SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

An Intelligent Facial Recognition Framework Using Stacked Autoencoders Integrated With Convolutional Neural Network Architectures

Dr. Sureshkumar Somanathan[1] , Dr M Praneesh[2] , Padmapriya M. R[3] , G.Kowsalya[4] , Dr.K.Sathishkumar[5] , Mrs.E.Pavithra[6] , R. Naveenkumar[7]

[1]Digital Tranformation Leader, Email: suresh.somanathan@gmail.com ORC ID: 0009-0006-3922-7503
[2]Assistant Professor, Department of Computer Science with Data Analytics, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamilnadu, India. Email: raja.praneesh@gmail.com Orcid id : 0000-0003-3691-1343.
[3]Department of CSE (Data Science), Dayananda Sagar College of Engineering, Bangalore, Karnataka, INDIA padmapriya-csds@dayanandasagar.edu https://orcid.org/0009-0002-0904-343X
[4]Independent Researcher, Erode, Tamilnadu, India. Email: kowsiphysics008@gmail.com https://orcid.org/0009-0000-0265-900X
[5]Assistant Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamilnadu, India. Email: sathishmsc.vlp@gmail.com https://orcid.org/0000-0002-7643-4791
[6]Assistant Professor, Department of Computer Science and Engineering(AIML), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai. Mailing address: pavithrae@veltech.edu.in ORCID ID: https://orcid.org/0009-0009-3421-1475?lang=en
[7]Dept of CSE, School of Engineering and Technology, CGC University Mohali-140307, Punjab India. drnk1983@gmail.com 0000-0001-9033-9400

Abstract: Facial Expression Recognition (FER) is an important study in the field of computer vision, which focuses on identifying human emotions through facial images automatically and, therefore, improving human-computer interaction. Nevertheless, the current FER methods are usually characterized by the weakness of reliance on facial landmarks, high-dimensional feature representation, and declining resistance to noise and environmental changes. To counter such issues, this research paper suggests a smart facial recognition algorithm that combines the Convolutional Neural Network (CNN) and Stacked Auto Encoder (SAE) to effectively learn features and classify them. First, the pre-processing of the images is carried out through Gaussian filtering to remove the noise to enhance the quality of the image. Then, a discriminative feature set is formed with the help of Grey Level Co-Occurrence Matrix (GLCM) and feature reduction is performed with the help of Singular Value Decomposition (SVD) in order to reduce redundancy. The refined features will be further learnt by the SAE-CNN architecture to obtain improved representation learning. Lastly, an optimized deep neural network is used to classify data that enhances the accuracy of prediction. Experimental analysis proves that the suggested model performs better than the current methods due to its accuracy, precision, and error rate. SAE + CNN integration greatly increases the feature discriminability and generalization abilities and therefore made the system applicable in real-time FER.

Keyword: Grey Level Co-occurrence Matrix (GLCM), Singular Value Decomposition (SVD), Dimensionality Reduction, Elman Neural Network, Hybrid Deep Learning Model, Feature Optimization, Emotion Classification, Computer Vision, Pattern Recognition, Nonlinear Feature Learning.

1. Introduction

Facial Expression Recognition (FER) is one of the fundamental areas of computer vision and affective computing, the problem of automatic recognition of human states of emotion based on the analysis of a facial image [1]. Facial expressions constitute important non-verbal communication tools that govern inter-personal communications and help to interpret the pattern of behavior. As a result, FER systems are critical in the human-computer interaction, healthcare monitoring, surveillance, and intelligent assistive system applications. Traditional methods of FER usually depend on geometric features as well as appearance-based features and feature extraction methods typically performed manually. And these techniques frequently rely heavily on some set of predefined face landmarks and reference models, subject to manual process and prone to changes due to illumination, pose, and occlusion. This dependency has a strong effect on their strength and scale in real-world applications.

As deep learning is rapidly developing, Convolutional Neural Networks (CNNs) have become an effective FER tool as they can learn hierarchical spatial representations directly with raw image data. CNN-based models do not require manual feature engineering and have a better performance in classification. Nonetheless, they are also effective, but CNNs produce high-dimensional features, which, in many cases, include unnecessary and redundant information, which increases computational complexity and can result in overfitting. Moreover, the current deep learning systems have lesser ability to capture nonlinear associations among variables and have poor ability to generalize and work across different datasets.

The limitations have been overcome by recent research into hybrid architectures, which combine deep learning and feature optimization methods. Nevertheless, most of these techniques do not have effective methods of feature refinement and dimensionality reduction that makes them optimum under noisy and dynamic scenarios. Moreover, the available models tend to show lesser discriminative ability in the case of dealing with subtle and complex emotional expressions [2].

