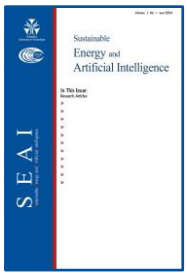
Hamedan University of Technology
Iranian Association of Electrical and Electronics Engineers

Sustainable Energy and Artificial Intelligence

Online ISSN: 3060-8015

Journal Homepage: www.enai.hut.ac.ir



Emotion Recognition in Video using a Hybrid of Supervised and Unsupervised Deep Neural Networks

Haniyeh Amirbeigi, Hamid Reza Shahdoosti*

Department of Electrical and Computer Engineering, Hamedan University of Technology, Hamedan, Iran.
** Corresponding Author: h.doosti@hut.ac.ir*

Article Info

Article type:
Original Article

Article history:
Received 2026-05-26;
Revised 2026-07-05;
Accepted 2026-07-07.

How to cite this article:
Amirbeigi, H and Shahdoosti, H.R. and. (2026). Emotion Recognition in Video Using a Hybrid of Supervised and Unsupervised Deep Neural Networks. *Sustainable Energy and Artificial Intelligence*, 2(3), 227-239.
DOI: 10.61186/seai.2605-1056

Abstract

Facial expressions constitute one of the most important forms of non-verbal communication, enabling humans to convey emotions and establish effective interpersonal interactions. In recent years, facial emotion recognition (FER) has significantly benefited from deep learning techniques, particularly architectures capable of modeling both spatial and temporal information in video sequences. Motivated by the dynamic nature of facial expressions, this study proposes and evaluates two hybrid deep learning frameworks for video-based emotion recognition: GRBM-ConvLSTM2D and GRBM-3DCNN. In both configurations, a Gaussian Restricted Boltzmann Machine (GRBM) is employed as a preliminary feature extraction layer to enhance spatiotemporal representation learning. Implemented in Python, the proposed models are designed to classify seven emotional categories: surprise, happiness, anger, sadness, fear, disgust, and contempt. The experimental evaluation was conducted on the benchmark CK+ dataset. The results demonstrate that both proposed architectures achieve strong recognition performance, with the GRBM-ConvLSTM2D model attaining a validation accuracy of 91.16% and the GRBM-3DCNN model achieving 88.80%, confirming the effectiveness and robustness of the proposed hybrid frameworks for video-based facial emotion recognition.

***Keywords*:** Emotion Recognition; Hybrid Neural Network; Convolutional Neural Network; Recurrent Neural Network; Gaussian Restricted Boltzman Machine.

Copyrights
© 2026 Licensee Hamedan University of Technology, Hamedan, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution –Non-Commercial 4.0 International (CC BY-NC 4.0) License (http://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

Human communication is fundamentally categorized into verbal and non-verbal modalities. Within this framework, facial expressions serve as the most prominent non-verbal indicator of emotional states. While body language encompassing postures, gestures, and eye contact contributes to interpersonal communication, facial analysis remains the most direct method for interpreting an individual's affective state [1,2]. Consequently, Facial Expression Recognition (FER) aims to automate the detection of emotional states through visual analysis, addressing a critical challenge in machine perception. Beyond enabling affective computing in robotics, FER holds significant potential for clinical and assistive technologies, including depression diagnosis, social support for autistic individuals, and sensory substitution devices such as smart glasses to facilitate communication for the visually impaired. To address the complexities of video-based emotion recognition, several deep learning paradigms are employed. Convolutional Neural Networks (CNNs) are extensively utilized for spatial feature extraction, proving effective in

computer vision tasks such as object classification and detection [3]. To model temporal dynamics inherent in video sequences, Recurrent Neural Networks (RNNs) are adopted; their recurrent loops leverage internal memory to retain information from previous frames, thereby enriching contextual analysis [4]. Finally, the Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM), a bipartite Markov random field, is integrated to facilitate unsupervised feature extraction. Similar to autoencoders, GRBMs excel at compressing and reconstructing continuous data, offering robust capabilities in dimensionality reduction, noise suppression, and image restoration [5–8].

2- Related Works

Facial expressions are among the most important nonverbal cues for revealing emotional states [9]. The automated interpretation of these expressions, known as Facial Expression Recognition (FER), has become a significant research area for estimating human affective and psychological conditions, with applications spanning behavioral research and clinical assessment [10]. In addition to healthcare, FER systems are increasingly used in human–computer interaction, social media analytics, personalized advertising, and counseling services [11]. A particularly challenging subfield is micro-expression analysis, where brief and involuntary facial movements may expose concealed emotions, making it highly relevant in forensic interrogation and deception detection [12, 13]. Micro-expressions are involuntary facial movements that reveal people's hidden feelings in high-stake situations and have practical importance in medical treatment, national security, interrogations and many human-computer interaction systems [14]. For this reason, many recent studies have adopted hybrid deep-learning frameworks, including CNN-RNN-based models, to improve recognition performance in security- and justice-oriented applications [15].

FER approaches are generally divided into static image-based methods and dynamic video-based methods. Although much of the existing literature focuses on still images [16, 17], video sequences contain richer information about facial motion and expression evolution over time [18]. A key limitation of standard Convolutional Neural Networks (CNNs) is that they are primarily designed for spatial feature learning and therefore do not adequately model temporal dependencies across video frames. To overcome this limitation, 3D CNNs have been introduced to jointly extract spatial and temporal features by applying convolution operations across both dimensions simultaneously [19–21]. Even so, 3D CNNs alone may still be insufficient for capturing the full complexity of spatio-temporal facial dynamics [18]. This has motivated the development of hybrid architectures that combine CNNs with Recurrent Neural Networks (RNNs), which are well suited for sequential modeling. In particular, Long Short-Term Memory (LSTM) networks are effective at learning temporal context in facial expression sequences. Unlike conventional RNNs, LSTMs use gating mechanisms that reduce the vanishing-gradient problem and preserve long-range dependencies in sequential data [22]. Since the vanilla fully-connected LSTM unrolls the image into a one-dimensional vector, it loses crucial spatial information, whereas Convolutional LSTM (ConvLSTM) performs LSTM operations directly through convolutions, thereby preserving spatial structure while modeling temporal dynamics [23].

Beyond supervised architectures, Restricted Boltzmann Machines (RBMs) have attracted considerable attention because of their ability to discover useful features in an unsupervised manner. Since standard RBMs are limited to binary visible units, Gaussian-Bernoulli RBMs (GRBMs) were developed to handle real-valued inputs such as images. However, most existing GRBM-based models rely solely on contrastive divergence to reconstruct each data point with minimal error, meaning their feature extraction process remains undirected and is conducted without any label-level guidance [24]. Through contrastive divergence learning, GRBMs transform nonlinear real-valued data into compact latent representations, making them effective feature extractors for FER tasks.

The rest of this paper is organized as follows. Section 3 describes the dataset, preprocessing steps, and the proposed hybrid neural architecture. Section 4 presents the experimental results, and Section 5 concludes the paper.

