2021) introduced a multimodal data fusion approach that combined CNN, Landmarks-based Neural Networks (LNN), and Histogram of Oriented Gradients-based Neural Networks (HNN). The inputs to these networks were face images, Histogram Oriented Gradient (HOG) features, and facial landmarks. By leveraging the complementarity of these modalities, improved performance in FER was achieved.

Authors in (Boughida et al., 2022) proposed the FER method that combined Gabor filters with a Genetic Algorithm (GA). Their model adopted facial landmarks to identify ROI on the face and extract Gabor features. The GA was used to select optimal features for the SVM classifier and optimize its hyperparameters for accurate classification. Jeong et

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

al. (Jeong et al., 2020) presented a practical FER method that utilized deep joint spatiotemporal features extracted using appearance and geometry. By employing three-dimensional (3D) convolution, the technique simultaneously captured spatial and temporal characteristics. The selection of (23) essential face landmarks using energy distribution analysis and the development of a joint fusion classifier effectively integrated the deep appearance and geometric features, leading to improved FER performance. A novel approach called Attention Mechanism-Fused Global and Local Features (AM-FGL) was proposed to integrate attention mechanisms for global-local fusion (Zhou et al., 2023). AM-FGL effectively captures global and local facial characteristics by leveraging a spatial channel model and a local feature CNN. The presented algorithms are specifically designed and optimized for recognizing tail expressions, partially obscured expressions, and facial images captured at wide angles. The proposed method addresses the challenges associated with these types of expressions.

The authors in (Zhu et al., 2021) proposed a FER cascade network incorporating spatial feature extraction, hybrid attention, and temporal feature extraction. It employed a residual neural network to extract spatial features and fused them using a hybrid attention module. A gate control loop unit was used to extract temporal features, enabling the network to capture the dynamic nature of facial expressions. The fully connected layer classified and recognized facial expressions based on these features. Suparshya Babu Sukhavasi et al. (Sukhavasi et al., 2022) presented a hybrid network architecture combining Deep Neural Networks (DNN) and SVM for driver expression prediction—the system aimed to enhance road safety by detecting driver stability and good driving behavior under diverse conditions. Gabor and Local Binary Patterns (LBP) features were used, and a combination of an SVM classifier and a CNN was employed for classification. A lightweight Multi-feature Fusion CNN (MFF-CNN) for image-based FER is proposed (Zou et al., 2022). The architecture included an Image Branch and a Patch Branch. The Patch Branch extracted local features from multiple image patches, while the Image Branch extracted mid-level and high-level global features from the input image. MFF-CNN employed the L2 norm to select discriminative local features, and joint fusion integrated the features from both branches, resulting in improved performance in image-based facial expression recognition. In another study, the authors (H. Wang et al., 2020) proposed an iterative fusion approach that combined Multi-Orientation Gradient (MO-HOG) features and deep-learned classifiers for FER. MO-HOG features were extracted from local facial images rich in expressions, while deep-learned features provided high-level semantic information. An optimization algorithm-based classifier fusion approach was used to find the best-fused classifier through iterative iterations. Yahui Nan et al. (Nan et al., 2022) introduced the A-MobileNet model, which focuses on enhancing the extraction of local features in facial expressions. This lightweight model caters to fast and light networking requirements in mobile internet applications. By optimizing the model's parameters using center-loss and softmax loss, the authors aimed to reduce intra-class distances while maximizing inter-class distances. This optimization strategy improved recognition accuracy compared to the original MobileNet series and other existing models. Notably, the proposed approach achieved these performance gains without additional model parameters.

While conventional FER systems have predominantly employed pixel-level analysis or geometric feature extraction in isolation, they frequently fail to capture the critical interdependencies between facial landmark positioning and appearance-based characteristics. This study systematically compares deep and traditional machine learning classifiers to bridge this methodological gap. A fusion approach is also employed to integrate geometric features based on the β-skeleton with appearance-based techniques. By leveraging this combined approach, the authors aim to overcome the limitations of previous methods and enhance the overall performance of FER systems.

The Proposed System Design

The proposed system structure consists of four stages, as shown in Figure 1. First, the implemented pipeline performs a pre-processing task on the input images (face detection, image normalization, landmark detection, and data augmentation). Once the images are pre-processed, appearance geometrical feature extraction and training can be used in the second stage. The two feature types are concatenated in the third stage of the proposed fusion technique. Finally, each image is classified by associating it with a label of the most probable facial expression in the proposed classifiers (deep learning or machine learning).

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

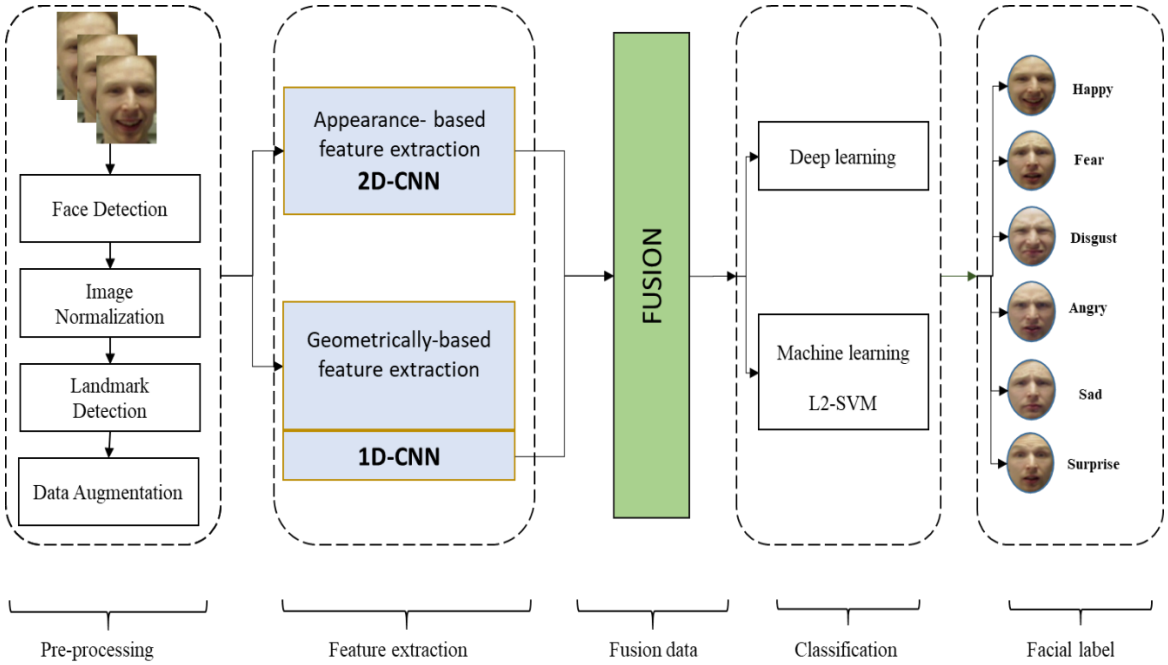


Fig. 1 Schematic diagram of the proposed Facial Expression Recognition (FER) system, illustrating the integration of fusion-based CNN features with an SVM classifier. The pipeline comprises four stages: pre-processing, geometric and appearance feature extraction, feature fusion, and classification.