In order to overcome these limitations, the suggested smart facial recognition system will combine Stacked Auto Encoder (SAE) and CNN to improve the features representation and classification performance. CNN element is used in deriving deep spatial features whereas SAE derives nonlinear feature compression and hierarchical representation learning to remove redundancy and enhance discriminability [3]. Also, preprocessing based on the Gaussian filtering is necessary to reduce noise, feature extraction with GLCM with SVD-based dimensionality reduction further optimizes feature space. Refined features are then categorised with the help of an optimised deep neural network which results in better accuracy and lower error rates. This combined framework is a powerful, scalable, and efficient solution to FER, which can deal with the variability of the real world and perform better than current solutions.

2. Related Works

As an established field of research in computer vision and affective computing, Facial Expression Recognition (FER) is intended to provide intelligent systems with the ability to automatically analyze human feelings in facial images. Early FER methods were mainly based on manual feature extraction models like geometric features and appearance models that are inherently sensitive to illumination changes, pose changes and noise. The FER performance has also been enhanced to a considerable extent with the emergence of deep learning and specifically the implementation of Convolutional Neural Networks (CNNs) which enable hierarchical learning of spatial features on raw images. Nevertheless, even with such developments, current deep models tend to be unable to represent high-dimensional features, handle redundancy or lack sufficient discriminative ability in dealing with fine-grained emotional patterns and variability in the real world.

Table 1. Comprehensive Analysis of Literature

Author(s) & Year	Methodology / Model	Key Components	Core Contribution	Limitations
Ko et al. (2018) [4]	Traditional vs Deep Learning FER	Feature extraction, classification; CNN + LSTM	Categorized FER approaches; hybrid CNN-LSTM captures spatial-temporal dependencies	Traditional methods lack end-to-end learning capability
Wu et al. (2018) [5]	Adaptive Feature Mapping (AFM)	Weighted-centre regression, feature distribution alignment	Improved generalization by mapping test features to training distribution	Neutral expressions may be misclassified due to wide feature distribution
Majumder et al. (2018) [6]	Autoencoder-based Feature Fusion	Geometric + appearance features, SOM classifier	Learned correlation between heterogeneous features using autoencoders	Conventional fusion techniques may still suffer from domain imbalance
Han Y et al. (2021) [7]	Hybrid Neural Network (HNN)	Latent Semantic Machines, Transfer Learning, Sparse Encoding	Enhanced interpretability and semantic feature learning for emotion classification	Increased model complexity

Li et al. (2017) [8]	Deep Fusion CNN (DF-CNN)	Geometry maps, normal maps, texture, curvature + CNN + SVM	High-dimensional facial attribute representation for improved FER	Requires complex 3D facial data processing
Derkach et al. (2017) [9]	Spectral Representation Model	Graph Laplacian, Fourier-like surface representation	Captures geometric and topological properties effectively	Computational complexity in spectral decomposition
General Observation	Hybrid & Feature-based Methods	Preprocessing, feature selection, classification	Emphasizes importance of noise reduction and dimensionality handling	Performance degradation on large-scale datasets

Critical evaluation of current literature shows that there are a number of research gaps that have existed in the literature that impede the strength and expansion of FER systems. The majority of the traditional and deep learning-based methods do not offer an effective balance between feature richness and compactness leading to a high level of computational complexity and overfitting. Moreover, the poor feature optimization and inability to obtain nonlinear correlations on features lower classification accuracy. The hybrid models that have been suggested in the previous literature also tend to lack an effective feature refinement and dimensionality reduction mechanism, in addition to being sensitive to noise and environmental change. In addition, most models have been shown to perform reasonably well on confined datasets and fail to generalize in large-scale and diversified real-life situations.

To overcome the constraints, the intelligent facial recognition system is proposed based on a Stacked Auto Encoder (SAE) and CNN to obtain improved feature representation and classification quality. CNN element is used in deep hierarchical features extraction, which obtains intricate spatial features of the face, SAE conducts nonlinear feature compression and trimming to remove redundancy and maximize discriminability. It is a hybrid design that can perform dimensionality reduction effectively, better noise and variation resistance, and improved dataset-wide generalization. In the proposed approach, which combines unsupervised features learning with the help of SAE and supervised learning with CNN, a good end-to-end approach is proposed that removes the existing weaknesses and provides a superior performance in FER tasks.