3. Materials and Methods

3-1. Dataset

To validate the performance of the proposed hybrid architecture, the CK+48 dataset was employed (Figure 1). This dataset serves as a specialized, cropped, and resized derivative of the Extended Cohn-Kanade (CK+) dataset, originally developed by Carnegie Mellon University (CMU) for advanced image processing and affective computing tasks [25]. The dataset consists of 981 grayscale images distributed across seven emotion

classes, where each class is characterized by three sequential frames at a 48×48 pixel resolution. While the dataset presents limitations regarding total sample volume and inherent video diversity, its structural consistency provides a significant benefit: the frontal-view captures and clean, unambiguous facial expression patterns facilitate the extraction of discriminative features during the learning process [26, 27].

3-2. Preprocessing

Due to the inherent class imbalance and the limited number of samples for specific emotional categories within the dataset, direct training poses a high risk of overfitting. To mitigate this issue and enhance the generalization capability of the network, a class-specific data augmentation strategy was implemented to balance the dataset. The primary spatial transformation operations included random rotations of up to 10°, horizontal and vertical translations limited to 10% of the image dimensions, a zoom range of 10% and horizontal flipping (Figure 2).

To address the severe underrepresentation of certain classes specifically **contempt, fear, and sadness** dynamic class-dependent augmentation factors were introduced. Rather than applying a uniform multiplication rate, higher augmentation coefficients (oversampling ratios) were selectively applied to these minority classes to bring their sample sizes on par with the majority classes. Following this targeted augmentation process, the dataset was successfully balanced, resulting in a total of 4,308 samples, with each emotional sequence standardized to 3 key temporal frames.

3-3. Proposed Method

The proposed methodology evaluates two distinct, self-contained hybrid deep learning configurations designed for spatiotemporal feature extraction and emotion classification:

1. **The GRBM-ConvLSTM2D Framework:** This configuration integrates a Gated Gaussian Restricted Boltzmann Machine (GRBM) with a unified Convolutional LSTM (ConvLSTM2D) network to jointly capture spatial features and temporal dynamics.
2. **The GRBM-3DCNN Framework:** This pipeline combines a Gated Gaussian Restricted Boltzmann Machine (GRBM) with a three-dimensional convolutional neural network (3D-CNN) to extract spatiotemporal representations directly from volumetric video frames.

The training pipeline is executed in two primary phases. In the first phase, unsupervised pre-training is performed. Each constituent network is trained independently without their classification heads using the augmented dataset. Notably, the training distribution for the GRBM component is deliberately augmented with a mixture of synthesized noisy and noise-free images to enhance the network's reconstruction robustness and feature representation. In the final phase, for each independent framework, the learned latent features are forwarded to a fully connected layer coupled with a 7-class Softmax activation function to perform the final emotion classification.

Fig. 1. Sample images from the CK+48 dataset.

Fig. 2. Sample images after applying data augmentation.

3-3-1. CNN-LSTM Hybrid Network

In this section, the **ConvLSTM2D** architecture was employed due to its effectiveness in handling video data and other temporally ordered inputs, where both spatial and temporal dependencies must be captured simultaneously. This model is particularly well suited for spatiotemporal feature extraction, making it an appropriate choice for facial expression analysis in video sequences.

Within each ConvLSTM2D layer, the **hard sigmoid** function was utilized as the gating activation, while the **tanh** function was adopted for generating the layer output (Figure 3). This combination enables the network to regulate temporal information flow while preserving nonlinear representational capacity.

Following each feature extraction block, a sequence of regularization and stabilization operations was applied. Specifically, a **MaxPooling** layer was included to reduce spatial dimensionality while retaining the most salient features, a **Dropout** layer with a rate of **0.3** was introduced to mitigate overfitting, and a **Batch Normalization** layer was employed to improve training stability and reduce the risk of vanishing gradients.

The **hard sigmoid** activation function can be regarded as a computationally efficient linear approximation of the standard sigmoid function. Although it offers faster execution and reduced complexity, it may still be affected by the vanishing gradient problem. Its mathematical formulation is given below.

$$Hardsigmoid(x) = \begin{cases} min\_val & x < \min\_val \\ \max\_val & x > \max\_val \\ x & otherwise \end{cases} \quad (1)$$

This is while, the tanh activation function projects its input values into the range [−1,+1]. Being symmetric around the origin, it yields zero-centered outputs, which facilitates smoother gradient flow during backpropagation and partially alleviates the vanishing gradient problem, formulated as follows:

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad (2)$$

In this configuration, the network excludes a Softmax classification layer, as the primary objective is strictly restricted to spatiotemporal feature extraction. The proposed feature extractor consists of six ConvLSTM2D layers and was trained independently using a learning rate of 10^{-4}.

3-3-2. 3D-CNN Feature Extractor

For this part, a customized 3D Convolutional Neural Network (3D-CNN) was used. Similar to the ConvLSTM2D, this model is highly suitable for video data or data with temporal sequences because it extracts spatiotemporal features by applying three-dimensional kernels over the input volume, preserving the relational information across the third dimension.

In each convolutional layer:

- **ReLU** is used as the activation function to introduce non-linearity and accelerate convergence.

After specific stages:

- A **MaxPooling3D** layer is applied to down-sample spatial dimensions while preserving the temporal/depth resolution,
- A **Dropout** layer with rates of 0.3, 0.4, or 0.5 is utilized to prevent overfitting,
- And a **BatchNormalization** layer is applied to stabilize the learning process and avoid vanishing gradients.

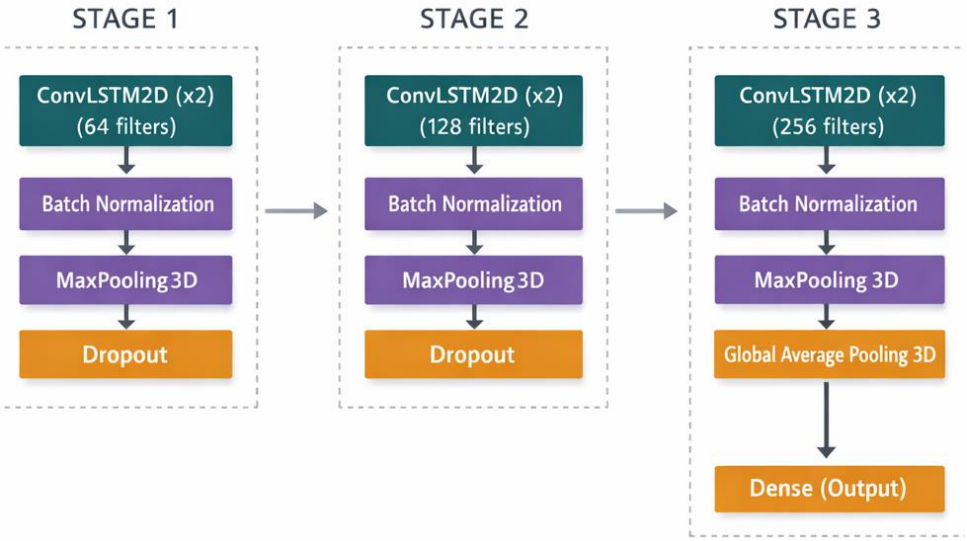


Fig. 3. Proposed ConvLSTM network architecture.

