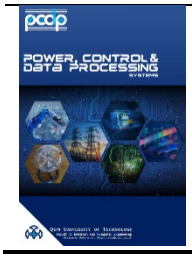
DOI: 10.30511/pcdp.2026.2087714.1069 - ISSN: 3060-8376 - Vol 3(3), 2026



POWER, CONTROL AND DATA PROCESSING SYSTEMS

Available Online at: https://pcdp.qut.ac.ir/

A New Method for Detecting Emotions in Video Using a Hybrid CNN-LSTM-RBM Network

ARTICLE INFO

Article Type
Original Research

Authors
Haniyeh Amirbeigi[1]
Hamid Reza Shahdoosti[2,*]

[1] Electrical and Computer Engineering Department, Hamedan University of Technology, Hamedan, Iran, haniyeamirbeygi@gmail.com
[2] Electrical and Computer Engineering Department, Hamedan University of Technology, Hamedan, Iran, h.doosti@hut.ac.ir

*** *Correspondence***
Address: Electrical and Computer Engineering Department, Hamedan University of Technology, Fahmideh Avenue, Hamedan, Iran. Postal Code: 65155579.
Phone: +988138411529
Fax: +988138380660
h.doosti@hut.ac.ir

Article History
Received: April 29, 2026
Accepted: July 01, 2026
ePublished: September 01, 2026

ABSTRACT

Facial expressions are one of the most important techniques of non-verbal communication that humans use to display their emotional states and establish non-verbal communication. In recent research in the field of emotion recognition based on facial expression changes, deep learning networks are used. In this regard, an appropriate network is selected according to the type of data. In the present research, video data is used, and for recognizing facial expressions, recurrent networks are employed. Subsequently, to achieve better results, a hybrid network consisting of Convolutional, Recurrent, and Restricted Boltzmann Machine networks has also been used. Finally, in this paper, using deep learning techniques and the Python language, an algorithm based on a hybrid CNN-LSTM-RBM neural network for emotion recognition through facial expressions in a video is presented. In this method, seven emotional states (happy, sad, disgust, excitement, contempt, fear, and anger) are detected. The CK+48 dataset has been used to train and test the proposed model. The experimental results demonstrate that the proposed Hybrid CNN-LSTM-GRBM model achieves a training accuracy of 99.06%. Crucially, the model attains 91.16% test accuracy and an F1-score of 89.27% on the test set**,** outperforming existing baseline methods and confirming its robustness in handling spatiotemporal facial features.

Keywords: Emotion Recognition; Hybrid Neural Network; Deep Learning; Convolutional Neural Network; Recurrent Network; Restricted Boltzmann Machine.

1 Introduction

Generally, communication can be divided into two types: verbal and non-verbal. The goal of emotion recognition is to establish non-verbal communication, which possesses the most since emotions. The most important technique of non-verbal communication is facial expressions for expressing emotions. Body language is a part of non-verbal communication formed by different parts of the body, such as body postures, hand and foot movements, facial signs and various facial expressions, and eye contact. As we know, to understand a person's mood, one can directly observe their face [1 , 2]. The goal of facial expression recognition is to detect a person's state of mind through analyzing their face. One of the challenges and weaknesses of machines is the correct recognition of a person's emotions based on received images of the person. Considering that analyzing facial expressions for emotion recognition can give robots emotions or, from another perspective, enable them for emotional reactions, it also has other important advantages and applications. This can be useful in areas such as assisting in diagnosing depression, helping autistic patients with communication, and even by employing this technology, blind people can be equipped with smart glasses or smart pupils that have the ability to recognize emotions for communication. One of the widely used tools in deep learning is the Convolutional Neural Network (CNN). The CNN model is typically used to solve computer vision problems such as object tracking, image classification, and object detection and recognition [3]. Another widely used tool in machine learning, especially for data with temporal sequences like video, is the Recurrent Neural Network (RNN). RNN converts image components into sequential data and improves image processing by preserving temporal information [4]. The most important feature of this method is that it has a recurrent loop. This recurrent loop, using internal memory, causes information obtained from previous moments to remain in the network. Although conventional Recurrent Neural Networks (RNNs) are capable of processing sequential data, they are fundamentally hindered by the vanishing gradient problem. To overcome this limitation, Long Short-Term Memory (LSTM) networks were employed. By integrating a dedicated cell state and sophisticated gating mechanisms, LSTMs effectively mitigate gradient instability, thereby enabling the network to capture complex long-term dependencies across extensive temporal sequences [5]. And finally, the GRBM (Gaussian-Bernoulli Restricted Boltzmann Machine), which is a Markov random field of the bipartite type, consists of two types of layers, namely visible and hidden layers, and can work with continuous data points [5 and 6]. This network extracts additional features in an unsupervised manner, like a simple autoencoder, compressing and reconstructing the input. This method is used in image reconstruction, colorizing black-and-white images, image denoising [7], and dimensionality reduction [8] of images.