3. Proposed Methodology

The presented intelligent facial recognition system is mathematically represented as a multi-stage transformation pipeline comprising of preprocessing, statistical features extraction, dimensionality reduction, deep representation learning, and optimized classification. To represent the input facial image, $I(a,b) \in \mathbb{R}^{M\times N}$ where a and b are spatial coordinates. The first process is preprocessing to reduce noise and increase structural details with the help of Gaussian filtering. The convolution of the image yields the smooth image as

$$I_s(a,b) = G_\sigma(a,b) * I(a,b) \qquad (1)$$

In which the Gaussian kernel is defined as

$$G_\sigma(a,b) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{a^2+b^2}{2\sigma^2}\right) \qquad (2)$$

In this case, σ is the standard deviation that regulates the effect of smoothing. The gradient operator is used on the smoothed image to capture the structural transitions.

$$\nabla I_s(a,b) = \nabla(G_\sigma(a,b) * I(a,b)) \qquad (3)$$

The boundary representation is calculated with the following formula:

$$M_Y(a,b) = \nabla(-G_\sigma(a,b) * I(a,b)) \qquad (4)$$

which is normalized to have a representation that is scale-invariant.

$$M_{NY}(a,b) = \frac{M_Y(a,b) - min(M_Y)}{max(M_Y) - min(M_Y)} \qquad (5)$$

A threshold $T \in [0,1]$ is then used to generate a binary boundary map

$$M_{YY}(a,b) = \begin{cases} 1, & M_{NY}(a,b) \geq T \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

After the preprocessing, the extraction of the texture-based features is carried out using the Grey Level Co-Occurrence Matrix (GLCM) which describes the statistical relationships between pixel values in the second order. Let p(l,m)represent the normalised co-occurrence likelihood of the gray level land m. Contrast is calculated as

$$\text{Contrast} = \sum_{l,m} (l-m)^2 p(l,m) \qquad (7)$$

Energy, the textural uniformity, is defined as

$$\text{Energy} = \sum_{l,m} p(l,m)^2 \qquad (8)$$

The entropy, the measure of randomness, is given as

$$\text{Entropy} = -\sum_{l,m} p(l,m)\, log_2\, p(l,m) \qquad (9)$$

The linear dependency or correlation is calculated as

$$\text{Correlation} = \sum_{l,m} \frac{(l-\mu_l)(m-\mu_m)}{\sigma_l \sigma_m} p(l,m) \qquad (10)$$

These attributes are combined to construct a feature vector $F \in \mathbb{R}^d$ and this feature may be full of redundant information and high dimensional information. To overcome this, Singular Value Decomposition (SVD) is implemented to reduce the dimensions. With a feature matrix D, the low-rank approximation of D is given by

$$D^{(k)} = \sum_{i=1}^{k} \sigma_i\, u_i v_i^T \qquad (11)$$

where $\sigma_i$are singular values, and u_i, $v_i$are orthogonal vectors. The decomposition satisfies

$$D^T D = V \Sigma^2 V^T \qquad (12)$$

and the left singular vectors are calculated to be

$$U = DV\Sigma^{-1} \qquad (13)$$

The low feature representation is then fed through a Stacked Auto Encoder (SAE) to learn nonlinear features. Encoding is determined as

$$h^{(1)} = f(W^{(1)}F + b^{(1)}) \qquad (14)$$

$$h^{(l)} = f(W^{(l)}h^{(l-1)} + b^{(l)}) \qquad (15)$$

where $h^{(l)}$denotes the latent representation at layer l, $W^{(l)}$and $b^{(l)}$are weights and biases, and $f(\cdot)$is a nonlinear activation function. As a result of the decoding process, the input is reassembled as

$$\hat{F} = g(W_d h^{(l)} + b_d) \qquad (16)$$

The minimization of the reconstruction loss is used to train the SAE

$$L = \| F - \hat{F} \|^2 \qquad (17)$$

which guarantees sparse and discriminatory feature presentation. The obtained learned features are then further refined with the aid of a Convolutional Neural Network (CNN) in which convolutional feature maps are calculated as

$$X^{(k)} = \sigma(W^{(k)} * X^{(k-1)} + b^{(k)}) \qquad (18)$$

where $*$denotes convolution and $\sigma(\cdot)$is an activation function such as ReLU.