Pre-Processing Stage

The pre-processing pipeline begins with the automated detection of the facial region, a crucial step in any facial expression recognition (FER) system. In this work, facial regions are localized using the Viola-Jones face detection algorithm [21], renowned for its real-time efficiency and high detection accuracy [11]. Upon detecting a face, the corresponding region of interest (ROI) is extracted by cropping the image using the top-left coordinate and the width and height of the bounding rectangle, thereby excluding extraneous background regions that could interfere with recognition.

Facial feature extraction is conducted following face detection, which is vital for robust FER performance. Improperly aligned or irrelevant facial data can degrade model performance, introducing noise and redundancy into the training process [22].

To standardize the input for CNN, two types of image normalization are performed:

1. **Spatial normalization**: All images are resized to a uniform resolution of 50×50 pixels, ensuring consistency across the dataset. This resizing facilitates compatibility with the CNN input layer and accelerates convergence during training [23].
2. **Intensity normalization**: Variations in lighting and contrast among images depicting the same expression can negatively impact recognition accuracy. To address this, histogram equalization is applied, enhancing contrast and improving the distinguishability of facial features. Although effective, this method may also accentuate noise, particularly in regions with limited grayscale variation [23].

Subsequently, facial landmark detection is carried out using the Dlib toolkit [24], which applies a robust Histogram of Oriented Gradients (HOG) filter in conjunction with Max-Margin Object Detection and an Ensemble of Regression Trees [25]. This method enables precise localization of 68 distinct facial landmarks, including those around the eyes, eyebrows, nose, and mouth, as illustrated in Figure 2.

Given the data-hungry nature of deep learning algorithms, data augmentation is employed to enhance model generalizability and mitigate overfitting. Leveraging the image data generator's parameters, a range of transformation techniques—such as mirroring, rotation, scaling, and translation—are applied to expand the dataset synthetically. This augmentation strategy introduces subtle facial appearance and positioning variations, enriching the training set and improving the CNN's robustness [4].

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

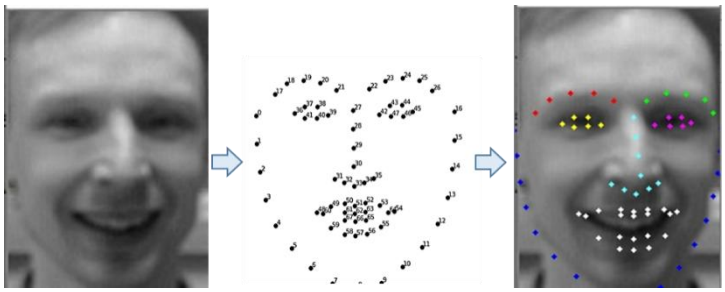


Fig. 2 Visualization of the facial landmark detection process using the Dlib toolkit, highlighting the 68 key landmarks localized around the eyes, eyebrows, nose, and mouth regions.

Geometrical-Feature Extraction using β-Skeleton Stage

Traditional graph-based neighborhood algorithms—such as the Nearest Neighbor Graph (NNG) and Gabriel Graph (GG)—primarily rely on classical graph-theoretical parameters for their formulation. While widely adopted, these methods face limitations in optimization due to their reliance on circular regions to define neighborhood boundaries [26]. Consequently, identifying meaningful connections between data points can be computationally expensive and inefficient. In contrast, the β-skeleton graph presents a more effective approach for neighborhood modeling by offering enhanced accuracy and spatial grouping. This graph constructs neighborhood structures by analyzing empty regions (or "vacant neighborhoods") around edges, thereby yielding a more refined connectivity scheme [27]. A key parameter in this method is β, which controls the size of the neighborhood region. A smaller β value tends to increase the number of edges by permitting tighter connections. A larger β value may lead to sparser graphs due to edge elimination in less-dense substructures [28].

Figure 3 depicts the influence of varying β values on the graph structure, and Equation (1) mathematically describes the corresponding angular criteria for edge formation following the β-skeleton formulation [26].

$$\theta = \begin{cases} \sin^{-1}\frac{1}{\beta} & \text{if } \beta \geq 1 \\ \pi - \sin^{-1} & \text{if } \beta < 1 \end{cases} \quad (1)$$

The proposed method enhances classification accuracy with a changeable neighborhood area but requires advanced angle parameter computations (Pourbahrami et al., 2020).

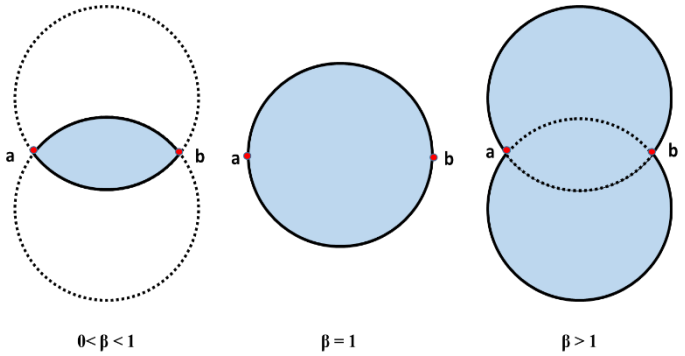


Fig. 3 Illustration of β-skeleton graph structures under varying β values, demonstrating the impact of β on neighborhood connectivity and edge formation. Adapted from [29].

Facial expression analysis incorporates the utilization of β-Skeleton-based Geometrical Feature Extraction. In this technique, β-skeletons represent facial landmarks as points in space (X. Liu et al., 2021). This approach aims to extract geometric features that provide insights into facial expressions, including landmarks' placement and spatial relationships. Each facial landmark is represented by a 2-D Cartesian coordinate (x, y). The facial landmarks are

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

categorized into seven groups by considering the positions of facial features like the eyes, eyebrows, mouth, and nose, allowing the analysis of their structure and interconnections (Jabbooree et al., 2023).
Consider the feature vector v1, which represents the left eyebrow, composed of five points (x17leb, y17leb), (x18le, y18le), ..., (x21leb, y21leb). Our previous paper [30] contains all the feature vectors that capture the specified qualities concerning other facial landmarks.