2 Related Works

As Facial images provide relevant information for understanding emotions. As humans, we can easily infer a person's emotional state from their facial movements [9]. Emotion recognition using facial expressions is an advancement in determining various human facial emotions to infer an individual's psychological state. This recognition framework has been applied in several fields but is most commonly used in the medical realm to determine mental health problems [10]. Facial Expression Recognition (FER) is a computer vision field useful for various techniques of human emotion detection from facial expressions. Researchers are interested in FER because understanding an individual's emotions can improve human-machine interaction, behavioral sciences, and clinical performance. Technological advancements and image classification techniques enable the development of more efficient FER systems. FER systems find applications in healthcare systems, social marketing, targeted advertising, the music industry, school counseling systems, and lie detection [11]. Micro-expressions, fleeting facial displays that reveal the genuine emotions an individual is trying to conceal, were first observed by Haggard and Isaacs in 1966. Ekman and Friesen later investigated micro-expressions and highlighted their importance through a study examining films of a psychiatric patient. Despite the patient consistently appearing content, they transiently revealed a hidden expression of sadness lasting only two frames (1/12 of a second). Research on micro-expressions emphasizes their crucial role in detecting deception, especially in high-stakes situations like police interrogations, where dishonesty may have severe consequences [12]. For instance, deploying FER in police investigations for lie detection is highly complex and requires deep understanding of psychology and body language. Involving experts in these fields is crucial for validating lie detection systems and minimizing error rates [13]. Similarly, FER used in the judicial sector often employs sophisticated FER methods, such as the study by Karimi et al., which developed hybrid CNN-RNN algorithms for lie detection purposes [14].

FER methods can be divided into two categories: video sequence-based (dynamic) methods and image-based (static) methods. Most prior FER studies focus on identifying facial expressions from static facial images [15, 16]. One of the key challenging issues in Facial Expression Recognition (FER) in video sequences is extracting spatio-temporal video features from video sequences [17]. Although these image-based methods can effectively extract spatial information from static images, they cannot capture temporal changes across consecutive frames in a video. As a dynamic event, classifying facial expressions from consecutive frames in a video is common, as video sequences provide much more information than static facial images for FER [17]. A major drawback of conventional CNNs is that they are capable of extracting spatial relationships of input images but cannot model their temporal relationships in a video sequence. To address this

problem, recently developed 3D CNNs may offer a viable solution [18]. 3D CNNs can extract spatio-temporal features in a video sequence by convolving over the temporal dimension of the input data as well as the spatial dimension simultaneously. In recent years, 3D CNNs have been used to learn the embedded features present in the spatio-temporal dimension of consecutive frames [19], [20]. Although these 3D-CNN-based methods have achieved effective performance for FER in video sequences, a significant problem remains in that these methods cannot simultaneously consider deep spatio-temporal features in their extraction process [20]. Consequently, efforts are directed towards using effective hybrid networks to solve such problems, which is why the practical RNN network can be mentioned, as its importance is precisely apparent in fields where temporal sequence is important. An RNN can be interpreted as a type of network that incorporates a loop as part of its feedforward mechanism. RNNs can be useful for sequential or time-dependent data. An advanced version of RNN, known as Long Short-Term Memory (LSTM), is used to capture the temporal and contextual complexities of facial expressions. Unlike traditional recurrent neural networks, LSTM addresses the vanishing gradient problem, where the gradient value becomes negligible, hindering information transfer in long sequences. LSTM introduces gates that regulate the flow of information, mitigating these issues [21].

Also, Restricted Boltzmann Machines (RBMs) have attracted significant research attention in recent years because they can discover latent representations in an unsupervised manner. Standard RBMs are only suitable for processing binary data. To overcome this limitation, the Gaussian-Bernoulli RBM (GRBM) was designed to model real-valued data, particularly images. A GRBM that aims to non-linearly map real-valued data into a latent space is typically trained using contrastive divergence learning [22-24].

Recent advancements in video-based emotion recognition have emphasized the integration of temporal dynamics. Prathwini and Prathyakshini [25] introduced a multi-phase framework that utilizes CNN-based feature extraction coupled with various classifiers like RNN, SVM, and KNN to analyze video sequences. While their research underscores the importance of temporal dependencies, our proposed model adopts a distinct architectural approach by leveraging a hybrid CNN-LSTM-GRBM framework. Unlike [25], which evaluates distinct classification pipelines, our architecture integrates a Gaussian Restricted Boltzmann Machine (GRBM) to extract latent, unsupervised features, and employs ConvLSTM to concurrently capture spatiotemporal dependencies. While recent studies [21] have explored various CNN-based architectures for emotion recognition, our approach uniquely combines the temporal modeling capabilities of LSTM with the feature extraction power of RBMs, achieving competitive results on the standard CK+48 dataset.

In parallel, other research efforts have addressed the complexities of video-based emotion recognition by shifting toward multi-stage hybrid or multimodal architectures. For instance, [26-28] utilizes a motion-aware framework with dynamic key-frame selection and YOLOv5 to isolate regions of interest before applying an Augmented-Attention Convolutional Transformer Fusion Network (AACTFN). Similarly, recent studies [29-30] have proposed multimodal ensemble systems that integrate CNNs for facial analysis, BERT for textual sentiment, and RNNs for temporal tracking to achieve a holistic emotional intelligence. While such approaches—whether through multi-stage fusion or multimodal integration—utilize complex modules to mitigate issues like occlusion and background noise, they often introduce significant computational overhead. This reliance on heavy feature-extraction pipelines or heterogeneous model ensembles contrasts with the proposed CNN-LSTM-GRBM architecture, which prioritizes lightweight, end-to-end representation learning to capture latent emotional dynamics through restricted Boltzmann machines.