The ReLU activation function is a piecewise linear function that outputs the input directly if it is positive, otherwise, it outputs zero. This helps mitigate the vanishing gradient problem during backpropagation and allows the network to learn faster. Its mathematical definition is presented below:

$$\mathrm{ReLU}(x) = \max(0, x) \quad (3)$$

The proposed network contains six 3D convolutional layers structured into three feature extraction stages, followed by global aggregation (using Global Average Pooling 3D) and a final dense layer (Figure 4). In this section, the network does not include a softmax layer because the goal is solely to extract spatiotemporal features. The final layer employs a **Linear** activation function to output continuous-valued feature embeddings. The Linear activation function (also known as the identity function) passes the input signal directly to the output without any modification:

$$\mathrm{Linear}(x) = x \quad (4)$$

The model is trained independently with a learning rate of 0.0001 to generate a representation vector of size 64, which is prepared for the subsequent fusion stage.

3-3-3. Gaussian RBM Network

In this study, we employ a Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM) as an unsupervised generative model. This framework operates by reconstructing flattened input images, thereby enabling the extraction of robust latent features without the need for label-based guidance (Figure 5). The integration of the GRBM module with the hybrid CNN–LSTM architecture is strategically designed to improve classification accuracy and simultaneously mitigate overfitting. The choice of the GRBM is justified by its inherent capacity to model continuous-valued data, specifically for inputs normalized within the [0,1] or [−1,1] range. Regarding the architectural configuration, the network utilizes 6,912 input neurons mapped to 512 hidden units. Within the GRBM module, the energy function incorporates Gaussian noise with a standard deviation of $\sigma = 0.1$ σ=0.1. This parameter plays a pivotal role in aligning the energy landscape with the normalized intensity distribution of input features, which reside in the [0,1] range. Empirical evidence suggests that σ=0.1 yields an optimal equilibrium for Contrastive Divergence (CD) training. Specifically, it prevents energy landscape saturation, thereby compelling the model to extract discriminative latent representations while maintaining stable gradient updates throughout the independent training phase of this pathway. The energy function of the GRBM, which extends the standard RBM framework with a quadratic term to accommodate continuous inputs, is formulated as follows:

$$E(V, h|\theta) = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{n_h} c_j h_j - \sum_{i=1}^{n_v} b_i v_i + \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} \quad (5)$$

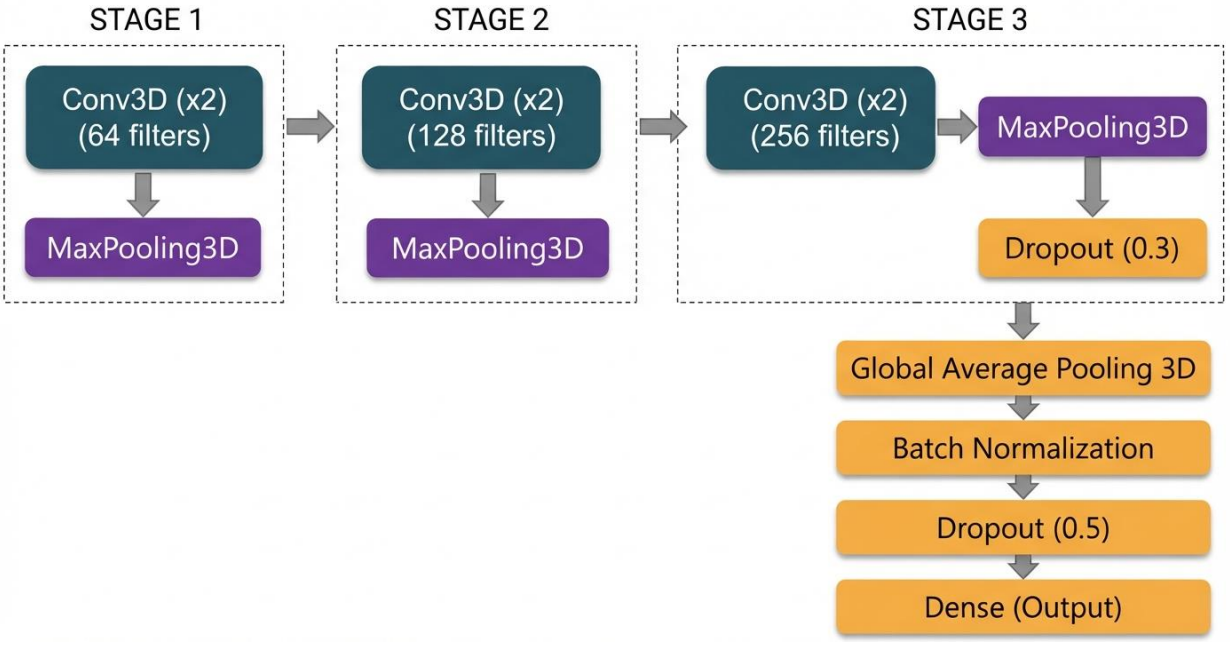


Fig. 4. Proposed CNN3D network architecture.

In most literature, the energy function of the **Gaussian–Bernoulli RBM** is commonly written as:

$$E(V,h|\theta) = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h} W_{ij}h_j\frac{v_i}{\sigma_i} - \sum_{j=1}^{n_h} c_j h_j + \sum_{i=1}^{n_v}\frac{(v_i - b_i)^2}{2\sigma_i^2} \quad (6)$$

In this formulation, the linear term $\sum_{i=1}^{n_v} b_i v_i$ is implicitly incorporated into the quadratic penalty. This simplification remains valid under the assumption that σ is treated as a fixed hyper parameter rather than a learnable variable [28, 29]. In our implementation, we specifically set σ=0.1 to refine the reconstruction process and enhance the model's sensitivity to continuous input variations. The parameters within these equations are defined as follows:

- v denotes the vector of visible units (input layer).
- h denotes the vector of hidden units (latent representation).
- W represents the shared weight matrix connecting the visible and hidden layers.
- b and ccc correspond to the bias vectors for the visible and hidden units, respectively.
- σ signifies the standard deviation of the Gaussian noise, which regulates the stochasticity of the visible units.

The probabilistic distributions inherent to the GRBM model facilitate effective conditional inference, allowing the network to perform stochastic sampling. Given the energy function defined above, the conditional probability distributions for the hidden and visible units are derived as follows:

1. Conditional probability of hidden units given visible units:

The hidden units (h_j) follow a Bernoulli distribution, where the probability of activation is computed using the sigmoid function:

$$p(h_i = 1|\mathrm{v}) = \sigma(W_i^T \mathrm{v} + c_i) \quad (7)$$

where $\sigma = \frac{1}{1+e^{-x}}$ represents the sigmoid activation function.