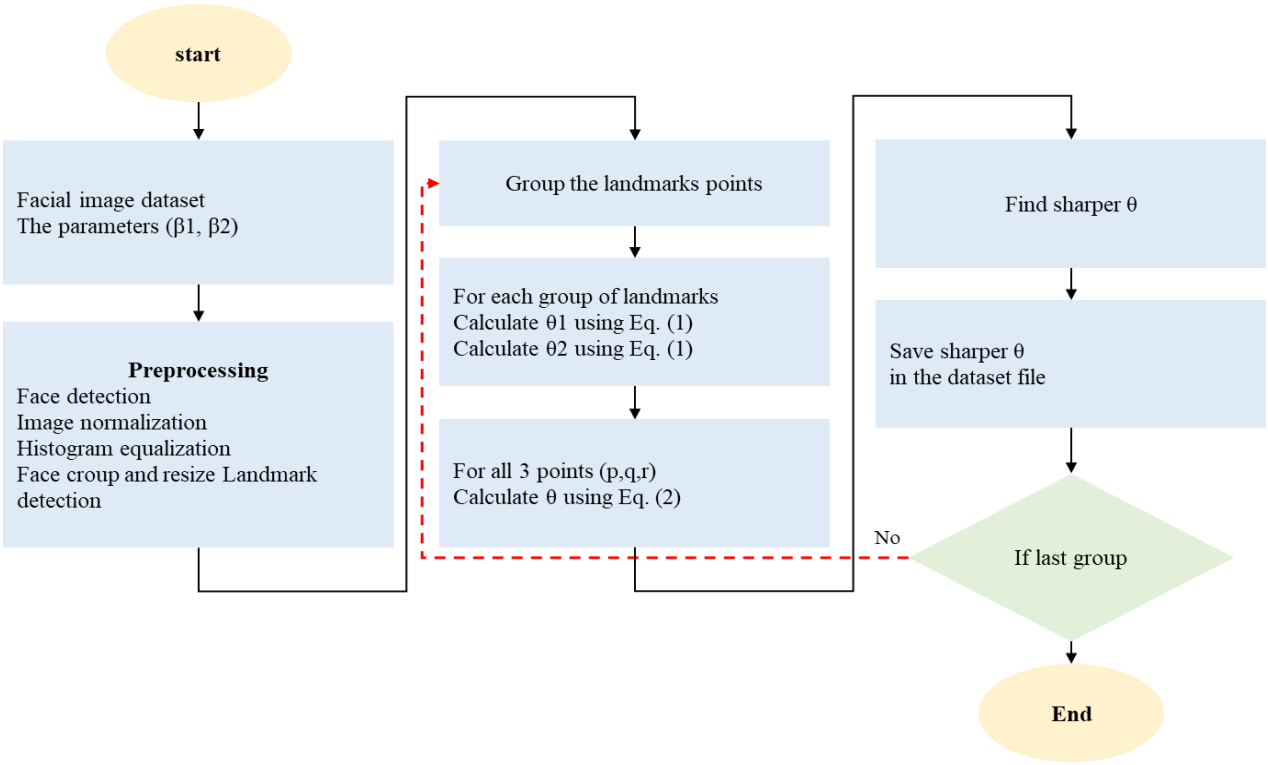


Fig. 4 Flowchart of the geometric feature extraction process using the β-skeleton graph method, detailing the steps from landmark grouping to angle computation and feature storage.

To calculate the values of θ1 and θ2 using Equation (1), we can apply the cosine law in Equation 2 to obtain the angle prq. In the equation, the angle of interest, denoted as (θ), corresponds to prq, where rpq represents a set of three points (Lagias et al., 2018). By comparing θ1 and θ2 with the angle (θ), we can determine if the angle is sharper and store the resulting data in a dataset file.

$$\theta = \cos^{-1}\left(\frac{a^2+b^2-c^2}{2*a*b}\right) \quad (2)$$

In Equation 2, a, b, and c represent the lengths of the three sides of a triangle formed by any three points in the feature set.
Figure 4 shows the flowchart for calculating the β-skeleton for each group of features, and Figure 5 shows the calculation results.

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

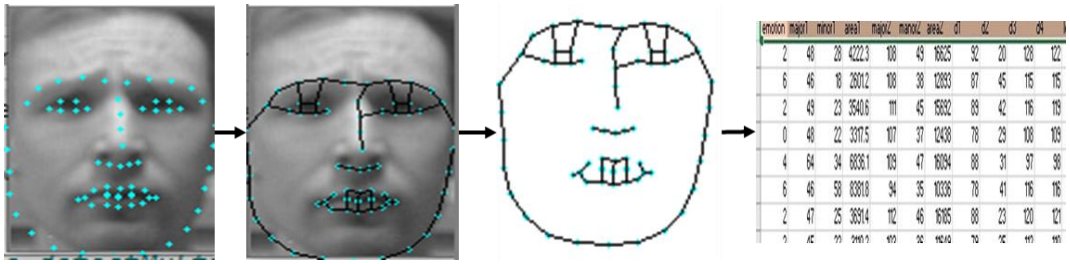


Fig. 5 Results of geometric feature extraction using β-skeleton calculations, showcasing the derived spatial relationships among facial landmarks for expression analysis.

Appearance Feature Extraction based on CNN Stage

CNN is a powerful deep learning model commonly used for interpreting information presented in a grid format, such as images or sequences. A typical CNN architecture consists of three main layers: convolutional layers, pooling layers, and fully connected layers (Yan & Zhou, 2022). Convolutional layers extract local features from the input data by applying a set of learnable filters. Each filter performs a convolution operation, which involves sliding over the input grid and computing dot products between the filter weights and local patches of the input. The output of this operation is a feature map highlighting relevant patterns and structures (Ozcan & Basturk, 2020)(Yan & Zhou, 2022). Pooling layers: After the convolutional layers, pooling layers are often employed to reduce the spatial dimensions of the feature maps while preserving important information. Pooling can be done using various techniques, such as maximum pooling, sum pooling, or average pooling (Yan & Zhou, 2022). These pooling operations summarize and compress the feature maps, allowing the network to focus on the most salient features while reducing computational complexity. Fully connected layers: A CNN's final layers are fully connected, typical layers found in traditional neural networks. These layers connect every neuron from the previous layer to each neuron in the next layer, allowing for high-level feature combinations and decision-making. The output of the fully connected layers is typically used for classification or regression tasks (Ozcan & Basturk, 2020)(Yan & Zhou, 2022).

This paper proposes a novel FER method that combines the outcomes of two distinct CNN models. The input image is pre-processed, and its geometric features are extracted before being fed into two channels. One channel employs a 2D-CNN to extract and train appearance features. At the same time, the other uses a 1D-CNN to train geometric features by merging the results of these smaller networks and inputting them into a classifier, as shown in Figure 1. The architecture of the suggested 2D-CNN model is outlined in Table 1 and visually represented in Figure 6 (a). The model consists of eight layers, including three convolution layers, three pooling layers, one dropout layer, and one flattening layer. It's worth noting that these layers are counted while including the input layer.