In this article, Section 3 introduces the dataset and the preprocessing applied to it, and also explains the hybrid neural network. Section 4 presents the simulation results, and Section 5 provides the conclusion.

3 Materials and methods

3.1 Dataset

To train and test the proposed method, the **CK+48** dataset was used. CK+48 is a cropped and downsized version of the original **Extended Cohn-Kanade (CK+)** dataset, which is designed for Facial Emotion Recognition. This dataset was developed by Carnegie Mellon University (CMU) and is mainly used for AI and image processing research [31-34]. The selected dataset contains 981 images and 7 emotions. It is worth mentioning that 3 frames are available for each emotion, and the images have a resolution of 48×48.In general, the images extracted from video frames do not have much diversity, and the total number of samples is relatively small compared to other datasets. However, the images are frontal-view, and the facial expression patterns are clean and distinguishable (see Figure 1) [23].

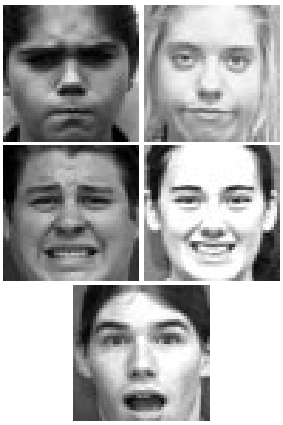


Figure 1: Sample images from the CK+48 dataset

3.2 Preprocessing

To ensure a robust evaluation and avoid data leakage, the original dataset, consisting of 981 images, was partitioned into training, test sets using a ratio of 80%, 20% respectively. Data augmentation techniques were applied exclusively to the training subset to address class imbalance and enhance model generalization. Consequently, the training set was expanded to 4,308 samples, while the test sets remained in their original, un-augmented state. The model's final performance was evaluated using the unseen 20% test set, as represented in the provided confusion matrix, ensuring that the reported accuracy reflects the model's performance on genuine, original facial images.

- Images were rotated by 10 degrees,
- Translated by 10% of their width and height,
- Zoomed by 10%,
- And horizontally flipped.

To address class imbalance, class-specific augmentation coefficients were employed. Classes with fewer samples, such as contempt, Fear, and Sadness, were subjected to higher augmentation factors or increased repetition to effectively balance the dataset. Consequently, the total number of images increased from 981 to 4,308, resulting in a more uniform distribution where each class contains three original frames plus their corresponding augmented versions. This expansion not only mitigated the risk of overfitting but also enabled the model to learn more robust features, making it resilient to geometric transformations such as rotation, translation, scaling, and horizontal flipping. Since all images in the dataset were grayscale, augmentation was applied to a single channel, maintaining pixel values within the [0, 255] or normalized [0, 1] range. Finally, to prevent data leakage, data augmentation was exclusively applied to the training set (Figure 2), ensuring that the validation and test sets remained pristine, thereby providing an unbiased assessment of the model's performance.

Figure 2: Sample images after applying data augmentation

3.3 Proposed method

The proposed model consists of a combination of three different networks:

- Two networks CNN and LSTM, combined into a Convolutional-LSTM network,
- And one Gaussian-RBM network.

First, both networks are trained without a classification layer using the generated dataset. It is important to note that the input dataset for the RBM is a combination of noisy and noise-free images. Finally, the outputs of the two networks are merged and fed into a SoftMax function for emotion classification.

3.3.1 CNN-LSTM Hybrid Network

For this part, the ConvLSTM2D model was used. This model is suitable for video data or data with temporal sequences because it extracts spatiotemporal features (Figure 3).
In each layer:

- hard sigmoid is used as the activation function for the gates,
- tanh is used as the activation function for the layer output.

After each filtered layer:

- A MaxPooling layer to preserve spatial dimensions,
- A Dropout layer with a rate of 0.3 to prevent overfitting,
- And a BatchNormalization layer to avoid vanishing gradients are applied.

The hard sigmoid activation function is a linear approximation of the sigmoid function, offering higher computational speed, but it still suffers from gradient vanishing. Its definition is presented below:

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

The tanh activation function maps its input to the range −1 to +1. It is symmetric around the origin, producing centered outputs, which helps gradients flow better during backpropagation and partially mitigates the vanishing gradient problem:

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

In this section, the network does not include a softmax layer because the goal is solely to extract spatiotemporal features. The proposed network contains six ConvLSTM2D layers and is trained independently with a learning rate of 0.0001.

3.3.2 Gaussian RBM Network

The network used here is the Gaussian RBM (GRBM). This network is an unsupervised model, where the inputs are the flattened image data and the outputs are reconstructions of those same inputs.