2. Conditional probability of visible units given hidden units:

The visible units (v_i) follow a Gaussian distribution, centered around the reconstruction mean:

$$p(V|\mathrm{h}) = \mathcal{N}(\mathrm{v}|\mathrm{Wh} + \mathrm{b}, \sigma^2\mathrm{I}) \quad (8)$$

Here, N(μ, σ2) denotes the Gaussian (normal) distribution, with the mean defined as the linear combination of the weighted hidden activations plus the bias b_i. In our model, σ_i is set to the fixed value of 0.1, ensuring consistent variance across the reconstruction process.

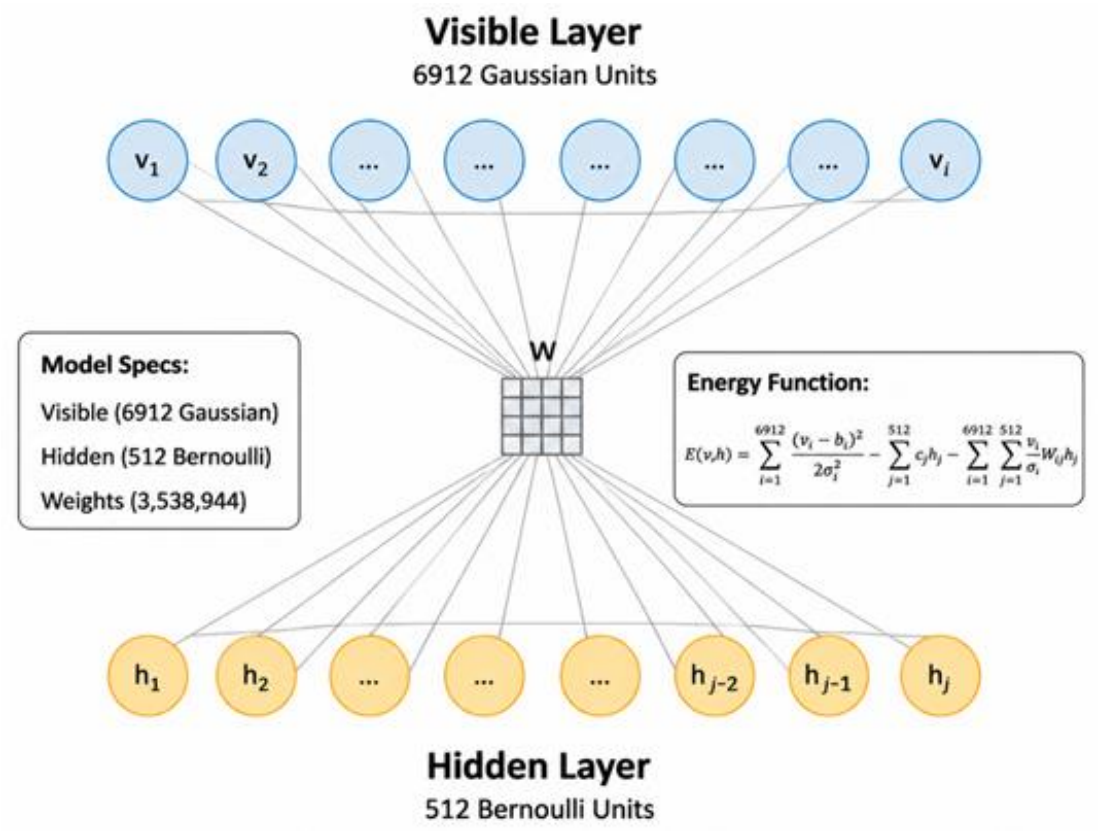


Fig. 5. Proposed GRBM network architecture.

3-3-4. Final Networks

The overall architecture of the proposed framework is presented in Figure 3. At the final stage, two separate fusion-based classification pipelines are investigated. In the first pipeline, the feature representation learned by the ConvLSTM2D network is fused with the representation generated by the GRBM, and the fused feature vector is then provided to the final classifier. In the second pipeline, the output of the CNN-3D network is independently fused with the GRBM representation, and the resulting combined feature vector is similarly passed to the same classification layer. In this way, the effectiveness of the proposed method is assessed in two different configurations, namely **GRBM + ConvLSTM2D** (Figure 6) and **GRBM + CNN-3D** (Figure 7).

For both configurations, the final classification stage is implemented using a single **Softmax layer**. This layer converts the fused input features into posterior probabilities over the seven target categories. By construction, the Softmax output forms a valid probability distribution, in which all components are non-negative and their sum equals 1. The class associated with the maximum probability is selected as the final prediction.

Mathematically, the Softmax function is expressed as:

$$p(y = j|x) = \frac{e^{x^t w_j}}{\sum_{k=1}^{k} e^{x^t w_k}} \quad (1)$$

where x is the fused input feature vector, w_j is the weight vector corresponding to class j, and K denotes the number of output classes. In the present study, K=7. The quantity $x^t w_j$ represents the inner product between the input vector and the weight vector of class j [30,31].

4. Simulation Results

- **Hardware and Software Environment**

The implementation and evaluation of the proposed hybrid architectures were conducted on a workstation featuring an Intel Core i7 processor, 16 GB of DDR4 RAM, and an NVIDIA GeForce MX350 GPU with 2 GB of dedicated memory. The software environment was built using Python, leveraging high-level deep learning libraries for the construction and optimization of the neural networks.

- **Training Protocol for Hybrid Architectures**

To ensure a rigorous comparative analysis, both proposed hybrid models **GRBM-ConvLSTM2D** and **GRBM-3DCNN** were trained and evaluated independently under identical experimental conditions.

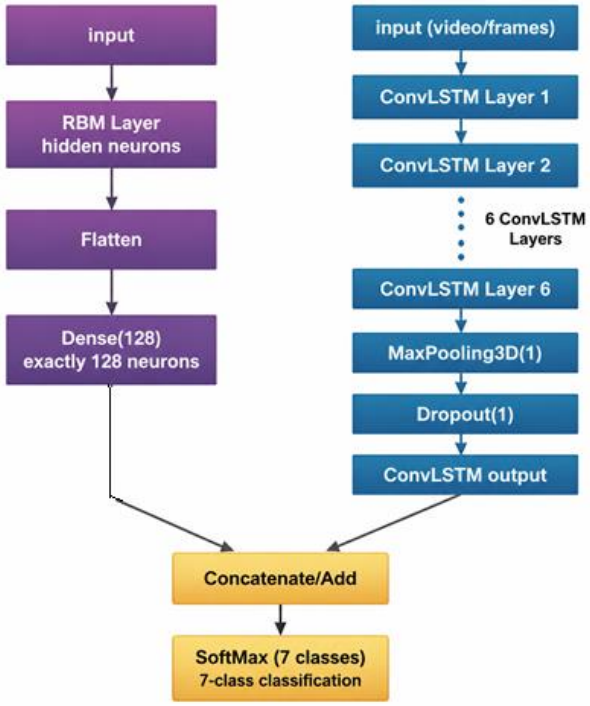


Fig. 6. Flowchart of the first model structure (ConvLstm2D+GRBM).