Table 1. The proposed 2D-CNN model architecture

Layer	**Dimension/ value**
Input Image	(50x50x1)
Convolution layer 1	(50x50x32)
Maxpool layer 1	(25x25x32)
Convolution layer 2	(25x25x64)
Maxpool layer 2	(12x12x64)
Convolution layer 3	(12x12x128)
Maxpool layer 3	(6x6x128)
Dropout	(0.5)
Flatten layer	

One-dimensional convolution kernels are useful in FER and find applications in other domains, such as voice recognition, human activity, and music genre classification (Allamy & Koerich, 2021). Compared to longer 2D-CNN, shorter 1D-CNN excels at processing one-dimensional data more efficiently. They require fewer layers to extract relevant features from the input, making them suitable for low-cost and real-time systems, including central processing

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

units(Mitiche et al., 2020). Unlike the conventional k×k size of the 2D-convolution kernel, a 1D-CNN uses a 1D-convolution kernel of size k×1 (Yan & Zhou, 2022). In this paper, we employ a 1D-CNN to extract phase information from geometric data. Figure 6 (b) illustrates the architecture of our 1D-CNN network, which comprises eight distinct layers. The architecture consists of three convolutional layers, three pooling layers, a dropout layer (rate=0.25), and a flattening layer. The output from this layer is then concatenated with features from the second channel before being fed into the final classifier for expression recognition.

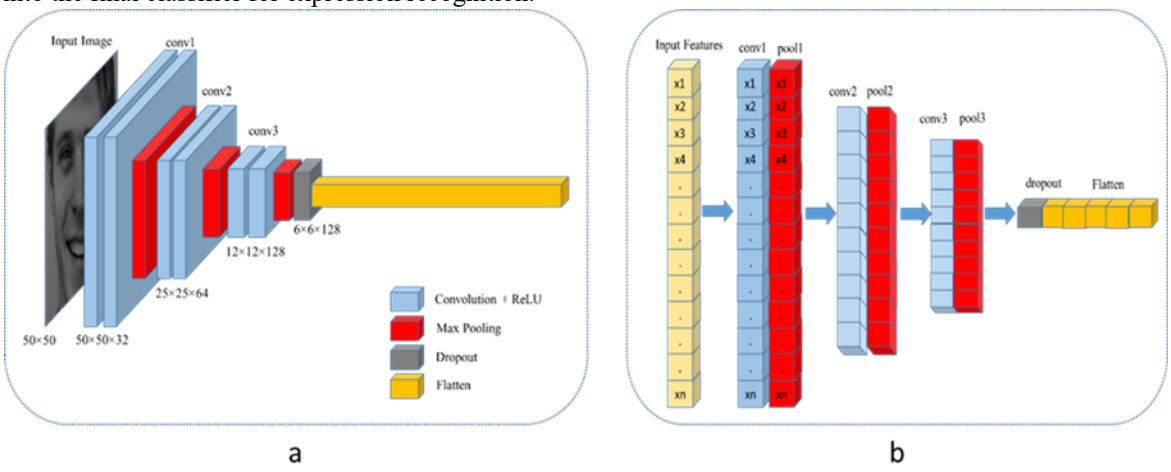


Fig. 6 Architectural diagrams of the proposed CNN models: (a) 2D-CNN for appearance feature extraction and (b) 1D-CNN for geometric feature processing, detailing layer configurations and dimensions."

The Fusion Stage

In facial expression recognition, fusion feature approaches aim to combine features obtained from different modalities or representations to enhance the accuracy and robustness of the recognition system (Han et al., 2021)(Park et al., 2021). Here, we present four commonly utilized fusion feature approaches in facial expression recognition. The first method involves extracting and normalizing features from multiple sources and combining the data into one vector. The classification process is then performed using this combined feature vector. The second fusion approach computes the arithmetic mean of prediction scores across all models, using this averaged value as the final classification output. The third strategy selects the fusion model with the highest score as the final classification result. The last approach utilizes the results from each model to determine the weight to assign to each. The weights are then used to combine the outputs, resulting in the final classification (Do et al., 2021). Our work adopts the first fusion method to merge the two channels' appearance and geometric features. By combining the data at the feature level, we incorporate more input information, which often leads to improved identification outcomes compared to previous fusion algorithms.

Support Vector Machine L2-SVM

The Support Vector Machine (SVM) is a robust classifier grounded in statistical learning theory, aiming to minimize structural risk rather than merely empirical risk [24]. It is known for its strong generalization ability and inherently sparse model representation, making it effective across various classification tasks.

Compared to traditional machine learning techniques, SVM offers several distinct advantages [37]:

1. Strong Generalization with Limited Data: SVMs are particularly effective when dealing with small datasets. They are designed to maximize the margin between classes, enhancing generalization performance.
2. Efficient Optimization Objective: For linearly separable data, SVM seeks to maximize the geometric margin between different class samples. In cases of nonlinearly separable data, SVM incorporates kernel functions to project the input data into a higher-dimensional feature space where linear separation becomes feasible. Notably, including kernel functions ensures that the algorithm's computational complexity remains largely independent of the dataset size.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

3. Robustness via Soft Margin Technique: SVMs utilize soft margin optimization to handle imperfect data distributions. This technique allows for a small number of misclassified samples, applying penalty terms to control the trade-off between margin maximization and classification error, thereby improving robustness.
4. Neural Network Analogy: Structurally, SVM can be likened to a three-layer feedforward neural network, where support vectors define the hidden layer nodes. The model convergence is assured by optimizing a global maximum, with both the weight vector and hidden nodes effectively determined by the data distribution.

Two variants of the standard SVM model have been proposed—L1-SVM and L2-SVM, each introducing different loss functions. The primary distinction lies in the differentiability of L2-SVM, which utilizes a quadratic loss function in contrast to the linear loss used in L1-SVM. This modification enhances the smoothness of the optimization landscape and provides better penalization for instances far from the decision boundary [38].

The Classification Stage

After the fusion stage is performed and two channels of features are concatenated into a single vector, it is used as input for the classifier. This paper proposes two classifiers based on deep neural networks (DNNs) and machine learning. DNNs are capable of automatically learning complex features from raw image data. They consist of multiple layers that extract hierarchical representations, allowing them to capture complex patterns and relationships within facial expressions. Deep learning models have achieved state-of-the-art performance in FER systems due to their ability to handle large datasets.

On the other hand, machine learning SVM is a classical algorithm for classification tasks. SVMs are particularly effective when dealing with smaller datasets with high-dimensional feature spaces. In the FER system, SVMs can be trained with handcrafted features extracted from facial images, such as geometric-based features. These features are carefully designed to capture relevant facial information about different expressions. SVMs aim to find an optimal hyperplane that separates other expression classes in the feature space, maximizing the margin between the classes and achieving good generalization performance. In this paper, we utilize the L2-SVM machine learning classifier.

Fully connected layers, also known as dense layers, are an integral part of deep learning models (Ozcan & Basturk, 2020). The input fully connected layer performs various operations and has different names depending on the context. In the five-layer deep learning model, the ReLU activation function is applied to the inputs from three fully connected layers with sizes 512, 256, and 128, respectively. These layers process the concatenated vector of outputs from the previous layer. Finally, the Softmax activation function is employed in the last fully connected layer. This activation function is commonly used for multi-class classification problems, producing a probability distribution over the classes. The Adam optimizer is utilized to optimize the training process and minimize the loss. The Adam optimizer is an adaptive optimization algorithm that adjusts the learning rate during training, leading to faster convergence and improved performance. Additionally, the Adam optimizer reduces the loss when cross-entropy is used as the loss function.