Lastly, it is classified under an Optimized Elman Deep Neural Network (OELDNN) that uses context memory in enhancing learning. The activation of the hidden layer is represented by

$$H_j = f\left(\sum_{i=1}^{N_1} w_{ji}^H\, OI_i - \theta_j^H\right) \qquad (19)$$

and the output is calculated as

$$O_s = g\left(\sum_{j=1}^{N_2} w_{sj}^O\, H_j - \theta_s^O\right) \qquad (20)$$

Minimization of the classification error is the minimization objective

$$min\, F_n = \sum_{e=1}^{E} \sum_{s=1}^{N_3} (T\, O_{es} - OO_{es})^2 \qquad (21)$$

where $TO_{es}$and OO_{es}denote target and observed outputs. The weight update mechanism is defined as

$$w = \sum cons(R - \frac{1}{2}) \qquad (22)$$

convergent nonlinear criterion. The state of context is maintained by

$$C^{(t)} = H^{(t-1)} \qquad (23)$$

which provides the ability to model the temporal dependency and enhances the stability of the classification.

Therefore, the suggested framework combines statistical feature modeling, dimensionality reduction, deep representation learning, and optimized recurrent classification in a single mathematical framework. This holistic method increases discriminability of features, decreases computational complexity and increases the accuracy of classification thus it is very relevant in the creation of robust and scalable facial expression recognition systems.

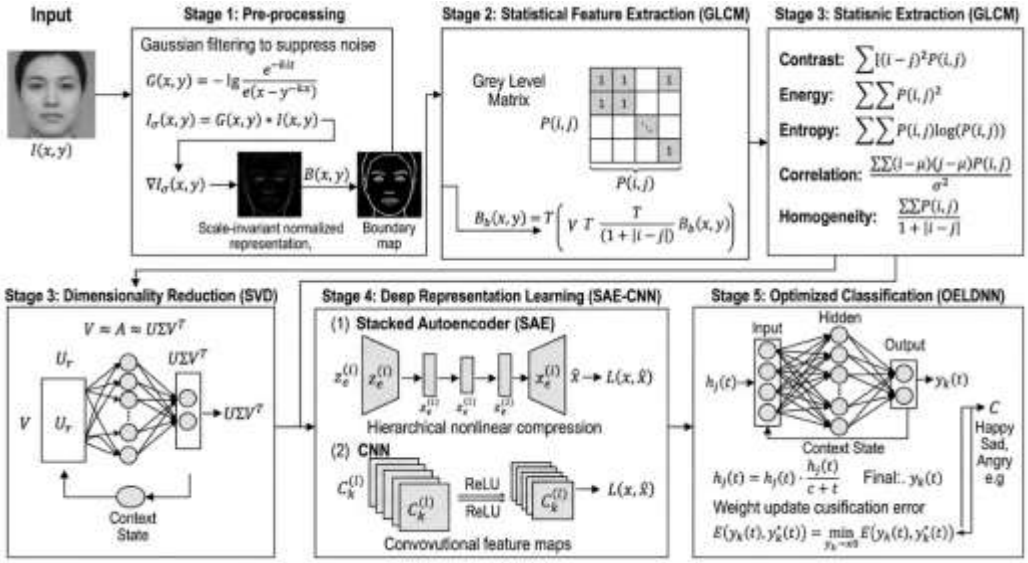

Figure 1. Architecture of optimized EDNN

Algorithm 1. Optimized Elman Deep Neural Network for Classification

Input:
Feature vector F, Weight matrices W, Bias vectors θ, Learning rate η, Context state C
Output:
Class label $CR \in$ {Emotion Classes}

Step 1: Initialize network parameters
Initialize input layer $OI = \{OI_1, OI_2, \ldots, OI_{N_1}\}$
Initialize hidden layer $H = \{H_1, H_2, \ldots, H_{N_2}\}$
Initialize output layer $O = \{O_1, O_2, \ldots, O_{N_3}\}$
Initialize context layer $C = \{C_1, C_2, \ldots, C_{N_2}\}$
Initialize weights W^H, W^O and biases θ^H, θ^O
Step 2: Set context state

$$C^{(0)} = 0$$

Step 3: For each training sample $F^{(e)}$, perform forward propagation
Step 3.1: Assign input features

$$OI_i^{(e)} = F_i^{(e)}, i = 1,2,\ldots,N_1$$

Step 3.2: Compute hidden layer input with context feedback

$$Input(H_j^{(e)}) = \sum_{i=1}^{N_1} w_{ji}^H \; OI_i^{(e)} + \sum_{k=1}^{N_2} w_{jk}^C \; C_k^{(e)} - \theta_j^H$$