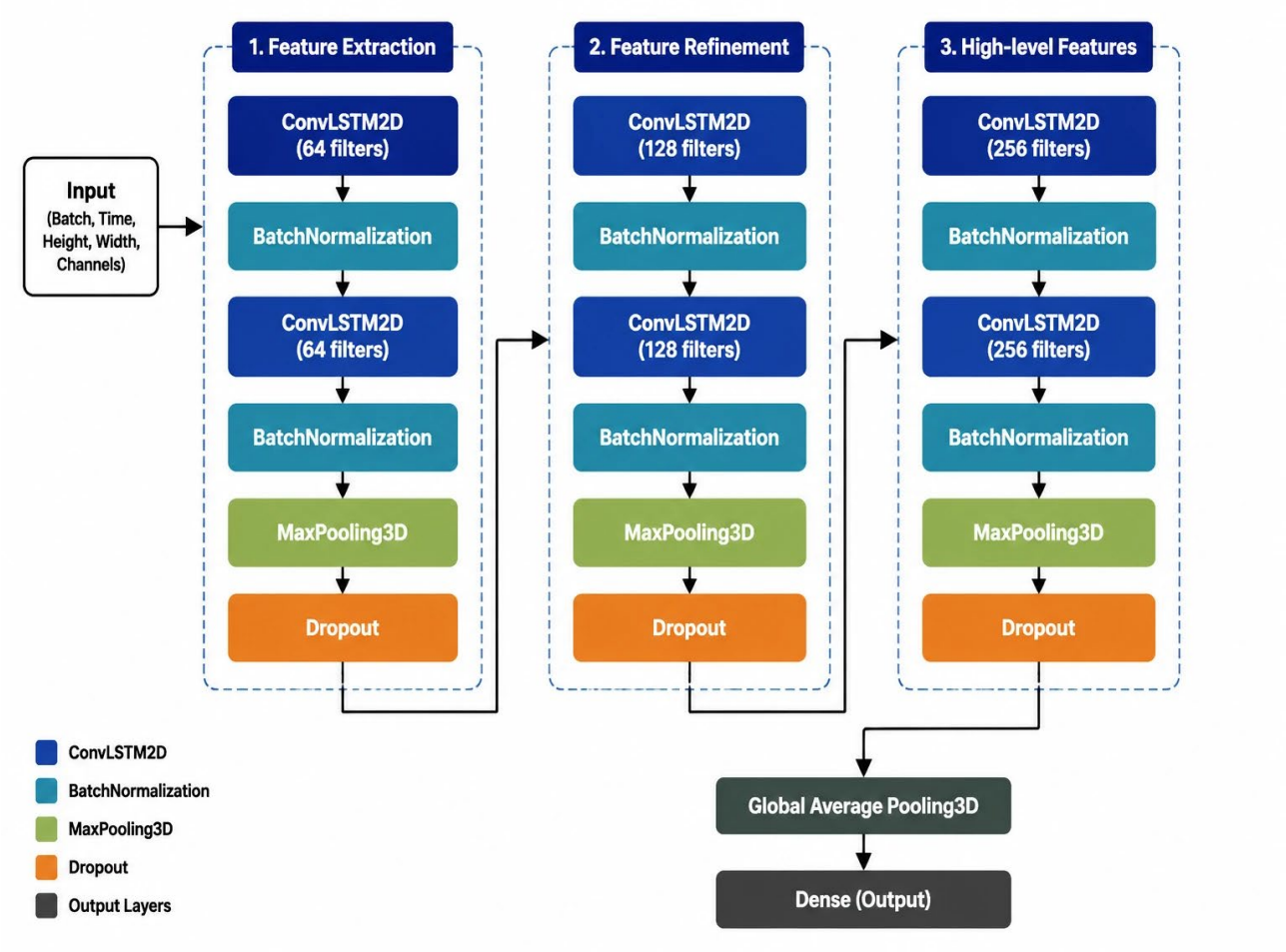


Figure 3: Proposed ConvLSTM network architecture

The purpose of using this network is to learn hidden features from the images without labels. Our final goal in combining this network with the CNN–LSTM network is to achieve higher accuracy and less overfitting. GRBM (Figure 4) is suitable for continuous and normalized data in ranges (0,1) or (−1,1).

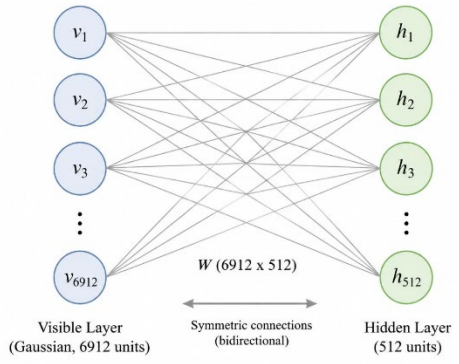


Figure 4: proposed GRBM network architecture

Since continuity is important in our input data, this network is appropriate for integration with the other network. The number of input neurons for this network is 6912, and the number of hidden neurons is 512. Gaussian noise with standard deviation **0.1** is used during reconstruction to help prevent overfitting. The energy function of the RBM, when extended with a quadratic term, yields the energy function of the Gaussian RBM:

$$E(V,h|\theta) = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{n_h} c_j h_j - \sum_{i=1}^{n_v} b_i v_i + \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} \quad \text{Eq 3}$$

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

$$E(V,h|\theta) = -\sum_{i=1}^{n_v}\sum_{j=1}^{n_h} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{n_h} c_j h_j + \sum_{i=1}^{n_v} \frac{(v_i - b_i)^2}{2\sigma_i^2} \quad \text{Eq 4}$$

In this form, the linear term $\Sigma_i (b_i v_i)$ is implicitly absorbed into the quadratic term. This simplification is valid when σ is considered a fixed parameter [26, 27]. In the GRBM module, the standard deviation of the Gaussian noise, σ=0.1, is employed within the energy function. This parameter is critical for aligning the energy landscape with the normalized intensity distribution of the input features [0,1], Empirically, we determined that σ=0.1 provides an optimal balance for