Both networks are trained on the same dataset to establish labels for each facial expression. Experimental tests have been conducted to verify the proposed method's robustness and efficiency, and it has been found to deliver satisfactory results.

3. RESULT AND DISCUSSIONS

Implementation Details

The proposed work was implemented in Python 3.11, utilizing an Intel Core i7-13700H CPU running at 5.00 GHz with 16GB of memory. The deep learning framework employed for this project was TensorFlow and Keras. Adam was utilized for 50 epochs to optimize the model, with the learning rate initialized to 0.0001. Throughout the training process, the learning rate (lr) exponentially decayed by 0.05 at the end of each epoch. The batch size was fixed at 32 to ensure efficient training. To address the issue of overfitting, a dropout ratio of 0.5 was incorporated into the CNN. The initial parameter values were set as $\beta 1 = 1.1$ and $\beta 2 = 0.9$ for the β-skeleton.

Datasets Description

This section presents the experimental results and analysis conducted on the CK+ dataset (Lucey et al., 2010) and the Oulu-CASIA dataset (Taini et al., 2008). These datasets were chosen due to their diverse image sizes and significant variations in facial expressions. The CK+ dataset comprises 123 patients with 593 images (Lucey et al., 2010)(X. Liu et al., 2021). The last frame of each image sequence is labeled with a list of action units. Additionally,

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

there are 327 videos featuring 118 individuals, from which 981 still images were extracted. The dataset covers a range of human emotions, including anger, disgust, fear, happiness, surprise, and sadness, as shown in Figure 7 (Lucey et al., 2010). This dataset is commonly used for expression analysis and serves as a benchmark in numerous articles.



Fig. 7 Example images from the CK+ dataset [12], displaying the six basic facial expressions (anger, disgust, fear, happiness, sadness, surprise) and neutral faces used for model training and evaluation.

The Oulu-CASIA dataset, a collaborative effort between the Chinese Academy of Sciences and the University of Oulu, was published (Taini et al., 2008)(Yu et al., 2018). Data were collected from 80 participants, including 50 Chinese individuals and 30 Finns aged 23 to 58. The imaging devices used in the dataset operated at a frame rate of 25 frames per second, capturing images with a size of 320 x 240 pixels. Participants were instructed to produce a series of successive facial expressions (Gera & Balasubramanian, 2021)(Jabbooree et al., 2023). The dataset comprises 1440 photos divided into six categories. Furthermore, the dataset includes image sequences captured under normal, weak, and no lighting, as shown in Figure 8.

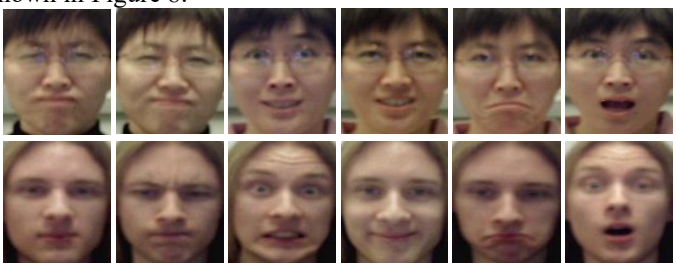


Fig. 8 Representative images from the Oulu-CASIA dataset [13], captured under varying illumination conditions (normal, weak, and dark), depicting the six prototypical facial expressions.

Experimental Results and Analysis

Experimentally, we adopt the proposed fusion machine learning method as the baseline model. To evaluate the impact of the β-skeleton method, we conducted a comparative analysis against the proposed method in multiple studies. Data augmentation techniques were applied to augment the size of smaller datasets, aiming to improve the classification performance of the proposed method. The dataset was randomly split into 80% for training and 20% for testing (M. Wang et al., 2020). Validation accuracy was used as the primary metric to assess the model's performance for evaluation purposes. Additionally, precision, recall, and F-measure were calculated at the level of individual expressions, providing further insights into the model's performance and ability to classify and identify specific expressions [41] accurately.

The CK+ Dataset Results

The observed 3.19% improvement in recognition accuracy compared to baseline models can be attributed to three key factors inherent to our fusion approach. First, the β-skeleton's geometric modeling effectively captures structural relationships between facial landmarks, particularly for expressions involving significant spatial distortions (e.g., the characteristic eyebrow arch in 'sadness' or lip corner depression in 'disgust'). Second, the parallel CNN pathway extracts complementary appearance-based features, such as transient wrinkle patterns around the forehead for 'anger' or nasolabial furrow deepening in 'happiness'. This dual-channel architecture addresses the fundamental limitation of prior works [6,7] that depended exclusively on handcrafted features, which often fail to capture geometric and textural

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

cues simultaneously. Third, the feature-level fusion strategy enables nonlinear interactions between these complementary modalities, allowing the model to learn expression-specific feature combinations that would be inaccessible to single-modality approaches. Figure 10 illustrates the accuracy curves during training and validation. Table 2 compares the two proposed methods and state-of-the-art models using the CK+ dataset.

Table 2. Comparative performance metrics (accuracy%) of the proposed methods and state-of-the-art models on the CK+ dataset, highlighting the superiority of the deep learning approach.

Methods	**Average accuracy**
GLFCNN with SVM in 2022 (Sukhavasi et al., 2022)	95.05
Gabor Filter, GA with SVM (Boughida et al., 2022)	94.26
LBP with CNN (Park et al., 2021)	96.23
Parabola method Lip with SVM (Barman & Dutta, 2021)	96.36
LBP with SVM (Kola & Samayamantula, 2021)	93.90
CDLLNET(T. Liu et al., 2021)	87.42
LBP, ORB features with SVM (Niu et al., 2021)	93.21
DGNN (Ngoc et al., 2020)	96.02
GA with SVM (X. Liu et al., 2021)	95.58
DGNN (Ngoc et al., 2020)	96.02
SACNN with ALSTM(C. Liu et al., 2021)	95.15
Multiple Instance Learning with SVM/HCR (Sikka et al., 2016)	95.10
The proposed method based on L2-SVM machine learning	**95.41**
The proposed method based on deep learning	**98.60**

Table 3. Precision, recall, F1-score, and accuracy metrics for each facial expression category using the machine learning-based L2-SVM classifier on the CK+ dataset.

Expression	**Precision**	**Recall**	**F1**	**Accuracy**
Anger	95.92	87.85	91.71	96.00
Contempt	95.88	93.00	94.42	93.00
Disgust	96.00	91.43	93.66	96.00
Fear	93.00	96.88	94.90	93.00
Happy	99.00	92.52	95.65	99.00
Sad	78.79	98.73	87.64	78.00
Surprise	97.03	98.00	97.51	98.00

Table 4. Precision, recall, F1-score, and accuracy metrics for each facial expression category using the deep learning-based classifier on the CK+ dataset.