Step 3.3: Apply activation function

$$H_j^{(e)} = f(Input(H_j^{(e)})), j = 1,2,\ldots,N_2$$

Step 3.4: Update context layer

$$C_j^{(e)} = H_j^{(e-1)}$$

Step 4: Compute output layer
Step 4.1: Output neuron input

$$Input(O_s^{(e)}) = \sum_{j=1}^{N_2} w_{sj}^{O}\ H_j^{(e)} - \theta_s^{O}$$

Step 4.2: Final output

$$O_s^{(e)} = g(Input(O_s^{(e)})), s = 1,2,\ldots,N_3$$

Step 5: Error computation

$$E = \sum_{e=1}^{E} \sum_{s=1}^{N_3} (T\ O_{es} - O_s^{(e)})^2$$

Step 6: Weight optimization (nonlinear constrained update)

$$w \leftarrow w + \eta \cdot cons \cdot (R - \frac{1}{2})$$

Update weights:

$$w_{ji}^{H} \leftarrow w_{ji}^{H} - \eta \frac{\partial E}{\partial w_{ji}^{H}}$$

$$w_{sj}^{O} \leftarrow w_{sj}^{O} - \eta \frac{\partial E}{\partial w_{sj}^{O}}$$

Step 7: Bias update

$$\theta_j^{H} \leftarrow \theta_j^{H} - \eta \frac{\partial E}{\partial \theta_j^{H}}$$

$$\theta_s^{O} \leftarrow \theta_s^{O} - \eta \frac{\partial E}{\partial \theta_s^{O}}$$

Step 8: Repeat Steps 3–7 until convergence
Stopping condition: $E < \epsilon$or maximum iterations reached
Step 9: Classification decision

$$CR = \begin{cases} \text{Class}_1, & 0 \le O_s < 0.4 \\ \text{Class}_2, & 0.4 \le O_s < 0.7 \\ \text{Class}_3, & O_s \ge 0.7 \end{cases}$$

Step 10: Return final classified output CR

4. Result and Discussion

Bosphorus database contains information about 105 subjects with various expressions. The 3D facial scans were acquired under good illumination conditions. It is mainly used for the recognition of faces and facial expressions. The facial data are self-occluded. The contents of Bosphorus Database are as follows. has 3D data packed as bnt file, 2D colour images of high resolution (cropped face regions) as png file, information about 24 2D and 3D coordinates of facial landmarks as lm2 and lm3 files respectively. Facial landmarks were manually annotated on 2D images if they were visible on the given facial scan and then using 3D – 2D correspondences, the 3D landmark co-ordinates were calculated. Figure 2 shows the landmarks that are annotated in a 3D facial scan.

There are 105 subjects in total, in which 60 are men and 45 are women. The race of most of the subjects is Caucasian. 29 actors and actress are used to obtain more realistic expression. Most of the subjects are in the age group of 25 to 35. Total Number of facial scans is 4666. For each subject, 54 facial scans are available. But, some subjects may have less scans due to fewer number of expressions.

Range images are the images with pixels that represent the distance. From the depth information available in the 3D facial scan, range image is constructed to represent the intensity variation based on depth. The depth value given by the z-coordinates of the 3D facial surface image are sorted and assigned intensity values to construct the range image. Range images representing all the seven facial expressions are given in Figure 3.