Contrastive Divergence (CD) training; it forces the model to learn discriminative latent representations by preventing energy landscape saturation while ensuring stable gradient updates during the independent training phase of this pathway. The probabilistic distributions in this model are as follows:

$$p(h_i = 1|\mathrm{v}) = \sigma(W_i^T \mathrm{v} + c_i) \quad \text{Eq 5}$$

$$p(V|\mathrm{h}) = \mathcal{N}(\mathrm{v}|\mathrm{Wh} + \mathrm{b}, \sigma^2 \mathrm{I}) \quad \text{Eq 6}$$

3.3.3 The Final network

The overall structure of the proposed network is shown in Figure 5. Finally, the outputs of the two mentioned networks are used as the input to the final network. Our final network consists solely of a SoftMax layer, which classifies the input data into 7 categories. Normalization in this case is applied with respect to the predicted output classes. This function generates a vector containing a probability distribution, where the components sum to 1.

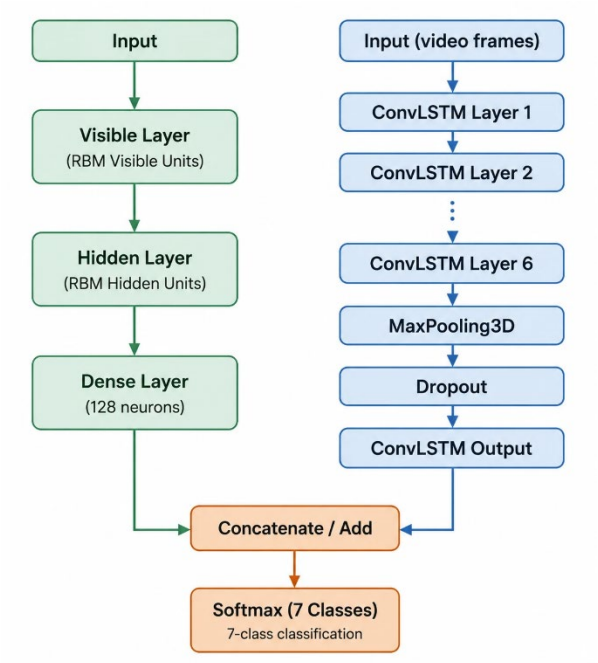


Figure 5: Flowchart of the overall model structure

The behavior of the SoftMax function is given by Equation (7). In multinomial logistic regression and linear discriminant analysis, if the input sample vector is x and the weight vector is w, the predicted probability for class j is:

$$p(y = j|x) = \frac{e^{x^t w_j}}{\sum_{k=1}^{k} e^{x^t w_k}} \quad \text{Eq 7}$$

Here, $x^T w_j$ is the inner product of the input vector and the weight vector [29]. To ensure the robustness of the emotion recognition model and to address the inherent class imbalance within the CK+48 dataset, a cost-sensitive learning approach was adopted during the training phase. Since some emotion categories (such as *fear* and *contempt*) contain fewer training samples compared to others, a class-weighting strategy was integrated into the loss function. This mechanism assigns higher penalties for misclassifying minority classes, thereby forcing the model to learn more representative features for these categories. The class weights were configured as follows: {0:1 ,1:1 ,2:1 ,3:3 ,4:5 ,5:1 ,6:5}, corresponding to the emotion labels: {surprise, happy, anger, sadness, fear, disgust, contempt}, respectively. This configuration significantly improved the model's generalization capabilities across all seven emotion classes.

4 Simulation Results

For training and testing the proposed model, a system equipped with a Core i7 processor, 16 GB RAM, 1 TB SSD storage, and an MX350 2GB GeForce GPU was used. The performance of the proposed model is as follows: the ConvLSTM2D and GRBM models were each trained separately using the CK+48 dataset. The outputs and extracted features of these two networks were then used as the input to a final layer equipped with a SoftMax function, which performs the desired classification. In the final layer, 70 epochs with a batch size of 8 were used. It should be noted that the images were pre-labeled with the classes: happy, surprise, neutral, sad, fear, anger, and disgust.

4.1 ConvLSTM2D Network Results

After applying the proposed model to the dataset, the output of the network does not include a classification layer. Therefore, for validating the network, the Mean Squared Error loss function was used, and the network was trained for 100 epochs. As a result, the training loss reached **0.008**, and the validation loss reached **0.0038**, which indicates the network's very good and successful performance. Figure 6 shows the performance graph of the network.

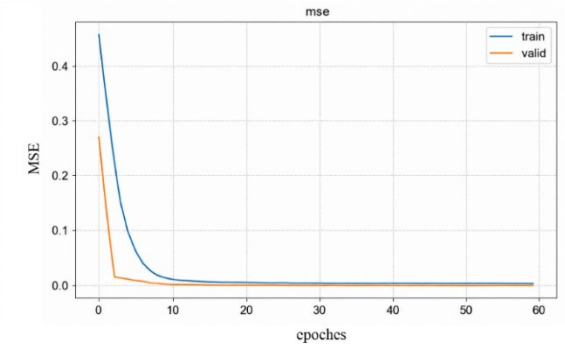


Figure 6: train and validation loss curve of the Conv-LSTM network

4.2 Gaussian RBM Network Results

Since the GRBM network is expected to extract hidden features, no classification layer was applied here either. The input data were fed into the network in flattened form, and the network was trained for 200 epochs. The training loss reached 0.05, and the validation loss reached 0.052. The obtained

results indicate that the network performs satisfactorily. The corresponding performance graph is presented in Figure 7.