Expression	**Precision**	**Recall**	**F1**	**Accuracy**
Anger	96.00	91.43	93.66	96.00
Contempt	95.00	76.00	84.44	95.00
Disgust	99.00	100.00	99.50	99.00
Fear	70.80	100.00	82.91	97.00
Happy	100.00	83.33	90.91	100
Sad	88.07	96.00	91.87	96.00
Surprise	99.00	100.00	99.50	99.00

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

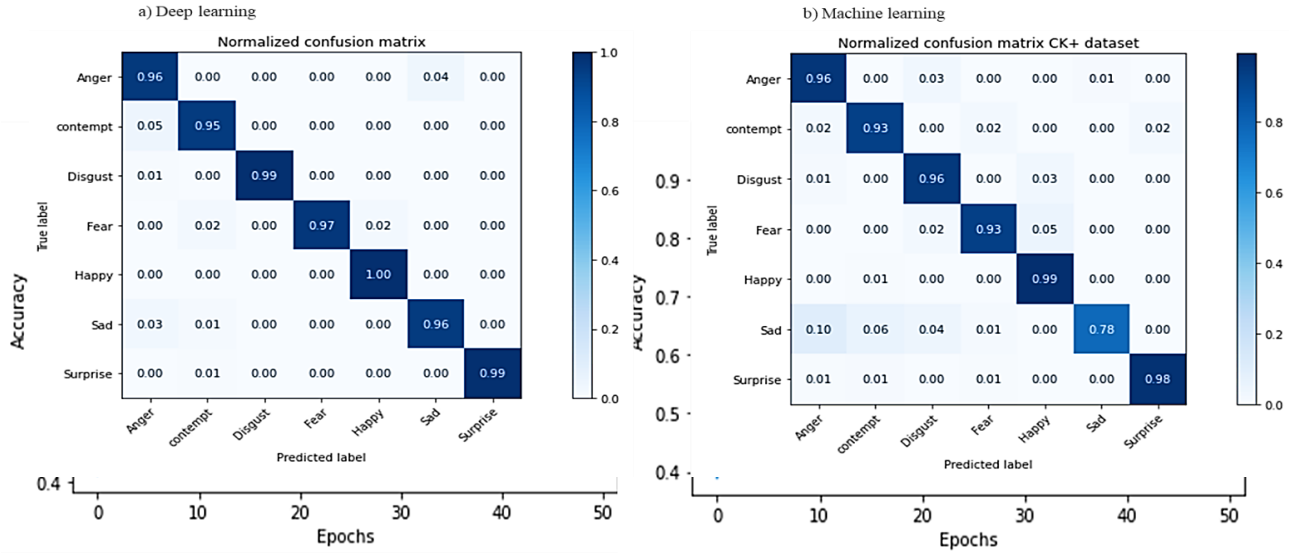


Fig. 9 Normalized confusion matrices for the CK+ dataset, comparing classification performance of (a) the deep learning model and (b) the L2-SVM classifier across seven expression categories.

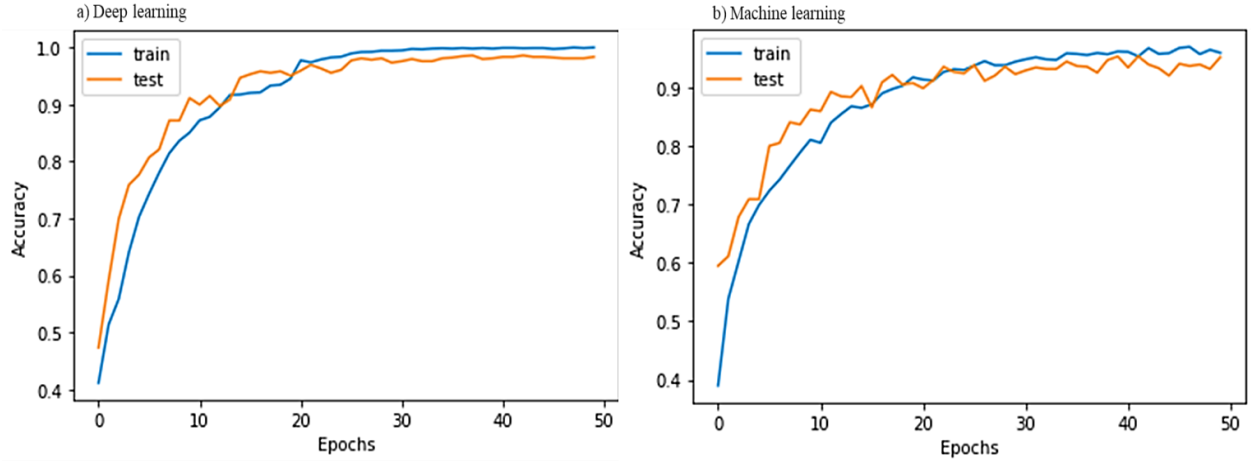


Fig. 10 Training and validation accuracy curves for the (a) deep learning and (b) L2-SVM models on the CK+ dataset, illustrating convergence behavior and generalization performance over 50 epochs.

The Oulu-CASIA Dataset Results

The system's superior performance (2.46% improvement over SCAN [40]) under challenging illumination conditions demonstrates the robustness conferred by geometric feature integration. While appearance-based methods like HOG [7] and LBP suffer from illumination-dependent texture variations, the β-skeleton's reliance on relative landmark positions maintains stability across lighting conditions. This is particularly evident in expressions involving eye region changes (e.g., 'fear' with upper eyelid elevation), where geometric relationships remain detectable even when local texture features degrade. The proposed architecture's illumination robustness suggests particular suitability for real-world applications like automotive systems, where lighting conditions are frequently suboptimal. Figure 12 illustrates the accuracy curves during training and validation. Table 5 compares the two proposed methods and state-of-the-art models using the Oulu-CASIA dataset.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

Table 5. Recognition accuracy (%) of the proposed methods and state-of-the-art models on the Oulu-CASIA dataset, highlighting the performance gains from feature fusion.

Classifier	**Accuracy %**
CDLLNET (T. Liu et al., 2021)	86.28
SCAN (Gera & Balasubramanian, 2021)	86.56
Hybrid Attention Cascade Network (Zhu et al., 2021)	87.31
FreNet (Tang et al., 2021)	88.33
Face2Net2ExpNet (Ding et al., 2017)	87.71
DCPN (Yu et al., 2018)	86.23
Multiple Instance Learning with SVM/HCR (Sikka et al., 2016)	82.10
The proposed method based on L2-SVM machine learning	**85.41**
The proposed method based on deep learning	**89.02**

Table 6. Detailed evaluation metrics (precision, recall, F1-score, and accuracy) for each expression class using the deep learning-based approach on the Oulu-CASIA dataset.