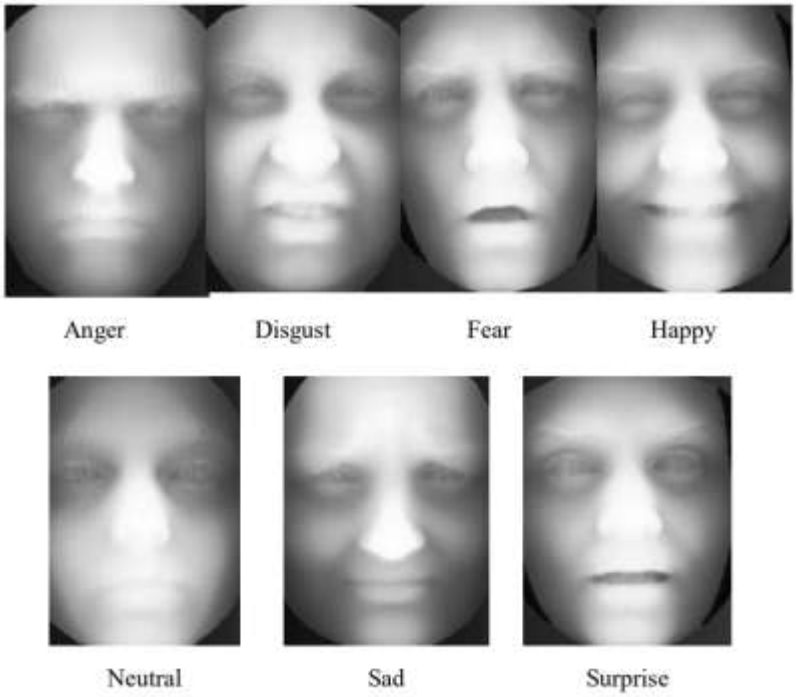

Figure 2. Diverse Image Ranges

The first measure of evaluation in classification systems is accuracy which measures the percentage of correctly predicted cases in the total samples. It is mathematically determined as the number of true positives/true negatives to the total population. Accuracy represents the general correctness of the model but can be misleading in unbalanced datasets in which the distributions of the classes are skewed. Hence, although high accuracy is considered to reflect good predictive ability, it must be coupled with various indexes like precision, recall, and F1-score to ensure good performance evaluation. The rate of accuracy is given n Figure 3 and Table 1. It's calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ ---------(21)}$$

Table 1. Comparison of Accuracy

Expression	Existing				Proposed
	CNN	LSTM	HNN	DFCNN	OELDNN
Anger	81	85	83	87	93
Disgust	83	87	86	89	95
Happy	84	88	88	90	96
Sad	85	89	89	91	97
Surprise	81.5	84	84	86	94
Fear	82	86	85	88	95
Neutral	85	88.5	89.5	91	97

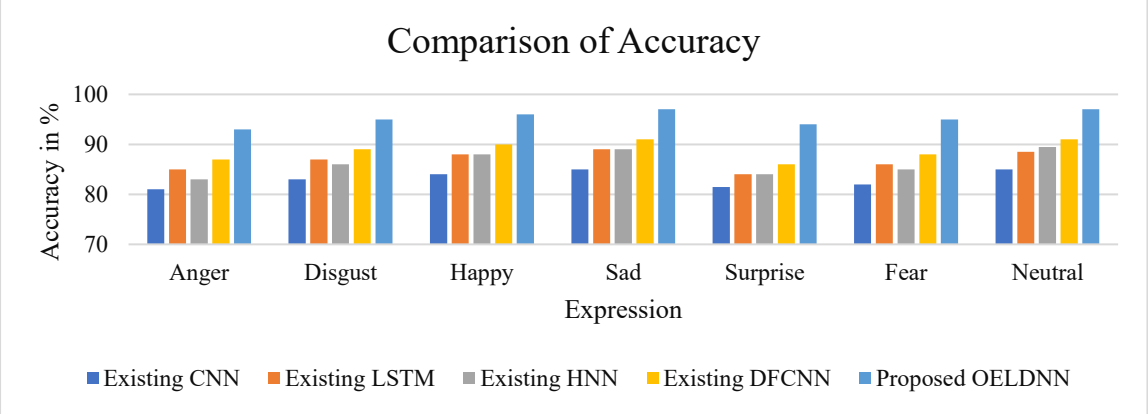

Figure 3. Comparison of Accuracy

Precision: The positive quantitative value or precision indicates the proximity of the observation and the relevance among the values discovered. The percentage rate of p The terms precision and accuracy are synonymous. It is calculated using True Positive (TP) and False Positive (FP) rates. The fraction of positive values in the entire population is related to precision. The accuracy estimate for a given issue in the classification stage is the number of real positive qualities (i.e. the count of the item correctly labelled as positive classes). The rate of precision is given in Table 2 and Figure 4. As a result, the algorithm with high precision creates more required data than unnecessary data.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \text{ ---------(22)}$$

Table 2. Comparison of Precision

Iteration Count	Existing				Proposed
	CNN	LSTM	HNN	DFCNN	OELDNN
Anger	83	85	83	86	93
Disgust	83.5	86	86	89	95
Happy	84	88	88	90	96
Sad	85	89	89	93	97.5
Surprise	83.5	84	84	86	94
Fear	83	83.4	86	85	85
Neutral	88	88.5	89.5	93	96