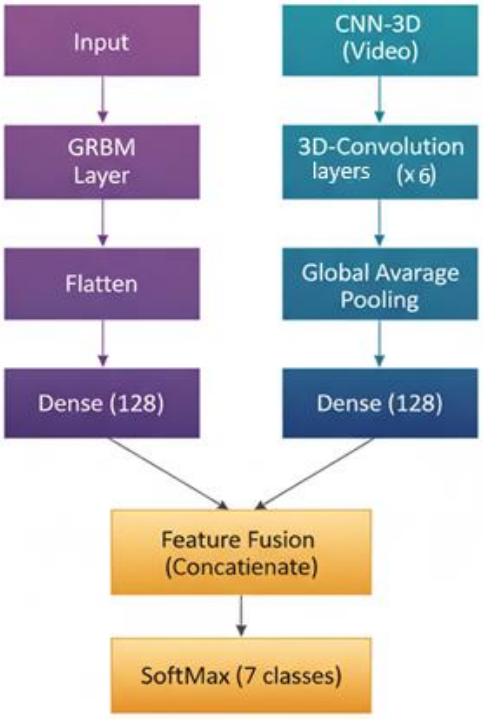


Fig. 7. Flowchart of the second model structure (CNN3D + GRBM).

1. **Feature Integration:** In both frameworks, the spatiotemporal features extracted by the recurrent (ConvLSTM2D) and volumetric

(3DCNN) components, in conjunction with the GRBM-refined representations, were fed into the final classification stage.

2. **Classification Layer:** The final layer of each model utilized a Softmax activation function to categorize the input video sequences into seven emotional classes: *happiness*, *sadness*, *disgust*, *excitement*, *contempt*, *fear*, and *anger*.
3. **Hyperparameters:** The classification heads were optimized over 35 training epochs with a consistent mini-batch size of 16. This configuration was selected to ensure stable convergence and to prevent overfitting, particularly given the balanced dataset provided by the augmentation process.

4-1. ConvLSTM2D Network Results

Following the application of the proposed ConvLSTM framework to the dataset, the network operated strictly as a feature extractor without an active classification layer. Consequently, the **Mean Squared Error (MSE)** loss function was adopted to evaluate the optimization process. The model was trained for **60 epochs**, yielding a final training loss of **0.0080** and a validation loss of **0.0038**. These training dynamics demonstrate stable convergence and the model's capacity to extract robust spatiotemporal representations. The corresponding training performance and optimization curves are illustrated in Figure 8.

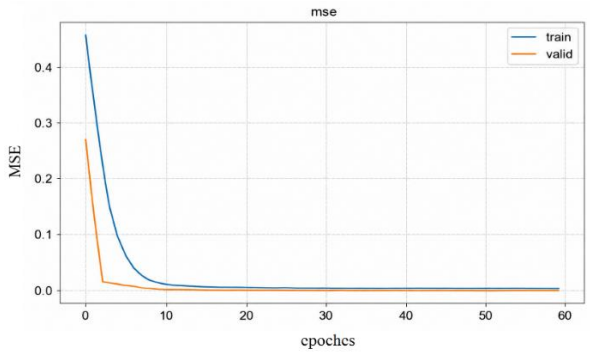


Fig. 8. Train and validation loss curve of the Conv-LSTM network.

4-2. CNN3D Network Results

The proposed **CNN-3D network**, trained independently as a feature extractor prior to GRBM integration, did not include a classification layer. Therefore, the **MSE loss** was adopted to monitor the learning process. After **60 epochs** of training with a learning rate of **0.001**, the model achieved a final training loss of **0.0023** and a validation loss of **0.0093**, demonstrating stable convergence and effective learning of spatiotemporal feature representations. The corresponding training curve is presented in Figure 9.

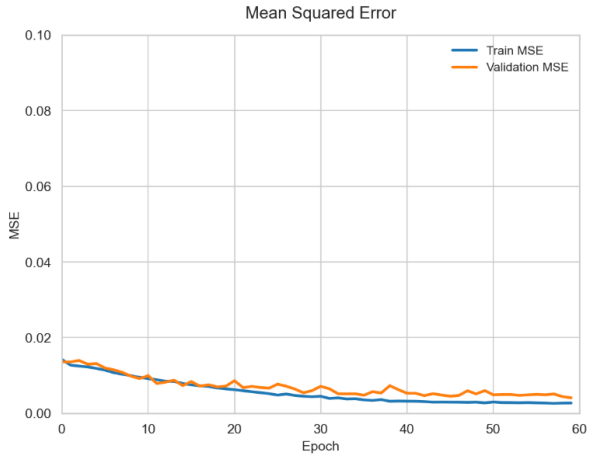


Fig. 9. Train and validation loss curve of the CNN3D network.

4-3. Gaussian RBM Network Results

Consistent with the objective of extracting latent feature representations, the **GRBM** module was implemented without a classification layer. The input data were reshaped into a flattened format before being fed into the network for unsupervised feature learning. The model was trained for **100 epochs**, achieving a final training loss of **0.0530** and a validation loss of **0.0550**. These metrics signify stable learning performance and the model's efficacy in reconstructing latent features from the input data. The detailed optimization profile and training curves are depicted in Figure 10.

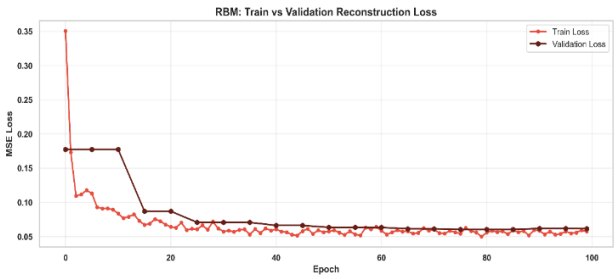


Fig. 10. Training and validation loss curve of the GRBM network.

4-4-1. Final Networks

Ultimately, the feature representations extracted from the individual sub-networks were fused and directed to the classification stage. The fundamental objective of integrating these distinct architectures is to synergize the latent representations from the GRBM with the spatiotemporal features learned by the deep sequential/volumetric blocks, thereby establishing a highly discriminative feature space.

Table 1. performance comparison between the two proposed hybrid networks.

Hybrid architecture	Training Accuracy	Validation accuracy	Training loss	Validation loss
GRBM+ConvLSTM2D	99.26%	91.16%	0.03	0.2
GRBM+CNN3D	100%	88.89%	0.02	0.54

The classification head consists of a dense layer coupled with a **Softmax** activation function, mapping the aggregated feature vectors into the seven target emotional classes. To evaluate the efficacy of the proposed fusion strategies, both hybrid networks were trained and validated under a batch size of 8 for 70 epochs.

- **GRBM+ConvLSTM2D Model:** This configuration converged efficiently, achieving an outstanding training accuracy of **99.06%** and a validation accuracy of **91.16%**. The corresponding optimization process yielded a training loss of **0.03** and a validation loss of **0.2**. The training convergence and loss progression curves are illustrated in Figures 11 and 12.
- **GRBM+3DCNN Model:** The volumetric hybrid model demonstrated competitive training characteristics, achieving a training accuracy of **100%** and a validation accuracy of **88.8%**. The final training loss converged to **0.002**, while the validation loss reached **0.54**. The optimization trajectories and classification performance curves for this architecture are provided in Figures 13 and 14.