Expression	**Precision**	**Recall**	**F1**	**Accuracy**
Anger	75.00	78.13	76.53	75.00
Disgust	89.90	83.18	86.41	89.00
Fear	92.08	91.18	91.63	93.00
Happy	98.98	97.00	97.98	97.00
Sad	86.00	86.87	86.43	86.00
Surprise	93.00	98.94	95.88	93.00

Table 7. Detailed evaluation metrics (precision, recall, F1-score, and accuracy) for each expression class using the machine learning-based L2-SVM classifier on the Oulu-CASIA dataset.

Expression	**Precision**	**Recall**	**F1**	**Accuracy**
Anger	58.00	78.38	66.67	58.00
Disgust	88.89	59.06	70.97	88.00
Fear	98.89	78.76	87.68	89.00
Happy	93.00	97.89	95.38	93.00
Sad	77.00	84.62	80.63	77.00
Surprise	87.00	100.00	93.05	87.00

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

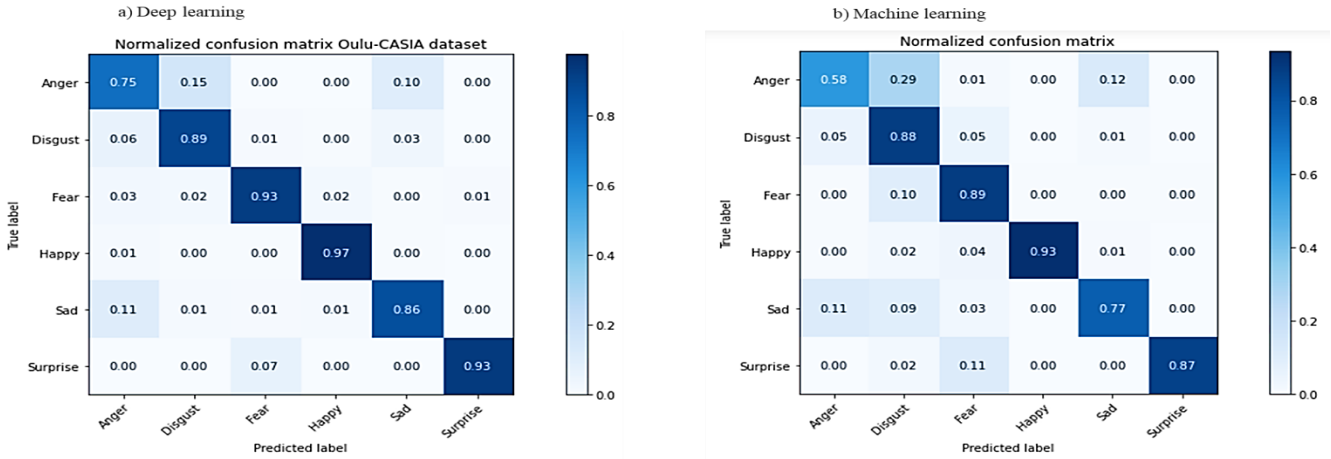


Fig 11. Normalized confusion matrices for the Oulu-CASIA dataset, comparing classification performance of (a) the deep learning model and (b) the L2-SVM classifier across seven expression categories.

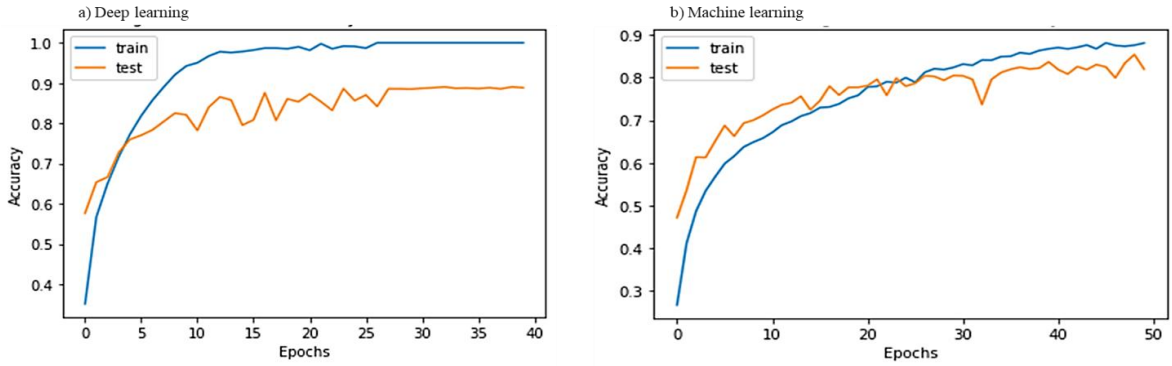


Fig.12 Training and validation accuracy curves for the (a) deep learning and (b) L2-SVM models on the Oulu-CASIA dataset, illustrating convergence behavior and generalization performance over 50 epochs.

4.CONCLUSION

This section summarizes the study's key findings, outlines its limitations, and proposes directions for future research to advance facial expression recognition systems. This study presents a comprehensive comparison of deep learning and machine learning classifiers for facial expression recognition (FER), leveraging a novel fusion of geometric (β-skeleton) and appearance-based (CNN) features. The experimental results demonstrate that the proposed deep learning approach significantly outperforms traditional methods, achieving accuracy improvements of 5.22% on the CK+ dataset and 3.07% on the Oulu-CASIA dataset. These findings highlight the efficacy of combining complementary feature representations through data fusion, which enhances the system's ability to capture structural relationships and texture patterns critical for accurate expression analysis. The success of this dual-channel architecture underscores its potential for advancing FER systems in both research and practical applications. Despite its promising results, this study has several limitations. First, the reliance on benchmark datasets, while standard in the field, introduces constraints related to sample diversity and demographic representation. The predominance of prototypical expressions in these datasets may not fully encapsulate the nuanced variations observed in real-world scenarios, where cultural differences and subtle emotional cues play a significant role. Second, the current framework assumes controlled imaging conditions, which may limit its robustness in environments with variable lighting, occlusion, or non-frontal poses. These factors could affect the system's performance in practical deployments, such as automotive safety systems or affective computing interfaces.Future research should focus on three key directions

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

to address these limitations and enhance the system's capabilities. First, validating the proposed approach on larger and more diverse datasets would improve its generalizability across different populations and real-world conditions. Second, exploring temporal dynamics through video-based analysis could refine recognition performance by capturing the evolution of facial expressions over time. Third, developing lightweight architectures optimized for edge devices would facilitate real-world deployment, particularly in resource-constrained applications like driver monitoring systems or mobile affective interfaces. Additionally, investigating hybrid models that integrate interpretable machine learning techniques with deep learning could enhance performance and transparency, ensuring the system's robustness while maintaining its interpretability. Addressing these challenges will be critical for advancing the practical utility of FER systems in diverse and dynamic environments.