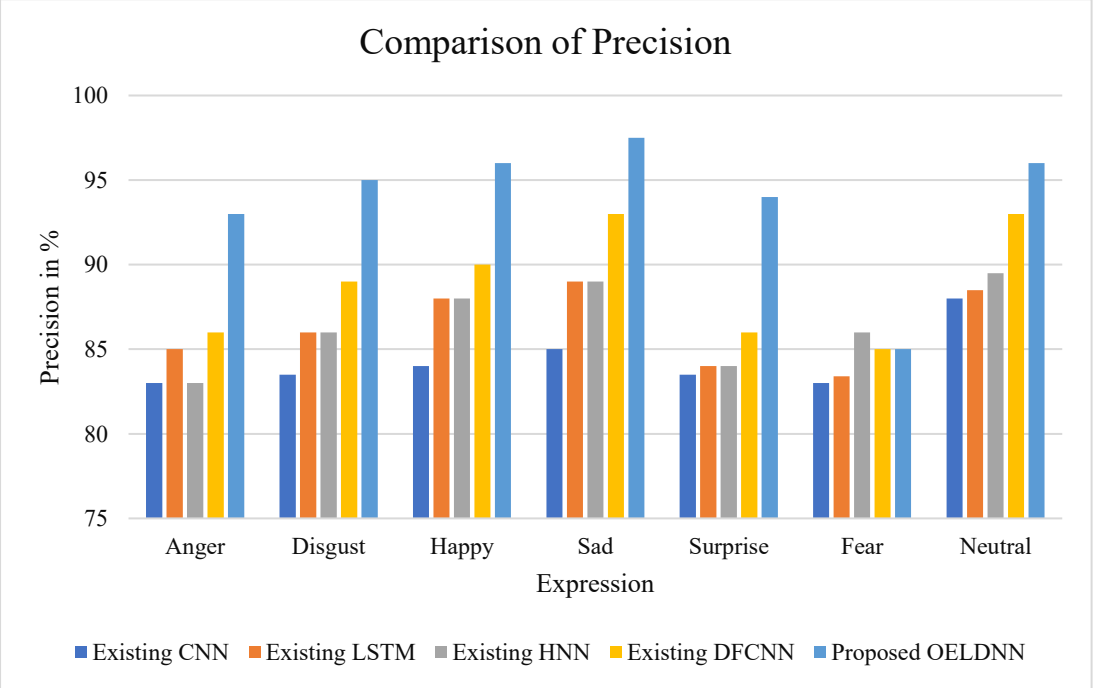

Figure 4. Comparison of Precision

Error Rate: Errors in digitalisation are caused by a variety of factors, including noise, distortion, and data processing disturbances. It's a ratio of competency rates. The rate of error is given in Figure 5 and Table 3.

Table 3. Comparison of Error Rate

Iteration Count	Existing				Proposed
	CNN	LSTM	HNN	DFCNN	OELDNN
Anger	0.578	0.561	0.567	0.551	0.498
Disgust	0.598	0.569	0.578	0.567	0.512
Happy	0.612	0.571	0.591	0.571	0.534
Sad	0.623	0.574	0.623	0.581	0.547
Surprise	0.561	0.541	0.557	0.561	0.488
Fear	0.588	0.549	0.558	0.577	0.492