Table 1 summarizes the performance comparison between the two proposed hybrid networks.

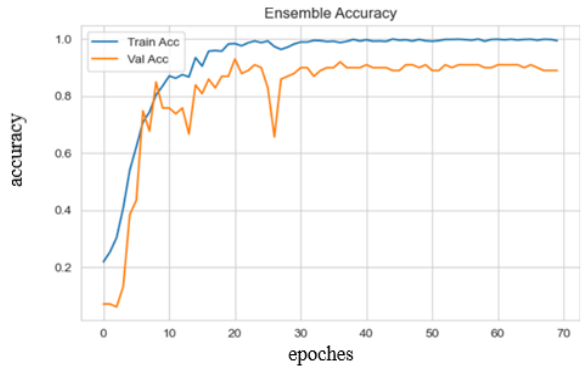


Fig. 11. Training and validation accuracy of the GRBM+ConvLSTM2D network.

To provide a granular and comprehensive diagnostic evaluation of the classification performance, Figures 15 and 16 illustrate the confusion matrices obtained during the testing phase for the two proposed hybrid architectures: **GRBM-ConvLSTM2D** and **GRBM-3DCNN**, respectively. These matrices elucidate the class-wise predictive dynamics, mapping out the precise classification rates and inter-class misclassifications (leakages) across all seven emotional categories.

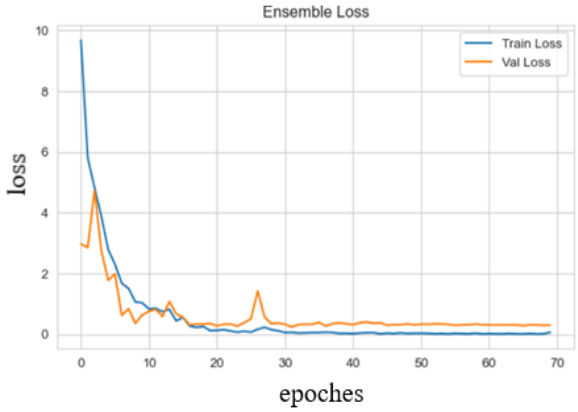


Fig. 12. Training and validation loss of the GRBM+ConvLSTM2D network.

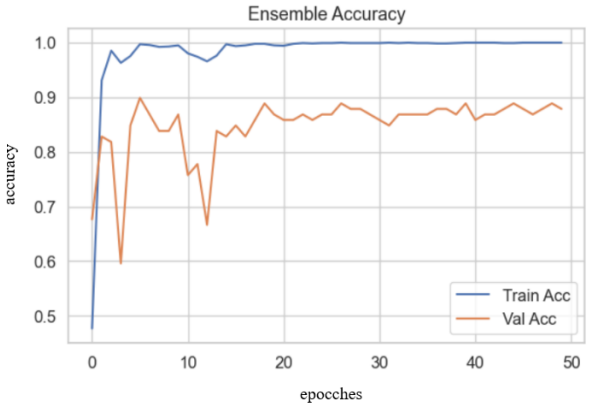


Fig. 13. Training and validation accuracy of the GRBM+3DCNN network.

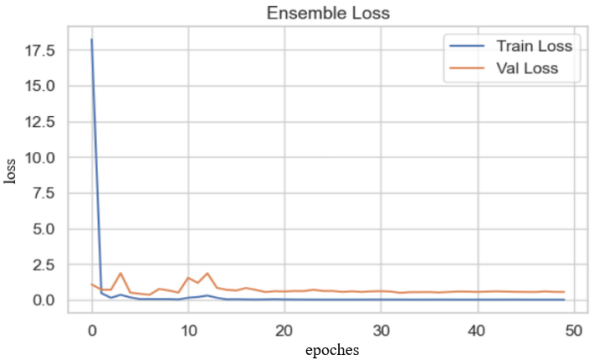


Fig. 14. Training and validation loss of the GRBM+3DCNN network.

A comparative analysis of the confusion matrices reveals that both models demonstrated relatively lower classification performance on

minority classes—specifically **Contempt, Fear, and Angry**—which were inherently constrained by a smaller volume of training samples. Conversely, dominant emotional categories exhibited significantly higher recognition accuracy. This discrepancy highlights the persistent challenges posed by class imbalance in facial emotion recognition (FER). Nevertheless, the overall classification performance across all emotional categories remains highly robust. This success is primarily attributed to the hybrid integration of the Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM), which effectively extracted low-level facial representations before temporal-spatial feature extraction.

To further validate and quantify the models' effectiveness, class-specific evaluation metrics, including **precision**, **Accuracy**, **recall**, **F1-score** and **Kappa**, were computed and are illustrated in Figure 17 (for the GRBM-ConvLSTM2D) and Figure 18 (for the GRBM-3DCNN).

A critical examination of the tabulated data yields the following insights:

- **GRBM-ConvLSTM2D Performance:** This configuration exhibited superior performance stability, achieving an F1-score of 95.24% for *Surprise* and 93.18% for *Happy*. Crucially, even for the highly challenging class of *Fear* (often characterized by high inter-class similarity with surprise and sadness), the model achieved a notable recall of 93.75%. The precision of 100% for the *Contempt* class further underscores the model's reliability in eliminating false positives for minority expressions.
- **GRBM-3DCNN Performance:** While the GRBM-3DCNN architecture maintained strong performance on prominent classes (such as a 100% precision for *Happy* and a 93.45% F1-score for *Surprise*), its capacity to generalize on minority classes was noticeably weaker. Specifically, the F1-scores for *Angry* (65.21%) and *Fear* (68.57%) indicate a higher susceptibility to classification degradation under data scarcity, compared to its ConvLSTM2D counterpart.

In summary, while both hybrid networks prove to be highly effective frameworks for FER, the **GRBM-ConvLSTM2D** exhibits greater resilience to class imbalance. This suggests that the sequential-temporal modeling of ConvLSTM2D, combined with the generative feature extraction of the GRBM, provides a more robust representation of subtle dynamic micro-expressions compared to the purely volumetric convolutions of the 3DCNN.

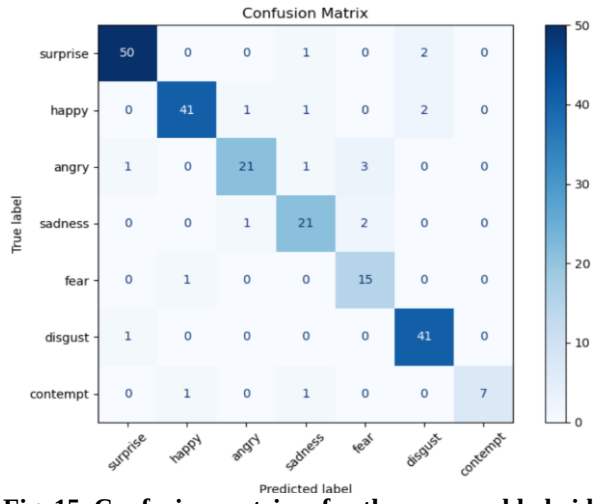


Fig. 15. Confusion matrices for the proposed hybrid model (GRBM-ConvLSTM2D).