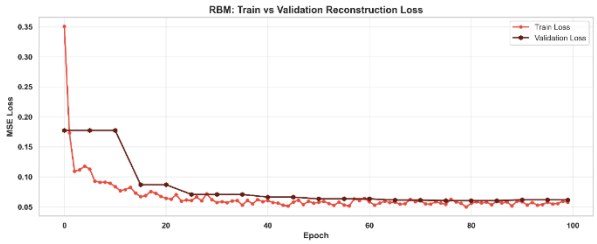


Figure 7: Training and validation loss curve of the GRBM network

4.3 Final network

Finally, the outputs of the previously mentioned networks were used as the input to the final network. The purpose of combining these networks is to extract the most important features in order to achieve an optimal result. The final network includes a single layer with a SoftMax activation function, which performs the classification of the extracted features. The network was trained for 70 epochs with a batch size of 8. The obtained training accuracy was **99.06%**, and the test accuracy was **91.16%**. The training loss reached **0.03**, and the validation loss reached **0.2**.

The graphs in Figures 8 and 9 illustrate the training process of the network. To provide a more thorough analysis of the results, Figure 10 illustrates how the proposed network performed for each emotion in the test phase.

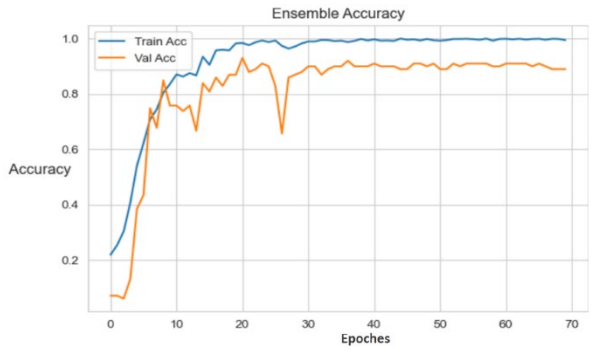


Figure 8: Training and validation accuracy of final network

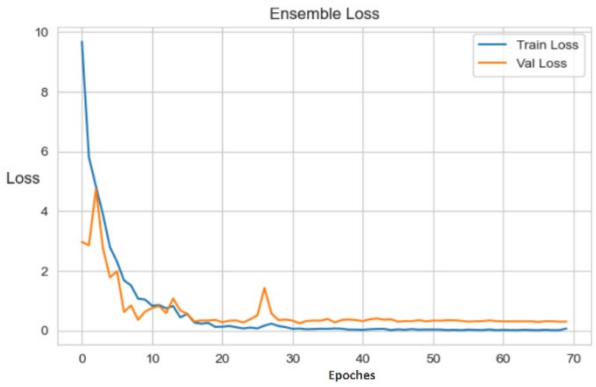


Figure 9: training and validation loss of final network

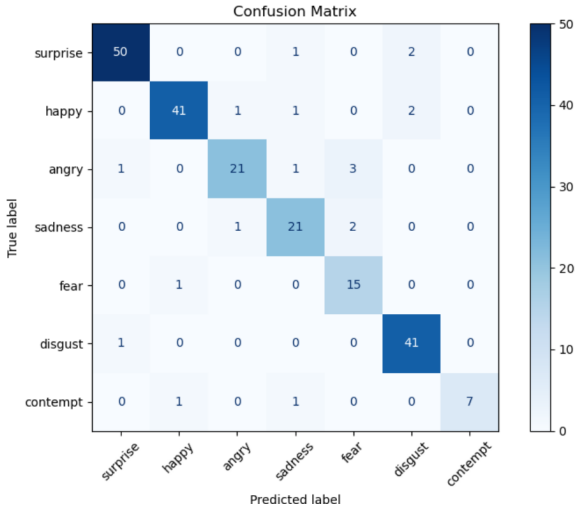


Figure 10: Confusion matrix of the network's tested recognition accuracy

The initial performance evaluation indicated a disparity in recognition accuracy, with emotions such as anger, sadness, and contempt showing lower metrics compared to dominant classes. This behavior is primarily attributed to the inherent class imbalance within the CK+ dataset. To mitigate this, we implemented a class-weighting strategy during the training phase, which assigned higher importance to minority classes, alongside optimizing the training process to 70 epochs with a batch size of 8. As demonstrated in Table 1, these modifications significantly enhanced the discriminability of the model for minority categories, leading to a more balanced and robust performance. While these classes still present a higher degree of complexity, the overall F1-scores reflect the effectiveness of the proposed hybrid architecture in handling spatiotemporal feature representation, even under data scarcity constraints.

Table1: performance metrics of proposed network for 7 emotions.