5.REFERENCES

Allamy, S., & Koerich, A. L. (2021). 1D CNN Architectures for Music Genre Classification. *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*.

Barman, A., & Dutta, P. (2021). Facial expression recognition using distance and shape signature features. *Pattern Recognition Letters*, *145*, 254–261.

Boughida, A., Kouahla, M. N., & Lafifi, Y. (2022). A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. *Evolving Systems*, *13*(2), 331–345.

Ding, H., Zhou, S. K., & Chellappa, R. (2017). FaceNet2ExpNet : Regularizing a Deep Face Recognition Net for Expression Recognition. *IEEE*, 118–126.

Do, L. N., Yang, H. J., Nguyen, H. D., Kim, S. H., Lee, G. S., & Na, I. S. (2021). Deep neural network-based fusion model for emotion recognition using visual data. *Journal of Supercomputing*, *77*(10), 10773–10790.

Gera, D., & Balasubramanian, S. (2021). Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *Pattern Recognition Letters*, *145*, 58–66.

González-Lozoya, S. M., de la Calleja, J., Pellegrin, L., Escalante, H. J., Medina, M. A., & Benitez-Ruiz, A. (2020). Recognition of facial expressions based on CNN features. *Multimedia Tools and Applications*, *79*(19–20), 13987–14007.

Han, Y., Wang, X., & Lu, Z. (2021). *Research on facial expression recognition based on Multimodal data fusion and neural network*.

Jabbooree, A. I., Khanli, L. M., Salehpour, P., & Pourbahrami, S. (2023). A novel facial expression recognition algorithm using geometry β –skeleton in fusion based on deep CNN. *Image and Vision Computing*, *134*, 104677.

Jeong, D., Kim, B. G., & Dong, S. Y. (2020). Deep joint spatiotemporal network (DJSTN) for efficient facial expression recognition. *Sensors (Switzerland)*, *20*(7).

Kola, D. G. R., & Samayamantula, S. K. (2021). A novel approach for facial expression recognition using local binary pattern with adaptive window. *Multimedia Tools and Applications*, *80*(2), 2243–2262.

Lagias, A. E., Lagkas, T. D., & Zhang, J. (2018). New RSSI-Based Tracking for Following Mobile Targets Using the Law of Cosines. *IEEE Wireless Communications Letters*, *7*(3), 392–395.

Liu, C., Hirota, K., Ma, J., Jia, Z., & Dai, Y. (2021). Facial Expression Recognition Using Hybrid Features of Pixel and Geometry. *IEEE Access*, *9*, 18876–18889.

Liu, T., Wang, J., Yang, B., & Wang, X. (2021). Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom. *Infrared Physics and Technology*, *112*(December 2020), 103594.

Liu, X., Cheng, X., & Lee, K. (2021). GA-SVM-Based Facial Emotion Recognition Using Facial Geometric Features. *IEEE Sensors Journal*, *21*(10), 11532–11542.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., & Ave, F. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *4*(July), 94–101.

Mitiche, I., Nesbitt, A., Conner, S., Boreham, P., & Morison, G. (2020). 1D-CNN based real-time fault detection system for power asset diagnostics. *IET Generation, Transmission and Distribution*, *14*(24), 5816–5822.

Nan, Y., Ju, J., Hua, Q., Zhang, H., & Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, *61*(6), 4435–4444.

Ngoc, Q. T., Lee, S., & Song, B. C. (2020). Facial landmark-based emotion recognition via directed graph neural

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6128

Submitted : 17 June 2025
Accepted : 23 June 2025
Published : 2 July 2025

network. *Electronics (Switzerland)*, *9*(5).
Niu, B., Gao, Z., & Guo, B. (2021). Facial Expression Recognition with LBP and ORB Features. *Computational Intelligence and Neuroscience- Hindawi*, *2021*, 1–10.
Ozcan, T., & Basturk, A. (2020). Static facial expression recognition using convolutional neural networks based on transfer learning and hyperparameter optimization. *Multimedia Tools and Applications*.
Park, S. J., Kim, B. G., & Chilamkurti, N. (2021). A robust facial expression recognition algorithm based on multi-rate feature fusion scheme. *Sensors*, *21*(21), 1–26.
Porcu, S., Floris, A., & Atzori, L. (2020). Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics (Switzerland)*, *9*(11), 1–12.
Pourbahrami, S., Khanli, L. M., & Azimpour, S. (2020). An automatic clustering of data points with alpha and beta angles on apollonius and subtended arc circle based on computational geometry. *2020 28th Iranian Conference on Electrical Engineering, ICEE 2020*.
Sajjad, M., Zahir, S., Ullah, A., Akhtar, Z., & Muhammad, K. (2020). Human Behavior Understanding in Big Multimedia Data Using CNN based Facial Expression Recognition. *Mobile Networks and Applications*, *25*(4), 1611–1621.
Sikka, K., Sharma, G., & Bartlett, M. (2016). LOMo: Latent ordinal model for facial analysis in videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 5580–5589.
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. *MDPI*.
Suraiya , Refat Khan , Munmun , Mayeen, M. R. I. (2020). Development of a Robust Multi-Scale Featured Local Binary Pattern for Improved Facial Expression Recognition. *Sensoe MDPI*, *20*(5391).
Taini, M., Zhao, G., Li, S. Z., & Pietikäinen, M. (2008). Facial expression recognition from near-infrared video sequences. *Proceedings - International Conference on Pattern Recognition*, 1–4.
Tang, Y., Zhang, X., Hu, X., Wang, S., & Wang, H. (2021). *Facial Expression Recognition Using Frequency Neural Network*. *30*, 444–457.
Wang, H., Wei, S., & Fang, B. (2020). Facial expression recognition using iterative fusion of MO-HOG and deep features. *Journal of Supercomputing*, *76*(5), 3211–3221.
Wang, M., Tan, P., Zhang, X., Kang, Y., Jin, C., & Cao, J. (2020). Facial expression recognition based on CNN. *Journal of Physics: Conference Series*, *1601*(5).
Yan, K., & Zhou, X. (2022). Chiller faults detection and diagnosis with sensor network and adaptive 1D CNN. *Digital Communications and Networks*, *June 2021*.
Yu, Z., Liu, Q., & Liu, G. (2018). Deeper cascaded peak-piloted network for weak expression recognition. *Visual Computer*, *34*(12), 1691–1699.
Zhou, J., Li, J., Yan, Y., Wu, L., & Xu, H. (2023). Mixing Global and Local Features for Long-Tailed Expression Recognition. *Information (Switzerland)*, *14*(2), 1–19.
Zhu, X., Ye, S., Zhao, L., & Dai, Z. (2021). Hybrid attention cascade network for facial expression recognition. *Sensors*, *21*(6), 1–16.
Zou, W., Zhang, D., & Lee, D. J. (2022). A new multi-feature fusion based convolutional neural network for facial expression recognition. *Applied Intelligence*, *52*(3), 2918–2929.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.