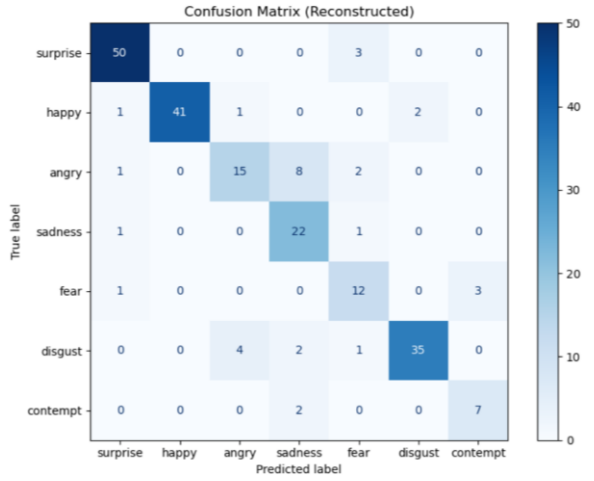


Fig. 16. Confusion matrices for the proposed hybrid model (GRBM-3DCNN).

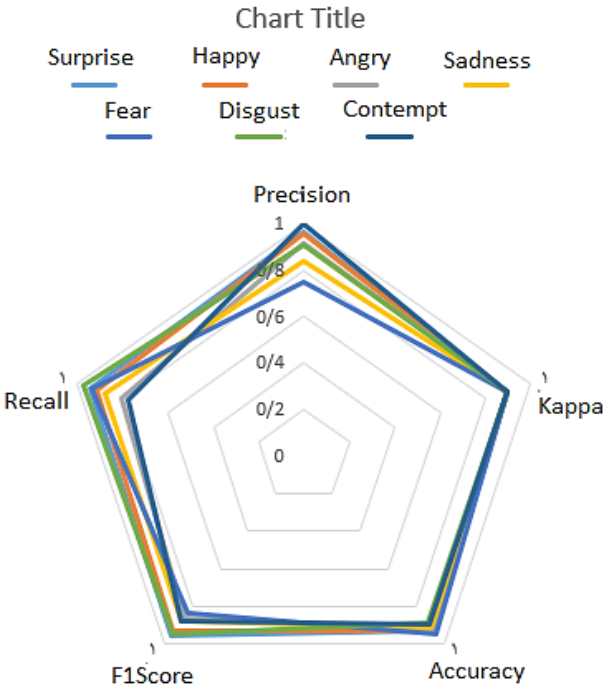


Fig. 17. Radar chart of different scores for the GRBM-ConvLSTM2D network.

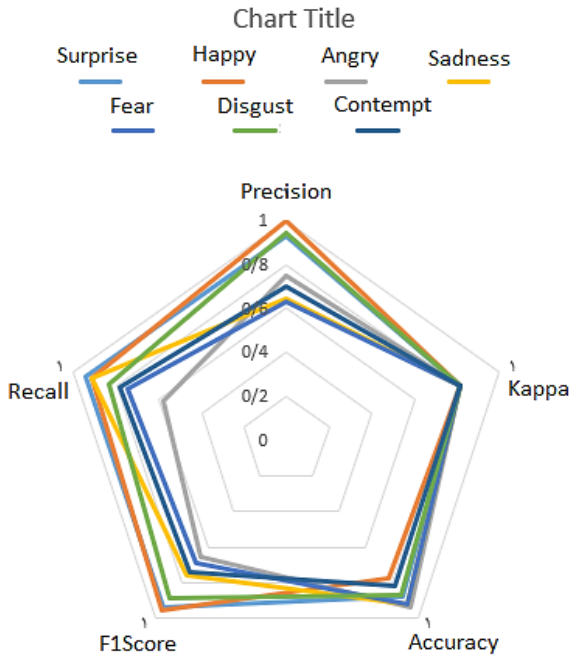


Fig. 18. Radar chart of different scores for the GRBM-3DCNN network.

To contextualize the performance of the proposed hybrid frameworks, Table 2 presents a comparative analysis between our models and several state-of-the-art methodologies evaluated on similar benchmark datasets.

As indicated in the comparative data, our proposed **GRBM-ConvLSTM2D** architecture achieves a superior validation accuracy of **91.16%** on the CK+ dataset, significantly outperforming recent deep learning approaches, including the EfficientNetB7-CNN cascade (78.72%) and traditional DBN fusion architectures (73.73%). Furthermore, our volumetric hybrid variant, the **GRBM-3DCNN**, demonstrates a highly competitive accuracy of **88.80%**.

Crucially, while single-stage transfer learning models or standard CNN-RNN architectures often suffer from performance degradation when subjected to the inherent class imbalances of datasets like CK+, both of our proposed hybrid models exhibit remarkable robustness. This stability is attributed to the initial unsupervised feature reconstruction phase performed by the GRBM, which effectively preserves critical facial topology before temporal and spatial feature extraction.

5. Conclusion

Facial Emotion Recognition (FER) remains a pivotal area of research within computer vision and human-computer interaction. In this study, we introduced and evaluated two novel hybrid deep learning architectures **GRBM-ConvLSTM2D** and **GRBM-3DCNN** designed to classify seven distinct facial expressions (Surprise, Happiness, Anger, Sadness, Fear, Disgust, and Contempt) from video sequences. The core paradigm of our approach relies on utilizing a Gaussian-Bernoulli Restricted Boltzmann Machine (GRBM) as a generative front-end to reconstruct and extract low-level spatial facial features, which are subsequently processed by temporal-spatial classifiers. The experimental simulations, implemented in Python, demonstrated the outstanding efficacy of the proposed hybrid systems on the benchmark CK+ dataset: 1-The **GRBM-ConvLSTM2D** network emerged as the top-performing configuration, achieving a peak validation accuracy of **91.16%**, with a highly stabilized training loss curve. This model proved highly resilient to the challenges of class imbalance, yielding balanced precision, recall, and F1-scores across both majority and minority emotional classes. 2- The **GRBM-3DCNN** configuration achieved a competitive validation accuracy of **88.80%**, proving its capability to capture volumetric spatial-temporal patterns, albeit with slightly higher sensitivity to classes constrained by data scarcity. Overall, the empirical findings confirm that coupling generative energy-based models (GRBM) with sequential deep networks (ConvLSTM2D/3DCNN) yields a synergistic effect that enhances feature discrimination in dynamic FER. Future research will focus on optimizing these hybrid pipelines for real-time edge-computing applications and validating their generalization capability on highly unconstrained, in-the-wild facial expression databases.

Table 2. Performance comparison of the proposed hybrid models against state-of-the-art methods.