	Precision	Recall	F1_score
Surprise	96.15%	94.34%	**95.24%**
Happy	95.35%	91.11%	**93.18%**
Angry	91.30%	80.77%	**85.71%**
Sadness	84.00%	87.50%	**85.71%**
Fear	75.00%	93.75%	**83.33%**
Disgust	91.11%	97.62%	**94.25%**
Contempt	100%	77.78%	**87.50%**

In table 2, an overall comparison is made between the proposed model and previous studies. While Yalçın and Alisawi [28] applied EfficientNetB7-CNN, a model known for its large parameter count, to emotion recognition, our proposed framework demonstrates that a lighter, task-specific hybrid architecture (CNN-LSTM+RBM) can achieve higher precision by better capturing the non-linear emotional nuances present in the CK+ dataset. In comparison with the multimodal transformer-based approach proposed in Deep-multi-modal-fusion-transformer-for-emotion-recognition [32], our method achieves highly competitive results using only facial video input. While the transformer model relies on multimodal EEG

and facial information with cross-attention fusion, our hybrid CNN framework reaches 91.16% accuracy on the CK+ dataset, demonstrating strong performance even in a unimodal setting. This indicates that, despite being simpler and more deployable, our model can outperform or match more complex multimodal architectures in practical facial emotion recognition scenarios.

Huang et al. [33] proposed a SE-ResNet-based facial emotion recognition framework for static facial images and reported 83.37% accuracy on RAF-DB after transfer learning from AffectNet. In contrast, our hybrid CNN-LSTM+GRBM model is designed for video-based emotion recognition and achieves 91.16% accuracy on CK+, indicating that temporal modeling and gated feature refinement can provide stronger performance in controlled video settings.

Table 2: Comparison of the proposed model with other models.

Models	Datasets	Refs.	Network	Accuracy
EfficientNetB 7CNN	CK+ +FER24	Nursel Yalçin a, Muthan a Alisawi [28]	EfficientNetB 7+CNN	**78.72%**
Fusion DBN	RML	SHIQING ZHANG 1, XIANZHANG PAN1, YUELI CUI1, XIAOMING ZHAO1, AND LIMEI LIU [17]	CNN+CNN+ DBN	**73.73%**
Transfer learning	Emotional Wearable 202	Haposan Vincentius Manalu, Ach mad Pratama Rifai [21]	CNN+RNN	**61.43%**
Deep multi model fusion	EEG & FE	Qian Zhang, Yifan Liu, Biaokai Zhu, Xun Han, Ruilin Zhang, Jun Xiao, Zhe Wang [32]	Multi model transformer	**81.04%**
Hybrid model	AffectNet, RAF-DB	Zi-Yu Huang 1, Chia-Chin Chiang 1, Jian-Hao Chen 2, Yi-Chian Chen 3*, Hsin-Lung Chung 1, Yu-Ping Cai 4 & Hsiu-Chuan Hsu[33]	ResNet+SE	**83.37%**
This study	CK+	Ours	CNNLSTM+ RBM	**91.16%**

5 Conclusion

Facial expression recognition is one of the noteworthy research areas. Although the limited number of samples in the CK+48 dataset (981 images) presents a challenge, our proposed model addresses this through the use of Gated RBM (GRBM) for robust feature extraction and sophisticated temporal modeling via LSTM, which effectively mitigates the risk of overfitting and enhances generalization. One of the widely used approaches for analyzing these expressions is the application of machine learning techniques, particularly those based on deep learning. Accordingly, in this study, using the Python programming language, an algorithm based on a CNN_LSTM_RBM neural network was presented for emotion recognition through facial expression changes in a video. The proposed model identifies seven emotions: happiness, sadness, disgust, excitement, neutral, fear, and anger. The simulation results demonstrated that the test accuracy of the network on the CK+48 dataset was equal to **91.16%**, and the test loss was **0.2.** In addition, considering the performance metrics for the proposed network, the F1-score accuracy was **89.27%**. All results indicate the effective and successful performance of the proposed method in recognizing facial expressions in video.

Disclosure of Potential Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Reference

[1] D. Joshi *et al.*, “Aesthetics and Emotions in Images,” *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, Sep. 2011, doi: 10.1109/msp.2011.941851.

[2] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended deep neural network for facial emotion recognition,” *Pattern Recognition Letters*, vol. 120, pp. 69–74, Apr. 2019, doi: 10.1016/j.patrec.2019.01.008.

[3] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, “Hybrid deep neural networks for face emotion recognition,” *Pattern Recognition Letters*, vol. 115, pp. 101–106, Nov. 2018, doi: 10.1016/j.patrec.2018.04.010.

[4] M. Nemanja. “Convolutions and convolutional neural networks,” *Introduction to Convolutional Neural Networks* vol. 11, pp. 435-479, Aug. 2020. doi: 10.1007/978-1-4842-5648-0_12.

[5] J.C. Baird, “Information theory and information processing,” *Information processing & management*, vol. 20, pp.373-381, Jun. 1984. doi: 10.1016/0306-4573(84)90068-2.

[6] G. E. Hinton, “Training Products of Experts by Minimizing Contrastive Divergence,” *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002, doi: 10.1162/089976602760128018.

[7] Y. Muneki, and X. Zhongren, “New learning algorithm of gaussian–bernoulli restricted boltzmann machine and its application in feature extraction," *IEICE Proceedings Series*, *76 (A3L-42)*. doi:10.34385/proc.76.A3L-42.

[8] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.

[9] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial

Movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019, doi: 10.1177/1529100619832930.