Neutral	0.601	0.589	0.573	0.595	0.499

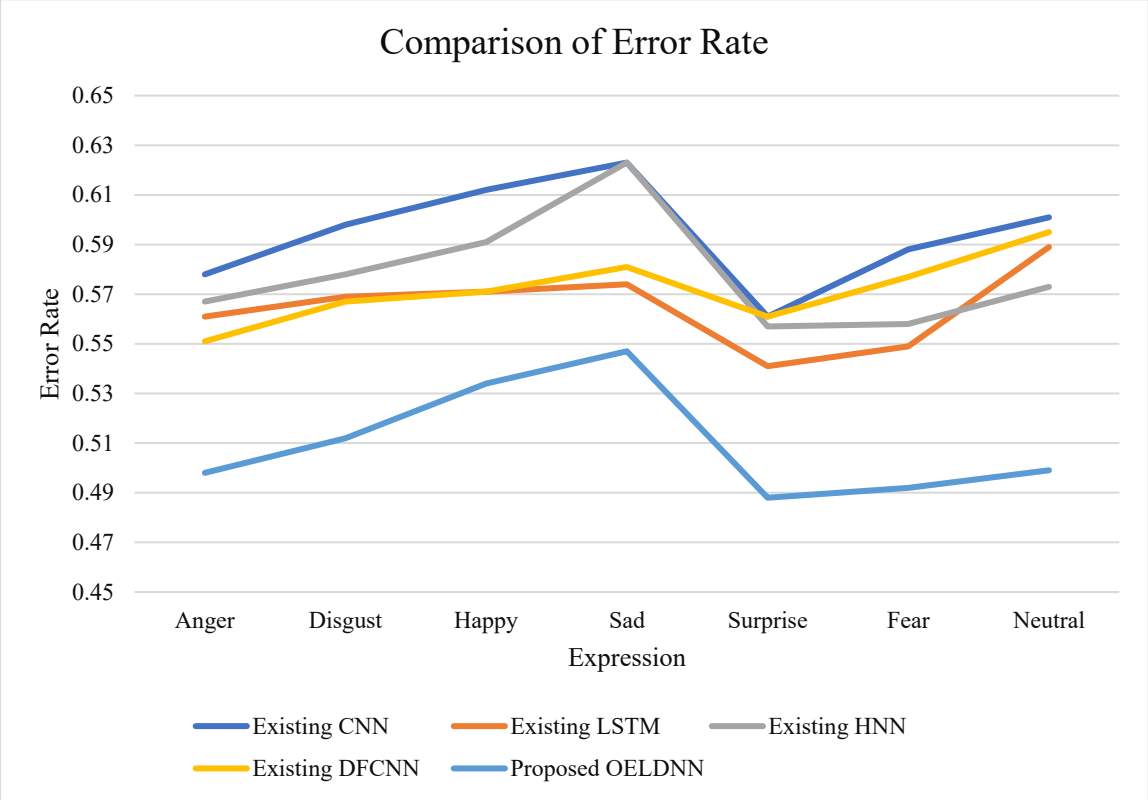

Figure 5. Comparison of Error Rate

Figure 6 indicates comparison of error rate for three different expressions. The proposed approach is compared with existing techniques namely CNN, LSTM, HNN, and DFCNN. The ISAE attains minimal error rate over other existing approaches that indicates the effectiveness of the proposed approach.

5. Conclusion

The concept of Facial Expression Recognition is also important in providing intelligent systems with an opportunity to comprehend the human emotions and improve interaction between the human beings and the machines. Intelligent FER system is proposed in this research to overcome the drawbacks of the traditional and the existing deep learning methods using an intelligent system that combines Stacked Auto Encoder and Convolutional Neural Network. The suggested framework operates effective preprocessing through the use of a Gaussian filtering technique, feature extraction through the adoption of the GLCM and dimensionality reduction through the use of SVD to build an optimized feature space. The union between SAE promotes nonlinear compression of features and the discriminative power of the obtained features, whereas CNN guarantees strong spatial feature learning. An optimized deep neural network is used to carry out the final classification and reduce the error of classification as well as enhancing predictive performance.

Experimental evidence shows that the method proposed performs better in terms of accuracy, precision, and fewer errors as compared to other currently used models like CNN, LSTM, HNN, and DFCNN. Another advantage is that the model is more resilient to noise and environmental changes, as well as offers greater generalization across datasets. Therefore, the suggested SAE-CNN architecture is an effective and scalable solution to the real-time facial expression recognition application and can be extended to more sophisticated intelligent systems.

Reference

1. Srinu, B., & Dharnasi, P. (2026). Smart Attendence System Using Facial Recognition for Staff using AI/ML. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 9(2), 513-519.

2. Zhou, D., Shao, F., Zhang, J., Zhang, Y., & Feng, S. (2026). Multi-scale orthogonal Gabor filters based ConvNets for illumination robust single sample face recognition. Engineering Applications of Artificial Intelligence, 163, 113143.
3. Tharini, V. J., & Shivakumar, B. L. (2024). A Canonical Particle Swarm Optimization (C-PSO) Approach to Identify High Utility Itemset. Journal of Computational Analysis & Applications, 33(5).
4. Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. sensors, 18(2), 401.
5. Wu, B. F., & Lin, C. H. (2018). Adaptive feature mapping for customizing deep learning based facial expression recognition model. IEEE access, 6, 12451-12461.
6. Majumder, A., Behera, L., & Subramanian, V. K. (2016). Automatic facial expression recognition system using deep network-based data fusion. IEEE transactions on cybernetics, 48(1), 103-114.
7. Han, Y., Wang, X., & Lu, Z. (2021). Research on facial expression recognition based on Multimodal data fusion and neural network. arXiv preprint arXiv:2109.12724.
8. Li, H., Sun, J., Xu, Z., & Chen, L. (2017). Multimodal 2D+ 3D facial expression recognition with deep fusion convolutional neural network. IEEE Transactions on Multimedia, 19(12), 2816-2831.
9. Derkach, D., & Sukno, F. M. (2017, May). Local shape spectrum analysis for 3D facial expression recognition. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 41-47). IEEE.