Models	Datasets	Refs.	Network	Accuracy
EfficientNetB7CNN	CK+ +FER24	Nursel Yalçin [a], Muthana Alisawi [30]	EfficientNetB7+CNN	78.72%
Fusion DBN	RML	SHIQING ZHANG, XIAN ZHANG [17]	CNN+CNN+DBN	73.73%
Transfer learning	Emotional Wearable 202	Haposan Manalu, Achmad Rifai [21]	CNN+RNN	61.43%
This study	CK+	Ours	CNN3D+RBM	88.8%
This study	CK+	Ours	CNNLSTM+RBM	91.16%

References

[1] Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.T., Wang, J.Z., Li, J. and Luo, J., (2011). Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5), 94-115.

[2] Jain, D.K., Shamsolmoali, P. and Sehdev, P., (2019). Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120, 69-74.

[3] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, 115, 101–106.

[4] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *In 2017 International Conference on Engineering and Technology (ICET)*, (pp. 1–6). Antalya: IEEE

[5] Smolensky, P., (2016). Information processing in dynamical systems: foundations of harmony theory," *Parallel distributed processing: Explorations in the microstructure of cognition*, 1, 194-205.

[6] Shavandi, M. and Afrakoti, I.E.P., (2019). Face recognition in thermal images based on sparse classifier. *International Journal of Engineering*, 32(1), 78-84.

[7] Yasuda, M. and Xiong, Z. (2023). New learning algorithm of Gaussian Bernoulli restricted Boltzmann machine and its application in feature extraction, *Proc. 2023 International Symposium on Nonlinear Theory and Its Applications*, 134-137.

[8] Hinton, G. and Salakhutdinov, R. (2021). Reducing the dimensionality of data with neural networks, *Science*, 313 (8), 504-507.

[9] Afrakoti, I.E.P., (2020). Speech emotion recognition using scalogram based deep structure. *International journal of engineering. Transactions B: Applications*. 33(2), 285-292.

[10] Rakshith, M. D., Harish, H. Kenchannavar, (2024), hybrid deep optimal network for recognizing emotions using facial expressions at real time. *intelligent systems and applications* 3, 47-58.

[11] Wahab, A., Nazir, M. N., A., Ren, A. T. Z., and Noor, M. H. M. (2019). Efficientnet-lite and hybrid CNN-KNN implementation for FER on raspberry Pi. *IEEE Access*, 7.

[12] Li, X., Pfister, T., Huang, X., Zhao, G., (2013), A Spontaneous Micro expression Database: Inducement, collection and baseline, *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, Shanghai, China*, 1-6

[13] Lakshan, I., Wickramasinghe, L., Disala, S., Chandrasegar, S., Prasanna S. (2019), real time deception detection for criminal investigation, *national information technology conference(NITC)*, 90-96. IEEE.

[14] Wei, L., Kauttonen, L, and Zhao, A., (2022), Deep Learning for Micro-Expression Recognition: A Survey, *IEEE Transactions on Affective Computing*, 13(4), 2028-2046.

[15] Karimi, H., Tang, J., & Li, Y. (2018). Toward end-to-end deception detection in videos. *In 2018 IEEE international conference on big data (Big Data)*, 1278-1283. IEEE.

[16] Martinez, B., Valstar, M.F., Jiang, B. and Pantic, M., (2017). Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*, 10(3), 325-347.

[17] Sariyanidi, E., Gunes, H. and Cavallaro, A., (2014). Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE transactions on pattern analysis and machine intelligence, 37(6), 1113-1133.

[18] Zhang, S., Pan, X., Cui, Y., Zhao, X. and Liu, L., (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*, 7, 32297-32304.

[19] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., (2015). Learning spatiotemporal features with 3d convolutional networks. *In Proceedings of the IEEE international conference on computer vision*, 4489-4497.

[20] Hasani, B. and Mahoor, M.H., (2017). Facial expression recognition using enhanced deep 3D convolutional neural networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 30-40.

[21] Zhang, S., Zhang, S., Huang, T., Gao, W. and Tian, Q., (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE transactions on circuits and systems for video technology*, 28(10), 3030-3043.

[22] Manalu, H.V. and Rifai, A.P., (2024). Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. *Intelligent systems with applications*, 21, 200339.

[23] Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A. and Singh, S., (2023). Facial expression recognition in videos using hybrid CNN & ConvLSTM. *International Journal of Information Technology*, 15(4), 1819-1830.

[24] Zhang, J., Wang, H., Chu, J., Huang, S., Li, T. and Zhao, Q., (2019). Improved Gaussian–Bernoulli restricted Boltzmann machine for learning discriminative representations. *Knowledge-Based Systems*, 185, 104911.

[25] Chaudhari, A., Bhatt, C., Krishna, A. and Mazzeo, P.L., (2022). ViTFER: facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4), 80.

[26] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I., (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *In 2010 ieee computer society conference on computer vision and pattern recognition-workshops*, 94-101. IEEE.

[27] Zhang, J., Wang, H., Chu, J., Huang, S., Li, T. and Zhao, Q., (2019). Improved Gaussian–Bernoulli

restricted Boltzmann machine for learning discriminative representations. *Knowledge-Based Systems*, 185, 104911.

[28] Hinton, G.E., 2012. A practical guide to training restricted Boltzmann machines. In Neural Networks: Tricks of the Trade: Second Edition (pp. 599-619). Berlin, Heidelberg: Springer Berlin Heidelberg.

[29] Cho, K., Raiko, T. and Ihler, A.T., (2011). Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In Proceedings of the 28th international conference on machine learning (ICML-11), 105-112.

[30] Yalçin, N. and Alisawi, M., (2024). Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques. *Heliyon*, 10(20).

[31] Goodfellow, A., Bengio, I. and Courville, Y., Softmax Units for Multinoulli Output *Distributions''Deep Learning, in '6.2. 2.3 Softmax Units for Multinoulli Output Distributions' Deep Learning*, 2016, 23-47.

Biography

Haniyeh Amirbeigi was born in Hamedan, Iran, in 1999. She received her B.Sc. degree in Electrical Engineering (specializing in Electronic) from Hamedan University of Technology in 2021. She continued her studies at the same institution, earning an M.Sc. degree in Micro and Nano Electric devices in 2026. Her research interests include Deep Learning, Artificial Intelligence, Image Processing, and Video Processing. Her master's thesis focused on video processing, a field in which she has also published several research papers.

Hamid Reza Shahdoosti was born in Hamedan, Iran, in 1984. He received the B.S.E.E., M.S.E.E, and Ph.D. degrees from the Amirkabir University of Technology, Sharif University of Technology, and Tarbiat Modares University, Tehran, Iran, in 2007, 2010, and 2014, respectively. Since 2014, he has been with the Department of Electrical Engineering, Hamedan University of Technology, Hamedan, Iran, where he is an associate professor of electrical engineering. He has received the annual award for the Best Researcher of the Year in 2018 in Hamedan Province. His current research interests focus on deep learning, artificial intelligence (AI), information fusion, energy management with AI, and internet of things (IOT). His research in these fields ended up in several journal and conference publications.