[10] R. M. D and H. H. Kenchannavar, "Hybrid Deep Optimal Network for Recognizing Emotions Using Facial Expressions at Real Time," *International Journal of Intelligent Systems and Applications*, vol. 16, no. 3, pp. 47–58, Jun. 2024, doi: 10.5815/ijisa.2024.03.04.

[11] M. N. Ab Wahab, A. Nazir, A. T. Zhen Ren, M. H. Mohd Noor, M. F. Akbar, and A. S. A. Mohamed, "Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi," *IEEE Access*, vol. 9, pp. 134065–134080, 2021, doi: 10.1109/access.2021.3113337.

[12] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," *In 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)* Apr. 2013, pp. 1-6. doi: 10.1109/fg.2013.6553717.

[13] I. Lakshan, L. Wickramasinghe, S. Disala, S. Chandrasegar, P. Haddela, "Real time deception detection for criminal investigation," *In2019 National information technology conference (NITC)* Oct. 2019, pp. 90-96. doi: 10.1109/nitc48475.2019.9114422.

[14] H. Karimi, J. Tang , Y. Li, "Toward end-to-end deception detection in videos," *In2018 IEEE international conference on big data (Big Data)* Dec. 2018, pp. 1278-1283. doi: 10.1109/bigdata.2018.8621909.

[15] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic Analysis of Facial Actions: A Survey," *IEEE Transactions on Affective Computing*, pp. 17–25, 2017, doi: 10.1109/taffc.2017.2731763.

[16] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015, doi: 10.1109/tpami.2014.2366127.

[17] S. Zhang, X. Pan, Y. Cui, X. Zhao, and L. Liu, "Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning," *IEEE Access*, vol. 7, pp. 32297–32304, Mar. 2019, doi: 10.1109/access.2019.2901521.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *InProceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497. doi: 10.1109/iccv.2015.510.

[19] B. Hasani, M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 30-40. doi: 10.1109/cvprw.2017.282.

[20] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features With a Hybrid Deep Model for Audio–Visual Emotion Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018, doi: 10.1109/tcsvt.2017.2719043.

[21] H. V. Manalu, and A. P. Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm," *Intelligent systems with applications*, 2024, pp. 330-339. doi: 10.1016/j.iswa.2024.200339.

[22] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial Emotion Recognition with Vision Transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, Aug. 2022, doi: 10.3390/asi5040080.

[23] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *In2010 ieee computer society conference on computer vision and pattern recognition-workshops,* Jun. 2010 pp. 94-101. doi: 10.1109/cvprw.2010.5543262.

[24] J. Zhang, H. Wang, J. Chu, S. Huang, T. Li, and Q. Zhao, "Improved Gaussian–Bernoulli restricted Boltzmann machine for learning discriminative representations," *Knowledge-Based Systems*, vol. 185, p. 104911, Dec. 2019, doi: 10.1016/j.knosys.2019.104911.

[25] O. Prathwini and L. Prathyakshini, "DeepEmoVision: Unveiling Emotion Dynamics in Video Through Deep Learning Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, 2024, doi: 10.14569/ijacsa.2024.0150388.

[26] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," *InNeural Networks: Tricks of the Trade*, Jan 2018, pp. 599-619. doi: 10.1007/978-3-642-35289-8_32.

[27] K. Cho, T. Raiko, A. T. Ihler, "Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines," *InProceedings of the 28th international conference on machine learning (ICML-11),* 2011, pp. 105-112. doi: 10.1162/neco_a_00397.

[28] N. Yalçin, M. Alisawi, "Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques," *Heliyon*. Vol. 30, no. 10, Oct. 2024. doi: 10.1016/j.heliyon.2024.e38913.

[29] J. R. Stevens, R. Venkatesan, S. Dai, B. Khailany and A. Raghunathan, "Softermax: Hardware/software co-design of an efficient softmax for transformers," *In 2021 58th ACM/IEEE Design Automation Conference (DAC)*, Dec. 2021, pp. 469-474. doi: 10.1109/dac18074.2021.9586134.

[30] N. Gupta, R. V. Priya, and C. K. Verma, "Video-based emotion recognition using motion-aware deep hybrid learning," *Cluster Computing*, vol. 29, no. 1, Nov. 2025, doi: 10.1007/s10586-025-05807-x.

[31] B. V. Gokulnath *et al.*, "Empowering emotional intelligence through deep learning techniques," *Scientific Reports*, vol. 16, no. 1, Dec. 2025, doi: 10.1038/s41598-025-29073-4.

[32] Q. Zhang, Y. Liu, B. Zhu, X. Han, R. Zhang, J. Xiao, Z. Wang, "Deep multi-modal fusion transformer for emotion recognition," *Proceeding in Engineering Applications of Artificial Intelligence*. Mar. 2026, pp. 116-129. doi: 10.1016/j.engappai.2026.113967.

[33] Z.Y. Huang, C. C. Chiang, J. H. Chen, "A study on computer vision for facial emotion recognition," *Proceeding in neural networks* Sep. 2023, pp.258-270. doi: 10.1038/s41598-023-35446-4.

[34] M. Luo, "Machine learning for time series analysis and forecasting,", *Master's thesis, Northeastern University*, 2023, pp.62-63. doi: 10.17760/d20487630.